Scaling View Synthesis Transformers

Evan Kim*
MIT
evnkim@mit.edu

Hyunwoo Ryu*
MIT
hwryu@mit.edu

Thomas W. Mitchel
Adobe
thomas.w.mitchel@gmail.com

Vincent Sitzmann
MIT
sitzmann@mit.edu

Abstract

Geometry-free view synthesis transformers have recently achieved state-of-the-art performance in Novel View Synthesis (NVS), outperforming traditional approaches that rely on explicit geometry modeling. Yet the factors governing their scaling with compute remain unclear. We present a systematic study of scaling laws for view synthesis transformers and derive design principles for training compute-optimal NVS models. Contrary to prior findings, we show that encoder–decoder architectures can be compute-optimal; we trace earlier negative results to suboptimal architectural choices and comparisons across unequal training compute budgets. Across several compute levels, we demonstrate that our encoder–decoder architecture, which we call the Scalable View Synthesis Model (SVSM), scales as effectively as decoder-only models, achieves a superior performance–compute Pareto frontier, and surpasses the previous state-of-the-art on real-world NVS benchmarks with substantially reduced training compute. https://www.evn.kim/research/svsm

1. Introduction

Given a set of images of a scene with known camera poses, the goal of Novel View Synthesis (NVS) is to render novel views of the scene from arbitrary viewpoints. Single scene approaches such as NeRF [15] and Gaussian Splatting [11] have achieved impressive fidelity by explicitly modeling 3D geometry and rendering. Feed-forward extensions of these frameworks train neural networks to reconstruct the 3D representation, achieving promising results [2, 3, 28, 30]. However, their formulation inherits handcrafted 3D structure, constraining their ability to scale and handle more complex artifacts such as reflections or transparency.

Typified by Large View Synthesis Model (LVSM) [9], a new class of view synthesis models have emerged which achieve state-of-the-art rendering quality using pure transformer architectures with fewer (if any) geometric inductive biases [8, 9, 16, 20]. However, this class of models is relatively new, and their design space remains unexplored. In particular, training a NVS model involves making design choices over a large number of variables, including the number of context and target views, camera pose parameterizations, attention mechanisms, etc., and there does not yet exist a rigorous investigation of how these choices affect the performance, training efficiency, and inference throughput. To this point, while there exist extensive scaling analyses in language modeling and 2D vision [6, 7, 10, 18, 29], there exists no analogue for 3D vision. Thus, the goal of this work is to provide such an analysis in addition to a compute-optimal training recipe for view synthesis transformers in terms of both architecture and training strategy.

In particular, we first challenge the necessity of the decoder-only architecture proposed in LVSM [9]. While powerful, it requires passing all context images through the entire transformer each time a *single* target image is decoded. It is a *bidirectional* model where both the target view tokens *and* context view tokens are updated in each layer of the network. While this allows the model to consider only information in the context images that is relevant to the target view, it incurs substantial computational cost due to repeated processing of context views.

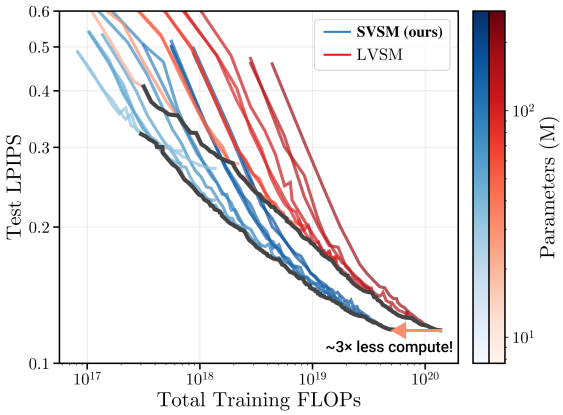

Figure 1. **Scaling Laws for View Synthesis Transformers.** Evaluated on RealEstate10K [33], our SVSM exhibits a 3× more compute-optimal Pareto frontier than LVSM while retaining the same scaling behavior (similar slope and curvature everywhere).

*Indicates equal contribution.

Instead, we advocate for an encoder-decoder design which produces an intermediate scene latent representation. This approach is potentially far more efficient: the computational cost of constructing the scene representation is amortized through repeated calls to the decoder, which efficiently extracts information from the representation via *unidirectional* cross attention from scene to target. However, the scene representation also represents an information bottleneck. Without a proper training strategy that maximally leverages the efficiency of the encoder-decoder design, it can be challenging for encoder-decoder models to outperform decoder-only models. Here, we identify that the key for unlocking the potential of encoder-decoder models lies in the way target views (*i.e.* the reconstruction targets) are utilized during training. The implicit standard practice employed by prior work [2, 9, 20, 30] has been to reconstruct multiple different target views from a single scene during training. However, the consequences of this approach have never been fully analyzed. To this point, we propose and empirically validate the *effective batch hypothesis*, which argues that reconstructing multiple target views per scene effectively multiplies the batch size.

These insights yield a principled transformer view synthesis model, which we call the *Scalable View Synthesis Model* (SVSM), that fully capitalizes on the rendering efficiency of a unidirectional encoder-decoder architecture, maximizing training throughput without compromising performance or scalability. We demonstrate that our unidirectional model scales as efficiently as bidirectional models, which aligns with the scalability of causal, unidirectional attention in large language models [10]. As part of this analysis, we also reveal scaling relationships within view synthesis that parallel those observed in the Chinchilla language model family [7]. Finally, we demonstrate that SVSM achieves state-of-the-art results in real-world NVS tasks with significantly reduced compute, challenging the previous understanding that bidirectional attention is critical to high-fidelity view synthesis [9].

Key Contributions:

- We provide the first rigorous scaling analysis for novel view synthesis transformers.
- We propose and empirically confirm the *effective batch size hypothesis* that unlocks compute-optimal training.
- We show that bidirectional decoding is not critical for scalable view synthesis, contrasting recent work [9].
- Based on this analysis, we present a compute-optimal model that achieves a new state-of-the-art in real-world NVS tasks with substantially reduced training compute.

2. Related Work and Preliminaries

Generalizable novel view synthesis. In generalizable novel view synthesis, we are given a set of V_C context images paired with known camera poses and intrinsics $\mathfrak{C} = \{(I_i, g_i, K_i) \mid i = 1, \ldots, V_C\}$. The typical objective is to synthesize an *unseen* view of the same scene given a target camera configuration g_T, K_T:

$$\tilde{I}_T = \text{Render}\big[\mathfrak{C}, g_T, K_T\big]. \tag{1}$$

One line of work attempts to solve this problem with neural network architectures that explicitly model aspects of 3D image formation, for instance, via differentiable rendering or using epipolar line constraints [2, 22, 24, 25, 28]. In contrast, *geometry-free* methods avoid explicit geometric modeling in favor of flexibility and generality [5, 19–21, 23]. Here, we are primarily interested in a recently proposed subclass of pure transformer architectures [8, 9, 16, 20]. In particular, we seek to study the "Large View Synthesis Model" (LVSM) [9], which achieves state-of-the-art NVS performance and serves as the prototypical instance of the view synthesis transformer. LVSM can be implemented in two ways: as either an encoder-decoder model or decoder-only model. The authors' proposed decoder-only variant is far more performant, so we primarily consider this architecture (pictured in Fig. 2, left) in our analysis.

Decoder-only LVSM consists of a single module: A decoder $\mathcal{D}$ which ingests the raw context $\mathfrak{C}$ along with a *single* target configuration, and aims to render a prediction of the target view $\tilde{I}_T = \mathcal{D}\big[\mathfrak{C}, g_T, K_T\big]$. This model is *bidirectional* as the context view tokens are updated with information about target pose and tokens in each layer. Thus, the processed context tokens cannot be reused and must be re-initialized and updated each time a new target view is rendered. As a consequence, rendering V_T target views requires V_T forward passes through the full network. Therefore, the FLOPs on a forward pass scale linearly with the number of target views V_T

$$\begin{aligned} \chi_{\texttt{MLP}}^{(\text{LVSM})} &\propto V_T \times (V_C + 1) \\ \chi_{\texttt{Attn}}^{(\text{LVSM})} &\propto V_T \times (V_C + 1)^2 \end{aligned} \tag{2}$$

where $\chi_{\texttt{MLP}}^{(\text{LVSM})}$ and $\chi_{\texttt{Attn}}^{(\text{LVSM})}$ are the FLOPs consumed by the MLP and attention in the decoder-only LVSM. In our studies, $\chi_{\texttt{MLP}}^{(\text{LVSM})}$ is the dominating factor (for more details, see Supp 8). In what follows, we will continue to use χ to denote the compute metric, typically measured in FLOPs.

Scaling Laws. As the scale of deep learning models continues to increase, it has become increasingly important to understand the relationship between performance and compute to ensure an efficient use of resources. To this end, scaling analyses have been conducted for language models [7, 10], vision transformers [29], and diffusion transformers [6, 18]. The general approach is straightforward: train models at different compute budgets and analyze performance as a function of compute.

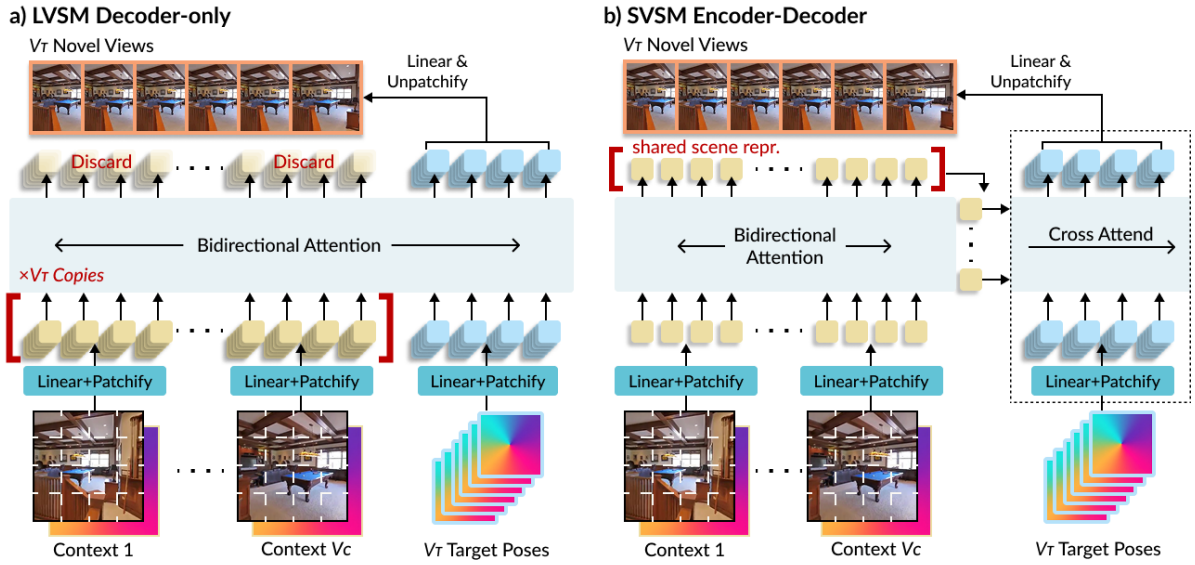

Figure 2. **Architectures of the current SOTA, the decoder-only LVSM [9] (a) and SVSM (ours, b).** Our cross-attention based decoder enables parallel rendering of multiple target views after a single scene encoding. Each target view is decoded independently given the shared scene representation, but the cross-attention allows these independent decodings to be executed in parallel.

Scaling studies are useful in two ways. First, they provide a predictable trend of performance with compute, describing a performance metric P as function of compute χ. For example, in language models, $P(\chi)$ has been found to approximately follow a power-law [10]. Second, they have revealed which hyperparameter choices are most effective as models scale. For instance, Chinchilla scaling laws [7] reveal the best way to trade-off between model size N, measured in parameter count, and the number of training samples used, D. Analysis is performed by sweeping across a wide range of N and D to discover for each compute budget χ the optimal N and D. Then, taking this paired data one can fit a power law relation between N_{opt} and D_{opt} and χ,

$$N_{\text{opt}}(\chi) \propto \chi^a, D_{\text{opt}}(\chi) \propto \chi^b. \tag{3}$$

Remarkably, experiments demonstrate $a \approx b$, suggesting N and D should scale proportionally. In our study, we will focus on the second of these two kinds of studies: replicating the Chinchilla study (Sec. 5) in the NVS domain and exploring other tradeoffs, such as the effective batch size (Sec. 4)

Extremely long context view synthesis. Recently, there has been a line of work aiming to develop view-synthesis and 3D-reconstruction models whose computational cost scales linearly with the number of context images [26, 32, 34]. Indeed, as in Eq. 2, as V_C grows, the quadratic cost of attention comes to dominate the compute and becomes infeasible. In that regime, such linear-cost models are promising alternatives. However, we restrict our focus to sparse to moderately sparse view synthesis, for which linear-cost models are currently not state of the art.

3. Encoder-Decoder View Synthesis

As discussed in Sec. 1 and Sec. 2, the decoder-only LVSM may not be the most compute-optimal due to the recomputation per-target view rendered. This motivates us to seek an alternative model with a scene representation that can be decoded in a *unidirectional* manner (*i.e.* via cross attention) to reduce the cost of rendering and avoid redundant reprocessing of context information. To this extent, we introduce the Scalable View Synthesis Model (SVSM), which can be viewed as a simple modification to encoder-decoder LVSM.

Specifically, our architecture implements Eq. 1 by first processing the context set $\mathfrak{C}$ with a transformer encoder, producing a set of latent tokens (a "scene representation") $\mathbf{z} = \mathcal{E}[\mathfrak{C}]$. The encoder $\mathcal{E}$ is standard transformer with full bidirectional self-attention. Unlike encoder-decoder LVSM, we do not employ a fixed-size scene representation, but instead take the set of encoded context image patch tokens as the scene representation to avoid introducing a bottleneck. To render a novel view, a cross-attention based decoder $\mathcal{D}$ ingests the target configuration and the fixed scene latent tokens $\mathbf{z}$ to render the target view, $\tilde{I} = D\big[\mathbf{z}, g_T, K_T\big]$. To render multiple images of the same scene, we only require encoding the context set once, reusing the scene embedding $\mathbf{z}$. As with LVSM, each novel view is decoded independently given $\mathbf{z}$ (i.e., there is no interaction between target views). However, because the decoder uses cross-attention, these independent target views can be decoded in parallel, without redundant recomputation of the scene representation.

To be more concrete, this architecture reduces the com-

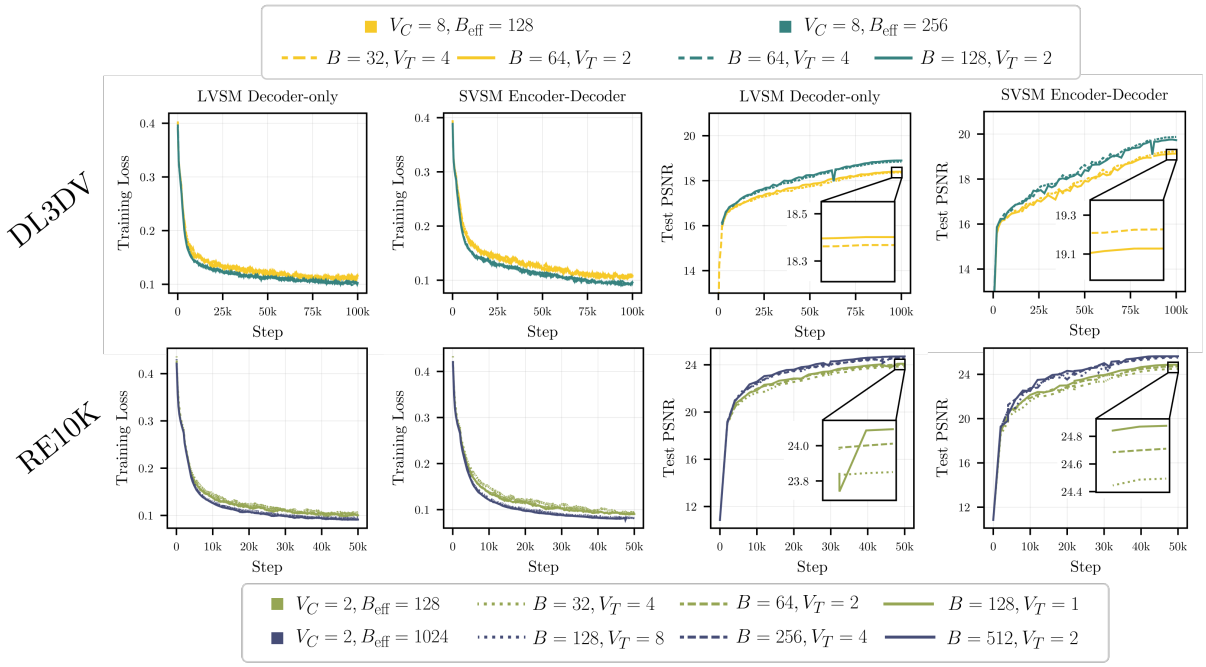

Figure 3. **Effective Batch Size**. Training loss (smoothed with a rolling-average) and test PSNR measured throughout training across various paired B and V_T runs provide evidence for our effective batch size hypothesis: Models trained with the same product of number of scenes in the batch B and number of reconstruction target views V_T, *i.e.* runs with the same *effective batch size* B_{eff}, perform the same and are colored identically. On $V_C = 8$ **(top)**, we sweep across $B_{\text{eff}} = 128, 256$ on DL3DV, and on $V_C = 2$ **(bottom)**, we sweep across $B_{\text{eff}} = 128, 1024$ on RealEstate10K.

plexity of rendering V_T targets to

$$\begin{aligned} \chi^{(\text{SVSM})}_{\texttt{MLP}} &\propto V_T + V_C \\ \chi^{(\text{SVSM})}_{\texttt{Attn}} &\propto V_C \times (V_T + V_C). \end{aligned} \tag{4}$$

In other words, taking the dominant factor to be MLP layers (see Supp 8), rendering V_T target views requires $\mathcal{O}(V_T + V_C)$ FLOPs. In the limit of inference where $V_T \gg V_C$, this reduces to $\mathcal{O}(V_T)$, in stark contrast to the $\mathcal{O}(V_T V_C + V_T)$ of LVSM (see Eq. 2). Further, the benefit of this paradigm extends beyond inference. As long as we are training with multiple target views, as is standard practice [2, 9, 20, 30], the parallel nature of unidirectional decoding can save substantial training compute.

Unidirectional decoding can be less expressive than bidirectional. Indeed, parameter count and training steps being equal, SVSM performs worse than LVSM. However, as we will show, SVMS's amortized rendering enables us to dramatically increase its size and training steps, such that when normalized by compute budget, SVSM significantly outperforms LVSM.

4. The Effective Batch Size for View Synthesis

As the cost of a forward pass scales both with the number of scenes (the batch size B) as well as the number of target views (V_T) that we render per scene, this introduces an additional hyperparameter into the training regime: What is the *optimal* trade-off between the number of target views and the number of different scenes? We study this question empirically and reveal that what matters is the *product* of target views and batch size, which we call the *effective batch size* of a NVS model.

Analysis Setup. We define effective bath size as $B_{\text{eff}} \equiv B \cdot V_T$, where B is the number of scenes in a training batch, and V_T is the number of rendering targets used per training scene. We train both decoder-only LVSM and the proposed SVSM models across two datasets — DL3DV [14] and RealEstate10K [33] (RE10K) — with $V_C = 8$ and $V_C = 2$ while holding B_{eff} constant and varying B and V_T. We test $B_{\text{eff}} = 128$ on both datasets, and additionally test $B_{\text{eff}} = 1024$ on RE10K and $B_{\text{eff}} = 256$ on DL3DV. For DL3DV we use the official test-train split, and for RE10K, we use the pixelSplat [2] test-train split. Further training details are outlined in Supp 9.

Effective Batch Size is What Matters. Remarkably, in all cases – across both models, both V_C settings, and all B_{eff} sets – the test metric and the training loss behavior remain approximately constant along a B_{eff}-level set (Fig. 3). This effect is especially clear in the $V_C = 8$ case, where the test PSNR varies by at most ± 0.1 and remains present in the $V_C = 2$ case, where the variation is at most ± 0.2 PSNR. Thus, since varying B and V_T within the same effective

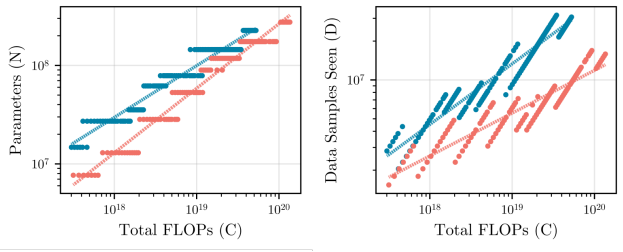

Figure 4. **Data and Model Scaling Plots.** While our model (blue) is optimal when sufficient data is available, decoder-only LVSM (red) performs better with less data. Our model is also less parameter-efficient, though the gap closes as we increase the training compute.

batch size B_{eff} does not make a significant difference, we treat B_{eff} as the true batch size.

SVSM Enables Compute-Optimal Tradeoff. How can we interpret this result through the lens of compute-optimality? For the LVSM decoder-only model, the training compute scales as

$$\chi^{(\text{LVSM})} \propto BV_T(V_C + 1) = B_{\text{eff}}(V_C + 1). \tag{5}$$

Thus, any training settings within constant B_{eff} not only achieve within-noise results (as per our effective batch result), but also require the same number of FLOPs. This means for the decoder-only model there is *no advantage to be gained by tuning* V_T. In contrast, for the SVSM model, training compute is proportional to

$$\chi^{(\text{SVSM})} \propto B(V_C + V_T) = B_{\text{eff}} + BV_C. \tag{6}$$

Therefore, by reducing B and increasing V_T, one can achieve the same effective batch size – and consequently, the same performance – with lower compute cost. This justifies our original motivation for a model design that efficiently decodes multiple V_T.

5. Scaling Laws for Stereo (V_C=2) NVS

Analysis Setup. We first experiment in the most classical setting for feed-forward Novel View Synthesis – stereo synthesis with two context views. All training and evaluation is done on RealEstate10K [33]. As before, we follow the evaluation framework of pixelSplat [2]. For training, we use $V_T = 6$ target views per training example, following the setup of [9], and a scene batch size of 256. We use the test LPIPS [31] loss as our primary performance metric, as it produces near linear trends on log-log plots against FLOPs.

Scaling Laws. We now follow the approach in language modeling [7] to answer the question: *for a given compute-budget, what is the optimal performance that can be attained?* For both model families, we sweep across a range of models from around 7M to 300M parameters, training each model for 3-4 different sample counts to densely cover the FLOP range [7]. Our training runs span a compute range of 10^3 magnitudes: 100 petaflops to 100 exaflops. From this data, we are then able to determine a mapping from compute budgets C to minimum test LPIPs – the Pareto frontier.

We plot results in Fig. 1, with their Pareto frontiers marked in dark gray. Plotted on a log-log scale, both models exhibit a consistent downward trend on test LPIPS with more compute. More significantly, the Pareto frontiers of both families have approximately the same slope at points of the same performance (see Supp 10), and SVSM's frontier is shifted left by a factor of 3. Thus, as an initial result, our scaling laws show us that our encoder-decoder architecture *scales exactly the same as the decoder-only LVSM while using* $3\times$ *less training-compute.* Qualitative results of this scaling are shown in Fig. 5, and we see that when FLOP-matched, SVSM has better rendering quality.

Optimal Model Choice. From our scaling experiments, we can further extract a compute-optimal training recipe for our view synthesis transformers, as demonstrated by Hoffmann et al. [7]. For each compute budget χ, we determine the corresponding optimal model size N and the amount of training data D used at that point. Then, plotting N and D against χ, we can extract the Chinchilla scaling equations (Eq. 3) by fitting lines onto the log-log plots in Fig. 4. The recovered coefficients are shown in Tab. 2, which inform how to train models which end on the frontier.

From these results it follows that for SVSM, if we increase our compute budget by a factor of $k\times$, it should approximately be equally allocated between increasing the model size by $\sqrt{k}$ and increasing data sample count by $\sqrt{k}$ as $a_{\text{SVSM}} \approx b_{\text{SVSM}}$, matching the findings of the Chinchilla scaling laws [7] for language models. For LVSM this relationship does not seem to hold exactly, but the fit still shows that requires significant scaling of data with respect to compute, though to a smaller power.

It should be noted that our data sample counts include *repeated scene data*, as we only have access to small, pose-labeled datasets. This differs from standard scaling practice, in which models are typically trained for less than one full epoch [7, 10, 29]. Although we have not yet seen evidence of overfitting in our experiments, we have shown that increasing scale requires increasing the number of training samples. Thus, having access to larger amounts of diverse posed data will be essential for developing large-scale generalizable NVS models.

SVSM-420M/740M Results. Finally, combining our scaling law findings, we train two separate models — SVSM-420M and SVSM-740M — to compare against the original results of LVSM's largest model on RealEstate10K. Due to compute-constraints, we train our models at a lower total budget of around 10^{21} FLOPs and a batch size of

Model	Scale Parameters			Reconstruction Quality			Rendering FPS (↑)		
	Model Size	Train Iters	Train FLOPs (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	V_C=2	V_C=4	V_C=8
LVSM Encoder-Decoder [9]	173M	100k	2.53 zflops	28.58	0.893	0.114	53.7	52.9	**52.7**
LVSM Decoder-Only [9]	171M	100k	1.60 zflops	29.67	0.906	0.098	37.9	19.5	8.6
SVSM Enc-Dec (**ours**, Iter-matched)	740M	100k	**0.74 zflops**	29.80	0.907	0.098	48.6	42.7	35.0
SVSM Enc-Dec (**ours**, Pareto-optimal)	416M	170k	**0.77 zflops**	**30.01**	**0.910**	**0.096**	**71.0**	**61.8**	49.7

Table 1. **Stereo ($V_C = 2$) NVS Results of the Largest Models.** All models use a patch size of 8 with input images at 256×256 resolution. Our models achieve the highest reconstruction metrics while using less than half of the training compute. The rendering FPS of both SVSM models is also much faster than that of the LVSM decoder-only model, though both are slower than the LVSM encoder-decoder when V_C is large.

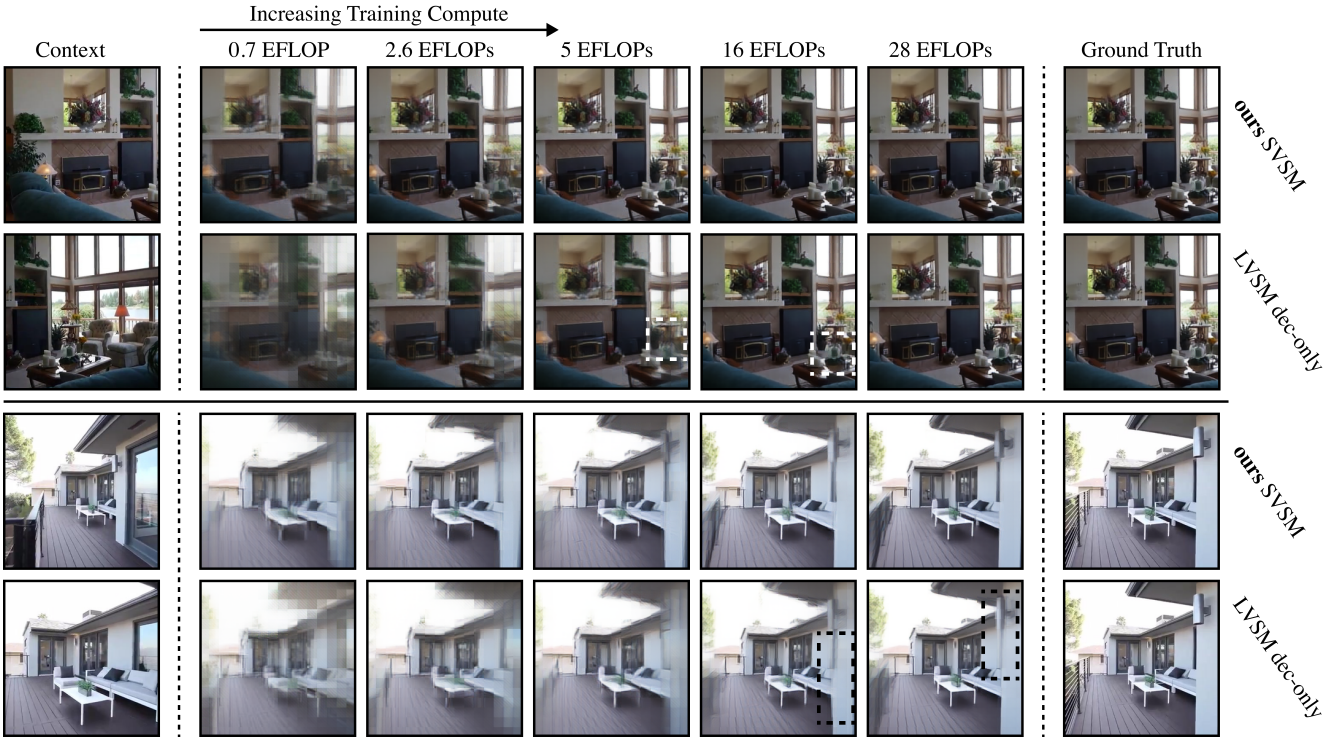

Figure 5. **Qualitative Scaling Behavior**, $V_C = 2$. From left to right, both models steadily increase in rendering quality until reaching near photo-realistic results. Compared vertically, for a given compute-budget, SVSM renderings consistently contain less artifacts.

Model	Parameter Coeff. a	Data Coeff. b
LVSM	0.65	0.33
SVSM	0.52	0.47

Table 2. **Parameter and Data Scaling Coefficients.** As regressed from the plots in Fig. 4, we find power law coefficients for scaling models and data with respect to compute.

256, approximately half the FLOPs and exactly half the batch size used by LVSM. We train two models under this budget: (1) a forward-pass-flop-matched model with the 24 layer LVSM model for a single training sample with V_C=2, V_T=6 and; (2) A model whose parameter count is given by plugging the budget[1] χ and the coefficients from Tab. 2 to Eq. 3.

[1]To be more specific, we plug in $\chi/4$ in order to adjust for the scaling law being derived off of 16×16 patch experiments, while this final budget is under 8×8 patches, which require roughly $4\times$ as much compute.

Model	Recon Quality		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)
pixelNeRF [28]	20.43	0.589	0.550
pixelSplat [2]	26.09	0.863	0.136
MVSplat [3]	26.39	0.869	0.128
GS-LRM [30]	28.10	0.892	0.114
SVSM Enc-Dec **(ours)**	**30.01**	**0.910**	**0.096**

Table 3. **Comparison to Geometry-Aware Methods.** Our method achieves a new state-of-the-art on RealEstate10K [33] with the set from Charatan et al. [2], outperforming not only LVSM, but also prior work with explicit 3D structure.

While we train with under half the compute, our scaling laws in Sec. 5 predict equal performance with three times less compute. Thus, both SVSM models outperform decoder-only LVSM (1). Notably, SVSM-420M, the model trained in accordance with our scaling laws performs the

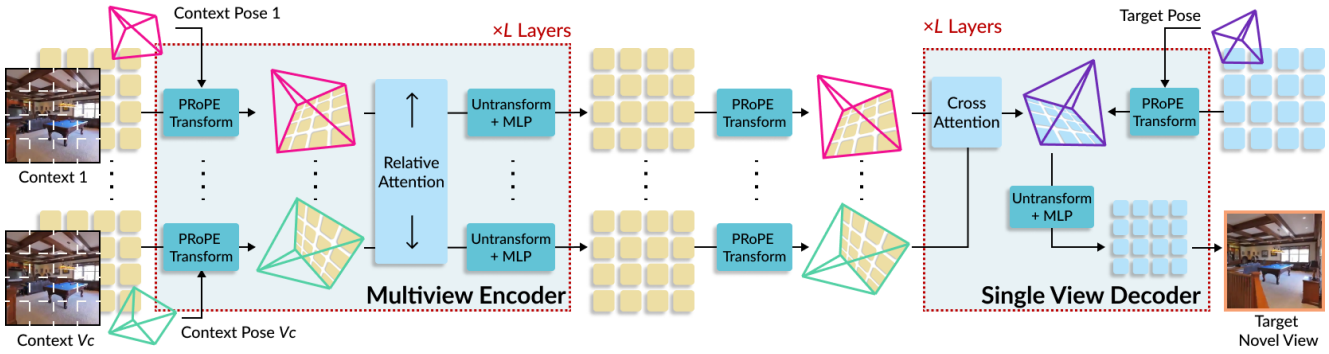

Figure 6. **Multiview PRoPE.** We find that multiview projective RoPE embeddings [12, 13, 17] are critical for our model to scale with compute and data in the multiview setting ($V_C > 2$). For each layer of the multiview transformer encoder, PRoPE embeddings use context camera poses to transform context view tokens into a common coordinate frames before the attention layer, and apply the inverse transformation before each MLP. To render, both context features and query tokens of the target view are transformed by PRoPE before cross-attention.

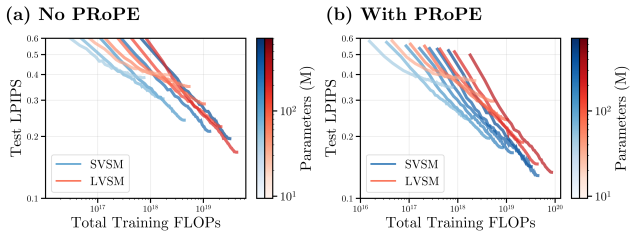

Figure 7. **Multiview Scaling Behavior.** Conducted on DL3DV [14]. **(a)** For V_C>2, without PRoPE, SVSM saturates and stops scaling much more quickly than LVSM. **(b)** When PRoPE is added, SVSM continues scaling with a better Pareto-frontier.

best. We also show reported results from prior work on this benchmark in Tab. 3.

Furthermore, we also benchmark the *rendering speed*, which is calculated with $V_T = 1$ to simulate real-time online rendering with respect to a stream of input poses. Additional details can be found in Supp 9.2. SVSM generally renders much faster than decoder-only LVSM (Tab. 1).

6. Scaling Laws for Multiview (V_C>2) NVS

Analysis Setup. We also experiment in the multiview paradigm (V_C>2). Specifically, we focus on $V_C = 4$. For training and evaluation, we choose DL3DV [14], a real-world dataset with wider baselines and more complex camera trajectories, making it more suitable for multiview experiments. We follow the official test-train split and use V_T=4 and, scene batch size of 64 for all experiments to save resources. All other settings follow those described in Sec. 5 and Supp 9.

Scaling Law Does Not Hold. Unfortunately, we find that naively extending our SVSM architecture to the multiview scenario does not result in a similar scaling trend. As can be seen from Figure 7a, the Pareto frontier of our unidirectional model saturates much quicker than bidirectional LVSM as we increase the train compute.

Relative Camera Attention Re-establishes Scaling Law. We hypothesize that this is not a fundamental problem of the encoder-decoder paradigm, but a problem caused by the way our model utilizes the pose information. Specifically, we find that adding a form of relative camera attention [12, 13, 17] resolves this issue. Relative attention mechanisms embed the pose information directly into the attention layers by transforming the query, key, and value vectors. This introduces an inductive bias which makes learning easier, particularly for our encoder-decoder architecture.

For our model, we adopt the recently proposed PRoPE [13] embedding as the relative camera attention mechanism. We illustrate SVSM's architecture with the incorporation of PRoPE embeddings in Fig. 6. After adding PRoPE embeddings to both LVSM and SVSM, we retrain all models. Results are shown in Fig. 7. The equivalent scaling is re-established, and SVSM again maintains a tighter Pareto-frontier. Qualitative scaling results for both models are shown in Fig. 8. Note that while both models benefit from PRoPE embeddings, the advantage is far more pronounced in our encoder-decoder SVSM.

Final Models. Equipped with PRoPE embeddings, we again train larger models to compare directly against the 24-layer LVSM Decoder-only model, also equipped with PRoPE. As we did in the stereo case, we train a naive forward pass-matched model along with a Pareto-optimal model from a $N(\chi)$ fit to the data. The performance of both models are listed in Tab. 4 and their test loss curves are plotted in Fig. 7. Again, the version of SVSM which follows the scaling laws outperforms both models, with significant 0.7 PSNR and -0.016 LPIPS gaps. Beyond the superior reconstruction quality, the efficiency of SVSM becomes clear in

Model	Scale Parameters			Reconstruction Quality			Rendering FPS (↑)		
	Model Size	Train Iters	Train FLOPs (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	V_C=4	V_C=8	V_C=16
LVSM Decoder-only + PRoPE [9, 13]	171M	100k	43 eflops	26.19	0.830	0.145	104.7	52.6	23.8
SVSM Enc-Dec (**ours**, ≈ Iter-matched)	711M	100k	32 eflops	26.29	0.835	0.141	280.4	261.2	230.4
SVSM Enc-Dec (**ours**, Pareto-optimal)	400M	233k	44 eflops	**26.87**	**0.853**	**0.129**	**411.1**	**381.1**	**333**

Table 4. **Multiview (V_C>2) NVS Results of the Largest Models.** Our compute-matched model achieves significantly better rendering quality (+0.68 PSNR, −0.016 LPIPS), while maintaining nearly four times the rendering speed at inference-time.

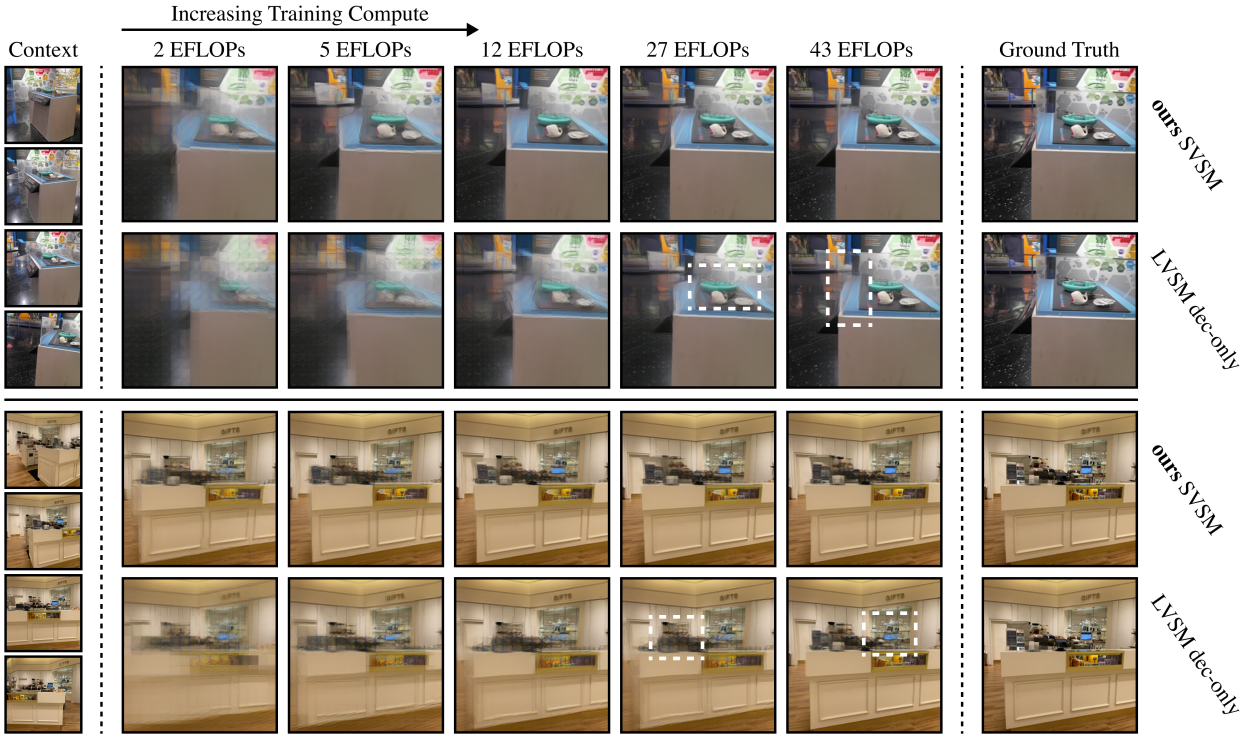

Figure 8. **Qualitative Scaling Behavior**, $V_C = 4$. The performance of both LVSM and our method steadily increases with compute (left to right). Compared vertically, for a given compute budget SVSM renderings are consistently less blurry.

the multi-view case, with a rendering FPS that is $4\times$ that of the decoder-only model, increasing to $14\times$ when extrapolated to larger context view counts.

7. Conclusion

In this work, we established a rigorous compute-normalized benchmark for transformer view synthesis models. Our empirical studies reveal the importance of the concept of *effective batch size* —the product of the number of scenes in a batch with the number of per-scene rendering target views — which redefines the notion of batch size for NVS training. Based on this insight, we propose the Scalable View Synthesis Model (SVSM), which features a unidirectional encoder-decoder architecture for favorable scaling with effective batch. We demonstrate that SVSM is dramatically more compute-efficient than the current SOTA architecture, LVSM, and consistently achieves the same performance with $2-3\times$ less training compute. We further demonstrate that relative camera pose embeddings in multi-view attention is the key to realizing favorable scaling behavior with increasing numbers of context views. In sum, our findings establish a new framework for evaluating the performance and effectiveness of transformer view synthesis models.

Acknowledgements. This work was supported by the National Science Foundation under Grant No. 2211259, by the Singapore DSTA under DST00OECI20300823 (New Representations for Vision, 3D Self-Supervised Learning for Label-Efficient Vision), by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) under 140D0423C0075, by the Amazon Science Hub, by the MIT-Google Program for Computing Innovation, and by a 2025 MIT Office of Research Computing and Data Seed Grant.

References

[1] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. *arXiv preprint arXiv:2309.16620*, 2023. 11

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024. 1, 2, 4, 5, 6

[3] Yuedong Chen, Haofei 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. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 1, 6

[4] 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. *arXiv preprint arXiv:2212.08051*, 2022. 13

[5] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360 (6394):1204–1210, 2018. 2

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 1, 2

[7] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1, 2, 3, 5

[8] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025. 1, 2

[9] 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. *arXiv preprint arXiv:2410.17242*, 2024. 1, 2, 3, 4, 5, 6, 8, 11, 12

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 2, 3, 5

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[12] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 7

[13] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025. 7, 8, 13

[14] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 4, 7

[15] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[16] Thomas W Mitchel, Hyunwoo Ryu, and Vincent Sitzmann. True self-supervised novel view synthesis is transferable. *arXiv preprint arXiv:2510.13063*, 2025. 1, 2

[17] Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. *arXiv preprint arXiv:2310.10375*, 2023. 7, 13

[18] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 2

[19] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. 2

[20] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 1, 2, 4

[21] Mehdi SM Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. Rust: Latent neural scene representations from unposed imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17297–17306, 2023. 2

[22] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 2

[23] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021. 2

[24] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 2

[25] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10208–10217, 2024. 2

[26] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3

[27] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks. *arXiv preprint arXiv:2310.02244*, 2023. 11

[28] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 6

[29] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 1, 2, 5

[30] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 1, 2, 4, 6

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[32] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025. 3

[33] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 1, 4, 5, 6

[34] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3