Canonical Factors for Hybrid Neural Fields

Brent Yi[1] Weijia Zeng[1] Sam Buchanan[2] Yi Ma[1]

[1]UC Berkeley [2]TTI-Chicago

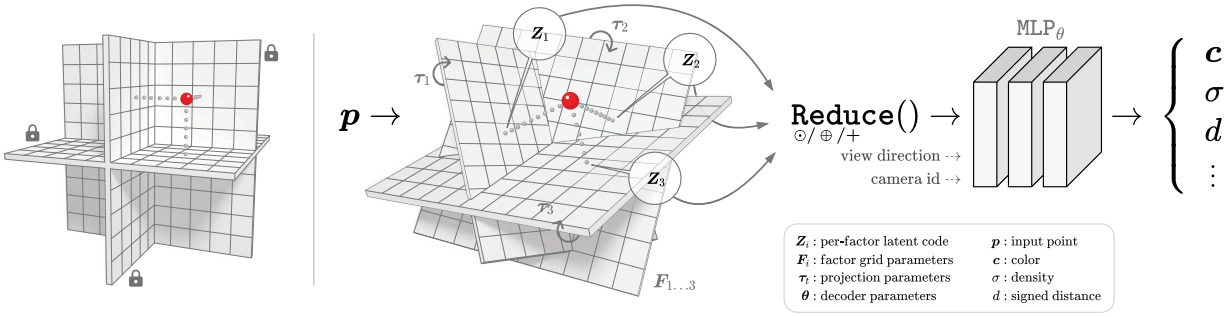


Figure 1: **Learned transforms for factored feature volumes.** Latent decompositions with fixed, axis-aligned projections (left) introduce biases for axis-aligned signals. A more robust, transform-invariant latent decomposition (TILTED) can be recovered by treating projections to feature grids as learnable functions, here parameterized by $\boldsymbol{\tau}_t$.

Abstract

Factored feature volumes offer a simple way to build more compact, efficient, and intepretable neural fields, but also introduce biases that are not necessarily beneficial for real-world data. In this work, we (1) characterize the undesirable biases that these architectures have for axis-aligned signals—they can lead to radiance field reconstruction differences of as high as 2 PSNR—and (2) explore how learning a set of canonicalizing transformations can improve representations by removing these biases. We prove in a simple two-dimensional model problem that a hybrid architecture that simultaneously learns these transformations together with scene appearance succeeds with drastically improved efficiency. We validate the resulting architectures, which we call TILTED, using 2D image, signed distance field, and radiance field reconstruction tasks, where we observe improvements across quality, robustness, compactness, and runtime. Results demonstrate that TILTED can enable capabilities comparable to baselines that are 2x larger, while highlighting weaknesses of standard procedures for evaluating neural field representations.

1. Introduction

Our physical world layers complexity on top of regularity. Tucked below the details that imbue our environments with character—the intricate fibers of a fine-grained veneer, or the light-catching specularities of everyday metal, plastic, and glass—one finds the simple geometric primitives and symmetries associated with built and natural structures. The challenge of representations for the world, such as those used for 3D reconstruction, anchors itself in the interaction between the two ends of this dichotomy: point clouds and voxel grids offer versatility, but their inability to capture structure results in resource usage that can grow too intractably to be useful for complex details; meshes harness the uniformity of surfaces for compactness, but still fail on entities with details that step outside of an acceptable regime—consider fog or deviations on curves.

In this work, we build on the idea that scalably capturing the details of a complex signal is only possible when a representation enables capture of its structure. We use this theme to study and improve state-of-the-art hybrid neural fields, which typically pair neural decoders with factored feature volumes [32, 51, 63, 66–68]. Aided by an ability to exploit sparse and low-rank structure, factorization is simple to implement and offers a host of advantages, such as compactness, efficiency, and interpretability. However, naive factorization also introduces the disadvantage of an implicit frame of representation, which is not guaranteed to be aligned with the structure of scenes or signals one aims to represent. Drawing on insights from both low-rank texture extraction [9] and implicit regularization in optimization methods for factorization [21, 42], we theoretically characterize the im-

portance of this alignment and then show how it can motivate practical improvements to neural field architectures that rely on factored feature volumes. Our contributions are as follows:

(1) We theoretically characterize the fragility of factored grids in a two-dimensional model problem, where resource efficiency on simple-to-capture structures can be undermined even by small planar rotations (Section 3). We prove that this fragility can be overcome by *jointly optimizing* over the parameters of the representation and a transformation of domain capturing pose, when the underlying structure is well-aligned in some frame of representation.

(2) We study how this same weakness affects practical neural field architectures, where it can lead to radiance field accuracy differences of as high as 2 PSNR. We propose optimization of more robust, transform-invariant latent decompositions (TILTED) via a modification of existing factored representations (Section 4). TILTED models recover *canonical factors* by jointly recovering factors of a decomposed feature volume with a set of canonicalizing transformations, which are simple to add to existing factorization techniques.

(3) We evaluate the TILTED family of models on three tasks: 2D image, signed distance field, and neural radiance field reconstruction (Section 5). Our experiments highlight biases in existing neural field architecture and evaluation procedures, while demonstrating TILTED's advantages across quality, robustness, compactness, and runtime. For real-world scenes, TILTED can simultaneously improve reconstruction, halve memory consumption, and accelerate training times by 25%.

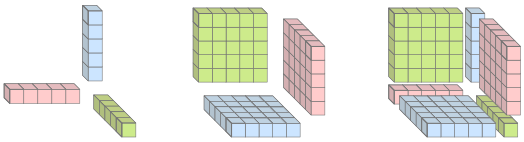
2. Related Work

2.1. Neural Fields

In its standard form, a neural field is implemented using an MLP that takes coordinates as input and returns a vector of interest. For example, a basic neural radiance field [29] with network parameters $\boldsymbol{\theta}$ maps spatial positions $\boldsymbol{p} = (p_x, p_y, p_z) \in \mathbb{R}^3$ to RGB colors $\boldsymbol{c} \in [0,1]^3$ and densities $\sigma \in \mathbb{R}_{\geq 0}$:

$$\boldsymbol{p} \xrightarrow{\texttt{MLP}_{\boldsymbol{\theta}}} (\boldsymbol{c}, \sigma). \tag{2.1}$$

This framework is highly versatile. Instead of only position, inputs can include additional conditioning information such as specularity-enabling view directions [29], per-camera appearance embeddings [38], or time [66]. Instead of radiance, possible outputs also include representations of binary occupancy [24, 25], signed distance functions [26, 27], joint representations of surfaces and radiance [39, 45, 46], actions [52, 62], and semantics [44, 53, 56, 71]. The core ideas behind TILTED are not tied to specific input or output modalities.



(a) CP (b) Tri-plane (c) Vector-matrix

Figure 2: **Tensor decompositions for 3D features volumes studied by prior work** [32, 51, 66]. Note that all assume a fixed, axis-aligned structure; TILTED instead proposes to learn transformations of this structure.

2.2. Hybrid Neural Fields

When a single MLP is used as a data structure, as in (2.1), all stored information needs to be encoded and entangled in the network weights $\boldsymbol{\theta}$. The result is expensive for both training and inference. To address this, several works have proposed forms of *hybrid neural fields*, which have two components: an explicit geometric data structure from which latent vectors are interpolated and a neural decoder [43, 55]. In the case of 3D coordinate inputs and radiance outputs, as in (2.1), these architectures can be instantiated as

$$\boldsymbol{p} \xrightarrow{\texttt{VoxelTrilerp}_{\boldsymbol{\phi}}} \boldsymbol{Z} \xrightarrow{\texttt{MLP}_{\boldsymbol{\theta}}} (\boldsymbol{c}, \sigma), \tag{2.2}$$

where $\texttt{VoxelTrilerp}_{\boldsymbol{\phi}}$ interpolates the 'latent grid' parameters $\boldsymbol{\phi}$ to produce a latent feature $\boldsymbol{Z} \in \mathbb{R}^d$, which is then decoded to standard radiance field outputs by an MLP with parameters $\boldsymbol{\theta}$.

Instead of implementing the latent feature volume $\boldsymbol{\phi}$ as a dense voxel grid, a common pattern is to decompose this tensor into lower-dimensional factors $\boldsymbol{\phi} = \{\boldsymbol{F}_1 \dots \boldsymbol{F}_F\}$. In this way, factored hybrid neural field approaches [32, 51, 59, 63, 64, 66] generalize (2.2) by (i) **projecting** input coordinates onto each of F lower-dimensional coordinate spaces, (ii) **interpolating** F feature vectors from the corresponding factors, and (iii) **reducing**—for example, by concatenation, multiplication, or addition—the set of latent features into the final latent $\boldsymbol{Z}$:

$$\begin{aligned}\boldsymbol{Z} = \texttt{Reduce}\big(&\big[\texttt{Interp}_{\boldsymbol{F}_1}(\texttt{Proj}_1(\boldsymbol{p}))\big], \dots, \\ &\big[\texttt{Interp}_{\boldsymbol{F}_F}(\texttt{Proj}_F(\boldsymbol{p}))\big]\big).\end{aligned} \tag{2.3}$$

Interpolating only on lower-dimensional feature grids $\boldsymbol{F}_1, \dots, \boldsymbol{F}_F$, which may be 1D or 2D when $\boldsymbol{p}$ is 3D or higher, provides efficiency advantages.

Hybrid neural fields offer a unique set of advantages. In contrast to techniques based on caching and distillation, which require a pretrained neural network [33–35, 41, 47, 57], hybrid neural field architectures accelerate both training and evaluation. They also offer unique opportunities in generation [32, 60, 65], real-time rendering [74, 78] up-

sampling [51], incremental growth [49, 69, 72, 79], interpretable regularization [66], anti-aliasing [68], exploiting sparsity [67], and dynamic scene reconstruction [63, 73, 75].

Existing latent grid factorization methods, which can be formalized under (2.3) (Appendix E), constrain the `Proj` operations to axis-aligned projections (Figure 2). Similar to what has been observed in axis-aligned positional encodings [30] (and pointed out by concurrent work [67]), this results in a bias for axis-aligned signals. TILTED proposes to learn a set of transforms that removes this bias.

2.3. Learning With Transformations of Domain

TILTED improves reconstruction performance via optimization over transformations of domain, a mathematical idea dating back to the earliest days of computer vision. A concrete example is the image registration problem [3, 4, 7, 8], where we seek a transformation $\boldsymbol{\tau}$ that deforms an observed image $\boldsymbol{Y}$ to match a target $\boldsymbol{X}$ via gradient descent. TILTED takes inspiration from many tried-and-tested techniques for robustly solving problems of this type, including coarse-to-fine fitting and other regularization schemes (e.g., [5, 6, 15]). Although this type of 'supervised' registration is studied in the context of neural fields [54], it is less relevant to learning neural implicit models like (2.2) and (2.3), where ground-truth is rarely available. Instead, we build TILTED around an insight of Zhang *et al.* [13]: *for scenes consisting of natural or built environments, the transformation that 'aligns' the scene with its intrinsic coordinate frame yields the most compact representation.* In the case of 2D images, Zhang *et al.* [13] instantiate this principle as a search for a transformation that minimizes the sum of the singular values of the image, a relaxation of the rank:

$$\min_{\boldsymbol{\tau}} \|\boldsymbol{Y} \circ \boldsymbol{\tau}\|_*. \tag{2.4}$$

TILTED combines this core insight with the emerging theoretical understanding of *implicit regularization* in various overparameterized matrix factorization problems [21, 42, 77], which implies that an implicit bias toward low-rank structures in factored grid representations learned with gradient descent obviates the explicit rank regularization of (2.4).

A parallel line of work seeks to imbue a broader family of neural network architectures with invariance or 'equivariance' to transformations or symmetries that the network should respect. These include parallel channel networks [10, 16, 18, 36], approaches based on pooling over transformations [12, 14, 22], and approaches with learnable deformation offsets [17, 19, 20, 48]. Other approaches aim to construct networks that are transformation-invariant by design [11, 23, 50]. With TILTED, we demonstrate how to combine the benefits of transformation invariance with a variety of hybrid neural field architectures—as we discuss in Section 3, naive factorizations can be severely limited.

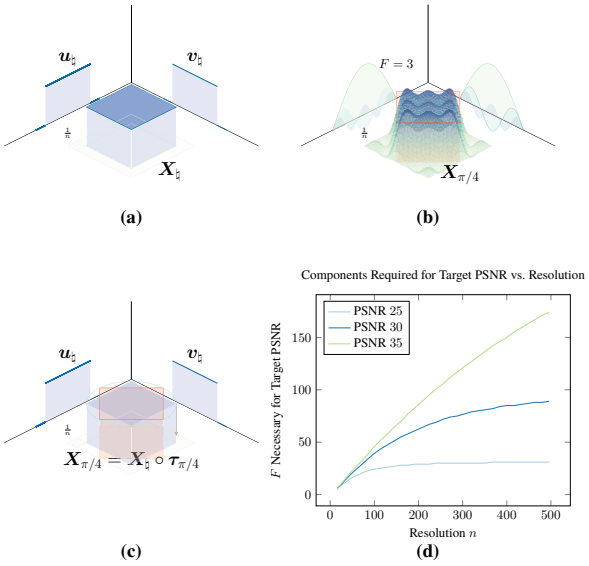


Figure 3: **Limitations of low-rank feature grids. (a)**: The square template $\boldsymbol{X}_\natural$ is axis-aligned, and has a maximally-compact (rank one) representation. **(b)**: After a rotation by $\pi/4$ radians, the square template (in red) only changes its orientation, but its approximability by a low-rank grid deteriorates dramatically. We draw the scaled eigenvectors and approximation for $F = 3$. **(c)**: By optimizing over transformations, a rank-one grid can be used to represent all rotations of $\boldsymbol{X}_\natural$. **(d)**: We plot the number of components needed to achieve varying PSNR levels as a function of image resolution for $\nu = \pi/4$. The number of components is always significantly larger than is necessary when transform optimization is used.

3. Low-Rank Grids Are Delicate Creatures

In this section, we theoretically justify the use of transform optimization in TILTED in a simple 2D image reconstruction problem. We omit the MLP decoder in (2.2) and focus only on the bottleneck imposed by the factored feature grid of (2.3). Note that the capacity of the MLP decoder is tightly constrained by performance considerations; both TensoRF [51] and K-Planes [66], for example, use only a single nonlinearity to decode density and proposal fields.

Concretely, let $\boldsymbol{X}_\natural \in \mathbb{R}^{n \times n}$ denote the grayscale image corresponding to the axis-aligned square pattern in Figure 3(a). We can decompose $\boldsymbol{X}_\natural$ as $\boldsymbol{X}_\natural = \boldsymbol{u}_\natural \boldsymbol{v}_\natural^*$, where $\boldsymbol{u}_\natural$ and $\boldsymbol{v}_\natural$ are one-dimensional square pulses aligned with the support of $\boldsymbol{X}_\natural$; $\boldsymbol{X}_\natural$ has rank one, and can be perfectly reconstructed by a maximally-compact low-rank feature grid. In contrast, consider exactly the same scene, but with an additional rotation by an angle of $\nu \in [0, \pi/4]$ applied to yield a transformed scene $\boldsymbol{X}_\nu = \boldsymbol{X}_\natural \circ \boldsymbol{\tau}_\nu$ (Figure 3(b)). As ν approaches its maximum value, the rank of the transformed

scene grows to a constant multiple of the resolution n, implying that *perfect* representation of $\boldsymbol{X}_\nu$ by a low-rank feature grid demands essentially as many components as a generic $n \times n$ matrix. Moreover, even *approximate* representation of the transformed scene by a pure low-rank grid is inefficient, as we prove:

Theorem 1 (informal version of Theorem F.1)**.** *There exist absolute constants $c_0, c_1 > 0$ such that for any target channel count $F \leq c_0 n^{1/6}$, every rank-F approximation $\hat{\boldsymbol{X}}$ to $\boldsymbol{X}_{\pi/4}$ satisfies*

$$\frac{1}{n^2}\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\pi/4}\right\|_{\mathrm{F}}^2 \geq \frac{c_1}{1+F}.$$

Theorem 1 asserts that a broad class of sublinear-rank approximations to $\boldsymbol{X}_\nu$ have mean squared error at least as large as the reciprocal number of components. Our proofs suggest this lower bound is tight up to logarithmic factors—in particular, as we illustrate numerically in Figure 3(d), target PSNR levels that are more stringent require larger grid ranks F as the image resolution grows. This situation stands in stark contrast to what can be achieved by capturing the structure of $\boldsymbol{X}_\natural$: there exists a transformed coordinate system in which $\boldsymbol{X}_\nu$ can be represented by a grid with $F = 1$, regardless of the image resolution. We prove that the $F = 1$ instantiation of TILTED successfully represents $\boldsymbol{X}_\nu$ by jointly optimizing over grid factors and transformations (Figure 3(c)):

Theorem 2 (informal version of Theorem F.2)**.** *The infinite-resolution limit of the optimization procedure*

$$\min_{\phi, \boldsymbol{u}} \left\|\boldsymbol{X}_{\pi/4} - (\boldsymbol{u}\boldsymbol{u}^*) \circ \boldsymbol{\tau}_\phi\right\|_{\mathrm{F}}^2 \tag{3.1}$$

solved with randomly-initialized constant-stepping gradient descent converges to the true parameters $(\pi/4, \boldsymbol{u}_\natural)$, up to symmetry, at a linear rate.

Theorem 2 provides theoretical grounding for TILTED's transformation optimization approach in an idealized setting: importantly, *there exist conditions under which the joint learning of the visual representation and pose parameters provably succeeds*. The proof of Theorem 2 reveals an important conceptual principle that underlies the success of this disentangled representation learning: there is a symbiotic relationship between the model's representation accuracy and its alignment accuracy, due to its constrained capacity (i.e., $F = 1$ feature channels). More precisely, incremental improvements to representation quality when the scene is aligned inaccurately help the model localize the scene content and create texture gradients that promote improvements to alignment; meanwhile, improvements to alignment allow the model to leverage its constrained capacity to more accurately represent the scene. In the remainder of the paper, we describe the necessary components to instantiate the optimization procedure (3.1) in practice.

4. TILTED

We design TILTED around two goals: **(1) Robustness.** TILTED aims for reconstruction ability that is invariant to rotations. As established theoretically in Section 3 and later empirically in Section 5, this does not hold for naively decomposed feature volumes. **(2) Generality.** TILTED does not attempt to re-invent the wheel; instead, it is designed to be compatible with and build directly upon existing approaches [51, 64, 66] for factoring feature volumes.

Rather than assuming that the projection functions $\texttt{Proj}_i$ in (2.3) are static and axis-aligned, the core idea of TILTED is to replace $\texttt{Proj}_i$ with learnable functions $\texttt{Proj}_{i,\boldsymbol{\tau}}$, where $\boldsymbol{\tau}$ is a set of learnable transformation parameters. By substituting into (2.3), the feature volume interpolation function then becomes:

$$\begin{aligned}\boldsymbol{Z} = \texttt{Reduce}\big(&\big[\texttt{Interp}_{\boldsymbol{F}_1}(\texttt{Proj}_{1,\boldsymbol{\tau}}(\boldsymbol{p}))\big], \ldots, \\ &\big[\texttt{Interp}_{\boldsymbol{F}_F}(\texttt{Proj}_{F,\boldsymbol{\tau}}(\boldsymbol{p}))\big]\big).\end{aligned} \tag{4.1}$$

The transformations $\boldsymbol{\tau}$ enable mapping from arbitrary scene coordinates to canonicalized domains for each factor $\boldsymbol{F}_i$. As llustrated in Figure 1, this can be interpreted as a spatial transformation of factors to a set of configurations that align best to the underlying structure of a signal or scene.

4.1. Applying Transformations

The design space for the parameterization of $\boldsymbol{\tau}$ and how it is applied to input coordinates $\boldsymbol{p}$ is large. We develop TILTED for the case where $\boldsymbol{\tau}$ is a set of T randomly initialized rotations $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1 \ldots \boldsymbol{\tau}_T\}$, parameterized by the unit circle $\mathbb{S}^1$ in 2D and $\mathbb{S}^3$ (the universal cover of the set of rotation matrices $\mathrm{SO}(3)$; i.e., unit quaternions) in 3D. We suffix variants with the value of T; TILTED-4, for example, refers to TILTED with 4 learned rotations.

All experiments build atop feature volumes studied in prior work: for 3D, the CP [1], vector-matrix [51], and K-Planes [66] decompositions, which are each detailed in Appendix E. K-Planes in 3D is equivalent to a tri-plane [32], but uses a multiplicative reduction. We characterize each decomposition architecture using the channel dimension d of its reduced latent vector $\boldsymbol{Z} \in \mathbb{R}^d$. We constrain T such that it evenly divides d, and apply rotations to the input coordinates such that each rotation $\boldsymbol{\tau}_t$ is used to compute d/T of the final output channels. This can be interpreted as a vectorized alternative to instantiating T instances of a given decomposition, each with channel count d/T, applying a different learned rotation to the input of each decomposition, and concatenating outputs. The resulting formulation has several desirable qualities:

Robustness. When $\boldsymbol{\tau}$ is defined by a family of transforms and optimized from a random initialization, we see two related advantages. First, as established in Section 3, the latent feature volume becomes able to represent signals

and scenes having non-axis-aligned poses with vastly improved parameter efficiency. Second, reconstruction quality becomes invariant to the transformation group encompassed by τ. When τ is constrained to rotations, a rotation applied to the scene becomes equivalent to a rotation applied to the random initialization of τ.

Convergence. Transformation optimization problems like camera registration are typically challenging and prone to local minima, but optimization in TILTED is better positioned to succeed. We initialize many transforms: for any given structure in a scene, only one of these many transforms needs to fall into the basin of attraction for success. Optimization of individual transforms is also highly symmetric. Consider rotation optimization over a 2D grid: each increment of 90 degrees results in a representation with equivalent structure. Our theoretical analysis, namely Theorem F.2, verifies that these properties are sufficient for optimization to succeed under idealized conditions.

Overhead. Finally, notice that rotations in this form are inexpensive both to store and apply. Standard hybrid neural fields can have on the order of millions of parameters; a library of geometric transformations requires only dozens. Because coordinate transformations reduce to simple matrix multiplications, the exerted runtime penalty is also small.

4.2. Coarse-to-Fine Optimization

When optimizing over transformations, high frequency signals produce undesirable local minima. We improve convergence via two coarse-to-fine optimization strategies.

Dynamic low-pass filtering. Similar to prior work [51, 61], we encode features interpolated in TILTED's RGB and SDF experiments with a Fourier embedding [30]. When these features are used, we adopt the coarse-to-fine strategy proposed for learning deformation in Nerfies [40] and for camera registration in BARF [37]. Given J frequencies, we weight the j-th frequency band via:

$$w_k^j(\eta_k) = \frac{1 - \cos(\pi \text{clamp}(\eta_k - j, 0, 1))}{2}$$

where k is the training step count and η_k is interpolated from a linear schedule $\in [0, J]$.

Two-phase optimization. Effective recovery of τ is coupled with the rank of latent decompositions. As rank is increased, high-frequency signals become easier to express and overfit to; as a target signal becomes more explainable without a well-aligned latent structure, optimizers have less incentive to push τ toward improved solutions.

To help disentangle the τ recovery from the capacity of latent feature volumes in radiance field experiments, we apply a two-phase strategy inspired by structure from motion, where procedures like the 8-point algorithm can be used to initialize Newton-based bundle adjustment. In the first phase, we train a hybrid field using a channel-limited CP decomposition, which has limited representational capacity. This produces "bottlenecked" MLP decoder parameters θ_{bneck}, feature grid parameters ϕ_{bneck}, and projection parameters τ_{bneck}. We discard all parameters but τ_{bneck}, and then simply set:

$$\tau_{\text{init}} = \tau_{\text{bneck}}$$

to initialize the final, more expressive neural field. Example reconstructions after each phase are displayed in Figure 4.

Figure 4: **Two-phase optimization.** Two TILTED neural fields are trained: the first trained using a rank-constrained *bottleneck* representation (left); all parameters are discarded except for the projection parameters τ_{bneck}, which are used for initialization of the final representation (right).

5. Experiments

5.1. 2D Image Reconstruction

To build intuition in a simple setting, we begin by studying TILTED for 2D image reconstruction with low-rank feature grids, analogous to our theoretical studies in Section 3. To evaluate sensitivity to image orientation, we evaluate two model variants—with an axis-aligned decomposition and with a TILTED decomposition—on four images rotated by angles sampled uniformly between 0 and 180, at 10 degree intervals. The setup of models can be interpreted as the 2D version of a CP decomposition-based neural field [51, 67]. In the axis-aligned case, latent grids are decomposed into $d = 64$ vector pairs, where each vector $\in \mathbb{R}^{128}$. The full latent grid can be computed by concatenating the outer products of each pair. In the TILTED variant, we introduce a set of $T = 8$ 2D rotations, each of which are applied to d/T vector pairs. Experiments are run by fitting a hybrid field with a 2-layer, 32-unit decoder to a randomly subsampled half of

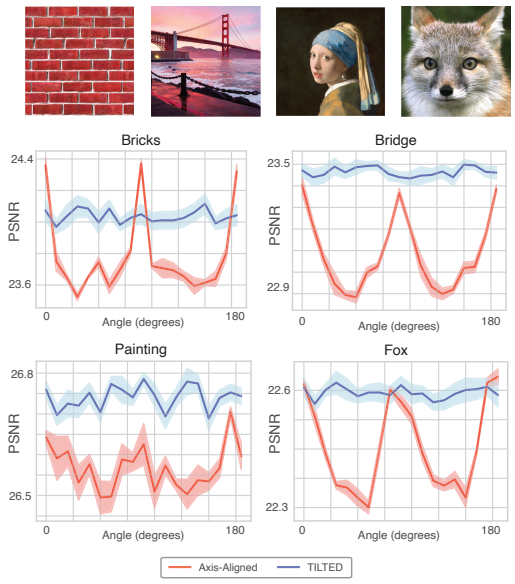


Figure 5: **Evaluation images and results for 2D image reconstruction.** We apply rotations to each input image, and plot holdout PSNR for a model trained at each angle. Axis-aligned feature decompositions are sensitive to transformations of the input, while TILTED retains a constant PSNR across angles.

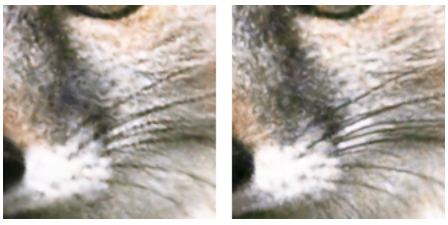


Figure 6: **Fine details without (left) and with (right) TILTED.** The TILTED reconstruction of the whiskers mitigates artifacts from axis-aligned factors.

the pixels in an image (in an approximation of radiance field reconstruction), and using the other half for evaluation. Results from this process over 5 seeds are reported in Figure 5. We observe from the experiments:

(1) TILTED improves robustness. When an axis-aligned decomposition is used, recovered PSNRs are more volatile, with a difference of as high as 1 PSNR for the *Bricks* test image. With the introduction of learned transforms, reconstruction quality becomes stable to input rotations.

(2) TILTED improves detail recovery. We qualitatively evaluate results by zooming into reconstructed images in Figure 6. TILTED improves reconstruction particularly in fine features like the whiskers, which are jagged and bottlenecked by the factorization in the axis-aligned case, but rendered with fewer artifacts when we apply TILTED.

IoU ↑	30	60	90
K-Planes	0.949±0.015	0.952±0.015	0.952±0.016
w/ TILTED	**0.989**±0.002	**0.990**±0.002	**0.991**±0.002
IoU ↑	45	90	135
Vector-Matrix	0.970±0.007	0.979±0.005	0.982±0.003
w/ TILTED	**0.982**±0.003	**0.989**±0.002	**0.988**±0.003

Table 1: **Aggregated metrics across models used for SDF experiments.** Three channel count variations are used for each latent decomposition structure. TILTED improves reconstructions consistently.

5.2. Signed Distance Field Reconstruction

Next, we study the impact of TILTED on reconstruction of signed distance fields. We follow the mesh sampling strategy used for studying signed distance fields in prior work [58, 64] to produce a set of 8M training points and 16M evaluation points, and then train hybrid fields based on both VM and K-Plane decompositions. Evaluation metrics are reported using intersection-over-union (IoU).

We sweep reconstructions based on both K-Planes and VM, with three channel counts for each architecture, on 8 different meshes. Each representation uses 3 resolutions—32, 128, and 256. For K-Planes, we use channel counts of 30, 60, and 90; for VM, we use channel counts of 45, 90, 135. All experiments use a 3-layer, 64-unit decoder and 5 transforms. We observe:

(1) Improved reconstructions across architectures and models. We report the average IoU for eight objects in Table 1. TILTED improves results for all decomposition and channel count variants. When we disaggregate results by model (Section B), TILTED outperforms its axis-aligned counterpart in all but one (of 48) examples.

(2) Implicit 3D regularization. To better understand how TILTED impacts SDF reconstruction, we apply marching cubes [2] to learned fields after training. Qualitative examples are shown in Figure 7. Renders reveal that the hybrid field architectures we use, which were proposed for and have not been extensively studied beyond the context of radiance fields, are prone to floating artifacts in recovered meshes. The typical solution for artifacts like these is to adjust the model size or regularization, for example to increase channel count or encourage spatial smoothness with total variation. We find that TILTED achieves a similar effect without expanding the factorization size or changing the optimized cost function.

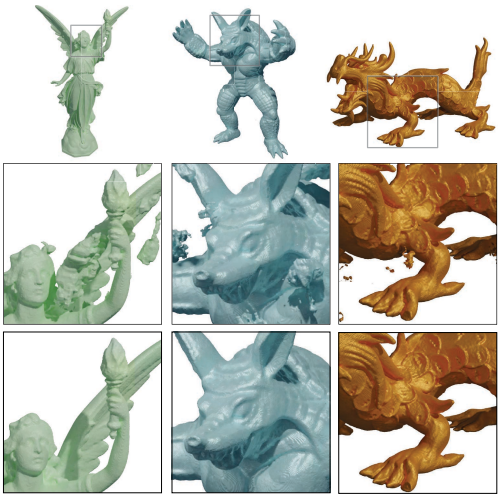


Figure 7: **Signed-distance field reconstruction before (above) and after (below) TILTED.** TILTED reduces floating artifacts without expressiveness-limiting regularization.

	K-Planes	VM
Lego	**35.31**$\pm$**0.02** $\rightarrow$ 33.29$\pm$0.11	**34.24**$\pm$**0.04** $\rightarrow$ 32.63$\pm$0.01
Avg.	**32.12**$\pm$**0.02** $\rightarrow$ 31.62$\pm$0.04	**31.30**$\pm$**0.03** $\rightarrow$ 30.76$\pm$0.03

Table 2: **PSNR decrease of prior methods, before and after random scene rotation.** Metrics are reported from NeRF-Synthetic (standard, axis-aligned) $\rightarrow$ NeRF-Synthetic$^{SO(3)}$(randomly rotated). Without TILTED, a simple rotation of the scene coordinate frame can lead to as high as a 2 PSNR drop in performance.

	K-Planes	w/ TILTED	VM	w/ TILTED
Lego	33.29$\pm$0.11	**34.35**$\pm$**0.07**	32.63$\pm$0.01	**33.90**$\pm$**0.06**
Avg.	31.62$\pm$0.04	**31.91**$\pm$**0.04**	30.76$\pm$0.03	**31.08**$\pm$**0.02**

Table 3: **PSNR improvement after incorporating TILTED, on the NeRF-Synthetic$^{SO(3)}$ dataset.** TILTED offers transform-invariant reconstruction quality and moderate PSNR improvements.

	8 transforms	4 transforms
Two-Phase	**34.35**$\pm$**0.07**	34.19$\pm$0.22
Without	33.95$\pm$0.15	33.83$\pm$0.08

Table 4: **Ablations on the Lego synthetic dataset.** Two-phase optimization and an increased number of transforms synergistically improve reconstruction quality. Similar but weaker trends can be found in less structured scenes. Reported metrics use the K-Planes model.

5.3. Neural Radiance Fields

5.3.1 Synthetic Study

We begin with a quantitative study using the NeRF-Synthetic [29] dataset. While this dataset is commonly used for evaluation of neural implicit architectures, it is unrealistic because objects are rendered in Blender and perfectly axis-aligned. The bricks of the Lego scene, for example, are exactly lined up with the coordinate system that camera poses are defined in. To better understand the robustness of representations, we compare NeRF-Synthetic against the randomly rotated variant proposed by [30]. We refer to this dataset as NeRF-Synthetic$^{SO(3)}$. In NeRF-Synthetic$^{SO(3)}$, an experiment for any given random seed begins training by applying a uniformly sampled $\mathrm{SO}(3)$ rotation to all camera poses. Robustness against this basic operation is critical for real-world data, where canonical orientations are rarely well-defined (let alone provided).

For each of the NeRF-Synthetic and NeRF-Synthetic$^{SO(3)}$ datasets, we train every combination of: **(1)** two decompositions: VM and K-Planes, **(2)** three parameterizations of τ: axis-aligned (baseline), 4 transforms, and 8 transforms, **(3)** eight scenes: chair, drums, ficus, hotdog, lego, materials, mic, and ship, and **(4)** three random seeds: we use 0, 1, and 2. To eliminate the possibility of bounding box clipping artifacts interfering with results, we use enlarged scene bounding boxes of $[-1.6, 1.6]$; this exerts a noticeable but uniform penalty on PSNR metrics relative to results with standard smaller bounding boxes. We additionally incorporate the proposal fields, histogram loss, and distortion losses proposed by MipNeRF-360 [31]. Our core conclusions are:

(1) Naive hybrid representations have strong axis-alignment biases. Results from the axis-aligned factorizations mirror our theoretical results in Section 3. When an axis-aligned decomposition is used, the quality of reconstructions becomes highly sensitive to the orientation of the target input. In Table 2, we observe as high as a 2 PSNR drop from scene rotation on the Lego dataset. In contrast, TILTED is designed with invariance in mind, and is thus robust to these transformations.

(2) TILTED improves reconstructions. On the NeRF-Synthetic$^{SO(3)}$ dataset, we observe performance increases from learned transforms, increasing the number of optimized transformations, and adopting two-phase optimization. Table 3 highlights how TILTED improves PSNRs for the NeRF-Synthetic$^{SO(3)}$ dataset, while Table 4 demonstrates how components of our method (multiple transforms and two-phase optimization) improve results.

5.3.2 Real-World Study

In our final set of experiments, we apply TILTED to 18 real-world scenes made available via Nerfstudio [76]. We

(a) Kitchen (b) Giannini

(c) Stump (d) Storefront

Figure 8: **Real-world radiance field comparisons, before (left) and after (right) TILTED.** For each scene, we arrange in three rows the outputs of **(i)** rendering RGB images, **(ii)** visualizing the structure-revealing ℓ^2-norm of interpolated features, and **(iii)** mapping the top three principal components of interpolated features to RGB. TILTED feature volumes result in better reconstruction quality, with more structured, interpretable, and expressive features. Results in this figure are from K-Planes.

use a nearly identical architecture and set of parameters as for experiments in the synthetic setting, but adopt (a) an ℓ^∞ norm-based scene contraction (Equation A.1) to handle the unbounded nature of real-world data, (b) camera pose optimization to account for noisy camera poses, and (c) NeRF-W-style appearance embeddings [38]. Once camera pose optimization and per-camera appearance embeddings are enabled, we lose the ability to reliably compute evaluation metrics [76]. Instead, we examine how incorporating TILTED impacts training PSNRs and qualitative results.

(1) On real-world data, TILTED can simultaneously halve the memory footprint of a model, accelerate training by 25%, and improve reconstructions. In Table 5, we compare standard factored neural field representations with two techniques for improving reconstructions: doubling the feature volume channel count and TILTED. Compared to an axis-aligned model of the same size, TILTED improves reconstruction performance on all scenes. It also outperforms axis-aligned models with 2x higher channel counts in most cases (72% of the time for VM, 56% for K-Planes), thus cutting parameter count by almost half while being 25% faster to train (11:04 vs 14:46 for 30k steps).

(2) Recovered transforms align factors to underlying scene geometry. In Figure 8, we visualize renders next to visualizations of underlying latent features. We display a norm-based approach, which involves volume rendering a map of feature norms using standard NeRF densities for each transform and then selecting the highest-valued map, and a PCA-based approach, which maps latent vectors to RGB. TILTED feature volumes interpretably align themself to the geometry of the scene, while enabling more detailed and expressive feature volumes.

(3) Standard evaluations incentivize axis-alignment biases. Despite significantly outperforming axis-aligned baselines on both real-world data and NeRF-Synthetic$^{SO(3)}$, we note that TILTED underperforms against baselines on the axis-aligned NeRF-Synthetic dataset. This hints at room for further performance optimizations of our method, while highlighting flaws in the way that radiance fields architectures are often evaluated. Concretely, these results suggest that the baselines are overfit to axis-aligned datasets, and that optimizing for standard evaluation metrics (like PSNR on the NeRF-Synthetic dataset) can end up undermining real-world capabilities.

Dataset	K-Plane / 2x / TILTED	VM / 2x / TILTED
Kitchen	25.95 / 26.91 / **27.12**	25.63 / 26.54 / **26.90**
Floating	24.58 / **25.17** / 25.06	24.03 / 24.70 / **25.04**
Poster	33.14 / 33.71 / **33.79**	32.84 / 33.49 / **33.61**
Redwoods	23.55 / 24.08 / **24.12**	23.22 / 23.81 / **23.85**
Stump	26.82 / **27.29** / 27.28	26.33 / 26.83 / **26.97**
Vegetation	21.62 / 22.10 / **22.10**	21.11 / 21.55 / **21.73**
BWW	24.64 / **25.06** / 24.95	24.22 / 24.75 / **24.80**
Library	25.24 / 25.68 / **25.78**	25.50 / 25.78 / **25.84**
Storefront	29.71 / **30.12** / 29.87	29.15 / 29.77 / **29.87**
Dozer	22.37 / **22.88** / 22.69	21.91 / **22.46** / 22.40
Egypt	20.69 / 21.10 / **21.12**	20.84 / **21.17** / 21.09
Person	24.83 / 24.93 / **25.36**	25.28 / 25.38 / **25.39**
Giannini	20.51 / **20.90** / 20.82	20.27 / **20.64** / 20.60
Sculpture	23.20 / 23.40 / **23.40**	22.86 / 23.07 / **23.28**
Plane	22.75 / 23.00 / **23.01**	22.53 / **22.84** / 22.74
Aspen	15.99 / 16.20 / **16.21**	15.98 / 16.15 / **16.20**
Desolation	22.14 / **22.40** / 22.25	21.88 / 22.11 / **22.12**
Campanile	24.27 / **24.64** / 24.37	23.97 / **24.35** / 24.19

Table 5: **For real-world data, TILTED improves PSNRs on all evaluated scenes, typically outperforming even much larger axis-aligned models.** We compare: standard hybrid neural fields (K-Plane, VM), axis-aligned fields with channel counts doubled (2x), and the fields with the original channel count but addition of TILTED (TILTED).

6. Conclusion

We demonstrate the importance of alignment for factored feature volumes via TILTED, an extension to existing hybrid neural field architectures based on the idea of *canonical factors*. For hybrid neural fields, canonicalizing factors via a learned set of transformations improves real-world reconstruction results, enabling improvements across reconstruction detail, compactness, and runtime. We also developed the theoretical foundations for this methodology; our analysis can be viewed as providing the first provable guarantee for explicit disentangled representation learning with visual data beyond spatial deconvolution (e.g., [28]), here disentangling appearance and pose.

Many directions exist for extending our work, both practically and theoretically. On the practical side, these include further studying and improving convergence characteristics and exploring more diverse families of transformations, particularly for 4D (dynamic scene) factorizations; on the theoretical side, they include extending our results to overparameterized models, MLPs, and scenes with visual clutter. We also note that our work compares TILTED neural fields only to their axis-aligned equivalents: while an abundance of prior work has shown the unique advantages of these representations over alternatives, many applications may still benefit from alternative techniques [58, 70].

Acknowledgements. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE 2146752. YM acknowledges partial support from the ONR grant N00014-22-1-2102, the joint Simons Foundation-NSF DMS grant 2031899, and a research grant from TBSI. The authors thank Justin Kerr, Chung Min Kim, Sara Fridovich-Keil, Druv Pai, and members of the Nerfstudio team for implementation references, technical discussions, and suggestions.

References

[1] J Douglas Carroll and Jih-Jie Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970. 4.

[2] William E Lorensen and Harvey E Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987. 6.

[3] Lisa Gottesfeld Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, Dec. 1992. 3.

[4] J B Antoine Maintz and Max A Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, Mar. 1998. 3.

[5] Martin Lefébure and Laurent D Cohen, "Image registration, optical flow and local rigidity," *Journal of mathematical imaging and vision*, vol. 14, no. 2, pp. 131–147, Mar. 2001. 3.

[6] M I Miller and L Younes, "Group actions, homeomorphisms, and matching: A general framework," *International journal of computer vision*, vol. 41, no. 1, pp. 61–84, Jan. 2001. 3.

[7] Simon Baker and Iain Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004. 3.

[8] Richard Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2007. 3.

[9] Zhengdong Zhang, Arvind Ganesh, Xiao Liang, and Yi Ma, "TILT: transform invariant low-rank textures," *CoRR*, vol. abs/1012.3216, 2010. arXiv: 1012.3216. 1.

[10] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2012, pp. 3642–3649. 3.

[11] Stéphane Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012. 3.

[12] Kihyuk Sohn and Honglak Lee, "Learning invariant representations with local transformations," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12, Edinburgh, Scotland: Omnipress, Jun. 2012, pp. 1339–1346. 3.

[13] Zhengdong Zhang, Arvind Ganesh, Xiao Liang, and Yi Ma, "TILT: Transform invariant Low-Rank textures," *International journal of computer vision*, vol. 99, no. 1, pp. 1–24, Aug. 2012. 3.

[14] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs, "Locally Scale-Invariant convolutional neural networks," Dec. 2014. arXiv: 1412.5104 [cs.CV]. 3.

[15] Elif Vural and Pascal Frossard, "Analysis of image registration with tangent distance," *SIAM journal on imaging sciences*, vol. 7, no. 4, pp. 2860–2915, Jan. 2014. 3.

[16] Sander Dieleman, Kyle W. Willett, and Joni Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, Apr. 2015. 3.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15, Montreal, Canada: MIT Press, 2015, pp. 2017–2025. 3.

[18] Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys, "Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016. 3.

[19] Chen-Hsuan Lin and Simon Lucey, "Inverse compositional spatial transformer networks," Dec. 2016. arXiv: 1612.03897 [cs.CV]. 3.

[20] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773. 3.

[21] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," Dec. 2017. arXiv: 1712.09203 [cs.LG]. 1, 3.

[22] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. 3.

[23] Thomas Wiatowski and Helmut Bölcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1845–1866, Mar. 2018. 3.

[24] Zhiqin Chen and Hao Zhang, "Learning implicit fields for generative shape modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5939–5948. 2.

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470. 2.

[26] 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*, 2019, pp. 165–174. 2.

[27] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2304–2314. 2.

[28] Han-Wen Kuo, Yuqian Zhang, Yenson Lau, and John Wright, "Geometry and symmetry in Short-and-Sparse deconvolution," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 1, pp. 216–245, Jan. 2020. 9.

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020. 2, 7.

[30] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *NeurIPS*, 2020. 3, 5, 7.

[31] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *arXiv*, 2021. 7.

[32] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *arXiv*, 2021. 1, 2, 4.

[33] Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang, "Differentiable surface rendering via non-differentiable sampling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6088–6097. 2.

[34] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 346–14 355. 2.

[35] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec, "Baking neural radiance fields for real-time view synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5875–5884. 2.

[36] Ylva Jansson and Tony Lindeberg, "Scale-invariant scale-channel networks: Deep networks that generalise to previously unseen scales," *CoRR*, vol. abs/2106.06418, 2021. arXiv: 2106.06418. 3.

[37] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey, "Barf: Bundle-adjusting neural radiance fields," in *IEEE International Conference on Computer Vision (ICCV)*, 2021. 5.

[38] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7210–7219. 2, 9.

[39] Michael Oechsle, Songyou Peng, and Andreas Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5589–5599. 2.

[40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874. 5.

[41] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger, "Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 335–14 345. 2.

[42] Dominik Stöger and Mahdi Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," Jun. 2021. arXiv: 2106.15013 [cs.LG]. 1, 3.

[43] Cheng Sun, Min Sun, and Hwann-Tzong Chen, *Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction*, 2021. arXiv: 2111.11215 [cs.CV]. 2.

[44] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth, "Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes," *arXiv preprint arXiv:2111.13260*, 2021. 2.

[45] 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.

[46] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman, "Volume rendering of neural implicit surfaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4805–4815, 2021. 2.

[47] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761. 2.

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable DETR: Deformable transformers for End-to-End object detection," in *International Conference on Learning Representations*, 2021. 3.

[49] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," *arXiv preprint arXiv:2112.12130*, 2021. 3.

[50] Sam Buchanan, Jingkai Yan, Ellie Haber, and John Wright, "Resource-Efficient invariant networks: Exponential gains by unrolled optimization," Mar. 2022. arXiv: 2203.05006 [cs.CV]. 3.

[51] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022. 1–5.

[52] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*, PMLR, 2022, pp. 158–168. 2.

[53] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," *arXiv preprint arXiv:2203.15224*, 2022. 2.

[54] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi, "Nerf2nerf: Pairwise registration of neural radiance fields," Nov. 2022. arXiv: 2211.01600 [cs.CV]. 3.

[55] Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra, "Relu fields: The little non-linearity that could," in *ACM SIGGRAPH 2022 Conference Proceedings*, ser. SIGGRAPH '22, Vancouver, BC, Canada: Association for Computing Machinery, 2022. 2.

[56] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser, "Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation," in *CVPR*, 2022. 2.

[57] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang, "Neurmips: Neural mixture of planar experts for view synthesis," *arXiv preprint arXiv:2204.13696*, 2022. 2.

[58] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *arXiv:2201.05989*, Jan. 2022. 6, 9.

[59] Anton Obukhov, Mikhail Usvyatsov, Christos Sakaridis, Konrad Schindler, and Luc Van Gool, "TT-NF: Tensor train neural fields," Sep. 2022. arXiv: 2209.15529 [cs.LG]. 2.

[60] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein, *3d neural field generation using triplane diffusion*, 2022. arXiv: 2211.16677 [cs.CV]. 2.

[61] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo, *Rodin: A generative model for sculpting 3d digital avatars using diffusion*, arXiv, Dec. 2022. 5.

[62] Thomas Weng, David Held, Franziska Meier, and Mustafa Mukadam, "Neural grasp distance fields for robot manipulation," *arXiv preprint arXiv:2211.02647*, 2022. 2.

[63] Ang Cao and Justin Johnson, "Hexplane: A fast representation for dynamic scenes," *arXiv preprint arXiv:2301.09632*, 2023. 1–3.

[64] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger, "Factor fields: A unified framework for neural fields and beyond," *arXiv preprint arXiv:2302.01226*, 2023. 2, 4, 6.

[65] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su, *Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction*, 2023. arXiv: 2304.06714 [cs.CV]. 2.

[66] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," *arXiv preprint arXiv:2301.10241*, 2023. 1–4.

[67] Quankai Gao, Qiangeng Xu, Hao Su, Ulrich Neumann, and Zexiang Xu, *Strivec: Sparse tri-vector radiance fields*, 2023. arXiv: 2307.13226 [cs.CV]. 1, 3, 5.

[68] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma, "Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields," in *ICCV*, 2023. 1, 3.

[69] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret, *Eslam: Efficient dense slam system based on hybrid representation of signed distance fields*, 2023. arXiv: 2211.11704 [cs.CV]. 3.

[70] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023. 9.

[71] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023. 2.

[72] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf, *Progressively optimized local radiance fields for robust view synthesis*, 2023. arXiv: 2303.13791 [cs.CV]. 3.

[73] Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang, *Temporal interpolation is all you need for dynamic neural radiance fields*, 2023. arXiv: 2302.09311 [cs.CV]. 3.

[74] Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman, *Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes*, 2023. arXiv: 2302.12249 [cs.CV]. 2.

[75] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu, *Tensor4d : Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering*, 2023. arXiv: 2211.11610 [cs.CV]. 3.

[76] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," *arXiv preprint arXiv:2302.04264*, 2023. 7, 9.

[77] Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma, "The power of preconditioning in overparameterized Low-Rank matrix sensing," Feb. 2023. arXiv: 2302.01186 [cs.LG]. 3.

[78] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P. Srinivasan, Richard Szeliski, Jonathan T. Barron, and Ben Mildenhall, *Bakedsdf: Meshing neural sdfs for real-time view synthesis*, 2023. arXiv: 2302.14859 [cs.CV]. 2.

[79] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R. Oswald, Andreas Geiger, and Marc Pollefeys, *Nicer-slam: Neural implicit scene encoding for rgb slam*, 2023. arXiv: 2302.03594 [cs.CV]. 3.