

Figure 5: **Qualitative visualization.** Concerto performs well across different point cloud inputs: a complete scene (top two rows) and an incomplete scene (bottom two rows).

309 Appendix

310 Concerto is a superior spatial representation point encoder capable of handling a wide range of scene
 311 types, including those with varying completeness in Fig. 5, video-lifted point clouds in Fig. 6, and
 312 the large scene in Fig. 7. Here, we further present the detailed implementation and results.

313 A Additional Implementation

314 We adopt the detailed parameters from Sonata [47] for intra-modal self-distillation and refer readers
 315 to the original Sonata paper for an in-depth description of its implementation. In this section, we
 316 provide a thorough explanation of the implementation for cross-modal joint embedding prediction.

317 A.1 Combination of Intra-Modal and Cross-Modal Learning

318 As in Sonata, we use 4 local views, 2 masked views, and 2 global views, with the first global
 319 view serving as the principal view. For cross-modal joint embedding prediction, we utilize the
 320 representations from the first masked view (based on the principal view) to predict the corresponding
 321 image representations. The cross-modal cosine similarity loss is computed at upcast level 3, while
 322 the online clustering cross-entropy loss for intra-modal self-distillation is calculated at upcast level 2.

323 A.2 Correspondence Between Pixels and Points

324 To establish reliable 3D point to 2D pixel correspondences across camera views, we employ a two-step
 325 approach: 3D-to-2D projection followed by depth-based visibility verification.

326 Let $\mathbf{p} = (X, Y, Z)^T$ denote a 3D point in world coordinates. Each camera c is defined by intrinsic
 327 matrix \mathbf{K} and extrinsic matrix $[\mathbf{R}|\mathbf{t}]$. The standard pinhole camera model projects the 3D point \mathbf{p} to
 328 2D pixel coordinates (x, y) and a projected depth d_{proj} :

$$d_{\text{proj}} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \quad (1)$$

329 To account for occlusions, we perform a visibility check comparing d_{proj} with the depth value
 330 $d_c = \mathbf{D}_c(x, y)$ retrieved from camera c 's depth map \mathbf{D}_c at the projected pixel coordinate (x, y) . The

Dataset	Source	Train	Val	Test	All
ScanNet [15]	real	25,472	7,092	2,802	35,366
ScanNet++ [50]	real	48,493	1,534	1,161	51,188
S3DIS [1]	real	10,977	3,668	0	14,645
ARKitScenes [7]	real	68,991	9,350	0	78,341
HM3D [31]	real	32,272	4,116	0	36,388
Structured3D [57]	synthesis	63,905	6,683	6,391	76,979
RE10K [58]	real	171,356	0	18,932	190,288
Concerto (ours)	mixed	421,466	32,443	29,286	483,195

Table 6: Image Data Source Collection.

Dataset	Source	Train	Val	Test	All
ScanNet [15]	real	1,201	312	100	1,613
ScanNet++ [50]	real	856	50	50	956
S3DIS [1]	real	204	68	0	272
ARKitScenes [7]	real	4,498	549	0	5,047
HM3D [31]	real	8,881	1,119	0	10,000
Structured3D [57]	synthesis	18,348	1,776	1,697	21,821
RE10K [58]	real	42,839	0	4,733	47,572
Concerto (ours)	mixed	76,827	3,874	6,580	87,281

Table 7: Point Cloud Data Source Collection.

point is considered visible if:

$$|d_c - d_{\text{proj}}| < \epsilon_{\text{depth}}, \quad (2)$$

where ϵ_{depth} is set to 0.01 in our experiments. Additionally, the correspondence is rejected if (x, y) falls outside image bounds or $D_c(x, y)$ contains invalid depth. This visibility check establishes a mapping between 3D points and corresponding 2D pixels, enabling direct correspondence between 3D points and ViT patches for cross-model joint embedding prediction mechanisms. Depending on the dataset, the depth map D_c is obtained in different ways:

- **RGBD datasets.** Depth maps are directly available as the depth channel of RGBD images, such as Structured3D [57].
- **Known ground truth mesh.** For datasets like ScanNet [15], ScanNet++ [50], S3DIS [1], and ARKitScenes [7], depth maps are rendered from the ground truth 3D mesh using camera parameters.
- **Pixel-aligned point clouds.** For video-lifted point clouds (e.g., using VGGT [41] on RealEstate10K [58]), per-view depth maps D_i are generated alongside point clouds \mathcal{P}_i . A point $p \in \mathcal{P}_i$ from camera i can be visible from camera j if it passes the visibility check.

For HM3D [31], which does not provide the raw images, we leverage Habitat-Sim [29] to simulate the scenes. For each navigatable room, we capture four images around the room with random initial camera orientations. The angular difference between consecutive images is 90 degrees. We record the camera parameters to compute the correspondence between points and pixels, as described previously. The total collections of our training data are shown in Tab. 6 and Tab. 7.

A.3 Image Augmentations

We implement the same point cloud augmentations as Sonata. For image augmentations, we initially adopt the process from DINOv2 [26], excluding geometric augmentations to simplify the alignment between pixels and points. Specifically, we apply color jittering, random grayscale, and Gaussian blur to the images, consistent with the settings used in DINOv2. This results in a slight drop in the mIoU on ScanNet semantic segmentation to 75.27%, compared to using the original images. Consequently, we continue to explore more suitable image augmentations. In the ablation study, we apply random color jittering, with the same intensity as the point cloud augmentations, along with Gaussian blur. This weaker augmentation improves Concerto’s performance, which is expected since the image encoder is currently frozen. Stronger augmentations may yield better results once both the image and point branches are unlocked for joint learning. Given the variability in image augmentation

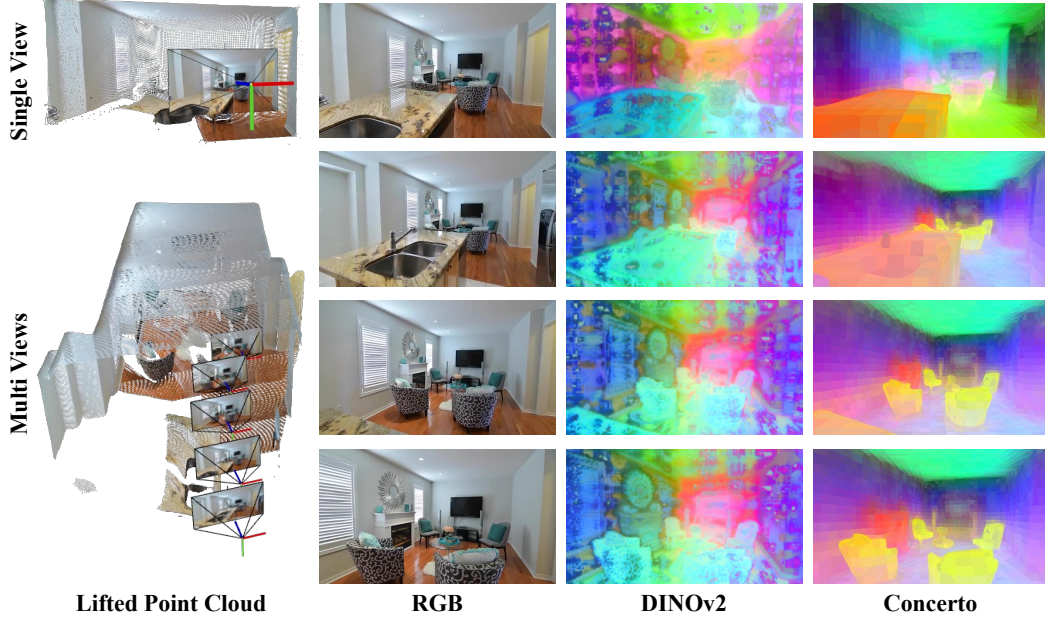


Figure 6: **Video perception.** Concerto can be applied to single-view (top row) and multi-view video-lifted data (bottom three rows). We visualize the PCA of one video in RE10K [58]. In the multi-view setting, the representations from all the frames are computed together for consistency.

combinations, we currently refrain from applying additional image augmentations to our pipeline, but this remains under investigation and may be updated in future iterations.

A.4 Experimental Setting

Software and hardware environment.

- CUDA version: 12.4
- PyTorch version: 2.4.1
- Python version: 3.10.15
- GPU: Nvidia H20 \times 16 for pretraining; Nvidia H20 \times 8 for evaluation.
- CPU: \times 360 for pretraining; \times 180 for evaluation.
- Memory: 3600GB for pretraining; 1800GB for evaluation.
- Time: 97h for pretraining; evaluation time is based on datasets and evaluation protocols.

Data license. We use the open-source datasets ScanNet [15], ScanNet++ [50], S3DIS [1], Structured3D [57], ARKitScenes [7], Habitat Matterport3D [31] and RealEstate10K [58] in latest versions. S3DIS, Structured3D, ScanNet, and ScanNet++ have custom licenses. RealEstate10K is licensed by Google LLC under a Creative Commons Attribution 4.0 International License. ARKitScenes is licensed by Apple Inc. HM3D is licensed by Matterport.

Training details. For pretraining, we leverage all train, val, and test splits to train the self-supervised model. For evaluation with linear probing, decoder probing, and full fine-tuning, we train on the train split and test on the val split of ScanNet, ScanNet++, ScanNet200, and Area 5 of S3DIS. We use AdamW as the optimizer, and cosine annealing policy as the scheduler. The learning rate is adjusted with the encoder depth and the max one is 0.004. The pertaining epoch is 100. For cross-modal joint embedding prediction, we set DINOv2 image encoder input resolution 518×518 and leverage DINOv2 L version currently. In the future, we will update to DINOv2.5 G version.

B Additional Results

B.1 Concerto with Video-Lifted Point Clouds

We utilize the current feed-forward reconstruction model VGGT [41] to lift RealEstate10K [58] video data to point cloud. Based on the camera poses, we heuristically select video clips with larger camera

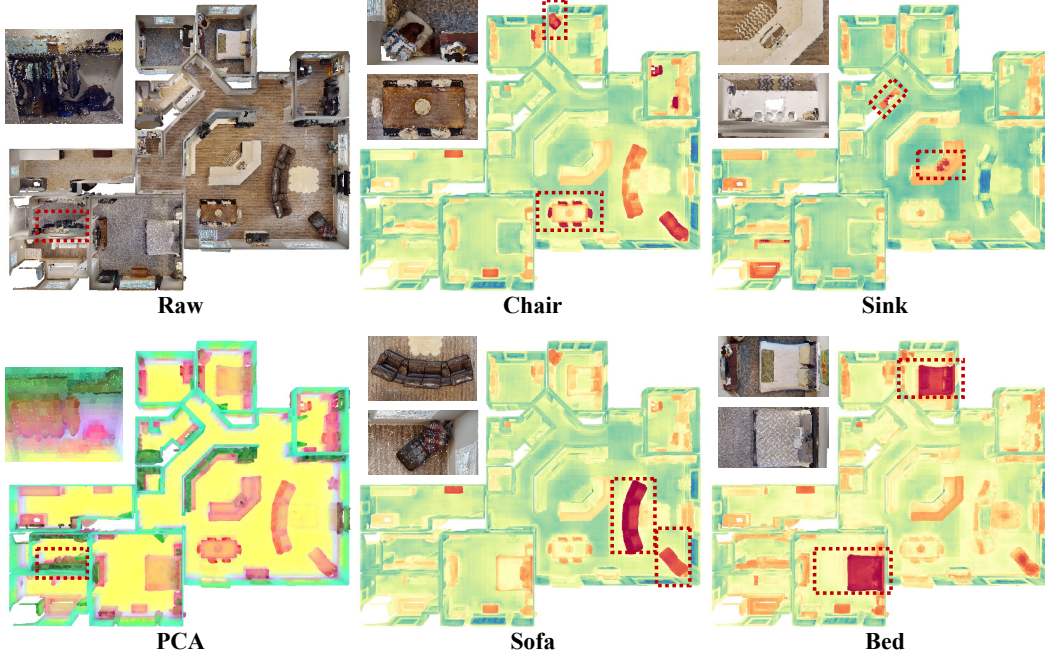


Figure 7: **Language locate.** We visualize the PCA of a large house scene from HM3D [31] along with the heatmap of zero-shot language-based object localization results. The upper-left part of the scene shows detailed local information. Given specific words, Concerto with text-aligned linear probing successfully locates objects in a zero-shot setting.

Model	ScanNet Val			ScanNet200 Val			ScanNet++ Val			S3DIS Area 5		
	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
img. enc.												
DINOv2 (lin.)	77.3	86.6	91.7	37.4	49.5	83.3	45.7	60.5	86.5	73.5	81.3	90.9
SigLIP2 (lin.)	76.3	86.0	91.4	36.7	48.9	82.7	45.8	61.4	86.8	72.3	79.4	91.0
RADIO (lin.)	73.5	84.0	90.3	31.0	42.3	81.8	42.7	57.2	85.3	72.9	80.5	90.9
DINOv2 (dec.)	79.5	87.6	92.6	37.8	50.5	84.1	48.3	62.3	87.7	75.5	84.2	92.3
SigLIP2 (dec.)	78.8	87.0	92.4	37.5	47.7	83.7	46.8	58.1	87.0	73.6	79.8	91.3
RADIO (dec.)	77.9	85.7	92.3	33.9	44.6	83.4	44.9	56.5	86.2	74.8	81.2	92.2
DINOv2 (full.)	80.7	87.4	93.1	39.2	50.2	85.0	50.7	63.3	87.9	77.4	85.0	93.2
SigLIP2 (full.)	79.7	86.9	92.7	38.4	49.9	83.9	50.0	62.0	88.2	75.0	80.2	92.5
RADIO (full.)	79.6	86.6	92.7	36.1	46.9	83.8	48.4	60.6	88.2	75.1	80.5	92.8

Table 8: **Segmentic segmentation of Concerto with different image encoders.** Concerto with DINOv2 based on self-distillation has the best performance in general.

pose transforms in comparison and abandon those with smaller camera pose transforms. With these video clips, we can build a video dataset with more completed scenes. In Fig. 6, we utilize Concerto to deal with single-view lifted data and multi-view lifted data. The visualizations show that Concerto adapts well to these two situations, suggesting that Concerto cannot only be applied to the offline video reconstruction but also the single view forward situation.

B.2 Concerto with Language Probing

We leverage a simple linear layer to translate the representations from Concerto to CLIP’s text space. During training, we force the linear probing output to align with the LSeg [25] image encoder’s output, which does not need the ground truth labels to supervise. In the aligning process, we do not use masks and crop augmentations. The visualization results are shown in Fig. 7.

Data Efficiency	Params		Limited Scenes (Pct.)					Limited Annotation (Pts.)				
Methods	Learn.	Pct.	1%	5%	10%	20%	100%	20	50	100	200	Full
Concerto (lin.)	0.02M	0.02%	48.2	69.1	73.6	75.0	77.3	73.9	75.2	76.2	76.3	77.3
Concerto (dec.)	16.3M	13.1%	44.6	67.9	73.7	74.6	79.5	72.6	74.6	76.7	77.6	79.5
Concerto (full.)	124.8M	100.0%	46.5	69.0	75.3	76.1	80.7	73.3	76.7	77.6	78.4	80.7
Concerto (lora)	0.3M	0.2%	48.4	70.2	74.9	76.8	79.8	75.1	77.2	78.3	78.7	79.8

Table 9: **Parameter Efficiency with LoRA.** Concerto with LoRA significantly improves the performance with a minimal number of learnable parameters, highlighting the reliability of pretrained Concerto representations and the effectiveness of LoRA fine-tuning.

LoRA	Params		ScanNet Val			ScanNet200 Val			ScanNet++ Val			S3DIS Area 5		
Methods	Learn.	Pct.	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
Concerto (lin.)	<0.2M	<0.2%	77.3	86.6	91.7	37.4	49.5	83.3	45.6	60.5	86.5	73.5	81.3	90.9
Concerto (dec.)	16.3M	13.1%	79.5	87.6	92.6	37.8	50.5	84.1	48.3	62.3	87.7	75.5	84.2	92.3
Concerto (full.)	124.8M	100.0%	80.7	87.4	93.1	39.2	50.2	85.0	50.7	63.3	87.9	77.4	85.0	93.2
Concerto (lora)	<0.5M	<0.5%	79.8	87.9	92.7	38.4	51.9	84.1	47.3	60.8	87.7	75.5	81.4	92.6

Table 10: **Semantic segmentation with LoRA.** We compare the LoRA fine-tuning method on Concerto across four semantic segmentation benchmarks, demonstrating LoRA’s remarkable capacity in general and the reliability of Concerto’s original pretrained representations.

398 B.3 Results with Different 2D Encoder

399 In this section, we compare the performance of different strong image encoders: DINOv2 [26],
400 SigLIPv2 [39], and RADIO [32]. We adopt DINOv2 L version with a resolution of 518×518,
401 SigLIPv2 So400m version with a patch size of 16 and resolution 512×512, and RADIOv2.5 L
402 version with a resolution of 768×768. For each model, we pretrain a variant of Concerto on 40k data,
403 excluding video-lifted data. We evaluate these models across four datasets on semantic segmentation,
404 as shown in Tab. 8. The results reveal that the Concerto model based on DINOv2, using self-
405 distillation, achieves the highest mIoU in general. This suggests that in our joint self-supervised
406 learning framework, the optimal synergy is achieved when representations from different domains
407 are derived through intra-modal self-distillation. RADIO, which incorporates distilled information
408 from multiple models, may damage the original self-distillation features from DINOv2, thus leading
409 to a decrease in performance.

410 B.4 Results with LoRA Finetuning

411 From the main results, we observe that linear probing outperforms full-finetuning in extreme data-
412 scarce scenarios. This suggests that training methods may benefit from shifting toward LoRA-based
413 fine-tuning. In this section, we present the results of LoRA fine-tuning. Specifically, we adapt LoRA
414 to the point encoder and evaluate it with linear probing. We set the LoRA rank to 8, the LoRA alpha
415 to 16, and the dropout rate to 0.1.

416 The results of ScanNet Data Efficiency are shown in Tab. 9. These results demonstrate that the
417 LoRA-based method outperforms both linear probing and full fine-tuning in terms of mIoU across
418 most scenarios, despite a small increase in learnable parameters compared to the original linear
419 probing. This suggests that LoRA is an effective fine-tuning approach, particularly when data is
420 limited. Notably, linear probing with LoRA achieves performance comparable to decoder probing in
421 the full evaluation and only a 0.9% performance drop compared to full fine-tuning on mIoU, while
422 offering significant improvements in training efficiency.

423 We also evaluate the LoRA fine-tuning on Concerto across four benchmarks, as shown in Tab. 10.
424 The results demonstrate that LoRA fine-tuning shows performance comparable to decoder probing,
425 even with relatively small learnable parameters. Overall, the LoRA fine-tuning demonstrates strong
426 efficiency and performance across various benchmarks, highlighting two key insights: Concerto
427 already yields reliable and generalizable representations, and leveraging pretrained representations
428 combined with LoRA fine-tuning is both efficient and effective for further task adaptation.

429 **C Broader Impact**

430 In this work, we introduce Concerto, a powerful model for spatial representation learning, achieving
431 SOTA performance in full fine-tuning and superior results in linear probing. Looking ahead, Concerto
432 holds great promise for extending multi-modal learning beyond 2D-3D, benefiting downstream tasks
433 such as autonomous driving, robotics, and mixed reality. However, if not properly trained, point cloud
434 encoders may learn biases from the data, reinforcing societal stereotypes. Researchers should be
435 aware of such potential negative social impacts and continuously monitor the training data for bias.

References

- [1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2, 5, 6, 7, 11, 12
- [2] Sergio Arnaud, Paul McVay, Ada Martin, Arjun Majumdar, Krishna Murthy Jatavallabhula, Phillip Thomas, Ruslan Partsey, Daniel Dugas, Abha Gejji, Alexander Sax, et al. Locate 3d: Real-world object localization via self-supervised learning in 3d. *arXiv:2504.14151*, 2025. 9
- [3] 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 *CVPR*, 2023. 9
- [4] 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 *CVPR*, 2023. 2
- [5] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *ECCV*, 2024. 2
- [6] Lawrence W Barsalou. Grounded cognition. *Annual Review of Psychology*, 2008. 2, 3
- [7] 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 *NeurIPS*, 2021. 2, 6, 11, 12
- [8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2
- [9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 4
- [10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, 2021. 4
- [11] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *CVPR*, 2023. 9
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 9
- [13] Zhimin Chen, Longlong Jing, Yingwei Li, and Bing Li. Bridging the domain gap: Self-supervised 3d scene understanding with foundation models. In *NeurIPS*, 2023. 9
- [14] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 6
- [15] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 3, 5, 6, 7, 8, 11, 12
- [16] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems*, 37:40212–40229, 2024. 9
- [17] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *ICLR*, 2023. 2
- [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 4, 9
- [19] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 5, 6
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

- [21] Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2d and 3d vision-language understanding. *arXiv:2503.10745*, 2025. 9
- [22] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *RAL*, 2020. 2
- [23] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022. 9
- [24] Yann LeCun. A path towards autonomous machine intelligence version 0.9.2. *OpenReview*, 2022. 2, 4
- [25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 2, 7, 13
- [26] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2, 3, 4, 8, 9, 11, 14
- [27] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 2
- [28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 9
- [29] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 11
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2
- [31] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *NeurIPS*, 2021. 2, 6, 11, 12, 13
- [32] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024. 14
- [33] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 3, 5, 6, 7
- [34] Ladan Shams and Aaron R Seitz. Benefits of multisensory learning. *Trends in cognitive sciences*, 2008. 2
- [35] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models. *arXiv:2406.10224*, 2024. 2
- [36] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2
- [37] Mukund Varma T., Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, and Ravi Ramamoorthi. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *CVPR*, 2024. 9
- [38] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *NeurIPS*, 2023. 9
- [39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 14

[40] Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, Bohao Peng, Hengshuang Zhao, and Jiaya Jia. Groupcontrast: Semantic-aware self-supervised representation learning for 3d understanding. In *CVPR*, 2024. 9

[41] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 2, 6, 9, 11, 12

[42] Pengfei Wang, Yuxi Wang, Shuai Li, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Open vocabulary 3d scene understanding via geometry guided self-distillation. In *ECCV*, 2024. 9

[43] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 6

[44] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, 2023. 2, 6, 7, 9

[45] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 1, 2, 4, 5, 6, 7, 8

[46] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *CVPR*, 2024. 6, 7

[47] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

[48] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 2, 6, 9

[49] 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 *CVPR*, 2022. 2

[50] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2, 5, 6, 7, 11, 12

[51] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 2

[52] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *ECCV*, 2024. 9

[53] Karim Abou Zeid, Kadir Yilmaz, Daan de Geus, Alexander Hermans, David Adrian, Timm Linder, and Bastian Leibe. Dino in the room: Leveraging 2d foundation models for 3d segmentation. *arXiv:2503.18944*, 2025. 9

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv:2203.03605*, 2022. 2, 9

[55] Michael Zhang, Aditi Raghunathan Wang, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 5

[56] Xiaoshuai Zhang, Zhicheng Wang, Howard Zhou, Soham Ghosh, Danushen Gnanapragasam, Varun Jampani, Hao Su, and Leonidas J. Guibas. CONDENSE: consistent 2d/3d pre-training for dense and sparse features from multi-view images. In *ECCV*, 2024. 9

[57] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, 2020. 2, 6, 8, 11, 12

[58] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 2, 11, 12

[59] Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, et al. Ponderv2: Pave the way for 3d foundaion model with a universal pre-training paradigm. *arXiv:2310.08586*, 2023. 9