Completing Explicit 3D Reconstruction via View Extrapolation with Diffusion Priors

Seunghoon Jeong[1], Eunho Lee[1], Myung-Hwan Jeon[2] and Ayoung Kim[3*]

***Abstract*— Completing 3D scenes from limited observations requires both optimization and generation, but existing methods often overfit to input views, making it difficult to produce realistic images for extrapolated viewpoints. To address this issue, we propose a pipeline that utilizes 2D diffusion prior explicitly with 3D foundation model for view extrapolation and scene completion. The key idea of this approach is harnessing the diffusion model's prior more directly than refining or inpainting the defective rendering results. A robust 3D reconstruction model, MASt3R, provides depth and normal maps from images, making it possible to create reliable warped images from reference views. Our diffusion model leverages the warped images as conditioning inputs for view extrapolation to ensure the generated images accurately align with the query poses. Furthermore, our method ensures geometric consistency across all views by adopting a divide-and-conquer strategy during the alignment process, incorporating newly generated information into the 3D scene and use it to create updated warped images. We validate our approach on multiple categories from the CO3D dataset, demonstrating superior extrapolation performance, realistic appearance, and enhanced 3D consistency compared to both 3D Gaussian Splatting-based and other diffusion-based baselines.**

I. Introduction & Related Works

3D reconstruction plays a critical role in bridging computer vision to real-world applications, and it is widely considered important in the fields of robotics [1, 2] and AR/VR [3, 4] for more practical uses. One notable advancement in this area is sparse view 3D reconstruction, which aims to generate a scene from only a few input images—unlike earlier methods such as Neural Radiance Fields (NeRF) [5] or 3D Gaussian Splatting (3DGS) [6] that require tens of images. In robotics, the principal challenge lies in scene completion, where occluded or unseen regions must be reliably generated to produce realistic extrapolated views and a 3D representation suitable for downstream tasks.

Many existing studies on sparse view 3D reconstruction have often been developed within the 3DGS framework, which is faster and lighter than NeRF while providing an explicit scene representation. These methods attempt to address the ill-posed nature of the problem by leveraging depth estimation models [7–11] or 3D foundation models [12, 13]. However, 3DGS-based methods are highly susceptible to overfitting in sparse input views, so they typically perform well only for interpolated viewpoints. If the 3D representation is optimized solely to match the input without generating unseen regions, any reconstruction derived from that representation will remain incomplete, significantly reducing its potential in robotics applications.

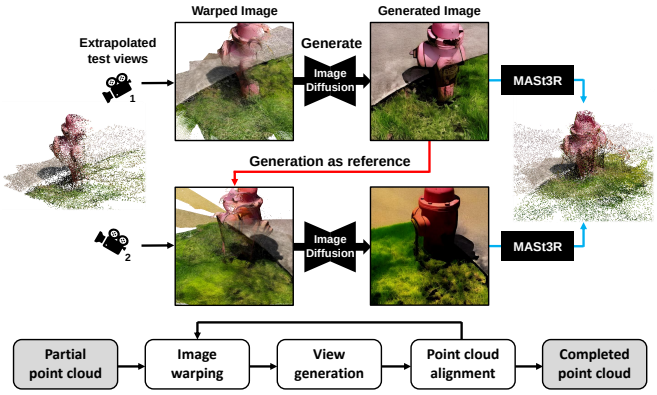

Fig. 1: Generating extrapolated views that move beyond the initial input views and still accurately depict the scene and objects is a challenging task in the sparse view setting. We tackle this problem through an iterative process of image diffusion and 3D reconstruction, ultimately aiming to produce a complete point cloud.

To handle this issue, the model and pipeline must effectively transform the information obtained from the input images into a 3D representation, while reliably generating the unobserved regions of the scene. Several approaches have utilized 2D diffusion models to generate extrapolated views—either by refining or inpainting the rendering results from 3DGS [14–16] or by generating gradients through Score Distillation Sampling (SDS) loss [10]. However, these methods have limitations in fully leveraging the diffusion prior, resulting in inadequate repairs due to the dominance of artifacts and floaters, as well as insufficient diversity in the generated regions. Some techniques such like [17, 18] utilize multiview transformer and synthesize novel views directly, but they suffer from slow rendering due to NeRF representation.

In this work, we propose a method for reconstructing 3D scenes from sparse views by directly utilizing the powerful prior of 2D diffusion model and the versatility of 3D foundation model as shown in Fig. 1. An image diffusion model generates extrpolated views from warped images created with depth maps of reference views. A 3D dense reconstruction model such as DUSt3R [19] and MASt3R [20] can estimate reliable depth maps regardless of the relative pose between images. By repeatedly using these two models, we can obtain a filled point cloud free from self-occlusion issues from any

[1]S. Jeong, E. Lee are with the Dept. of Interdisciplinary Program in Artificial Intelligence, SNU, Seoul, S. Korea [shoon0602, eunho1124]@snu.ac.kr

[2]M-. Jeon is with the Kinetic Intelligent Machine Lab (KIMLAB), University of Illinois Urbana-Champaign, Champaign, IL, 61801, USA mhjeon@illinois.edu

[3]A. Kim is with the Dept. of Mechanical Engineering, SNU, Seoul, S. Korea ayoungk@snu.ac.kr

viewing direction.

Warped image offers partial guidance for certain pixels when generating novel views, making it extremely useful. In video diffusion models, for example, many approaches such as [21, 22] condition on the warped image from the first frame to generate subsequent frames with respect to a given camera trajectory. We similarly leverage this rich information in image diffusion to ensure the generated images align with the target camera pose without translation or rotation errors. The use of image diffusion also preserves the advantages of lower memory usage and computational efficiency, making it particularly well-suited for robotics applications.

The query view images, generated based on the structure of the warped images, are aligned with the reference views using 3D reconstruction model again. Obtained depth maps are subsequently merged with the existing global point cloud to reflect newly generated regions. A 360-degree 3D scene consistent with the original reference images is constructed using a divide-and-conquer strategy, in which previously generated view images also become reference views for subsequent steps. Our method generates images corresponding to the desired query pose while preserving 3D consistency not only among the original images, but also among the newly generated images. The key contributions of our work include:

- **3D Reconstruction with Divide-and-conquer Strategy**: We propose a pipeline that leverages MASt3R's ability to provide an explicit and dense 3D scene representation with a divide-and-conquer strategy to progressively generate new views.
- **Warped Image-conditioned Image Generation Model**: We utilize a category-specific image diffusion model which can reliably produce extrapolated views including the object and background from the warped images.
- **Experimental Evaluation and Discussion**: We observed that our method can generate realistic, consistent images for extrapolated views and complete the scene, which are typically challenging for various 3DGS-based and other diffusion-based approaches.

II. Methods

Our goal is to obtain a reliable 3D point cloud for object-centric scenes given a sparse set of input images and poses. After performing an initial 3D reconstruction from the input images, we repeat the following steps: (1) create a warped image with respect to selected query pose, (2) generate an image conditioned on the warped image, and (3) align the newly generated image with the global map to update the 3D scene.

A. Warped Image Generation

A warped image can be derived from the reference image's depth D_{ref}, the reference/query camera poses P_{ref}/P_{query}, and the camera intrinsics K, using the following:

$$I_{\text{warp}} = H_{warp} I_{ref} = K\, P_{\text{query}}\, P_{\text{ref}}^{-1}\, D_{\text{ref}}\, K^{-1}\, I_{\text{ref}}, \quad (1)$$

where the image depth can be obtained from MASt3R, and the pose and intrinsics are known.

However, when using warped images, a major issue arises from self-occlusion: surfaces that should not be visible in the query pose can erroneously appear. To mitigate this, we calculate normal vectors for each point in the point cloud using Open3D [23]. Only pixels whose normals form an angle of 90 degree or more with the query camera ray are warped. Additionally, for pixels with multiple projected points, we apply Softmax Splatting [24] based on depth, so the closer point with smaller depth dominates. This approach yields cleaner edges and more realistic views than simple averaging. When assembling warped images from multiple reference images, we similarly use depth-aware weighting.

B. Warped Image-conditioned with Diffusion

In our method, which incrementally generates new views and integrates them into the global map, selecting the next view to generate is a critical step. If a candidate view is too close to the existing views, the model can easily generate it, but it will provide little additional 3D information. Conversely, if it is too far from the existing views, producing a reliable image becomes more challenging. We experimentally found that for viewpoint changes of roughly 30 degrees, the model can produce sufficiently reliable images and also provide enough new information.

To generate novel views from the warped images, we adapted the ZeroNVS [25] model by adding extra inputs and fine-tuning it. ZeroNVS effectively reduces scale ambiguity by normalizing the pose with camera locations, and our model also utilizes this relative pose representation. Additionally, we incorporate warped images obtained from multiple reference views, while using a single reference view (the one closest to the query) to preserve fine details. This reference view is concatenated with the warped images as an input; its CLIP [26] embedding is also obtained and used alongside the pose as a conditioning signal.

A warped image supplies structural information about the scene; however, using only RGB is insufficient for establishing pixel-level correspondences. Therefore, we concatenate the sinusoidal positional encoding for each pixel of the reference view image and its warped counterpart (aligned to the query pose) as additional inputs.

The model is trained separately for each category and used accordingly. By training the model separately for each category, it acquires a sufficient prior specific to that category. Furthermore, we added an additional LoRA [27] fine-tuning step tailored to each individual scene. Given N input images, each image can be warped from the perspective of every other image, generating up to $N(N-1)$ pairs. A few minutes of training on these pairs adapts the model to the specific scene, thereby enhancing its per-scene performance.

C. Incremental 3D Reconstruction

To obtain depth maps for newly generated images, we run MASt3R again. Because MASt3R is a model that takes a pair of images as input, aligning multiple images typically

Fig. 2: Synthesis results of novel view extrapolation for several categories. We also visualized warped image used in our model.

requires running it on every edge of the complete graph, then building a global map via a minimum spanning tree. To reduce unnecessary computation, we skip creating edges between already existing images. Moreover, since we already know all of the poses and intrinsics of the images and the depth maps for the original and previously generated images, we freeze those values and optimize only the depth map for the newly generated images.

III. EXPERIMENT

A. Experimental Setup

1) Dataset: For model training, we filtered 360-degree, object-centric sequences from 6 categories in CO3D [28]—apple, book, chair, cup, hydrant, and teddybear, and created sets of image pairs. We selected categories spanning a broad spectrum of object variance—from those with significant intra-category diversity to those with relatively little. Each pair is taken from the same sequence, and the model learns to generate one image from the other's warped image. To filter too easy or difficult pairs to train and enhance the quality of training, the pairs are greedily chosen according to a score function below:

$$s(R_1, R_2) = \cos\alpha(1 - \cos\alpha), \tag{2}$$

where α is the angle between two camera poses, which is similar to the image pair selection method of CroCo v2 [29]. Additionally, to resolve the misalignment between the warped image and the ground truth image caused by depth errors, we used MINIMA [30], a robust image matching model as a preprocessing step. This ensures that the partial observations from the warped image align to the correct positions in the ground truth image.

For each category, we split its sequences into train and test sets, training the model solely on the image pairs derived from the training sequences. Each category includes 4-5 test sequences, and we performed farthest point sampling on the camera positions to select 9 view images that were maximally distant from each other. The baselines and our model were then evaluated by using 3 or 6 of these images for input and attempting to reconstruct the remaining images. Note that within the sparse view inputs, some sequences are appropriately spaced around the 360-degree view, while others cluster on just one side, providing no information about the opposite side.

2) Metric: We evaluated and compared the novel view synthesis results of our method and various baselines from multiple perspectives. While PSNR and SSIM provide an intuitive, pixel-level measure of error, these metrics can be misleadingly high for blurry or unrealistic images. Additionally, these metrics do not adequately address the issue of generating unknown regions. Therefore, we evaluated Masked PSNR and Masked SSIM from MegaScenes [31], which is calculated with pixels available in the mask derived from the warped images between reference views.

We also measure LPIPS [32] and DISTS [33] to assess the generated images in a manner more closely aligned with human perception. LPIPS evaluates how visually similar the image remains even if the diffusion model introduces slight transformations, thereby gauging how faithfully the "feel" of the original view is preserved. DISTS is designed to be more sensitive to structural differences rather than texture, focusing on how geometrically consistent the synthesized view is.

We measured each method's perform time to compare their computational efficiency. Perform time denotes the image preprocessing and optimization duration for 3DGS-based methods, and the per-image generation and alignment duration for diffusion-based methods. 3DGS-based methods render new views very quickly but require a long optimization stage, whereas our method aligns views rapidly but spends more time on view generation.

3) Baseline: Our baselines include the standard 3DGS and its sparse view variants—FSGS [8], InstantSplat [12],

TABLE I: Rendering/Generation results for novel view synthesis. All metrics are reported as averages over each category. Except for the perform time column, whose measurement differs by method, the best result is shown in **bold**, and the second-best is shown in underline.

		LPIPS ↓		DISTS ↓		Masked PSNR ↑		Masked SSIM ↑		Perfrom Time ↓	
		3-view	6-view	3-view	6-view	3-view	6-view	3-view	6-view	3-view	6-view
3DGS based	3DGS	0.707	0.640	0.385	0.351	12.759	14.342	0.583	0.621	291.7	335.7
	FSGS	0.666	0.596	0.384	0.338	14.784	15.218	0.684	0.673	108.8	118.8
	InstantSplat	0.588	**0.481**	0.327	0.276	**16.792**	**18.149**	**0.750**	**0.753**	71.1	103.6
	DNGaussian	0.723	0.644	0.417	0.354	13.759	14.463	0.683	0.677	159.5	141.4
Diffusion based	ZeroNVS	0.694	0.652	0.339	0.311	12.641	12.812	0.606	0.569	22.6	37.0
	MegaScenes	0.603	0.526	0.279	**0.245**	14.352	14.783	0.646	0.631	38.0	59.3
	Ours	**0.577**	0.525	**0.266**	0.246	14.397	14.694	0.645	0.616	38.5	59.4

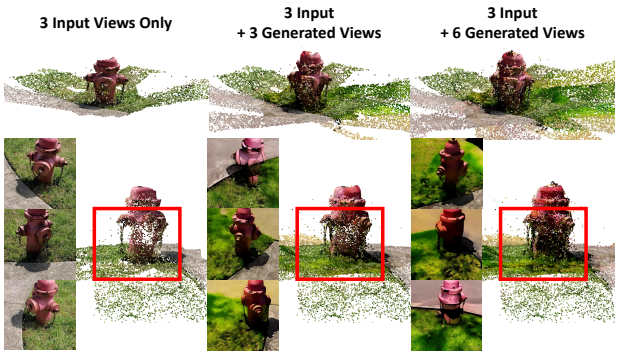

Fig. 3: 3D Reconstruction results of our method. Generated images by the categorical diffusion model are well-aligned, progressively filling in the scene's empty regions. Notably, the rear surfaces of the hydrant which were not visible in the input views are also properly reconstructed.

and DNGaussian [7]. All 3DGS-based methods underwent the necessary preprocessing steps and were run using the default hyperparameters of each method. We also adapted ZeroNVS [25], originally designed for single view NVS, to select the nearest reference view at inference time, enabling it to operate in sparse view scenarios. Finally, we compare against the model introduced from MegaScenes, which fine-tunes ZeroNVS model to accept a single reference-view warped image as an additional input.

4) Implementation Details: Our model was trained using a modified version of the publicly available MegaScenes code, initialized with ZeroNVS weights. When creating the warped image, up to three nearest reference views were used, and DDIM [34] was employed by the diffusion model to generate novel views. Training on a single NVIDIA A6000 GPU took about six hours with batch size 4.

B. Qualitative & Quantitative Results

As shown in Fig. 2, the qualitative results demonstrate that warped images can be highly advantageous for novel view extrapolation. 3DGS-based methods often overfit to the sparse input images, leading to unrealistic outcomes with even slight viewpoint change. Meanwhile, the warped images generated using MASt3R's 3D reconstruction offer reliable information for moderate extrapolations, enabling our model to produce more realistic images. In contrast to MegaScenes, training categorical model on object-centric scenes and incorporating additional positional encoding appears to help the model more effectively understand the scene.

The quantitative results in Table. I compare novel view synthesis performance, including interpolation and extrapolation scenarios. The 3DGS-based methods perform well on query poses that are interpolated within the input views, yielding high pixel-based metric scores. However, these methods exhibit lower scores on LPIPS (assessing realism) and DISTS (focusing on consistency), especially when generating extrapolated views. The strong performance of our method in DISTS indicates that our method preserves the geometric structure well, thanks to the warped images. When compared with other diffusion-based approaches, our method consistently generates images that are perceptually superior.

When examining how performance varies with the number of input views, our method shows a slight improvement compared to 3DGS-based approaches. This is because, as more input views are provided, overall coverage increases and the need for novel view extrapolation decreases.

Fig. 3 illustrates how unseen parts of the scene are gradually filled in by the alignment of the newly generated views. The point cloud of hydrant is completed using the categorical diffusion model's prior, learned from other object-centric scenes within the same category.

IV. Conclusion

Although many methods have been proposed for reconstructing 3D scenes using only a small number of images, they often fail to ensure consistency or require substantial resources. To address this, we propose a pipeline that iteratively performs novel view generation and 3D reconstruction, leveraging the categorical diffusion model's prior to extrapolate views and fill in unobserved regions. By utilizing the geometric information obtained from warped images, we generate realistic and consistent images. The newly filled areas are then integrated into the global point cloud through MASt3R. Furthermore, by reusing the newly generated images as references in subsequent iterations, we maintain consistency across all generated images.

Despite the advantages, our approach heavily depends on both the generation and alignment steps. Since each generation outcome influences the subsequent one, accumulated errors from MASt3R can result in improperly produced warped images, and the diffusion model may fail to generate appropriate image. As future work, we plan to use the confidence output from MASt3R to detect misalignments in the depth map and apply refinement to the global point cloud at each iteration, aiming for a cleaner overall completion.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) No.2022-0-00480, Development of Training and Inference Methods for Goal-Oriented Artificial Intelligence Agents

REFERENCES

[1] O. Shorinwa, J. Tucker, A. Smith, A. Swann, T. Chen, R. Firoozi, M. Kennedy III, and M. Schwager, "Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting," *arXiv preprint arXiv:2405.04378*, 2024.

[2] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu *et al.*, "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," *IEEE Robotics and Automation Letters*, 2024.

[3] M.-D. Yang, C.-F. Chao, K.-S. Huang, L.-Y. Lu, and Y.-P. Chen, "Image-based 3d scene reconstruction and exploration in augmented reality," *Automation in Construction*, vol. 33, pp. 48–60, 2013.

[4] M. Cao, L. Zheng, W. Jia, H. Lu, and X. Liu, "Accurate 3-d reconstruction under iot environments and its applications to augmented reality," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2090–2100, 2020.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[7] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 775–20 785.

[8] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "Fsgs: Real-time few-shot view synthesis using gaussian splatting," in *European conference on computer vision*. Springer, 2024, pp. 145–163.

[9] H. Huang, Y. Wu, C. Deng, G. Gao, M. Gu, and Y.-S. Liu, "Fatesgs: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[10] H. Xiong, *SparseGS: Real-time 360° sparse view synthesis using Gaussian splatting*. University of California, Los Angeles, 2024.

[11] A. Paliwal, W. Ye, J. Xiong, D. Kotovenko, R. Ranjan, V. Chandra, and N. K. Kalantari, "Coherentgs: Sparse novel view synthesis with coherent 3d gaussians," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–37.

[12] Z. Fan, K. Wen, W. Cong, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "Instantsplat: Sparse-view sfm-free gaussian splatting in seconds," *arXiv preprint arXiv:2403.20309*, 2024.

[13] Y. Tang, Y. Guo, D. Li, and C. Peng, "Spars3r: Semantic prior alignment and regularization for sparse 3d reconstruction," *arXiv preprint arXiv:2411.12592*, 2024.

[14] C. Yang, S. Li, J. Fang, R. Liang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Gaussianobject: Just taking four images to get a high-quality 3d object with gaussian splatting," *arXiv e-prints*, pp. arXiv–2402, 2024.

[15] S. Paul, C. Wewer, B. Schiele, and J. E. Lenssen, "Sp2360: Sparse-view 360○ scene reconstruction using cascaded 2d diffusion priors," in *ECCV 2024 Workshop on Wild 3D*. OpenReview. net, 2024.

[16] A. Paliwal, X. Zhou, W. Ye, J. Xiong, R. Ranjan, and N. K. Kalantari, "Ri3d: Few-shot gaussian splatting with repair and inpainting diffusion priors," *arXiv preprint arXiv:2503.10860*, 2025.

[17] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole *et al.*, "Reconfusion: 3d reconstruction with diffusion priors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 551–21 561.

[18] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole, "Cat3d: Create anything in 3d with multi-view diffusion models," *arXiv preprint arXiv:2405.10314*, 2024.

[19] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[20] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[21] N. Müller, K. Schwarz, B. Rössle, L. Porzi, S. R. Bulò, M. Nießner, and P. Kontschieder, "Multidiff: Consistent novel view synthesis from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 258–10 268.

[22] K. Liu, L. Shao, and S. Lu, "Novel view extrapolation with video diffusion priors," *arXiv preprint arXiv:2411.14208*, 2024.

[23] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[24] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5437–5446.

[25] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun *et al.*, "Zeronvs: Zero-shot 360-degree view synthesis from a single image," *arXiv preprint arXiv:2310.17994*, 2023.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[28] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 901–10 911.

[29] P. Weinzaepfel, T. Lucas, V. Leroy, Y. Cabon, V. Arora, R. Brégier, G. Csurka, L. Antsfeld, B. Chidlovskii, and J. 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*, 2023, pp. 17 969–17 980.

[30] X. Jiang, J. Ren, Z. Li, X. Zhou, D. Liang, and X. Bai, "Minima: Modality invariant image matching," *arXiv preprint arXiv:2412.19412*, 2024.

[31] J. Tung, G. Chou, R. Cai, G. Yang, K. Zhang, G. Wetzstein, B. Hariharan, and N. Snavely, "Megascenes: Scene-level view synthesis at scale," in *European Conference on Computer Vision*. Springer, 2024, pp. 197–214.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
[33] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.
[34] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.