Multi-task View Synthesis with Neural Radiance Fields

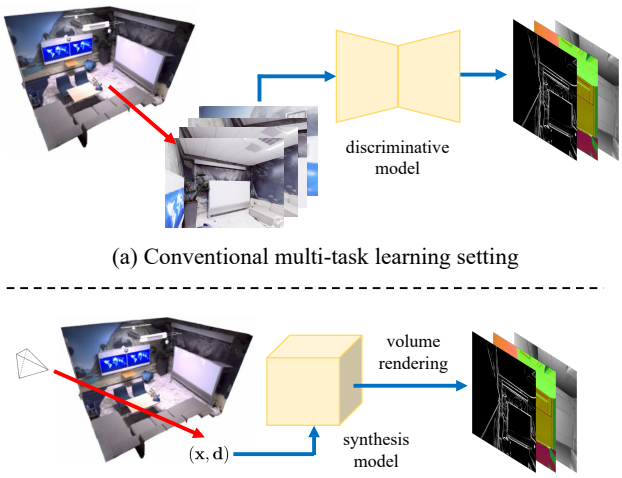
Shuhong Zheng*,1 Zhipeng Bao*,2 Martial Hebert[2] Yu-Xiong Wang[1]
[1]University of Illinois Urbana-Champaign [2]Carnegie Mellon University
{szheng36, yxw}@illinois.edu {zbao, hebert}@cs.cmu.edu

Abstract

Multi-task visual learning is a critical aspect of computer vision. Current research, however, predominantly concentrates on the multi-task dense prediction setting, which overlooks the intrinsic 3D world and its multi-view consistent structures, and lacks the capacity for versatile imagination. In response to these limitations, we present a novel problem setting – multi-task view synthesis (MTVS), which reinterprets multi-task prediction as a set of novel-view synthesis tasks for multiple scene properties, including RGB. To tackle the MTVS problem, we propose MuvieNeRF, a framework that incorporates both multi-task and cross-view knowledge to simultaneously synthesize multiple scene properties. MuvieNeRF integrates two key modules, the Cross-Task Attention (CTA) and Cross-View Attention (CVA) modules, enabling the efficient use of information across multiple views and tasks. Extensive evaluations on both synthetic and realistic benchmarks demonstrate that MuvieNeRF is capable of simultaneously synthesizing different scene properties with promising visual quality, even outperforming conventional discriminative models in various settings. Notably, we show that MuvieNeRF exhibits universal applicability across a range of NeRF backbones. Our code is available at https://github.com/zsh2000/MuvieNeRF.

1. Introduction

When observing a given scene, human minds exhibit a remarkable capability to mentally simulate the objects within it from a novel viewpoint in a *versatile* manner [40]. It not only includes hallucinating the colors of objects, but also extends to numerous associated scene properties such as surface orientation, semantic markings, and edge patterns. Prompted by this, a burgeoning interest has emerged, seeking to equip modern robotic systems with similar capabilities for handling multiple tasks. Nevertheless, contemporary research [33, 66, 65] has primarily centered on the *multi-task dense prediction* setting, which employs a conventional discriminative model to concurrently predict multiple pixel-level scene properties using given RGB images (refer to Figure 1(a)). Yet, the methodologies arising from this context often demonstrate practical limitations, chiefly due to their tendency to treat each image as a separate entity, without constructing an explicit 3D model that adheres to the principle of multi-view consistency. Even more critically, they lack the ability to "imagine" – they are incapable of inferring scene properties from an *unseen* perspective, as these models invariably require RGB images.



(a) Conventional multi-task learning setting

(b) Our multi-task view synthesis pipeline

Figure 1. Comparison between (a) the conventional multi-task learning scheme and (b) our multi-task view synthesis setting. The conventional "discriminative" multi-task learning makes predictions for single images while multi-task view synthesis aims to render visualizations for multiple scene properties at novel views.

To circumvent these constraints, we propose a novel approach that reconsiders multi-task learning (MTL) [5] from a *synthesis* perspective. This leads to a flexible problem setting that reinterprets multi-task visual learning as a collection of novel-view synthesis problems, which we refer to as *multi-task view synthesis* (MTVS) (refer to Figure 1(b)). As an illustration, the task of predicting surface

*Equal contribution

normals for a given image could be reframed as visualizing a three-channel "image" with the given pose and camera parameters. With the achievements of Neural Radiance Fields (NeRF) [32], the implicit scene representation offers an effective solution to synthesize scene properties beyond RGB [68]. Importantly, this scene representation takes multi-view geometry into account, which consequently enhances the performance of all tasks.

Extending from [68], we make the exploration that mining multi-task knowledge can simultaneously enhance the learning of different tasks, extending beyond discriminative models [66, 45] to include synthesis models as well. Furthermore, we argue that the alignment of features across multiple reference views and the target view can reinforce cross-view consistency, thereby bolstering the implicit scene representation learning. Informed by this insight, we propose MuvieNeRF, a unified framework for the MTVS task, which incorporates ***Mu***lti-task and cross-***vie***w knowledge, thus enabling the simultaneous synthesis of multiple scene properties through a shared implicit scene representation. MuvieNeRF can be applied to an arbitrary conditional NeRF architecture and features a unified decoder with two key modules: *Cross-Task Attention (CTA) module*, which investigates relationships among different scene properties, and *Cross-View Attention (CVA) module*, which aligns features across multiple views. The integration of these two modules within MuvieNeRF facilitates the efficient utilization of information from multiple views and tasks, leading to better performance across all tasks.

To demonstrate the effectiveness of our approach, we first instantiate our MuvieNeRF with GeoNeRF [22], a state-of-the-art conditional NeRF model, and conduct comprehensive evaluations on both synthetic and real-world benchmarks. The results illustrate that MuvieNeRF is capable of solving multi-task learning in a synthesis manner, even outperforming several competitive discriminative models in different settings. Moreover, we ablate the choice of conditional NeRF backbones to illustrate the broad applicability of our framework. We further validate the individual contributions of the CVA and CTA modules by building and comparing different variants of MuvieNeRF. Finally, we demonstrate the broader applications and analysis of MuvieNeRF, such as generalization on out-of-distribution datasets.

In summary, **our contributions** are three-fold: **(1)** We pioneer a novel problem definition, multi-task view synthesis (MTVS), which reconsiders multi-task visual learning as a set of view synthesis tasks. The introduction of MTVS paves the way for robots to emulate human-like mental simulation capabilities by utilizing the implicit scene representation offered by Neural Radiance Fields (NeRF). **(2)** We present MuvieNeRF, a unified framework that employs Cross-Task Attention (CTA) and Cross-View Attention (CVA) modules to leverage cross-view and cross-task information for the MTVS problem. **(3)** Comprehensive experimental evaluations demonstrate that MuvieNeRF shows promising results for MTVS, and greatly outperforms conventional discriminative models across diverse settings.

2. Related Work

In this work, we propose the MuvieNeRF model which leverages both *multi-task* and *cross-view* information for *multi-task view synthesis*. We review the most relevant work in the areas below.

View Synthesis aims to generate a target image with an arbitrary camera pose by referring to source images [50]. Numerous existing methods have delivered promising results in this area [44, 63, 57, 34, 2]. However, unlike these conventional approaches, MTVS endeavors to synthesize multiple scene properties, including RGB, from novel viewpoints. In pursuit of a similar goal, another group of methods seeks to render multiple annotations for novel views, following a *first-reconstruct-then-render* strategy [13, 19, 26, 12]. These methods typically collect or construct a 3D scene representation (*e.g.*, mesh or point cloud) and subsequently render multiple scene properties using 3D-to-2D projection. In contrast, our work constructs an *implicit* 3D scene representation using a NeRF-style model based on 2D data. This approach is more computationally efficient and, importantly, our implicit representation provides an opportunity to further model task relationships, an advantage the aforementioned methods do not possess.

Neural Radiance Fields are originally designed for synthesizing novel-view images with ray tracing and volume rendering technologies [32]. Follow-up work [3, 35, 10, 41, 18, 27, 56, 36, 38, 29, 39, 14, 49, 58] further improves the image quality, optimization, and compositionality. In addition, several approaches [64, 6, 22, 48], namely conditional NeRFs, encode the scene information to enable the conditional generalization to novel scenes, which are more aligned with our setting. Our MuvieNeRF takes the encoders from these conditional NeRFs as backbones. Some work has also paid attention to synthesizing other properties of scenes [37, 60, 53, 69, 68, 11]. Among them, Semantic-NeRF [69] extends NeRF from synthesizing RGB images to additionally synthesizing semantic labels. SS-NeRF [68] further generalizes the NeRF architecture to simultaneously render RGB and different scene properties with a shared scene representation. Panoptic 3D volumetric representation [43] is introduced to jointly synthesize RGB and panoptic segmentation for in-the-wild images. Different from them, we tackle the novel MTVS task and leverage both *cross-view* and *cross-task* information.

Multi-task Learning aims to leverage shared knowledge across different tasks to achieve optimal performance on

all the tasks. Recent work improves multi-task learning performance by focusing on better optimization strategies [7, 8, 20, 21, 28, 1, 15] and exploring more efficient multi-task architectures [23, 47, 54, 4].

Cross-task Relationship is an interesting topic in multi-task learning, which aims to explore the underlying task relationships among different visual tasks [24]. Taking the task relationship into consideration, cross-stitch networks [33] adopt a learnable parameter-sharing strategy for multi-task learning. Taskonomy and its follow-up work [66, 45, 65] systematically study the internal task relationships and design the optimal multi-task learning schemes accordingly to obtain the best performance. Inspired by them, we also investigate how to better model multi-task learning but in a *synthesis* framework with our model-agnostic MuvieNeRF.

3. Method

In this section, we first describe our novel multi-task view synthesis problem in Section 3.1. Next, we briefly review conditional neural radiance fields (NeRFs) and volume rendering in Section 3.2. In Section 3.3, we explain the proposed MuvieNeRF (as shown in Figure 2) in detail. Finally, we discuss how we handle a more challenging setting without access to source-view annotations at test time in Section 3.4.

3.1. Multi-task View Synthesis Problem

Different from conventional multi-task learning settings, our goal is to jointly synthesize multiple scene properties including RGB images from *novel* views. Therefore, we aim to learn a model Φ which takes a set of V source-view task annotations with camera poses as reference, and predicts the task annotations for a novel view given camera pose **(Inference Setting I)**:

$$\mathbf{Y}_T = \Phi\left(\{(\mathbf{Y}_i, \mathbf{P}_i)\}_{i=1}^{V}, \mathbf{P}_T\right), \tag{1}$$

where $\mathbf{Y}_i = \left[\mathbf{x}_i, \mathbf{y}_i^1, \cdots, \mathbf{y}_i^K\right]$ denotes RGB images $\mathbf{x}_i$ and K other multi-task annotations $\{\mathbf{y}_i^j\}_{j=1}^K$ in the i^{th} source view. $\mathbf{P}_i$ is the i^{th} source camera pose, and $\mathbf{P}_T$ is the target camera pose. During evaluation, Φ is *supposed to be generalized to novel scenes that are not seen during training*.

For the evaluation of those novel scenes, we also provide a more challenging setting lacking source-view annotations during the inference time with the assumption that the model may not get access to additional annotations other than RGB during inference **(Inference Setting II)**:

$$\mathbf{Y}_T = \Phi\left(\{(\mathbf{x}_i, \mathbf{P}_i)\}_{i=1}^{V}, \mathbf{P}_T\right). \tag{2}$$

With the above two settings, **Inference Setting I** allows us to better evaluate the task relationships in our synthesis framework in a cleaner manner, so it is *the focused setting* in our paper; **Inference Setting II** is more aligned with real practice, for which we also propose a solution which is discussed in Sections 3.4 and 4.5.

3.2. Preliminary: Conditional Neural Radiance Fields and Volume Rendering

Neural radiance fields (NeRFs) [32] propose a powerful solution for implicit scene representation, and are widely used in novel view image synthesis. Given the 3D position of a point $\mathbf{q} = (x, y, z)$ in the scene and the 2D viewing direction $\mathbf{d} = (\theta, \phi)$, NeRFs learn a mapping function $(\mathbf{c}, \sigma) = F(\mathbf{q}, \mathbf{d})$, which maps the 5D input $(\mathbf{q}, \mathbf{d})$ to RGB color $\mathbf{c} = (r, g, b)$ and density σ.

To enhance the generalizability of NeRFs, **conditional NeRFs** [64, 22, 6, 48] learn a scene representation across multiple scenes. They first extract a feature volume $\mathbf{W} = E(\mathbf{x})$ for each input image $\mathbf{x}$ of a scene. Next, for an arbitrary point $\mathbf{q}$ on a camera ray with direction $\mathbf{d}$, they are able to retrieve the corresponding image feature on $\mathbf{W}$ by projecting $\mathbf{q}$ onto the image plane with known pose $\mathbf{P}$. We treat the above part as the *conditional NeRF encoder*, which returns:

$$f_{\text{scene}} = F_{\text{enc}}(\{\mathbf{x}_i, \mathbf{P}_i\}_{i=1}^{V}, \mathbf{q}). \tag{3}$$

We have $f_{\text{scene}} \in \mathbb{R}^{V \times d_{\text{scene}}}$, which contains the scene representation from V views. Next, the conditional NeRFs further learn a decoder $(\mathbf{c}, \sigma) = F_{\text{dec}}(\mathbf{q}, \mathbf{d}, f_{\text{scene}})$ to predict the color and density.

Given the color and density of 3D points, NeRFs render the 2D images by running **volume rendering** for each pixel with ray tracing. Every time when rendering a pixel in a certain view, a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ which origins from the center $\mathbf{o}$ of the camera plane in the direction $\mathbf{d}$ is traced. NeRFs randomly sample M points $\{t_m\}_{m=1}^{M}$ with color $\mathbf{c}(t_m)$ and density $\sigma(t_m)$ between the near boundary t_n and far boundary t_f. The RGB value of the pixel is given by:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{m=1}^{M} \hat{T}(t_m)\alpha(\delta_m \sigma(t_m))\mathbf{c}(t_m), \tag{4}$$

where δ_m is the distance between two consecutive sampled points ($\delta_m = \|t_{m+1} - t_m\|$), $\alpha(d) = 1 - \exp(-d)$, and

$$\hat{T}(t_m) = \exp\left(-\sum_{j=1}^{m-1} \delta_j \sigma(t_j)\right) \tag{5}$$

denotes the accumulated transmittance. The same technique can be used to render an arbitrary scene property $\mathbf{y}^j$ by:

$$\hat{\mathbf{Y}}^j(\mathbf{r}) = \sum_{m=1}^{M} \hat{T}(t_m)\alpha(\delta_m \sigma(t_m))\mathbf{y}^j(t_m). \tag{6}$$

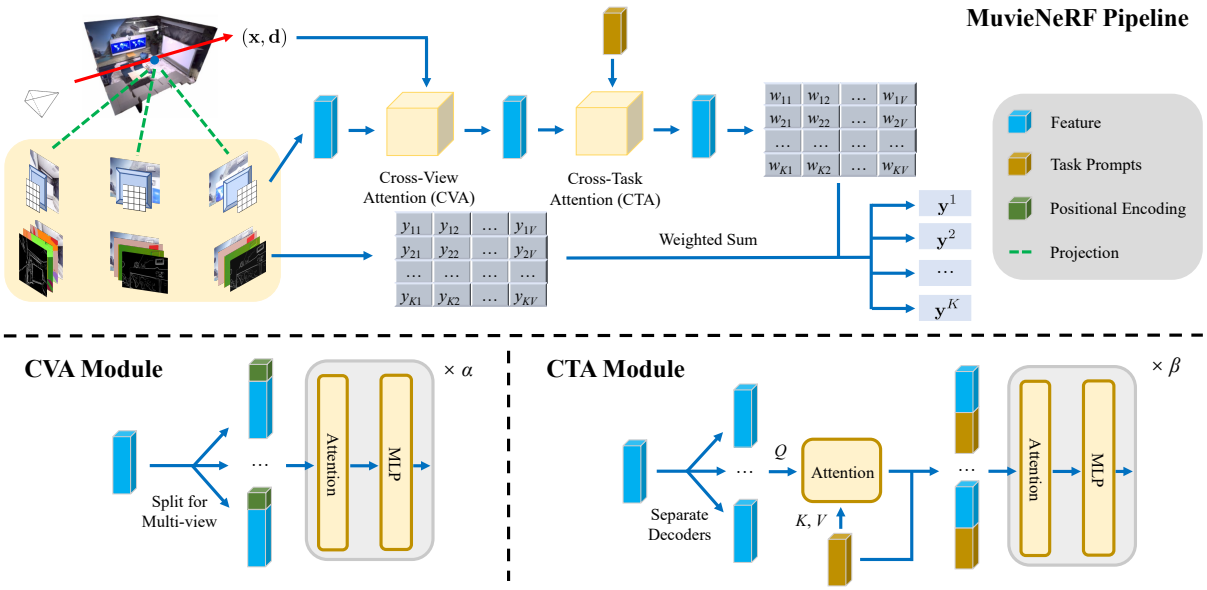


Figure 2. Model architecture. MuvieNeRF is a unified framework for multi-task view synthesis equipped with Cross-View Attention (CVA) and Cross-Task Attention (CTA) modules. It predicts multiple scene properties for arbitrary 3D coordinates with source-view annotations.

3.3. MuvieNeRF

As illustrated in Figure 2, MuvieNeRF first fetches the scene representation f_{scene} from the conditional NeRF encoder, then predicts multiple scene properties $\left[\mathbf{x}(\mathbf{q}), \mathbf{y}^1(\mathbf{q}), \cdots, \mathbf{y}^K(\mathbf{q})\right]$ for arbitrary 3D coordinate $\mathbf{q}$. The final annotations are rendered by Equation 6. We explain how to predict multiple scene properties with f_{scene} and source annotations $[(\mathbf{Y}_1, \mathbf{P}_1), \cdots, (\mathbf{Y}_V, \mathbf{P}_V)]$ as follows. The full detailed architecture is included in the supplementary.

3.3.1 Cross-view Attention Module

The cross-view attention (CVA) module (Figure 2 bottom left) leverages the multi-view information for MuvieNeRF. To start, we first concatenate f_{scene} with a positional embedding derived from the target ray and the source-view image plane: $f^{\text{pos}}_{\text{scene}} = [f_{\text{scene}}; \gamma(\theta_{n,v})]$, where $\gamma(\cdot)$ is the positional encoding proposed in [32], and $\theta_{n,v}$ is the angle between the novel camera ray $\mathbf{r}$ and the line that connects the camera center of view v and the point $\mathbf{q}_n$ in the queried ray, which measures the similarity between the source view v and the target view.

Next, α CVA modules are used to leverage the cross-view information. Concretely, in each module, we have one self-attention union followed by a multi-layer perceptron (MLP): $f_{\text{CVA}} = \text{MLP}_{\text{CVA}}(f^{\text{pos}}_{\text{scene}} + \text{MHA}(f^{\text{pos}}_{\text{scene}}, f^{\text{pos}}_{\text{scene}}))$, where $\text{MHA}(a, b)$ denotes multi-head attention [52] with a as query and b as key and value.

After these processes, we apply $(K+1)$ different MLPs (corresponding to K vision tasks in multi-task view synthesis plus the RGB synthesis task) to broadcast the shared feature, leading to the $(K+1)$-branch feature $f_{\text{task}} \in \mathbb{R}^{(K+1)\times V\times d_{\text{task}}}$.

3.3.2 Cross-task Attention Module

In order to simultaneously benefit all the downstream tasks, we propose a novel cross-task attention (CTA) module (Figure 2 bottom right) to facilitate knowledge sharing and information flow among all the tasks. The CTA module has two attention components with shared learnable task prompts [62], $p_t \in \mathbb{R}^{(K+1)\times d_t}$, where d_t is the dimension of task prompts. The first attention component applies cross-attention between features from each branch and the task prompts $f_{\text{stage1}} = f_{\text{task}} + \text{MHA}(f_{\text{task}}, p_t)$. In this stage, we run K MHA individually for each task branch with the shared task prompts. After the cross-attention, we further concatenate f^j_{stage1} for task T_j and the corresponding task prompt p^j_t to obtain $f_{\text{stage1}'}$.

Next, we apply the second component to use β self-attention modules for all the branches jointly to leverage the cross-task features. The final feature representation is obtained by: $f_{\text{stage2}} = \text{MLP}_{\text{CTA}}(f_{\text{stage1}'} + \text{MHA}(f_{\text{stage1}'}, f_{\text{stage1}'}))$.

Finally, to predict the task annotations of the target view, we adopt the formulation of GeoNeRF [22]. The prediction $\hat{\mathbf{y}}^j$ of task T_j on the target view is the weighted sum of the source views:

$$\hat{\mathbf{y}}^j = \sum_{i=1}^{V} \mathbf{w}[j,i] \cdot \mathbf{y}[j,i], \tag{7}$$

Evaluation Type		Training scene evaluation						Testing scene evaluation					
Task		RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)	RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
Replica	Heuristic	29.60	0.0272	0.0482	0.0214	0.0049	0.9325	20.86	0.0395	0.0515	0.0471	0.0097	0.8543
	Semantic-NeRF	33.60	0.0211	0.0403	0.0128	0.0037	0.9507	27.08	0.0221	0.0418	0.0212	0.0055	0.9417
	SS-NeRF	33.76	0.0212	0.0383	0.0116	0.0035	0.9533	27.22	0.0224	**0.0405**	0.0196	0.0053	0.9483
	MuvieNeRF	**34.92**	**0.0193**	**0.0345**	**0.0100**	**0.0034**	**0.9582**	**28.55**	**0.0201**	0.0408	**0.0162**	**0.0051**	**0.9563**
SceneNet RGB-D	Heuristic	22.66	0.0496	-	0.0521	0.0093	0.8687	22.02	0.0394	-	0.0525	0.0124	0.8917
	Semantic-NeRF	28.29	0.0248	-	0.0212	0.0050	0.9152	28.85	0.0186	-	0.0198	0.0051	0.9417
	SS-NeRF	28.93	0.0244	-	0.0216	0.0050	0.9175	29.18	0.0182	-	0.0197	0.0052	0.9510
	MuvieNeRF	**29.29**	**0.0237**	-	**0.0207**	**0.0049**	**0.9190**	**29.56**	**0.0173**	-	**0.0189**	**0.0050**	**0.9556**

Table 1. Averaged performance of MuvieNeRF on Replica [46] and SceneNet RGB-D [30] datasets on both training scenes and testing scenes. Full results with multiple runs are provided in the supplementary, our model consistently outperforms both the single-task Semantic-NeRF baseline and multi-task SS-NeRF baseline, owing to the proposed CVA and CTA modules.

where the matrix $\mathbf{y}$ is made of input view annotations $\{\mathbf{Y}_i\}_{i=1}^{V}$ and $\mathbf{w}$ is obtained by an additional MLP layer which processes f_{stage2}.

3.3.3 Optimization

For the set of K tasks $\mathcal{T} = \{T_1, T_2, \cdots, T_K\}$ including the RGB colors, we apply their objectives individually and the final objective is formulated as

$$\mathcal{L}_{\text{MT}} = \sum_{T_j \in \mathcal{T}} \lambda_{T_j} \mathcal{L}_{T_j}, \tag{8}$$

where λ_{T_j} is the weight for the corresponding task T_j. For each task, $\mathcal{L}_{T_j}$ is formulated as:

$$\mathcal{L}_{T_j} = \sum_{\mathbf{r} \in \mathcal{R}} \mathcal{L}_j(\hat{\mathbf{y}}^j(\mathbf{r}), \mathbf{y}^j(\mathbf{r})), \tag{9}$$

where $\mathbf{y}^j(\mathbf{r}), \hat{\mathbf{y}}^j(\mathbf{r})$ are the ground-truth and prediction for a single pixel regarding task T_j. $\mathcal{R}$ is the set of rays $\mathbf{r}$ for all training views. $\mathcal{L}_j$ is chosen from L_1 loss, L_2 loss, and cross-entropy loss according to the characteristics of the tasks.

3.4. Tackling without Source-view Annotations

The proposed model is based on the assumption that source-view task annotations are available during *inference* time. The assumption rules out the influence from inaccurate source-view task information, which sets a cleaner environment to excavate multi-task synergy in a synthesis framework for the MTVS problem. However, from the real application perspective, traditional discriminative models only take RGB images as input without any task annotations. To demonstrate that our model is able to be applied in real scenarios, we introduce the more challenging **Inference Setting II** formulated by Equation 2 and provide a solution by incorporating a U-Net [42] shaped module F_{UNet} into our MuvieNeRF architecture. The detailed architecture of F_{UNet} is shown in the supplementary.

Conceptually, F_{UNet} takes RGB images from the V source views $\{\mathbf{x}_i\}_{i=1}^{V}$ as input and produces the corresponding multi-task annotations $\{\tilde{\mathbf{Y}}_i\}_{i=1}^{V}$, where $\tilde{\mathbf{Y}}_i = [\tilde{\mathbf{y}}_i^1, \cdots, \tilde{\mathbf{y}}_i^K]$. Next, similar to the conditional NeRF encoder, we retrieve the corresponding multi-task annotations $\{\tilde{\mathbf{Y}}_i(\mathbf{q})\}_{i=1}^{V}$ for an arbitrary point $\mathbf{q}$ by projection.

During training time, F_{UNet} is trained with pixel-wise task-specific losses. Concretely, for task T_j, we have:

$$\mathcal{L}_{U_j} = \sum_{\mathbf{r} \in \mathcal{R}} \sum_{i=1}^{V} \mathcal{L}_j(\tilde{\mathbf{y}}_i^j(\mathbf{r}), \mathbf{y}_i^j(\mathbf{r})). \tag{10}$$

The final loss becomes $\mathcal{L}_{\text{final}} = \sum_{T_j \in \mathcal{T}} \lambda_{T_j}(\mathcal{L}_{T_j} + \mathcal{L}_{U_j})$, for which we take the ground-truth multi-task annotations to learn the weights during training. However, we instead use the predictions produced by F_{UNet} for inference:

$$\hat{\mathbf{y}}^j = \sum_{i=1}^{V} \mathbf{w}[j,i] \cdot \tilde{\mathbf{y}}[j,i], \tag{11}$$

4. Experimental Evaluation

In this section, we start with main evaluations from Section 4.1 to 4.3, including experimental setting, quantitative and qualitative results, and comparison to conventional discriminative multi-task models. Next, we make further explorations in Section 4.4 and 4.5, including ablation studies and additional explorations. Finally, we discuss the limitations and future work in Section 4.6.

4.1. Experimental Setting

Model Instantiation: As illustrated in Section 3, our model can build upon arbitrary conditional NeRF encoders. For the main evaluation, we instantiate our model with state-of-the-art GeoNeRF [22]. We use $\alpha = 4$ and $\beta = 2$ for the number of self-attention unions in the CVA and CTA modules. We additionally show the performance with other NeRF encoders in the ablation study.

Benchmarks: We take two benchmarks for our main evaluation. **Replica** dataset [46] is a commonly-used indoor

Model	RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
MuvieNeRF$_{\text{w/o SH}}$	28.26	0.0204	-	0.0171	**0.0051**	0.9557
MuvieNeRF$_{\text{w/o KP}}$	27.96	0.0212	0.0423	0.0181	-	0.9519
MuvieNeRF	**28.55**	**0.0201**	**0.0408**	**0.0162**	**0.0051**	**0.9563**

Table 2. Additional test scene evaluation for our variants without SH (MuvieNeRF$_{\text{w/o SH}}$) and KP (MuvieNeRF$_{\text{w/o KP}}$) tasks on the Replica dataset. These two tasks work as a role of auxiliary tasks.

scene dataset containing high-quality photo-realistic 3D modelling of 18 scenes. Following the data acquisition method as [69], we manage to collect 22 scene sequences each containing 50 frames at a resolution of 640×480. **SceneNet RGB-D** dataset [30] is a large-scale photorealistic indoor scene dataset expanding from SceneNet [16]. We include 32 scenes with 40 frames of each at a resolution of 320×240 in our evaluation. Besides the above two datasets, we further evaluate zero-shot adaptation on four out-of-distribution datasets: LLFF [31], TartanAir [55], ScanNet [9] and BlendedMVS [59].

Task Selection: We select six representative tasks to evaluate our method following previous multi-task learning pipelines [68, 45]. The tasks are Surface Normal Prediction **(SN)**, Shading Estimation **(SH)**, Edges Detection **(ED)**, Keypoints Detection **(KP)**, Semantic Labeling **(SL)**, together with the **RGB** synthesis. For the SceneNet RGB-D dataset, we drop the SH task due to missing annotations.

Evaluation Set-up: For the Replica dataset, we divide the 22 scenes into 18, 1, and 3 for training, validation, and testing, respectively. For SceneNet RGB-D, we split 26 scenes for training, 2 for validation, and 4 for testing. For each scene, we hold out every 8 frames as testing views.

For these held-out views, we provide two types of evaluations: *Training scene evaluation* is conducted on novel views from the training scenes; *Testing scene evaluation* runs on novel scenes and is used to evaluate the generalization capacity of the compared models.

Evaluation Metrics: For RGB synthesis, we measure Peak Signal-to-Noise Ratio (PSNR) for evaluation. For semantic segmentation, we take mean Intersection-over-Union (mIoU). For the other tasks, we evaluate the L_1 error.

Baselines: We consider synthesis baselines for the main evaluation. **Semantic-NeRF** [69] extends NeRF for the semantic segmentation task. We further extend this model the same way for other tasks, which only considers single-task learning in a NeRF style. **SS-NeRF** [68] considers multi-task learning in a NeRF style, but ignores the cross-view and cross-task information. We equip both models with the same GeoNeRF backbone as our model. Following [68], we also include a **Heuristic** baseline which estimates the annotations of the test view by projecting the source labels from the nearest training view to the target view.

Implementation Details: We set the weights for the six chosen tasks as $\lambda_{\text{RGB}} = 1$, $\lambda_{\text{SN}} = 1$, $\lambda_{\text{SL}} = 0.04$, $\lambda_{\text{SH}} = 0.1$, $\lambda_{\text{KP}} = 2$, and $\lambda_{\text{ED}} = 0.4$ based on empirical observations. We use the Adam [25] optimizer with an initial learning rate of 5×10^{-4} and set $\beta_1 = 0.9, \beta_2 = 0.999$. During training, each iteration contains a batch size of 1024 rays randomly sampled from all training scenes.

More details about our encoder architectures, dataset processing, out-of-distribution analysis, and implementations are included in the supplementary.

4.2. MuvieNeRF Is Capable of Solving MTVS

In Table 1, we present the average results derived from the held-out views across both the training and testing scenes. The key observations are: Firstly, it is clear that our problem statement is non-trivial as evidenced by the notably inferior performance exhibited by the simple heuristic baseline when compared to the other models. Secondly, SS-NeRF registers a minor performance, surpassing Semantic-NeRF on average, indicating the contribution of multi-task learning. Lastly, our model consistently surpasses all the baselines, reaffirming that cross-view and cross-task information are invariably beneficial within our framework.

Interestingly, we noted that MuvieNeRF exhibited a performance closely comparable to the two NeRF baselines on novel scenes for the KP and SH tasks. To decipher the underlying reason, we carried out an additional evaluation on two variants of our model without the two tasks, MuvieNeRF$_{\text{w/o KP}}$ and MuvieNeRF$_{\text{w/o SH}}$, on the test scenes on Replica in Table 2. Our findings indicate that the KP and SH tasks indeed enhance the learning of other tasks, serving as effective auxiliary tasks. This conclusion aligns with previous studies on traditional multi-task learning models as reported by [45].

Figure 3 showcases a comparative analysis of the qualitative results. It is evident that our predictions supersede those of other baselines in terms of precision and clarity. This superior performance can be attributed to the additional information provided by shared cross-view and cross-task knowledge, which proves beneficial for the target tasks.

4.3. MuvieNeRF Beats Discriminative Models

Although conventional discriminative models fall short in addressing the proposed MTVS problem, we have explored several hybrid settings to facilitate a comparison between our MuvieNeRF and these discriminative models.

Hybrid Set-up: We provide *additional* RGB images from novel views to the discriminative models under three settings with different choices of RGB images. (1) We train on GT pairs and evaluate on novel view images generated by a NeRF (*NeRF's Images (No Tuned)*); (2) We additionally fine-tune the discriminative models with paired NeRF's images and corresponding GT (*NeRF's Images (Tuned)*); (3) We evaluate on GT images from novel views as the perfor-

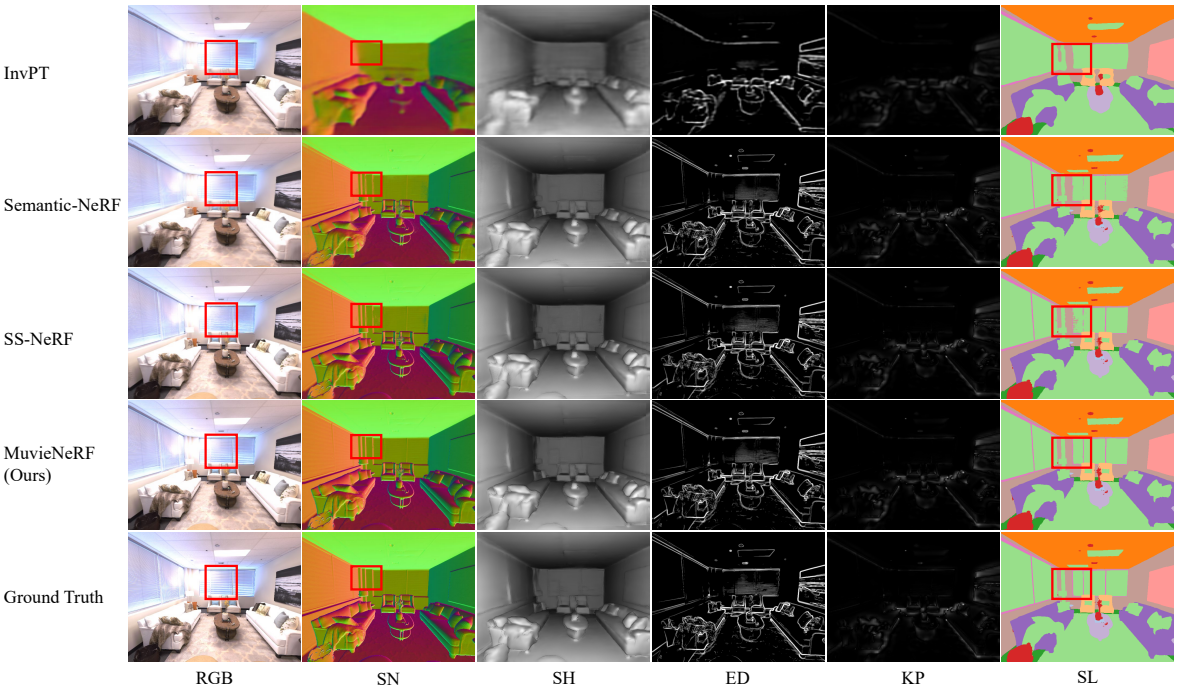


Figure 3. Visual comparisons of our model and baselines on a test scene of Replica dataset [46]. Our predictions are sharper and more accurate compared with other baselines. The underlying reason is that shared cross-view and cross-task knowledge can provide additional information for the target tasks.

Model	NeRF's Images (No Tuned)					NeRF's Images (Tuned)					GT Images (Upper Bound)				
	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
Taskgrouping	0.0568	0.0707	0.0408	0.0089	0.5361	0.0530	0.0677	0.0423	0.0090	0.5590	0.0496	0.0607	0.0298	0.0060	0.6191
MTI-Net	0.0560	0.0636	0.0418	**0.0078**	0.5440	0.0486	**0.0549**	0.0389	0.0078	0.6753	0.0422	0.0498	**0.0281**	**0.0050**	0.7196
InvPT	**0.0479**	**0.0618**	**0.0400**	0.0091	**0.7139**	**0.0474**	0.0587	**0.0328**	**0.0074**	**0.7084**	**0.0409**	**0.0484**	0.0282	0.0055	**0.8158**
Ours	**0.0201**	**0.0408**	**0.0162**	**0.0051**	**0.9563**	-	-	-	-	-	-	-	-	-	-

Table 3. Comparison to the discriminative models for the test scenes on Replica [46] dataset. MuvieNeRF clearly beats all the discriminative models in all three settings, indicating that our model is more capable of both performance and generalizability.

Model	RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
MuvieNeRF$_{\text{w/o CTA}}$	27.55	0.0214	0.0424	0.0198	0.0056	0.9501
MuvieNeRF$_{\text{w/o CVA}}$	28.25	0.0206	**0.0407**	0.0170	0.0052	0.9557
MuvieNeRF	**28.55**	**0.0201**	0.0408	**0.0162**	**0.0051**	**0.9563**

Table 4. Ablation study with CTA and CVA modules on Replica [46] dataset. MuvieNeRF$_{\text{w/o CTA}}$ is the variant without CTA module; MuvieNeRF$_{\text{w/o CVA}}$ is the variant without CVA module. The CTA module is more crucial compared to the CVA module while combining them leads to the best performance.

Backbone	RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
PixelNeRF + SS-NeRF	23.21	0.0328	0.0457	0.0376	0.0074	0.8620
PixelNeRF + Ours	24.14	0.0302	0.0420	0.0342	0.0068	0.8961
GNT + SS-NeRF	21.51	0.0405	0.0497	0.0420	0.0096	0.8302
GNT + Ours	22.67	0.0333	0.0455	0.0384	0.0076	0.8686
MVSNeRF + SS-NeRF	25.27	0.0261	0.0419	0.0248	0.0061	0.9294
MVSNeRF + Ours	25.73	0.0248	0.0408	0.0227	0.0056	0.9303
GeoNeRF + SS-NeRF	27.22	0.0224	**0.0405**	0.0196	0.0053	0.9483
GeoNeRF + Ours	**28.55**	**0.0201**	0.0408	**0.0162**	**0.0051**	**0.9563**

Table 5. Ablation study with different choices of conditional NeRF encoders. The proposed MuvieNeRF is universally beneficial to different encoders, owing to the proposed CTA and CVA modules.

Model	SN (↓)	ED (↓)	KP (↓)
Discriminative	0.0778	0.0355	0.0086
MuvieNeRF$_D$	**0.0605**	**0.0230**	**0.0074**

Table 6. Results for the setting with unknown nearby-view annotations. MuvieNeRF$_D$ still outperforms the hybrid discriminative model with a similar backbone.

mance upper bound (*GT Images (Upper Bound)*). For all the settings, we train the discriminative models on both training and testing scenes (training views only) to make sure that they get access to the same number of data as MuvieNeRF.

Discriminative Baselines: As the comparison to discriminative models is not our main goal, we select three representative baselines of different architectures. **Taskgrouping** [45] leverages an encoder-decoder architecture with a shared representation. **MTI-Net** [51] adopts a CNN-based backbone and enables multi-scale task interactions. **InvPT** [61] takes a transformer-based architecture that en-

Evaluation Type		Training scene evaluation						Testing scene evaluation					
Tasks		RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)	RGB (↑)	SN (↓)	SH (↓)	ED (↓)	KP (↓)	SL (↑)
Replica	SS-NeRF	33.76	0.0212	0.0383	0.0116	0.0035	0.9533	27.22	0.0224	**0.0405**	0.0196	0.0053	0.9483
	MuvieNeRF	**34.92**	**0.0193**	**0.0345**	**0.0100**	**0.0034**	**0.9582**	**28.55**	**0.0201**	0.0408	**0.0162**	**0.0051**	**0.9563**
	SS-NeRF (Enlarged)	34.20	0.0210	0.0388	0.0107	0.0035	0.9557	27.37	0.0226	**0.0405**	0.0186	0.0053	0.9498
SceneNet RGB-D	SS-NeRF	28.93	0.0244	-	0.0216	0.0050	0.9175	29.18	0.0182	-	0.0197	0.0052	0.9510
	MuvieNeRF	**29.29**	**0.0237**	-	**0.0207**	**0.0049**	**0.9190**	**29.56**	**0.0173**	-	**0.0189**	**0.0050**	**0.9556**
	SS-NeRF (Enlarged)	29.03	0.0244	-	0.0215	**0.0049**	0.9186	29.46	0.0182	-	0.0191	**0.0050**	0.9520

Table 7. Comparison between MuvieNeRF and an enlarged version of SS-NeRF. our model still significantly outperforms SS-NeRF (Enlarged), indicating that the good performance of MuvieNeRF is not simply achieved by a heavier decoder, and demonstrating the effectiveness of our module designs.

Model	Runtime (Per Training Iter.)	Num. of Params.	FLOPs
SS-NeRF	1×	1.21M	4.57×10^{11}
MuvieNeRF	1.22×	1.30M	5.84×10^{11}

Table 8. Computational cost for SS-NeRF and MuvieNeRF on Replica dataset. Our CTA and CVA modules are light-weight designs.

courages long-range and global context to benefit MTL.

Table 3 details the averaged results and Figure 3 offers a visual comparison. It is discernible that our MuvieNeRF outstrips all the discriminative models, illustrating that these models struggle to effectively tackle the MTVS problem, despite fine-tuning or utilizing GT images. We surmise that this shortcoming is rooted in the evaluation of novel scenes, wherein the generalization capacity of our model noticeably outperforms that of discriminative ones.

4.4. Ablation Study

We consider the test scene evaluation on the Replica dataset for the following ablation and exploration sections.

Contributions of CTA and CVA: We dissect the individual contributions of the proposed CTA and CVA modules in Table 4. An examination of the results reveals a more significant impact from the CTA module when compared to the CVA module. We postulate that this occurs because the NeRF encoder and volume rendering have already mastered an implicit 3D representation, albeit falling short in modeling task relationships. Nevertheless, the integration of both modules yields further enhancements.

Choice of Condition NeRF: We also ablate the conditional NeRF encoders with PixelNeRF [64], MVSNeRF [6], GNT [48] and GeoNeRF [22] in Table 5. Note that we only adopt the encoder part defined in Section 3, causing a variation in performance from the full model. Each variant under our design surpasses its respective SS-NeRF baselines, affirming the universal advantage of our proposed CTA and CVA modules across different conditional NeRF encoders.

Effectiveness of CTA and CVA Modules: By integrating CVA and CTA modules, we concurrently increase the number of trainable parameters. To truly assess the advancements brought by the proposed CVA and CTA modules, we introduce an expanded version of SS-NeRF, designated as

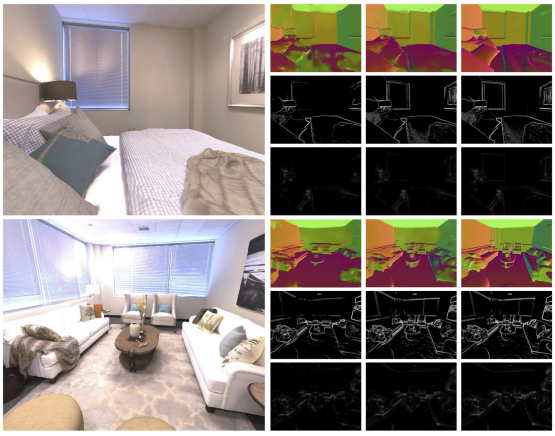


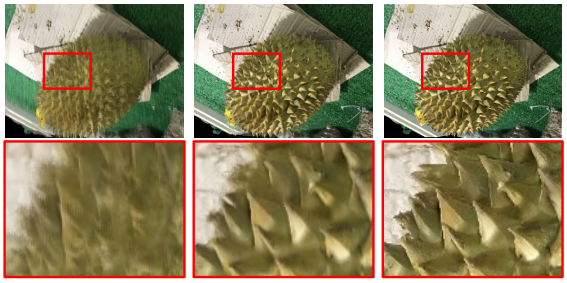
Figure 4. Visualizations for the setting without nearby-view annotations. our MuvieNeRF$_D$ markedly surpasses the discriminative baseline model [45] and can generate results closely paralleling the ground truth, indicating MuvieNeRF$_D$ is capable of tackling the more challenging settings.

SS-NeRF (Enlarged). This variant features a doubled latent dimension, comprising 2.09M parameters and 6.18×10^{11} FLOPs. The outcomes, as illustrated in Table 7, confirm that MuvieNeRF's excellent performance is not simply the result of a more complex decoder, thereby highlighting the effectiveness of our module designs. Moreover, we further list the computational cost for our model and SS-NeRF in Table 8 to show that our CTA and CVA modules are indeed lightweight, but effective designs.

4.5. Additional Explorations

4.5.1 Tackling with Unknown Nearby-view Labels

In this section, we apply the variant discussed in Section 3.4 to tackle the more challenging problem setting with unknown nearby-view annotations. In Table 6, we show the quantitative comparisons for this variant, **MuvieNeRF$_D$** and a hybrid baseline, **Discriminative** [45], which shares almost the same architecture as our discriminative module. We use pre-trained weights from Taskonomy [66] to initialize weights for both models. Limited by the computational constraint, we select three tasks, SN, ED, and KP with the



GeoNeRF MuvieNeRF (Ours) Ground Truth

Figure 5. An out-of-distribution comparison on BlendedMVS dataset [59] for GeoNeRF and our MuvieNeRF. Our model offers superior visual quality and preserves sharp contours. More visualizations about other OOD benchmarks are included in the supplementary.

closest relationships [45] for demonstration in this setting.

In conjunction with the visualizations illustrated in Figure 4, it is clear that our MuvieNeRF$_D$ markedly surpasses the discriminative baseline model [45] and can generate results closely paralleling the ground truth. This observation indicates our model is capable of tackling more challenging settings. We conjecture the reason is that the weighted sum format (Equation 11) enhances the fault tolerance of the predictions.

4.5.2 Out-of-distribution Generalization

We demonstrate how the multi-task information learned from one dataset can effectively be utilized to benefit other datasets by performing a zero-shot adaption on out-of-distribution datasets with our MuvieNeRF trained on the Replica. This takes a step further from investigating the generalization ability of unseen testing scenes in previous sections. We considered four datasets in total: LLFF [31], TartanAir [55], ScanNet [9] and BlendedMVS [59] containing indoor, outdoor and even object-centric scenes.

As an illustrative example, we showcase the comparison between conditional NeRF backbone, GeoNeRF [22], and our MuvieNeRF for novel-view RGB synthesis task in Table 9 and Figure 5. Evidently, our model surpasses GeoNeRF by a significant margin, offering superior visual quality and retaining sharp contours, likely a result of the edge and surface normal information absorbed during the multi-task training. These outcomes substantiate that augmenting the model with more tasks, as part of our multi-task learning strategy, dramatically bolsters the generalization capacity, thereby showcasing its immense potential for real-world applications.

4.6. Limitations and Future work

Limitations: One major limitation of this work is the reliance on data. MuvieNeRF requires images from dense views, a requirement not fulfilled by most multi-task benchmarks. To circumvent this limitation, techniques that allow NeRF to learn from sparse views [35, 67] could be adopted.

Model	ScanNet	TartanAir	LLFF	BlendedMVS
GeoNeRF	31.71	26.51	20.68	16.27
MuvieNeRF	**32.76**	**30.21**	**22.91**	**20.97**

Table 9. Averaged PSNR of out-of-distribution RGB synthesis task. Our model surpasses the GeoNeRF baseline by a large margin, affirming that the inclusion of additional tasks, as part of our approach, contributes to a substantial enhancement in the model's generalization capacity.

Task Relationships: As elaborated in Section 4.2, SH and KP function as auxiliary tasks within the system. A deeper exploration into the relationships between tasks and the corresponding geometric underpinnings within our synthesis framework offers intriguing avenues for future research.

Extension to Other Synthesis Models: We have demonstrated that incorporating cross-view geometry and cross-task knowledge can enhance multi-task learning for synthesis models. We anticipate that similar strategies could be extended to 3D synthesis models other than NeRF, such as point clouds [57] and meshes [17, 26].

5. Conclusion

This paper introduces the novel concept of Multi-Task View Synthesis (MTVS), recasting multi-task learning as a set of view synthesis problems. Informed by this new perspective, we devise MuvieNeRF, a unified synthesis framework enriched with novel Cross-View Attention and Cross-Task Attention modules. MuvieNeRF enables the simultaneous synthesis of multiple scene properties from novel viewpoints. Through extensive experimental evaluations, we establish MuvieNeRF's proficiency in addressing the MTVS task, with performance exceeding that of discriminative models across various settings. Our model also demonstrates broad applicability, extending to a variety of conditional NeRF backbones.

Acknowledgement

We thank Kangle Deng, Derek Hoiem, Ziqi Pang, Deva Ramanan, Manolis Savva, Shen Zheng, Yuanyi Zhong, Jun-Yan Zhu, and Zhen Zhu for their valuable comments.

This work was supported in part by NSF Grant 2106825, Toyota Research Institute, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014 and CIS230012 from the ACCESS program.

References

[1] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *ICML*, 2022. 3

[2] Zhipeng Bao, Yu-Xiong Wang, and Martial Hebert. Bowtie networks: Generative modeling for joint few-shot recognition and novel-view synthesis. In *ICLR*, 2021. 2

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[4] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. MulT: An end-to-end multitask learning transformer. In *CVPR*, 2022. 3

[5] Rich Caruana. Multitask learning. *Machine Learning*, 1997. 1

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 2021. 2, 3, 8

[7] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 3

[8] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *NeurIPS*, 2020. 3

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 6, 9

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 2

[11] Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 3D-aware conditional image synthesis. In *CVPR*, 2023. 2

[12] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In *ICCV*, 2021. 2

[13] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007. 2

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *ICLR*, 2022. 2

[15] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *ICML*, 2020. 3

[16] Ankur Handa, Viorica Pătrăucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. SceneNet: Understanding real world indoor scenes with synthetic data. In *CVPR*, 2016. 6

[17] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3D sheet for view synthesis from a single image. In *ICCV*, 2021. 9

[18] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. EfficientNeRF: Efficient neural radiance fields. In *CVPR*, 2022. 2

[19] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view PointNet for 3D scene understanding. In *ICCVW*, 2019. 2

[20] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *ICML*, 2022. 3

[21] Adrián Javaloy and Isabel Valera. RotoGrad: Gradient homogenization in multitask learning. In *ICLR*, 2022. 3

[22] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing NeRF with geometry priors. In *CVPR*, 2022. 2, 3, 4, 5, 8, 9

[23] Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. In *ECCV*, 2020. 3

[24] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[26] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vMAP: Vectorised object mapping for neural field SLAM. In *CVPR*, 2023. 2, 9

[27] Andreas Kurz, Thomas Neff, Zhaoyang Lv, Michael Zollhöfer, and Markus Steinberger. AdaNeRF: Adaptive sampling for real-time rendering of neural radiance fields. In *ECCV*, 2022. 2

[28] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021. 3

[29] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 2

[30] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet RGB-D: 5M photorealistic images of synthetic indoor trajectories with ground truth. In *ICCV*, 2017. 5, 6

[31] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *TOG*, 2019. 6, 9

[32] 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, 3, 4

[33] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 1, 3

[34] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019. 2

[35] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 2, 9

[36] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[37] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 2

[38] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2

[39] 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 *ICCV*, 2021. 2

[40] Etienne Pelaprat and Michael Cole. "Minding the gap": Imagination, creativity and human cognition. *Integrative Psychological and Behavioral Science*, 2011. 1

[41] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In *ICCV*, 2021. 2

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[43] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3D scene understanding with neural fields. In *CVPR*, 2023. 2

[44] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3D feature embeddings. In *CVPR*, 2019. 2

[45] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, 2020. 2, 3, 6, 7, 8, 9

[46] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5, 7

[47] Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. Task switching network for multi-task learning. In *ICCV*, 2021. 3

[48] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all NeRF needs? In *ICLR*, 2023. 2, 3, 8

[49] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *CVPR*, 2022. 2

[50] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 2

[51] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 7

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[53] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 2

[54] Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *CVPR*, 2022. 3

[55] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IROS*, 2020. 6, 9

[56] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 2

[57] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 2, 9

[58] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. In *ECCV*, 2022. 2

[59] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 6, 9

[60] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 2

[61] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 7

[62] Hanrong Ye and Dan Xu. TaskPrompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023. 4

[63] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 2

[64] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 3, 8

[65] Amir R. Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 1, 3

[66] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1, 2, 3, 8

[67] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for

sparse-view 3D reconstruction in the wild. In *NeurIPS*, 2021. 9

[68] Mingtong Zhang, Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Beyond RGB: Scene-property synthesis with neural radiance fields. In *WACV*, 2023. 2, 6

[69] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2, 6