DeLiRa: Self-Supervised Depth, Light, and Radiance Fields

Vitor Guizilini[1] Igor Vasiljevic[1] Jiading Fang[2] Rares Ambrus[1] Sergey Zakharov[1]
Vincent Sitzmann[3] Adrien Gaidon[1]

[1]Toyota Research Institute (TRI), Los Altos, CA
[2]Toyota Technological Institute of Chicago (TTIC), Chicago, IL
[3]Massachusetts Institute of Technology (MIT), Cambridge, MA

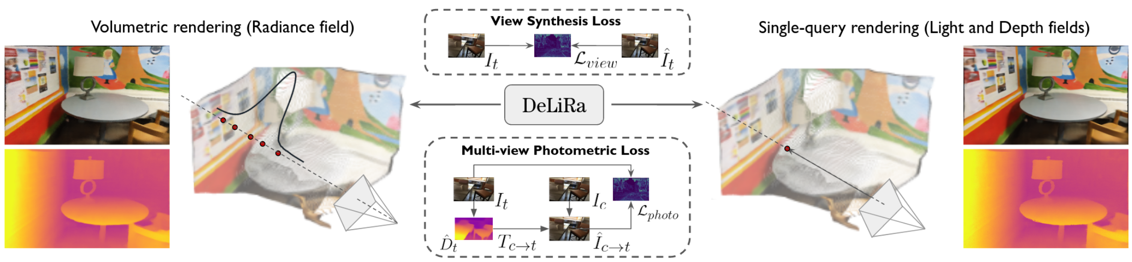


Figure 1: **DeLiRa augments volumetric view synthesis with the multi-view photometric objective**, as a regularizer to improve novel view and depth synthesis in the limited viewpoint setting. We use this implicit representation to jointly learn depth, light, and radiance fields from a shared latent space in a synergistic way.

Abstract

Differentiable volumetric rendering is a powerful paradigm for 3D reconstruction and novel view synthesis. However, standard volume rendering approaches struggle with degenerate geometries in the case of limited viewpoint diversity, a common scenario in robotics applications. In this work, we propose to use the multi-view photometric objective from the self-supervised depth estimation literature as a geometric regularizer for volumetric rendering, significantly improving novel view synthesis without requiring additional information. Building upon this insight, we explore the explicit modeling of scene geometry using a generalist Transformer, jointly learning a radiance field as well as depth and light fields with a set of shared latent codes. We demonstrate that sharing geometric information across tasks is mutually beneficial, leading to improvements over single-task learning without an increase in network complexity. Our DeLiRa architecture achieves state-of-the-art results on the ScanNet benchmark, enabling high quality volumetric rendering as well as real-time novel view and depth synthesis in the limited viewpoint diversity setting. Our project page is https://sites.google.com/view/tri-delira.

1. Introduction

Inferring 3D geometry from 2D images is a cornerstone capability in computer vision and computer graphics. In recent years, the state of the art has significantly advanced due to the development of neural fields [47], which parameterize continuous functions in 3D space using neural networks, and differentiable rendering [40, 23, 36], which enables learning these functions directly from images. However, recovering 3D geometry from 2D information is an ill-posed problem: there is an inherent ambiguity of shape and radiance (i.e. the *shape-radiance ambiguity* [54]). These representations thus require a large number of diverse camera viewpoints in order to converge to the correct geometry. Alternatively, methods that explicitly leverage geometric priors at training time, via the self-supervised multi-view photometric objective, have achieved great success for tasks such as depth [6, 9, 45, 8], ego-motion [38, 37], camera geometry [41, 4, 7], optical flow [10], and scene flow [10, 14].

In this work, we combine these two paradigms and introduce the multi-view photometric loss as a complement to the view synthesis objective. Specifically, we use depth inferred via volumetric rendering to warp images, with the photometric consistency between synthesized and original images serving as a self-supervisory regularizer to scene structure. We show through experiments that this explicit regularization facilitates the recovery of accurate geometry in the case of low viewpoint diversity, without requiring additional data. Because the multi-view photometric objective is unable to model view-dependent effects (since it assumes a Lambertian scene), we propose an attenuation schedule that gradually removes it from the optimization, and show that our learned scene geometry is stable, leading to further improvements in view and depth synthesis.

We take advantage of this accurate learned geometry and propose **DeLiRa**, an auto-decoder architecture inspired by [16] that jointly estimates **De**pth [12], **Li**ght [35], and **Ra**diance [23] fields. We maintain a shared latent representation across task-specific decoders, and show that this increases the expressiveness of learned features and is beneficial for all considered tasks, improving performance over single-task networks without additional complexity. Furthermore, we explore other synergies between these representations: volumetric predictions are used as pseudo-labels for the depth and light fields, improving viewpoint generalization; and depth field predictions are used as guidance for volumetric sampling, significantly improving efficiency without sacrificing performance.

To summarize, our contributions are as follows. In our first contribution, we show that the **multi-view photometric objective is an effective regularization tool for volumetric rendering**, as a way to mitigate the shape-radiance ambiguity. To further leverage this geometrically-consistent implicit representation, in our second contribution we propose a **novel architecture for the joint learning of depth, light, and radiance fields**, decoded from a set of shared latent codes. We show that jointly modeling these three fields leads to improvements over single-task networks, without requiring additional complexity in the form of regularization or image-space priors. As a result, **our proposed method achieves state-of-the-art view synthesis and depth estimation results on the ScanNet benchmark**, outperforming methods that require explicit supervision from ground truth or pre-trained networks.

2. Related Work

2.1. Implicit Representations for View Synthesis

Our method falls in the category of auto-decoder architectures for neural rendering [27, 47], which directly optimize a latent code. Building on top of DeepSDF [27], SRN [36] adds a differentiable ray marching algorithm to regress color, enabling training from a set of posed images. The vastly popular NeRF [23] family regresses color and density, using volumetric rendering to achieve state-of-the-art free view synthesis. CodeNeRF[18] learns instead the variation of object shapes and textures, and does not require knowledge of camera poses at test time.

Despite recent improvements, efficiency remains one of the main drawbacks of volumetric approaches, since rendering each pixel requires many network calls. To alleviate this, some methods have proposed better sampling strategies [13, 43]. This is usually achieved using depth priors, either from other sensors [29], sparse COLMAP [32] predictions [3], or pre-trained depth networks [46, 31, 25]. Other methods have moved away from volumetric rendering altogether and instead generate predictions with a single forward pass [35, 12, 42]. While much more efficient, these methods require substantial viewpoint diversity to achieve the multi-view consistency inherent to volumetric rendering. Light field networks [35] map an oriented ray directly to color, relying on generalization to learn a multi-view consistency prior. DeFiNe [12] learns *depth* field networks, using ground truth to generate virtual views via explicit projection. R2L [42] uses a pre-trained volumetric model, that is distilled into a residual light field network.

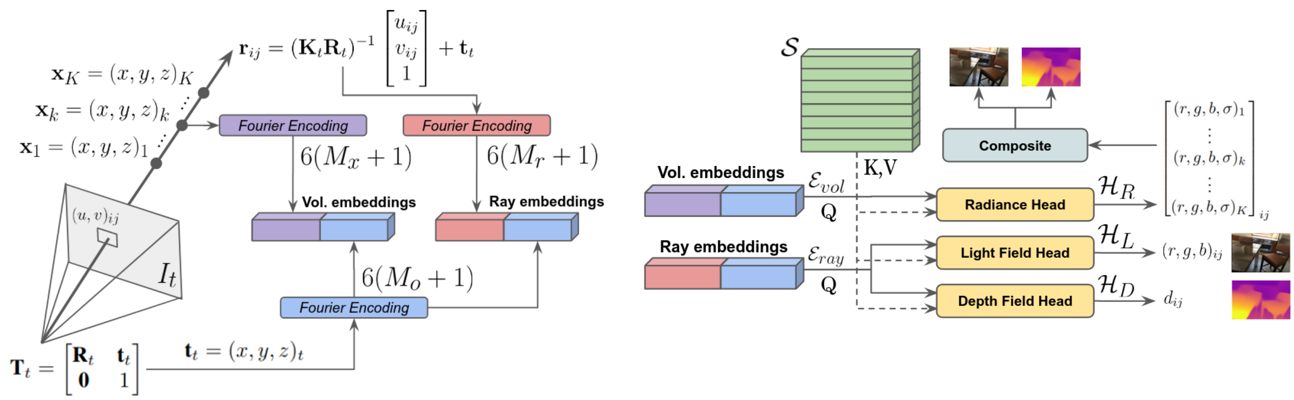
Our proposed method combines these two directions into a single framework. Differently from DeFiNe and R2L, it does not require ground truth depth maps or a pre-trained volumetric model for distillation. Instead, we propose to *jointly* learn self-supervised depth, light, and radiance fields, decoded from the same latent space. A radiance decoder generates volumetric predictions that serve as additional multi-view supervision for light and depth field decoders. At the same time, predictions from the depth field decoder serve as priors to improve sampling for volumetric rendering, decreasing the amount of required network calls.

2.2. Self-Supervised Depth Estimation

The work of Godard *et al.* [5] introduced self-supervision to the task of depth estimation, by framing it as a view synthesis problem: given a target and context images, we can use predicted depth and relative transformation to warp information between viewpoints. By minimizing a photometric objective between target and synthesized images, depth and relative transformation are learned as proxy tasks. Further improvements to this original framework, in terms of losses [6, 33, 11], camera modeling [7, 41, 19, 4] and network architectures [9, 8, 50], have led to performance comparable to or even better than supervised methods. We extend this self-supervised learning paradigm to the volumetric rendering setting, introducing it as an additional source of regularization to the single-frame view synthesis objective. Our argument is that, by enforcing explicit multi-view photometric consistency in addition to implicit density-based volumetric rendering, we avoid degenerate geometries during the learning process.

2.3. Structure Priors for Volumetric Rendering

A few works have recently started exploring how to incorporate depth and structure priors in the volumetric rendering setting [25, 3, 46, 31]. Most use structure-from-motion (e.g., COLMAP [32]) pointclouds, predicted jointly with camera poses, as "free" supervision to regularize volumetric depth estimates. However, because these pointclouds are noisy and very sparse ($< 0.1\%$ of valid projected pixels), substantial post-processing is required. RegNeRF [26] uses a normalizing-flow-based likelihood model over image patches to regularize predictions from unobserved views. DS-NeRF [3] uses reprojection error as a measure of uncer-



(a) Geometric embeddings, including camera center $\mathbf{t}_t$, viewing rays $\mathbf{r}_{ij}$, and sampled 3D points $\mathbf{x}_k$ from each viewing ray.

(b) Multi-task depth, light, and radiance decoding. Geometric embeddings are used as queries Q, and cross-attended with keys K and values V from $\mathcal{S}$.

Figure 2: **Diagram of our proposed DeLiRa architecture**. In (a) we show how the various geometric embeddings are calculated from camera information, and in (b) we show depth, light, and radiance decoding from the same latent space $\mathcal{S}$.

tainty, and minimizes the KL-divergence between volumetric and predicted depth. NerfingMVS [46] trains a separate depth prediction network, used for depth-guided sampling. DDP-NeRF [31] goes further and trains a separate depth *completion* network, that takes SfM pointclouds as additional input to generate dense depth maps for supervision and sampling guidance. SinNeRF [48] uses a single RGB-D image to learn a radiance field. Like our method, they use multi-view photometric warping as additional supervision, but crucially they rely on *ground truth* depth, and thus the two objectives (volume rendering and warp-based view synthesis) are decoupled. MonoSDF [53] uses a pre-trained depth network to supervise SDF-based volume rendering, achieving impressive improvements in depth estimation, albeit at the expense of novel view synthesis performance.

Importantly, all these methods require additional data to train their separate networks, in order to generate the depth priors used for volumetric rendering. NerfingMVS uses a network pre-trained on 170K samples from 4690 COLMAP-annotated sequences [20]. DDP-NeRF uses a network trained on 94K RGB-D in-domain samples (i.e., from the same dataset used for evaluation). In these two methods (and several others [2, 51]), supervision comes from "free" noisy COLMAP pointclouds. Drawing from the self-supervised depth estimation literature, we posit that geometric priors learned with a multi-view photometric objective are a stronger source of "free" supervision.

3. Methodology

Our goal is to learn an implicit 3D representation from a collection of RGB images $\{I_i\}_{i=0}^{N-1}$. For each camera, we assume known intrinsics $\mathbf{K}_i \in \mathbb{R}^{3\times3}$ and extrinsics $\mathbf{T}_i \in \mathbb{R}^{4\times4}$, obtained as a pre-processing step [32]. Note that we assume neither ground-truth [25] nor pseudo-ground truth [25, 46, 3, 31] depth supervision.

3.1. Shared Latent Representation

Our architecture for the joint learning of depth, light, and radiance fields (DeLiRa) stores scene-specific information as a latent space $\mathcal{S} \in \mathbb{R}^{N_l \times C_l}$, composed of N_l vectors with dimensionality C_l. Cross-attention layers are used to decode queries, containing geometric embeddings (Fig. 2a), into task-specific predictions. To self-supervise these predictions, we combine the view synthesis objective on RGB estimates and the multi-view photometric objective on depth estimates. We also explore other cross-task synergies, namely how volumetric predictions can be used to increase viewpoint diversity for light and depth field estimates, and how depth field predictions can serve as priors for volumetric importance sampling. A diagram of DeLiRa is shown in Fig. 2b, and below we describe each of its components.

3.2. Geometric Embeddings

We use geometric embeddings to process camera information, that serve as queries to decode from the latent space $\mathcal{S}$. Let $\mathbf{u}_{ij} = (u, v)$ be an image coordinate in target camera t, with assumed known pinhole intrinsics $\mathbf{K}_t$, resolution $H \times W$, and extrinsics $\mathbf{T}_t = \left[\begin{smallmatrix}\mathbf{R} & \mathbf{t}\\ \mathbf{0} & 1\end{smallmatrix}\right]$ relative to camera T_0. Its origin $\mathbf{o}_t$ and direction $\mathbf{r}_{ij}$ vectors are given by:

$$\mathbf{o}_t = -\mathbf{R}_t\mathbf{t}_t \quad , \quad \mathbf{r}_{ij} = \left(\mathbf{K}_t\mathbf{R}_t\right)^{-1}[u_{ij}, v_{ij}, 1]^T \tag{1}$$

Note that direction vectors are normalized to unitary values $\bar{\mathbf{r}}_{ij} = \frac{\mathbf{r}_{ij}}{||\mathbf{r}_{ij}||}$ before further processing. For volumetric rendering, we sample K times along the viewing ray to generate 3D points $\mathbf{x}_k = (x, y, z)$ given depth values z_k:

$$\mathbf{x}_{ij}^k = \mathbf{o}_t + z_k\bar{\mathbf{r}}_{ij} \tag{2}$$

In practice, z_k values are linearly sampled between a $[d_{min}, d_{max}]$ range. These vectors, $\mathbf{o}_t$, $\bar{\mathbf{r}}_{ij}$ and $\mathbf{x}_{ij}^k$, are then

Fourier-encoded [52] to produce geometric embeddings $\mathcal{E}_o$, $\mathcal{E}_r$ and $\mathcal{E}_x$, with a mapping of:

$$x \mapsto [x, \sin(f_1\pi x), \cos(f_1\pi x), \ldots, \sin(f_M\pi x), \cos(f_M\pi x)] \tag{3}$$

where M is the number of Fourier frequencies used (M_o for camera origin, M_r for viewing ray, and M_x for 3D point), equally spaced in the interval $[1, \frac{\mu}{2}]$. Here, μ is a maximum frequency parameter shared across all dimensions. These embeddings are concatenated to be used as queries by the cross-attention decoders described below. Ray embeddings are defined as $\mathcal{E}_{ray} = \mathcal{E}_o \oplus \mathcal{E}_r$ and volumetric embeddings as $\mathcal{E}_{vol} = \mathcal{E}_o \oplus \mathcal{E}_x$, where $\oplus$ denotes concatenation.

3.3. Cross-Attention Decoders

We use task-specific decoders, with one cross-attention layer between the $N_q \times C_q$ queries and the $N_l \times C_l$ latent space $\mathcal{S}$ followed by an MLP that produces a $N_q \times C_o$ output (more details in the supplementary material). This output is processed to generate estimates as described below.

The radiance head $\mathcal{H}_R$ decodes volumetric embeddings $\mathcal{E}_{vol}$ as a 4-dimensional vector $(\mathbf{c}, \sigma)$, where $\mathbf{c} = (r, g, b)$ are colors and σ is density. A sigmoid is used to normalize colors between $[0, 1]$, and a ReLU truncates densities to positive values. To generate per-pixel predictions, we composite K predictions along its viewing ray [23], using sampled depth values $Z_{ij} = \{z_k\}_{k=0}^{K-1}$. The resulting per-pixel predicted color $\hat{\mathbf{c}}_{ij}$ and depth $\hat{d}_{ij}$ is given by:

$$\hat{\mathbf{c}}_{ij} = \sum_{k=1}^{K} w_k \hat{\mathbf{c}}_k \quad , \quad \hat{d}_{ij} = \sum_{k=1}^{K} w_k z_k \tag{4}$$

Per-point weights w_k and accumulated densities T_k, given intervals $\delta_k = z_{k+1} - z_k$, are defined as:

$$w_k = T_k \Big(1 - \exp(-\sigma_k \delta_k)\Big) \tag{5}$$

$$T_k = \exp\Big(- \sum_{k'=1}^{K} \sigma_{k'} \delta_{k'}\Big) \tag{6}$$

The light field head $\mathcal{H}_L$ decodes ray embeddings $\mathcal{E}_{ray}$ as a 3-dimensional vector $\hat{\mathbf{c}}_{ij} = (r, g, b)$ containing pixel colors. These values are normalized between $[0, 1]$ with a sigmoid.

The depth field head $\mathcal{H}_D$ decodes ray embeddings $\mathcal{E}_{ray}$ as a scalar value $\hat{d}_{ij}$ representing predicted pixel depth. This value is normalized between a $[d_{min}, d_{max}]$ range.

3.4. Self-Supervised Losses

We combine the traditional volumetric rendering view synthesis loss $\mathcal{L}_s$ (Sec. 3.4.1) with the multi-view photometric loss $\mathcal{L}_p$ (Sec. 3.4.2), using α_p as a weight coefficient:

$$\mathcal{L} = \mathcal{L}_s + \alpha_p \mathcal{L}_p \tag{7}$$

3.4.1 Single-View Volumetric Rendering

We use Mean Squared Error (MSE) to supervise the predicted image $\hat{I}_t$ (where $\hat{I}_t(i,j) = \hat{\mathbf{c}}_{ij}$, see Eq. 4), relative to the target image I_t.

$$\mathcal{L}_s = ||\hat{I}_t - I_t||^2 \tag{8}$$

This is the standard objective for radiance-based reconstruction [23]. Importantly, this is a *single-view* objective, since it directly compares prediction and ground truth without considering additional viewpoints. Therefore, multi-view consistency must be learned implicitly by observing the same scene from multiple viewpoints. When such information is not available (e.g., forward-facing datasets with limited viewpoint diversity), it may lead to degenerate geometries that do not properly model the observed 3D space.

3.4.2 Multi-View Photometric Warping

To address this limitation, we introduce the self-supervised multi-view photometric objective [6, 9] as an additional source of self-supervision in the volumetric rendering setting. For each pixel (u, v) in target image I_t, with predicted depth $\hat{d}$ (e.g., see Eq. 4), we obtain the projected coordinates (u', v') with predicted depth $\hat{d}'$ in a context image I_c via a warping operation, defined as:

$$\hat{d}' \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \mathbf{K}_c \mathbf{R}_{t\to c} \left(\mathbf{K}_t^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \hat{d} + \mathbf{t}_{t\to c} \right) \tag{9}$$

To produce a synthesized target image, we use grid sampling with bilinear interpolation [15] to place information from the context image onto each target pixel, given their corresponding warped coordinates. The photometric reprojection loss between target I_t and synthesized $\hat{I}_t$ images consists of a weighted sum with a structure similarity (SSIM) term [44] and an L1 loss term:

$$\mathcal{L}_p(I_t, \hat{I}_t) = \alpha \, \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1 - \alpha) \, \|I_t - \hat{I}_t\| \tag{10}$$

Strided ray sampling Due to the large amount of network calls required for volumetric rendering, it is customary to randomly sample rays at training time [23]. This is possible because the volumetric view synthesis loss (Eq. 8) can be calculated at a per-pixel basis. The photometric loss (Eq. 10), however, requires a dense image, and thus is incompatible with random sampling. To circumvent this, while maintaining reasonable training times and memory usage, we use *strided ray sampling*. Fixed horizontal s_w and vertical s_h strides are used, and random horizontal $o_w \in [0, s_w - 1]$ and vertical $o_h \in [0, s_h - 1]$ offsets are selected to determine the starting point of the sampling process, resulting in $s_h \times s_w$ combinations. The rays can be

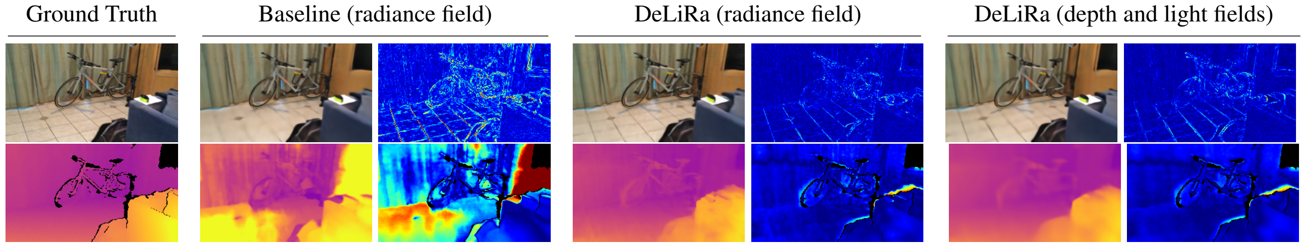


Figure 3: **Qualitative depth and view synthesis results** from unseen viewpoints, using different DeLiRa decoders. As a baseline, we show predictions obtained from a model trained without our contributions, leading to a degenerate learned geometry due to shape-radiance ambiguity (i.e., accurate view synthesis with poor depth predictions).

arranged to produce a downsampled image I'_t of resolution $\lfloor \frac{H-o_h}{s_h} \rfloor \times \lfloor \frac{W-o_w}{s_w} \rfloor$, with corresponding predicted image $\hat{I}'_t$ and depth map $\hat{D}'_t$. Note that the target intrinsics $\mathbf{K}'$ have to be adjusted accordingly, and context images do not need to be downsampled.

3.5. Light and Depth Field Decoding

We take advantage of the general nature of our framework to produce novel view synthesis and depth estimates in two different ways: *indirectly*, by compositing predictions from a volumetric decoder (Eq. 4); and *directly*, as the output of light and depth field decoders. Because light and depth field predictions [12] lack the multi-view consistency inherent to volumetric rendering [12, 42, 35], we augment the amount of available training data by including virtual supervision from volumetric predictions.

This is achieved by randomly sampling virtual cameras from novel viewpoints at training time, and using volumetric predictions as pseudo-labels to supervise the light and depth field predictions. Virtual cameras are generated by adding translation noise $\boldsymbol{\epsilon}_v = [\epsilon_x, \epsilon_y, \epsilon_z]_v \sim \mathcal{N}(0, \sigma_v)$ to the pose of an available camera, selected randomly from the training dataset. The viewing angle is set to point towards the center of the original camera, at a distance of d_{max}, which is also disturbed by $\boldsymbol{\epsilon}_c = [\epsilon_x, \epsilon_y, \epsilon_z]_c \sim \mathcal{N}(0, \sigma_v)$. We can use information from this virtual camera to decode a predicted volumetric image $\hat{I}_v$ and depth map $\hat{D}_v$, as well as a predicted image $\hat{I}_l$ and depth map $\hat{D}_d$ from the light and depth field decoders. We use the MSE loss (Sec. 3.4.1) to supervise $\hat{I}_l$ relative to $\hat{I}_v$, as well as the L1-log loss to supervise $\hat{D}_d$ relative to $\hat{D}_v$, resulting in the virtual loss:

$$\mathcal{L}_v = \left\| \log(\hat{D}_r) - \log(\hat{D}_d) \right\|_1 + \left(\hat{I}_r - \hat{I}_l \right)^2 \quad (11)$$

Note that the self-supervised losses from Sec. 3.4 are also applied to the original light and depth field predictions.

3.6. Depth Field Volumetric Guidance

In the previous section we described how volumetric predictions can be used to improve light and depth field estimates, by introducing additional supervision in the form of virtual cameras. Here, we show how depth field predictions can be used to improve the efficiency of volumetric estimates, by sampling from areas near the observed surface. Although more involved strategies have been proposed [3, 31, 25], we found that sampling from a Gaussian distribution $\mathcal{N}(\hat{D}_d, \sigma_g)$, centered around $\hat{D}_d$ with standard deviation σ_g, provided optimal results. Importantly, all these strategies require additional information from pre-trained depth networks or sparse supervision, whereas ours use predictions generated by the same network, learned from scratch and decoded from the same representation.

3.7. Training Schedule

Since all predictions are learned jointly, we use a training schedule such that depth field estimates can reach a reasonable level of performance before serving as guidance for volumetric sampling. In the first 400 epochs (10% of the total number of steps), we consider K ray samples, and depth field guidance (DFG) is not used. Afterwards, K_g samples are relocated to DFG, and drawn instead from $\mathcal{N}(\hat{D}_d, \sigma_g)$. After another 400 epochs, we once again reduce the amount of ray samples by K_g, but this time without increasing the number of depth field samples, which decreases the total number of samples used for volumetric rendering. This process is repeated every 400 epochs, and at $K = 0$ ray sampling is no longer performed, only DFG with K_g samples.

Moreover, we note that the multi-view photometric objective is unable to model view-dependent artifacts, since it relies on explicit image warping. Thus, we gradually remove this regularization, so that our network can first converge to the proper implicit scene geometry, and then use it to further improve novel view synthesis. In practice, after every 400 epochs we decay $\mathcal{L}_p$ by a factor of 0.8, and completely remove it in the final 800 epochs.

4. Experimental Results

4.1. Dataset

The primary goal of DeLiRa is to enable novel view synthesis in the limited viewpoint diversity setting, which is very challenging for implicit representations (prior work on few-shot NeRF [17, 26] is limited to synthetic or tabletop settings). Thus, following standard protocol [3, 46, 31] we use the ScanNet dataset [1] as our evaluation benchmark. This is a challenging benchmark, composed of real-

Method		Superv.	*Lower is better*				*Higher is better*		
			Abs.Rel.	Sq.Rel.	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
COLMAP [32]		–	0.462	0.631	1.012	1.734	0.481	0.514	0.533
ACMP [49]		D	0.194	0.171	0.455	0.306	0.731	0.881	0.942
DELTAS [34]		D	0.100	0.032	0.207	0.128	0.862	0.992	0.999
DeepV2D [39]		D	0.082	0.023	0.171	0.109	0.941	0.991	0.998
Atlas [24]		D	0.078	0.063	0.244	0.269	0.929	0.954	0.959
NeRF [23]		–	0.393	1.485	1.090	0.521	0.489	0.732	0.828
CVD [22]		–	0.099	0.030	0.194	0.127	0.901	0.988	0.997
NerfingMVS [46] (w/o filter)		C	0.063	**0.014**	0.145	0.094	0.954	0.991	0.999
NerfingMVS [46] (w/ filter)		C	0.061	**0.014**	0.134	0.086	0.960	**0.995**	0.999
DeLiRa	R	–	0.055	**0.014**	**0.131**	**0.082**	**0.970**	0.994	0.999
	D	–	**0.054**	0.028	0.138	0.088	0.966	0.992	**1.000**

Table 1: **Average depth synthesis results** on *ScanNet-Frontal*. *Superv.* indicates the source of depth supervision: *D* for ground truth, and *C* for COLMAP predictions. DeLiRa outperforms all other methods, despite not requiring supervision. Moreover, our depth field results (D) are on par with radiance predictions (R), and can be generated with a single query.

Method		*Scene 0000*		*Scene 0079*		*Scene 0158*		*Scene 0316*		*Scene 0521*		*Scene 0553*		*Scene 0616*		*Scene 0653*	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NSVF [21]		23.36	0.823	26.88	0.887	31.98	0.951	22.29	0.917	27.73	0.892	31.15	0.947	15.71	0.704	28.95	0.929
SVS [30]		21.39	0.914	25.18	**0.923**	29.43	0.953	20.63	0.941	27.97	**0.924**	30.95	0.968	**21.38**	**0.899**	27.91	0.965
NeRF [23]		18.75	0.751	25.48	0.896	29.19	0.928	17.09	0.828	24.41	0.871	30.76	0.950	15.76	0.699	30.89	0.953
NerfingMVS [46]		22.10	0.880	27.27	0.916	30.55	0.948	20.88	0.899	28.07	0.901	32.56	0.965	18.07	0.748	31.43	0.964
DeLiRa	R	**25.88**	**0.919**	28.01	0.916	34.68	0.969	**23.31**	**0.948**	28.97	0.909	**36.32**	**0.981**	20.27	0.851	33.70	0.967
	L	25.34	0.907	**28.48**	0.920	**35.77**	**0.980**	23.18	0.937	**29.22**	0.916	35.94	0.974	19.18	0.832	**34.63**	**0.973**

Table 2: **Per-scene view synthesis results**, on *ScanNet-Frontal*. DeLiRa improves view synthesis results (PSNR) in all considered scenes (+9.8% ± 4.1%), relative to the previous state of the art [46].

Method		Superv.	*Depth*	*View Synthesis*		
			RMSE↓	PSNR↑	SSIM↑	LPIPS↓
NeRF [23]		-	1.163	19.03	0.670	0.398
DS-NeRF [3]		C	0.423	20.94	0.721	0.330
NerfingMVS [46]†		C	0.469	16.45	0.641	0.488
DDP-NeRF [31]†		C	0.504	20.71	0.719	0.337
DDP-NeRF [31]†		D+C	0.229	21.02	0.742	**0.289**
DeLiRa	R	-	0.215	**21.64**	**0.761**	0.302
	DL	-	**0.213**	21.26	0.748	0.305

Table 3: **Depth and view synthesis results** on *ScanNet-Rooms*. The *superv.* column indicates the source of depth supervision: D denotes ground-truth depth maps, and C denotes COLMAP predictions. The symbol † indicates the use of separate depth networks, pre-trained on additional data.

world room-scale scenes subject to motion blur and noisy calibration [31]. For a fair comparison with other methods, we consider two different training and testing splits: *ScanNet-Frontal* [46], composed of 8 scenes, and *ScanNet-Rooms* [31], composed of 3 scenes. For more details, please refer to the supplementary material.

4.2. Volumetric Depth and View Synthesis

First, we evaluate the performance of DeLiRa focusing on volumetric predictions. Improvements in depth synthesis are expected to lead to improvements in view synthesis, which we validate in the following section. In Tab. 1 we compare our depth synthesis results on *ScanNet-Frontal* with several classical methods [49, 34, 39, 24], all of which require ground-truth depth maps as a source of supervision. We also consider NeRF [23], that only optimizes for volumetric rendering, as well as CVD [22] and NerfingMVS[46], that use a pre-trained depth network fine-tuned on in-domain sparse information. Even though DeLiRa operates in the same setting as NeRF (i.e., only posed images), it still achieves substantial improvements over all other methods. In particular, DeLiRa improves upon NeRF by 86.3% in absolute relative error (AbsRel) and 88.0% in root mean square error (RMSE), as well as improving upon the previous state of the art [46] by 14.2% in Abs.Rel. and 9.6% in RMSE. Similar trends are also observed for view synthesis, where DeLiRa improves upon [46] in PSNR by 11.1%, as shown in Tab. 2. We attribute this to the fact that our regularization leverages dense direct self-supervision from the environment, rather than relying on sparse noisy samples to fine-tune a pre-trained network.

In Tab. 3 we show a similar evaluation on *ScanNet-Rooms*, which constitutes a more challenging setting due to a larger area coverage (an entire room, as opposed to a local region). In fact, as shown in [31], most methods struggle in this setting: NeRF [23] generates "floaters" due to limited viewpoint diversity, DS-NeRF [3] is prone to errors in sparse depth input, and the error map calcula-

	Version	Decoder	Depth↓ AbsRel	Depth↓ RMSE	View Synth.↑ PSNR	View Synth.↑ SSIM
A	Monodepth (self-sup.)	-	0.096	0.268	—	—
B	COLMAP sup.	R	0.134	0.401	26.17	0.862
C	DeLiRa (w/o self-sup.)	R	0.245	0.570	27.64	0.895
D	MLP (w/o self-sup.)	R	0.238	0.546	27.50	0.882
E	MLP (self-sup.)	R	0.068	0.163	27.51	0.887
F	DeLiRa (1-MLP)	DL	0.057	0.133	27.34	0.889
G	Volumetric-only	R	0.059	0.142	27.87	0.891
H	Depth / Light-only	DL	0.123	0.344	23.56	0.757
I	Distilled (separate $\mathcal{S}_{DL}$)	DL	0.063	0.152	27.52	0.881
J	Distilled (shared $\mathcal{S}_R$)	DL	0.066	0.169	27.33	0.868
K	Distilled (shared $\mathcal{S}_R$, ~~VCA~~)	DL	0.076	0.181	26.16	0.849
L	Separate $\mathcal{S}_R$ and $\mathcal{S}_{DL}$	R	0.058	0.140	28.04	0.898
		DL	0.061	0.152	27.77	0.892
M	Without VCA	R	0.077	0.210	27.74	0.877
		DL	0.079	0.227	27.25	0.855
N	Without DFG	R	0.059	0.145	28.27	0.908
		DL	0.055	0.139	28.73	0.924
O	Without vanishing $\mathcal{L}_p$	R	0.056	0.135	28.59	0.922
		DL	0.055	0.144	28.74	0.918
	DeLiRa	R	0.055	**0.131**	28.98	0.934
		DL	**0.054**	0.138	**29.10**	**0.932**

Table 4: **Ablative analysis** of the various components of our proposed DeLiRa method, on *ScanNet-Frontal*. *R*, *D*, and *L* indicate radiance, depth, and light field predictions.

tion of NerfingMVS [46] fails when applied to larger areas. DDP-NeRF [31] circumvents these issues by using an additional uncertainty-aware depth completion network, trained on RGB-D data from the same domain. Even so, our simpler approach of regularizing the volumetric depth using a multi-view photometric objective leads to a new state of the art, both in depth and novel view synthesis.

4.3. Light and Depth Field Performance

In addition to volumetric rendering, DeLiRa also generates light and depth field predictions, that can be efficiently decoded with a single network forward pass. We report these results in the same benchmarks, achieving state-of-the-art performance comparable to their corresponding volumetric predictions. In the supplementary material we explore the impact of using different decoder architectures, noticing that deeper networks yield significant improvements in view synthesis, which is in agreement with [42]. Interestingly, we did not observe similar improvements in depth synthesis, which we attribute to the lack of higher frequency details in this task.

4.4. Ablative Analysis

Here we analyse the various components and design choices of DeLiRa, to evaluate the impact of our contributions in the reported results. Our findings are summarized in Tab. 4, with qualitative results in Figs. 3 and 4.

Multi-view Photometric Objective. Firstly, we ablate the multi-view photometric loss, used as additional regularization to the single-frame view synthesis loss. By removing this regularization (**C**), we observe significantly worse depth results (0.245 vs 0.054 AbsRel), as well as some view synthesis degradation (27.64 vs 28.96 PSNR). This is evidence that volumetric rendering is unable to properly learn scene geometry with low viewpoint diversity, and that accurate view synthesis can be obtained with degenerate geometries. Alternatively, we trained a self-supervised monocular depth network [6] using the same data, and achieved substantially better performance than volumetric rendering (**A**), however still worse than DeLiRa (0.096 vs 0.054 AbsRel, with qualitative results in the supplementary material). These indicate that our hybrid approach improves over any single objective: photometric warping explicitly enforces multi-view consistency, while volumetric rendering implicitly learns 3D geometry. We also experimented with replacing the multi-view photometric objective with COLMAP supervision (**B**). As pointed out in other works [46, 3, 31], these predictions are too sparse and noisy to be used without pre-trained priors, leading to worse results (0.134 vs 0.054 AbsRel, 26.17 vs 28.96 PSNR).

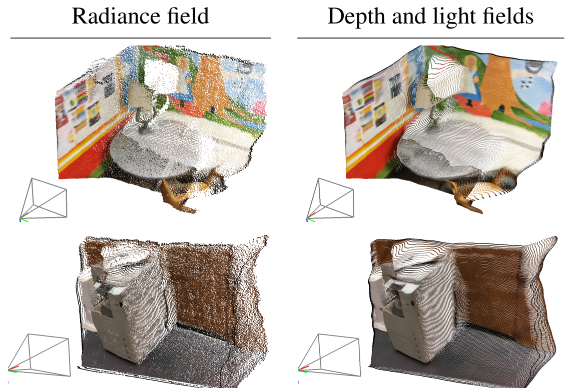


Figure 4: **Reconstructed pointclouds** from novel views, using color and depth predictions from different decoders.

Architecture. Next, we compare our auto-decoder architecture with a standard NeRF-style MLP [23] (**D**). Instead of decoding from the latent space, we map volumetric embeddings $\mathcal{E}_{vol}$ directly into $(\mathbf{c}, \sigma)$ estimates (Sec. 3.3). As we can see, this approach leads to worse results in depth synthesis (0.068 vs 0.054 AbsRel) and view synthesis (27.50 vs 28.96 PSNR), however it still benefits from photometric regularization (0.068 vs 0.238 AbsRel when removed) (**E**). This is in accordance with [28], in which, for radiance and light field networks, an auto-decoder with a learned latent representation outperformed MLP baselines.

Moreover, replacing our residual light field decoder [42] with a single MLP (**F**) degrades novel view synthesis (27.34 vs 28.96 PSNR), due to a lack of high frequency details.

Joint decoding. Here we consider the joint learning of depth, light, and radiance field predictions from the same latent space. We evaluate models capable of only volumetric rendering (**G**), or only light and depth field synthesis (**H**). As expected, depth and light field-only predictions

Version	Decoder	Inference	
		Speed (FPS)	*Memory (GB)*
NeRF	-	0.35	36.54
DeLiRa (w/o DFG)	R	0.82	29.98
DeLiRa (1-MLP)	L	351.1	4.34
DeLiRa	D	378.4	4.21
	L	118.2	6.88
	R	37.3	11.75

Table 5: **Efficiency analysis** for different DeLiRa decoders, at inference time (with 192×320 resolution). *w/o DFG* indicates the removal of depth field guidance, and *1-MLP* indicates a single linear layer in the light field decoder.

greatly degrade without the view diversity from virtual cameras, that leads to overfitting (0.123 vs 0.054 AbsRel, 23.56 vs 28.96 PSNR). Interestingly, volumetric-only predictions also degraded, which we attribute to the use of a shared latent space, that is optimized to store both representations.

Distillation. We also investigated augmenting a pre-trained volumetric representation to also decode depth and light field predictions. Three scenarios were considered: separate latent spaces (**I**), and shared latent spaces with (**J**) and without (**K**) virtual camera augmentation (VCA). When separate latent spaces are used, we observe a substantial improvement in depth and light field performance over single-task learning (0.123 vs 0.063 AbsRel, 27.52 vs 28.96 PSNR). We attribute this behavior to VCA, since this is the only way these two representations interact with each other. Interestingly, a similar performance is achieved using shared latent spaces (0.066 vs 0.063 AbsRel, 27.52 vs 27.33 PSNR), even though $\mathcal{S}_R$ is no longer optimized. This is an indication that the radiance latent space can be repurposed for depth and light field decoding without further training. Moreover, removing VCA in this setting did not degrade performance nearly as much as when separate latent spaces were used (0.076 vs 0.123 AbsRel, 26.16 vs 23.56 PSNR). This is further evidence that radiance representation provides meaningful features for depth and light field decoding, including the preservation of implicit multi-view consistency.

Joint training. Here we evaluate the benefits of jointly training volumetric, depth, and light fields under different conditions. Three settings were considered: the use of separate latent spaces $\mathcal{S}_R$ and $\mathcal{S}_{DL}$ (**L**), as well as the removal of VCA (**M**) or depth field guidance (DFG) (**N**). A key result is that the use of a shared latent space improves results (0.061 vs 0.054 AbsRel, 27.77 vs 28.96 PSNR), as further evidence that both representations produce similar learned features. Moreover, the removal of VCA or DFG leads to overall degradation. Finally, we show that vanishing $\mathcal{L}_p$ (Sec. 3.7) improves novel view synthesis, by enabling the proper modeling of view-dependent artifacts (**O**). Interestingly, it does not degrade depth synthesis, indicating that our learned geometry is stable and will not degrade if the photometric regularization is removed. However, it is fundamental in the initial stages of training, as shown in (**C**) when it is removed altogether (0.245 vs 0.054 AbsRel).

4.5. Computational Efficiency

In Tab. 5 we report inference times and memory requirements using different DeLiRa decoders and variations (for hardware details please refer to the supplementary material). Two different components are ablated: depth field guidance (DFG), as described in the Sec. 3.6, and the number of MLP layers in the light field decoder. As expected, depth and light field predictions are substantially faster than volumetric predictions. Furthermore, volumetric prediction efficiency can be greatly improved using DFG (0.82 to 37.3 FPS) to decrease the number of required ray samples (note that this improvement includes the additional cost of evaluating the depth field decoder). Interestingly, even without DFG our auto-decoder architecture is roughly 2 times faster than a traditional NeRF-style MLP [23]. Moreover, using a single MLP layer for light field decoding speeds up inference by roughly 3 times (118.2 to 378.4 FPS), at the cost of some degradation in novel view synthesis (Tab. 4, **F**).

5. Limitations

Our method still operates on a scene-specific setting, and thus has to be retrained for new scenes. Our method also does not address other traditional limitations of NeRF-like approaches, such as extrapolation to unseen areas and unbounded outdoor scenes. However, our contributions (i.e., the shared latent representation and photometric regularization) can be directly used to augment methods that focus on such scenarios. Finally, DeLiRa requires overlap between images to enable multi-view photometric self-supervision, and thus is not suitable for very sparse views.

6. Conclusion

This paper introduces the multi-view photometric objective as regularization for volume rendering, to mitigate shape-radiance ambiguity and promote the learning of geometrically-consistent representations in cases of low viewpoint diversity. To further leverage the geometric properties of this learned latent representation, we propose DeLiRa, a novel transformer architecture for the joint learning of depth, light, and radiance fields. We show that these three tasks can be encoded into the same shared latent representation, leading to an overall increase in performance over single-task learning without additional network complexity. As a result, DeLiRa establishes a new state of the art in the ScanNet benchmark, outperforming methods that rely on explicit priors from pre-trained depth networks and noisy supervision, while also enabling real-time depth and view synthesis from novel viewpoints.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 5

[2] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. 3

[3] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 3, 5, 6, 7

[4] Jiading Fang, Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Greg Shakhnarovich, Adrien Gaidon, and Matthew Walter. Self-supervised camera self-calibration from video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 1, 2

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017. 2

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 7

[7] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[8] Vitor Guizilini, Rares Ambrus, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[9] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4

[10] Vitor Guizilini, Kuan-Hui Lee, Rares Ambrus, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. In *Robotics and Automation Letters (RA-L)*, 2022. 1

[11] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. In *Robotics and Automation Letters (RA-L)*, 2022. 2

[12] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus, Greg Shakhnarovich, Matthew Walter, and Adrien Gaidon. Depth field networks for generalizable multi-view scene representation. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5

[13] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12902–12911, 2022. 2

[14] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv:1506.02025*, 2015. 4

[16] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 2

[17] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021. 5

[18] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[19] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[20] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[21] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 6

[22] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4), 2020. 6

[23] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 1, 2, 4, 6, 7, 8

[24] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 6

[25] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton S. Kaplanyan, and Markus Steinberger. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance

Fields using Depth Oracle Networks. *Computer Graphics Forum*, 2021. 2, 3, 5

[26] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2, 5

[27] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[28] Daniel Rebain, Mark J Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields. *arXiv preprint arXiv:2209.10684*, 2022. 7

[29] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. *arXiv preprint arXiv:2111.14643*, 2021. 2

[30] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 6

[31] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 3, 5, 6, 7

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 6

[33] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[34] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Deltas: Depth estimation by learning triangulation and densification of sparse points. In *ECCV*, 2020. 6

[35] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5

[36] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[37] Jiexiong Tang, Rares Ambrus, Vitor Guizilini, Sudeep Pillai, Hanme Kim, Patric Jensfelt, and Adrien Gaidon. Self-Supervised 3D Keypoint Learning for Ego-Motion Estimation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2020. 1

[38] Jiexiong Tang, Hanme Kim, Vitor Guizilini, Sudeep Pillai, and Rares Ambrus. Neural outlier rejection for self-supervised keypoint learning. In *International Conference on Learning Representations*, 2020. 1

[39] Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 6

[40] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 1

[41] Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural ray surfaces for self-supervised learning of depth and ego-motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 1, 2

[42] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, 2022. 2, 5, 7

[43] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 4

[45] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021. 1

[46] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5, 6, 7

[47] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 1, 2

[48] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 3

[49] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 6

[50] John Yang, Le An, Anurag Dixit, Jinkyu Koo, and Su Inn Park. Depth estimation with simplified transformer, 2022. 2

[51] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision:

Learning dense object descriptors from neural radiance fields. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6496–6503. IEEE, 2022. 3

[52] Wang Yifan, Carl Doersch, Relja Arandjelović, João Carreira, and Andrew Zisserman. Input-level inductive biases for 3D reconstruction. *arXiv preprint arXiv:2112.03243*, 2021. 4

[53] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 3

[54] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 1