Abstract

Factored feature volumes offer a simple way to build more compact, efficient, and interpretable 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 \( \theta \) maps spatial positions \( p = (p_x, p_y, p_z) \in \mathbb{R}^3 \) to RGB colors \( c \in [0, 1]^3 \) and densities \( \sigma \in \mathbb{R}_{\geq 0} \):

\[
p \xrightarrow{\text{MLP}_\theta} (c, \sigma).
\] (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.

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 \( \theta \). 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

\[
p \xrightarrow{\text{VoxelTrilerp}_\phi} Z \xrightarrow{\text{MLP}_\theta} (c, \sigma),
\] (2.2)

where VoxelTrilerp_\phi interpolates the ‘latent grid’ parameters \( \phi \) to produce a latent feature \( Z \in \mathbb{R}^d \), which is then decoded to standard radiance field outputs by an MLP with parameters \( \theta \).

Instead of implementing the latent feature volume \( \phi \) as a dense voxel grid, a common pattern is to decompose this tensor into lower-dimensional factors \( \phi = \{ F_1 \ldots F_F \} \). 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 \( Z \):

\[
Z = \text{Reduce}([\text{Interp}_{F'_1}(\text{Proj}_1(p))], \ldots, [\text{Interp}_{F'_F}(\text{Proj}_F(p))]).
\] (2.3)

Interpolating only on lower-dimensional feature grids \( F_1, \ldots, F_F \), which may be 1D or 2D when \( p \) 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 $\tau$ that deforms an observed image $Y$ to match a target $X$ 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 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.

\[
\min_{\tau} \| Y \circ \tau \|_a. \tag{2.4}
\]

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 $X_2$ 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 $X_2$. (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 $X_2 \in \mathbb{R}^{n \times n}$ denote the grayscale image corresponding to the axis-aligned square pattern in Figure 3(a). We can decompose $X_2$ as $X_2 = u_2 v_2^T$, where $u_2$ and $v_2$ are one-dimensional square pulses aligned with the support of $X_2$; $X_2$ 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 $X_\nu = X_2 \circ \tau_\nu$ (Figure 3(b)). As $\nu$ approaches its maximum value, the rank of the transformed
scene grows to a constant multiple of the resolution \( n \), implying that perfect representation of \( X_p \) 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 \( \hat{X} \) to \( X_{\pi/4} \) satisfies

\[
\frac{1}{n^2} \left\| \hat{X} - X_{\pi/4} \right\|_F^2 \geq \frac{c_1}{1 + F}.
\]

Theorem 1 asserts that a broad class of sublinear-rank approximations to \( X_p \) 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 \( X_p \): there exists a transformed coordinate system in which \( X_p \) 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 \( X_p \) 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

\[
\min_{\phi, \mathbf{u}} \left\| X_{\pi/4} - \left( \mathbf{u} \mathbf{u}^* \right) \circ \tau_\phi \right\|_F^2 \tag{3.1}
\]

solved with randomly-initialized constant-stepping gradient descent converges to the true parameters \( (\pi/4, \mathbf{u}_t) \), 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 \( \text{Proj}_j \) in (2.3) are static and axis-aligned, the core idea of TILTED is to replace \( \text{Proj}_j \) with learnable functions \( \text{Proj}_{j, \tau} \), where \( \tau \) is a set of learnable transformation parameters. By substituting into (2.3), the feature volume interpolation function then becomes:

\[
Z = \text{Reduce} \left( \left[ \text{Interp}_{F_1} (\text{Proj}_{j_1, \tau}(p)), \ldots, \text{Interp}_{F_T} (\text{Proj}_{j_T, \tau}(p)) \right] \right) \tag{4.1}
\]

The transformations \( \tau \) enable mapping from arbitrary scene coordinates to canonicalized domains for each factor \( F_t \). As illustrated 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 \( \tau \) and how it is applied to input coordinates \( p \) is large. We develop TILTED for the case where \( \tau \) is a set of \( T \) randomly initialized rotations \( \tau = \{ \tau_1, \ldots, \tau_T \} \), parameterized by the unit circle \( S^1 \) in 2D and \( S^3 \) (the universal cover of the set of rotation matrices 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 \( Z \in \mathbb{R}^d \). We constrain \( T \) such that it evenly divides \( d \), and apply rotations to the input coordinates such that each rotation \( \tau_t \) 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 \( \tau \) 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 $\tau$. When $\tau$ is constrained to rotations, a rotation applied to the scene becomes equivalent to a rotation applied to the random initialization of $\tau$.

**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:

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

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

**Two-phase optimization.** Effective recovery of $\tau$ 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 $\tau$ toward improved solutions.

To help disentangle the $\tau$ 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 $\theta_{bneck}$, feature grid parameters $\phi_{bneck}$, and projection parameters $\tau_{bneck}$. We discard all parameters but $\tau_{bneck}$, and then simply set:

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

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

### 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 4: Two-phase optimization.](image-url)

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 $\tau_{bneck}$, which are used for initialization of the final representation (right).
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.

<table>
<thead>
<tr>
<th>IoU ↑</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Planes</td>
<td>0.949 ± 0.015</td>
<td>0.952 ± 0.015</td>
<td>0.952 ± 0.016</td>
</tr>
<tr>
<td>w/ TILTED</td>
<td><strong>0.989 ± 0.002</strong></td>
<td><strong>0.990 ± 0.002</strong></td>
<td><strong>0.991 ± 0.002</strong></td>
</tr>
<tr>
<td>IoU ↑</td>
<td>45</td>
<td>90</td>
<td>135</td>
</tr>
<tr>
<td>Vector-Matrix</td>
<td>0.970 ± 0.007</td>
<td>0.979 ± 0.005</td>
<td>0.982 ± 0.003</td>
</tr>
<tr>
<td>w/ TILTED</td>
<td><strong>0.982 ± 0.003</strong></td>
<td><strong>0.989 ± 0.002</strong></td>
<td><strong>0.988 ± 0.003</strong></td>
</tr>
</tbody>
</table>

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.

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-SyntheticSO(3). In NeRF-SyntheticSO(3), an experiment for any given random seed begins training by applying a uniformly sampled 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-SyntheticSO(3) datasets, we train every combination of: (1) two decompositions: VM and K-Planes, (2) three parameterizations of \( \tau \): 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 bundling boxes 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-SyntheticSO(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-SyntheticSO(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].

<table>
<thead>
<tr>
<th></th>
<th>K-Planes</th>
<th>VM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lego</td>
<td>35.31±0.02</td>
<td>34.24±0.04</td>
</tr>
<tr>
<td>Avg.</td>
<td>32.12±0.02</td>
<td>31.30±0.03</td>
</tr>
</tbody>
</table>

Table 2: PSNR decrease of prior methods, before and after random scene rotation. Metrics are reported from NeRF-Synthetic (standard, axis-aligned) → NeRF-SyntheticSO(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.

<table>
<thead>
<tr>
<th></th>
<th>K-Planes</th>
<th>VM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lego</td>
<td>33.29±0.11</td>
<td>32.63±0.01</td>
</tr>
<tr>
<td>Avg.</td>
<td>31.62±0.04</td>
<td>30.76±0.03</td>
</tr>
</tbody>
</table>

Table 3: PSNR improvement after incorporating TILTED, on the NeRF-SyntheticSO(3) dataset. TILTED offers transform-invariant reconstruction quality and moderate PSNR improvements.

<table>
<thead>
<tr>
<th></th>
<th>8 transforms</th>
<th>4 transforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-Phase</td>
<td>34.35±0.07</td>
<td>34.19±0.22</td>
</tr>
<tr>
<td>Without</td>
<td>33.95±0.15</td>
<td>33.83±0.08</td>
</tr>
</tbody>
</table>

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.
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 $\ell^2$-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 \( \ell_\infty \) 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 themselves 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-SyntheticSO(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.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>K-Plane / 2x / TILTED</th>
<th>VM / 2x / TILTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>25.95 / 26.91 / 27.12</td>
<td>25.63 / 26.54 / 26.90</td>
</tr>
<tr>
<td>Floating</td>
<td>24.58 / 25.17 / 25.06</td>
<td>24.03 / 24.70 / 25.04</td>
</tr>
<tr>
<td>Poster</td>
<td>33.14 / 33.71 / 33.79</td>
<td>32.84 / 33.49 / 33.61</td>
</tr>
<tr>
<td>Redwoods</td>
<td>23.55 / 24.08 / 24.12</td>
<td>23.22 / 23.81 / 23.85</td>
</tr>
<tr>
<td>Stump</td>
<td>26.82 / 27.29 / 27.28</td>
<td>26.33 / 26.83 / 26.97</td>
</tr>
<tr>
<td>BWW</td>
<td>24.64 / 25.06 / 24.95</td>
<td>24.22 / 24.75 / 24.80</td>
</tr>
<tr>
<td>Library</td>
<td>25.24 / 25.68 / 25.78</td>
<td>25.50 / 25.78 / 25.84</td>
</tr>
<tr>
<td>Storefront</td>
<td>29.71 / 30.12 / 29.87</td>
<td>29.15 / 29.77 / 29.87</td>
</tr>
<tr>
<td>Dozer</td>
<td>22.37 / 22.88 / 22.69</td>
<td>21.91 / 22.46 / 22.40</td>
</tr>
<tr>
<td>Giannini</td>
<td>20.51 / 20.90 / 20.82</td>
<td>20.27 / 20.64 / 20.60</td>
</tr>
<tr>
<td>Sculpture</td>
<td>23.20 / 23.40 / 23.40</td>
<td>22.86 / 23.07 / 23.28</td>
</tr>
<tr>
<td>Plane</td>
<td>22.75 / 23.00 / 23.01</td>
<td>22.53 / 22.84 / 22.74</td>
</tr>
<tr>
<td>Aspen</td>
<td>15.99 / 16.20 / 16.21</td>
<td>15.98 / 16.15 / 16.20</td>
</tr>
<tr>
<td>Campanile</td>
<td>24.27 / 24.64 / 24.37</td>
<td>23.97 / 24.35 / 24.19</td>
</tr>
</tbody>
</table>

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).

We also note that our work compares TILTED neural fields to 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


