Learn Your Scales: Towards Scale-Consistent Generative Novel View Synthesis

Fereshteh Forghani[1] Jason J. Yu[1] Tristan Aumentado-Armstrong[1,3]
Konstantinos G. Derpanis[1,2,3] Marcus A. Brubaker[1,2,4]

[1]York University [2]Vector Institute for AI [3]Samsung AI Centre Toronto [4]Google DeepMind

Abstract

Conventional depth-free multiview datasets are captured using a moving monocular camera without metric calibration. The scales of camera positions in this monocular setting are ambiguous. Previous methods have acknowledged scale ambiguity in multiview data via various ad-hoc normalization pre-processing steps, but have not directly analyzed the effect of incorrect scene scales on their application. In this paper, we seek to understand and address the effect of scale ambiguity when used to train generative novel view synthesis methods (GNVS). The generative nature of these models captures all aspects of uncertainty, including any uncertainty of scene scales, which act as nuisance variables for the task. We study the effect of scene scale ambiguity in GNVS when sampled from a single image by isolating its effect on the resulting models and, based on these intuitions, define new metrics that measure the scale inconsistency of generated views. We then propose a framework to estimate scene scales jointly with the GNVS model in an end-to-end fashion. Empirically, we show that our method reduces the scale inconsistency of generated views without the complexity or downsides of previous scale normalization methods. Further, we show that removing this ambiguity improves generated image quality.

1. Introduction

A central task of visual perception is interpreting the 3D structure of 2D images, as geometric information is lost in the perspective projection process. Multiview vision can recover substantial 3D structure, but cannot obtain *metric* geometry, due to the "scale ambiguity" inherent to images [10, 12] (see Fig. 1). Similarly, an important task in visual perception for robotics, monocular visual odometry (VO), also struggles with scale ambiguity [4, 11, 21]. The scale ambiguity problem, therefore, naturally appears in most real datasets using visual methods for camera pose estimation.

Most multiview image datasets used by learning-based methods, e.g., RealEstate10K [38], obtain camera poses via monocular SLAM systems, such as ORBSLAM [20] or COLMAP [26], yielding consistent relative poses but ambiguous absolute scale. To handle the scale inconsistencies across scenes, the authors "scale-normalized" each sequence by scaling a specific depth quantile to a reasonable value for their method. This ambiguity complicates many 3D computer vision tasks, such as novel view synthesis (NVS), where uncalibrated data introduces variability in the perceived sizes and distances of objects.

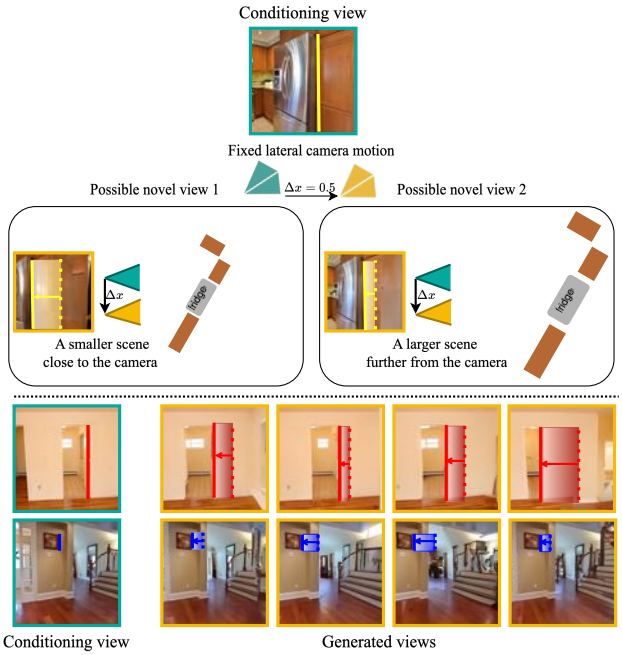

Figure 1. **Scale ambiguity and inconsistency in GNVS.** (Top) Two novel views are independently sampled using the same conditioning and camera motion, Δx. Samples exhibit different disparities due to uncertainty over scene scale. Here, we depict the samples' plausible top down scene layouts in the boxes. (Bottom) Additional samples and scenes in the same setting are shown, where a salient edge is highlighted to show the different disparities in the generated views.

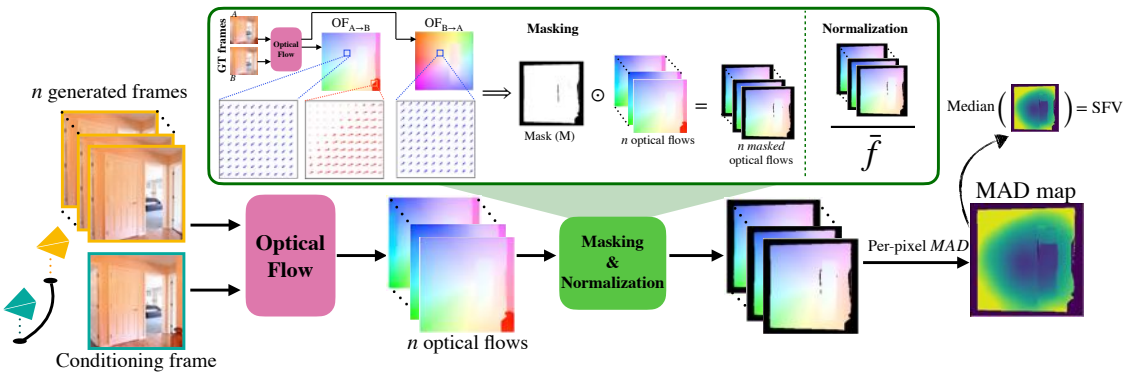

Figure 2. **SFV metric overview.** From n generated frames, we compute optical flows between the conditioning image and each output. Patches outlined by squares illustrate masking: red vectors (masked) show pixels leaving the view; blue vectors (unmasked) indicate cycle-consistent motion. Masked flows are then normalized by the mean magnitude of unmasked flows, $\bar{f}$.

NVS, a long-standing challenge in computer vision and graphics (e.g., [1, 7, 15]), can be a highly ambiguous task due to unobserved elements within scenes. To handle such uncertainties, *generative* NVS (GNVS) models have been devised (e.g., [16, 22, 23, 32]), where most recent methods use diffusion models [5, 9, 17, 18, 30, 34–36]. A GNVS model, unlike NeRFs [2, 19, 36] and 3DGS [14], faces an under-determined scenario: given a single observed image and a trajectory of camera parameters, it must generate novel views for that camera path. Inconsistent scales in the training data create uncertainty since the model can no longer assume a fixed size for objects. Existing GNVS models show this with high entropy in generated views: e.g., two samples of a novel viewpoint can place the same object at different positions despite same camera movements.

Existing methods, e.g., [25, 29, 31] perform scale estimation or normalization as *a fixed preprocessing step,* based on heuristics applied to the scene and camera geometries. As a result, errors in the initial scales cannot be corrected, valuable training data is potentially sacrificed, and the resulting GNVS model learns to reproduce this uncertainty in scale. In contrast, our approach *learns* scene scales during the training process, adapting and correcting them via the GNVS task loss. In terms of evaluation, NVS metrics neither capture the statistical entropy induced by the scale variability nor measure the inconsistencies in scale among the generated views. FID [13] compares image distributions without considering metric scale, while reconstruction metrics such as LPIPS[37] and PSNR are quickly dominated by other sources of error, including differences in semantic content. In contrast, Structure-from-Motion (SfM)-based metrics [31] and TSED [34] are inherently scale-invariant.

In this work, we (i) identify and quantify the issue of scale uncertainty in the GNVS task, and (ii) propose a simple learning-based method to curtail its impact by optimizing a per-scene scale jointly with the generative model.

2. Methods

2.1. Scale Learning

Let $S = \{s_i\}_{i=1}^{N}$ be the per-scene scales for N scenes, which we aim to learn. These scene scales modify the camera extrinsics for each frame in a scene by scaling the translation component of the pose. That is, for each frame belonging to scene i, the translation vector is scaled by the corresponding scene scale s_i. We assume the original camera translations, as determined by the applied calibration and normalization procedures, provide a reasonable initialization. To ensure scales remain positive, we use an exponential parameterization. This formulation allows the model to adjust scene-wise scales within a moderate and controlled range, centered around the assumed base scale.

The scales S are learned jointly with the parameters, θ, of a latent diffusion model. We use PolyOculus [35], a GNVS model based on a latent diffusion process [24]. The training loss follows the standard form used in denoising diffusion models, where noise is predicted from noisy latent codes and corresponding scaled camera poses.

2.2. Quantifying Scale Variability

Sample Flow Variability. An overview of the SFV metric is given in Fig. 2. For a given condition image, we generate a set of n images conditioned on the same frame with the same camera motion between the conditioned and generated frames. We compute the forward optical flow from the conditioned image and each generated image. We use RAFT [28] to compute optical flow. To avoid using parts of the image where correspondences cannot be made (occlusions and disocclusions), we compute a mask, M, of the image by checking for cycle consistency [3, 6, 27] of the forward and backward optical flows. To avoid our metric from being dominated by other sources (geometry, camera motion), we normalize the optical flow of each scene. Specifically, we normalize the forward optical flows with the average mag-

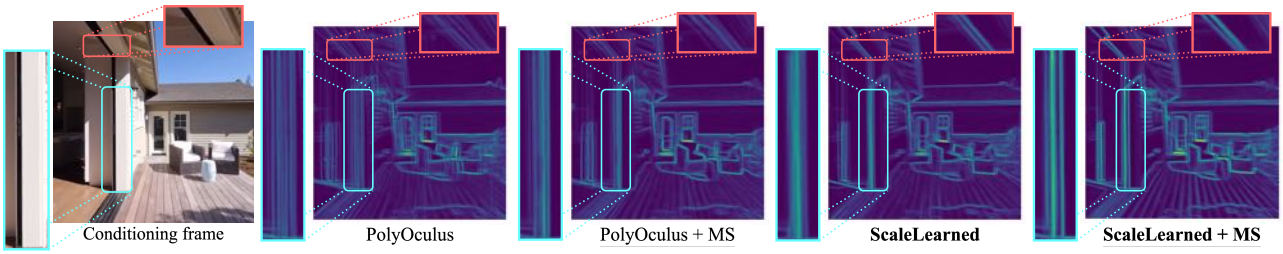

Figure 3. **Edge heatmaps.** We visualize the average Sobel filter responses of multiple samples generated with the same conditioning information to highlight distinct regions of the scene structure. Consistency in edge locations results in more clarity in the edge heatmaps, which indicates a reduction in randomness caused by scale ambiguity. Note that PolyOculus samples are quite noisy in terms of edge locations, and scale learning helps stabilize edge locations.

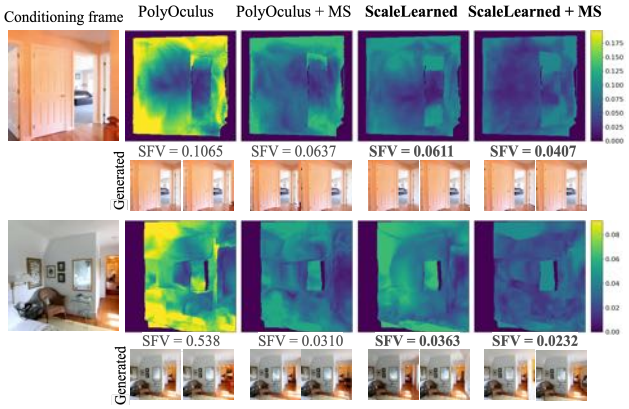

Figure 4. **Examples of per-pixel optical flow MAD maps.** The darker the pixel, the lower the variation of optical flow in that pixel, indicating a more consistent scale among the generated samples. The entropy in scale can also be seen by comparing the generated frames, *e.g.*, the width of the door in the second row.

nitude of flows at unmasked pixels, $\bar{f}$.

Finally, to be robust to outliers in optical flow, we use the per-pixel median absolute deviation (MAD) across all n masked normalized flows. The SFV of a single conditioning image and desired camera pose is, SFV $=$ median$_{p \in M}$MAD$[p]$, where the median is computed over pixel locations, p, and M is the cycle consistency mask. We average the SFV of different conditioning images and camera poses to create a final metric. Visualizations of the MAD maps as heatmaps and SFV values for different scenes are shown in Fig. 4.

Scale-Sensitive Thresholded Symmetric Epipolar Distance. TSED measures geometric inconsistencies between pairs of images using epipolar geometry. By design, TSED is insensitive to changes in scene scale between a generated and conditioning view because 2D correspondences move along epipolar lines in respect to changes in scene scale, yielding no detectable errors.

Instead, we make TSED able to detect scale inconsistencies by measuring the TSED between two independently generated views that move in perpendicular directions. During generation, the model can "choose" different scales, which creates an inconsistency in the epipolar geometry between the generated views. By evaluating TSED only on the generated views, TSED becomes sensitive to scene scale.

To measure this, we compute a metric as follows: for each axis (x, y, z), we apply a fixed-magnitude camera translation in a random direction and generate a corresponding image. Only translations are used, since rotations are unaffected by scale. We then sample pairs of generated views along the various axes and compute the percentage of consistent pairs using TSED [34]. Averaging this across many conditioning views yields the *Scale-Sensitive Thresholded Symmetric Epipolar Distance*, or SS-TSED score.

3. Experiments

We explore a variety of methods to evaluate the reduction of scale uncertainty in our scale learning method. We use SFV and SS-TSED as metrics for evaluating the amount of scale variability a model exhibits in its generations. Next, we evaluate the effect of scale learning on image quality in GNVS using reconstruction metrics.

3.1. Experimental Setup

Datasets. We train and evaluate on the RealEstate10K (RE10K) [38] dataset. In addition to the raw camera poses provided by the dataset, we also consider poses calibrated using metric monocular depth estimators (MS) [33] as a representative method for existing ad-hoc scale estimation methods [31]. We set the scale to 1.0 for scenes with unreliable scale estimates, resulting in $\sim$30% of the metric scales staying the same as those provided by RE10K and use both sets of poses to act as *reference scales* from which we apply our scale learning.

GNVS Model. We use PolyOculus [35] as our baseline GNVS model. To compare the effect of scale learning, we train two models for each reference scale, one with and one without scale learning. All models are trained from scratch for 900 epochs. For models using scale learning, we initialize the learnable parameters, s_i, to one. We use different optimizers for the scale factors and the denoising network.

Method	LPIPS ↓				PSNR ↑			
	1	2	3	4	1	2	3	4
PolyOculus	0.037	0.045	0.053	0.061	29.98	27.38	26.32	25.21
PolyOculus + MS	0.037	0.046	0.052	0.060	29.10	27.61	26.56	25.56
ScaleLearned	0.035	0.042	0.049	0.056	29.50	28.13	27.17	26.17
ScaleLearned + MS	**0.034**	**0.041**	**0.048**	**0.055**	**29.62**	**28.22**	**27.24**	**26.24**

Table 1. **Reconstruction metrics.** Computed on the test set for one to four frames ahead. Scale learning shows improvements in reconstrunction metrics.

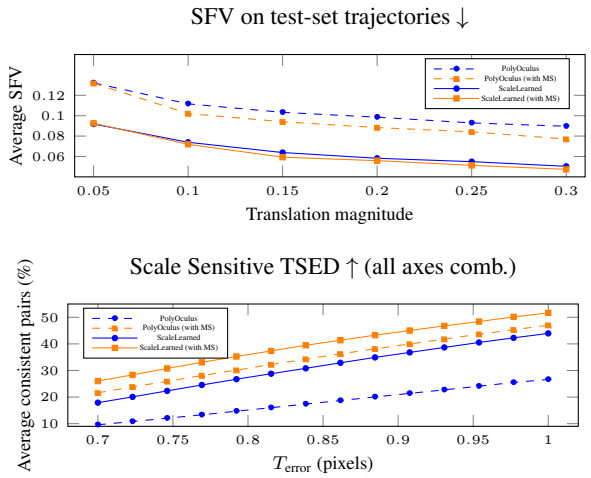

Figure 5. **SFV and SS-TSED evaluations.** (Top) Average SFV values for different translation magnitudes. (Bottom) SS-TSED consistent pair percentages across T_{error} thresholds. Scale learning improves performance with both metrics, further enhanced by combining with metric scales.

3.2. Scale Variability as Apparent Motion

The variance in motion among novel views with the same generation parameters can be made more salient by averaging Sobel filter [8] outputs from multiple samples to form heatmaps. In Fig. 3, as expected, we see that the PolyOculus baseline model produces samples with large amounts of scale variability, where magnified regions show edges that become blurred, appearing in multiple locations between the samples. In contrast, scale learned models show sharper heatmaps. Qualitative inspection of the MAD heatmaps in fig. 4 also show the ScaleLearned models have darker maps with lower SFVs due to lower per-pixel MAD, reflecting lower optical flow variation.

For quantitative evaluations, we compute SFV using 200 randomly selected test images. For each image, we select poses at various distances from the observed view along ground-truth camera trajectories to evaluate scale variation at multiple magnitudes of camera motion. We select views with distances that match the closest to this set of magnitudes: $T = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. The magnitudes chosen are relatively small to avoid introducing additional variance when generating new scene content. Finally, for each generated and observed view, we draw 10 samples to evaluate models' scale uncertainty.

As shown in Figure 5, models incorporating scale learning outperform those without it in terms of SFV across a range of translation magnitudes, demonstrating that learning scale during training reduces motion variation between the conditioning and generated frames. While metric-depth scaled camera poses help reduce sample scale variance (PolyOculus + MS), ScaleLearned model achieves significantly better performance, indicating that the scale-learning approach is more effective than metric-depth estimation in preserving scale consistency. Further improvements are observed when using metric-depth scales as reference scales in scale learning (ScaleLearned + MS).

3.3. Scale Variability as Epipolar Errors

We compute SS-TSED over 200 images in the test set and for each image, we sample 100 pairs. The results in Fig. 5 demonstrate that incorporating scale learning significantly improves epipolar consistency across both reference scale cases. Using metric-depth estimated scales also enhances SS-TSED, and further performance gains are achieved by combining scale learning with this approach.

3.4. Scale Variability as Reconstruction Errors

Reconstruction errors in GNVS can be mostly attributed to the misalignment of scene content due to different scene scales. To mitigate the issue and make reconstruction metrics more sensitive to scale variability, we propose a test-time scale estimation procedure. Specifically, we freeze the model weights and *only* learn the scales of 1500 random test scenes with the diffusion loss used in training. This procedure should align the scale of each test scene with each models' internal expectation of scale, thus mitigating reconstruction errors caused by misaligned content. For these reasons, we apply test-time scale estimation for all models, even those that do not use scale learning during training. The same trend found in our previous experiments is also found here in Table 1. Scale learning consistently improves reconstruction, and performs slightly better when applied in conjunction with metric-depth calibrated reference scales.

4. Conclusion and Discussion

In this paper, we addressed the challenge of scale ambiguity in GNVS models. We first demonstrated the problem, showing that scale ambiguities present in multiview datasets manifest in the conditional image distributions learned by GNVS models. To quantify scale variability in a GNVS model, we defined two metrics based on optical flow and epipolar geometry. We further introduced a new method to learn scene scales by optimizing them during GNVS model training. We empirically showed that our learning-based method effectively reduces the scale ambiguities in a trained GNVS model. Further, we showed that image quality metrics were improved. Our approach offers a simple yet effective general solution that requires no preprocessing, and our metrics evaluate scale consistency in GNVS models.

References

[1] Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *CVPR*, pages 1034–1040, 1997. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 2

[3] Michael J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 1996. 2

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1

[5] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, pages 4194–4206, 2023. 2

[6] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *ICCV*, 2021. 2

[7] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of SIGGRAPH*, pages 279–288, 1993. 2

[8] David A. Forsyth and Jean Ponce. *Computer Vision - A Modern Approach, Second Edition*. Pitman, 2012. 4

[9] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. *NeurIPS*, 2024. 2

[10] Sanjib K. Gosh. Analytical methods and instruments (chapter 6). In *History of photogrammetry*. ISPRS, 1981. Laval University, Canada. 1

[11] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 1

[12] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 1

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 2

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 2

[15] Stéphane Laveau and Olivier D. Faugeras. 3D scene representation as a collection of images. In *ICPR*, pages 689–691, 1994. 2

[16] Andrew Liu, Ameesh Makadia, Richard Tucker, Noah Snavely, Varun Jampani, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, pages 14438–14447, 2021. 2

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, 2023. 2

[18] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[19] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[20] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015. 1

[21] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 1

[22] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing A consistent long-term 3D scene video from A single image. In *CVPR*, pages 3553–3563, 2022. 2

[23] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3D priors. In *ICCV*, pages 14336–14346, 2021. 2

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[25] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 2

[26] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[27] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*, pages 438–451, 2010. 2

[28] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 2

[29] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, pages 548–557, 2020. 2

[30] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *CoRR*, abs/2210.04628, 2022. 2

[31] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. Controlling space and time with diffusion models. *CoRR*, abs/2407.07860, 2024. 2, 3

[32] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, pages 7465–7475, 2020. 2

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. *NeurIPS*, 2024. 3

[34] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, pages 7071–7081, 2023. 2, 3

[35] Jason J. Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. PolyOculus: Simultaneous multi-view image-based novel view synthesis. *ECCV*, pages 433–451, 2024. 2, 3

[36] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-Splatting: Alias-free 3D Gaussian splatting. In *CVPR*, 2024. 2

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 2

[38] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 37(4):65, 2018. 1, 3