Scene Representations for Generalizable Novel View Synthesis

by

Youssef Abdelkareem

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Novel view synthesis involves generating novel views of a scene when seen from different viewpoints. It offers numerous applications in computer vision domains such as telepresence, virtual reality, re-cinematography, etc. Recent literature work in the field successfully achieved remarkable photo-realistic synthesis results, however, they require per-scene optimization settings and densely sampled input views which is not easily attainable in practice. Developing lightweight generalizable view synthesis systems with sparse input views would make them more applicable for direct consumer usage. Novel view synthesis poses several difficulties, such as addressing obscured regions, broadening the range of viewing directions, sidestepping suboptimal per-scene optimization configurations, and depicting intricate multi-human scenarios. Tackling those challenges depends on the representation used to model the 3D structure of the scenes. Explicit 3D representations utilize different techniques to explicitly model the scene structure. One example is multi-plane images (MPIs) that segment the scene into a set of parallel planes giving it the ability to effectively handle occlusions. Implicit neural representations, such as Neural Radiance Fields, enable the encapsulation of 3D scene structure within the weights of a neural network, thereby facilitating a 360-degree range of viewing directions and photorealistic synthesis results. However, a promising avenue of research would be to explore the combination of implicit and explicit representations in order to harness their advantages and address more challenging scenarios.

In this thesis, we focus on layered scene representations that blend explicit and implicit properties at either the pixel or object level in a generalizable manner. One example of the pixel-level representation is Multi-plane Neural Radiance Fields (MINE), which combines multi-plane images with Neural Radiance Fields for efficient and generalizable novel view synthesis. However, current literature only examines single-view settings for MINE, which limits its viewing range. Our work conducts a thorough technical analysis of the capabilities of single-view MINE and proposes a new Multi-plane NeRF architecture that accepts multiple views to improve synthesis results and expand the viewing range. Additionally, existing methods for handling complex multi-human scenes rely on per-scene optimization settings, making them impractical for real-world use. To address this, we propose a novel object-level layered scene representation named GenLayNeRF that can generate novel views of scenes with close human interactions while generalizing to new human subjects and poses. Furthermore, there is a scarcity of open-source datasets for multi-human view synthesis. To fill this gap, we create two new datasets, ZJU-MultiHuman and DeepMultiSyn, which contain scenes with close human interactions. These datasets are used to evaluate our performance against generalizable and per-scene baselines. The

results indicate that our proposed approach outperforms generalizable and non-human per-scene NeRF methods while performing at par with layered per-scene methods without test time optimization.

Acknowledgements

I would like to thank everyone who made this thesis possible. Specifically, I would like to express my gratitude to my family for their constant support. I also want to thank Prof. Karray for his guidance and professional support. Lastly, I would like to give a special thanks to Dr. Shady Shehata for his technical help throughout all the projects done in this thesis.

Dedication

This is dedicated to my sisters.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CNN Convolutional Neural Network 7, 8, 10, 12

GenLayNeRF Generalizable Layered Scene Representations for Multi-human Novel View Synthesis 3, 4, 14, 19, 26, 27

LDI Layered Depth Images 5, 6, 8, 12

MINE Multi-plane Neural Radiance Fields 3, 4, 11, 13–17, 26, 27, 36

MLP Multi-layer Perceptron 9

MPI Multi-plane Images 3, 5–8, 12, 14

MV-MINE Multi-view Multi-plane Neural Radiance Fields 4, 14, 16, 26, 27, 36, 50

NeRF Neural Radiance Fields x, 3, 4, 9–11, 14–16, 21, 22, 36, 47

SMPL Skinned Multi-Person Linear Model 20–22, 47, 51

Chapter 1

Introduction

1.1 Preface

Photography has evolved in recent years and most consumer phones contain high-quality cameras allowing any user to be a photographer. Every second, a vast amount of photos are recorded. When confronted with an intriguing situation, a user would typically document it by utilizing a camera to snap as many photographs from as many different perspectives as possible. The more perspectives and photographs gathered, the greater the story and experience. Novel view synthesis algorithms strive to improve a user's experience even further by allowing him or her to reproduce the imaging by synthesizing a novel picture from a different viewpoint at the scene. Novel view synthesis offers a wide range of applications, including re-cinematography [25], producing material for virtual reality [7], and synthesizing scenes to improve computer vision algorithms [63]. It is also used to enable very high frame-rate movies in multi-lens camera array systems [74].

1.2 Motivations

The research community has developed methods that achieve photo-realistic view synthesis results on a variety of synthetic and real-world scenes [42, 45, 47, 55, 1]. However, most of the methods are constrained to per-scene optimization settings [42, 45]. In other words, models are trained once per scene and need to be re-trained from scratch for each novel scene at inference time. In real-world applications, it would be highly ineffective to retrain models for every new scene. In addition, some of the methods require a densely sampled set

of input views to generate plausible novel views [10, 15, 23]. This requires a combination of synchronized camera rig systems which is not easily achievable or cost-effective in practice. In recent years, few methods started presenting generalizable view synthesis approaches that require sparse input view to increase the applicability to real-world scenarios [72, 80, 34]. We believe that creating light view synthesis systems that generalize to unseen scenarios while requiring a small number of views will highly impact many real-world applications. It can enable consumers to participate in immersive reality experiences using their handheld consumer phone cameras without requiring expensive setups. As more multi-view photos and videos are captured, the generalization capabilities of the view synthesis systems will eventually grow and elevate the realistic properties of the virtual experiences.

1.3 Problems And Challenges

Our target in this thesis is to explore novel view synthesis approaches that generalize to new scenes at test time while requiring a small number of input views. The main challenge to be tackled is how to effectively model the 3D structure of the scene. The difficulty lies in the ability to infer accurate geometric structures and texture details from sparse 2D input images. This is particularly evident in complex scenes with occluded areas, lighting effects, non-smooth surfaces, etc. View synthesis systems need to generate 3D scene representations that can model occlusions and predict the content of the hidden regions, while also taking into consideration the effect of lighting on the object colors when seen from novel views. The synthesis task becomes even more challenging when the target viewpoints are considered significantly far away from the input views which means researchers need to consider the viewing range capability when creating their systems. Another main challenge is handling scenes with human subjects. Such scenes are characterized by containing non-static subjects with complex deformations for the different body parts, fine-grained texture and body geometry details, and self-occlusions that occur during the motion of the subject. The difficulty of the problem increases when multiple human subjects are interacting together which introduces additional inter-human occlusions. In general, it remains a challenge to develop systems that are also both memory and computation efficient to enable effective real-world deployment.

1.4 Proposed Solutions

Different 3D scene representations were proposed in the literature to tackle the challenges mentioned for novel view synthesis. Explicit 3D approaches aim to portray the scene structure using volumetric representations [57, 1, 65]. Those types of representations help in explicitly modeling occluded areas and lighting effects making them optimal for scenes with many complex overlapping objects. Multi-plane Images (MPI) are an example of volumetric representations that divide the scene into a set of parallel planes with pixels containing color and transparency values. On the other hand, implicit neural representations [80, 47] such as Neural Radiance Fields (NeRF) [42], tend to encapsulate the 3D scene structure within neural network weights enabling a 360° viewing direction range and photo-realistic synthesis results. A promising direction would be to study the combination of implicit and explicit representations to make use of their advantages and handle challenging scenarios.

In this thesis, we focus on layered scene representations that combine the explicit and implicit properties either at the pixel level or object level in a generalizable manner. Multi-plane neural radiance fields (MINE) [34] was proposed as a pixel-level layered representation that combines multi-plane images with neural radiance fields for generalizable and efficient novel view synthesis. However, the current literature work only studies single-view settings for MINE which limits the viewing range capability of the method. In our work, we carry out an in-depth technical analysis of the capabilities of single-view MINE in terms of performance, generalization, and efficiency. We then propose a novel multi-plane NeRF architecture that accepts arbitrary multi-view input to enhance the synthesis results and the viewing range. Regarding complex multi-human scenes, the existing work [55] only handles per-scene optimization settings making them inefficient to use in practice. For that reason, we additionally propose a novel object-level layered scene representation, GenLayNeRF, that can generate novel views of scenes with close inter-human interactions while generalizing to unseen human subjects and poses at inference time. There is a lack of open-source multi-human view synthesis datasets. For that reason, we also create two new datasets, ZJU-MultiHuman and DeepMultiSyn, which contain scenes with close human interactions. We used the two datasets to compare our performance against generalizable and per-scene baselines.

1.5 Contributions

The main contributions of this thesis are summarized as follows:

- We provide in-depth technical analysis on the performance, generalization, and efficiency of single-view multi-plane neural radiance fields for novel view synthesis.
- We propose, MV-MINE, an architecture merging generalizable neural radiance fields and multi-plane images with a multi-view input setting.
- We propose an attention-based feature fusion module for effectively aggregating multi-view input for multi-plane neural radiance fields.
- We propose a generalizable object-level layered scene representation, GenLayNeRF, with attention-aware feature fusion for the free-viewpoint rendering of real-world multi-human scenes from sparse input views while operating on novel human subjects and poses.
- We construct multi-human view synthesis datasets that are used for evaluation. The datasets can be considered as a benchmark for any comparison between the relevant multi-human methods.
- Our approach, GenLayNeRF, surpasses state-of-the-art generalizable and non-human per-scene NeRF methods while performing at par with the multi-human per-scene methods without requiring long per-scene training procedures.

1.6 Organization Of Thesis

The rest of the thesis will be organized as follows. Chapter 2 will present the background and literature review regarding the relevant novel view synthesis approaches. In Chapter 3, we present the proposed methodology of the technical analysis of single-view MINE along with the architecture of proposed architecture MV-MINE and GenLayNeRF. Chapter 4 will discuss the experimental setup and results for assessing the proposed approaches and analysis points, which is followed by a discussion in Chapter **??**. Lastly, in Chapter 5, we present the summary of this thesis and discuss its related future research directions.

Chapter 2

Background And Literature Review

In this chapter, we will discuss the classical (Section 2.1) and learning-based approaches (Section 2.2) for view synthesis, based on explicit and implicit utilization of scene geometry. Explicit 3D approaches, presented in Section 2.2.1, model the camera frustum directly through different representations to better model occluded areas. Multi-plane explicit representations (MPI) project parallel RGB-α planes that can be warped and used to generate new views, but suffer from incomplete 3D scene representation due to depth discretization. Layered Depth Images (LDI) offer a more memory and space-efficient approach by allowing each pixel to have arbitrary layers at different depths. Implicit 3D representations, discussed in Section 2.2.2, model the 3D scene structure within neural network weights and can be per-scene optimization methods or generalizable approaches that handle unseen scenes at inference time. We also go over the human-based approaches that can handle complex subjects with deformations and self-occlusions. Recent approaches that combine implicit and explicit representations on a pixel or object level are mentioned in Section 2.2.3. Finally, we will also explore different attention mechanisms available in computer vision tasks in Section 2.3.

2.1 Classical View Synthesis Approaches

It can be helpful to categorize traditional novel view synthesis algorithms based on how much they utilize explicit scene geometry [56]. Initial work for novel view synthesis utilized the concept of light fields [33, 21] which require a dense sample of input frames that are organized on a regular grid, and novel views are sampled by slicing the sampled light field from existing views. Such methods lie at one extreme of the spectrum by completely

not exploiting the scene geometry. Other approaches tend to explicitly predict a global mesh representation of the scene which is then reprojected and blended with input views to generate novel views [12]. In between the two extremes, several more recent methods [5, 27] aim to predict the local geometry of each input view which are then projected and blended to generate the novel views, however, they still rely on interpolation and fail to render occluded areas leading to implausible synthesis results in those areas.

2.2 Learning-based View Synthesis Approaches

Recent progress has been made in utilizing end-to-end learning approaches for predicting novel views. We categorize the approaches based on the 3D scene representation which can be explicit, implicit, or a combination of both.

2.2.1 Explicit 3D Representations

Volumetric approaches aim towards learning explicit representations of the camera frustum, which opens the door for modeling occluded regions and non-Lambertian effects. The representations include 3D voxel grids [57, 77], textured meshes [1, 75], point clouds [65] , layered depth images (LDI) [68, 53] and MPI [64, 87, 41]. In this section, we focus on the layer-based approaches which are LDI and MPI.

Multi-plane Images (MPI)

MPI approaches represent the scene as a set of discretized RGB-α front-parallel planes representing the elements of the scene at different depths, as shown in Figure 2.1. One way to generate MPIs from input images would be to use a gradient-based approach to optimize its parameters with the target of re-creating the input images after projecting the predicted MPI. This requires a large number of input views or the addition of regularization terms to reach the optimal parameters and avoid overfitting making it ineffective in practice. Zhou et al. [87] utilized a feed-forward neural network to directly predict the MPI from input images. They rendered novel views of scenes through forward projecting and alpha compositing of the planes. However, the method suffers from the inability to intrinsically model the geometric visibility between input images and predicted MPI as they are implicitly learned within the network weights. This issue presents itself evidently when distant scene elements occlude each other, which requires extremely large network weights

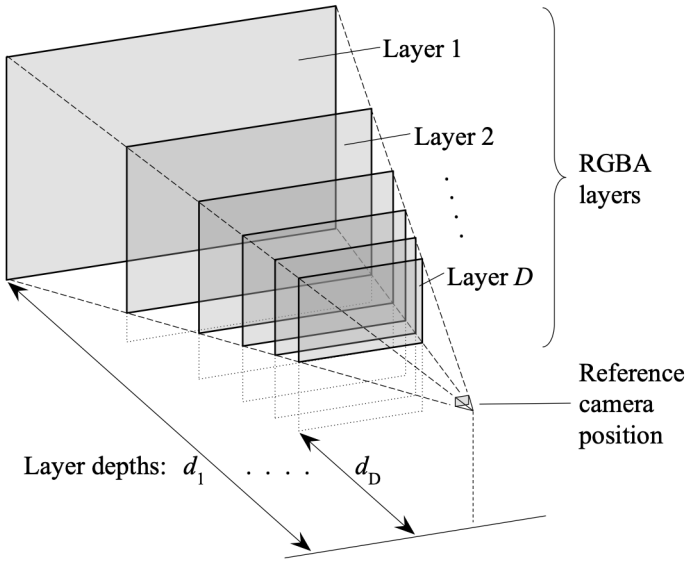

Figure 2.1: Visualization of the concept of multi-plane images [87].

to be effectively represented. Flynn et al. [16] proposed a solution by combining direct parameter optimization with network learning. They use an iterative algorithm where the current MPI is improved by computing gradients with respect to the input images and processing these gradients using a convolutional neural network (CNN) to compute an enhanced MPI. This allows for avoiding overfitting while modeling occlusions without dense network connections. Srinivasan et al. [59] enhances the MPI prediction results by allowing an increase in the number of planes at test-time using 3D CNN. They also introduce a two-step MPI prediction approach to enforce the plausibility of the textures and structure of disoccluded regions.

Most of the MPI approaches mentioned previously rely on multi-view input for novel view prediction. A recent approach [67] proves the potential of utilizing MPI in the single-view setting for high-quality view synthesis. They estimate the planes using a deep CNN and introduce a scale-invariant synthesis approach to solve the scale ambiguity problem for single-view settings. The main drawback is that the planes are predicted at discrete depths which constrains the ability to model the 3D space at any depth value continuously.

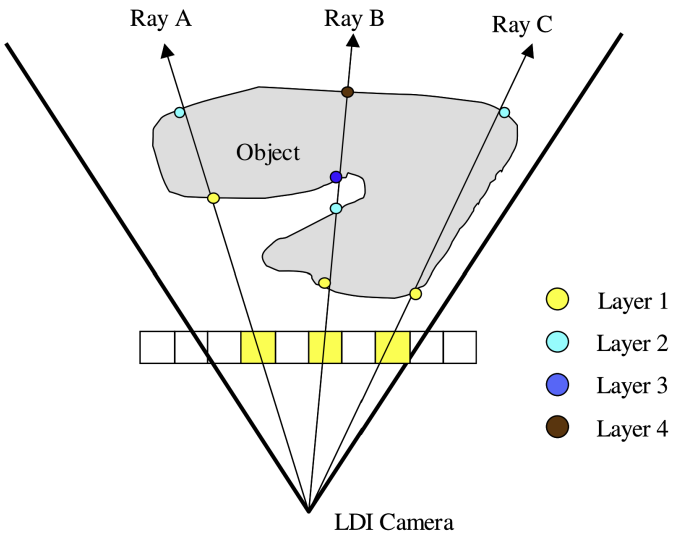

Figure 2.2: Visualization of the concept of layered depth images [79].

Layered Depth Images (LDI)

LDI [52] are sparse 3D representations of the scene modeled as a pixel lattice where each pixel location contains a set of pixels. Each pixel is represented with a color and a depth value. A visualization is shown in Figure 2.2. In smooth regions, all pixels are directly connected to their 4 neighboring pixels. One line of approach [13, 68] uses a variant of the LDI which has fixed layers (depths) for all pixels. Specifically, layers in every pixel are sorted from nearest to farthest. This leads to problems around areas with discontinuous depths due to abrupt changes in depth which lead to poor locality representation. Other approaches [25, 26] explicitly store the connectivity information within each pixel to allow for the arbitrary number of layers and they possess no connections in areas of depth discontinuities. Those representations have grabbed the attention of researchers due to their adaptability to scenes with complex depths with the help of the arbitrary number of layers per pixel and their memory and space-efficient properties compared to MPI. Quite recently, [53] proposed an LDI-based approach with explicit connectivity storage along with a novel learning-based in-painting approach that predicts the color and depth of occluded regions to synthesize texture and structures. Their algorithm is recursive in nature as it carries out local in-painting on spatial contexts with standard CNN until all depth edges are traversed. Even though the results were impressive, the algorithm's recursive nature makes it inefficient for real-world applications.

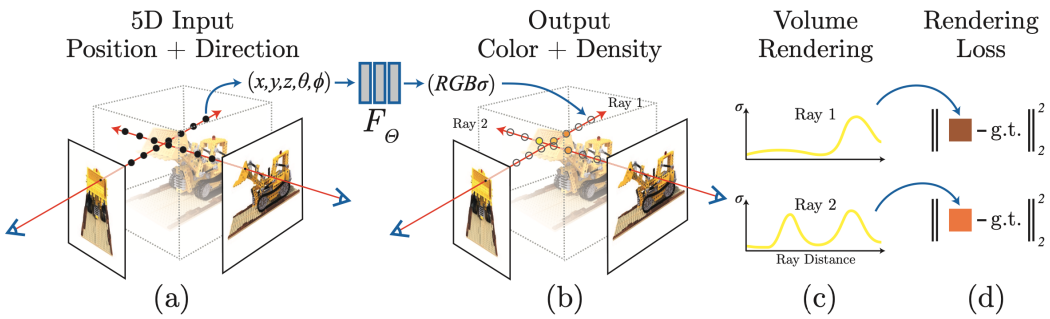

Figure 2.3: Visualization of the neural radiance fields (NeRF) architecture [42].

2.2.2 Implicit Neural Representations

Recent work has been made in implicitly modeling the geometry, structure, and texture of the 3D scene within neural network weights. We categorize the approaches into per-scene optimization methods, generalizable approaches, and human-based approaches.

Per-Scene Optimization Approaches

NeRF [42] revolutionized the concept of novel view synthesis by encapsulating the full continuous 5D radiance field of scenes inside a Multi-Layer Perceptron (MLP). This enables the representation of the whole continuous 3D space of a scene inside the MLP weights. For each pixel in the input view images, a ray is transmitted across the 3D space with respect to the camera, and 3D points are sampled. Each 3D point along with the ray viewing direction passes through an MLP to produce the radiance and density of the point. The predicted radiance fields are aggregated across all points on the ray using volumetric rendering to produce the final pixel color per ray. This operation is repeated for all pixels in the target image until the image is fully rendered.

The method achieved photo-realistic results but failed to work on highly deformable scenes with non-static subjects. Deformable NeRF methods [43, 47] modeled the dynamic subjects by training a deformation network that transforms 3D points to a canonical space before querying the MLP. However, one of the main drawbacks is being confined to per-scene optimization and lacking any generalization capabilities to novel scenes.

Generalizable Approaches

Per-scene optimization NeRF methods [45, 42, 47, 55] need to be trained from scratch on each scene with a moderately high number of source views which is often impractical due to the large time and computational costs. Generalizable NeRF methods [66, 80, 72] offered a possible solution by conditioning NeRF on image features. Specifically, the network architecture consists of a feature encoder CNN for generating feature planes from input images and a NeRF network for predicting novel views from the target viewpoints. For each 3D point, the methods extract pixel-aligned features for each input view and aggregate the multi-view feature vectors with pooling operations. pixelNeRF [80] used a simple averaging strategy, while IBRNet [72] carried out globally and locally conditioned weighted pooling operations. Those approaches have the ability to generate novel views of scenes not seen during training while only relying on sparse views as input. Utilization of image features opened the door for implicitly learning strong priors from the diverse training scenes. Despite the promising generalization capabilities, the methods suffered from blur artifacts with human subjects due to the large degree of self-occlusions and complex motions.

Human-based Approaches

Handling scenes with human subjects that have relatively complex deformations is a challenging task. NeuralBody [45] offered a solution by anchoring NeRF with a deformable human model [38] to provide a prior over the human body shape and correctly render self-occluded regions. This resulted in high-quality synthesis output for single-human scenes but was constrained to the per-scene optimization setting. Recently, NHP [32] combined the 3D human mesh with image features to accurately represent complex body dynamics and generalize to novel human subjects and poses. They carried out feature aggregation across multiple views and timesteps using cross-attention. The work in HumanNeRF [85] enhanced the quality of the results by incorporating efficient fine-tuning procedures and neural appearance blending techniques. However, the blending module only operates on pre-scanned synthetic data with accurate depth maps and cannot be extended to real-world data. The limitation of state-of-the-art generalizable human view synthesis methods [32, 85] lies in the inability to be directly extended to multi-human scenes which impose extra challenges due to the inter-human occlusions and the complex human interactions.

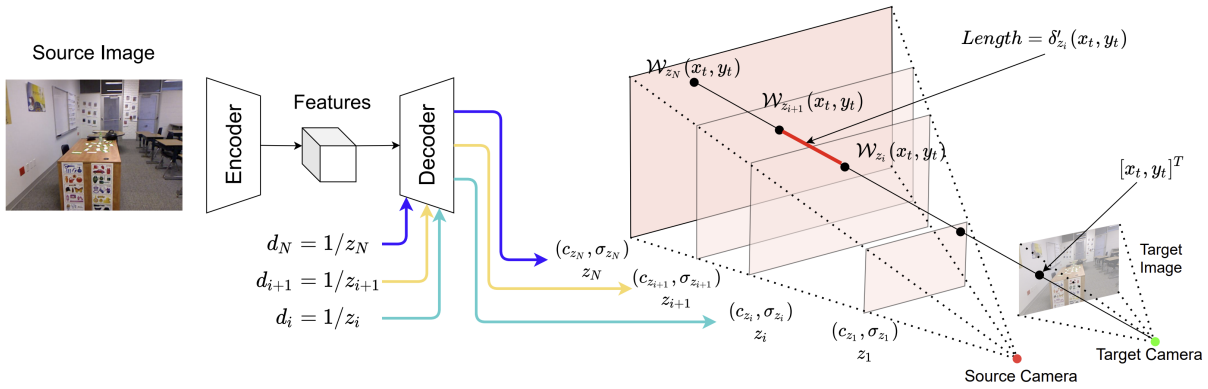

Figure 2.4: Full architecture of the single-view multi-plane neural radiance field [34] architecture.

2.2.3 Combination of Implicit And Explicit Representations

There are a few approaches in the literature that offer a combination of both implicit and explicit 3D representations to make use of their advantages. We divide them into pixel-level representations and object-level representations.

Pixel-level Representations

MINE [34] was proposed as a pixel-level combination of implicit and explicit representations for novel view synthesis with single-view input. They utilized an encoder-decoder architecture, shown in Figure 2.4, to predict front-parallel planes consisting of 4D radiance fields (RGB and volume density) [42] for each pixel. They sample the planes at arbitrary depth values throughout the training allowing the method to possess continuous representations of the depth dimension. This is followed by homography warping [67] and volumetric rendering [42] to render the target frame. MINE [34] proves the promising capability of marrying the concepts of NeRF with multi-plane images for high-quality synthesis results. However, being limited to a single-view setting, the method is constrained to a narrow viewing direction angle range.

Object-level Representations

Object-level layered scene representations were proposed to handle complex scenes with multiple subjects. Each layer represents a single entity in the scene which can be a human,

object, or background. [39] manipulated the timing of the subjects' motions in a video by decomposing each frame into a set of RGBA layers. ST-NeRF [83] modeled each dynamic human layer using a deformable model similar to D-NeRF [47] to achieve editable free-viewpoint rendering. Recently, [55] extended ST-NeRF by modeling the human subjects using NeuralBody [45] and predicted human segmentation masks as part of the network training. They were able to rely on more sparse input views by using 8 viewpoints instead of 16 and achieved a wider view range by covering a 360° range instead of 180°. The restriction of both methods is requiring lengthy per-scene training procedures for learning, yielding them inefficient to use.

2.3 Attention Mechanisms

The attention mechanism has gained significant focus in recent years for its impressive performance in natural language processing [69]. It has also proved to have a great impact on computer vision tasks like image classification [71], segmentation [81, 82], multi-view stereo [40], and hand-pose estimation [29]. Generally, the mechanism aims to explore dependencies and similarities between input and query vectors and then carries out weighted averaging to generate a contextual feature representation. In particular, AttsMVS [40] uses an attention-aware network embedded with a regularization module to robustly fuse multi-view information. In the object detection task, the DETR [2] framework combines a 2D CNN with an attention module to detect objects in parallel as a sequence of output tokens. In image classification, the ViT [14] model demonstrates the ability of the attention mechanism to learn global contexts without relying on CNN features, which are more suited to local concepts.

2.4 Summary

In this chapter, we discussed the classical view synthesis approaches based on the extent of explicit utilization of the scene geometry. We then presented the learning-based approaches according to the 3D scene representations. Explicit 3D approaches tend to directly model the camera frustum through different representations which enable better modeling of occluded areas. MPIs consist of parallel RGB-α planes that can be warped and projected to render novel views. The problem lies in the discretization of the depth of the planes leading to an incomplete 3D scene representation. LDIs are a more general representation that allows each pixel to have an arbitrary number of layers at different depths which is

more memory and space efficient. Implicit 3D representations model the 3D scene structure within the weights of neural networks. They are divided into per-scene optimization methods which require re-training for novel scenes and generalizable approaches that handle unseen scenes at inference time. There are also human-based approaches that handle the complexity of human subjects in terms of deformations and self-occlusions. Recent methods propose a combination of implicit and explicit representations either on a pixel or object level. The pixel-level combination takes the form of multi-plane neural radiance fields (MINE), while the object-level combination revolves around representing each object in the scene with an independent neural radiance field. Lastly, we go through the different attention mechanisms available in computer vision tasks. In the next chapter, we will discuss our proposed methodology which includes a detailed analysis of single-view MINE in terms of performance, generalization, and efficiency. In addition, we present a detailed overview of our newly proposed multi-view MINE architecture for elevating the performance of MINE through the utilization of multi-view information. With regard to the object-level representations, we present the details of our proposed approach, GenLayNeRF, which utilizes generalizable layered scene representations for multi-human novel view synthesis from sparse input views.

Chapter 3

Proposed Methodology

Layered scene representations offer a promising combination of explicit and implicit properties to achieve high-quality novel view synthesis. Such a layered combination exists at a pixel level or object level. MINE offer a pixel-level blend of multi-plane images [67] (MPI) and neural radiance fields (NeRF) [42] to achieve generalizable novel view synthesis. Existing literature work [34] in the domain is constrained to a narrow viewing direction range due to the single-input view setting. This chapter presents our proposed methodology for analyzing the capabilities of single-view MINE and enhancing its representational capacity by allowing multi-view input settings. In addition, object-level layered scene representations offer an effective solution to model scenes with complex multi-human subjects. However, existing literature work [55, 83] in the field is constrained to per-scene optimization settings making them inefficient for practical usage. In this chapter, we address the current research gap by proposing an object-level layered representation, GenLayNeRF, for achieving generalizable novel view synthesis for multi-human scenes using sparse input views while operating on novel subjects and poses at test time.

3.1 Multi-plane Neural Radiance Fields (MINE)

In this section, we carry out an in-depth technical analysis of single-view MINE [34] for novel view synthesis. We additionally explain the proposed architecture, MV-MINE, to utilize multi-view information for enhancing multi-plane radiance fields.

3.1.1 Analysis of Single-view MINE

We aim to assess three main aspects of the single-view MINE [34] architecture which is grouped into the following categories: Performance, Generalization, and Efficiency.

Regarding performance, we train the network on the ShapeNet dataset [4] which is a challenging dataset used by various state-of-the-art generalizable NeRF methods [66, 80] to assess their degree of generalization through various distribution of objects in training and testing. Additionally, we carry out ablation studies to test the impact of some NeRF [42] concepts on the results of MINE. As mentioned in Section 2.2.2, for each pixel in the target image, a 3D ray r is projected into the scene. 3D points are then sampled across the ray using a specific sampling technique. Fixed-depth sampling involves sampling points at rigid depth values across all training runs which limits the representational capacity of the depth dimension. Stratified sampling involves randomly sampling points at different depth locations across the projected rays. As points are sampled in random depth locations in every training run, the method achieves a continuous depth representation by the end of all training runs. In our ablation studies, we test the performance of MINE with fixed-depth sampling and stratified sampling. To produce the final color $\hat{C}(r)$ per ray r, two methods exist in the literature to fuse the predicted colors of all points i sampled on the ray. Alpha compositing involves carrying out an *over* operation [46] to aggregate the colors c_i for each point i based on their alpha value α_i, such that,

$$\hat{C}(r) = \sum_{i=1}^{N} (c_i \alpha_i \prod_{j=i+1}^{N} (1 - \alpha_j)), \tag{3.1}$$

On the other hand, volumetric rendering involves weighing all the RGB colors c_i by the density σ_i and the depth difference δ_i of each point $i \in [1, N]$, such that,

$$\hat{C}(r) = \sum_{i=1}^{N} (T_i (1 - \exp(-\sigma_i \delta_i)) c_i), \text{ where, } T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right) \tag{3.2}$$

This formulation enables a more intuitive representation of occlusions in the scene. In other words, if a 3D point i is occluded by points appearing before it across the ray the transmittance T_i will be low and the point will contribute less to the final color $\hat{C}(r)$ of the pixel. We carry out an ablation study to compare the effects of alpha compositing and volumetric rendering on the results of MINE.

With reference to generalization, we aim to validate the degree of generalization of single-view MINE [34] to new scenes that were not seen during training. The network uses

an encoder-decoder network allowing the decoder to be locally conditioned on the image features extracted per pixel from the encoder. The network learns features about the scene that serves as a strong prior when presented with frames from novel scenes leading to the generalization ability. To validate this ability, we feed the model with new scenes that were not seen during training and qualitatively judge the quality of the novel views produced by the model.

In terms of efficiency, MINE [34] is characterized to be more efficient than some of the implicit neural representation counterparts [80, 66] as it models only the frustum of the source camera, while the other synthesis methods represent the whole 3D space. During inference, MINE only produces N planes corresponding to N depth values from the source view to render a new view which is one single forward pass through the network. On the other hand, [80] needs to query a multi-layer perceptron for each point across a ray per pixel leading to $D \times H \times W$ forward passes through the network, where H and W are the height and width of the images respectively and D is the number of points sampled per ray. We aim to quantify such speed-up to verify the efficiency hypothesis, while also contributing a quantitative baseline time to compare with other NeRF-variants that offer an increase in speed just like MINE.

3.1.2 Proposed Multi-view MINE Architecture

Reliance on single-view input hinders the ability of MINE [34] to render target views that are far from the source view. We explored the extension of the architecture to a multi-view input setting to leverage the rich information seen from different views for better performance on more challenging datasets, while also opening the doors to comparing with state-of-the-art multi-view synthesis methods. The following section gives an overview of the proposed architecture, MV-MINE, along with the modules used for multi-view feature fusion.

Problem Formulation

Given a synchronized set Ω of frames I taken from B sparse input viewpoints of a scene such that $\Omega = \{I_1, .., I_B\}$, our target is to synthesize a novel view frame $\{I_q\}$ of the scene from a query viewing direction $\mathbf{q}$ with respect to a source view $\mathbf{s}$. Each input viewpoint b is represented by the corresponding camera intrinsics K, and camera rotation R and translation t, where $b = \{K_b, [R_b|t_b]\}$. For each input frame $I_w \in \mathbf{R}^{H \times W \times 3}$ with height H and width W, we extract a multi-scale feature pyramid using a ResNet50 [24]

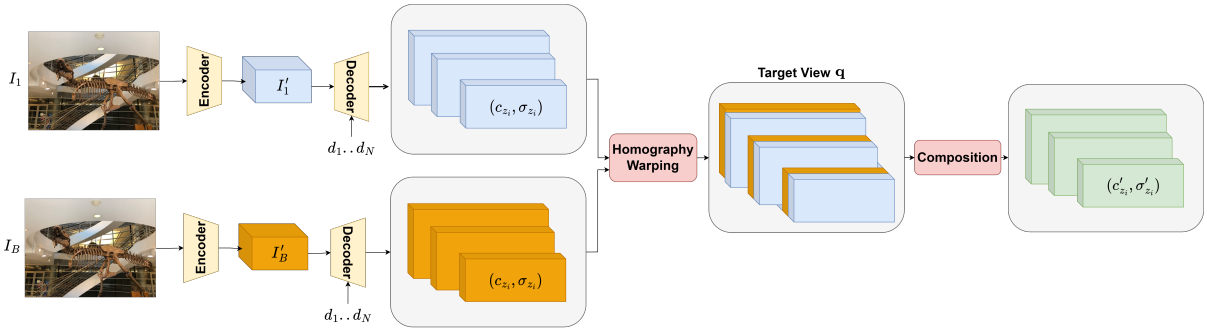

Figure 3.1: Full architecture of the proposed post-decoder fusion architecture design.

encoder network, pre-trained on ImageNet. The operation is carried out for all input views b in $\{1, .., B\}$ to produce the multi-scale feature planes for each view, defined as $\{I'_b \in \mathbf{R}^{H_b \times W_b \times C_b}\}$.

Similar to MINE [34], a decoder network with Monodepth2 [20] architecture takes the encoded feature maps and a disparity value $d_i = 1/z_i$ to produce the radiance field plane (c_{z_i}, σ_{z_i}), where c_{z_i}, σ_{z_i} represent the color and volume density at depth z_i, respectively. Homography warping is then utilized to retrieve the radiance field plane $(c'_{z_i}, \sigma'_{z_i})$ at the target camera $\mathbf{q}$. Lastly, volumetric rendering uses the predicted volume densities to aggregate the colors at different depth values producing the final target image $I_\mathbf{q}$.

We experiment with different architecture designs to fuse the multi-view image feature planes $I'_{1..B}$. The designs include doing the fusion before or after the decoder network. We discuss both designs in the following sections.

Post-Decoder Fusion

Figure 3.2 shows the full post-decoder fusion architecture design. For each view b, the multi-scale feature planes $\{I'_b\}$ are passed along with N disparity values retrieved with stratified sampling [42] to produce N radiance field planes $(c^b_{z_i}, \sigma^b_{z_i})$ at different depth values. We then warp the radiance field planes from each source view to the target view using homography warping producing a set of planes $(c'^{1:B}_{z_i}, \sigma'^{1:B}_{z_i})$ aligned with the target camera frustum. To compose the radiance field planes, we can carry out basic averaging across all views such that $(c'_{z_i}, \sigma'_{z_i}) = \frac{1}{B}\sum_b (c'^b_{z_i}, \sigma'^b_{z_i})$. However, such formulation could lead to hallucinations as equal weight is given to all input views. To overcome this, we experiment with doing weighted averaging based on the distance between the source view b and the target view q giving higher weight to views that are closer to the target view.

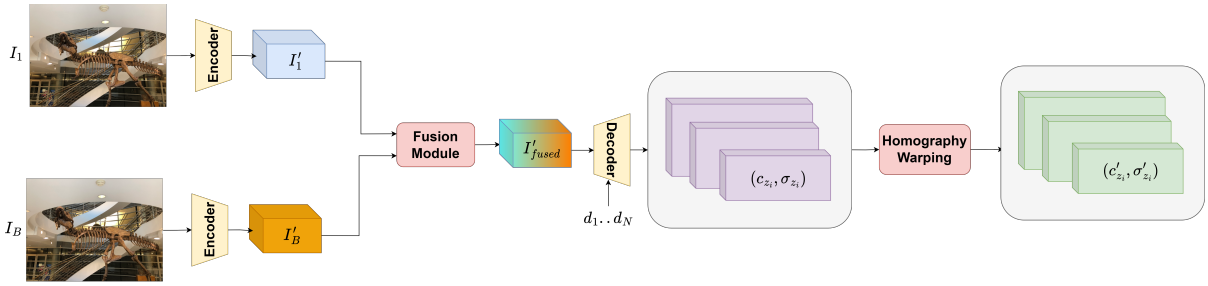

Figure 3.2: Full architecture of the proposed Pre-Decoder Fusion architecture design.

Pre-Decoder Fusion

Compositing the radiance field planes after passing through the decoder for each input view is considered highly inefficient. Specifically, the decoder is invoked $N \times B$ times. A more efficient solution would fuse the multi-view feature planes before passing through the decoder leading to N decoder invocations instead. We propose two fusion modules to aggregate the multi-view feature planes $I'_{1:B}$ with respect to a source view $\mathbf{s}$. The fused multi-view features I'_{fused} are then passed to the decoder to predict the radiance field planes.

Fixed View Fusion Module. In this module, we assume that the architecture accepts a fixed number of B input views. We start by concatenating each input feature plane with their corresponding viewing direction $b_{1:B}$. All feature planes are then concatenated and passed through channel-wise fusion layers $Conv_{1\times1}$, composed of 1×1 convolution layers with non-linear activation, to fuse the multi-view features per pixel. This is followed by 3×3 convolution $Conv_{3\times3}$ for learning spatially fused features. The final fused features are derived by adding the source view $\mathbf{s}$ features, such that,

$$I'_{fused} = Conv_{3\times3}(Conv_{1\times1}([I'_1; \gamma(b_1)] \oplus .. \oplus [I'_B; \gamma(b_B)])) + I'_{\mathbf{s}} \tag{3.3}$$

Attention-based View-agnostic Fusion Module. To increase the flexibility of our architecture with multi-view input, we propose an attention-based fusion module that accepts an arbitrary number B of input views throughout training and inference. Figure 3.3 shows the architecture of the module. Each input view feature $I'_{1:B-1}$ is concatenated with generated source view features and passed through a soft-attention masking module. To create a soft mask, the input is down-sampled using max pooling to widen the receptive field, then the features are refined using residual units, up-sampled to their original size, and the mask is normalized to the [0-1] range using a sigmoid function. The learned attention mask highlights areas of the input views that contain complementing features with respect to

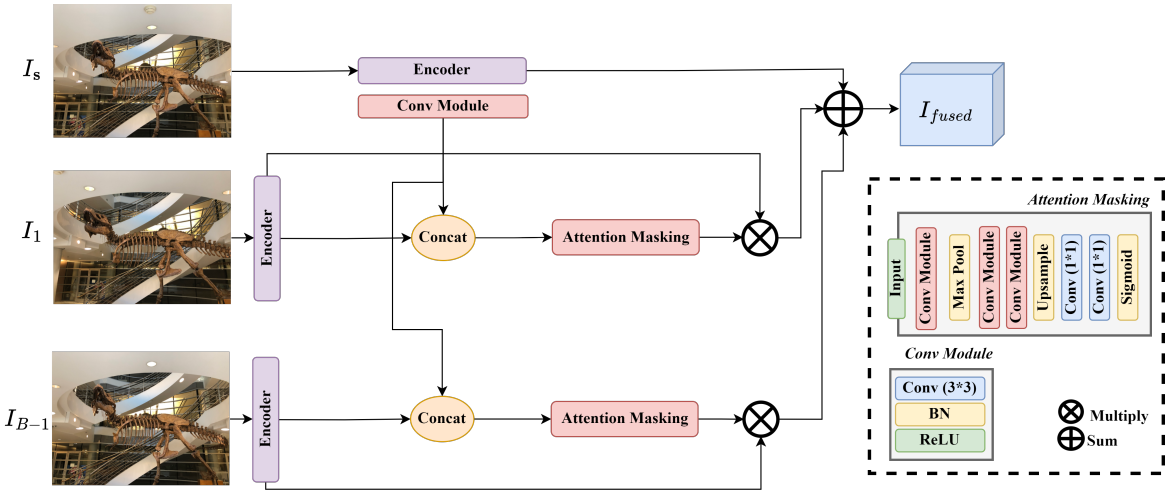

Figure 3.3: Full architecture of the proposed view-agnostic attention module. $(K * K)$ denotes a convolution layer with $K * K$ filter size.

the source view. Input view features are multiplied by their soft mask and added to the source view features generating the final fused features I'_{fused}.

3.2 GenLayNeRF: Generalizable Layered Scene Representations for Multi-human Novel View Synthesis

Novel view synthesis of scenes with close interactions between multiple humans impose challenges due to the complex inter-human occlusions. Object-level layered scene representations [55, 83] effectively handle some of the complexities by dividing the scene into multi-layered radiance fields, however, they are mainly constrained to per-scene optimization settings making them inefficient to use in practice. On the other hand, generalizable human view synthesis methods [32, 85] combine the 3D human meshes with image features to generalize to novel human subjects and poses, yet they are mainly designed to operate on single-human scenes. In this section, we propose, GenLayNeRF, a generalizable object-level layered scene representation for the free-viewpoint rendering of multiple human subjects which requires no per-scene optimization and very sparse views as input.

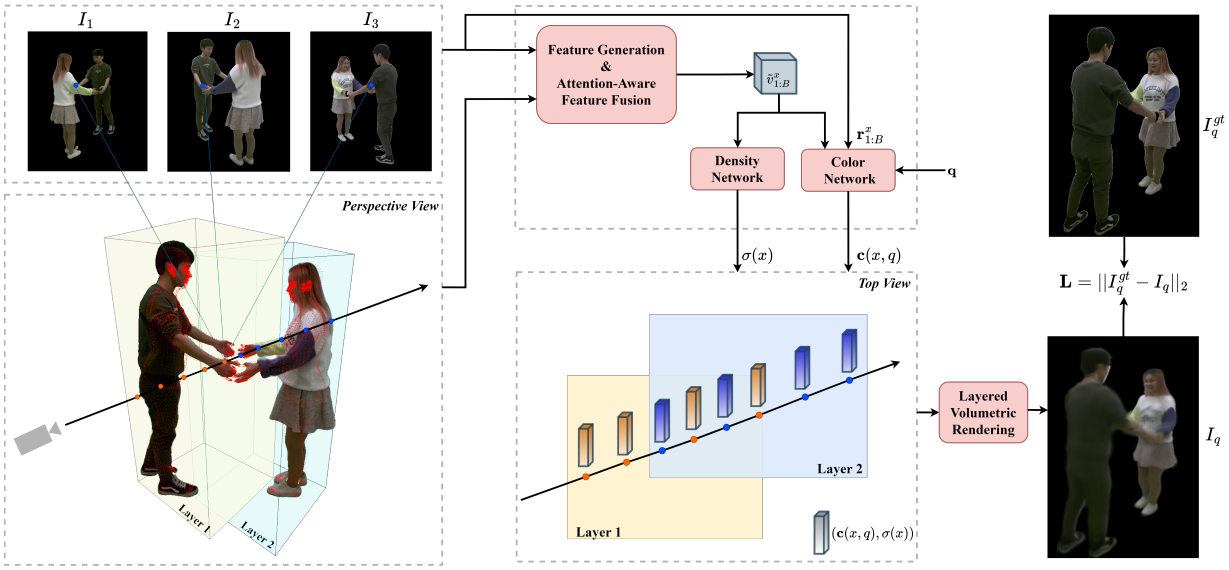

Figure 3.4: Overview of the GenLayNeRF approach.

3.2.1 Overview

As seen in Figure 3.4, we consolidate a layered scene representation where each human subject is modeled using the SMPL-X model (shown as red dots in the perspective view). We render the target image I_q from the query viewing direction $\mathbf{q}$ by projecting rays through the scene layers and sampling per-layer 3D points within the intersections of the rays with the layers (shown in the top view). Image-aligned and human-anchored features are then generated and effectively fused using self-attention and cross-attention modules to output the final fused features $\tilde{v}^x_{1:B}$. The generated features are passed to the density network to predict the volume density $\sigma(x)$, whereas the color network additionally uses the raw RGB values $\mathbf{r}^x_{1:B}$ and $\mathbf{q}$ to predict the color $\mathbf{c}(x, \mathbf{q})$.

3.2.2 Problem Definition

Given a synchronized set Ω of frames I taken from B sparse input viewpoints of a scene with N arbitrary number of humans , such that $\Omega = \{I_1, .., I_B\}$, our target is to synthesize a novel view frame $\{I_q\}$ of the scene from a query viewing direction $\mathbf{q}$. Each input viewpoint b is represented by the corresponding camera intrinsics K, and camera rotation R and translation t, where $b = \{K_b, [R_b|t_b]\}$. The N pre-fitted 3D human body meshes are given

for each input frame. Our system should be trained on multiple scenes and generalize to novel poses and human subjects at test time. A full overview of the proposed system is shown in Figure 3.4.

3.2.3 Layered Scene Representation

Scenes with multiple humans suffer from inter-human occlusions that become evident when subjects closely interact together. A practical solution to handle complex multi-human scenarios is dividing the scene into distinct layers where each layer models an entity using a neural radiance field [83, 39]. Entities can be humans, objects, or backgrounds. Our proposed approach focuses mainly on human layers and represents each layer using the SMPL-X model. SMPL-X [44] are deformable skinned models that are vertex-based where each model for a human h consists of 10,475 vertices, such that $s_h \in \mathbf{R}^{10,475\times 3}$. These models are responsible for preserving the local geometry and appearance of humans making it possible to model their complex deformations and occluded areas.

Our target is to render the full novel view image I_q from a query viewpoint $\mathbf{q}$. To achieve that, we first use the camera-to-world projection matrix, defined as $\mathbf{P}^{-1} = [R_q|t_q]^{-1}K_q^{-1}$, to march 3D rays across the multi-layered scene. In practice, we have a ray for each pixel p in the final image, where the ray origin $r_0 \in \mathbf{R}^3$ is the camera center and the ray direction is given as $d = \frac{\mathbf{P}^{-1}p - r_0}{||\mathbf{P}^{-1}p - r_0||}$.

3D points x are sampled across the rays at specific depth values z, where $x = r(z) = r_0 + zd$. Since we have several human layers in the scene, we determine the intersection areas of the rays with the humans using the 3D bounding box around each layer defined by the minimum and maximum vertex points of the SMPL-X meshes. We then sample depth values within the n_p intersecting areas only such that $z \in [[z_{near_1}, z_{far_1}], .., [z_{near_{n_p}}, z_{far_{n_p}}]]$. This guarantees that the sampled points lie within areas that contain the relevant human subjects as clear in the top view shown in Figure 3.4. Our proposed method is capable of implicitly representing the true contents of the points lying inside the ambiguous intersection areas between layers. This is achieved using the multi-view aggregated features and the layered volumetric rendering module.

3.2.4 Feature Generation And Attention-Aware Feature Fusion

The original NeRF architecture [42] directly uses the sampled points' locations and the query viewing direction to predict the color and density of each point. Conditioning the

NeRF predictor on extracted image features proved to achieve impressive generalization capabilities [80, 72] by implicitly learning strong priors from the training scenes. In our proposed approach, we extract multi-view image features for each query point x and effectively merge them using attention-based fusion modules to derive the needed spatially-aligned feature vectors. This strategy allows the system to extrapolate to novel human subjects and poses beyond what it saw during training by learning implicit correlations between the independent human layers.

Image-aligned Feature Generation

Given an input view image $I_w \in \mathbf{R}^{H \times W \times 3}$ with height H and width W, we extract a multi-scale feature pyramid using a ResNet34 [24] backbone network f, pre-trained on ImageNet. The multi-scale feature planes have the following dimensionality $[(64 \times \frac{H}{2} \times \frac{W}{2}), (64 \times \frac{H}{4} \times \frac{W}{4}), (128 \times \frac{H}{8} \times \frac{W}{8})]$. The feature maps are concatenated into a shape $(C \times \frac{H}{2} \times \frac{W}{2})$ after being bilinearly upsampled to the highest resolution, which is $(\frac{H}{2} \times \frac{W}{2})$ producing a feature map $I_b^{'} \in \mathbf{R}^{H \times W \times C}$ with C output channels. The operation is carried out for all input views b in $\{1, .., B\}$. We then project the point x on all input feature maps $I_b^{'}$ to collect the corresponding image-aligned features for each view b denoted as p_b^x.

Human-anchored Feature Generation

Existing layered scene representations [55] follow the approach of NeuralBody [45] by encoding the vertices of human layers using learnable embeddings that are unique to each layer in each training scene. In our approach, we embed the vertices with general features instead by projecting the world-coordinate vertices s_h on the multi-view feature maps extracted from the input images, such that, $v_{h,b} = I_b^{'}[K_b((R_b s_h^T) + t_b)]$. $v_{h,b} \in \mathbf{R}^{10,475 \times C}$ represents the features of the vertices projected on feature map $I_b^{'}$ for layer h.

We query the radiance field predictor using continuous 3D sampled points. For that reason, the sparse human vertices need to be diffused into a continuous space that can be queried at any location. We incorporate the SparseConvNet [22, 45] architecture which utilizes 3D sparse convolution to diffuse the vertex features into different nearby continuous spaces for every layer. A detailed description of the architecture of the employed SparseConvNet network is shown in Table 3.1. Before diffusion, the vertex locations of each layer are transformed to their SMPL-X coordinate system to make sure that the diffused spaces are independent of the humans' world locations. To effectively anchor the network on the available SMPL-X body priors, we transform x to the SMPL-X coordinate space

	Layer Description	Output Dim.
	Input volume	$D' \times H' \times W' \times 64$
1-2	[F=(3,3,3),K=64,S=1]×2	$D' \times H' \times W' \times 64$
3	[F=(3,3,3),K=64,S=2]	$\frac{D'}{2} \times \frac{H'}{2} \times \frac{W'}{2} \times 64$
4-5	[F=(3,3,3),K=64,S=1]×2	$\frac{D'}{2} \times \frac{H'}{2} \times \frac{W'}{2} \times 64$
6	[F=(3,3,3),K=64,S=2]	$\frac{D'}{4} \times \frac{H'}{4} \times \frac{W'}{4} \times 64$
7-9	[F=(3,3,3),K=64,S=1]×3	$\frac{D'}{4} \times \frac{H'}{4} \times \frac{W'}{4} \times 128$
10	[F=(3,3,3),K=128,S=2]	$\frac{D'}{8} \times \frac{H'}{8} \times \frac{W'}{8} \times 128$
11-13	[F=(3,3,3),K=128,S=1]×3	$\frac{D'}{8} \times \frac{H'}{8} \times \frac{W'}{8} \times 128$
14	[F=(3,3,3),K=128,S=2]×3	$\frac{D'}{16} \times \frac{H'}{16} \times \frac{W'}{16} \times 128$
15-17	[F=(3,3,3),K=128,S=1]×3	$\frac{D'}{16} \times \frac{H'}{16} \times \frac{W'}{16} \times 128$
	Resize & Concat layers 5,9,13, and 17	$\frac{D'}{16} \times \frac{H'}{16} \times \frac{W'}{16} \times 384$

Table 3.1: The architecture of SparseConvNet. The layers consist of 3D sparse convolution, batch normalization, and ReLU activation. "F" denotes filter size, "K" denotes the number of kernels, and "S" denotes stride.

of its corresponding human layer. Trilinear interpolation is then utilized to retrieve the corresponding human-anchored features v_b^x from the diffused layer spaces of each view b.

Attention-Aware Feature Fusion

There are different possible strategies to fuse the feature representations $(v_{1:B}^x, p_{1:B}^x)$ for point x; one simple strategy is a basic averaging approach [48, 49]. This usually leads to smoother output and ineffective utilization of the information seen from distinct views. To learn effective cross-view correlations, we employ a self-attention module that attends between all the multi-view human-anchored features $v_{1:B}^x$ where each feature in one view is augmented with the extra features seen from the other views. This is done with a weighted average between the view features based on their similarity. Each view feature is first concatenated ($\oplus$) with its corresponding viewing direction $\mathbf{d}_b$.

Specifically, the self-attention weights mv_self^x and the view-aware human-anchored

features $\hat{v}^x_{1:B}$ are calculated as:

$$\begin{aligned} v^x_{1:B} &= v^x_{1:B} \oplus \mathbf{d}_{1:B}, \\ mv\_self^x &= soft(\frac{1}{\sqrt{d_{k_1}}} query(v^x_{1:B}) \ . \ key(v^x_{1:B})^T), \\ \hat{v}^x_{1:B} &= mv\_self^x \ . \ val_1(v^x_{1:B}) + val_2(v^x_{1:B}), \\ mv\_self^x \in \mathbf{R}^{B\times B}&, \ \hat{v}^x_{1:B} \in \mathbf{R}^{B\times C}, \mathbf{d}_{1:B} \in \mathbf{R}^{B\times 3} \end{aligned} \tag{3.4}$$

where key, $query$, and (val_1, val_2) represent the key, query, and value embeddings of the corresponding argument features respectively, and d_{k_1} denotes the dimensionality of the key embedding and is set to 128. $soft$ denotes the softmax operation.

We additionally make use of the rich spatial information in the image-aligned features by carrying out cross-attention from the view-aware human-anchored features to the image-aligned features. The similarity between the multi-view image features and the per-view vertex features is used to re-weigh the image features and embed them with the vertex features. The fused features $\tilde{v}^x_{1:B}$ are calculated with the same formulation in Equation 3.5. The arguments of $query$ and val_2 are replaced by $\hat{v}^x_{1:B}$, while the arguments of key and val_1 are replaced by $p^x_{1:B}$, such that,

$$\begin{aligned} \hat{v}^x_{1:B} &= \hat{v}^x_{1:B} \oplus \mathbf{d}_{1:B}, \\ p^x_{1:B} &= p^x_{1:B} \oplus \mathbf{d}_{1:B}, \\ mv\_cross^x &= soft(\frac{1}{\sqrt{d_{k_1}}} query(\hat{v}^x_{1:B}) \ . \ key(p^x_{1:B})^T), \\ \tilde{v}^x_{1:B} &= mv\_cross^x \ . \ val_1(p^x_{1:B}) + val_2(\hat{v}^x_{1:B}), \\ mv\_cross^x \in \mathbf{R}^{B\times B}&, \ \tilde{v}^x_{1:B} \in \mathbf{R}^{B\times C} \end{aligned} \tag{3.5}$$

Afterward, we carry out view-wise averaging, such that $\tilde{v}^x = \frac{1}{B}\sum_b \tilde{v}^x_b$, to generate the final fused feature representation for x.

3.2.5 Radiance Field Predictor

The radiance field predictor, shown in Figure 3.5, consists of a color network to predict the RGB color $\mathbf{c}$ of a point x, and a density network to predict the volume density σ.

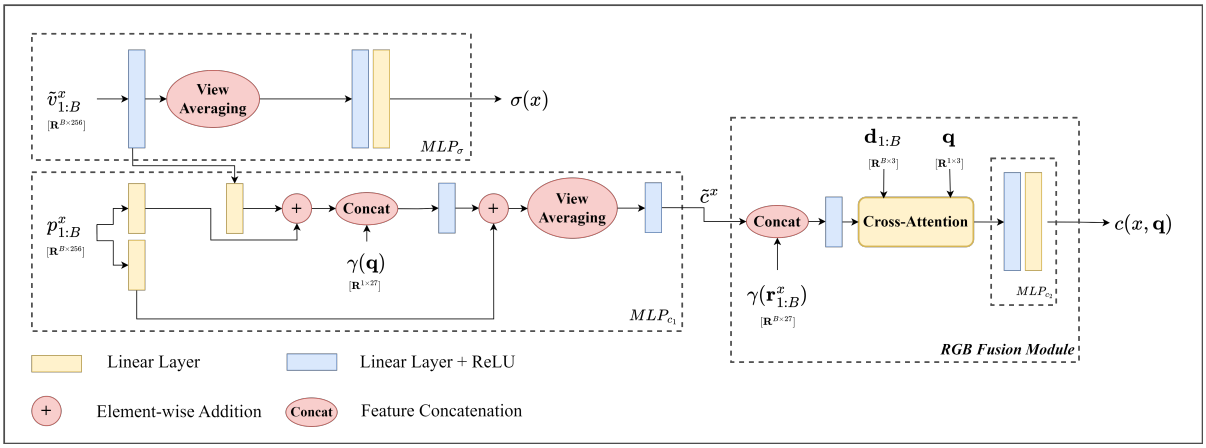

Figure 3.5: The architecture of the Radiance Field Predictor.

Color Network

For the prediction of the color $\mathbf{c}$ of point x, we use of the query viewing direction $\mathbf{q}$ to model the view-dependent effects [42]. In addition, we explicitly augment the fused high-level features with low-level pixel-wise information to leverage the high-frequency details in the images. This has been achieved with an RGB fusion module which concatenates the high-level features with the encoded raw RGB pixel values $\mathbf{r}^x_b$ for each view b. RGB values from closer input views are assigned higher weights by cross-attending $\mathbf{q}$ with the input viewing directions $\mathbf{d}_{1:B}$ such that,

$$\begin{aligned} \tilde{c}^x &= MLP_{c_1}(\tilde{v}^x_{1:B}; \gamma(\mathbf{q}); p^x_{1:B}), \\ \hat{c}^x_{1:B} &= \{[\tilde{c}^x \oplus \gamma(\mathbf{r}^x_1)], ..., [\tilde{c}^x \oplus \gamma(\mathbf{r}^x_B)]\}, \\ rgb\_att^x &= soft(\frac{1}{\sqrt{d_{k_2}}} query(\mathbf{q}) \ . \ key(\mathbf{d}_{1:B})^T), \\ \mathbf{c}(x, \mathbf{q}) &= MLP_{c_2}(rgb\_att^x \ . \ val_1(\hat{c}^x_{1:B})), \\ rgb\_att^x &\in \mathbf{R}^{1 \times B}. \end{aligned} \tag{3.6}$$

Density Network

We predict volume density $\sigma(x)$ for point x using the fused feature $\tilde{v}^x$, such that:

$$\sigma(x) = MLP_\sigma(\tilde{v}^x), \tag{3.7}$$

, where MLP_σ, MLP_{c_1}, and MLP_{c_2} consist of fully connected layers described in the supplementary material. $\gamma : \mathbf{R}^3 \rightarrow \mathbf{R}^{(6\times l)+3}$ denotes a positional encoding [42] with $2 \times l$ basis functions and d_{k_2} is set to 16.

3.2.6 Layered Volumetric Rendering

Layered volumetric rendering is used to accumulate the predicted RGB and density for all points across all human layers. The points in intersecting areas n_p across all human layers are sorted based on their depth value z before accumulation. The synthesized image I_q is calculated as follows,

$$\begin{aligned} I_q(p) &= \sum_{i=1}^{n_p} \int_{z_{near_i}}^{z_{far_i}} \mathbf{T}(z)\sigma(r(z))\mathbf{c}(r(z),\mathbf{q})dz \\ , where \quad \mathbf{T}(z) &= exp\left(-\int_{z_{near_i}}^{z} \sigma(r(s))ds\right) \end{aligned} \tag{3.8}$$

We sample 64 points per ray and approximate the internal integral using the quadrature rule [42]. Given a ground truth novel view image I_q^{gt}, all of the network weights are supervised using the traditional L2 Norm photo-metric loss, such that $\mathbf{L} = ||I_q^{gt} - I_q||_2^2$.

3.3 Summary

In this chapter, we presented the methodology used to explore the boundaries and capabilities of the combination between neural radiance fields and multi-plane images. Specifically, we discussed the performance, generalization, and efficiency aspects for the analysis of single-view (MINE) [34]. Furthermore, we showcased the architecture of multi-view multi-plane neural radiance field architecture, MV-MINE, which effectively utilizes information from different viewpoints to enhance the view synthesis performance. Lastly, we presented the architecture of the proposed object-level layered scene representation, GenLayNeRF, that successfully handles the complexity of multi-human scenes for novel view synthesis while working with unseen subjects and poses at test time. In the next chapter, we will discuss the details of our experiments and present the results both quantitatively and qualitatively.

Chapter 4

Experimental Results

This chapter presents the experiments done to evaluate our proposed methodologies and architectures. Section 4.2 presents the experimental setup and results regarding the analysis points for single-view MINE. In Section 4.3, we show the results of the comparison of our proposed MV-MINE architecture with baseline methods and the effect of different feature fusion modules. Lastly, regarding our proposed GenLayNeRF approach, we present the datasets used, comparison settings, experiment results, and ablation studies in Section 4.4.

4.1 Evaluation Metrics

We utilize NeRF [42] in the evaluation metrics for novel view synthesis which are peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [73], and the learned perceptual image patch similarity (LPIPS) [84] in all experiments.

4.2 Single-view MINE Experiments

We present the experimental details and results of the technical analysis of single-view MINE [34] in terms of performance, generalization, and efficiency as mentioned in Section 3.1.1. Figure 4.1 shows an overview of all the experiments made in this section.

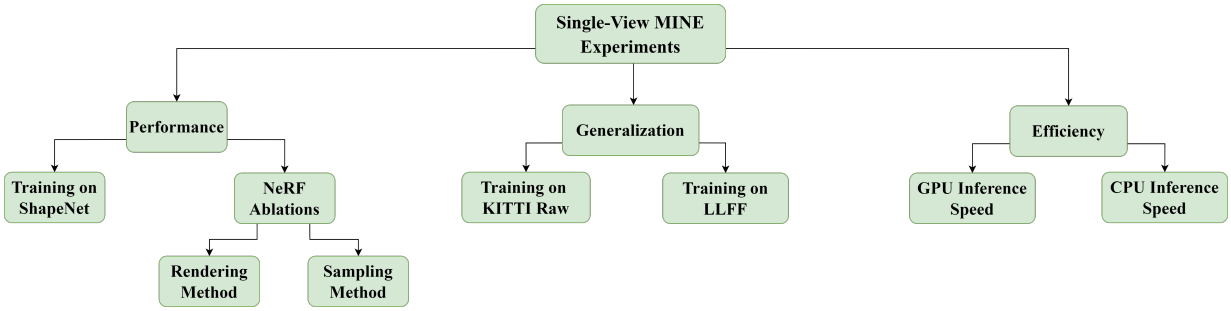

Figure 4.1: Overview of the experiments made for analyzing the performance, generalization, and efficiency of single-view MINE.

4.2.1 Performance

Regarding performance, we present the experimental setup and results used for training on the ShapeNet dataset [4], and the ablation studies made.

Setup

Training on ShapeNet. We train MINE on specific subsets of the ShapeNet dataset [4] to have a fair performance comparison with pixelNeRF [80] which is a generalizable single-view NeRF method. Specifically, we focus on using the Category Agnostic ShapeNet experiments [80] which train on single-view images of 13 categories of objects. Each category has multiple objects and each object has 24 views. Following [80] we sample one random view for training and 23 other views as target views. The train-test split is composed of 156,877 and 45,586 source and target pairs for training and validation respectively. We trained on 4 V100 GPUs with batch size 4 and a 0.001 learning rate for the encoder and decoder. Training for one epoch takes about 6 hours and validation takes about 3 hours.

Effect of Continuous Depth & Volumetric Rendering. The continuous depth reconstruction proposed by NeRF [42] allowed MINE [34] to generalize the discretized depth representation of MPI [67]. We verify this hypothesis by training on the LLFF [41] dataset from scratch with the fixed depth sampling approach from MPI [67] and the stratified sampling approach from NeRF [80]. In addition, using the volumetric rendering technique applied by NeRF [80] instead of alpha compositing [67] is one of the factors contributing to enhancing the results of MINE [34]. To verify that, we train on the LLFF dataset with both volumetric rendering and alpha compositing. The LLFF dataset [41] contains real-world images taken by phone camera at views lying in an equally spaced grid of a specific size. There are 8

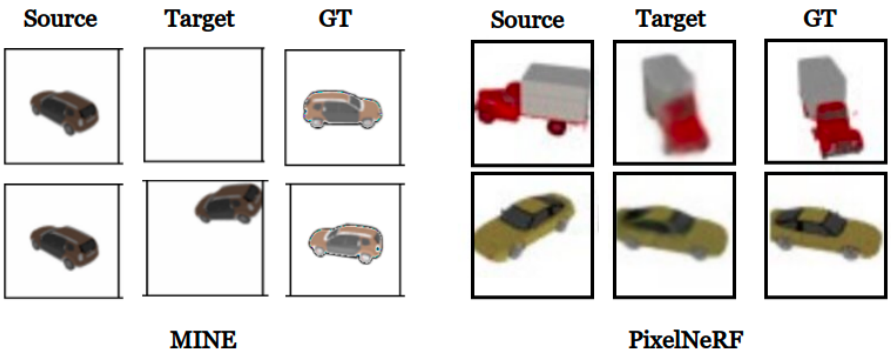

Figure 4.2: Output of MINE after training on Shapenet [4] using the same preprocessing used by pixelNeRF [80]. "GT" denotes the ground truth target view, "Target" denotes the output target view, and "Source" denotes the input view to the network. Distortion in GT of MINE is due to normalizing the images by 0.5.

scenes available with each scene having around 20-50 views available. The scenes available are of the following objects: fern, flower, fortress, horns, leaves, orchids, room, and trex. During the training, for each view in the scene, a random view is taken as the target view. The sparse disparity loss is included and the scale is calculated using 3D point clouds estimated for the images using COLMAP [50, 51]. The training was done on 4 V100 GPUs and took around 4 hours. We used a decaying learning rate starting at 0.001 and decaying by 0.1 every 50 epochs for 200 epochs and a batch size of 2.

Results

Training on ShapeNet. We carried out a qualitative analysis to check the plausibility of results returned by MINE compared to pixelNeRF [80] with single-view input on ShapeNet [4], shown in Figure 4.2. The first row shows that MINE failed to render the target object within the boundaries of the image plane since the target viewing direction is very far from the source viewing direction. In the second row, the object was rendered within the image plane and the car's structure was retained appropriately since the two viewing directions are closer, in this case, however, the car location is still inaccurate. On the other hand, pixelNeRF is able to correctly render the target view object in an accurate location within the image plane regardless of how far the source and target views are.

Effect of Continuous Depth & Volumetric Rendering. Table 4.1 shows the results after

Sampling	Compositing	LPIPS ↓	SSIM ↑	PSNR ↑
Stratified	Volumetric	0.397	0.5244	18.12
Fixed	Volumetric	**0.389**	**0.5331**	**18.20**
Stratified	Alpha	0.448	0.4870	17.78

Table 4.1: Training results of MINE [34] on the LLFF dataset [41] with fixed, stratified sampling, volumetric rendering, and alpha compositing.

training MINE on LLFF with fixed disparity taken at equally spaced locations, with a random stratified sampled disparity in each training step, and with volumetric rendering and alpha compositing for aggregating the colors from the radiance field planes. It can be seen that the usage of stratified sampling did not enhance the results, yet the fixed disparity yielded slightly better performance on all metrics. However, the usage of volumetric rendering led to significantly better results than alpha compositing.

4.2.2 Generalization

Regarding generalization, we present the experimental setup and results of evaluating MINE [34] on novel scenes from the LLFF [41] and KITTI Raw [18] datasets.

Setup

Generalization on LLFF. In this experiment, we leave out the "fortress" scene from the LLFF dataset during training and evaluate it. This setting is considered challenging as the novel scene differs highly from the scenes seen during training. We follow the same experimental setup of the ablation studies mentioned in Section 3.1.1.

Generalization on KITTI Raw. We utilize samples of scenes from the KITTI Raw [18] dataset which were not seen during training (specifically scenes dated 2011_09_26 scenes 0104, 0106, 0113, and 0117). The model is tested on each image in the scenes individually. The GPU used for this experiment is NVIDIA GTX1070 8GB.

Results

Generalization on LLFF. The results of the generalization experiment on the fortress scene on the LLFF dataset [41] are shown in Figure 4.3. In the second column, it is clear that the

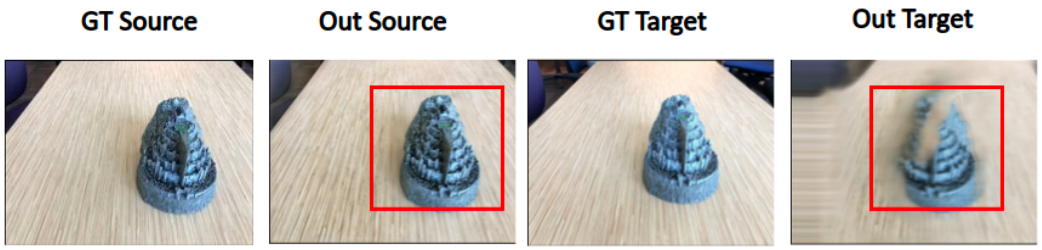

Figure 4.3: Output of MINE after training it on 7 LLFF [41] categories and evaluating on the fortress scene. "GT" denotes ground truth and "Out" denotes the output of the model.

model was successful in rendering the geometric structure of the source image accurately. However, regarding the target novel views in the fourth column, it is clear the model failed to render the geometric structure of the whole object properly showing a lot of distortions.

Generalization on KITTI Raw. The results of testing the generalization on KITTI Raw [18] made us consider two main divisions of the problems encountered, the division of global problems which are visible in almost all of the pictures tested, and local problems which are visible in specific frames of the scenes. The first global problem is edge distortion where the edges of the videos while moving along the Z-axis are highly distorted. This happens due to duplicating the edge pixels to in-paint parts which were occluded in the source image, as visible in Figure 4.4. Another global problem is rendering pixels where an object is behind another object which is visible clearly in Figure 4.4 samples 1-3. Specifically, in sample 1, it is visible in the sign at the front, when trying to render the car behind it. In sample 2, it is visible on the right of the motorcycles, where motorcycles are getting distorted and rendered unsuccessfully due to small barriers in front of them. In sample 3, it is visible when looking at the car on the right, when the camera moves the car shape changes. Lastly, the heads of pedestrians show large distortions as visible in Figure 4.4 samples 1 and 4, or have a ghost-like effect, as seen in samples 2 and 3. Locally, we highlight areas in Figure 4.5 where pedestrians, traffic signs, and buildings suffer from splitting distortions, ghost-like effects, and the incorrect representation of the geometric structure.

4.2.3 Efficiency

Regarding efficiency, we present the experimental setup and results of comparing the inference speed of MINE [34] and pixelNeRF [80].

Figure 4.4: Global problems encountered in the KITTI Raw [19] generalization experiment.

Setup

We fixed the input frame shape to (128, 128) and the number of planes in MINE to 32 to be the same as the number of points sampled per ray in pixelNeRF. We used the pre-trained models and the code published for both pixelNeRF and MINE to run the experiments. The GPU used for the experiment is NVIDIA GTX1070 8GB, and the CPU is Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz with 8 cores and 32 GB RAM. To obtain an accurate time per frame, we ran 150 frames and got the average time per frame.

Results

Table 4.2 presents the results of this experiment. We were able to validate that MINE is more efficient in inference than pixelNeRF [80]. Particularly, MINE renders a single (128, 128) target frame in 0.77 seconds on GPU, while pixelNeRF takes 1.24 seconds which is approximately 38% speed enhancement. Regarding CPU, MINE shows 45% enhancement over pixelNeRF.

Figure 4.5: Local problems encountered in the KITTI Raw [19] generalization experiment.

Method	GPU Time	CPU Time
Single-view MINE (32 Planes)	0.77s	8.43s
pixelNerf (32 Coarse Points)	1.24s	15.45s

Table 4.2: Results of comparing rendering time per frame for pixelNeRF[80] and MINE[34].

4.3 Multi-view MINE Experiments

Our experiments in this section focus on evaluating the performance of the proposed architecture designs for MV-MINE, described in Section 3.1.2, and comparing them against baseline NeRF methods. We discuss the experimental setup, while also presenting the results both quantitatively and qualitatively.

4.3.1 Experimental Setup

The experimental setup involves the training and testing details and the datasets used in all experiments. The experiments are split into an evaluation of the proposed modules and a comparison with baseline methods.

Comparison Of Fusion Techniques

We train the fusion modules on all the scenes of the LLFF [41] dataset. Validation is done on unseen target views. The fixed-view pre-decoder module 3.1.2 was trained and

evaluated on 5 input views, while other modules used a range of 3-7 input views for training and were evaluated on 5 input views for a fair comparison.

Comparison With Baseline Methods

Regarding per-scene methods, we evaluate our approach against NeRF [42] and SRN [58]. Regarding generalizable methods, we include pixelNeRF [80] in our baselines. In addition, we provide the results of LLFF [41] as an MPI method. Our proposed method and pixelNeRF were both trained on a collection of the LLFF [41], Spaces [16], IBR-collected [72], and RealEstate-4k [87] datasets. Training samples for each epoch are drawn with the following probabilities 0.4, 0.15, 0.35, and 0.1 respectively. Evaluation is done on novel target views of the LLFF dataset. NeRF was trained on each scene of the LLFF dataset separately.

4.3.2 Results

We present the results of our experiments for fusion modules and baseline methods comparison.

Comparison Of Fusion Techniques

Table 4.3 presents the performance of the original MINE method with single view inputs along with our proposed fusion modules operating on 5 input views. It could be seen that the post-decoder fusion with averaging leveraged multi-view information to enhance results compared to single-view MINE. Introducing weighted averaging led to better utilization of features from close views and significantly enhanced results on all metrics. Implicit feature aggregation introduced in the fixed-view pre-decoder fusion notably elevated the performance. Lastly, shifting to view-agnostic attention-aware fusion shows the best overall performance on all metrics. This validates the impact of the learned soft masks in highlighting important features in the input views with respect to the source view. Qualitatively, it could be seen in Figure 4.6 that MINE suffers from strong hallucinations around image borders. The post-decoder module solves that issue yet still contains strong blur artifacts. The pre-decoder modules show the best synthesis quality, especially with the attention-aware module in terms of lighting and colors.

Figure 4.6: Comparison of the proposed multi-view fusion modules. We include the original MINE [34] method operating with single input views. All fusion modules were tested with 5 input views.

Method	LPIPS ↓	SSIM ↑	PSNR ↑
Single-View MINE	0.397	0.5244	18.12
Post-Decoder Fusion (Averaging)	0.354	0.601	19.56
Pre-Decoder Fusion (Averaging)	0.321	0.621	20.10
Post-Decoder Fusion (Weighted Averaging)	0.298	0.652	20.43
Fixed-View Pre-Decoder Fusion	0.232	0.761	24.08
Attention-based Pre-Decoder Fusion	**0.223**	**0.803**	**24.43**

Table 4.3: Quantitative comparison of the performance of the proposed multi-view fusion modules using 5 input views and MINE using a single input view.

Comparison With Baseline Methods

Table 4.4 shows the results of our attention-aware fusion module 3.1.2 compared to the baseline view synthesis methods. Regarding per-scene methods, it could be seen that our method significantly surpasses SRN on all metrics, while performing better than NeRF on the LPIPS metric without per-scene training. Regarding the generalizable methods, we show comparable performance to both LLFF and pixelNeRF, while performing better than pixelNeRF on the LPIPS and SSIM metrics. We also introduce slight improvements over LLFF on the SSIM and PSNR metrics.

Method	LPIPS ↓	SSIM ↑	PSNR ↑
SRN (P)	0.378	0.668	22.84
NeRF (P)	0.250	0.811	26.50
LLFF (G)	0.212	0.798	24.13
pixelNeRF (G)	0.224	0.802	24.61
Ours (G)	0.218	0.808	24.56

Table 4.4: Comparison of our attention-based view-agnostic fusion module, with baseline view synthesis methods. "P" denotes per-scene optimization methods, while "G" denotes generalizable methods.

4.3.3 Discussion

Regarding the single-view MINE analysis, we concluded that MINE is limited only to render novel views that are close to input source views, and in the current setting would fail to give 360° views of a scene like other NeRF variants [66, 80]. We believe that the reason behind that is having only a single image as input, so the model doesn't get exposed to several views to enhance its novel view prediction on far target poses. Moreover, homography warping could be another reason why the model has limited capability to render a wide range of views since the decoder is only producing a feature plane representation that is conditioned on the source image, and transforming the output planes by a large amount is considered ill-posed and would cause the distortion and incorrect results shown previously. In addition, it could be concluded that MINE cannot generalize to areas around the edges of the images since it will need to in-paint the content of areas that it hasn't seen before from the single-view input. In the output, we saw that the model does nearest neighbor interpolation in those areas instead of correctly predicting their structure and color. The method also failed to appropriately render the fortress scene due to its disparate distribution compared to the training scenes which highlights the weak generalization ability of single-view MINE.

Regarding the proposed MV-MINE experiments, it was clear that the multi-view information contributed to the enhancement of the synthesis quality compared to its single-view counterpart. In complex scenes with severe occlusions, the utilization of a single-input image to render a novel view increases the difficulty of predicting the structure of the scene in areas that were not visible in the source input view. On the other hand, if a model utilizes multiple views as input, it can solve the ambiguities in the occluded areas by reasoning about the information from different views. We carried out fusion using basic averaging techniques and attention modules. Our proposed attention modules achieved the best synthesis quality as seen in Figure 4.6 compared to the averaging and post-decoder modules

which highlights its effectiveness in leveraging multi-view information. Specifically, the modules learned the appropriate soft masks to successfully highlight the complementary features present in the input views with respect to the source views and aggregated the features properly into a final fused representation.

4.4 GenLayNeRF: Generalizable Layered Scene Representations for Multi-human Novel View Synthesis

In this section, we introduce the experimental details regarding our proposed GenLayNeRF architecture, described in Section 3.2. This includes a discussion of the datasets used in the experiments, the training details, the baselines used in the comparison, the experimental results, and the ablation studies.

4.4.1 Datasets

The existence of readily-available open-source multi-human view synthesis datasets is limited. To solve this challenge, we construct two new datasets, ZJU-MultiHuman and DeepMultiSyn, for our evaluation and comparison purposes. Both datasets will be published to act as a benchmark for multi-human view synthesis methods.

DeepMultiSyn

The DeepMultiSyn dataset is an adaptation of the 3D reconstruction dataset published by DeepMultiCap [86]. We take the raw real-world multi-view sequences and process them for novel view synthesis. There exist 3 video sequences of scenes containing 2 to 3 human subjects captured from 6 synchronized cameras. The number of frames in the sequences ranges from 756 to 1976. We use EasyMoCap [54] to fit the SMPL-X human models for all the subjects in all available frames. Additionally, we predict the human segmentation masks following [35] to separate the humans from the background. This dataset is considered challenging due to the existence of close interactions and complex human actions such as boxing, and dancing activities.

	Method	DeepMultiSyn		ZJU-MultiHuman	
		PSNR	SSIM	PSNR	SSIM
(a) Seen Models, Seen Poses					
S	NeRF	15.49	0.497	16.42	0.525
	D-NeRF	17.08	0.702	18.53	0.748
	L-NeRF*	23.79	0.845	24.72	0.898
	Ours$_{ft}$	**24.77**	**0.873**	**24.85**	**0.906**
G	PixelNeRF	14.81	0.534	19.74	0.629
	SRF	20.39	0.724	17.87	0.657
	IBRNet	19.45	0.741	20.03	0.766
	NHP*	20.91	0.698	21.75	0.813
	Ours	**23.61**	**0.847**	**24.61**	**0.893**
(b) Seen Models, Unseen Poses					
S	L-NeRF*	21.37	0.810	22.84	0.867
G	PixelNeRF	14.14	0.520	16.88	0.560
	SRF	18.07	0.663	17.93	0.680
	IBRNet	18.01	0.710	19.84	0.772
	NHP*	20.26	0.677	20.64	0.791
	Ours	**22.19**	**0.826**	**23.04**	**0.873**
(c) Unseen Models, Unseen Poses					
G	PixelNeRF	13.12	0.457	Not Applicable	
	SRF	13.95	0.548		
	IBRNet	18.80	0.672		
	NHP*	19.51	0.678		
	Ours	**20.43**	**0.787**		

Table 4.5: Comparison with generalizable and per-scene NeRF methods on the DeepMultiSyn and ZJU-MultiHuman Datasets. "G" and "S" denote generalizable and per-scene methods, respectively. "*" refers to human-based methods. PSNR and SSIM metric values are the greater the better. "ft" refers to finetuning.

ZJU-MultiHuman

The ZJU-MultiHuman dataset consists of one video sequence with 600 frames taken from 8 uniformly distributed synchronized cameras. The video sequence was published online [54] along with the camera calibration files. The captured scene contains 4 different human

subjects with simple action poses such as standing, sitting, walking, and swapping seats. Similar to DeepMultiSyn, we predict the SMPL-X models and segmentation masks utilizing [54, 35].

ZJU-MoCap

Our method can work with any arbitrary number of humans in the scene. For that reason, we utilize a subset of the ZJU-MoCap dataset [45] which is a single-human view synthesis benchmark consisting of 10 human scenes captured from 23 synchronized cameras in order to increase the diversity in the human subjects used for training. We rely on 5 human scenes that have their pre-computed 3D body priors and masks available.

4.4.2 Training Details

Our model is implemented using Pytorch. It is trained using the Adam [30] optimizer with a decaying learning rate that starts at 5e-4 and decays by 0.1 every 300 epochs. We sample 1,024 rays per image during training from within the bounding box of the humans in the scene. We optimize our network on a single Nvidia V100 GPU with 32 GB RAM.

Regarding the train-test splits, the DeepMultiSyn training split is comprised of 301 frames on average per scene for each camera view after excluding the 18% inaccurate frames. This sums up to 5,418 frames for three scenes and 6 camera views. The ZJU-MoCap training split consists of 50 frames on average per scene for each camera view, which sums up to 3,450 frames for three scenes and 23 camera views. The ZJU-MultiHuman training split consists of 400 frames for each camera view which sums up to 3,200 frames for 8 camera views. Therefore, the total number of training frames is around 12,068 frames. The training was done with three input views chosen randomly for each frame.

Each generalization setting has its own test split. Two target camera views are fixed for testing on the DeepMultiSyn and the ZJU-MoCap dataset, while three views are fixed for the ZJU-MultiHuman dataset. In all settings, three camera views are fixed as input for all the datasets. The "Seen Models, Seen Poses" setting is tested on the same training poses which sum up to 1,806 and 1,200 testing frames on all the testing views for the DeepMultiSyn and the ZJU-MultiHuman datasets, respectively. The "Seen Models, Unseen Poses" is tested on novel poses which sum up to 846, 750, and 916 frames for the DeepMultiSyn, ZJU-MultiHuman, and ZJU-MoCap datasets, respectively. Lastly, the "Unseen Models, Unseen Poses" setting is tested on novel subjects and poses which sum up to 242 and 263 testing frames on the DeepMultiSyn and ZJU-MoCap datasets, respectively.

Figure 4.7: Comparison with generalizable NeRF methods on **seen models** and **unseen poses** for the DeepMultiSyn Dataset. We include the top two performing generalizable methods, NHP [32] and IBRNet [72], in the qualitative comparison.

4.4.3 Baselines

We compare our proposed approach with generalizable and per-scene NeRF methods that are human and non-human based.

Comparison With Generalizable NeRF Methods

Generalizable human-based NeRF methods [32, 85] operate only on scenes with single humans. For a fair comparison, we adjust the NHP [32] method to work on multi-human scenes. We make use of the per-human segmentation masks to render a separate image for each individual in the scene. We then superimpose the human images based on their depth to render the novel view image. Regarding non-human methods, PixelNeRF [80] is the first NeRF method to incorporate novel scene generalization by conditioning NeRF on pixel-aligned features. IBRNet [72] merges concepts from image-based rendering and NeRF to aggregate the sparse multi-view information. SRF [9] utilizes stereo correspondences in the input images along with NeRF to perform well on unseen scenes. All methods were trained on all the human scenes of the datasets simultaneously.

Figure 4.8: Comparison with per-scene NeRF methods on **seen models, and seen poses** for the DeepMultiSyn Dataset. The red boxes highlight areas where our method is better at representing the texture details compared to L-NeRF [55].

Comparison With Per-scene Methods

The first baseline is the multi-human layered scene representation approach [55], denoted as L-NeRF. We reimplemented [55] since the code was not publicly available at the submission

GT Ours NHP IBRNet

Figure 4.9: Comparison with generalizable NeRF methods on **seen models** and **unseen poses** for the ZJU-MultiHuman Dataset.

Method	(a) Seen Models, Unseen Poses		(b) Unseen Models, Unseen Poses	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
NHP	26.19	0.869	24.63	0.872
Ours	**28.01**	**0.905**	**25.36**	**0.886**

Table 4.6: Performance evaluation on single-human scenes on the ZJU-MoCap dataset.

time. We also compare against D-NeRF [47] and the original NeRF [42] method. All of the mentioned approaches are trained on each scene separately.

4.4.4 Experimental Results

Our evaluation spans three settings that test different degrees of generalization as follows:

Seen Models, Seen Poses

In this setting, we test on the same human subjects and poses that the model is trained on. Table 4.5a indicates the results in terms of the per-scene and generalizable baselines. Regarding the generalizable approaches, our method exhibits the best overall performance on both datasets on all metrics. Figure 4.8 shows the qualitative comparison between our method and the per-scene NeRF methods on seen models and seen poses. The highlighted areas indicate that our proposed method is capable of representing more texture details compared to L-NeRF [55]. D-NeRF [47] shows highly blurred results that impact the representation of the main human features, while NeRF [42] cannot handle dynamic scenes, hence, it renders the average of all the training frames.

cross_att	self_att	rgb_att	# V.	PSNR ↑	SSIM ↑
			3	20.92	0.7860
✓			3	21.75	0.8045
✓	✓		3	21.98	0.8093
✓	✓	✓	3	**22.19**	**0.8260**
✓	✓	✓	1	20.48	0.7800
✓	✓	✓	2	21.47	0.8060
✓	✓	✓	4	22.42	0.8316

Table 4.7: Ablation study results on **seen models** and **unseen poses** for the DeepMultiSyn dataset. "# V." denotes the number of views.

GT Ours L-NeRF

Figure 4.10: Comparison with a per-scene multi-human method [55] on **seen models**, and **unseen poses** on the DeepMultiSyn Dataset.

Pose Generalization

We additionally test all approaches on the same human subjects seen during training, but with novel poses. L-NeRF is a human-based method that generalizes to novel poses, therefore, it is included in this comparison. On both datasets, Table 4.5b shows that our proposed approach highly outperforms all the generalizable NeRF methods on all metrics. Our performance is relatively close to L-NeRF on the ZJU-MultiHuman dataset as the novel poses have a similar distribution to the training poses. On the other hand, L-NeRF lags behind our method on the DeepMultiSyn dataset due to the complex novel poses available which validate the pose generalization ability of our method on challenging motions. Qualitatively, Figures 4.7 and 4.9 show that IBRNet fails to properly model the full body of the human subjects. NHP can moderately render each individual subject solely, however it fails to represent areas of occlusions where subjects highly overlap. On the other hand, our method successfully models the body shapes and can handle overlapping areas which vali-

Figure 4.11: Qualitative comparison on **seen models, unseen poses** on the DeepMultiSyn dataset. We include the results of our proposed approach using a single input view and compare it to the generalizable NeRF methods that take 3 views as input.

Figure 4.12: Qualitative comparison on **unseen models, unseen poses** on the ZJU-MoCap dataset.

date the effectiveness of the layered scene representation in the generalizable multi-human setting. Figure 4.10 shows how L-NeRF fails to properly render the appearance of subjects when presented with complex unseen poses. Figure 4.11 shows the results of our proposed approach using a single input view compared to the generalizable NeRF methods that take 3 views as input for seen models, and unseen poses. Our single-view results show better performance than NHP in the overlapping areas among persons, while also representing the main features of the human subjects better than SRF [9] and IBRNet [72]. Our 3-view results show enhancements by removing some appearance artifacts in the rendered image.

Figure 4.13: Comparison with generalizable NeRF methods on **unseen models** and **unseen poses** for the DeepMultiSyn dataset.

Human Generalization

A challenging setting would be to test on human subjects and poses that were not seen during training. This was only done on the DeepMultiSyn dataset by leaving out one scene for testing. Table 4.5c validates that our method has the best generalization capability as it outperforms all other methods by a large margin. 4.13 shows that our method better represents the main body features of the novel human subjects. IBRNet fails to fully render some body parts like the legs, while NHP suffers from more blur artifacts, especially in overlapping areas.

4.4.5 Ablation Studies

Performance on Single-Human Scenes

We evaluate our performance on single-human scenes on the ZJU-MoCap dataset compared to the state-of-the-art method, NHP. The method was trained on the single and multi-human training split detailed in the supplementary material. Table 4.6a shows that our method surpasses NHP by a large margin on pose generalization on both metrics. For

the human generalization in Table 4.6b, we also show a noticeable enhancement which demonstrates the effectiveness of our method in handling scenes with as low as one human subject. Figure 4.12 demonstrates a qualitative comparison with NHP [32] on single-human scenes for unseen human subjects on the ZJU-MoCap dataset. Our method shows fewer overall appearance artifacts in the rendering results.

Effect of Fusion modules

We assess the effect of different fusion modules on the synthesis results. From Table 4.7, the second row uses the cross-attention module (cross_att) in Section 3.2.4 and it shows a noticeable improvement over doing basic average pooling in the first row. This indicates the effectiveness of the correlation learned between the vertex and image features. The addition of the self-attention module (self_att) in Section 3.2.4 in the third row led to the incorporation of multi-view aware features and achieved a slight enhancement on both metrics. The last row adds the raw RGB fusion module (rgb_att) in the Color Network presented in Section 3.2.5. It enhances the performance, especially on the SSIM metric, validating the importance of utilizing low-level information.

Effect of Number of Views

We evaluate the performance of our proposed approach when given a different number of input views at test time. Table 4.7 indicates that using 4 views leads to an enhancement in both metrics due to the extra information available. Decreasing the number of views gradually degrades the performance. However, using only one input view, our method outperforms all the generalizable NeRF methods in Table 4.5 that use 3 input views.

4.4.6 Discussion

One of the main contributions of this method revolves around the problem that is being solved, how non-trivial it is, and how existing literature work does not sufficiently solve the problem which is creating a generalizable multi-human view synthesis method that works with very sparse input views.

Existing generalizable human-based methods [32, 85] cannot be extended directly to multi-human settings. Comparisons with NHP [32] for multi-human scenes in Figure 4.7 show our superior performance where NHP fails to render overlapping areas of humans. Regarding layered scene representations[55, 83], they are mainly constrained to the per-scene

training settings and cannot operate properly on novel subjects/poses. The comparison with L-NeRF [55] in Figure 4.10 shows how they fail to properly render the human subjects when complex novel poses are given, unlike our method. Another drawback of existing methods is demanding a higher number of input views. ST-NeRF [83] requires 16 input views and can only render a 180° viewing range. L-NeRF utilizes 8 input views, while our method is designed to work with 3 input views and achieves 360° free-viewpoint rendering. We even provide adequate quality results with as low as 1 input view, as shown in Figure 4.11. As a result, a clear research gap exists in the literature for having a high-quality sparse-view multi-human method that requires no per-scene training and we offer an effective solution to fill the gap which highlights our contribution and position in the literature work.

Existing attention-aware feature fusion methods suffer from limited utilization of cross-view information and low-level frequency details. Our proposed approach utilizes a novel and unique collection of three attention modules (cross-att, self-att, and rgb-att) to generate view-aware and pixel-aware human features augmented with encoded low-level RGB values for retaining high-frequency details. They jointly allow our method to have superior multi-human performance compared to the baseline methods. The enhancement over other collections of modules was also fairly proved for single-human settings in Table 4.6 against NHP [32], which uses temporal and multi-view attention modules.

Regarding the per-scene NeRF methods, Table 4.5 show that NeRF [42] exhibits a significantly low performance since it cannot handle dynamic scenes. D-NeRF [47] has a dynamic object modeling ability, yet suffers in representing complex human motions leading to degraded performance. Our proposed method performs at par with the state-of-the-art per-scene baseline (L-NeRF [55]), while effectively saving computational and time resources. Specifically, L-NeRF takes a total of 144 hours to converge on all three scenes one at a time, while our method needs a total of 50 hours to converge on all the scenes simultaneously. After per-scene finetuning, our method surpasses L-NeRF on both datasets. Our method is more robust to SMPL-X inaccuracies due to the usage of low and high-level image features that complement mesh errors.

4.5 Summary

In this chapter, we discuss the experimental setup and results for evaluating the analysis criteria of single-view MINE [34]. Regarding performance, the results of training on the ShapeNet dataset [4] showed the inability of single-view MINE to render objects from far away views, unlike the implicit 3D scene representation method, pixelNeRF [80]. The

ablation studies on NeRF concepts proved the effectiveness of volumetric rendering for enhanced synthesis results, whereas stratified sampling did not contribute positively to the overall performance. Additionally, we showed the weak generalization ability of single-view MINE to novel scenes, especially on the LLFF [41] dataset. We also proved the significant inference efficiency of MINE compared to pixelNeRF. Moreover, the experiments for the proposed MV-MINE architecture show the effectiveness of utilizing the attention-based pre-decoder fusion for high-quality results compared to other proposed modules. We also show comparable performance with state-of-the-art implicit and explicit novel view synthesis baseline methods.

Regarding the GenLayNeRF architecture, we discussed the details of the datasets proposed and the baseline methods. Experimental results showed that our method outperforms state-of-the-art generalizable NeRF methods in different generalization settings and performs at par with layered per-scene optimization methods on all metrics without requiring long per-scene optimization runs and high computational resources. The ablation studies highlighted the superior performance of our method on single-human scenes compared to NHP [32] and the effectiveness of the proposed attention modules to enhance the synthesis results. In the next chapter, we will provide a detailed discussion of the limitations of our proposed approaches along with the possible future research directions.

Chapter 5

Conclusion

In this thesis, we went over the existing literature work in the field of novel view synthesis. We examined the classical view synthesis approaches and presented learning-based approaches based on 3D scene representations. Explicit 3D approaches directly model the camera frustum through different representations, allowing better modeling of occluded areas. Multi-plane images (MPI) [67] are explicit representations that are composed of parallel RGB-α planes that can be warped and projected to render novel views. However, MPI has the drawback of incomplete 3D scene representation due to the discretization of the depth of the planes. Layered Depth Images (LDI) [52] are a more memory and space-efficient explicit representation that allows each pixel to have an arbitrary number of layers at different depths. Implicit 3D representations model the 3D scene structure within the weights of neural networks. These representations can be classified as per-scene optimization methods that require re-training for novel scenes, and generalizable approaches that can handle unseen scenes during inference. There are also human-based approaches that handle the complexity of human subjects in terms of deformations and self-occlusions. Recent methods propose a combination of implicit and explicit representations, either on a pixel or object level. The pixel-level combination takes the form of multi-plane neural radiance fields (MINE) [34], while the object-level combination represents each object in the scene with an independent neural radiance field [55]. Finally, we provided an overview of the different attention mechanisms available in computer vision tasks.

Furthermore, we tackled several challenges with regard to novel view synthesis which includes handling occluded areas, expanding the viewing direction range, avoiding in-efficient per-scene optimization settings, and representing complex multi-human scenes. This was done by exploring the capabilities of combining explicit [67, 55] and implicit [72, 32] 3D scene representations in the form of scene layers at the pixel level or the object level. For

the pixel-level representations, we presented an in-depth technical analysis of single-view MINE to evaluate their boundaries in terms of performance, generalization, and efficiency. Performance was evaluated through training on a novel challenging dataset [4] and comparing the effect of different rendering and sampling techniques borrowed from NeRF [42] on the quality of the results. Generalization was assessed by evaluating the network on novel scenes from the KITTI Raw [18] and LLFF [41] datasets. Efficiency was assessed through a quantitative time comparison with pixelNeRF [80] on both GPU and CPU. We concluded from our experiments that single-view MINE demonstrated weak generalization and a small viewing direction range, which might be due to single-view input or homography warping, although it had faster inference times than pixelNeRF [80]. We also deduced that volumetric rendering plays a more important role than stratified sampling in achieving better results for single-view MINE. Furthermore, we proposed a novel multi-view MINE architecture, MV-MINE, that utilizes a novel attention-based module to effectively fuse multi-view features and enhance the synthesis quality for MINE. Experiments indicated that our proposed attention module performs better than other proposed fusion techniques. In addition, our method shows competitive performance compared to baseline novel view synthesis approaches. One main limitation of the proposed MV-MINE method is the reliance on homography warping to render the novel views. Even though we proved the strength of utilizing neural radiance planes along with volumetric rendering to predict the novel views. The requirement of having a rigid warping mechanism between the source and target planes still limits the viewing direction range to be rendered as warping becomes more ill-posed whenever the distance between the source and target views increases. Future work could look into ways to render novel views from the predicted MPIs directly without the need to carry out warping from source to target views.

As to the object-level representations, we introduced a generalizable layered scene representation, GenLayNeRF, for the free-viewpoint rendering of multi-human scenes using very sparse input views while operating on unseen poses and subjects without test time optimization. We divide the scene into a set of multi-human layers and generate multi-view image features and human-anchored features. We then utilize a combination of cross-attention and self-attention modules that effectively fuse the information seen from different viewpoints. In addition, we introduce an RGB fusion module to embed low-level pixel values into the final color prediction for higher-quality results. We assess the efficacy of our approach on two newly proposed multi-human datasets. Experimental results show that our method outperforms state-of-the-art generalizable NeRF methods in different generalization settings and performs at par with layered per-scene optimization methods on all metrics without requiring long per-scene optimization runs and high computational resources. Our proposed method has the potential of performing direct inference on novel

human subjects to suit real-world applications when trained on larger datasets. Several enhancements to our proposed method could be investigated further. As our two proposed datasets were sufficient to show the generalization capability of our method, there is room for improvement by elevating the diversity in terms of the number of scenes, camera views, distinct humans, and complex actions. This would lead to better generalization capabilities on broader challenging scenarios. Furthermore, inaccuracies in the estimation of the SMPL-X models highly hinder performance. A possible research direction could explore the optimization of the SMPL-X parameters as part of the network training. Lastly, our method suffers from blur artifacts in the representation of human clothing details. One could experiment with integrating a deformation model to represent small deformations such as textured clothing.

References

[1] Kara-Ali Aliev, Dmitry Ulyanov, and Victor S. Lempitsky. Neural point-based graphics. *ArXiv*, abs/1906.08240, 2020.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

[3] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM SIGGRAPH 2003 Papers*, 2003.

[4] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015.

[5] Gaurav Chaurasia, Sylvain Duchêne, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Trans. Graph.*, 32:30:1–30:12, 2013.

[6] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. *ArXiv*, abs/2112.04312, 2021.

[7] Shenchang Eric Chen. Quicktime vr: an image-based approach to virtual environment navigation. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995.

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

[9] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34:1 – 13, 2015.

[11] Paul E. Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[12] Paul E. Debevec, Camillo Jose Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.

[13] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognit. Lett.*, 125:333–340, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[15] Mingsong Dou, S. Khamis, Yu.G. Degtyarev, Philip L. Davidson, S. Fanello, Adarsh Kowdle, Sergio Orts, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4d: real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35:114:1–114:13, 2016.

[16] John Flynn, Michael Broxton, Paul E. Debevec, Matthew DuVall, Graham Fyffe, Ryan S. Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2362–2371, 2019.

[17] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *ArXiv*, abs/2012.05903, 2020.

[18] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[20] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3827–3837, 2018.

[21] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.

[22] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9224–9232, 2018.

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), nov 2019.

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[25] Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. Casual 3d photography. *ACM Transactions on Graphics (TOG)*, 36:1 – 15, 2017.

[26] Peter Hedman and Johannes Kopf. Instant 3d photography. *ACM Transactions on Graphics (TOG)*, 37:1 – 12, 2018.

[27] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel J. Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (TOG)*, 37:1 – 15, 2018.

[28] Sascha Hilgenfeldt, Michael P Brenner, Siegfried Grossmann, and Detlef Lohse. Analysis of rayleigh-plesset dynamics for sonoluminescing bubbles. *Journal of Fluid Mechanics*, 365:171–204, 1998.

[29] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[31] Wayne Kreider, Lawrence A Crum, Michael R Bailey, and Oleg A Sapozhnikov. A reduced-order, single-bubble cavitation model with applications to therapeutic ultrasound. *The Journal of the Acoustical Society of America*, 130(5):3511–3530, 2011.

[32] Youngjoon Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *NeurIPS*, 2021.

[33] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[34] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021.

[35] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[36] Tianye Li, Miroslava Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, S. Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *ArXiv*, abs/2103.02597, 2021.

[37] Detlef Lohse. Bubble puzzles. *Physics Today*, 56(2):36–41, 2003.

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[39] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, D. Salesin, William T. Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *ACM Transactions on Graphics (TOG)*, 39:1 – 14, 2020.

[40] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhu Chen, and Yawei Luo. Attention-aware multi-view stereo. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1587–1596, 2020.

[41] 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. *ACM Transactions on Graphics (TOG)*, 2019.

[42] 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 *European conference on computer vision*, pages 405–421. Springer, 2020.

[43] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *https://arxiv.org/abs/2011.12948*, 2020.

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9050–9059, 2021.

[46] Thomas K. Porter and Tom Duff. Compositing digital images. *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, 1984.

[47] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *https://arxiv.org/abs/2011.13961*, 2020.

[48] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[49] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020.

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[52] Jonathan Shade, Steven J. Gortler, Li wei He, and Richard Szeliski. Layered depth images. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998.

[53] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8025–8035, 2020.

[54] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Easymocap - make human motion capture easier. Github, 2021.

[55] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[56] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing*, 2000.

[57] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2441, 2019.

[58] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *ArXiv*, abs/1906.01618, 2019.

[59] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane

images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–184, 2019.

[60] Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2262–2270, 2017.

[61] Carsten Stoll, Juergen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM SIGGRAPH Asia 2010 papers*, 2010.

[62] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020.

[63] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J. Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *European Conference on Computer Vision*, 2018.

[64] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32:45–61, 1998.

[65] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[66] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. *arXiv preprint arXiv:2010.04595*, 2020.

[67] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.

[68] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *European Conference on Computer Vision*, 2018.

[69] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[70] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[71] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[72] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4688–4697, 2021.

[73] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[74] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Marc Levoy, and Mark Horowitz. High-speed videography using a dense camera array. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2:II–II, 2004.

[75] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1679–1688, 2020.

[76] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9416–9426, 2021.

[77] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *ArXiv*, abs/1612.00814, 2016.

[78] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal of Computer Vision*, 128(1):53–73, Aug 2019.

[79] Seung-Uk Yoon, Eun-Kyung Lee, Sung-Yeol Kim, and Yo-Sung Ho. A framework for multi-view video coding using layered depth images. pages 431–442, 11 2005.

[80] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4576–4585, 2021.

[81] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[82] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[83] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40:1 – 18, 2021.

[84] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[85] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei-Ying Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. *ArXiv*, abs/2112.02789, 2021.

[86] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6219–6229, 2021.

[87] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ArXiv*, abs/1805.09817, 2018.