

One View Is Enough: In-the-Wild Monocular Pretraining for Novel View Generation

Adrien Ramanana Rahary^{1,2} Nicolas Dufour¹ Patrick Pérez¹ David Picard²

¹Kyutai ²LIGM, ENPC, IP Paris, CNRS, UGE

{adrienrr, nicolas.dufour, patrick}@kyutai.org david.picard@enpc.fr

Abstract

Monocular novel-view synthesis has long required multi-view image pairs for supervision, limiting training to a narrow set of purpose-built datasets. We propose in-the-wild monocular pretraining: a frozen depth estimator lifts each source image into 3D and reprojects under sampled poses to yield pseudo-target views; masked losses restrict supervision to valid regions and an adversarial objective covers disoccluded areas. Scaled to 30 million uncurated images, this produces OVIE, requiring only a source image and target pose at inference. Prior work trains without multi-view data but needs a depth estimator at inference, or drops this dependency but requires multi-view training pairs; OVIE is the first to require neither. Without multi-view supervision, OVIE rivals in-domain baselines on RealEstate10K and surpasses all on DL3DV; brief multi-view fine-tuning outperforms all geometry-free methods on their training domain. At 116 FPS, it is over 600× faster than the fastest baseline. Code and models will be open-sourced.

1. Introduction

Trillions of photographs document the world, yet monocular novel view synthesis, the task of generating what a scene looks like from a new camera pose given a single image, trains on a tiny fraction of them. The bottleneck is a data-type constraint: the task requires large volumes of video from static environments to extract posed multi-view pairs, a condition most of the visual world cannot meet. Paintings, historical photographs, and any dynamic scene simply have no multi-view counterpart. The community has converged on a handful of purpose-built datasets, such as RealEstate10K [80] and DL3DV [32], which cover only a narrow set of visual domains. Models trained on this data generalize within their training domains but fail under domain shift, and the reason is structural.

The natural approach to sidestep multi-view requirements synthesizes training pairs from single images via depth esti-

mation and reprojection under a sampled pose. Such pseudo-targets omit disoccluded regions, making naive supervision harmful. Methods that take this route, such as Infinite Nature-Zero [31], still require a depth model at inference, while geometry-free models [16, 53, 73] drop this dependency by training on posed multi-view datasets. No prior work achieves both.

We show this tradeoff can be broken with large-scale monocular pretraining. A frozen monocular depth estimator constructs training pairs from any single image by estimating metric depth, lifting the scene into a point cloud, and reprojecting under a sampled camera pose. A masked training formulation restricts losses to valid reprojected pixels; an adversarial objective covers disoccluded regions. Depth errors are absorbed by boundary masking, an L2 penalty that down-weights residuals, and training scale. At inference, the depth estimator is discarded and the model generates the novel view in a single forward pass (Figure 1).

Trained on 30 million in-the-wild images from ImageNet-21K [52], Places [78], Open Images [29], and OSV5M [1] without any multi-view supervision, OVIE already rivals specialized in-domain baselines on RealEstate10K [80] despite never training on it, and surpasses all baselines on DL3DV [32] where all methods face the same domain shift. A brief fine-tuning stage on RealEstate10K then suffices to surpass all geometry-free methods on pixel fidelity and perceptual similarity, establishing monocular pretraining as an efficient foundation for in-domain adaptation. OVIE generates novel views at 116 FPS on a single H100 GPU, more than 600× faster than the fastest baseline.

Our contributions are threefold:

- *We introduce a monocular pretraining paradigm that produces geometry-free models without multi-view supervision.* A masked formulation restricts losses to valid reprojected regions and an adversarial objective covers disoccluded areas, together making depth estimation a training scaffold discarded at inference.
- *We show that monocular pretraining is an efficient foundation for in-domain adaptation.* Brief fine-tuning on a single multi-view dataset surpasses all geometry-free in-

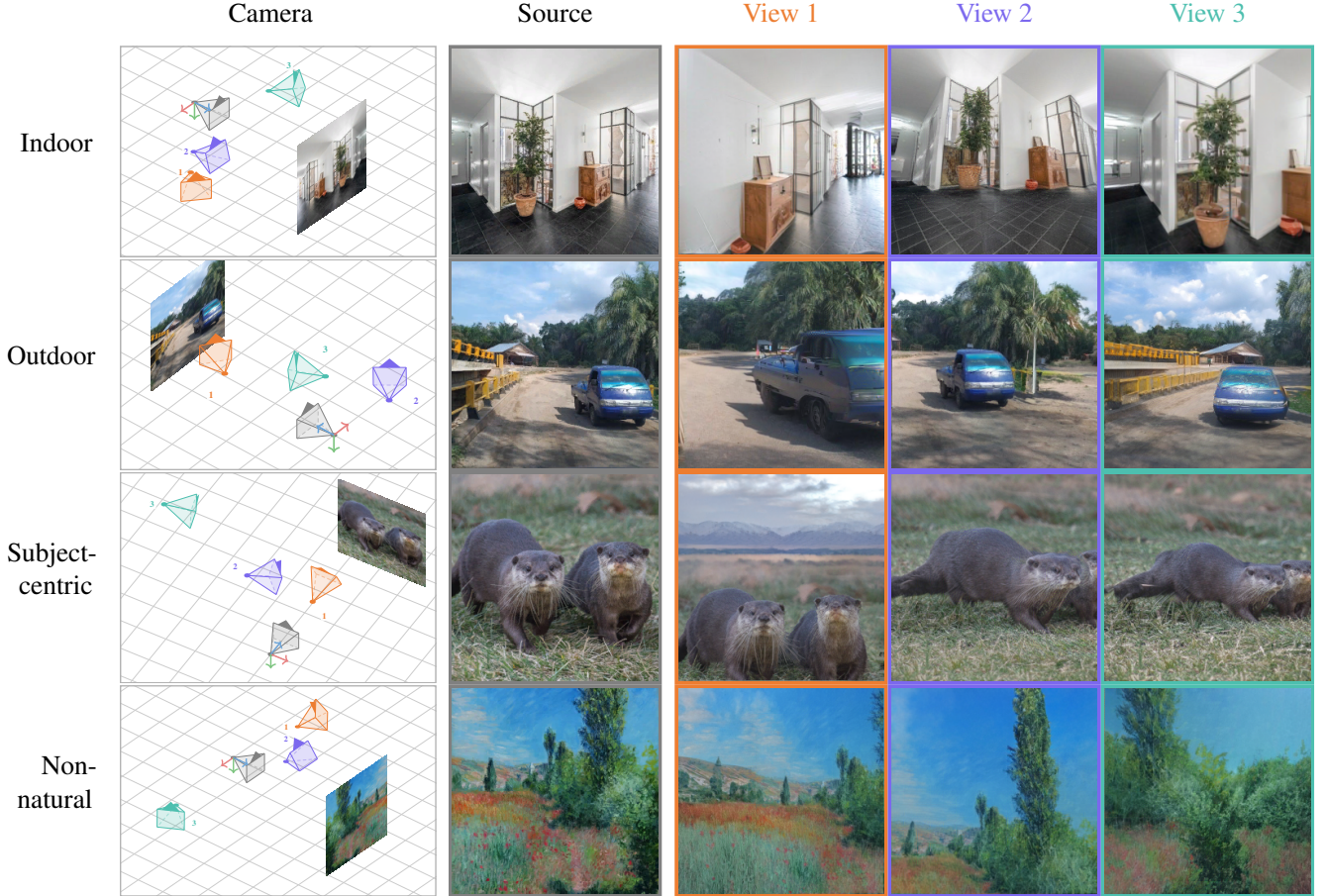


Figure 1. **OVIE generates novel views from a single image in real time** across diverse domains given a source image (gray) and target poses (colored), regardless of content or style.

domain methods, demonstrating effective transfer from large-scale unpaired pretraining.

- We characterize the effect of data scale and diversity on *monocular novel view synthesis*. Scale is the primary driver of performance, improving consistently from 3K to 30M training images; diverse mixing provides a complementary gain at fixed budget.

2. Related Work

Problem setting. Novel view synthesis spans tasks differing significantly in input, output, and generalization scope. Per-scene optimization methods such as NeRF [38] and Gaussian Splatting [28] fit a scene representation to tens or hundreds of posed views and generalize only within that scene. Multi-view feed-forward methods [8, 10, 11, 25, 64, 79] generalize across scenes but require multiple source images at inference. Feed-forward reconstruction methods [37, 60, 61, 70, 72] predict 3D representations from a single image, yet remain largely object-centric. We address *monocular scene-level novel view synthesis*: given a *single* image and a target pose, synthesize the target view in one feed-forward pass with a

model that generalizes across scenes and domains.

Monocular novel view synthesis with multi-view supervision. The dominant paradigm trains feed-forward models on posed multi-view collections, from which geometric correspondences can be extracted as supervision. SynSin [68] established this paired-data approach; NViST [24] scaled it to MVImgNet [75]. Geometry-free, pose-conditioned image-to-image models (SRT [54], GeoGPT [53], PhotoNVS [73], and VIVID [16]) represent the family most related to ours, synthesizing the target view directly in pixel space without an explicit 3D representation. All share a structural bottleneck: they require posed, static multi-view datasets (e.g., RealEstate10K [80], DL3DV [32], ScanNet [13]), which cover only a narrow visual domain. OVIE replaces this with monocular pretraining on 30 million single images from domains these benchmarks do not reach.

Monocular novel view synthesis without multi-view supervision. Learning novel view synthesis without posed pairs has been explored primarily through 3D-aware generative models trained on unposed image collections [6, 7, 19, 40, 41, 56]. Applying these to a real input requires test-time

GAN inversion [82], which is slow, per-image, and confined to the training category distribution; Ramirez et al. [47] and G3DR [50] build novel view synthesis frameworks on this paradigm but remain category-specific. Cycle-consistency approaches [83] do not generalize to 3D viewpoint change; Infinite Nature-Zero [31] extends this idea to view generation by coupling cycle consistency with monocular depth warping, but then requires explicit depth warping at inference and is limited to a single domain. Single-image supervision via 3D bounding box annotations [36] trades one scarce signal for another. All methods in this family share the same structural bottleneck: a depth estimator is load-bearing at inference. OVIE eliminates this through monocular pretraining on 30 million unconstrained single images, with no category prior and no depth estimator at inference. Additionally, we show that models trained under this constraint, scaled to large in-the-wild collections, reach state-of-the-art geometry-free performance with minimal fine-tuning on multi-view data.

Large generative priors for novel view synthesis. An alternative strategy adapts large pretrained generative models to novel view synthesis via fine-tuning on posed multi-view data: Zero-1-to-3 [34] adapts Stable Diffusion [18] on Objaverse [14] for object-level view synthesis; ViewCrafter [74] conditions DynamiCrafter [69] on DUS3R [67] renders for scene-level pose control; Stable Virtual Camera [79] fine-tunes a video diffusion model with Plücker ray conditioning; and PE-Field [2] adapts Flux.1 Kontext [30] with depth-aware 3D encodings. These methods use a proxy pretraining task (image or video generation) and inherit the large, compute-heavy architectures of their base models, making inference expensive by design. OVIE takes a different route: pretraining directly on the novel view synthesis task, scaled to 30 million in-the-wild images, with a compact feed-forward architecture. Task-aligned pretraining proves more efficient: OVIE runs at 116 FPS, more than $7,000\times$ faster than ViewCrafter, while surpassing it on RealEstate10K.

Monocular depth estimation. Learning-based depth estimation splits into two branches. The *relative* branch (MiDaS [4, 49], DPT [48], Depth Anything [71], Marigold [26, 27], MoGe [65]) yields affine-invariant predictions. The *metric* branch (ZoeDepth [3], UniDepth [45, 46], MoGe-2 [66]) also recovers absolute scale. OVIE exploits this: metric depth lets us build pseudo-pairs with metric translations at training time, providing geometric supervision without manual annotation.

Warping-based novel view synthesis. Warping-based methods unproject the source image into 3D and inpaint disoccluded regions via geometry-based pop-up models [20, 21], layered depth inpainting [42, 58], soft point-cloud rendering [23, 68], MPI blending [62, 81], or diffusion-based pipelines such as GenWarp [57] and LucidDreamer [12]. Infi-

nite Nature [33] extends this to perpetual view generation but relies on posed video sequences for training. MultiDiff [39] conditions a video diffusion model on depth-warped inputs, tying quality to depth accuracy. Because the depth estimator is load-bearing at inference and pipelines are iterative, these methods are fragile to depth failures and computationally slow; OVIE requires no depth estimate at inference.

Geometry-free methods [16, 53, 73] require multi-view data from the start; methods without multi-view supervision, such as Infinite Nature-Zero [31], still need depth at inference. No prior work has explored monocular pretraining as an initialization for geometry-free novel view synthesis: training on large unpaired image collections to produce a depth-free model that transfers to competitive performance via brief multi-view fine-tuning. OVIE fills this gap through a masked training formulation that confines supervision to valid reprojected pixels and covers disoccluded regions adversarially, reducing depth to a training scaffold discarded at inference.

3. Method

Given a source image I_0 and a camera transformation $T_{0\rightarrow 1} \in SE(3)$, OVIE generates a target view $\hat{I}_1$. Monocular pretraining on large single-image collections builds on two components: on-the-fly pseudo-pair generation and partial supervision. A frozen monocular depth network lifts each source image into 3D; reprojecting the point cloud under sampled camera poses yields sparse pseudo-targets I_1^* with binary validity masks M , which a pose-conditional image-to-image network f_θ uses as ground truth (Fig. 2, Fig. 3). At inference, OVIE discards the depth estimator and synthesizes the novel view in a single forward pass. The pretrained model transfers to a specific domain via brief fine-tuning on posed multi-view data.

3.1. Annotation-Free Training Pair Construction

Depth-Based Scene Lifting. A pretrained monocular depth network processes the source image I_0 , estimating a depth map $D \in \mathbb{R}^{H \times W}$ and surface normals $N \in \mathbb{R}^{H \times W \times 3}$. Together, these quantities define a point cloud $\mathcal{P} \in \mathbb{R}^{HW \times 3}$ in the source camera coordinate system via standard unprojection.

Viewpoint Sampling and Reprojection. We sample $T_{0\rightarrow 1} \in SE(3)$ from a mixture of six strategies assigned stochastically: identity, pure translation, pure rotation, combined rotation and translation, normal-derived (camera placed along an estimated surface normal), and frontal hemisphere (camera orbiting an anchor point). Translation magnitudes scale with median scene depth; rotations are bounded by the estimated field of view to maintain sufficient overlap. Reprojecting the transformed $\mathcal{P}$ yields a pseudo-ground-truth view I_1^* and a binary mask $M \in \{0, 1\}^{H \times W}$, where $M_{ij} = 0$ marks

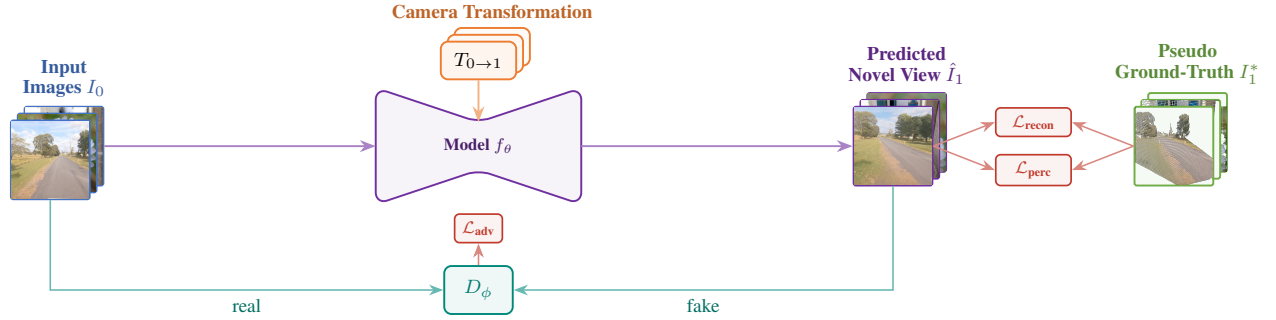
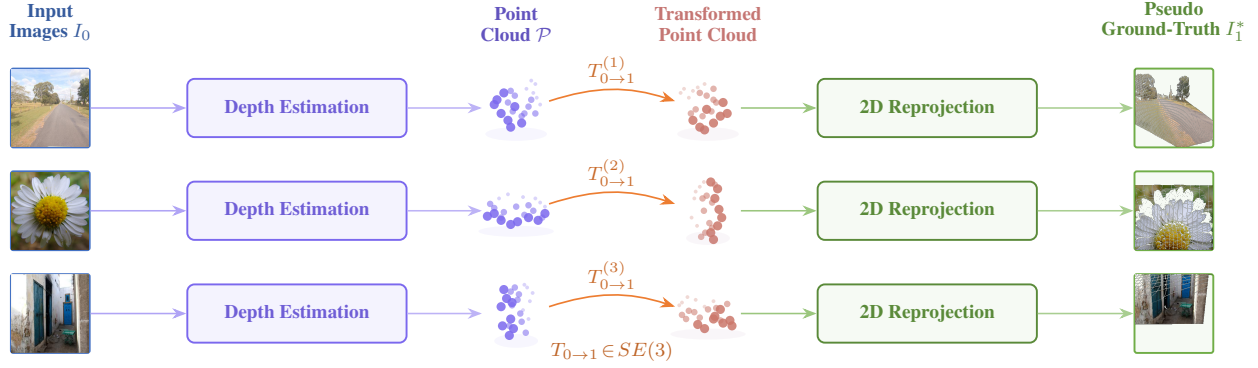


Figure 2. **Monocular pretraining method overview.** **Top:** A frozen depth estimator lifts web-sourced images I_0 into point clouds $\mathcal{P}$; sampled transformations $T_{0 \rightarrow 1} \in SE(3)$ and reprojection yield pseudo-target views I_1^* . **Bottom:** Feed-forward model f_θ takes I_0 conditioned on $T_{0 \rightarrow 1}$ and predicts $\hat{I}_1$. Training uses a masked reconstruction loss $\mathcal{L}_{\text{recon}}$ and a perceptual loss $\mathcal{L}_{\text{perc}}$ to compare I_1^* and $\hat{I}_1$, and an adversarial loss $\mathcal{L}_{\text{adv}}$ where discriminator model D_ϕ distinguishes I_0 from $\hat{I}_1$.

disocclusions, backface-culled regions, and out-of-frame content. Full details and sampling weights appear in the supplementary material (Section C).

Metric-Scale Supervision. When the depth model is metric (e.g., MoGE-2 [66]), the resulting pairs encode metric-scale changes, a supervision signal far scarcer than standard pose annotations.

3.2. Training Objective

We use a multi-term objective that enforces geometric accuracy, semantic consistency, and textural realism. Because the pseudo-target I_1^* contains missing content, the reconstruction and perceptual losses are restricted to observed regions via the mask M ; the adversarial loss addresses disoccluded regions through a different mechanism, described below.

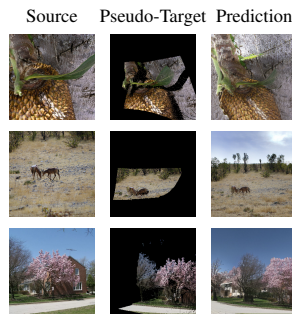


Figure 3. **Pseudo-pairs and predictions.**

Geometric Consistency. The main supervision is a masked reconstruction loss. Depth estimation errors propagate into the pseudo-targets, but three complementary mechanisms limit their impact: the mask M eliminates gross failures at depth discontinuities and occlusion boundaries; the L2 penalty is everywhere differentiable, providing a stable gradient signal even near masked boundaries; and training on 30 million images ensures no erroneous target dominates the gradient signal. The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \frac{\|M \odot \hat{I}_1 - M \odot I_1^*\|_2^2}{\|M\|_1 + \epsilon}, \quad (1)$$

where $\epsilon > 0$ prevents division by zero.

Semantic Preservation. We mask both the prediction and target prior to feature extraction, restricting the loss to valid regions. Following prior work [35], we combine LPIPS [76] with P-DINO, a patch-level loss derived from activations of a pretrained DINO model [5, 43, 59]:

$$\begin{aligned} \mathcal{L}_{\text{perc}} = & \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}(M \odot \hat{I}_1, M \odot I_1^*) \\ & + \lambda_{\text{DINO}} \mathcal{L}_{\text{P-DINO}}(M \odot \hat{I}_1, M \odot I_1^*), \end{aligned} \quad (2)$$

where $\mathcal{L}_{\text{P-DINO}}(\mathbf{x}, \mathbf{y}) = 1 - \cos(\text{DINO}(\mathbf{x}), \text{DINO}(\mathbf{y}))$, and $\lambda_{\text{LPIPS}}, \lambda_{\text{DINO}} > 0$ are scalar weights balancing the two perceptual terms.

Disocclusion Supervision and High-Frequency Realism. In disoccluded regions ($M_{ij} = 0$), neither $\mathcal{L}_{\text{recon}}$ nor $\mathcal{L}_{\text{perc}}$ contributes a gradient, so the model must hallucinate plausible content. A PatchGAN adversarial objective [22] covers this gap: because the incomplete I_1^* cannot serve as a reliable real sample, I_0 acts as the real distribution instead, an approximation that holds since viewpoint sampling bounds parallax and keeps patch statistics consistent across views:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{I_0}[\log D(I_0)] + \mathbb{E}_{\hat{I}_1}[\log(1 - D(\hat{I}_1))]. \quad (3)$$

The discriminator D_ϕ follows the StyleGAN-T [55] design, using a pretrained representation backbone. The adaptive weight w_{adap} , introduced in VQ-GAN [17], balances reconstruction and adversarial terms automatically. Details appear in the supplementary material (Section D.1).

Total Objective. The total loss combines all three terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{perc}} + \lambda_a \cdot w_{\text{adap}} \cdot \mathcal{L}_{\text{adv}}, \quad (4)$$

where $\lambda_a > 0$ controls the contribution of the adversarial term. All hyperparameters are specified in the supplementary material (Section D.3).

Fine-Tuning. When fine-tuning on a posed multi-view dataset, the real target view I_1 replaces the pseudo-target: $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{perc}}$ are computed against I_1 without masking, and I_1 also serves as the real sample for the discriminator, replacing I_0 . The objective and architecture remain unchanged.

3.3. Model Architecture

A convolutional encoder downsamples the source image I_0 into a dense feature map; a stack of pose-conditioned Transformer blocks then processes this representation; a mirrored convolutional decoder finally upsamples it back to the source resolution, producing $\hat{I}_1$.

Pose Conditioning. We encode the relative transformation $T_{0 \rightarrow 1}$ as a 7D vector $\mathbf{p} \in \mathbb{R}^7$ (3D translation and unit quaternion), following VGGT [63] but omitting camera intrinsics: MoGe-2 [66] supplies intrinsic estimates at training time, and camera geometry is read implicitly from the source image’s perspective cues, with 30 million diverse training images providing robustness to intrinsic variation at inference. We omit intrinsics from the conditioning vector to keep the inference interface minimal. A linear layer projects $\mathbf{p}$ into a conditioning token $\mathbf{c} = W\mathbf{p} \in \mathbb{R}^d$, which modulates each Transformer block via Adaptive Layer Normalization (AdaLN) [44].

4. Experiments

We evaluate two model variants: OVIE, trained exclusively on unpaired single images via monocular pretraining, and OVIE-ft, fine-tuned on RealEstate10K from the pretrained checkpoint. Evaluation covers five axes: qualitative generalization (Figure 4), quantitative comparison on RealEstate10K, where geometry-free baselines are in-domain and OVIE is out-of-domain, and on DL3DV, where all geometry-free methods face the same domain shift (Table 1), throughput showing OVIE runs at real-time speeds more than $600\times$ faster than the fastest baseline (Figure 6), and ablations of loss design and training data (Table 2 and Figure 7).

4.1. Experimental Setup

Training Data. OVIE is trained on 30 million in-the-wild images from ImageNet-21K [52], Open Images [29], OpenStreetView5M (OSV5M) [1], and Places [78], with no multi-view supervision; MoGe-2 [66] generates pseudo-pairs on-the-fly. OVIE-ft is fine-tuned on RealEstate10K [80].

Benchmarks and Metrics. We evaluate on RealEstate10K [80] and DL3DV [32]; OVIE has never been trained on either. PSNR and SSIM measure pixel-level fidelity, LPIPS [76] perceptual similarity, and FID distributional realism. Following prior work [2, 51], we sample a starting frame from each of 750 scenes, generate 14 novel views at a stride of 3 (RealEstate10K) or 1 (DL3DV), and average metrics across scenes, with FID computed over all generated and ground-truth images. Since benchmark poses are derived from SLAM and are scale-ambiguous, we perform a per-scene scale sweep for each method, selecting the value that maximizes its own performance.

Baselines. OVIE is compared against GeoGPT [53], PhotoNVS [73], and VIVID [16], geometry-free image-to-image methods sharing our problem formulation, all trained exclusively on RealEstate10K. This creates a deliberate asymmetry: these baselines are *in-domain* on RealEstate10K while OVIE is *out-of-domain*; OVIE-ft, fine-tuned on RealEstate10K, is *in-domain*. All geometry-free methods are out-of-domain on DL3DV, enabling a fair comparison. ViewCrafter [74] is included as a geometry-based reference: it offloads geometric reasoning to DUST3R [67] and builds on a large video diffusion foundation model ($\sim 10\times$ more parameters), trained on both benchmarks. An additional comparison with Infinite Nature-Zero [31], a geometry-based method trained with single images, appears in the supplementary material (Section G).

Implementation Details. OVIE uses a convolutional encoder, a ViT-B bottleneck [15], and a convolutional decoder, trained for 2M steps at batch size 512; ablations use 250K steps. Full architectural details and hyperparameters are in the supplementary material (Section D).

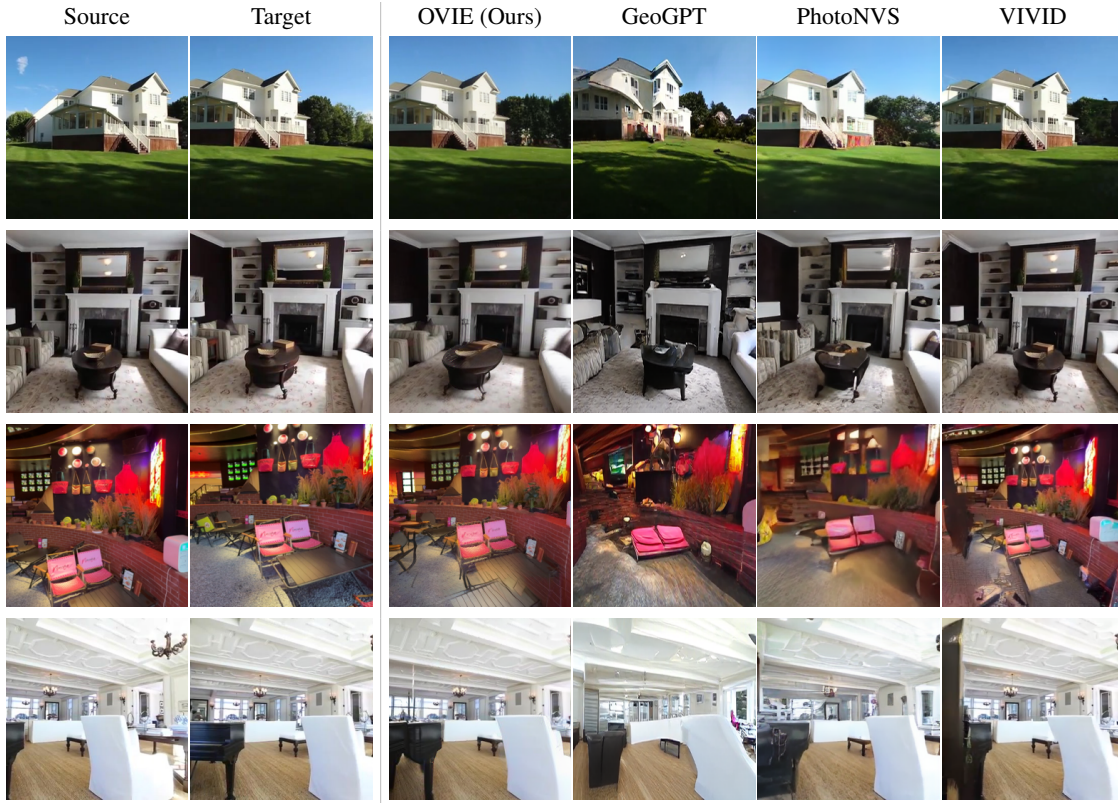


Figure 4. **Qualitative comparison with baseline geometry-free methods.** Pretrained on single images only, OVIE produces sharp, geometrically consistent novel views.

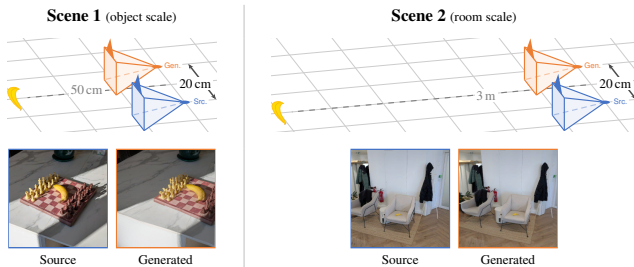


Figure 5. **Metric scale awareness.** The same translation applied to a close-up scene (left) produces a larger apparent displacement than on a room-scale scene (right), consistent with metric parallax.

4.2. Qualitative Results

Figure 1 shows OVIE generalizing across indoor, outdoor, object-centric, and non-photographic scenes. Figure 4 compares OVIE against baseline methods: despite training on none of their evaluation data, OVIE matches or exceeds their quality. Figure 5 qualitatively illustrates metric scale awareness inherited from MoGe-2: the same camera translation produces larger apparent displacement for nearby objects than distant ones.

4.3. Comparison with State of the Art

RealEstate10K: competing at a disadvantage. OVIE outperforms three of the four in-domain baselines (Table 1), despite never training on RealEstate10K data; the remaining gap with VIVID reflects the domain disadvantage, as DL3DV confirms. Fine-tuning OVIE for 2% of the pre-training budget (OVIE-ft) achieves state of the art among geometry-free methods, demonstrating that single-image pretraining provides an efficient foundation for in-domain adaptation. OVIE-ft leads all baselines on pixel fidelity and perceptual similarity, over $600\times$ faster than any competitor.

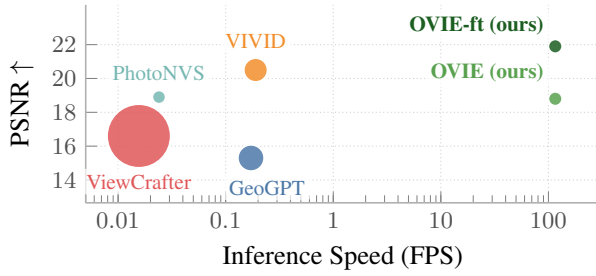
DL3DV: outperforming all baselines under domain shift. When all methods face the same domain shift on DL3DV, OVIE outperforms all baselines on all metrics. This result shows that the diversity of OVIE’s large-scale training data delivers robustness to distribution shifts that in-domain specialization cannot match. Notably, OVIE outperforms OVIE-ft on DL3DV across all metrics, confirming that fine-tuning on a single domain trades generalization for in-domain performance.

4.4. Real-Time Novel View Synthesis

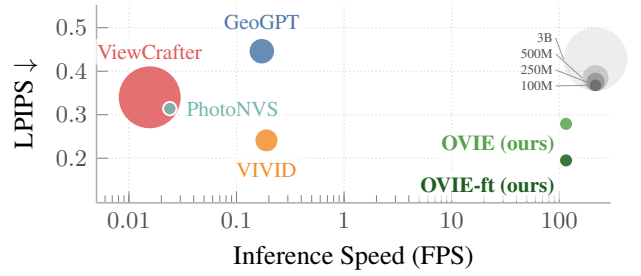
Figure 6 plots throughput against generation quality (PSNR and LPIPS) on a single H100 GPU on RealEstate10K. OVIE

Table 1. **In-distribution and out-of-domain robustness on RealEstate10K and DL3DV.** $\uparrow/\downarrow$: higher/lower is better. **Bold**: best; underline: second best. OOD: not trained on the benchmark. *Trained in-domain on DL3DV: scores greyed out and excluded from rankings for fairness.

Method	RealEstate10K [80]					DL3DV [32]				
	OOD	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	OOD	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
GeoGPT [53]	$\times$	15.3	0.480	0.446	18.0	$\checkmark$	13.1	0.339	0.560	35.9
ViewCrafter* [74]	$\times$	16.6	0.559	0.340	9.98	$\times$	15.4	0.419	0.425	10.50
PhotoNVS [73]	$\times$	18.9	0.601	0.314	10.6	$\checkmark$	13.8	0.349	0.525	37.6
VIVID [16]	$\times$	<u>20.5</u>	<u>0.661</u>	<u>0.241</u>	4.26	$\checkmark$	<u>14.5</u>	<u>0.362</u>	<u>0.471</u>	<u>18.0</u>
OVIE (ours)	$\checkmark$	18.8	0.602	0.279	6.74	$\checkmark$	14.8	0.369	0.464	13.6
OVIE-ft (ours)	$\times$	21.9	0.695	0.195	<u>5.59</u>	$\checkmark$	14.1	0.351	0.531	24.66



(a) PSNR vs. FPS (upper-right is better)



(b) LPIPS vs. FPS (lower-right is better)

Figure 6. **Quality vs. Inference speed tradeoff on RealEstate10K.** Bubble radius indicates parameter count. OVIE-ft achieves competitive quality at drastically higher FPS than baselines.

Table 2. **Loss ablation studies on RealEstate10K and DL3DV.** Each group varies one design axis while keeping all others at the default configuration (**bold**).

			RealEstate10K [80]				DL3DV [32]			
P-DINO	LPIPS	GAN	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Loss component ablation										
✓	✓	✓	18.9	0.596	0.284	7.12	15.0	0.373	0.468	14.3
	✓	✓	19.0 ^{+0.1}	0.599 ^{+0.003}	0.288 ^{+0.004}	8.34 ^{+1.22}	15.1 ^{+0.1}	0.377 ^{+0.004}	0.472 ^{+0.004}	15.7 ^{+1.4}
✓		✓	18.7 ^{−0.2}	0.592 ^{−0.004}	0.297 ^{+0.013}	8.43 ^{+1.31}	14.9 ^{−0.1}	0.371 ^{−0.002}	0.478 ^{+0.010}	<u>15.3</u> ^{+1.0}
		✓	18.7 ^{−0.2}	0.584 ^{−0.012}	0.367 ^{+0.083}	18.7 ^{+11.6}	14.9 ^{−0.1}	0.368 ^{−0.005}	0.540 ^{+0.072}	27.0 ^{+12.7}
✓	✓		<u>19.2</u> ^{+0.3}	<u>0.598</u> ^{+0.002}	0.301 ^{+0.017}	13.4 ^{+6.28}	<u>15.4</u> ^{+0.4}	0.375 ^{+0.002}	0.496 ^{+0.028}	48.5 ^{+34.2}
			19.6 ^{+0.7}	0.627 ^{+0.031}	0.416 ^{+0.132}	62.1 ^{+55.0}	15.7 ^{+0.7}	0.441 ^{+0.068}	0.627 ^{+0.159}	111.0 ^{+96.7}
Reconstruction loss										
L2			18.9	0.596	0.284	7.12	15.0	0.373	0.468	14.3
L1			18.5 ^{−0.4}	0.594 ^{−0.002}	0.297 ^{+0.013}	8.57 ^{+1.45}	14.5 ^{−0.5}	0.367 ^{−0.006}	0.477 ^{+0.009}	14.3 ^{−0.0}
Charbonnier			18.5 ^{−0.4}	0.592 ^{−0.004}	0.296 ^{+0.012}	8.35 ^{+1.23}	14.5 ^{−0.5}	0.367 ^{−0.006}	0.476 ^{+0.008}	14.1 ^{−0.2}

achieves 116 FPS (8.6 ms), versus 0.19 FPS for VIVID (50 diffusion steps), 0.17 FPS for GeoGPT (autoregressive), 0.024 FPS for PhotoNVS (2000 diffusion steps), and 0.016 FPS for ViewCrafter, which additionally requires DUST3R reconstruction and video diffusion sampling (50 steps). OVIE runs more than 600 $\times$ faster than the fastest

baseline. OVIE-ft matches this throughput while surpassing all baselines on RealEstate10K. With keyboard and mouse inputs mapped to extrinsic changes, this enables real-time scene navigation (see supplementary material, Section F.4).

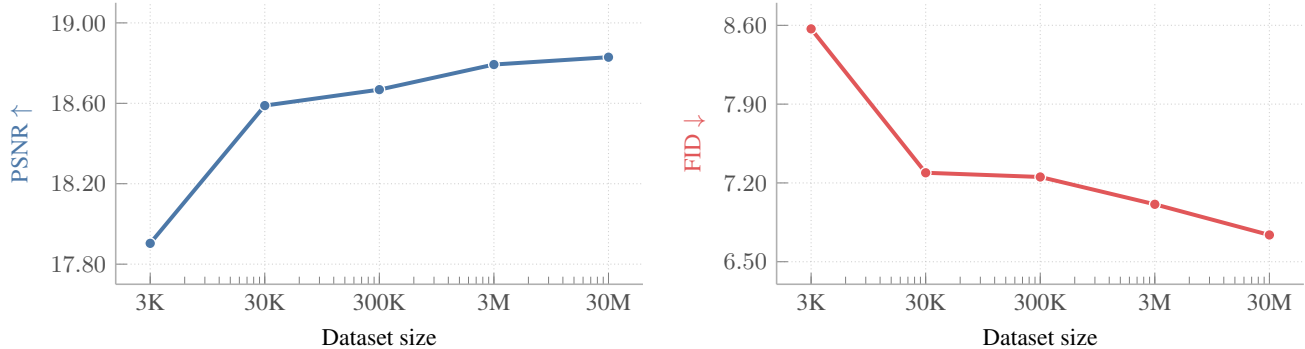


Figure 7. **Scaling with dataset size.** PSNR and FID improve as data volume increases.

Table 3. **Comparison of data coverage on model performance** on RealEstate10K and DL3DV. **Bold**: best; underline: second best. Differences are relative to the Mix baseline.

Dataset	Domain	RealEstate10K [80]				DL3DV [32]			
		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Mix	Mixed	18.8	<u>0.595</u>	0.284	7.08	15.0	0.372	0.467	14.2
OSV5M [1]	Street View	<u>18.2</u> -0.6	0.566 -0.029	0.318 $+0.034$	8.59 $+1.51$	15.0	0.369 -0.003	0.481 $+0.014$	18.9 $+4.7$
ImageNet21K [52]	Objects	18.8	0.593 -0.002	0.289 $+0.005$	7.32 $+0.24$	<u>14.9</u> -0.1	<u>0.370</u> -0.003	0.472 $+0.005$	15.9 $+1.7$
Places [78]	Scenes	18.8	0.596 $+0.001$	<u>0.286</u> $+0.002$	7.41 $+0.33$	14.8 -0.2	0.367 -0.006	0.477 $+0.010$	20.3 $+6.1$
OpenImages [29]	General	18.8	0.593 -0.002	0.287 $+0.003$	<u>7.20</u> $+0.12$	<u>14.9</u> -0.1	<u>0.370</u> -0.003	<u>0.469</u> $+0.002$	<u>14.8</u> $+0.6$

4.5. Ablation Studies

Loss terms. Table 2 shows each term’s distinct role. Keeping only $\mathcal{L}_{\text{recon}}$ maximizes PSNR and SSIM while driving FID to 62.1 / 111.0, illustrating why pixel metrics alone are insufficient. P-DINO and LPIPS contribute complementary perceptual supervision: removing both raises FID to 18.7 / 27.0. The adversarial loss is critical for generalization, with its impact especially pronounced out-of-domain (FID +34.2 on DL3DV vs. +6.3 on RealEstate10K).

Reconstruction loss. L2 outperforms L1 and Charbonnier [9] across both benchmarks, despite the intuition that robust losses better handle erroneous pseudo-targets.

4.6. Data Scaling and Diversity

Effect of data scale. Performance improves rapidly in the first decade of data (3K to 30K images) and continues across each subsequent decade (Figure 7), well beyond the scale of dedicated multi-view datasets such as RealEstate10K and DL3DV (10K scenes).

Effect of data diversity. To isolate diversity from scale, we train four single-source models and a mixed model on 2M images each. The mixed model matches or exceeds the best single source on both benchmarks (Table 3), confirming that diverse mixing yields a modest but reliable gain.

5. Conclusion

We introduce monocular pretraining for novel view synthesis: a frozen depth estimator constructs pseudo-pairs from any single image, masked losses restrict supervision to valid regions, and an adversarial objective covers disoccluded areas, reducing depth to a training scaffold discarded at inference. Trained on 30 million uncured internet images without multi-view supervision, OVIE rivals specialized in-domain baselines under domain shift at 116 FPS. Brief fine-tuning suffices to surpass all geometry-free methods, establishing monocular pretraining as a strong initialization for novel view synthesis. Internet-scale single-image collections, long underexplored for this task, are a viable foundation for the field. Domains that produce no multi-view counterpart, such as historical photographs or paintings, are now viable training sources.

Limitations. Performance is bounded by the depth estimator: failures on reflective surfaces, transparent materials, or low-light conditions propagate as missing regions that masking cannot recover. Disoccluded content is hallucinated adversarially, which is perceptually plausible but not geometrically grounded. OVIE is designed for moderate viewpoint changes.

Acknowledgments. We thank Robin Courant for proof-reading, and Eloi Alonso, Mathieu Aubry, Antoine Guédon, Anthony Hu, Loïc Landrieu, Vincent Lepetit, Vincent Micheli, Manu Orsini, Amélie Royer, and Václav Volhejn for interesting discussions.

References

- [1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronsohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. OpenStreetView-5M: The many roads to global visual geolocation. *CVPR*, 2024.
- [2] Yunpeng Bai, Haoxiang Li, and Qixing Huang. Positional encoding field. *arXiv*, 2025.
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023.
- [4] Reiner Birkel, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv*, 2023.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
- [6] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021.
- [7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022.
- [8] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.
- [9] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 1997.
- [10] Yuedong Chen, Haoifei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv*, 2024.
- [11] Yuedong Chen, Chuanxia Zheng, Haoifei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS*, 2024.
- [12] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Lucidreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv*, 2023.
- [13] 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.
- [14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *CVPR*, 2023.
- [15] 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. *ICLR*, 2021.
- [16] Noam Elata, Bahjat Kavar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. *CVPR*, 2025.
- [17] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.
- [18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, 2024.
- [19] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: a generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022.
- [20] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, 2005.
- [21] Ronghang Hu, Nikhila Ravi, Alexander C. Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3D sheet for video synthesis from a single image. In *ICCV*, 2021.
- [22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.
- [23] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *ICCV*, 2021.
- [24] Wonbong Jang and Lourdes Agapito. Nvst: In the wild new view synthesis from a single image with transformers. In *CVPR*, 2024.
- [25] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2025.
- [26] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.
- [27] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE TPAMI*, 2025.
- [28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *SIGGRAPH*, 2023.
- [29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

- [30] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv*, 2025.
- [31] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *ECCV*, 2022.
- [32] 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 *CVPR*, 2024.
- [33] 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*, 2021.
- [34] 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.
- [35] Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss. *arXiv*, 2026.
- [36] Léopold Maillard, Tom Durand, Adrien Ramanana Rahary, and Maks Ovsjanikov. Laconic: A 3d layout adapter for controllable image creation. In *ICCV*, 2025.
- [37] Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, Yanghai Tsing, Stephan R. Richter, and Vladlen Koltun. Sharp monocular view synthesis in less than a second. In *ICLR*, 2026.
- [38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [39] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kotschieder. Multidiff: Consistent novel view synthesis from a single image. In *CVPR*, 2024.
- [40] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019.
- [41] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021.
- [42] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 2019.
- [43] 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. Di-nov2: Learning robust visual features without supervision. *arXiv*, 2023.
- [44] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *ICCV*, 2023.
- [45] 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.
- [46] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeels, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *arXiv*, 2025.
- [47] Pierluigi Zama Ramirez, Alessio Tonioni, and Federico Tombari. Unsupervised novel view synthesis from a single image. *arXiv*, 2021.
- [48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.
- [49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2022.
- [50] Pradyumna Reddy, Ismail Elezi, and Jiankang Deng. G3dr: Generative 3d reconstruction in imagenet. *CVPR*, 2024.
- [51] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025.
- [52] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv*, 2021.
- [53] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *ICCV*, pages 14356–14366, 2021.
- [54] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. In *CVPR*, 2022.
- [55] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: unlocking the power of gans for fast large-scale text-to-image synthesis. In *Proc. ICML*, 2023.
- [56] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [57] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *arXiv*, 2024.
- [58] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020.

- [59] 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. *arXiv*, 2025.
- [60] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, 2024.
- [61] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *3DV*, 2025.
- [62] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020.
- [63] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.
- [64] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021.
- [65] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025.
- [66] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025.
- [67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.
- [68] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020.
- [69] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2025.
- [70] Haoifei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depth-splat: Connecting gaussian splatting and depth. In *CVPR*, 2025.
- [71] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.
- [72] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021.
- [73] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023.
- [74] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, 2024.
- [75] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023.
- [76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [77] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.
- [78] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv*, 2016.
- [79] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishtha, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv*, 2025.
- [80] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *SIGGRAPH*, 2018.
- [81] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 2018.
- [82] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.
- [83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

A. Summary of Supplementary Material

This supplementary document provides additional details and results that complement the main paper. It is organized as follows:

- **Section B – Additional Quantitative Results.** We report supplementary SSIM and LPIPS curves for the data-scaling (B.1) and SSIM and FID curves for the throughput analysis (B.2).
- **Section C – Camera Sampling Details.** We describe the distribution from which relative camera transformations are sampled during training, including the parameterization (C.1, C.2), the geometry-aware reprojection (C.3) and pseudo-views sampling hyperparameters (C.4).
- **Section D – Additional Implementation Details.** We provide extended details on architecture (D.1), pose-conditioning (D.2), and training details and hyperparameters (D.3).
- **Section E – Training Resources and Broader Impact.** We report compute requirements and discuss the broader societal and environmental implications of the work.
- **Section F – Additional Qualitative Results.** We illustrate OVIE’s robust generalization by synthesizing additional novel views from diverse out-of-distribution images, including non-realistic source inputs such as paintings (F.1). We compare OVIE-generated views against their corresponding pseudo-ground-truth supervision targets (F.2). We present more qualitative comparisons on RealEstate10K [80], contrasting OVIE’s outputs with the source, the ground-truth novel view, and baseline methods (F.3). To highlight generation consistency and responsiveness, we provide animated side-by-side navigations (F.4) and real-time interactive screen recordings, driven by mouse and keyboard inputs, all generated continuously from a single initial frame (F.5).
- **Section G – Additional Comparison to InfiniteNature-Zero [31].** We compare OVIE to InfiniteNature-Zero. While omitted from the baselines because its inference requires geometry-based depth estimation and sky segmentation, it shares OVIE’s approach of training entirely on single-view images.

B. Additional Quantitative Results

B.1. Effect of Dataset Size on SSIM and LPIPS

Figure 8 shows SSIM and LPIPS as functions of dataset size, confirming the trends observed for PSNR and FID in Figure 7 of the main paper: both metrics improve consistently with more training data.

B.2. Throughput analysis

Figure 9 demonstrates that OVIE is $600\times$ faster than the fastest baseline while also achieving improved SSIM and FID scores on RealEstate10K [80]. This complements the

PSNR and LPIPS results discussed in Figure 6 of the main paper.

B.3. Domain-specific finetuning on DL3DV

In the main paper we report OVIE-ft, the base model finetuned on RealEstate10K [80]. For completeness, we apply the identical protocol to finetune the base OVIE model on DL3DV [32]. Finetuning markedly improves in-domain reconstruction on DL3DV relative to the base model, raising PSNR from 14.8 to 17.07, SSIM from 0.369 to 0.433, and lowering LPIPS from 0.464 to 0.368; FID, however, rises from 13.6 to 17.24, suggesting that domain-specific finetuning favors faithful per-view reconstruction at a slight cost to global distributional realism. More notably, this DL3DV-finetuned variant retains strong performance on RealEstate10K even though that benchmark is now out-of-distribution: it matches or slightly surpasses the base model on every reconstruction metric (PSNR 19.34 vs. 18.8, SSIM 0.608 vs. 0.602, LPIPS 0.260 vs. 0.279), with only a marginal increase in FID (7.14 vs. 6.74). This indicates that finetuning sharpens in-domain fidelity without eroding, and in this case slightly improving, cross-domain generalization.

C. Camera Sampling Details

During training, we generate novel views of the input image to serve as pseudo-ground-truth targets by reprojecting the scene’s 3D point cloud into newly sampled camera viewpoints. To achieve this, a monocular depth estimator first predicts both the absolute depth and the camera’s horizontal field of view (Θ_h), together enabling reconstruction of the point cloud in a true metric, real-world scale. Because this geometry possesses accurate physical dimensions, the distributions used to sample new camera poses can be defined directly in actual scene units (e.g., meters) rather than an arbitrary coordinate space. This ensures that camera displacements and distances scale consistently with the specific geometry of each scene. Next, a routing module stochastically assigns each batch element to one of six sampling methods based on fixed prior weights. All methods ultimately produce a world-to-camera extrinsic matrix $[\mathbf{R} \mid \mathbf{t}] \in \mathbb{R}^{3 \times 4}$.

Throughout this section, we denote continuous uniform and normal distributions as $\mathcal{U}$ and $\mathcal{N}$, respectively, and let $\hat{\mathbf{y}} = [0, 1, 0]^\top$ and $\hat{\mathbf{z}} = [0, 0, 1]^\top$ represent the canonical up and forward directional unit vectors.

C.1. Sampling methods.

The six strategies span a range of transformation types, from trivial (identity) to geometry-grounded (normal-derived, frontal hemisphere), ensuring the model is trained on diverse yet plausible viewpoint changes.

- **Identity.** No transformation is applied. The extrinsic is set to $[\mathbf{I} \mid \mathbf{0}]$.

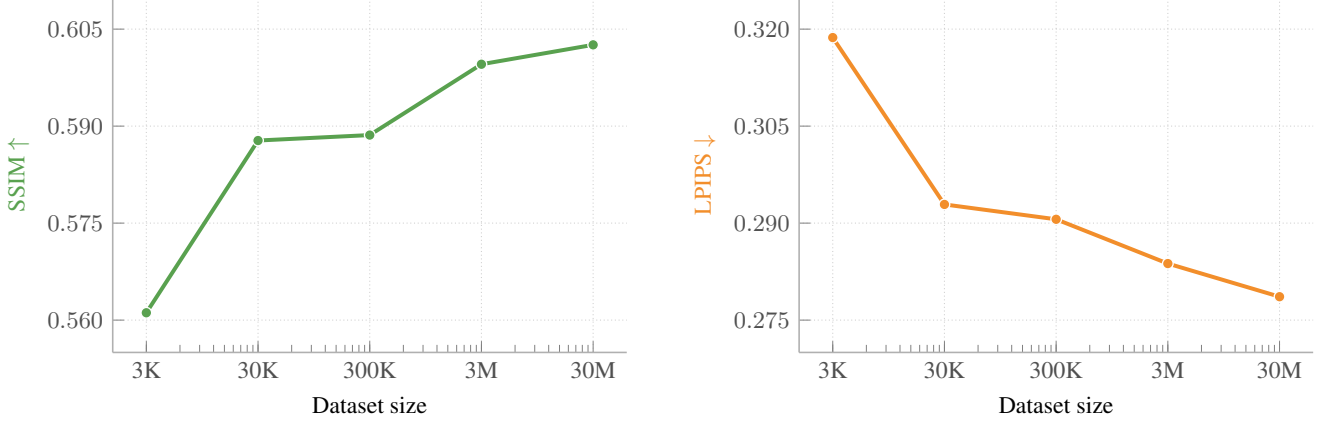


Figure 8. **Scaling with dataset size – SSIM and LPIPS.** Complementary to Figure 7 in the main paper, SSIM and LPIPS on RealEstate10K [80] follow the same monotonic improvement as data volume increases.

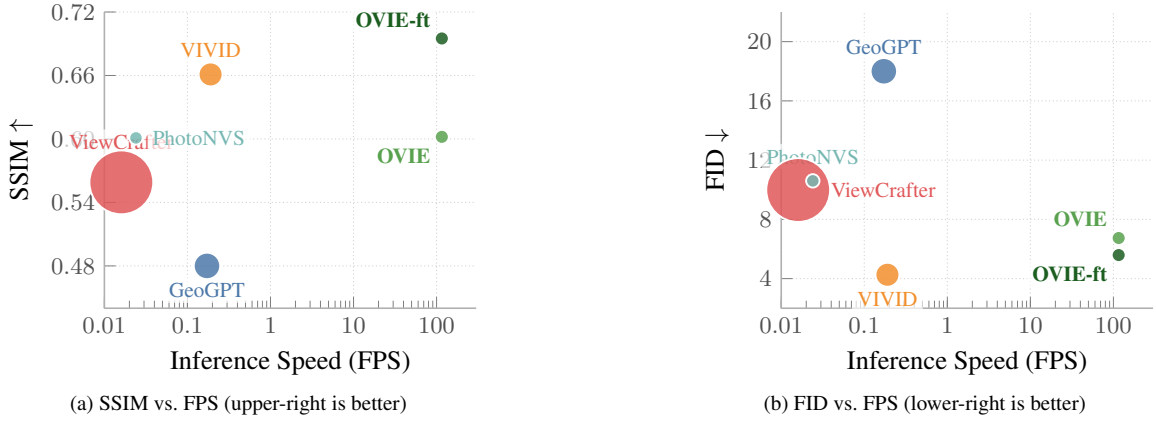


Figure 9. **Quality vs. Inference tradeoff on RealEstate10K – SSIM and FID.** Complementary to Figure 6 in the main paper. Bubble radius indicates parameter count. The same trend holds: OVIE is faster and better-performing than baseline methods.

- **Pure translation.** The camera is shifted relative to the scene without any rotation, with the shift magnitude tied to the spatial extent of the point cloud. The rotation is fixed to $\mathbf{R} = \mathbf{I}$. The translation is sampled as $\mathbf{t} \sim \mathcal{U}[-\alpha_t \boldsymbol{\sigma}, +\alpha_t \boldsymbol{\sigma}]$, where $\alpha_t \in \mathbb{R}^+$ is a scaling hyper-parameter and $\boldsymbol{\sigma} \in \mathbb{R}^3$ is the per-axis standard deviation of the point cloud. To prevent points from passing behind the camera, the z-component of the translation is clamped to $t_z \leq \min_i z_i$, where z_i is the depth (z-coordinate) of the i -th point.
- **Pure rotation.** The camera rotates in place, with the maximum rotation angle bounded by the field of view. The translation is fixed to $\mathbf{t} = \mathbf{0}$. A forward direction is sampled by rotating the canonical forward axis $\hat{z}$ by polar angle $\theta \sim \mathcal{U}[0, \alpha_r \Theta_h]$ (where $\alpha_r \in \mathbb{R}^+$ is a rotation scaling factor and Θ_h is the horizontal field of view estimated by the monocular depth estimator) and azimuth $\phi \sim \mathcal{U}[0, 2\pi)$. This direction is then orthonormalized against the canonical up vector $\hat{y}$ to form $\mathbf{R}$.
- **Combined rotation and translation.** The camera is both shifted and rotated, combining the two previous strategies.

A translation $\mathbf{t}_t$ and rotation $\mathbf{R}_r$ are sampled independently as above and composed as $\mathbf{R}_{\text{hybrid}} = \mathbf{R}_r$, $\mathbf{t}_{\text{hybrid}} = \mathbf{R}_r \mathbf{t}_t$.

- **Normal-derived.** The camera is placed above a randomly selected surface point, looking at it from along its normal direction, as estimated by the monocular depth estimator. An anchor point $\mathbf{p} \in \mathbb{R}^3$ is sampled with probability $\propto \|\mathbf{p}\|^{-1}$, restricted to points whose surface normal $\hat{\mathbf{n}} \in \mathbb{R}^3$ satisfies $|n_y| < \tau$, where n_y is the y-component of the normal and τ is a filtering threshold. The camera is placed at $\mathbf{c} = \mathbf{p} + s \hat{\mathbf{n}}$, where the distance multiplier s is drawn from a log-uniform distribution $s \sim \log \mathcal{U}(d_{\min} \|\mathbf{p}\|, d_{\max} \|\mathbf{p}\|)$, with $d_{\min}$ and $d_{\max}$ representing the minimum and maximum distance bounds. Log-uniform sampling is used here to ensure that exponentially large distances are not overrepresented. Finally, $\mathbf{R}$ is set by a look-at from $\mathbf{c}$ to $\mathbf{p}$. Elements for which no valid normal survives filtering fall back to identity.
- **Frontal hemisphere.** The camera orbits around a randomly selected scene point, staying roughly frontal with a limited angular deviation. An anchor $\mathbf{p}$ is sampled with probability $\propto \|\mathbf{p}\|^{-1}$ and jittered as $\tilde{\mathbf{p}} = \mathbf{p} + \epsilon$,

$\epsilon \sim \mathcal{N}(\mathbf{0}, (\sigma_{\text{anchor}} \|\mathbf{p}\|)^2 \mathbf{I})$, where σ_{anchor} is a hyperparameter controlling the variance of the jitter. The reference direction $\hat{\mathbf{r}} = -\tilde{\mathbf{p}}/\|\tilde{\mathbf{p}}\|$ is perturbed by azimuth and elevation each drawn from $\mathcal{U}[-\delta, \delta]$, where δ bounds the maximum angular deviation, to obtain a new viewing direction $\hat{\mathbf{d}}$. The camera is placed at $\mathbf{c} = \tilde{\mathbf{p}} + z\hat{\mathbf{d}}$, where $z = \|\mathbf{p}\| \cdot s$ and $s \sim \log \mathcal{U}(d_{\min}, d_{\max})$. As previously mentioned, the log sampling of the distance multiplier s ensures that large distances are not overrepresented. Finally, $\mathbf{R}$ is set by a look-at from $\mathbf{c}$ to $\tilde{\mathbf{p}}$.

C.2. Look-at construction.

Given camera position $\mathbf{c}$ and target $\mathbf{p}$, we compute the forward vector $\hat{\mathbf{f}} = (\mathbf{p} - \mathbf{c})/\|\mathbf{p} - \mathbf{c}\|$, the right vector $\hat{\mathbf{r}} = (\hat{\mathbf{y}} \times \hat{\mathbf{f}})/\|\hat{\mathbf{y}} \times \hat{\mathbf{f}}\|$, the true up vector $\hat{\mathbf{u}} = \hat{\mathbf{f}} \times \hat{\mathbf{r}}$, and set $\mathbf{R} = [\hat{\mathbf{r}} \mid \hat{\mathbf{u}} \mid \hat{\mathbf{f}}]$.

C.3. Geometry-aware reprojection.

Given a target viewpoint, source colors are reprojected by mapping each 3D point into the new camera’s image plane. Formally, each source point $\mathbf{p}_i \in \mathbb{R}^3$, derived from the monocular depth estimator, with normal $\hat{\mathbf{n}}_i$ is projected to 2D pixel coordinates $\mathbf{q}_i$. This is expressed in homogeneous coordinates as $\mathbf{q}_i \sim \mathbf{K}(\mathbf{R}\mathbf{p}_i + \mathbf{t})$, where $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ is the known camera intrinsic matrix.

To handle occlusions, we apply a strategy akin to backface culling in computer graphics, discarding points that face away from the camera, *i.e.*, those satisfying $\hat{\mathbf{n}}_i^\top (\mathbf{c} - \mathbf{p}_i) \leq 0$, where $\mathbf{c} = -\mathbf{R}^\top \mathbf{t}$ is the target camera center in world coordinates. When multiple valid points project onto the exact same discrete pixel, a z-buffer resolves the collision by assigning the pixel the color of the point with the minimum projected depth. If no points project onto a given pixel, it remains black.

Finally, a visibility mask is computed to indicate these valid, populated pixels; this mask is later applied during the computation of perceptual losses as discussed in the main paper.

C.4. Hyperparameters.

All sampling hyperparameters and model settings are summarized in Table 4.

D. Implementation Details

D.1. Architecture

OVIE consists of a convolutional encoder ($8 \times$ spatial down-sampling), a ViT-B bottleneck, and a symmetric convolutional decoder. Input images (256×256) are compressed to a $32 \times 32 \times 512$ feature map via three ResNet stages (Group-Norm, SiLU), then patchified into 1024 tokens (1×1 patches) and linearly projected to $D=768$ dimensions. A 12-layer ViT-B ($D=768$, 12 heads, RMSNorm, SwiGLU) processes

Table 4. Camera sampling hyperparameters and model settings used during training.

Parameter	Symbol	Value
<i>Sampling probabilities</i>		
Identity	–	0.15
Pure translation	–	0.10
Pure rotation	–	0.10
Combined rotation & translation	–	0.35
Normal-derived	–	0.05
Frontal hemisphere	–	0.25
<i>Translation & rotation</i>		
Translation scaling factor	α_t	1.0
Rotation scaling factor	α_r	1.0
<i>Normal-derived & frontal hemisphere</i>		
Distance range	$[d_{\min}, d_{\max}]$	$[0.75, 1.5]$
Max perturbation angle	δ	25°
Anchor jitter scale	σ_{anchor}	0.02
<i>Model settings</i>		
Depth estimator	–	moge-2-vitl-norma

the tokens, after which they are unpatchified and decoded symmetrically, concluding with a 1×1 convolution and Sigmoid activation.

D.2. Camera Conditioning via AdaLN

Camera extrinsics $\mathbf{p} \in \mathbb{R}^7$ (translation and a quaternion for rotation) are projected to a conditioning embedding $\mathbf{c} \in \mathbb{R}^D$ via a single linear layer. This embedding modulates the ViT-B bottleneck via adaLN-Zero [44]: for each transformer block, a two-layer MLP regresses, from $\mathbf{c}$, dimension-wise scale γ , shift β , and residual gate α for both the MSA and SwiGLU sub-layers:

$$\mathbf{x}' = \mathbf{x} + \alpha_{\text{msa}} \odot \text{MSA}(\gamma_{\text{msa}} \odot \text{RMSNorm}(\mathbf{x}) + \beta_{\text{msa}}), \quad (5)$$

$$\mathbf{x}'' = \mathbf{x}' + \alpha_{\text{mlp}} \odot \text{SwiGLU}(\gamma_{\text{mlp}} \odot \text{RMSNorm}(\mathbf{x}') + \beta_{\text{mlp}}). \quad (6)$$

Following adaLN-Zero, the MLP’s final linear layer is zero-initialized so the conditioning path contributes nothing at the start of training.

D.3. Optimization and Training

OVIE is trained for 2,000,000 steps with a global batch size of 512 using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.05). We apply gradient clipping with a maximum norm of 1.0. The learning rate follows a cosine decay schedule, annealing from a peak of 2×10^{-4} down to a minimum of 2×10^{-5} , which is preceded by a linear warmup phase over the first 0.625% of training ($\sim 12.5\text{k}$ steps). Finally, we maintain an exponential moving average

(EMA) of the generator weights with a decay rate of 0.999 for inference.

Loss. The training objective combines L2 reconstruction loss, LPIPS ($\lambda_{\text{LPIPS}}=1.0$), and a P-DINO perceptual loss ($\lambda_{\text{P-DINO}}=0.5$) extracted from a pretrained DINOv3-ViT-B/16 model [59].

Adversarial Training. A PatchGAN discriminator [22] sharpens high-frequency detail. Following Representation Autoencoders (RAE) [77], we adopt a frozen DINO-S/8 backbone (inputs resized to 224×224) over the standard DINO-S/16, which reduces adversarial patch artifacts. The discriminator uses standard batch normalization, a convolutional head (kernel size 9) with Spectral Normalization, a hinge loss for discriminator updates, and a non-saturating loss for the generator. To balance the scale of the reconstruction and adversarial gradients, we employ a dynamic adaptive weighting scheme originally introduced in VQ-GAN [17]. At each training step, the adaptive weight λ is computed as:

$$\lambda = \text{clamp} \left(\frac{\|\nabla_{\mathbf{w}_L} \mathcal{L}_{\text{recon}}\|_2}{\|\nabla_{\mathbf{w}_L} \mathcal{L}_{\text{GAN}}\|_2 + \epsilon}, 0, \lambda_{\text{max}} \right),$$

where $\nabla_{\mathbf{w}_L}$ denotes the gradient with respect to the weights of the last convolutional layer of the decoder, $\mathcal{L}_{\text{recon}}$ is the L2 reconstruction loss, $\mathcal{L}_{\text{GAN}}$ is the generator’s adversarial loss, and $\epsilon = 10^{-6}$ ensures numerical stability. The weight is clamped to a maximum limit of $\lambda_{\text{max}} = 10,000$ to prevent gradient explosion. The final adversarial penalty added to the total training objective is scaled by $\lambda_{\text{adv}} \lambda \mathcal{L}_{\text{GAN}}$, where $\lambda_{\text{adv}} = 0.75$ is a fixed scalar. Finally, discriminator updates and adversarial penalties are delayed until 37.5% and 40% of total steps, respectively, to prevent early collapse.

Hyperparameters. All hyperparameters are reported in Table 5. For all ablation studies and supplementary experiments, we use the identical hyperparameter configuration but reduce the total training duration to 250,000 steps, keeping the absolute number of warmup steps constant.

E. Training Resources and Broader Impact

Training resources. All models are trained on 32 H100 GPUs. Ablation models complete in approximately 29 hours (250K steps); the full model completes in approximately 9.5 days (2M steps).

Broader impact. Training on unpaired single images opens novel view synthesis to domains where multi-view capture is impossible, including cultural heritage and archival photography. The ability to synthesize views from a single photograph also carries misuse risk for disinformation.

Table 5. Architecture, optimization, and loss hyperparameters.

Hyperparameter	Value	Hyperparameter	Value
<i>Generator Architecture</i>		<i>Losses</i>	
Resolution	256×256	Reconstruction loss	L2 (MSE)
Base channels	128	LPIPS weight λ_{LPIPS}	1.0
Channel multipliers	[1, 2, 4]	P-DINO model	DINOv3-ViT-B/16
Downsampling factor	$8 \times$	P-DINO weight λ_{DINO}	0.5
ViT bottleneck		Adversarial weight λ_{adv}	0.75
# Layers	12	Max adaptive weight limit	10,000
# Heads	12	<i>Discriminator & Adversarial Training</i>	
Hidden dimension	768	Backbone	Frozen DINO-S/8, 224×224
Normalization	RMSNorm	Head	Conv 9×9 , BN, SN
Activation	SwiGLU	Augmentation probability	1.0
Camera embedder	Linear($\mathbb{R}^7 \rightarrow \mathbb{R}^D$)	Discriminator loss	Hinge
<i>Generator Optimisation</i>		Generator loss	Non-saturating (Vanilla)
Batch size	512	Optimizer	AdamW
Training steps	2,000,000	Optimizer betas (β_1, β_2)	(0.9, 0.95)
Optimizer	AdamW	Peak learning rate	2×10^{-4}
Optimizer betas (β_1, β_2)	(0.9, 0.999)	Minimum learning rate	2×10^{-5}
Peak learning rate	2×10^{-4}	Learning rate scheduler	Cosine with warmup
Minimum learning rate	2×10^{-5}	Warmup ratio	5%
Learning rate scheduler	Cosine with warmup	Weight decay	0.0
Warmup ratio	0.625% ($\sim 12.5k$ steps)	D_ϕ update start	37.5% ($\sim 750k$ steps)
Weight decay	0.05	$\mathcal{L}_{\text{adv}}$ start	40.0% ($\sim 800k$ steps)
Gradient clip (max norm)	1.0		
EMA decay	0.999		

On the environmental side, the one-time training cost (9.5 days on 32 H100 GPUs) is comparable to similar-scale generative models. At inference, however, OVIE operates at 116 FPS in a single forward pass, more than $600 \times$ faster than the fastest baseline. This makes its per-image carbon footprint orders of magnitude smaller than diffusion-based competitors, whose iterative sampling (50–2000 steps) dominates inference cost at deployment scale.

F. Additional Qualitative Results

F.1. Out-of-distribution novel views

Figure 10 illustrates novel views synthesized from out-of-domain, non-realistic source images (*e.g.*, paintings). Notably, training on such artistic domains would be unfeasible using standard monocular novel-view synthesis methods reliant on multi-view datasets.

F.2. Comparison between training pseudo-targets and generated views

Figures 11–14 show examples of pseudo-targets used for supervision during training, along with the views generated from the source image and the target camera pose.

The grid-like patterns in the pseudo-target images (middle) stem from the point cloud’s regular spatial structure. Because each 3D point is a back-projected source pixel, the points inherit the original image’s grid layout. When rendered from a novel viewpoint, the spacing between these points becomes visible as a grid that varies with depth and angle.

F.3. Comparison to baseline methods

Figure 15–17 show qualitative comparisons between GeoGPT [53], PhotoNVS [73], VIVID [16], and OVIE on the RealEstate10K [80] dataset.

F.4. Side-by-side navigation clips

Please refer to the accompanying supplementary .zip archive for .gif files demonstrating OVIE’s performance

on continuous trajectories. These animations utilize sequences from the RealEstate10K [80] dataset. In each `.gif`, the left panel displays the ground-truth sequence, while the right panel shows the corresponding novel views generated by our approach. The generated sequence is synthesized by conditioning solely on the first image of the sequence, without utilizing any subsequent ground-truth frames. Unlike diffusion or autoregressive models, OVIE is a deterministic mapping: fixed inputs always produce the same output, with no stochastic sampling. Even in disoccluded regions, the source image serves as a fixed conditioning anchor: rather than drawing from stochastic noise as diffusion models do, OVIE maps the same input to the same hallucinated content at each forward pass. Temporal coherence over a trajectory then follows from the neural network’s smooth response to continuously varying pose inputs: nearby viewpoints produce nearby outputs, including in unobserved regions.

F.5. Real-time interactive navigation clips

Please refer to the accompanying supplementary `.zip` archive for `mp4` screen recordings showcasing the real-time interactive navigation capabilities of OVIE. To achieve this, we map standard mouse and keyboard actions to small, incremental changes in the camera’s position and rotation, similar to the control mechanics found in first-person video games. Based on these inputs, we continuously update the camera extrinsics to generate a new image on the fly, conditioned strictly on the initial source image.

G. Additional Comparison to InfiniteNature-Zero

InfiniteNature-Zero [31] learns perpetual view generation from single-view image collections, using a geometry-based render-refine-repeat pipeline that relies on monocular depth estimation and sky segmentation at inference. Like OVIE, it requires no posed multi-view data during training, making it a conceptually relevant comparison despite its fundamentally different, geometry-dependent approach.

To ensure a fair comparison, we use the public checkpoint of InfiniteNature-Zero and adapt its inference to our single-step evaluation protocol, generating each target view independently from the source frame rather than autoregressively. This grants the model access to the ground-truth source image and its depth estimate at every step, an inherent advantage over OVIE and other geometry-free methods.

Despite this advantage, InfiniteNature-Zero’s generation quality lags behind OVIE’s. On RealEstate10K, it achieves an LPIPS of 0.395 and FID of 28.8, compared to OVIE’s 0.279 and 6.74. On DL3DV, it scores an LPIPS of 0.472 and FID of 45.7 versus OVIE’s 0.464 and 13.6, the highest FID among all compared methods on both benchmarks. As shown in Figure 18, InfiniteNature-Zero produces blurry

outputs, which explains its poor LPIPS and FID scores. Notably, its PSNR and SSIM remain comparable to OVIE’s (19.4 and 0.642 on RealEstate10K versus 18.8 and 0.602), a discrepancy attributable to the insensitivity of these metrics to perceptual sharpness, as discussed in Section 4.5.



Figure 10. **Qualitative results on out-of-distribution images.** Each pair shows the input source image followed by the generated novel view. Source views are, from left to right and top to bottom: *Gas* by Edward Hopper, *Untitled* by Ralambo, *A Sunday on La Grande Jatte* by Georges Seurat, *Nighthawks* by Edward Hopper, *Portrait of an Artist (Pool with Two Figures)* by David Hockney, and *The Sea of Ice* by Caspar David Friedrich.

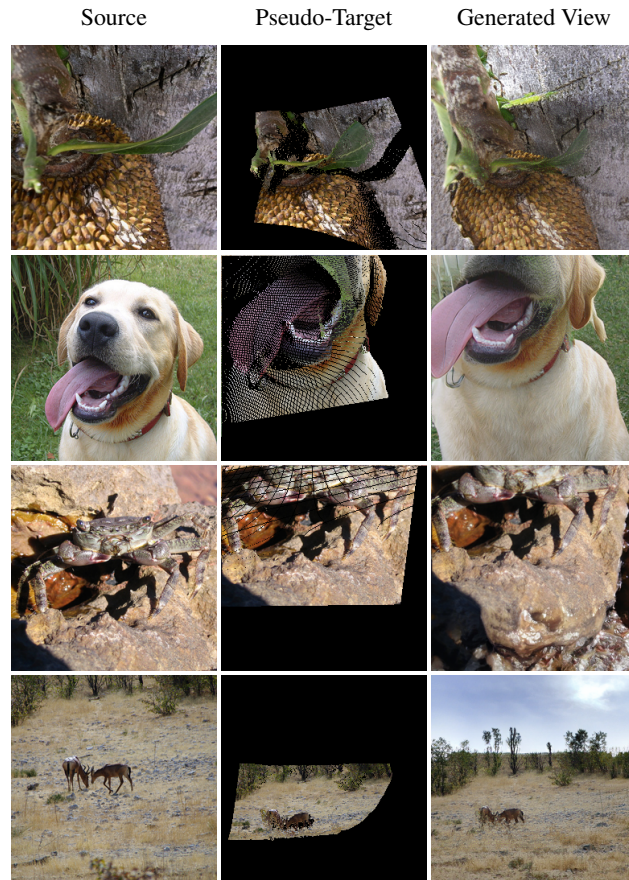


Figure 11. **Comparison of source inputs, training pseudo-targets, and generated views.** During training, OVIE is supervised on pseudo-targets (middle) created by depth-lifting the source image (left) to a sampled pose. At inference, it generates novel views (right) from a source image and target pose. Here, the generated views are rendered at the same poses as their corresponding pseudo-targets.



Figure 12. **Comparison of source inputs, training pseudo-targets, and generated views.** During training, OVIE is supervised on pseudo-targets (middle) created by depth-lifting the source image (left) to a sampled pose. At inference, it generates novel views (right) from a source image and target pose. Here, the generated views are rendered at the same poses as their corresponding pseudo-targets.



Figure 13. **Comparison of source inputs, training pseudo-targets, and generated views.** During training, OVIE is supervised on pseudo-targets (middle) created by depth-lifting the source image (left) to a sampled pose. At inference, it generates novel views (right) from a source image and target pose. Here, the generated views are rendered at the same poses as their corresponding pseudo-targets.

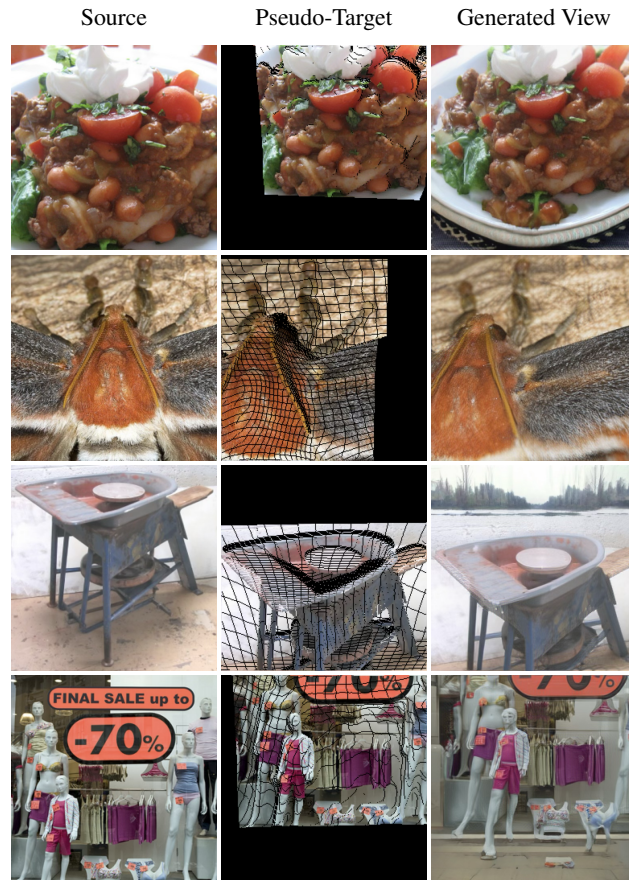


Figure 14. **Comparison of source inputs, training pseudo-targets, and generated views.** During training, OVIE is supervised on pseudo-targets (middle) created by depth-lifting the source image (left) to a sampled pose. At inference, it generates novel views (right) from a source image and target pose. Here, the generated views are rendered at the same poses as their corresponding pseudo-targets.

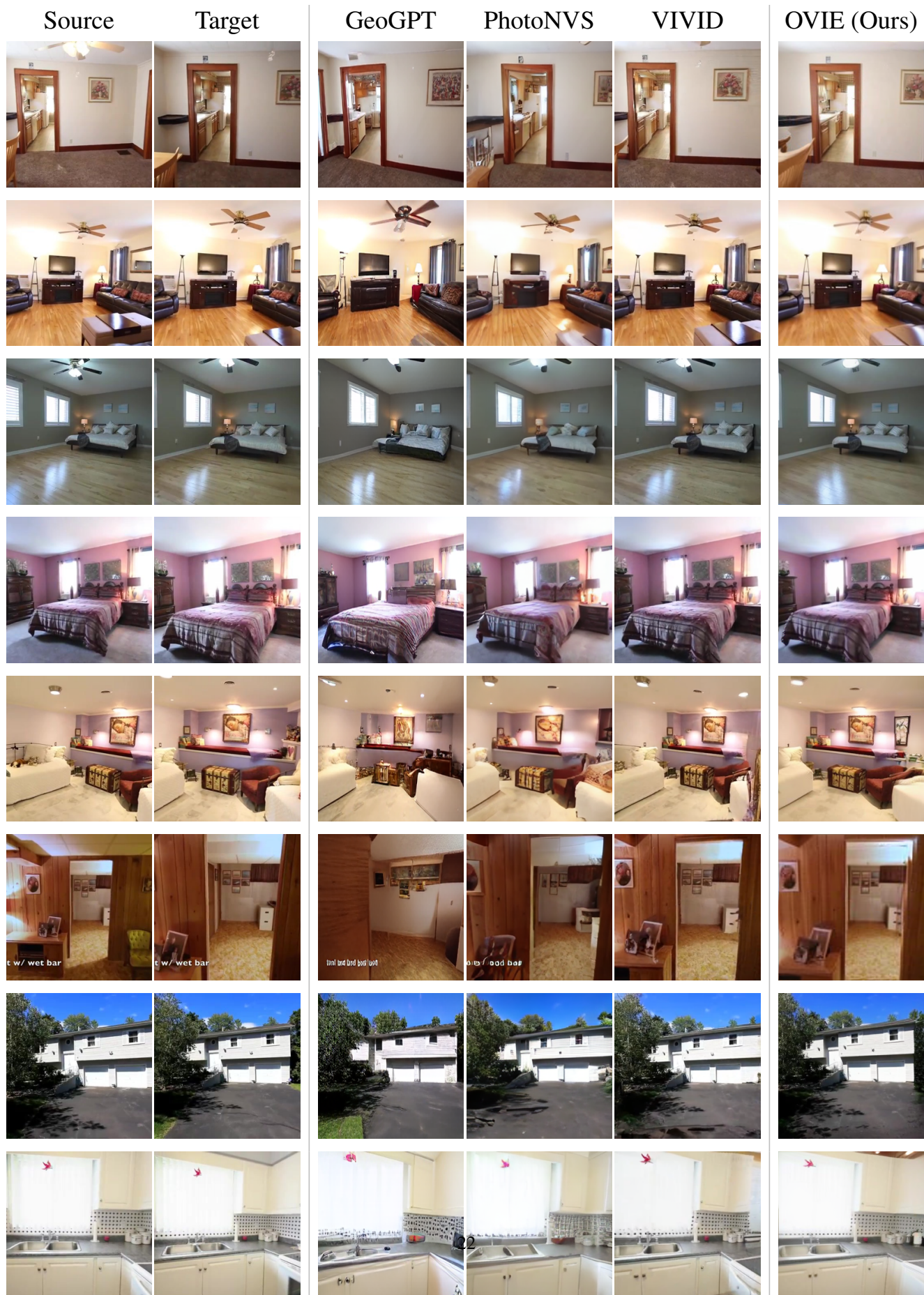


Figure 15. **Qualitative comparison with state-of-the-art methods.** Given a source image and target camera pose, each method synthesizes a novel view. Despite training on no multi-view data, OVIE generates novel views that match or exceed the quality of concurrent methods.

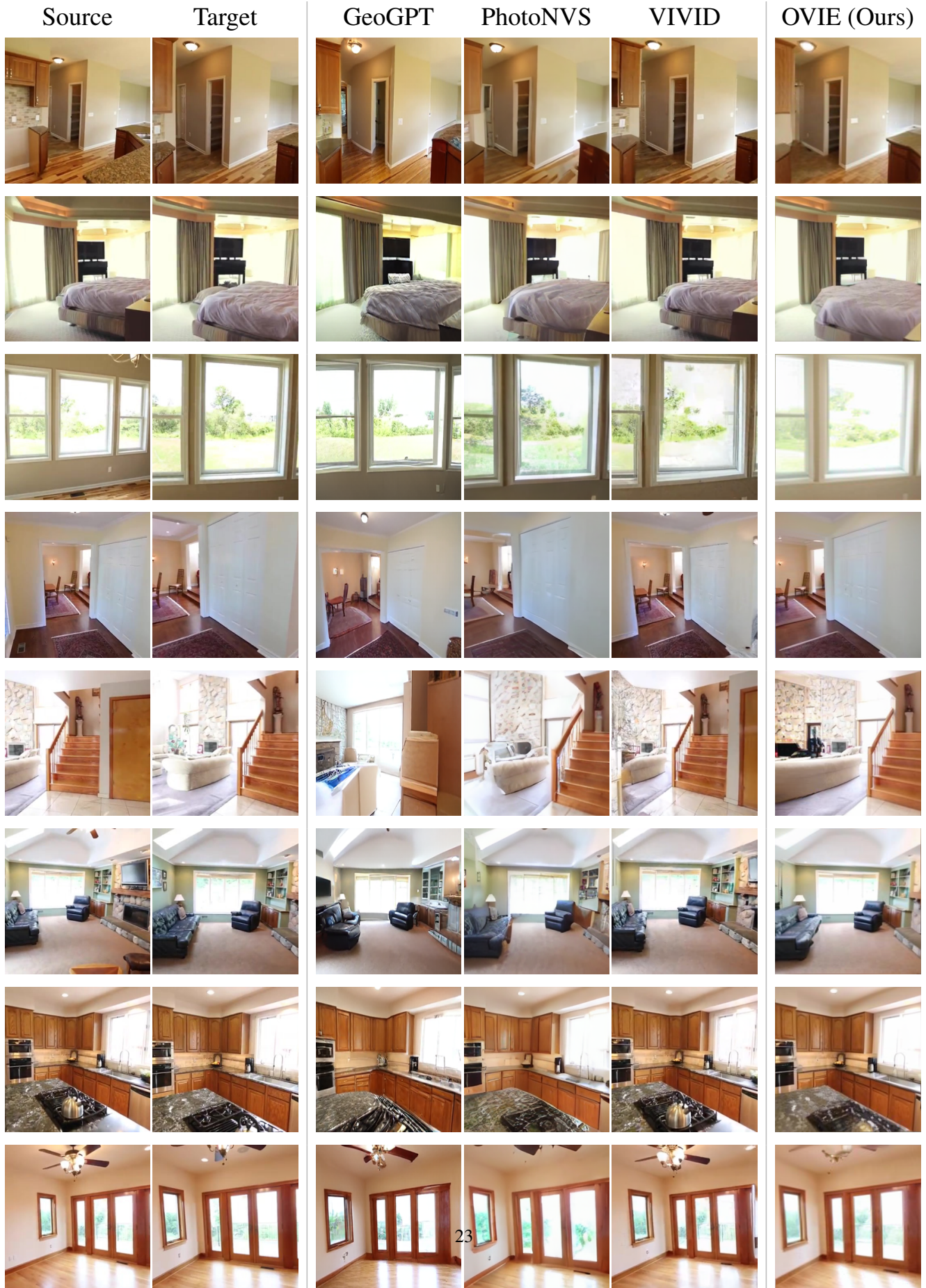


Figure 16. **Qualitative comparison with state-of-the-art methods.** Given a source image and target camera pose, each method synthesizes a novel view. Despite training on no multi-view data, OVIE generates novel views that match or exceed the quality of concurrent methods.

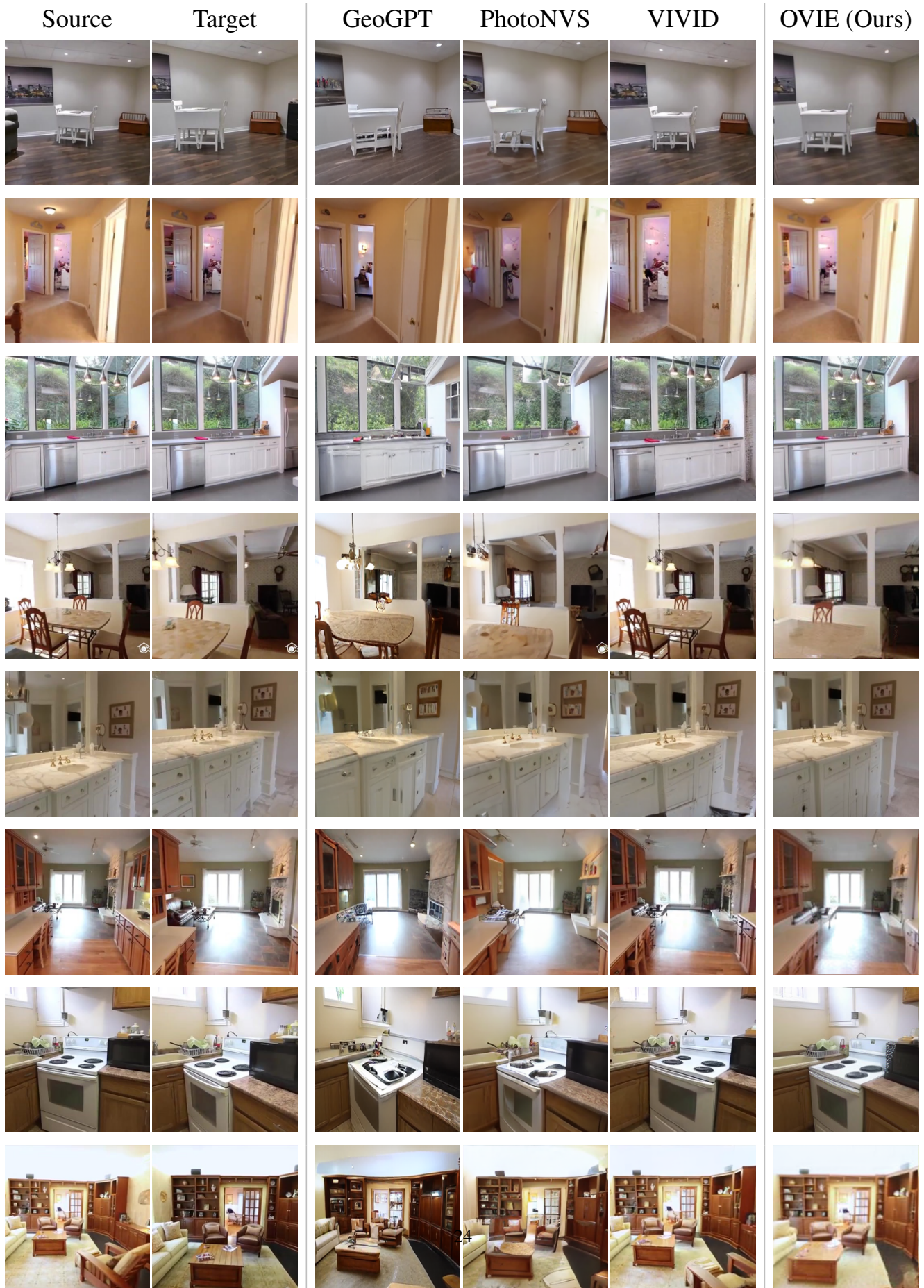


Figure 17. **Qualitative comparison with state-of-the-art methods.** Given a source image and target camera pose, each method synthesizes a novel view. Despite training on no multi-view data, OVIE generates novel views that match or exceed the quality of concurrent methods.



Figure 18. **Qualitative comparison with InfiniteNature-Zero [31].** Given a source image (left), we show the ground-truth target view, the view generated by OVIE (ours), and the view generated by InfiniteNature-Zero. OVIE’s generated views are sharper and more realistic than those of InfiniteNature-Zero.