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 fundamental problem in computer vision. However, current research has primarily focused on the multi-task dense prediction setting, which fails to account for the underlying 3D world with multi-view consistent structures and lacks the ability to hallucinate. In this work, we introduce a novel problem setting called multi-task view synthesis (MTVS), which reformulates multi-task prediction as a set of novel-view synthesis tasks for multiple scene properties, including RGB. To tackle this problem, we propose MuvieNeRF, a unified framework equipped with our novel Cross-View Attention (CVA) and Cross-Task Attention (CTA) modules. With these modules, MuvieNeRF is able to facilitate the interaction among the bottom-up signals from different downstream tasks and different source views, thereby enabling the flow of knowledge sharing across all the tasks. MuvieNeRF is capable of simultaneously synthesizing different scene properties with promising visual quality, outperforming conventional discriminative models in various settings.

1. Introduction

When observing a scene, humans are able to mentally simulate how the objects within it would look like from a novel viewpoint and in a *versatile* manner, hallucinating not only the color of the objects but also various associated scene properties, such as their surface orientation, semantic marks, and edge patterns [35]. Motivated by this, there has been a growing interest in equipping modern robots with similar capabilities to solve multiple tasks. However, current research [28, 58, 57] has primarily focused on the *multi-task dense prediction* setting, which involves using a conventional discriminative model to jointly predict multiple pixel-level scene properties with the given RGB images (see Figure 1(a)). Approaches developed for this setting are restrictive in practice, because they often treat each image individually, without constructing an explicit model of the 3D world that adheres to the principle of multi-view consistency. More importantly, they lack the ability to "imagine" – they are not able to infer scene properties from an *unseen* view as the RGB images are always required.

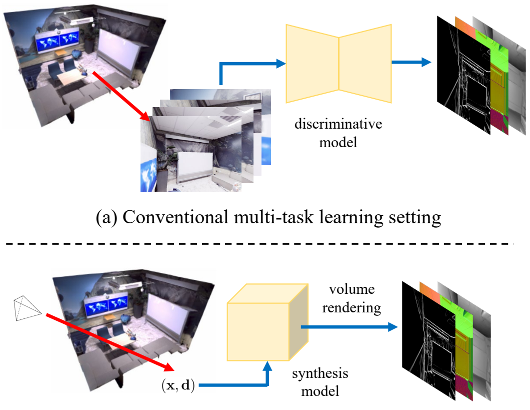
discriminative model
(a) Conventional multi-task learning setting
volume rendering
(x, d)
synthesis model
(b) Our multi-task view synthesis pipeline

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 takes single images and makes predictions for different visual tasks. Multi-task view synthesis aims to render visualizations for multiple scene properties at novel views.

To address these limitations, this work revisits multi-task learning (MTL) [5] from a novel *synthesis* perspective and proposes a new, more flexible problem setting that formulates multi-task visual learning as a set of novel-view synthesis problems. We refer to this setting as *multi-task view synthesis* (MTVS) (see Figure 1(b)). For example, predicting surface normals for a given image can be treated as visualizing a 3-channel "image" with the pose and camera parameters of the input image. One key question that may arise in solving this problem is *whether a synthesis model is capable of rendering multiple scene properties*, given that conventional discriminative models are insufficient. The great success of neural radiance fields (NeRF) [27] has shown that the answer to this question is **yes** – fortunately! NeRF's implicit scene representation makes it possible to extend to other scene properties beyond RGB [60]. Moreover, this scene representation considers multi-view geometry, which is consequently beneficial for all the tasks.

With this insight, we introduce a unified framework called MuvieNeRF, which leverages ***Mu***ti-task and cross-***vie***w knowledge so that can simultaneously synthesize multiple scene properties with a shared implicit scene representation. The proposed MuvieNeRF can be applied to an arbitrary conditional NeRF architecture and features a unified decoder architecture with two key modules: *cross-view attention (CVA) module* and *cross-task attention (CTA) module*. The CVA module leverages and aligns the features among multiple reference views and the target view to enforce cross-view consistency. The CTA module explores relationships among different scene properties, which have been widely studied but within discriminative models [58, 39] to achieve better performance. Incorporating these two modules within MuvieNeRF enables effective leveraging of information from multiple views and across multiple tasks, leading to improved performance across all tasks.

To demonstrate the efficacy of our approach, we instantiate our MuvieNeRF with a state-of-the-art conditional NeRF model, GeoNeRF [20], and conduct a comprehensive evaluation on both synthetic and real-world benchmarks. Our results show that MuvieNeRF is capable of solving the MTVS task, and even outperforms several competitive discriminative models in different settings.

2. Method

We describe our novel multi-task view synthesis setting and the proposed MuvieNeRF (as shown in Figure 2) in this section.

2.1. Multi-task View Synthesis Set-up

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 annotations with camera poses as a reference, and predicts the annotations for a novel view given camera pose:

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

where $\mathbf{Y} = \left[\mathbf{x}, \mathbf{y}^1, \cdots, \mathbf{y}^K\right]$ denotes RGB images $\mathbf{x}$ and K other multi-task annotations $\{\mathbf{y}^j\}_{j=1}^{K}$. $\mathbf{P}_i$ is the i^{th} source camera pose, and $\mathbf{P}_T$ is the target camera pose.

We evaluate the model Φ in the scene level as it requires a few paired annotations from the same scene (see Equation 1). However, Φ is supposed to learn the implicit scene representation during training as well so that it is able to generalize to novel scenes that are not seen during training. Thus, our proposed MuvieNeRF is built upon conditional NeRF backbones. Conditional NeRFs [61, 56, 20, 42] learn scene representation across multiple scenes during training and are capable of generalizing to novel scenes.

2.2. MuvieNeRF

As demonstrated by Figure 2, MuvieNeRF first fetches the scene representation f_{scene} from the conditional NeRF encoder, then predicts multiple scene properties $\left[\mathbf{x}_q, \mathbf{y}_q^1, \cdots, \mathbf{y}_q^K\right]$ for arbitrary 3D coordinate $\mathbf{q}$. We illustrate how to predict multiple scene properties with f_{scene} and source annotations $\{(\mathbf{Y}_i, \mathbf{P}_i)\}_{i=1}^{V}$ as follows.

2.2.1 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 the f_{scene} with a positional embedding derived from the target ray and the source-view image plane: $f_{\text{scene}}^{\text{pos}} = [f_{\text{scene}}; \gamma(\theta_{n,v})]$, where $\gamma(\cdot)$ is the sinusoidal positional encoding proposed in [27], 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 novel ray, which measures the similarity between the source view v and the target view.

Next, α Cross-View Attention 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): $f_{\text{CVA}} = \text{MLP}_{\text{CVA}}(f_{\text{scene}}^{\text{pos}} + \text{MHA}(f_{\text{scene}}^{\text{pos}}, f_{\text{scene}}^{\text{pos}}))$, where $\text{MHA}(a, b)$ denotes multi-head attention [46] with a as query and b as key and value.

After these processes, we apply K different MLPs to broadcast the shared feature to K downstream tasks, leading to the K-branch feature $f_{\text{task}} \in \mathbb{R}^{K \times V \times c'}$.

2.2.2 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 [54], $p_t \in \mathbb{R}^{K \times c_t}$. The first attention component applies cross-attention between features from each branch and the task prompts $f_{s1} = f_{\text{task}} + \text{MHA}(f_{\text{task}}, p_t)$. 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_{s1}^j for task T_j and the corresponding task prompt p_t^j to obtain $f_{s1'}$.

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: $f_{s2} = \text{MLP}_{\text{CTA}}(f_{s1'} + \text{MHA}(f_{s1'}, f_{s1'}))$.

Finally, to predict the annotations of the target view, we adopt the formulation of GeoNeRF [20]. The prediction $\hat{\mathbf{y}}^j$ of task T_j on the target view is the weighted sum of the source views:

$$\hat{\mathbf{y}}^j = \sum_{i=1}^{V} \mathbf{w}[j, i] \cdot \mathbf{y}[j, i], \quad (2)$$

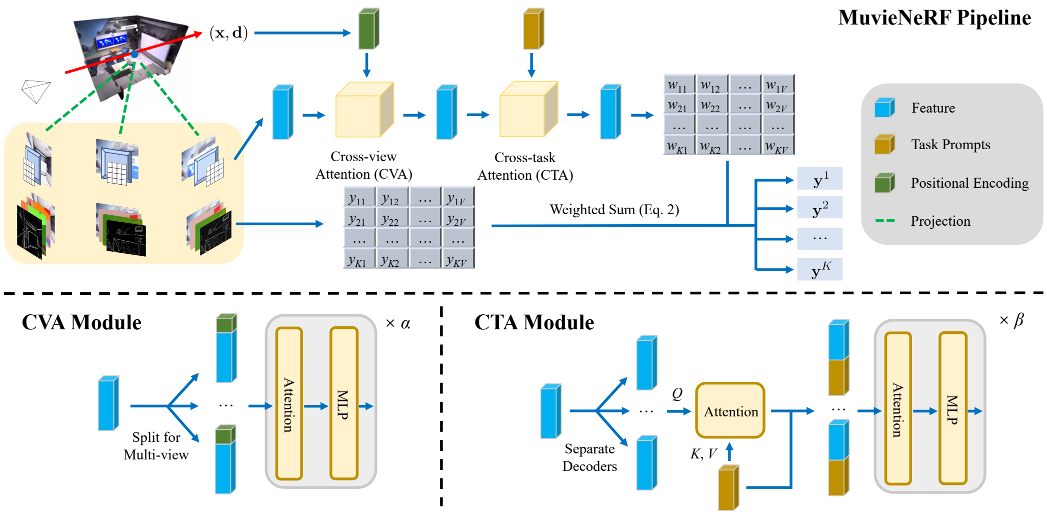

Figure 2. Model architecture. MuvieNeRF is a unified framework for multi-task view synthesis equipped with Cross-View Attention (CVA) and Cross-Task Attention (CTA) modules. It predicts multiple scene properties for arbitrary 3D coordinates with nearby-view annotations.

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_{s2}.

2.2.3 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 $\mathcal{L}_{\mathrm{MT}} = \sum_{T_j \in \mathcal{T}} \lambda_{T_j} \mathcal{L}_{T_j}$, where λ_{T_j} is the weight for the corresponding task T_j. For each task, $\mathcal{L}_{T_j}$ is formulated as:

$$\mathcal{L}_{T_j} = \sum_{\mathbf{r} \in \mathcal{R}} \mathcal{L}_j(\hat{\mathbf{y}}^j(\mathbf{r}), \mathbf{y}^j(\mathbf{r})), \tag{3}$$

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 $\mathbf{r}$ in each batch. $\mathcal{L}_j$ is chosen from L_1 loss, L_2 loss, and cross-entropy loss according to the characteristics of the tasks.

3. Experimental Evaluation

We show the quantitative and qualitative results, and comparison to conventional discriminative models in this section.

3.1. Experimental Setting

Set up. For the main evaluation, we instantiate our model with state-of-the-art GeoNeRF [20]. We set $\alpha = 4$ and $\beta = 2$ for the number of self-attention unions in the CVA and CTA modules. We pick six representative tasks for evaluation: Surface Normal Prediction **(SN)**, Shading Estimation **(SH)**, Edges Detection **(ED)**, Keypoints Detection **(KP)**, Semantic Labeling **(SL)**, together with the **RGB** synthesis.

Benchmarks: We take two benchmarks for our main evaluation. For **Replica** dataset [40], we manage to collect 22 scene sequences each containing 50 frames at a resolution of 640 × 480. For **SceneNet RGB-D** dataset [26], we include 32 scenes with 40 frames of each at a resolution of 320 × 240 in our evaluation.

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* is used to evaluate the generalization capacity of the compared models to novel scenes.

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** [61] 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** [60] 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 [60], 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.

3.2. MuvieNeRF Is Capable of Solving MTVS

We report the average results on the held-out views of both training and testing scenes in Table 1 and Figure 3. We

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 [40] and SceneNet RGB-D [26] 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.

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 2. Comparison to the discriminative models for the test scenes on Replica [40] dataset. MuvieNeRF clearly beats all the discriminative models in all three settings, indicating that our model is more capable of both performance and generalizability.

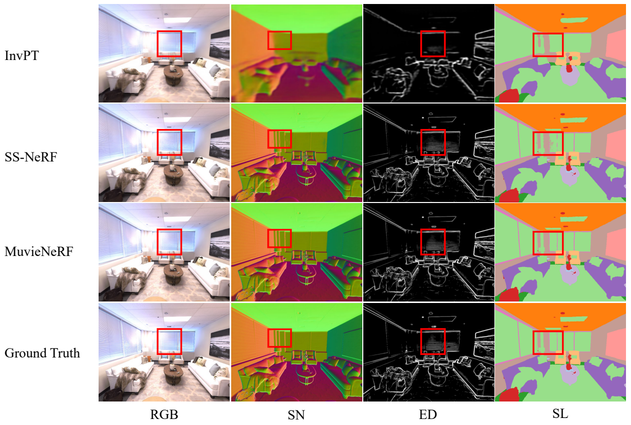

Figure 3. Visual comparisons of our model and baselines. Our predictions are sharper and more accurate.

have the following observations: First, the simple heuristic baseline has significantly worse performance compared with other models, showing that our problem setting is non-trivial. Next, SS-NeRF outperforms Semantic-NeRF marginally on average, indicating the contribution of multi-task learning. Finally, our model consistently outperforms all the baselines, demonstrating that the cross-view and cross-task information is universally helpful.

3.3. MuvieNeRF Beats Discriminative Models

Though the conventional discriminative models are not capable of solving the proposed MTVS problem, we do provide several hybrid settings for comparison.

Hybrid Set-up: The high-level idea is to provide *additional* RGB images from novel views to the discriminative models. We provide three different 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 performance 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 our proposed MuvieNeRF.

We select three representative baselines of different architectures: **Taskgrouping** [39], **MTI-Net** [45] and **InvPT** [53]. The averaged results are reported in Table 2 and a visual comparison is shown in Figure 3. Our MuvieNeRF clearly beats all the discriminative models and it is clear to find that the discriminative models do not work well for the MTVS problem, even if after fine-tuning or with ground-truth images. We think the reason lies in the evaluation of novel scenes – the generalization capacity of discriminative models is not as good as our model.

4. Discussions

Limitations: One major limitation of this work is the reliance on data. MuvieNeRF requires images from dense views, while most multi-task benchmarks do not satisfy. To address this limitation, some techniques enabling NeRF to learn from sparse views [30, 59] can be applied.

Task Relationships: As discussed in Section 3.2, SH and KP tasks are working as a role of auxiliary tasks. Further comprehensive explorations on the task relationships and the underlying geometric reasons within our synthesis framework are interesting directions for future work.

Extension to Other Synthesis Models: The motivation of this work is that the cross-view geometry and shared knowledge across tasks can facilitate multi-task learning, not only for discriminative models but for synthesis models as well. We believe similar strategies can be applied to other formats of synthesis models for 3D scene representations, such as point clouds [50] and meshes [15, 22].

References

[1] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *ICML*, 2022. 7

[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. 7

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for antialiasing neural radiance fields. In *ICCV*, 2021. 7

[4] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. MulT: An end-to-end multitask learning transformer. In *CVPR*, 2022. 7

[5] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 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. 7

[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. 7

[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. 7

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 7

[10] Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 3D-aware conditional image synthesis. In *CVPR*, 2023. 7

[11] 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. 7

[12] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007. 7

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *ICLR*, 2022. 7

[14] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *ICML*, 2020. 7

[15] 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. 4

[16] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. EfficientNeRF: Efficient neural radiance fields. In *CVPR*, 2022. 7

[17] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view PointNet for 3D scene understanding. In *ICCVW*, 2019. 7

[18] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *ICML*, 2022. 7

[19] Adrián Javaloy and Isabel Valera. RotoGrad: Gradient homogenization in multitask learning. In *ICLR*, 2022. 7

[20] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing NeRF with geometry priors. In *CVPR*, 2022. 2, 3, 7

[21] 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. 7

[22] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vMAP: Vectorised object mapping for neural field SLAM. In *CVPR*, 2023. 4, 7

[23] 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. 7

[24] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021. 7

[25] 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. 7

[26] 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. 3, 4

[27] 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. 1, 2, 7

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

[29] 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. 7

[30] 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. 4, 7

[31] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 7

[32] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 7

[33] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 7

[34] 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. 7

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

[36] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In *ICCV*, 2021. 7

[37] 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. 7

[38] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3D feature embeddings. In *CVPR*, 2019. 7

[39] 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, 4

[40] 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. 3, 4

[41] 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. 7

[42] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all NeRF needs? In *ICLR*, 2023. 2, 7

[43] 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. 7

[44] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 7

[45] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 4

[46] 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. 2

[47] 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. 7

[48] 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. 7

[49] 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. 7

[50] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 4, 7

[51] 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. 7

[52] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 7

[53] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 4

[54] Hanrong Ye and Dan Xu. TaskPrompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023. 2

[55] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 7

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 7

[57] 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

[58] 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

[59] 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. 4

[60] Mingtong Zhang, Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Beyond RGB: Scene-property synthesis with neural radiance fields. In *WACV*, 2023. 1, 3, 7

[61] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2, 3, 7

Supplementary Material

In this supplementary material, we first present the related works of our proposed MTVS problem in Section A. Then, in Section B we introduce the preliminary of conditional neural radiance fields. Next, in Section C we provide additional full qualitative results on all the modelled tasks for the two main datasets Replica and SceneNet RGB-D.

A. Related Work

In this work, we propose a *NeRF* model which leverages *multi-task* and *cross-view* Info for *multi-task view synthesis*. We review the most relevant works in these areas below.

View synthesis aims to generate a target image with an arbitrary camera pose from given source images [44]. There have been a lot of existing methods, with implicitly or explicitly multi-view constraints, showing promising results for this task [38, 55, 50, 29, 2]. Different from these approaches, we aim to synthesize multiple scene properties including RGB for novel views.

There is another group of methods aiming to render multiple annotations for novel views in a *first-reconstruction-then-render* manner [12, 17, 22, 11]. Concretely, they first collect or build the 3D scene representation (*e.g.* mesh or point cloud) and then render multiple scene properties with 3D-to-2D projection. Different from these works, we build *implicit* 3D scene representation with a NeRF-style model based on 2D data, which is more computationally efficient. Moreover, our implicit representation enables the possibility to further model task relationships while they cannot.

Neural Radiance Fields is originally designed for synthesizing novel-view images with ray tracing and volume rendering technology [27]. Follow-up works [3, 30, 9, 36, 16, 23, 49, 31, 33, 25, 34, 13, 43, 51] further improve the image quality, optimization, and compositionality. Besides these works, several approaches [56, 6, 20, 42], namely conditional NeRFs, encode the scene information to enable the conditional generalization to novel scenes, which are more satisfied with our setting. Our MuvieNeRF takes the encoders from these conditional NeRFs as backbones.

Some works also have paid their attention to synthesizing other properties of scenes [32, 52, 47, 61, 60, 10]. Among them, Semantic-NeRF [61] extends NeRF from synthesizing RGB images to additionally synthesizing semantic labels. SS-NeRF [60] further generalizes the NeRF architecture to simultaneously render RGB and different scene properties with a shared scene representation. [37] proposes a panoptic 3D volumetric representation for the joint synthesis of RGB images 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 works improve multi-task learning performance by focusing on better optimization strategies [7, 8, 18, 19, 24, 1, 14] and exploring more efficient multi-task architectures [21, 41, 48, 4].

B. Preliminary: Conditional Neural Radiance Fields and Volume Rendering

Neural radiance fields (NeRF) [27] proposes a powerful solution of implicit scene representation, and is widely used in novel view image synthesis. Given the 3D position of the point $\mathbf{q} = (x, y, z)$ in the scene and 2D viewing direction $\mathbf{d} = (\theta, \phi)$, NeRF learns 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 NeRF, **Conditional NeRFs** [56, 20, 6, 42] learn 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, 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 will return:

$$f_{\text{scene}} = F_{\text{enc}}(\{\mathbf{x}_i, \mathbf{P}_i\}_{i=1}^{V}, \mathbf{q}). \tag{4}$$

We have $f_{\text{scene}} \in \mathbb{R}^{V \times c}$, 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, NeRF renders 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. NeRF randomly samples 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:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{m=1}^{M} \hat{T}(t_m)\alpha(\delta_m \sigma(t_m))\mathbf{c}(t_m), \tag{5}$$

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

$$\hat{T}(t_m) = \exp\left(-\sum_{j=1}^{m-1} \delta_j \sigma(t_j)\right) \tag{6}$$

denotes the accumulated transmittance.

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

$$\hat{\mathbf{Y}}^j(\mathbf{r}) = \sum_{m=1}^{M} \hat{T}(t_m)\alpha(\delta_m \sigma(t_m))\mathbf{y}^j(t_m). \tag{7}$$

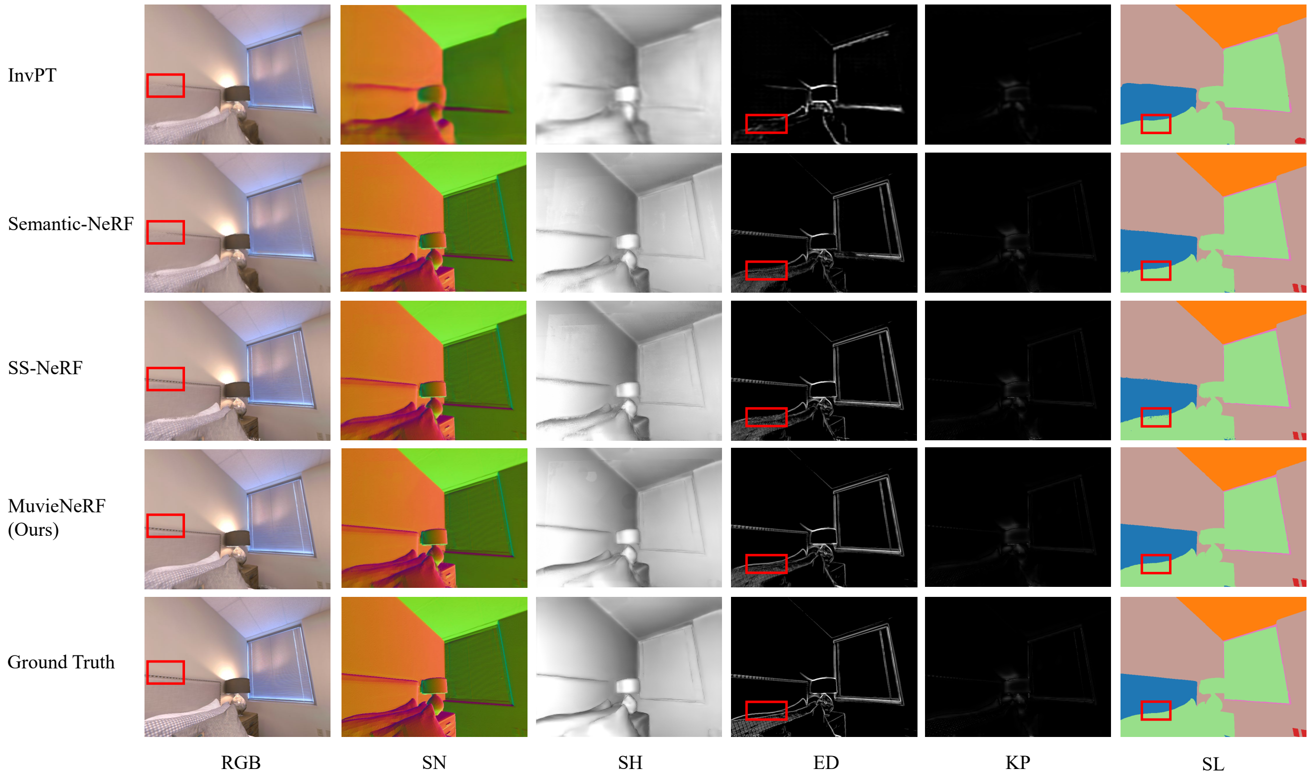

Figure A. Additional qualitative results on one testing scene in the Replica dataset. Our proposed MuvieNeRF outperforms other methods with more accurate predictions and sharper boundaries, which demonstrates the effectiveness of the multi-task and cross-view information modeled by the CTA and CVA modules. **Zoom in to better see the comparison.**

C. More Visualizations

Full qualitative comparisons for all the compared methods in the Replica and SceneNet RGB-D datasets are shown in Figure A-B and Figure C, respectively. Our MuvieNeRF outperforms other methods with clearer and more accurate contours of the objects in scenes. This is because MuvieNeRF utilizes the CTA and CVA modules to better take advantage of the shared knowledge across different downstream tasks and the cross-view information.

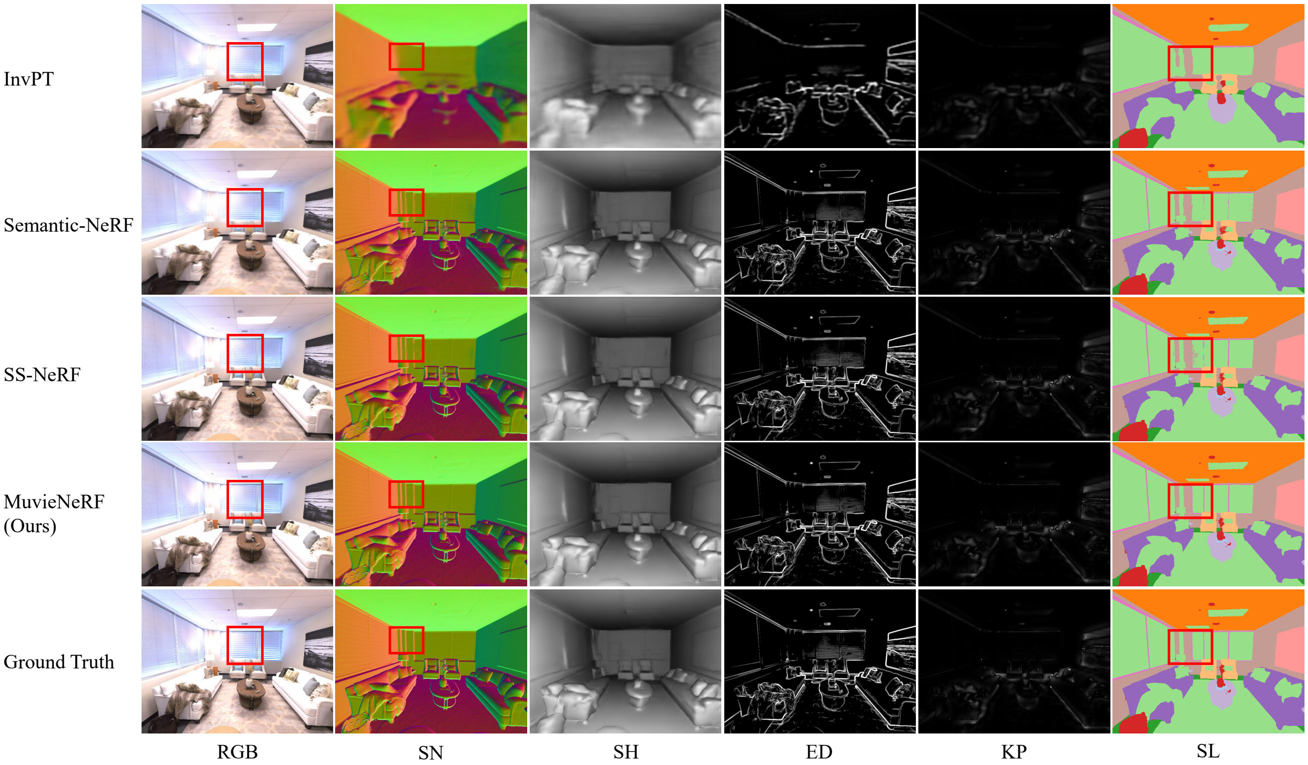

Figure B. Full results on all scene properties and compared methods of the testing scene shown in Fig. 3. Our proposed MuvieNeRF outperforms other methods with more accurate predictions and sharper boundaries, which demonstrates the effectiveness of the multi-task and cross-view information modeled by the CTA and CVA modules. **Zoom in to better see the comparison.**

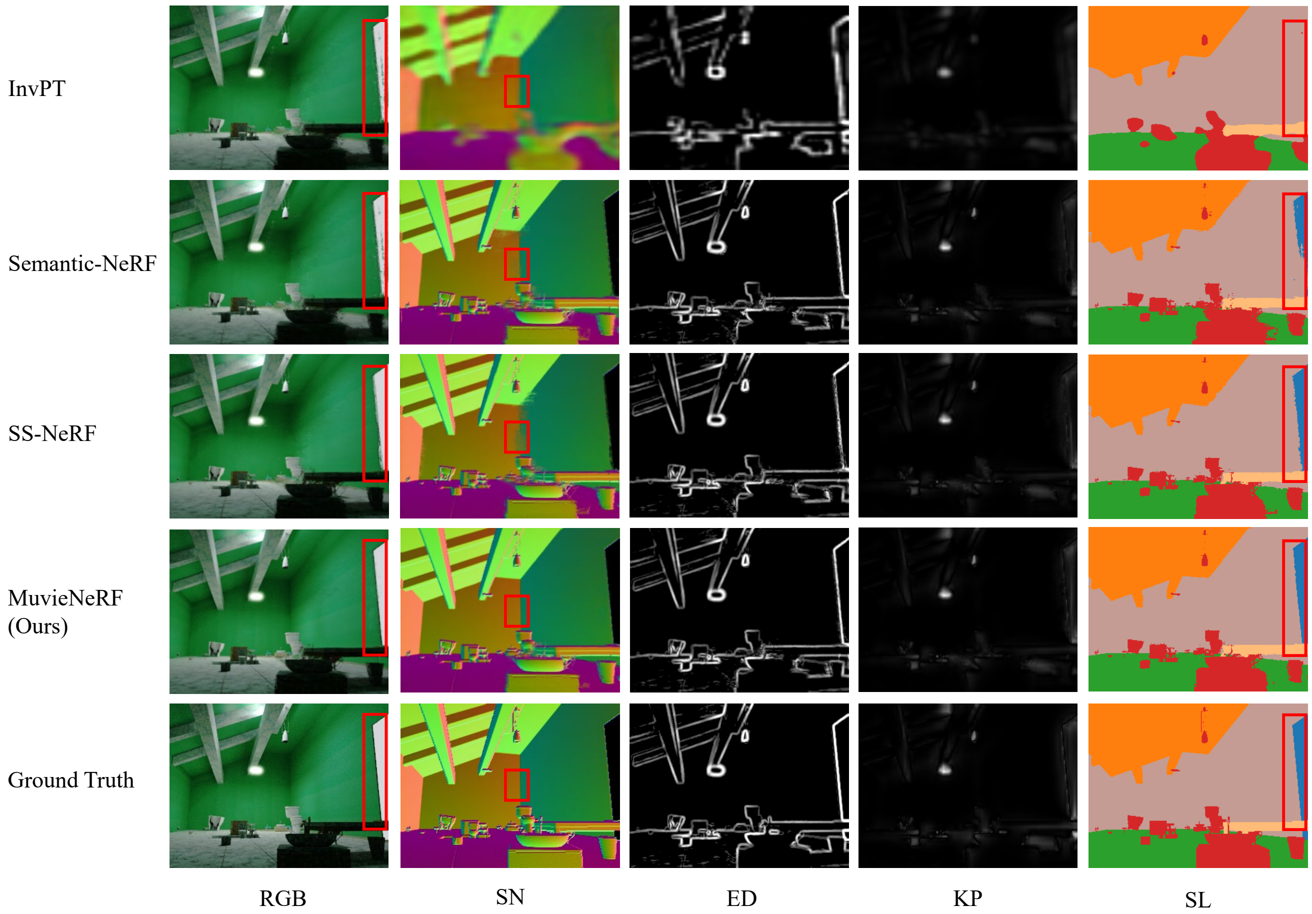

Figure C. Additional qualitative results on one testing scene in the SceneNet RGB-D dataset. Our proposed MuvieNeRF outperforms other methods, indicating that our model benefits from the multi-task and cross-view information with the designed CTA and CVA modules. The black regions in the surface normal visualizations are due to the missing depth values in those regions. **Zoom in to better see the comparison.**