NeRF based SLAM for Datasets with Illumination Changes

Andreya Ware, Che Chen, Swetha Sakunthala Subbiah, Ved Abhyankar, Tiancheng Zhang

***Abstract*— This project aims to improve NeRF based SLAM methods to make them robust against illumination changes. Datasets are prone to global and local illumination variations but these methods prove to be unable to reproduce accurate 3D reconstructed scenes when these variations are present. Global changes can occur due to change in environmental conditions over a large time span as well as poor camera settings while local illumination changes can occur due to sources of light such as lamps or flares. We incorporated affine transformations in Vox-Fusion [1] SLAM method to deal with global changes. Local changes are treated as outliers in the mapping process. This method was tested Replica dataset with illumination changes demonstrating noticeable improvement over previous methods.**

I. INTRODUCTION

In this work we extend a NeRF based SLAM method, Vox-Fusion [1], to make it more robust to illumination changes. The VOX-Fusion report explored the fusion of the volume rendering problem and dense SLAM. Volume rendering of a scene can be accomplished from input 2D images using the neural radiance field technique presented in the original NeRF paper [2] or similar extensions of the method. Dense Visual SLAM is a phenomena in the field of computer vision which is most commonly applied in Virtual Reality, Autonomous Driving and Robotics. This method involves simultaneous localization and mapping of a collection of RGB-D data, which details depth and color factors of a scene. Illumination changes can be local, caused by a blinking light on desk for example, or global, caused by changes in time of day or weather.

For our contribution to this work we propose two changes, namely adding an affine transformation to the color space and removal of outliers (based on pixel loss) to improve the mapping robustness of the method against both global and local illumination changes respectively. Results are presented in both qualitative and quantitative form that show our method improves the output 3D reconstructed scene.

II. LITERATURE REVIEW

Relevant work includes those involving methods for novel view synthesis. Most relevant to our implementation is NeRF[2] which develops neural radiance fields for scene representations and a subsequent work building upon NeRF includes Urban Radiance Fields [6], Block NeRF [7] and VOX-Fusion [1]. Each of these works contribute a specific improvement to NeRF based on their application goals.

A. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[2]

Mildenhall et. al present a method to generate novel views of complex scenes using a sparse set of input view. In this method, the underlying continuous space is represented in 5D as

$$(x, y, z, \theta, \phi)$$

i.e the 3 position coordinates and 2 view-direction coordinates, based on camera rays $\mathbf{r}$. This space is then sampled using stratified sampling. In stratified sampling the range of the ray (between far and near coordinates is divided into N bins and then one sample is drawn from each bin uniformly at random.

$$t_i \sim \mathcal{U}[t_n + (i-1) * (t_f - t_n)/N, t_n + i * (t_f - t_n)/N] \quad (1)$$

Further, positional encoding is used to lift the 5D input into higher dimensional space.

Multilayer perceptrons i.e. fully connected neural networks without convolution, are used to map the encoded input to obtain the color and volume density at that position. The positional encoding helps the network to better fit high frequency variation data One MLP is used to obtain the volume density (based only on position) and another MLP is used to obtain color using viewing direction coordinates and the feature vector obtained from previous MLP.

Hierarchical volume sampling is used during volume rendering to compose images using MLP output. This is done to efficiently sample the high frequency representation obtained from the MLPs. 2 networks are optimised, a course network and a fine network. The fine network performs an informed sampling based on output of the course network by defining a PDF by normalising the weights of the MLP color output. The loss function used is the total sum of squares of error of both, the course and fine network.

$$\mathcal{L} = \Sigma_{\mathbf{r} \in \mathcal{R}} [\|\hat{\mathbf{C}}_c(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|^2 + \|\hat{\mathbf{C}}_f(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|^2] \quad (2)$$

where $\hat{\mathbf{C}}_c(\mathbf{r})$ is the prediction by coarse network, $\hat{\mathbf{C}}_f(\mathbf{r})$ is the prediction by the fine network and $\mathbf{C}(\mathbf{r})$ is the ground truth.

B. VOX-FUSION: Dense Tracking and Mapping with Voxel-based Neural Implicit Representation[1]

Vox-Fusion is a tracking and mapping system that combines neural implicit representations with traditional volume rendering techniques and dense SLAM methods. The camera is propagated around the scene and as data is collected, the output 3D reconstruction map of the scene is incrementally updated making the voxels dynamic, able to be added as new parts of the map are observed. This project represents 3D scenes as a collection of N-dimensional sparse voxel embeddings. These embeddings are shared by neighboring embeddings allowing for better representation of border

artifacts. During volume rendering the scene is sampled using a efficient point sampling method which only considers the rays which intersect voxels, enforcing a limit on the maximum sampling distance. The voxels in the scene and the ray which intersects them are are fed through a implicit signed distance function decoder F_θ with optimizable parameter θ. This outputs color and signed distance for each voxel. The volume rendering process is utilized in both mapping and tracking.

The camera pose is represented in 6 DoF as it moves through the space. As seen in the Vox-Fusion framework image below (Fig 1), the model begins with the input RGB-D stream and frames are created. After stratified ray sampling of the space, the column rendering process mentioned above is utilized. After volumes are rendered the pose is then updated and this information is backpropagated through the model. Rather than mapping the space on every iteration, the map is updated through a keyframe selection process. Keyframes are selected when a given key frame candidate has more observable voxels then the previous keyframe. After volume rendering the keyframe in the final mapping step, the camera pose and map are optimized jointly.

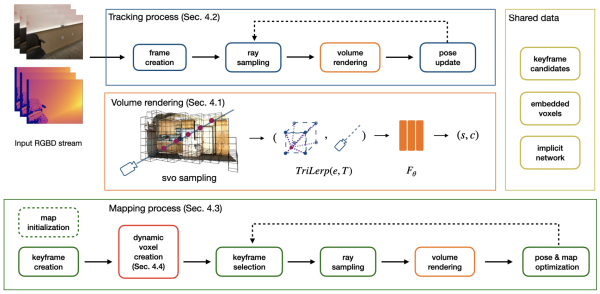


Fig. 1: Vox-Fusion Framework

For surface renderings, volume rendering similar to that in NeRF is used to render color and depth. However the formulas to calculate the color and depth along any camera rays are modified to be applicable for feature embeddings (obtained from MLP) instead of global coordinates. The color is calculated as a weighted average of the color output (c_j) of the MLP while the weights are calculated as the weighted average of the depth values (d_j) along any camera ray. The weights (w_j) are a sigmoid function of the ratio of SDF values (s_j) predicted by the MLP and a truncation factor *tr*.

Four different loss functions are used to train the network. 2 cover the RGB and depth losses respectively as the average sum of absolute differences between the ground truth and predicted values:

$$\mathcal{L}_{RGB} = \frac{1}{|P|} \cdot \Sigma_{i=0}^{|P|} \|C_i - C_i^{gt}\| \tag{3}$$

$$\mathcal{L}_{depth} = \frac{1}{|P|} \cdot \Sigma_{i=0}^{|P|} \|D_i - D_i^{gt}\| \tag{4}$$

Here P is the set of sampled points. Then a free-space loss is defined so that the MLP learns the truncation factor of any point that lies between the camera center and the truncation region of the surface. Finally a SDF-loss is used so that the MLP learns accurate surface representations within the surface truncation area. These losses are defined as follows:

$$L_{fs} = \frac{1}{|P|} \cdot \Sigma_{p \in P} \left(\frac{1}{S_p^{fs}} \Sigma_{s \in S_p^{fs}} (D_s - tr)^2\right) \tag{5}$$

$$L_{sdf} = \frac{1}{|P|} \cdot \Sigma_{p \in P} \left(\frac{1}{S_p^{tr}} \Sigma_{s \in S_p^{tr}} (D_s - D_s^{gt})^2\right) \tag{6}$$

C. Urban Radiance Fields (URF)

[6] This work presents a system that extends on the NeRF framework by implementing neural implicit networks to map urban areas. Specifically, the authors for URF explore usage of data from mobile scanning platforms for reconstructing street level 3D scenes. In their input stream, in addition to RGB signals generated from camera panoramas of the scene, they also incorporate LiDAR data. This LiDAR data is modelled with the assumption that the LiDAR actively emits rays and surfaces it detects are stationary, giving each ray a origin and termination. They optimize their method using a photometric-based loss function.

An important contribution of the Urban Radiance Field method for our project is their work to improve the exposure compensation of NeRF based SLAM. In order to better consider the auto white balance and auto exposure photos acquired by mapping systems, they propose an affine transformation for mapping of the radiance predicted by the shared network. This transformation is a 3x3 matrix decoded on a per image basis based on a per-image latent code. This gives white balance and exposure variations a more restrictive function and thus leads to less artifacts when optimizing the scene radiance parameter θ and exposure mappings β. We replicate this transformation in our work to improve Vox-Fusion.

D. Block-NERF: Scalable Large Scene Neural View Synthesis

[7] This is another real world application of NeRF that aims to make a robust version of it to reconstruct large scale scenes. The authors of this paper utilize smaller individual "Block-NeRFs" to make the original NeRF more adaptable to larger scenes. They also made improvements to each Block-NeRF's appearance and tracking capabilities during the training phases. Finally they considered methods to simplify the reconstruction when regenerating models from each Block-NeRF. Each Block is also initialized to have a 50% overlap which allows for easy appearance alignment during reconstruction. They utilize a target view radius to determine which Block-NeRFs are needed for scene reconstruction. This filtering typically reduces the view synthesis down to only one to three Block-NeRFs. Their model was able to produce the largest 3D reconstruction to date, using their modified NeRF SLAM method and a collected dataset in San Francisco sized at approximately 960m x 570m.

III. Methodology

A. Overall Structure

We based the project on Vox-Fusion. Vox-Fusion stores embeddings within sparse octrees. The data structure allows for efficient memory usage as well as dynamic map expansion. Trilinear interpolations are performed during sampling to ensure a continuous space representation. A decoder is used to decode embeddings stored in the voxels into the signed distance field (SDF) and the color output.

B. Dataset Generation

Datasets with global as well as local illumination changes were generated using the Replica dataset [8] as baseline. Room 0 and Office 4 datasets were modified to include a brightness factor changing as a sine wave for global illumination change (Fig 2), and the local illumination change was introduced as a circular mask of random brightness factor in a random location on the scene (Fig 3).

Fig. 2: Global Illumination Change

Fig. 3: Local Illumination Change

C. Global Illumination

To handle possible global illumination changes that can happen because of camera settings change or weather conditions, we adopt the method described in [6] into our Vox-Fusion [1] algorithm, which utilizes a single MLP $\mathbf{\Gamma}$ and a per-image embedding β_i to produce an affine transformation in color space. The MLP and per-image embeddings are jointly optimized with the scene representation.

$$\Gamma(\beta_i) : \mathbb{R}^B \rightarrow \mathbb{R}^{3\times 3} \tag{7}$$

and the produced matrix is applied on the RGB value of the sample to perform an affine transformation in color space.

$$\hat{\mathbf{C}}_i = \Gamma(\beta_i)\mathbf{C}_i \tag{8}$$

After transformation, the transformed color is used to calculate the modified rgb loss term,

$$\mathbb{L}_{rgb} = \frac{1}{P}\sum_{i=0}^{P} \|\hat{\mathbf{C}}_i - \mathbf{C}_i^{gt}\| \tag{9}$$

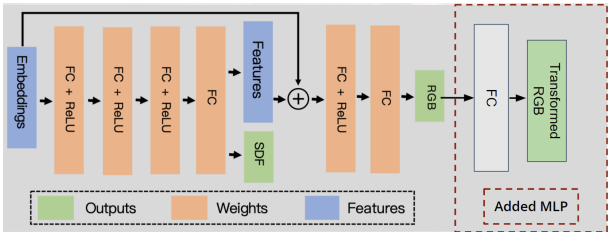


Fig. 4: MLP Structure for Our Improved Vox-Fusion Method

However, during the experiment, we found that it's often hard for the MLP output to converge to a sensible value. Based on the prior knowledge that the output transformation matrix should be approximately an identity matrix, we modified the method to model the residual part of the color transformation (Fig 4).

$$\hat{\mathbf{C}}_i = [I + \Gamma(\beta_i)]\mathbf{C}_i \tag{10}$$

The new representation turns out to have a better performance than the one proposed in the original paper.

D. Local Illumination

Due to the transient and local characteristics of the local illumination change, we can just treat them as outliers during the mapping process. For simplicity we ignore some pixels with a very large difference to ground truth (10x median) when calculating loss.

IV. Evaluation and Results

A. Criterion and Dataset

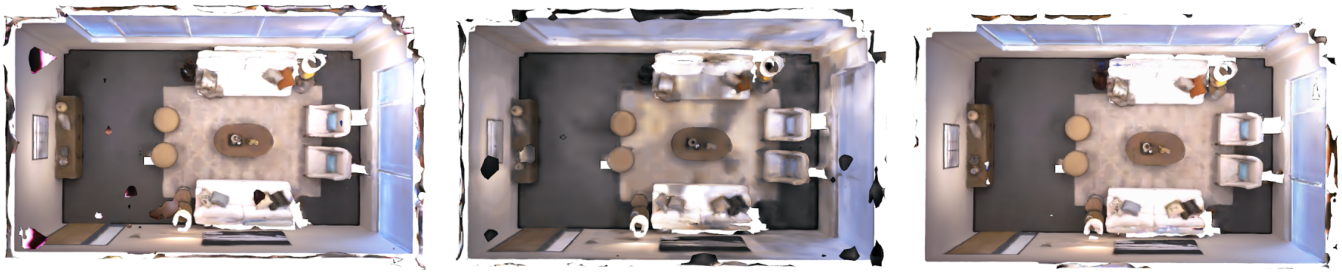
In this section, we present the quantitative and qualitative evaluation of our algorithm. Our goal is to demonstrate the effectiveness of our approach in generating high-quality 3D map reconstruction results that are both accurate and complete when the illumination changes are applied, compared to the original algorithm. We employ four quantitative metrics for the mapping reconstruction: Chamfer Distance (CD), F1-score, Completeness, and Complete Ratio. Also, for the tracking error, RMSE is applied. Additionally, we provide a visual comparison between the original and the new dataset for qualitative assessment. The dataset that is used for evaluation is the Replica: Room0 and Office4 [8] as well as its augmented illumination change dataset.

B. Qualitative Result

1) Global Illumination Change: The qualitative result mainly focuses on the visual quality comparison between the original Vox-fusion model and our robust model on both room0 and the office4 datasets. (Fig 5 and 6)

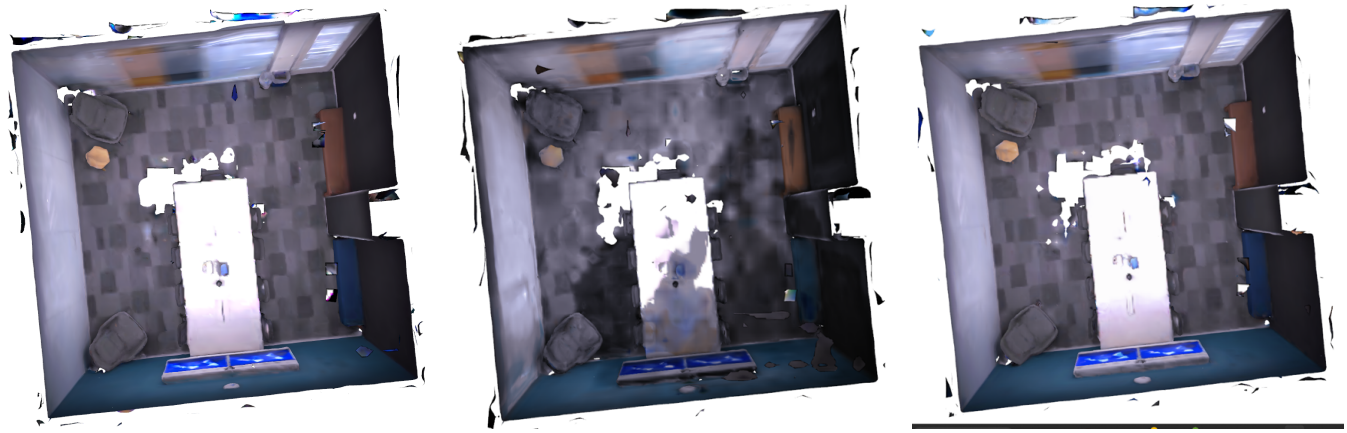
Compared to the non-illumination change result, the original model reconstruction of the mapping renders abnormal voxel rgb rendering and unnatural split between the voxels. The total exposure setting of the original model is not similar to what no-illumination status looks like.

Instead our robust model recovers the illumination status of the reconstruction scene and resets the exposure setting to the one that is compatible with the original non-illumination mapping reconstruction result.



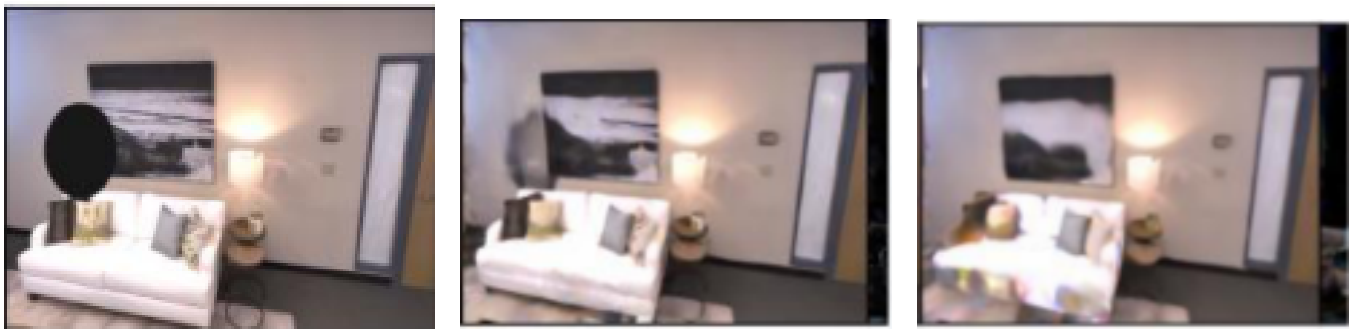
(a) No illumination change (b) Original Model for illumination change (c) Robust Model for illumination change

Fig. 5: Room0 original model and global model comparison



(a) No illumination change (b) Original Model for illumination change (c) Robust Model for illumination change

Fig. 6: Office4 original model and global model comparison



(a) Ground truth (b) Original Model RGB (c) Robust Model RGB

Fig. 7: Office4 original model and global model comparison

2) Local Illumination Change: With local illumination (the black mask) on the figure, if the original model is applied to the dataset, the black spot persists in the figure (Fig. 7), while using the robust model, the black mask will be erased completely. However, it results in poor rendering due to loss of details and color. Therefore, the robust model can result in local illumination erasure but does not offer the best performance.

C. Quantitative Result

This section shows the result of the robustness method applied to the Replica room0 and office4 dataset, and evaluation is carried out on the reconstruction completeness and similarity to the ground truth, the result is shown in Table I.

		room0	**office4**
Orig	**CD**	0.056	0.058
	F1	0.535	0.536
	Comp.	0.029	0.027
	Comp. Ratio	0.929	0.893
Robust	**CD**	0.046	0.055
	F1	0.534	0.541
	Comp.	0.027	0.026
	Comp. Ratio	0.926	0.893

TABLE I: **Reconstruction Mapping Evaluation**

According to the result shown in the Table, for both room0 and office4, the completeness of the rendering is almost the same while the chamfer distance and F1-score of the robust model outweigh the original model in the global illumination change case. When it comes to the tracking result, the RMSE for the tracking distance is shown in Table II.

		room0	**office4**
Orig	**Avg. Err**	0.83	0.68
	RMSE	0.90	0.76
Robust	**Avg. Err**	0.59	0.69
	RMSE	0.63	0.77

TABLE II: **Tracking Evaluation**

According to the result shown in Table II, both room0 and office4 renders the compatible tracking result in the global illumination case, and room0 outperforms the original model in the tracking result. In summary, the robust model outperforms

V. Conclusion and Future Work

Our model adds robustness against global illumination changes to the Replica dataset. It gives better performance in the mapping task and comparable performance in tracking to the original Vox-Fusion methods. Both the quantitative and qualitative results presented show this. Our proposed method currently leads to data loss when dealing with local illumination changes. More efficient means of handling local changes need to be explored. These methods can also be extended to real world datasets.

References

[1] X. Yang, H. Li, H. Zhai, Y. Ming, Y. Liu, and G. Zhang, "Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation," in 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, oct 2022. [Online]. Available: https://doi.org/10.1109%2Fismar55827.2022.00066

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," 2020. [Online]. Available: https://arxiv.org/abs/2003.08934

[3] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," 2021. [Online]. Available: https://arxiv.org/abs/2112.12130

[4] Z. Zhu, S. Peng, V. Larsson, Z. Cui, M. R. Oswald, A. Geiger, and M. Pollefeys, "Nicer-slam: Neural implicit scene encoding for rgb slam," 2023. [Online]. Available: https://arxiv.org/abs/2302.03594

[5] S. Park, T. Schops, and M. Pollefeys, "Illumination change robustness in direct visual slam," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 4523–4530

[6] K. Rematas et al., "Urban Radiance Fields," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 12922-12932, doi: 10.1109/CVPR52688.2022.01259.

[7] Tancik et al., "Block-NeRF: Scalable Large Scene Neural View Synthesis," (2022). Available: https://doi.org/10.48550/arXiv.2202.05263

[8] J. Straub et al., "The Replica Dataset: A Digital Replica of Indoor Spaces," (2019). Available: https://doi.org/10.48550/arXiv.1906.05797