Pedestrian Intent Estimation Through Visual Attention and Time and Memory Conscious U-Shaped Networks for Training Neural Radiance Fields

by

Abiramy Kuganesan

B.A.Sc., University of Waterloo, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

October 2022

© Abiramy Kuganesan, 2022
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

**Pedestrian Intent Estimation Through Visual Attention and Time and Memory Conscious U-Shaped Networks for Training Neural Radiance Fields**

submitted by Abiramy Kuganesan in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

**Examinining Committee:**

Helge Rhodin, Assistant Professor, Department of Computer Science, UBC  
*Supervisor*

Alan Kingstone, Professor, Department of Psychology, UBC  
*Supervisory Committee Member*

James J. Little, Professor, Department of Computer Science, UBC  
*Supervisory Committee Member*

Renjie Liao, Professor, Department of Electrical and Computer Engineering, UBC  
*Supervisory Committee Member*
Abstract

When people cross the street, they make a series of movements that are indicative of their attention, intention, and comprehension of the roadside environment. These patterns in attention are linked to individual characteristics which are often neglected by autonomous vehicle prediction algorithms. We make two strides towards more personalized pedestrian modelling. First, we design an outdoor data study to collect behavioural signals such as pupil, gaze, head, and body orientation from an ego-centric human point of view. We gather this data over a range of diverse variables including age, gender, geographical context, crossing type, time of day, and the presence of a companion. In order for simulation engines to be able to leverage such dense data, efficient 3D human and scene reconstruction algorithms must be available. The increased resolution and model-free nature of Neural Radiance Fields for large scene reconstruction and human motion synthesis come at the cost of high training times and excessive memory requirements. Little has been done to reduce the resources required at training time in a manner that supports both dynamic and static tasks. Our second contribution takes the form of an efficient method which provides a reduction of the memory footprint, improved accuracy, and reduced amortized processing time both during training and inference. We demonstrate that the conscious separation of view-dependent appearance and view-independent density estimation improves novel view synthesis of static scenes as well as dynamic human shape and motion. Further, we show that our method, UNeRF, can be used to augment other state-of-the-art reconstruction techniques to further accelerate and enhance the improvements which they present.
Lay Summary

Research has shown that there are distinct patterns of variation in crossing behaviour across different classes of pedestrians. Factors such as age, gender, geographical context, crossing type, time of day, presence of a companion, and much more may contribute to individual crossing behaviour. Often, autonomous vehicle algorithms consider pedestrians as a uniform entity when predicting their future trajectory. We attempt to break this bias by introducing a dataset which highlights the behavioural patterns across several of the aforementioned factors. We also present a memory and time efficient method for training models which can reconstruct the scene and human motion from videos such as the ones gathered in this dataset.
Preface

The written contents of this thesis are the original work of Abiramy Kuganesan under the guidance of Helge Rhodin, Alan Kingstone, and James J. Little. All software and all figures presented in this work were created by Abiramy Kuganesan except where otherwise noted. All experiments were implemented and performed by Abiramy Kuganesan with the help of hardware resources provided by Compute Canada. The Pupil Labs hardware used in this work was made available by the Brain, Attention, and Reality (BAR) Lab. The Protruly V10s device, which made the 360 video capture possible, was generously provided by Robert Xiao. The data collection study presented in this work is covered by the Certificate of Ethical Approval: Harmonized Minimal Risk Behavioural Study which holds an REB Number of H10-00527.
# Table of Contents

Abstract ................................................................. iii

Lay Summary ............................................................ iv

Preface ................................................................. v

Table of Contents .................................................... vi

List of Tables ........................................................ ix

List of Figures ......................................................... xi

Glossary ............................................................... xv

Acknowledgments ..................................................... xvii

1 Modelling Pedestrians in a 3D Environment from Video ............ 1

2 PIETVA: Pedestrian Intent Estimation Through Visual Attention .... 2
   2.1 Introduction .................................................... 2
   2.2 Related Work .................................................. 4
      2.2.1 Attention as a Precursor to Action ....................... 4
      2.2.2 Factors Influencing Pedestrian Crossing Behaviour .... 5
      2.2.3 Evaluation Strategies in Socio-Psychological Studies ... 7
      2.2.4 Real Life Gaze Allocation Studies ....................... 8
      2.2.5 Existing Datasets ......................................... 8
<table>
<thead>
<tr>
<th>Section</th>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3</td>
<td>Method</td>
<td>10</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Study Overview</td>
<td>10</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Ethics Approval</td>
<td>10</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Participants</td>
<td>11</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Stimuli and Design</td>
<td>12</td>
</tr>
<tr>
<td>2.3.5</td>
<td>Apparatus</td>
<td>12</td>
</tr>
<tr>
<td>2.3.6</td>
<td>Procedure</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>Results</td>
<td>16</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Dataset Overview</td>
<td>16</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Data Transfer</td>
<td>17</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Data Processing</td>
<td>17</td>
</tr>
<tr>
<td>2.5</td>
<td>Discussion</td>
<td>20</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Limitations</td>
<td>20</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Future Work</td>
<td>20</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusion</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>UNeRF: Time and Memory Conscious U-Shaped Network for Training Neural Radiance Fields</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>Related Work</td>
<td>25</td>
</tr>
<tr>
<td>3.2.1</td>
<td>NeRF and its Limitations</td>
<td>25</td>
</tr>
<tr>
<td>3.2.2</td>
<td>The Performance Trifecta: Memory, Time, Perceptual Quality</td>
<td>25</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Discrete Data Structure Acceleration</td>
<td>26</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Optimized Sampling Strategies</td>
<td>26</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Architectural and Hardware Optimization</td>
<td>27</td>
</tr>
<tr>
<td>3.2.6</td>
<td>Training Environment Optimization</td>
<td>27</td>
</tr>
<tr>
<td>3.3</td>
<td>Preliminary Exploration</td>
<td>29</td>
</tr>
<tr>
<td>3.3.1</td>
<td>NeRF Overview</td>
<td>29</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Frequency-based Quantization</td>
<td>31</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Capacity-Performance Trade-Offs with NeRF</td>
<td>31</td>
</tr>
<tr>
<td>3.4</td>
<td>Method</td>
<td>33</td>
</tr>
<tr>
<td>3.4.1</td>
<td>UNeRF-Conv</td>
<td>33</td>
</tr>
</tbody>
</table>
List of Tables

| Table 2.1 | Pedestrian crossing variable analysis. Inference of pedestrian behaviour from individualized and circumstantial variables. | 6 |
| Table 2.2 | Dataset overview. This table summarizes the data gathered as part of this data collection study and details the variables of interest for which this study was designed. | 16 |
| Table 3.1 | Performance summary on the Blender dataset. Images are rendered at 400 \( \times \) 400 resolution. Both UNeRF variants deliver comparable or even better image quality with improved memory consumption. UNeRF-Conv presents a 12\% improvement in training time. | 40 |
| Table 3.2 | Quantitative comparisons on the LLFF dataset. The following experiments were evaluated across 100k iterations. Our UNeRF-Fine-Sub-Mem baseline produces comparable results to NeRF with a slight performance reduction. UNeRF-Conv struggles with the reconstruction of this data. | 41 |
| Table 3.3 | Quantitative comparisons on rendering dynamic human motion from the SURREAL dataset. Images are rendered at 512 \( \times \) 512 resolution. UNeRF-Conv yields better image quality as well as memory consumption. | 43 |
Table 3.4  **Quantitative performance metrics of memory-limited NeRF baselines on the Blender dataset.** These models were evaluated across 300k iterations. Reduction of the importance samples results in the least loss of visual quality and the fastest time per iteration for a relatively low maximum memory allocation. 44

Table 3.5  **Ablation study on convolution kernel size on the Blender dataset.** Images are rendered at 400 × 400 resolution. The larger UNeRF-Conv kernel size improves rendering outcomes. 45

Table 3.6  **Kernel Size Ablation Study of UNeRF-Conv on the SURREAL Test Dataset using A-NeRF.** These experiments are evaluated across 150k iterations. UNeRF-Conv with a kernel size of 3 × 1 outperforms kernel sizes of 2 and 1 on the SURREAL dataset. 45

Table 3.7  **Larger kernel size ablation study of UNeRF-Conv on the Blender Lego test dataset.** These experiments are evaluated on a single NVIDIA GeForce RTX 3090 GPU. UNeRF-Conv with a kernel size of 3 × 1 provides the best tradeoff between performance, time, and memory when compared to the larger kernel sizes. 46

Table 3.8  **Ablation study on interpolation mechanisms with UNeRF-Sub on the Fern LLFF dataset.** The following experiments are evaluated across 100k iterations. UNeRF-Sub with our position-aware interpolation mechanism achieves a benchmark VAL PSNR faster than the two simpler baselines. 46

Table 3.9  **Ablation study on interpolation mechanisms with UNeRF-Conv on the Blender dataset.** The following experiments were evaluated across 100k iterations. Our position-aware interpolation strategy outperforms nearest neighbour and average interpolation on most datasets. 47

Table 3.10  **Application of UNeRF-Conv to Mip-NeRF.** These results are evaluated on a single NVIDIA GeForce RTX 3090 GPU. With the UNeRF-Conv optimization on Mip-NeRF, we achieve a much better trade-off between performance, time, and memory. 48
List of Figures

Figure 2.1  **Endsley model.** Adaptation of the Endsley Model of situational awareness which relates the performance of actions through decision making informed by situational awareness at three levels: perception, comprehension, and projection [11].

Figure 2.2  **Participant statistics.** The distribution of participants by gender and age is shown on the left. Participant responses to the question "Which continent were you raised in?" is illustrated on the right.

Figure 2.3  **Route overview.** The study route is annotated with the different types of crossings that the participant is expected to incur. Google Maps estimates that this route should take approximately 10 minutes to complete.

Figure 2.4  **Pupil labs camera footage processing.** The Pupil Player interface shows a participant approaching the first zebra crossing on the study route. The magenta line shows the detected scanpath in the previous few milliseconds. The individual green circles show individual gaze positions superimposed on the scene. The top left window shows the eye camera which performs pupil (red/orange) and eye (blue) detection.
Figure 2.5 **Data calibration and handling.** Left: The pupil data panel where the eye and pupil detection may be performed. Centre: The gaze calibration panel where the pupil direction and position is aligned with the scene through the use of the calibration targets. Right: The 360° footage captured during the distracted pedestrian section of the study.

Figure 3.1 **NeRF architecture.** The NeRF network architecture is composed of 8 fully connected layers and subsequent ReLU activation functions. The blue boxes convey input vectors which are concatenated with their respective positional encoding. An initial input vector of 3D sample points concatenated with their respective positional encoding \( \gamma(x) \) are fed into the network. This positional encoding \( \gamma(x) \) is also concatenated with the output of layer 4 before being passed into layer 5. The portion of the network from layers 1 - 8 inform the density function \( \sigma(x) \). The positional encoding corresponding to view direction is concatenated with the output of layer 8 to inform the view-dependant RGB values at each position. [35].

Figure 3.2 **UNeRF architecture.** UNeRF improves computational and memory efficiency by sharing feature computations between adjacent 3D samples. Starting with positional encoded sample points, the first 5 layers process and output the feature vector with progressively smaller spatial resolution (Sections 3.4.1 and 3.4.2). We then apply position-aware linear interpolation (Section 3.4.3) to upsample and add the interpolated features to the earlier ones to recover the feature in full spatial resolution, analogous to UNet [45].
Feature vector manipulation. Layers 0 and 1 process the features (Figure 3.3a) corresponding to each sample along the ray. Downsampling using convolutions (Section 3.4.1) or sub-sampling (Section 3.4.2) results in a downsampled feature set (Figure 3.3b) that corresponds to a sparser set of samples. Position-aware linear interpolation (Section 3.4.3) brings the feature set back to its original resolution through iterative application. The interpolated feature set (Figure 3.3c) accounts for the features for samples that were omitted from MLP invocations.

Intermediate feature vectors are interpolated using a linear approximation between the two adjacent anchor feature vectors queried at the intermediate point.

Qualitative comparisons on the Blender dataset. Top row: Rendered testset images from UNeRF-Conv with a $3 \times 1$ kernel. Bottom row: Difference maps between UNeRF-Conv, NeRF and the ground truth. Pixels highlighted more prominently in blue indicate regions where UNeRF-Conv more closely aligns with the ground truth than NeRF, and red the vice versa. The pixel color difference is normalized to $[0, 1]$. While NeRF and UNeRF-Conv typically exhibits comparable results when describing low frequency details, UNeRF-Conv has an edge on depicting high frequency detail more accurately.

Qualitative results of UNeRF on the LLFF dataset. Top row: novel views rendered by NeRF at 100k iterations. Bottom row: novel views rendered by UNeRF-Fine-Sub-Mem at 100k iterations. Our method produces comparable results to NeRF while consuming less time and memory during training time.

Dynamic human motion rendering on the SURREAL dataset. The 3D human pose and view are both unseen during training time. UNeRF-Conv preserves better finger structure than the other architectures and achieves better results than the full capacity NeRF baseline when fully trained.
Figure A.1  **Study protocol.** The study protocol serves as a quick checklist highlighting all the key steps of the study. . . . . . . . . . . . . 64

Figure A.2  **Questionnaire QR code.** The QR code used to share the questionnaire with participants. This allows the participants to perform the questionnaire on their own device with ease. . . . . . 65

Figure A.3  **Demographic questionnaire.** This questionnaire is administered to participants after their participation in the PIETVA data collection study. . . . . . . . . . . . . . . . . . . . . . . 66

Figure A.4  **Study debriefing.** The debriefing form outlines a brief description of the prior art, the objective of the study, including the hypothesis and study variables, the nature of the experiment, and relevant contact information. . . . . . . . . . . . . 67
Glossary

AUTOINT automatic integration

DERF decomposed radiance fields

FPS frames per second

GPU graphics processing unit

HSP human subject pool

JAAD joint attention in autonomous driving

LPIPS learned perceptual image patch similarity

MLP multilayer perceptron

ND neutral density

NERF neural radiance fields
NSVFS  neural sparse voxel fields

OS  operating system

PIE  pedestrian intent estimation

PIETVA  pedestrian intent estimation through visual attention

PSNR  peak signal to noise ratio

SSIM  structural similarity index measure

TAMNATAR  toward a more natural approach to attention research

UI  user interface
Acknowledgments

I would like to express my sincere gratitude to everyone who supported me intellectually, emotionally, and financially throughout the past two years. First and foremost, I am indebted to my supervisor Dr. Helge Rhodin whose humble expertise, invaluable encouragement, and endless enthusiasm to extend lab experiences beyond pure research, have created such a welcoming environment. I must thank Dr. Alan Kingstone for the countless hours of discussion, experience-driven feedback, connections to experts in the field, and access to resources which made the Psychology and Computer Science collaboration possible. I am extremely grateful to have collaborated with Dr. James J. Little whose positivity, research insights, and willingness to make time for me even after his retirement have motivated my desire to pursue academia further. Thank you to Dr. Renjie Liao who offered his time and insight in being the second reader for this thesis.

I must also extend many thanks to Dr. Robert Xiao for offering me with access to hardware without which we could not have gathered such meaningful data, Dr. Amir Rasouli for spearheading the social pedestrian intent research domain especially within Engineering and Computer Science, Dr. Nicola Anderson for answering all of my questions regarding the psychology of it all, and Shih-yang Su for serving as my NeRF encyclopedia. My labmates Frank Yu, Daniel Ajisafe, Chunjin Song, Xingzhe He, and Bastian Wandt were huge pillars of support during countless late nights at the lab. Dr. Laura Ballay and the DFP Community have provided an incredible support system continue to keep my human-centric design spirit alive.

This work was sponsored in part by the Natural Sciences and Engineering Research Council of Canada and the generous GPU server resources from Compute
To Manuel Oliveira, Sneha Sambandam, Julia Fast, Dominic Janus, and Brandon Chan, I thank you for being the Vancouver family that kept me sane through it all. I thank Jasmeet Munday, Varjitt Jeeva, Shuruthii Thiyagasoruban, and Raegina Jeyaranjan for always believing in me and for keeping my spirits high.

Finally, to my supportive parents Kuganesan and Thevagi, my loving sister Kashthuraey, and my beloved grandmothers Kanagambal and Thevanai, who cherished each of my little academic achievements as if they were this one, without all of your sacrifices, inspiration, and continuous support, this milestone could not have been possible. A special thank you to my dearest partner Anojan who continues to be my biggest cheerleader in all of my endeavours.
Chapter 1

Modelling Pedestrians in a 3D Environment from Video

The main goal of this thesis is to encourage better machine modelling of pedestrians in a 3D environment from video. This large open-ended research problem was approached from two angles. First, through exploring new recording modalities and the subsequent study design for curating training data for such a task (Chapter 1). Chapter 2 follows up on this by presenting a new algorithm to reconstruct the 3D world and human motion in a more resource-efficient manner. At last, Chapter 3 ties both of these themes together in an attempt to provide tangible motivation for this line of work.
Chapter 2

PIETVA: Pedestrian Intent Estimation Through Visual Attention

2.1 Introduction

Research suggests that the elimination of the human driver from the driving equation would result in the reduction of human error due to fatigue, misperception or inattention in addition to 93.5% of road accidents [59]. Algorithms supporting autonomous vehicles have evolved at an unprecedented rate over the past decade. Policymakers have remained skeptical of the feasible implementation of such research efforts—safety during the interaction of autonomous vehicles with other road agents has been a heavily debated topic between researchers and policy makers [7]. As such, it is critical for computational systems in autonomous vehicles to predict and accommodate for human behaviour in the surrounding environment.

Modelling pedestrians as points in space, bounding boxes, and skeletons is limiting in the sense that human perception and prediction of pedestrian trajectories involve consideration for several socio-psychological factors. Psychology literature has identified many relations between various classes of pedestrians and the consequentially exhibited roadside behaviour. For instance, one study found that
crossing speeds can be attributed to both gender and age. While the outcomes of this study may be biased by geographical location, the average crossing speed for males was found to be 1.85\(\text{m/s}\) compared to 1.67\(\text{m/s}\) for females. The same study found that children are more likely to cross at higher speeds whereas adults and the elderly population had no significant variance in crossing speeds [24].

Street crossings in dynamic environments involve complex decision making which requires a thorough understanding of the surrounding conditions. Situational awareness is defined as the "the perception of the elements in the environment within a volume of time and space, the comprehension of their meanings, and the projection of their status in the near future" [16]. While this phenomenon is often considered as a precursor to decision making, additional factors such as experience gained and time resource constraints often inform the ultimate decision to cross the road. From a visual perspective, much of this can be informed by human pose, head orientation, and patterns in gaze allocation [52]. These factors have yet to be integrated into road agent intent estimation algorithms in a manner that autonomous vehicles can effectively employ during inference.

Prior to deployment of such skilled algorithms on the road, it is important to gain the trust of the general public and policymakers. Simulation engines lend themselves to this task by demonstrating model behaviour without the high stakes. The barrier to modelling 3D dynamic pedestrian motion in simulation presents itself in one primary way.

There is a lack of ego-centric pedestrian data which captures human gaze allocation, head orientation, body pose, and the surrounding environment. While datasets from an upper-angle perspective [15, 64] and from vehicle-mounted cameras are available [33, 42, 51], they are limited in the duration of interactions or are unable to capture gaze location and head orientation in a manner that is necessary for effective modelling.

We address this concern through the design of a pedestrian-centric, real world data collection study to inform intention via attention. We address this as the Pedestrian Intent Estimation Through Visual Attention (PIETVA) dataset. Through detailing the design and execution of this experiment in this thesis, we stipulate that this data has the potential to inform more holistic models of pedestrian crossing behaviour.
2.2 Related Work

In the following section, we provide a socio-psychological synopsis of pedestrian behaviour through discussion of relevant methodologies and findings which contribute to the domain of Social Pedestrian Intent Estimation.

2.2.1 Attention as a Precursor to Action

Coeugnet et al. decompose situational awareness into 3 levels through discussion in the context of pedestrian crossing behaviour:

- **Level 1: Perception** This level captures perception of elements to cross such as street-crossing infrastructure, traffic, other pedestrians, etc.

- **Level 2: Comprehension** The pedestrian’s experience shapes their compre-
hension of the situation — this may take the shape of evaluating speed and distance of vehicles, determining whether their behaviour is risky, or considering time pressure, weather constraints, etc.

• **Level 3: Projection** This level factors the time to arrival, task objectives such as crossing safely or rapidly, and individual factors such as accompaniment of a child, physical limitations or strong emotions.

This work further details an observation study accompanied by an interview to understand how time pressure, emotions, and social context influence street crossing behaviour [11]. This study is guided by the Endsley model of situational awareness [16] which is presented in Figure 2.1. Building on this theme, Zito et al. performed an observational study to assess the exhibited behaviour and the role age plays in the decision to cross a street. The observational study involved a younger and older age group, measured head and eye movements in varying crossing scenarios, and assessed their behaviour. The study found that older pedestrians struggled to gauge their crossing speeds and assess the speeds of oncoming vehicles. Older pedestrians also had a tendency to look at their feet more to plan precise steps which resulted in less attention to the traffic scene. The study was conducted in a VR environment [69].

### 2.2.2 Factors Influencing Pedestrian Crossing Behaviour

This section aims to provide a brief summary of factors and variables that could improve safety and prediction around pedestrian crossing behaviour. By nature, human decision making is a relatively complex phenomenon which is shaped by environmental sensitivities, personal motives, demographic characteristics, norms attributed to geographical location, etc. As discussion regarding autonomous vehicles becomes more and more prevalent, increasing concerns have risen regarding the safety and predictability of human interactions with such autonomous agents. Perhaps the following statistic will aid in materializing the motivation that drives this analysis: of the traffic accident related deaths in India in 2013, 65% of road accident related deaths were those of pedestrians; 35% of which were children [24]. To understand the predictors that contribute to decision making in crossing behavior of individual pedestrians, it is necessary to study and understand the literature...
Table 2.1: Pedestrian crossing variable analysis. Inference of pedestrian behaviour from individualized and circumstantial variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Impact</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age lends itself to the classification of pedestrians as law-abiding (Type I) or opportunistic ones (Type II). Seniors are generally classified as law abiding pedestrians. The compliance rate is typically attributed to age, where younger pedestrians are found to be less compliant than older ones.</td>
<td>[63]</td>
</tr>
<tr>
<td>Gender</td>
<td>Males tend to exhibit more hazardous crossing behaviour than females. This finding was characterized by less waiting time shown by males over females at uncontrolled intersections prior to crossing.</td>
<td>[54]</td>
</tr>
<tr>
<td>Crossing Speeds</td>
<td>Crossing speeds can be attributed to both gender and age. The average crossing speed for males was found to be 1.85m/s compared to 1.67m/s for females. The outcomes of this study may be biased by geographical location and context. Children tend to cross at higher speeds than other categories and there is no significant variance in crossing speed between the adult and elderly population.</td>
<td>[24]</td>
</tr>
<tr>
<td>Pedestrian Group Size</td>
<td>Children have shown to be more likely to cross as a group than individually. Another study finds that drivers tend to yield to groups more than single pedestrians. Heavy pedestrian flow density is shown to encourage improper crossing activity. This allows pedestrians to gain priority at the crossing as well as motivates stragglers who would otherwise not cross in high risk situations to cross, at times, causing vehicular delays.</td>
<td>[22][9][24]</td>
</tr>
<tr>
<td>Behaviour of Other Pedestrians</td>
<td>One study finds that pedestrians are inclined to walk across a crossing at unlawful times when fewer pedestrians are waiting at crosswalk. et al. suggest that pedestrians are more inclined to violate rules when other pedestrians partake in the violation as well.</td>
<td>[46] [68]</td>
</tr>
</tbody>
</table>
as it pertains to various classes of pedestrians in controlled crossing scenarios. Table 2.1 aims to highlight studies that have identified a correlation between various individual and environmental variables to corresponding crossing behaviour.

### 2.2.3 Evaluation Strategies in Socio-Psychological Studies

Several of the socio-psychological studies present one of the following evaluation strategies for pedestrian crossing behaviour:

- **Observation using Virtual Reality** Virtual reality studies are common for gap acceptance criteria to vary the variables and see how pedestrians respond when crossing at uncontrolled intersections [69]. Virtual reality studies have been criticized for not being representative of true crossing behaviour due to biases from perceived safety concerns, skewed location, size and distance perception [23].

- **Informed Behaviour Through Questionnaires** Previous studies regarding pedestrian decisions have assessed their conformity and non-conformity in specific roadside scenarios through questionnaires [17, 68]. These studies often apply the Theory of Planned Behaviour [6] which highlights 3 main influences on an individual’s intention to engage in a particular behaviour. They include the individual’s perceived outcome (positive or negative) of engaging in the behaviour, followed by the perceived social pressure to perform the behaviour, and the perception of control over performing the behaviour [6, 17]. While questionnaire-based comprehension of decision making factors is efficient, even anonymous questionnaires may suffer from Social Desirability Bias or Response Bias where the participant responds in agreement to social ‘norms’ to please the researchers [26].

- **Observation in Urban Environments** These types of studies typically involve a drafted route with fixed crossing types. Researchers may be situated at the crossing to take notes of pedestrian behaviour. There may also be a camera mounted at the intersection to observe the pedestrian’s behaviour. Unconstrained pre-recorded clips may also be used for observation which may result in more naturalistic behaviour at the cost of an uncontrolled vari-
2.2.4 Real Life Gaze Allocation Studies

Observations in urban environments, especially as it pertains to attention, can be difficult due to the unconstrained nature of the task. One primary signal in understanding visual cues which contribute to individuals’ mental model of a scene is eye gaze. Task-based intention has been shown to embed itself in change detection [55], the processing nature of specific object features [20], and gaze itself [47]. Patterns in eye movements may be quantified through pupil and head orientation. Eye movements measured in lab environments have limited predictive power to expectations in real world environments [18] and thus it is important to study pedestrian gaze allocation in context. A naturalistic study of eye and head movements in relation to real-world gaze allocation found that when walking, head movements are typically limited to infrequent and coarse reorientation. Eye-in-head orientation is attributed to gathering necessary information for immediate action and informational demands [52].

2.2.5 Existing Datasets

Several pedestrian-centric datasets exist with various intended purposes. The Pedestrian Intent Estimation (PIE) dataset [42] offers annotations for crossing actions such as “walking”, “standing”, “looking”, “not looking”, “crossing”, and “not crossing” for a diverse range of crosswalk types. Similarly the Joint Attention in Autonomous Driving (JAAD) dataset [41] offers 240 hours worth of diverse pedestrian and vehicle interactions from Canada, Ukraine, Germany, and the USA. These 5-15s clips hone in on these interactions in urban driving settings which makes them a better fit for pedestrian tracking applications. Other datasets such as Daimler’s Pedestrian Path Prediction dataset [48] and Waymo’s Open Motion dataset [51] provide rich annotations of the surrounding environment, accompanying roadside agents, and pedestrian action types from vehicle-mounted cameras. Of remarkable note is Honda’s TITAN dataset [33] as it is a huge step towards studying personalized pedestrian characteristics and specific behaviour types such as
communicative contexts, pedestrian age categorizations, and distractive instances. These datasets are captured from a vehicle’s point of view and thus pedestrian footage is limited to a few seconds at a time—making it difficult to study pedestrian behaviour at scale.
2.3 Method

2.3.1 Study Overview

We explore whether idiosyncratic tendencies extend to how people orient their whole attention (including eye, head, and body movement) when interacting with distinct crossing types in real life. In this experiment, participants were asked to walk a route which contained two zebra crossings, two stop sign crossings, and two traffic signal crossings. Participants were asked to walk this route while accompanied by a companion, by themselves, and with a mobile device in hand. Afterward, they were asked a few questions about their age, gender identity, and the geographical context in which they grew up.

We hypothesize that people with similar age, gender, and places of origin will display similar patterns in the way they orient their body, head, and eyes when crossing at different intersections when crossing alone. Further, we hypothesize that all pedestrians will exhibit similar patterns in the way they orient their body, head, and eyes when crossing with a companion and that there will be even fewer variations in behaviour when participants a mobile device in hand.

The independent variables in this study are your age, gender, the norms associated with the geographical context within which participants were raised, as well as the crossing type (zebra, stop sign, traffic light). The dependent variables measured by this study include scanpath similarity with individuals with similar demographics, average duration of fixations, crossing speed and crossing motion type, average scan times before crossing, and potential objects in the scene that were the subject of fixation.

2.3.2 Ethics Approval

This study received ethics clearance from the UBC Behavioural Research Ethics Board under the parent study Toward a more natural approach to attention research (TAMNATAR). This individual experiment is the 199th study in this group.
2.3.3 Participants

Twenty-six participants were recruited using convenience sampling through the University of British Columbia (UBC) Department of Psychology’s human subject pool (HSP) and through Computer Science Department general advertisement. Of these participants, 18 were females and 8 were males. Our recruitment was primarily through HSP which led to a female-biased recruitment pool. We attempted to offset this by strategically recruiting within the Computer Science department. Participants predominantly originated from North America (46%) and Asia (46%) accompanied by 1 participant from South America and 1 participant from Europe. As glasses would interfere with data collection, participants, we limited the study to those who wore contact lenses or did not require glasses.

For this study, the ideal case would be to collect data from 3 different age classifications of pedestrians with a 1:1 ratio between males and females. We split our demographic into children (10-17), adults (18-64), and older adults (64+). Ideally participants would not be too close to the edges of the age spectrum. When collecting data from children, a higher than required sample size would be necessary to allow for any noisy measurements to be comfortably discarded. Considerations should be given to the senior population where any handheld apparatus may be
difficult to manage. Due to ease of access, these initial experiments were run on a subset of the adult population. For future amendments to this experiment, it would be interesting to extend to these populations.

2.3.4 Stimuli and Design

To understand how perceptual patterns can influence future action prediction, we revisit the Endsley Model (Figure 2.1) as it pertains to pedestrian roadside behaviour. We focus on how Level 1: the perception of elements in the current situation maps to Level 3: Prediction of future status and its carry through. Level 1 can be informed through a forward facing camera coupled with a camera which studies eye gaze movement. Together, this footage provides an intuition as to what the individual is perceiving in the current scene. Instances of high dwell time on certain objects in the scene and various other gaze patterns may indicate the comprehension of the scene (Level 2). Level 3, may be informed by the specific crossing action that the individual has been instructed to perform. This level of behaviour can be captured using 360° camera footage and a forward facing camera. As such, we hold the future action as a known in this study so that we can focus on the perceptual indicators that inform this status.

2.3.5 Apparatus

Pupil Labs Eye Glasses

A monocular eye gaze tracker and forward facing camera encompassed in the form of glasses were used to capture the eye movement and scene. The scene camera on the Pupil Labs glasses was highly sensitive to lighting conditions. As such, it was important to minimize the amount of light captured by recording at a frame rate of 60 frames per second (FPS). This unfortunately was not enough. We retrofitted an neutral density (ND) filter on top of this forward facing camera in order to filter out most of the light for a more comprehensive video sequence. To minimize the amount of glare on the pupil during capture, participants were asked to wear a hat depending on the weather condition. The eye tracker was connected to a Motorola phone for storing the data using the Pupil Labs mobile application.
Protruly V10s Mobile Phone

The 360° camera on the Protruly V10s mobile phone was used to capture the 360° footage of the participant during the distracted state stage. This footage allowed for the capture of the surrounding scene in addition to the participants’ head and body poses.

2.3.6 Procedure

Participants signed up for an experiment slot and were made aware of general study information on the HSP portal. Once the participants arrived at the meetup location in front of the ICICS building, they were greeted, made aware of the tasks that they would be performing, as well as provided with the consent form to read and sign (See Appendix A.1.1). The Pupil Labs glasses were fitted on the participant after receiving their consent to do so and the raw footage was analyzed on the Pupil Labs mobile application to ensure correct placement of the hardware. In cases where there was excessive sun exposure, the participant was asked to wear a hat or adjustments were made to the ND-filter. All channels on the mobile application were activated prior to hitting record. The Motorola phone was then placed in the participant’s pocket.

Calibration of the eye tracker and the forward facing camera is performed in a shaded area. A target is used to synchronize the pupil gaze with the pixel location in the forward facing capture. The participant is asked to hold their head still for the first section of calibration. During this time, the researcher holds the calibration target in various fixed positions for approximately 3s at a time and the participant is asked to follow the target with just their eyes. For the second stage aimed at head orientation calibration, the participant is asked to keep their eyes in the centre of their head and follow the target using modifications to their head orientation. The general study and calibration protocol can be found in Appendix A.1.2.

For the first stage of the experiment, the participant is accompanied by the researcher so that the participant can become familiar with the route shown in Figure 2.3 and specific crossing instructions. The second and third stages of the experiment will be a "solo crossing" and a "distracted crossing" at a randomly chosen order to isolate any bias from the order of crossing stages. Each of the 3 stages
of the experiment are preceded and followed by the calibration step. At the end of every stage, the corresponding data recording is stopped and restarted. The "solo crossing" involves the participant repeating the initial stage unattended and is aimed at capturing the case where a pedestrian crosses alone. The "distracted crossing" stage requires that the participant hold the Protruly V10s phone in their hands at all times during the crossing. The participant is asked to look at the phone occasionally, ensuring that the 360° video on the phone is continuing to record throughout the entirety of this stage. Participants are reminded that they should be aware of their surroundings and are encouraged to pay close attention to the road and their surroundings while actively crossing at intersections. This stage is intended to capture a distracted pedestrian’s behaviour.

Following these three crossing stages, the hardware is removed from the participant and all recordings are paused. The participant is then asked to complete a debriefing survey https://ubc.ca1.qualtrics.com/jfe/form/SV_6sSljU50XG4JMUu on their mobile device. The full questionnaire can also be found in Appendix A.1.3. This survey also collects anonymized demographic data which is integral to the

Figure 2.3: Route overview. The study route is annotated with the different types of crossings that the participant is expected to incur. Google Maps estimates that this route should take approximately 10 minutes to complete.
study. Once this is complete, the participant is handed a physical copy of the study debriefing (See Appendix A.1.4) and is thanked for their time and participation. It is generally recommended that a 20 minute break be provided between scheduled study slots as the hardware must be sanitized and the data must be backed up.
2.4 Results

2.4.1 Dataset Overview

This dataset contains approximately 19.5 hours of pupil data and 6.5 hours of 360° video data. The weather conditions captured in this dataset include light rain, sunny, cloudy, and partly cloudy weather conditions. The geographical context is North America during the spring with the study being hosted on the University of British Columbia Vancouver campus. A general summary of the dataset is provided in Table 2.2.

Each participant receives an identifier which is used to classify their data in an anonymous manner. This convention takes a participant number, participant gender (F for female, M for male, O for other), age classification (A for adult, C for child, E for elderly person), geographical context (AS for Asia, NA for North America, SA for South America, EU for Europe and AF for Africa), and location (C for campus) to format a string identifier such as 01FAASC.

Table 2.2: Dataset overview. This table summarizes the data gathered as part of this data collection study and details the variables of interest for which this study was designed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variants Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Adults (18-28)</td>
</tr>
<tr>
<td>Camera Modality</td>
<td>Pupil and eye (19.5h), forward facing scene (19.5h), and 360° scene and body (6.5h)</td>
</tr>
<tr>
<td>Crossing Type</td>
<td>Zebra Crossing (2), Light Controlled Intersection (2), Stop Sign (2)</td>
</tr>
<tr>
<td>Gender</td>
<td>Female (18), Male (8)</td>
</tr>
<tr>
<td>Most Familiar Geographical Context</td>
<td>Asia (12), Europe (1), North America (12), and South America (1)</td>
</tr>
<tr>
<td>Task</td>
<td>Crossing Alone, Accompanied Crossing, and Crossing in Distracted State</td>
</tr>
<tr>
<td>Time of Day</td>
<td>Morning, Afternoon, and Evening</td>
</tr>
<tr>
<td>Weather Conditions</td>
<td>Sunny, Cloudy, Partly Cloudy, Light Rain</td>
</tr>
</tbody>
</table>
Figure 2.4: Pupil labs camera footage processing. The Pupil Player interface shows a participant approaching the first zebra crossing on the study route. The magenta line shows the detected scanpath in the previous few milliseconds. The individual green circles show individual gaze positions superimposed on the scene. The top left window shows the eye camera which performs pupil (red/orange) and eye (blue) detection.

2.4.2 Data Transfer

After each study session, it is best practice to backup the data. The files should be transferred from the Motorola and Protruly V10s phones periodically to avoid data loss. Each session will carry approximately 60 minutes worth of data. The files should be deleted from the phones after verification of the backup to make space for new data.

2.4.3 Data Processing

It is recommended that the following data processing steps be performed in a Windows operating system (OS) as the software is most compatible with this OS.
Pupil Labs Glasses Footage

For processing the pupil labs glasses footage, the Pupil Player [25] application is required. Further information on the software can be found here: https://docs.pupil-labs.com/core/software/pupil-player/#visualization-plugins. The instructions for processing the pupil labs glasses footage are detailed below:

1. Open the Pupil Player desktop application and configure the application to occupy 1/2 of your screen real estate.

2. Open to file directory containing the pupil footage.

3. Drag and drop the capture file bundle which is labelled by a timestamp (i.e., 20190930035324132) onto the Pupil Player desktop application.

4. Three windows should be available as part of the Pupil Player at this time. The first window resembles a linux terminal and outputs processing logs. The second window is the main user interface (UI) which contains the controls and scene footage (see Figure 2.4). The final window shows a focused view of the eye and pupil detection.

Figure 2.5: Data calibration and handling. Left: The pupil data panel where the eye and pupil detection may be performed. Centre: The gaze calibration panel where the pupil direction and position is aligned with the scene through the use of the calibration targets. Right: The 360° footage captured during the distracted pedestrian section of the study.
5. Allow the eye and pupil detection to run in the background until it parses through the entire video. Minimize interactions with the UI at this time as this can be a resource intensive task.

6. Once this task is complete, orange points will appear on the bottom portion of the main UI where the labels Pupil Diameter 3D and Pupil Confidence reside. These points show the pupil detections throughout the capture and the confidence in the accuracy of the detection. It is evident that monocular glasses were used here as only data for the right eye is available (orange points). If a binocular camera was used, purple points would also be superimposed on this timeline. Pupil detection can be performed again using the Pupil Data panel on the right hand side of the main UI shown in Figure 2.5a.

7. The panel directly below the Pupil Data panel is the Gaze data panel. On this panel, select the Calculate all Calibrations and Mapping button. This will detect the bulls-eye calibration targets pictured in Figure 2.5b. The calibration portion of the capture will then be easily visible on the timeline at the bottom of the screen with white boxes. This process aligns the pupil direction and position with the forward facing camera to calibrate the correlation between the pupil position and the focus location in the scene.

8. Scroll down on this panel and click Calculate under the Gaze Mapper header to generate the scanpath and gaze map shown in Figure 2.4).

9. To launch the exporter, press CTRL+E. You will know when this process is complete through the messaging in the log window. This will output a video of the scene as well as fixation and gaze data.

**Protruly V10s Footage**

It is recommended that the 360° footage be opened for viewing using the Movies & TV Windows application [1]. A 360° button at the bottom right corner of the screen allows for panning and re-orientation of the scene. This allows for viewing the footage from any desired angle. Manipulation of the footage for export from a particular viewpoint may be performed using Adobe Premiere Pro [2] using the GoPro FX Reframe plugin [5].
2.5 Discussion

2.5.1 Limitations

As this study was performed at the University of British Columbia, participants were restricted to the student population which may introduce some bias in behaviour. The age range was thus limited to 18-28 based on participant availability. A broader age range would aid in making this dataset more versatile for the general learning case. While this dataset also encapsulates several different crossing types, it is performed in one specific geographical context. It is important to gather data from multiple different locations with varying pedestrian and vehicle patterns.

2.5.2 Future Work

This study would greatly benefit from being extended to the children and elderly population as outlined in Section 2.3. Additional diversity in the adult participant group including participants older than 28 and more individuals with prolonged exposure to African, South American, and European geographical norms would result in a more balanced dataset. In addition, possible extensions to this experiment could include data capture at different locations to improve data diversity. Due to time limitations, we were unable to perform any in-depth analysis of the capture, surrounding scene, and the participants head, body and pupil poses. This analysis would provide a wealth of data which could be used to inform pedestrian behavioural models which factor individual characteristics. A promising direction would be to use the collected 360° footage for 3D pedestrian and scene reconstruction as discussed in the next chapter.
2.6 Conclusion

This chapter provided an overview of an outdoor data collection study performed at the University of British Columbia Vancouver campus. We proposed a method for collecting roadside behavioural data of pedestrians from an ego-centric point of view. We discussed the pros and cons of various data collection methods in the literature and designed our study methodology to maximize the benefits of participant-driven data collection. Through controlled variables such as a fixed route, we are able to acquire data able to better understand how the pupil, gaze, head, and body inform attention during crossing actions. As with many outdoor experiments, outdoor data collection comes with a range of considerations including weather and noise from the hardware. We outlined solutions for overcoming such obstacles and executed the experiment with the help of 26 participants between the ages of 18-28. This dataset managed to span 3 different crossing types (zebra, light-controlled, stop sign) in 4 different weather conditions (light rain, cloudy, sunny, partly cloudy) during 3 different times of the day (morning, afternoon, evening). We hope that this dataset and extensions of this work may inform more personalized pedestrian roadside modelling.
Chapter 3

UNeRF: Time and Memory
Conscious U-Shaped Network for Training Neural Radiance Fields

3.1 Introduction

Various graphics primitives and strategies to support neural volumetric representations have contributed to the exponentially growing domain of model-free reconstruction and novel-view synthesis [62]. A key stepping stone in this line of research was the introduction of Neural Radiance Fields (NERF) [35] for rendering novel photorealistic views from static images for the purpose of 3D object and scene representation. Model-free scene encoding with a neural network succeeds without spatial discretization, thereby avoiding the excessive memory consumption of voxel grids [28] and the difficulty of optimizing topological changes for meshes. However, the underlying neural network still has a large memory footprint and requires thousands of evaluations to produce each image. For complex scenes, the process of training a NeRF model easily spans days and requires multiple high-end graphics processing units (GPUS) [35], making experimentation expensive, slow or even inaccessible for researchers without adequate resources. Large institutions may be able to afford this, but the environmental damage of applying large models
at scale is monumental.

There exists some work on improving the runtime of neural fields by using octrees [65] or Voronoi partitions [43]. Yet these acceleration structures are mostly targeted to solely improve rendering time, with the imposed structure typically inducing an additional toll on training time. A very promising recent approach is to use hashing [36], which improves both training and test time at the expense of massive memory consumption [28] and without applicability to dynamic cases where the camera position and properties of the scene continuously evolve. Other methods focused on training time optimizations rely on hardware parallelization [8, 13] which may not be accessible to many researchers.

Our goal is to replace the multilayer perceptron (MLP) used in NeRF and many related neural field approaches with a more efficient network that exploits local continuity while not enforcing additional constraints on the scene. In the original NeRF, the MLP represents the scene as a radiance field that stores the amount of light emitted at every point and direction in space, paired with a density field which models opacity. An image is formed by classical ray marching, by casting a ray from the camera through every pixel and integrating the light emitted towards the camera step-by-step while accounting for occlusions using density. The costly part is evaluation of the MLP for every step along the ray, with the number of samples usually in the order of hundreds.

Our underlying idea is to share computation between neighboring samples, building upon the fact that these are nearby and only high-frequency details should change between them. Simple solutions such as processing lower frequencies at lower spatial resolutions did not help in our preliminary experiments (see Section 3.3). The challenge is to create a structure that enables the network to learn what can and cannot be shared as well as to maintain beneficial invariances of the original NeRF, such as not revealing the orientation to early layers of the network and not learning occlusion effects.

In our work, we take inspiration from the UNet architecture [45], used for image segmentation, which gradually reduces the spatial resolution in the first half of the network and subsequently upsamples the resulting high-level features in the second half through the use of skip connections and deconvolution. In the same vein, we gradually subsample along the ray using convolution and perform in-
terpolation instead of deconvolution for the upsampling procedure. This reduces training time, inference time, and the memory footprint, all while retaining reconstruction accuracy.

We experiment with different instantiations of our UNeRF method, showing that the proposed network consistently outperforms NeRF and other simpler baselines on the most widely used benchmarks while consuming significantly fewer resources and maintaining rendered image quality. By applying it to learning the human body shape from videos [49] and variants of NeRF with improved runtime [8], we further demonstrate the versatility and plug-in nature of our approach.
3.2 Related Work

In this section, we review the algorithmic advances in the literature for minimizing NeRF’s imminent resource demand and contrast them to our proposed architectural modifications to NeRF.

3.2.1 NeRF and its Limitations

NeRF inspired a significant shift in learning implicit volumetric representations of scenes and objects using a limited set of static images. Key to the success of NeRF and other neural field methods is the use of an MLP to represent the density and radiance at individual points in space [35, 38] which makes these methods applicable to scenes of arbitrary topology [61]. With the traditional implementation of NeRF, thousands of rays and dozens of samples are often used, resulting in tens of thousands of invocations per network layer. Millions of invocations across the network are summarized using integration along individual rays to produce a value for each pixel. Monte Carlo sampling is used to approximate this integration which requires hundreds of additional forward passes through the neural network for each of these rays. This computational burden translates to a large strain on the memory resources required for training such models. The sheer number of invocations required to optimize NeRF models also contributes to the excessive amount of time spent during experimentation. For instance, training a single scene for 100k to 200k iterations may take up to 1-2 days on an NVIDIA V100 GPU [35].

3.2.2 The Performance Trifecta: Memory, Time, Perceptual Quality

The main obstacle with current methods is the trade-off between compute time, memory, and perceptual quality especially during training. Perceptual quality can be quantified by peak signal to noise ratio (PSNR), the AlexNet [27] learned perceptual image patch similarity (LPIPS) [67], and the structural similarity index measure (SSIM) [58]. For our purposes, memory is profiled as the mean memory allocated during the forward pass and the maximum memory allocated during the entire training cycle. We quantify compute time as the time per iteration $\times$ the number of iterations required to achieve a fixed validation PSNR value.
3.2.3 Discrete Data Structure Acceleration

An alternative to volumetric scene representation through ray marching is the use of explicit volumetric grids [10, 31, 32, 50] to encode position in a multi-scale higher-dimensional space spanned by sine-wave functions. Although the use of discretized grids during training time may significantly reduce inference times [10, 32], output image resolutions are compromised due to excessive memory consumption. For instance, Neural Volumes [32] and Deep Reflectance Volumes [10] are limited to a maximum volume size of $128^3$. Decomposed Radiance Fields (DeRF) [43] divides the integration domain into Voronoi cells to produce a 3x rendering speedup with improved perceptual quality scores. Unfortunately, DeRF models require more time to train than models without decomposition. These voxelized alternatives often exhibit poorer results when applied to scenes with arbitrary topology and cannot be applied to dynamic scenes. The dynamic reconstruction task may be defined as a case where the camera position and the subject of the scene are not fixed as in the more common NeRF reconstruction task. The scene and camera positions exhibit some dynamic motion and the reconstruction task is expected to render this. Classical acceleration structures such as octrees reduce computation by avoiding sampling empty space, either by directly storing radiance [19, 60, 65] or by subdividing the space that needs to be sampled [31, 53].

3.2.4 Optimized Sampling Strategies

Computational efficiency can also be improved by relying on the tendency for natural scenes to be volumetrically sparse. Maintenance of an occupancy grid indicating empty regions for which computations do not need to be performed benefits from this natural phenomenon. Samples can be reduced or picked along each ray intelligently using depth supervision to inform effective placement of samples near possible surfaces or participating media [14, 39]. Another approach is to train a separate depth network using ground-truth depth maps to inform these surface locations [37]. This method makes the ultimate trade-off between improved rendering time (15 FPS) for half of the visual quality of NeRF on 800^2 images. Neural Sparse Voxel Fields (NSVFS) [31] and Neural Geometric Level of Detail [53] propose dynamically constructing an occupancy octree structure. Despite taking less
than 4 seconds to render an 800² image, NSVF s suffer from a reduction of image quality as measured through decreased PSNR values and still require many hours to train due to the deep MLP required for its representation.

3.2.5 Architectural and Hardware Optimization

Work centered around explicit volumetric representations has attempted to parallelize and optimize this complex MLP for a reduction in rendering time. [19, 44, 65]. KiloNeRF [44] factorizes the MLP into a large set of very small MLPs. FastNeRF [19] restructures the MLP into two MLPs supported by caching which introduces a speedup during rendering. Following the recent popularity of caching, PlenOctrees [65] replaces the MLP entirely with an octree that can be queried for direction-dependant radiance by using position as a key. However, all of these methods require a conversion precursor from a trained explicit model such as NeRF and thus experience lengthier training times especially for end-to-end models.

Along this vein, Automatic Integration (AUTOINT) [30] reduces the process of Monte Carlo sampling along thousands of rays to just two evaluations. While the method focuses on render-time speedups, their custom framework for automatic integration improves training speed per iteration by a factor of 1.8 and reduces memory consumption by 15%. With a model that matches the memory consumption of the original NeRF method, they are only able to achieve 86% of the perceptual quality.

3.2.6 Training Environment Optimization

While numerous methods have addressed the concern of high compute time during rendering, much less interest has been shown in reducing compute time and memory consumption during training. Depth-Supervised NeRF (DSNeRF) [14] relies on external 3D point clouds as input to minimize the number of samples required during training. This method is able to train 2-6 times faster than the original NeRF method through effective sample placement. However, this method relies on an additional input which may not be readily available for all datasets, may require additional preprocessing to infer, and introduces increased memory consumption.

More recently, a hierarchical hash table of training feature vectors [36] has
reduced training to seconds and rendering to milliseconds by compressing the network to two fully connected layers. Although this method’s hash table encoding is task agnostic, the exceptional memory burden placed on storing a multiresolution hash table at this scale hardly justifies the means. This is illustrated by the 16.8x increase in memory consumption when comparing Instant-NGP to the vanilla implementation of NeRF [28].

The stellar improvement of Instant-NGP [36] and MipNeRF [8] is twofold: conceptually by using acceleration structures, but also by using custom cuda/JAX optimizations [13] which are too rigid for development. We demonstrate our method on top of MipNeRF’s PyTorch variant on a single GPU which can also be parallelized in the future. Their speedups, gained by replacing networks in NeRF partially or fully with efficient lookup functions, come with image quality degradation. Furthermore, multi-GPU environments may only be attainable by a limited subset of researchers, making research inaccessible to many.

We seek a robust method which introduces reduced compute time and lower memory consumption with a minimal reduction in perceptual image quality—one that is as task agnostic as the original NeRF method without the aforementioned consequences.
3.3 Preliminary Exploration

This section details fundamental themes and preliminary exploration which shaped the method presented in Section 3.4.

3.3.1 NeRF Overview

![NeRF Architecture Diagram]

**Figure 3.1: NeRF architecture.** The NeRF network architecture is composed of 8 fully connected layers and subsequent ReLU activation functions. The blue boxes convey input vectors which are concatenated with their respective positional encoding. An initial input vector of 3D sample points concatenated with their respective positional encoding $\gamma(x)$ are fed into the network. This positional encoding $\gamma(x)$ is also concatenated with the output of layer 4 before being passed into layer 5. The portion of the network from layers 1 - 8 inform the density function $\sigma(x)$. The positional encoding corresponding to view direction is concatenated with the output of layer 8 to inform the view-dependant RGB values at each position. [35].

NeRF establishes view synthesis with volume rendering using numerous camera rays which serve as viewpoints from varying virtual cameras. Points $x$ sampled along each ray define the ray in 3D discrete space accompanied by a viewing direction $d$. Uniform or nonuniform sampling strategies may be employed depending on whether the model seeks a coarse or fine estimate of the scene. Often, a coarse model with uniform sampling will inform the refinement of samples along the ray for a fine model through a process referred to as importance sampling [35]. Stratified sampling along individual camera rays produce a series of 6D input coordi-
nates which are passed through an input encoding function. Formerly introduced in the context of attention for transformer architectures [57], this encoding function can be seen as a transformation between the spatial and Fourier domains across a spectrum of $f$ frequencies:

$$\gamma(c) = (\sin(2^0 \pi c), \cos(2^0 \pi c), \ldots, \sin(2^f \pi c), \cos(2^f \pi c)). \quad (3.1)$$

Applied individually to each component of the 3D input position $x$ and 3D Cartesian unit vector $d$, such an input encoding function projects each component into a higher dimensional space $\mathbb{R}^{6f}$ using 2 periodic operators $[\sin(\cdot), \cos(\cdot)]$ to mitigate for any bias towards lower frequency features [40].

An 8-layer fully connected neural network with 256 channels per layer and ReLU activations transforms the positional encoding $\gamma(x)$ to a 256-dimensional feature vector and volume density $\sigma(x)$ (see Figure 3.1). Concatenated with the encoded camera viewing direction $\gamma(d)$, the feature vector is passed through another fully connected layer with 256 channels and a ReLU activation function before a subsequent fully connected layer with 128 channels and a ReLU activation function produce the view-dependent color $C(r)$.

The outputted volume density $\sigma(x)$ is representative of the probability that the camera ray terminates at position $x$ along the ray. This may also be interpreted as the differential probability that the camera ray hits a visible surface at this position. The view-dependent color $C(r)$ is a continuous function representing the RGB value at the visible surface. Without prior knowledge of the location of this visible surface, $C(r)$ becomes an integral approximating RGB values through the transmittance of various samples along the ray.

With individual rays being casted through corresponding pixels onto their volumetric representation and samples along each ray serving as keys for querying outputs using the MLP, numerous values along each ray are aggregated to produce a single output. The main disadvantage is that hundreds of complex MLP invocations per pixel are required to produce such impressive visual results. As a result, this method exhibits undesirable outcomes for memory consumption and time required during training.
3.3.2 Frequency-based Quantization

An initial attempt to reduce the capacity of the network through frequency-based quantization yielded unfavourable results. The premise for this exploration is the idea that not all of the $f$ frequencies in Equation 3.1 are required at every layer of the network to reconstruct a scene. For instance, we hypothesized that earlier layers may benefit from focusing on lower frequency high level details while higher frequencies may only be required near the end of the network. With the hypothesis that some high frequency components of the positional encoding may still result in effective quantization for natural scenes, we modified the architecture to drop the higher frequency bands at the model’s input and consequently the computations associated with these bands throughout the MLP. While this resulted in reduced training time per iteration and reduced memory costs, the performance also suffered by the loss of this high frequency information.

A subsequent attempt aimed at understanding which frequency bands contributed most to a specific dataset through holding out frequency bands and observing the visual quality degradation during rendering time. This experiment revealed that the visual quality contribution from different frequency bands was not uniform across multiple viewpoints and thus would be a difficult path to follow. Eliminating certain frequency bands may have little impact on test PSNR for one camera position but a drastic reduction on PSNR for another. See Section 3.5 for further details.

3.3.3 Capacity-Performance Trade-Offs with NeRF

The inspiration for UNeRF lies within a memory-limited ablation study investigating the perceptual quality degradation incurred by reducing the model’s capacity in various directions. Starting from the default NeRF recommendation [35], we impose memory requirements along the following dimensions in order to achieve our target of 9.06GB:

- **Fewer Channels**: The width of each layer is reduced to 202 channels.
- **Fewer Layers**: The depth of the network is reduced to 6 layers, each with 249 channels.
• **Fewer Coarse Samples:** The number of samples for the coarse network is limited to 39.

• **Fewer Importance Samples:** A total of 78 importance samples are fed to the fine network.

As NeRF with fewer importance samples produces the least reduction in visual quality for the most benefits along the time and memory axes (See Section 3.5.5), it is evident that the redundancy in the importance samples may be exploited favourably. This lends opportunity for the manipulation of features corresponding to particular samples, which in turn, motivates our method, UNeRF. We use this naive method for memory reduction as a baseline for comparison with the proposed UNeRF method.
3.4 Method

We aim to improve the efficiency of NeRF by sharing some computations between adjacent samples along the ray direction since samples in close proximity share smoothly changing field properties. To this end, we propose two variants of UNeRF which both reduce resource consumption by downsampling the spatial resolution of parts of the network. The first variant is an adaptation of the UNet [45] for a 1D convolutional configuration along the samples of a ray. It departs from NeRF’s use of implicit bias to ensure view invariances, thereby opening the door for data-driven learning of dependencies between neighboring samples. The latter uses subsampling instead of convolutions to reduce the resolution of feature vectors through shared computation while maintaining view invariances. Both variants utilize position-aware interpolation to account for the non-uniform sampling structure in comparison to the regular 2D grids in the original UNet application.

Figure 3.2: UNeRF architecture. UNeRF improves computational and memory efficiency by sharing feature computations between adjacent 3D samples. Starting with positional encoded sample points, the first 5 layers process and output the feature vector with progressively smaller spatial resolution (Sections 3.4.1 and 3.4.2). We then apply position-aware linear interpolation (Section 3.4.3) to upsample and add the interpolated features to the earlier ones to recover the feature in full spatial resolution, analogous to UNet [45].

3.4.1 UNeRF-Conv

Starting from the original 10-layer MLP architecture used by NeRF, we replace the six middle layers with a variant of the 2D UNet adapted to 1D. Instead of processing sample points $x_1, \ldots, x_N$ independently as in NeRF, information between
neighboring features is exchanged with convolutions along the ray direction. Layers 2, 3, and 4 operate at a reduced spatial resolution by employing a $k \times 1$ kernel with a stride of size 2. Each of these convolutions is followed by a ReLU activation function to mimic the functional equivalent of a fully connected layer, coupled with a subsampling operation, followed by a ReLU activation function. Subsequently, the outputs of layers 4, 5, and 6 are upsampled with position-aware linear interpolation, creating the U shape in Figure 3.2. These up-scaling layers are fully connected followed by a ReLU activation function. Borrowing inspiration from ResNet [21], we make use of skip connections which help pass high-frequency information through the spatial bottleneck. This mitigates the otherwise common bias of MLPs towards lower frequency features [40] and ensures that computations can be shared across samples – increasing the model capacity without consuming additional memory.

Neighboring points along the ray involuntarily encode the view direction since samples are exclusively chosen along these rays during training and inference. The restrictive $k$-dimensional window of the 1D convolutions and irregular sampling limit its violation of the implicit bias on view invariance, allowing the 3D shape and appearance to be reconstructed using this UNeRF-Conv variant.

### 3.4.2 UNeRF-Sub

In order to share information while minimizing the violation of view invariance, we propose UNeRF-Sub, which applies subsampling instead of the strided convolution in UNeRF-Conv. The subsampling operation is followed by the MLP’s fully connected layer and subsequent ReLU activation function at layers 2, 3, and 4. Characterized by their scalar distances along the ray from the origin, points which are passed through the network are referred to as anchor points $x'$ while points that are dropped during subsampling are referred to as intermediate points $x''$. Figure 3.3b explains how this subsampling reduces the computational burden and memory footprint while the subsequent upsampling between anchor values with linear interpolations shares information (see Figure 3.3c). Following layer 4, we seek an incremental strategy to bring the feature resolution back to its original resolution.
Figure 3.3: Feature vector manipulation. Layers 0 and 1 process the features (Figure 3.3a) corresponding to each sample along the ray. Downsampling using convolutions (Section 3.4.1) or subsampling (Section 3.4.2) results in a downsampled feature set (Figure 3.3b) that corresponds to a sparser set of samples. Position-aware linear interpolation (Section 3.4.3) brings the feature set back to its original resolution through iterative application. The interpolated feature set (Figure 3.3c) accounts for the features for samples that were omitted from MLP invocations.

3.4.3 Position-aware Linear Interpolation

While pixels in the UNet method are inherently uniform, NeRF samples are chosen at irregular intervals due to importance sampling. Hence, upsampling the latter half of UNeRF as shown in Figure 3.2 by a simple average neglects the relative positioning of samples. Points along a ray can be parameterized by their scalar distances $x$ from the origin of the ray and feature maps $f(x)$. Given an intermediate point $x_1$, its neighbouring anchor points $x_0$ and $x_2$, and their corresponding feature vectors $f(x_0)$ and $f(x_2)$, our interpolation function $I(x_1)$ performs a weighted average on these anchor feature vectors to approximate the intermediate feature vector $\hat{f}(x_1)$. Figure 3.4 portrays this interpolation technique on a subset of sample points.
\[ I(x_1) = \hat{f}(x_1) \approx f(x_0) + \frac{(f(x_2) - f(x_0))(x_1 - x_0)}{x_2 - x_0}. \] (3.2)

Interleaving the interpolated feature vector set \( \hat{f}(x'') \) with the anchor feature vectors \( f(x') \) results in a feature set \( \hat{f}(x) \) of double the resolution. This interpolation operation is visualized in Figure 3.3c and its application is shown in Figure 3.2. Simpler interpolation strategies are discussed in Section 3.5.

### 3.4.4 Implementation

Our implementation of UNeRF builds on top of a PyTorch implementation of NeRF [29] which can be found on Github (https://github.com/yenchenlin/nerf-pytorch). This pre-existing repository holds an MIT license.
3.5 Evaluation

In this section, we provide an overview of the baselines and benchmarks against which we evaluate UNeRF. We perform experiments to show that our method is applicable to both dynamic and static scenes for 3D scene reconstruction. Additional experiments demonstrate UNeRF’s ease of substitution for NeRF in other downstream tasks such as human motion reconstruction while exhibiting a smaller memory footprint, reduced optimization time, and better performance than NeRF itself. Ablation studies justify the various design choices which are discussed in Section 3.4.

3.5.1 Datasets

Blender This dataset group contains 3D renderings of synthetic objects where viewpoints are sampled densely from the upper hemisphere for the majority of the datasets and from all directions for 2 datasets. As in NeRF, 100 of these views are used for training, 100 for validation, and 200 are left for testing. The version of the dataset used in this work was downloaded from https://drive.google.com/drive/folders/1JDdLGDrUGNWXnM1eqY1FNL9PISjaKWf. We follow suit in evaluating on 7 of the static synthetic Blender scenes presented by NeRF [35] where each synthetic object is processed at a resolution of 512 × 512. These 3D renderings include the Chair, Drums, Ficus, Lego, Materials, and Ship datasets.

LLFF This dataset captures real scenes from a handheld phone camera pointed at the environment with only forward facing angles, leading to a less uniform sampling of training viewpoints. The amount of images ranges between 20-62 for each of the LLFF datasets which is much less than the Blender scenes, making it a more difficult benchmark. This dataset can be found at https://drive.google.com/drive/folders/14bol-o5hGO9srnWaaogTU5-jl7wkx2S7. A subset of 5 complex forward facing camera scenes, made available by the Local Light Field Fusion paper [34], are used. These static scenes include the Fern, Horns, Room, Trex, and Flower datasets. An eighth of these 1008 × 756 handheld phone captures are used as the test split. The data split for this dataset also follows from NeRF [34, 35].
The SURREAL dataset captures dynamic synthetic human motion. It is the only dataset in use in this work which contains humans. Due to its synthetic nature, no personally identifiable elements are present. The dataset was accessed at https://drive.google.com/drive/folders/1DEF6xi2XTrxBx1IYbMwvSIDcMkFmlKYL. This dynamic synthetic dataset encompasses 10,800 training images from 1,200 3D human body poses rendered from 9 cameras. The test set is composed of 300 body poses from 5 different scene cameras which amounts to 1,500 test images. Although this dataset was first presented by the Learning from Synthetic Humans paper [56], it was later adopted by A-NeRF [49].

### 3.5.2 Metrics

**Visual Quality** We report the peak signal to noise ratio (PSNR), the AlexNet [27] learned perceptual image patch similarity (LPIPS) [67], and the structural similarity index measure (SSIM) [58] to quantify visual image quality. For results that pertain to A-NeRF experiments, we also present the PSNR and SSIM values for the foreground to quantify the margin our method achieves with respect to the baseline on the subject of interest. In some cases, we report mean PSNR values across similar classes of datasets to summarize model performance.

All difference maps presented in this work are generated by running models across 300k iterations. Both the UNeRF-Conv models and the NeRF models used for these results were run with $N_{rays} = 4096$, $N_{samples} = 64$, and $N_{importance} = 128$. Normalized pixel differences $\hat{D}$ between UNeRF-Conv (A), NeRF (B), and the ground truth (GT) of pixel $p = (p_r, p_g, p_b)$ are computed using

$$\hat{D}(p) = \frac{\log(|D(p_A) - D(p_B)| + 1)}{\text{MAX}(|D(p_A) - D(p_B)| + 1))}, \text{ with}$$

$$D(p_A) = \sqrt{(GT p_r - A p_r)^2 + (GT p_g - A p_g)^2 + (GT p_b - A p_b)^2},$$

$$D(p_B) = \sqrt{(GT p_r - B p_r)^2 + (GT p_g - B p_g)^2 + (GT p_b - B p_b)^2}.$$
Memory Profile  The maximum memory allocated during training and the median memory consumed during the forward pass are used to quantify the memory footprint. We leverage the PyTorch Profiler API for benchmarking the forward pass memory consumption [4]. The maximum memory allocation is gauged using PyTorch’s `torch.cuda.max_memory_allocated()` function [3].

Optimization Time  The primary time metric of interest is the amount of time it takes for a model to reach NeRF quality in terms of a fixed validation PSNR on an NVIDIA V100 GPU. We also document the median time per iteration where relevant on the same GPU model.

3.5.3 Comparable Models and Baseline

The default configuration suggested by NeRF [35] is an 8 layer network with 64 coarse samples, 128 importance samples, and 4096 rays. As our objective is to present a method that optimizes memory, training time, and visual performance, all quantities influence one other, so we compare models where the forward pass consumes a constant 9.06GB of GPU memory usage to enable direct comparisons. We fine-tune the number of importance samples to meet the memory target for the NeRF baseline and our UNeRF variants as our preliminary ablations demonstrate that this reduces visual fidelity the least compared to varying layer depth or width. The best-performing model variants are:

- NeRF-Mem: An ablation of the NeRF architecture with only 78 importance samples.
- UNeRF-Fine-Sub-Mem: Our UNeRF-Sub architecture applied to only the fine network with 112 importance samples.
- UNeRF-Conv: Our UNeRF-Conv architecture applied to both the coarse and fine networks.
Table 3.1: Performance summary on the Blender dataset. Images are rendered at 400 × 400 resolution. Both UNeRF variants deliver comparable or even better image quality with improved memory consumption. UNeRF-Conv presents a 12% improvement in training time.

<table>
<thead>
<tr>
<th>Model</th>
<th>Memory (GB) ↓</th>
<th>Time ↓</th>
<th>PSNR (dB) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Forward</td>
<td>Max</td>
<td>Per Iteration (s)</td>
</tr>
<tr>
<td>NeRF</td>
<td>11.27</td>
<td>11.76</td>
<td>0.42</td>
</tr>
<tr>
<td>NeRF-Mem</td>
<td>9.06</td>
<td>9.47</td>
<td>0.34</td>
</tr>
<tr>
<td>UNeRF-Fine-Sub-Mem</td>
<td>9.07</td>
<td>9.62</td>
<td>0.36</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>9.06</td>
<td>9.73</td>
<td>0.59</td>
</tr>
</tbody>
</table>

1 Time to achieve a VAL PSNR of 34.29dB on the chair dataset.
2 Average Test PSNR across 7 Blender scenes.

3.5.4 Object, Scene, and Human Motion Reconstruction

Static Synthetic Scenes

The NeRF-Mem baseline using fewer importance samples converges quicker but degrades the mean visual quality. Our memory limited UNeRF-Fine-Sub-Mem network improves over this straight-forward baseline achieving 99.68% of NeRF’s visual performance at full capacity. It struggles to achieve higher resolution results beyond a fixed threshold on volumetrically dense synthetic scenes due to the compression caused by subsampling. Furthermore, UNeRF-Conv can match the full capacity NeRF baseline while consuming 24% less memory per forward pass, 21% less overall memory, and taking 3.86 fewer hours to converge to NeRF performance on the Chair dataset. Fully trained, the UNeRF-Conv method even outperforms NeRF by 0.93dB across 7 Blender scenes after 300k iterations. Our method achieves better performance in all 3 categories, showing that the benefits of shared computation and using convolutions outweigh the benefits of strict view-invariances in NeRF. Table 3.1 provides a quantitative overview of these results and Figure 3.5 supplies visual comparisons of the improved reconstruction quality, especially for fine details.
Figure 3.5: Qualitative comparisons on the Blender dataset. Top row: Rendered testset images from UNeRF-Conv with a $3 \times 1$ kernel. Bottom row: Difference maps between UNeRF-Conv, NeRF and the ground truth. Pixels highlighted more prominently in blue indicate regions where UNeRF-Conv more closely aligns with the ground truth than NeRF, and red vice versa. The pixel color difference is normalized to $[0, 1]$. While NeRF and UNeRF-Conv typically exhibit comparable results when describing low frequency details, UNeRF-Conv has an edge on depicting high frequency detail more accurately.

Static Real Forward Facing Scenes

UNeRF-Conv struggles against the full capacity baseline, NeRF-Mem, and UNeRF-Fine-Sub-Mem on real forward facing static scenes. We speculate that this is due to the volumetric sparsity of the scenes in this dataset and the added difficulty in

Table 3.2: Quantitative comparisons on the LLFF dataset. The following experiments were evaluated across 100k iterations. Our UNeRF-Fine-Sub-Mem baseline produces comparable results to NeRF with a slight performance reduction. UNeRF-Conv struggles with the reconstruction of this data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fern</th>
<th>Flower</th>
<th>Horns</th>
<th>Room</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeRF</td>
<td>26.89</td>
<td>27.90</td>
<td>28.50</td>
<td>31.73</td>
<td>28.76</td>
</tr>
<tr>
<td>UNeRF-Fine-Sub-Mem</td>
<td>26.82</td>
<td>27.80</td>
<td>28.31</td>
<td>30.87</td>
<td>28.45</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>26.34</td>
<td>27.82</td>
<td>27.95</td>
<td>30.54</td>
<td>28.16</td>
</tr>
</tbody>
</table>
Figure 3.6: Qualitative results of UNeRF on the LLFF dataset. Top row: novel views rendered by NeRF at 100k iterations. Bottom row: novel views rendered by UNeRF-Fine-Sub-Mem at 100k iterations. Our method produces comparable results to NeRF while consuming less time and memory during training time.

Dynamic Human Motion

The following experiments investigate UNeRF’s applicability to dynamic scenes capturing non-rigid human motion.

Application to the dynamic human motion problem is a critical one; most prior
Table 3.3: Quantitative comparisons on rendering dynamic human motion from the SURREAL dataset. Images are rendered at $512 \times 512$ resolution. UNeRF-Conv yields better image quality as well as memory consumption.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Max Memory (GB) ↓</th>
<th>Overall PSNR ↑</th>
<th>Overall SSIM ↑</th>
<th>Foreground PSNR ↑</th>
<th>Foreground SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeRF</td>
<td>6.270</td>
<td>34.58</td>
<td>0.986</td>
<td>25.25</td>
<td>0.889</td>
</tr>
<tr>
<td>UNeRF-Sub</td>
<td>5.221</td>
<td>33.94</td>
<td>0.984</td>
<td>24.62</td>
<td>0.875</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>5.235</td>
<td><strong>34.96</strong></td>
<td><strong>0.987</strong></td>
<td><strong>25.53</strong></td>
<td><strong>0.897</strong></td>
</tr>
</tbody>
</table>

methods [43, 50, 66] with interest in training resource optimization have not been applicable to dynamic scenes. The hyperparameter settings for these experiments come directly from the A-NeRF recommendation for the SURREAL dataset [49], which demonstrates our plug-in nature. The models are configured with 2048 rays, 64 coarse samples, and 16 importance samples. For evaluation, we use human poses and camera positions that are unseen during training. We compare the performance of the full NeRF model to UNeRF-Conv and UNeRF-Sub which both exhibit comparable maximum memory expenditure (see Table 3.3).

Figure 3.7: Dynamic human motion rendering on the SURREAL dataset. The 3D human pose and view are both unseen during training time. UNeRF-Conv preserves better finger structure than the other architectures and achieves better results than the full capacity NeRF baseline when fully trained.

UNeRF-Sub applied to both the coarse and fine networks suffers from aliasing as can be seen in the fingers of the subject in Figure 3.7c. With a maximum memory consumption of 5.24GB in contrast to 6.27GB, UNeRF-Conv consumes 20%
less memory than NeRF while exhibiting stronger foreground and overall PSNR and SSIM performance across 150k runs. This demonstrates that the benefit seen for static scenes translates directly to dynamic ones, which is in contrast to most existing approaches that improve efficiency only for static scenes.

3.5.5 Ablation Studies

In this section, we experiment with different settings and alternatives functions to defend our chosen design as discussed in sections 3.3 and 3.4.

Parameter Tuning and Trade-offs

The experiments presented in Table 3.4 motivated the decision to manipulate samples to satisfy the performance trifecta (see Section 3.3). Reducing samples and importance samples leads to a reduction in the time per iteration, reduced max memory costs, and the achieves the highest PSNR values. Reducing the number of channels to meet our desired forward pass memory threshold resulted in the most degradation in terms of PSNR.

Convolution Size

Theoretically, UNeRF-Conv with kernel size $1 \times 1$ and stride 2 is the functional equivalent of UNeRF-Sub where every other sample is dropped during the sub-

<table>
<thead>
<tr>
<th>Table 3.4: Quantitative performance metrics of memory-limited NeRF baselines on the Blender dataset. These models were evaluated across 300k iterations. Reduction of the importance samples results in the least loss of visual quality and the fastest time per iteration for a relatively low maximum memory allocation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>NeRF</td>
</tr>
<tr>
<td>Fewer Layers</td>
</tr>
<tr>
<td>Fewer Channels</td>
</tr>
<tr>
<td>Fewer Samples</td>
</tr>
<tr>
<td>Fewer Importance Samples</td>
</tr>
</tbody>
</table>
Table 3.5: Ablation study on convolution kernel size on the Blender dataset. Images are rendered at 400 × 400 resolution. The larger UNeRF-Conv kernel size improves rendering outcomes.

<table>
<thead>
<tr>
<th>Model</th>
<th>Kernel</th>
<th>Chair</th>
<th>Drums</th>
<th>Ficus</th>
<th>Hotdog</th>
<th>Lego</th>
<th>Materials</th>
<th>Ship</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeRF</td>
<td>-</td>
<td>35.39</td>
<td>25.62</td>
<td>29.94</td>
<td>36.57</td>
<td>32.36</td>
<td>29.81</td>
<td>29.65</td>
<td>31.33</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>3x1</td>
<td>35.69</td>
<td>26.11</td>
<td>31.63</td>
<td>37.75</td>
<td>33.30</td>
<td>31.73</td>
<td>29.62</td>
<td>32.26</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>2x1</td>
<td>35.53</td>
<td>25.68</td>
<td>31.08</td>
<td>37.29</td>
<td>33.04</td>
<td>31.60</td>
<td>30.47</td>
<td>32.10</td>
</tr>
<tr>
<td>UNeRF-Conv</td>
<td>1x1</td>
<td>34.90</td>
<td>25.41</td>
<td>29.58</td>
<td>36.26</td>
<td>31.59</td>
<td>29.71</td>
<td>29.62</td>
<td>31.01</td>
</tr>
<tr>
<td>UNeRF-Sub</td>
<td>-</td>
<td>34.90</td>
<td>25.36</td>
<td>29.65</td>
<td>36.58</td>
<td>31.31</td>
<td>29.78</td>
<td>29.60</td>
<td>31.03</td>
</tr>
</tbody>
</table>

sampling operation. As the 1D convolution doubles as a fully connected layer and subsampling operation, when trailed by a ReLU activation function, the UNeRF-Conv 1 × 1 is the equivalent of the UNeRF-Sub implementation. As such, it makes sense that UNeRF-Sub and UNeRF Conv 1 × 1 match in PSNR performance with some noise. We perform experiments with kernel sizes of 1, 2, and 3 to show that a kernel size of 3 × 1 produces superior results at a reduced memory capacity (see Table 3.5).

In order to emphasize that this configuration is not dataset-specific, we also demonstrate that a kernel size of 3 × 1 is most favourable on the SURREAL dataset through application of UNeRF to A-NeRF (see Table 3.6).

Questions regarding whether larger kernel sizes produce more favourable results may arise. To address these concerns, we performed an ablation study to

Table 3.6: Kernel Size Ablation Study of UNeRF-Conv on the SURREAL Test Dataset using A-NeRF. These experiments are evaluated across 150k iterations. UNeRF-Conv with a kernel size of 3 × 1 outperforms kernel sizes of 2 and 1 on the SURREAL dataset.

<table>
<thead>
<tr>
<th>Kernel Size</th>
<th>Full</th>
<th>Foreground</th>
<th>Full</th>
<th>Foreground</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 × 1</td>
<td>34.96</td>
<td>25.53</td>
<td>0.987</td>
<td>0.897</td>
</tr>
<tr>
<td>2 × 1</td>
<td>34.51</td>
<td>25.25</td>
<td>0.986</td>
<td>0.889</td>
</tr>
<tr>
<td>1 × 1</td>
<td>33.81</td>
<td>24.50</td>
<td>0.983</td>
<td>0.871</td>
</tr>
</tbody>
</table>
Table 3.7: Larger kernel size ablation study of UNeRF-Conv on the Blender Lego test dataset. These experiments are evaluated on a single NVIDIA GeForce RTX 3090 GPU. UNeRF-Conv with a kernel size of $3 \times 1$ provides the best tradeoff between performance, time, and memory when compared to the larger kernel sizes.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Memory (GB) ↓</th>
<th>Time ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Forward Pass</td>
<td>Max</td>
</tr>
<tr>
<td>$3 \times 1$</td>
<td>9.06</td>
<td><strong>9.729</strong></td>
</tr>
<tr>
<td>$4 \times 1$</td>
<td>9.06</td>
<td>9.735</td>
</tr>
<tr>
<td>$5 \times 1$</td>
<td>9.06</td>
<td>9.744</td>
</tr>
<tr>
<td>$6 \times 1$</td>
<td>9.06</td>
<td>9.750</td>
</tr>
</tbody>
</table>

benchmark larger kernel sizes along our 3 primary performance directions. As shown in Table 3.7, larger kernel sizes inevitably lead to a slight memory increase and significantly longer training time; this works against our goal of being more performant. Nevertheless, such a hyperparameter choice could be beneficial when higher image quality is desired.

Table 3.8: Ablation study on interpolation mechanisms with UNeRF-Sub on the Fern LLFF dataset. The following experiments are evaluated across 100k iterations. UNeRF-Sub with our position-aware interpolation mechanism achieves a benchmark VAL PSNR faster than the two simpler baselines.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR (dB) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nearest Neighbor</td>
</tr>
<tr>
<td>UNeRF-Sub</td>
<td>26.25</td>
</tr>
<tr>
<td>Time Per Iteration (s)</td>
<td><strong>0.57</strong></td>
</tr>
<tr>
<td>Time to 26dB VAL PSNR (h)</td>
<td>10.7</td>
</tr>
</tbody>
</table>
Interpolation Strategy

The aforementioned position-aware linear interpolation mechanism maintains the non-uniform nature of samples. Simpler interpolation strategies such as nearest neighbour and average interpolation are substituted for this upsampling operation to show its merit. Nearest neighbour interpolation suffers the most in terms of perceptual quality while lending a 0.02s speedup in median time per iteration (0.57s) over our position-aware linear interpolation (0.59s) (see Table 3.8). Experiments using averaged interpolation result in lower PSNR values than our interpolation strategy (see Table 3.9).

Table 3.9: Ablation study on interpolation mechanisms with UNeRF-Conv on the Blender dataset. The following experiments were evaluated across 100k iterations. Our position-aware interpolation strategy outperforms nearest neighbour and average interpolation on most datasets.

<table>
<thead>
<tr>
<th>Interpolation</th>
<th>Chair</th>
<th>Drums</th>
<th>Ficus</th>
<th>Hotdog</th>
<th>Lego</th>
<th>Materials</th>
<th>Ship</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>33.79</td>
<td>25.54</td>
<td>29.92</td>
<td>35.83</td>
<td>31.43</td>
<td>30.41</td>
<td>29.50</td>
<td>30.89</td>
</tr>
<tr>
<td>Average</td>
<td>33.80</td>
<td>25.63</td>
<td>30.44</td>
<td>36.16</td>
<td>31.72</td>
<td>30.26</td>
<td>29.05</td>
<td>31.01</td>
</tr>
<tr>
<td>Position-Aware</td>
<td>34.07</td>
<td>25.73</td>
<td>30.58</td>
<td>36.49</td>
<td>31.78</td>
<td>30.22</td>
<td>28.91</td>
<td>31.11</td>
</tr>
</tbody>
</table>

3.5.6 Application

Ever since the widespread adoption of NeRF, several methods [8, 36] have built on top of this architecture to optimize for various considerations. The orthogonal, plug-in nature of UNeRF enables us to optimize such methods as well. Equipping Mip-NeRF [8] with our UNeRF method reduces the max memory consumption during training by 21%, while also achieving a higher validation PSNR value when training for the same amount of time as the baseline, and reaching higher test scores (see Table 3.10). It is notable that much of Mip-NeRF’s performance advantages to the original NeRF can be attributed to optimized hardware use with JAX [13] which is complementary to our advancements.

Additionally, Table 3.3 shows UNeRF’s application on more complex forms of NeRF, such as A-NeRF, where the scene topology and structure of the inputs varies.
Table 3.10: Application of UNeRF-Conv to Mip-NeRF. These results are evaluated on a single NVIDIA GeForce RTX 3090 GPU. With the UNeRF-Conv optimization on Mip-NeRF, we achieve a much better trade-off between performance, time, and memory.

<table>
<thead>
<tr>
<th>Mip-NeRF Model</th>
<th>Kernel Size</th>
<th>Max Memory (GB)</th>
<th>Test PSNR* ↑</th>
<th>Test SSIM* ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>12.72</td>
<td>33.23</td>
<td>0.966</td>
</tr>
<tr>
<td>With UNeRF-Conv</td>
<td>3 × 1</td>
<td>10.56</td>
<td>33.31</td>
<td>0.967</td>
</tr>
</tbody>
</table>

dramatically from the typical application.
3.6 Conclusion

We present a method for training photorealistic scene rendering applications with a reduced optimization time and a significant reduction in the memory footprint compared to state-of-the-art methods while achieving higher fidelity results. Our UNeRF method with subsampling processes samples sparsely yet independently, allowing for it to achieve comparable results to NeRF especially on volumetrically sparse scenes with limited data and camera direction diversity. Although UNeRF-Conv shares information between adjacent samples along a ray, its restrictive neighbourhood allows for the method to disentangle geometry and view so long a sufficient number of training views is available. We show that our UNeRF-Conv method outperforms NeRF with respect to visual quality while reducing the memory footprint and by producing faster optimization times on both dynamic and static synthetic scenes. To the best of our knowledge, this is the first neural radiance fields optimization method to report such an improvement in both memory and training time consumption while retaining its applicability to both static and dynamic scenes as well as presenting no image quality degradation.

Our proposal of dropping the strict assumption of view invariances for the sake of sharing computations across samples is of conceptual nature and orthogonal to several state-of-the-art methods. We reinforce this by integrating our UNeRF approach into MipNeRF, showcasing a 21% improvement in memory while also exceeding their image quality on top of their already optimized model within identical time constraints. The largest advantage of working with UNeRF is its task agnostic nature, working on dynamic scenes, where existing acceleration structures do not apply. UNeRF performs particularly well on dynamic human motions as already a single camera recording of strongly articulated motion shows body parts from all sides, acting as many virtual views, which we believe eases learning view invariances in a data-driven way. The application of UNeRF to the articulated neural radiance fields method A-NeRF validates its plug-in nature to other methods including dynamic ones. This is an important step towards training both time and memory conscious networks for photorealistic novel scene reconstruction.
3.6.1 Impact

Even with our proposed method, training NeRF and related neural radiance fields models on GPUs may span several hours if not days, imposing significant environmental harm. We acknowledge the carbon footprint of training such work and encourage further research in the domain of reducing the resource demands of such methods. As memory availability can be a crucial bottleneck for training such methods, we hope the memory savings introduced in this work increase access for researchers with limited resources.
Chapter 4

Tying it Together

This line of work was inspired by the long term goal of building more accurate pedestrian simulation engines which will hopefully not require manual coding to achieve realistic roadside behaviour in the future. In Chapter 1, we procured a new pedestrian-centric dataset which captures 26 adult pedestrians’ behaviour and attention patterns in varying weather conditions, crossing types, and times of the day. Chapter 2, aims at providing a mode for training dynamic human motion reconstruction models and static scene models in a more efficient manner than the current state-of-the-art by optimizing along all directions of the performance tri-recta. A promising next step would be to leverage this dataset to train both pedestrian and scene reconstruction models using UNeRF. With additional behavioural and attention modelling, this could result in much more personalized models which consider individual pedestrian characteristics in their predictions. We hope that this work serves as a stepping stone in furthering the push towards seeing autonomous vehicles on the road while also encouraging more environmentally friendly and accessible research considerations in the photo-realistic scene reconstruction domain.
Bibliography


53


54


55


Appendix A

Supporting Materials

A.1 PIETVA Study Documentation

A.1.1 Consent Form
The consent form must be presented to the participant at the very beginning of the study. The study may be described to the participant in vague terms such as the tasks they will be performing but the hypothesis or objective of the study must not be disclosed. The consent form used for this study is attached below.
Consent Form
Toward a More Natural Approach to Attention Research 1-200

Principal Investigator: Dr. A. Kingstone, Department of Psychology, University of BC

Introduction and Purpose
Think of the last time you were in a classroom before the start of class. There were probably a number of different conversations going on, the classroom door was opening and closing as students filtered into the room, books were being opened, and pencils and pens taken out. As you sat in your seat, you probably found that you focused on some people and some conversations while other people and conversations went unnoticed. That act of selecting and attending to some items in the environment at the cost of not attending to others, is what will be studied in the present investigation. In particular we are interested in understanding attention in the context of natural everyday tasks.

Study Procedures
If you agree to participate, the experiment will take about between 45 to 60 minutes of your time depending on length of the experiment. If you are unsure of the length of the experiment please ask. Depending on the particular experiment you will be asked to perform a simple task. This will involve responding to stimuli (e.g., images) presented on a computer. Responses may include manual (e.g., pressing a button, or moving a mouse), eye movements, audio/video recordings, and questionnaires. The recording of eye movements will not involve wearing any equipment, however your chin may be positioned in a chin rest. You will receive practice with specific instructions for the task before you begin. If you are not sure about any instructions, or wish to have more practice, do not hesitate to ask. Your data will only be used for research purposes. In some cases, this will involve other participants viewing your data.

Risks
Due to the in-person nature of this experiment, there is the risk of COVID-19 transmission. To minimize this risk, all researchers who interact with participants are fully vaccinated. All participants are screened for COVID-19 symptoms and
their potential risk of exposure to the virus when they register for the study, and are screened again when they arrive at the lab on the day of the study. Additionally, all participants will be required to provide proof of full COVID-19 vaccination via the BC vaccination card prior to participation. Any participant not meeting COVID-19 health precautions and requirements will not be allowed to participate. Additionally, researchers and participants must wear masks, engage in safe physical distancing at all times, and utilize sanitization sprays, wipes, and/or gels before, during, and after the study. Masks will be provided for participants who do not have one, and all equipment will be sanitized after each participant. There are no risks associated with participating in this experiment, other than those related to COVID-19 and outlined above.

COVID-19 contact tracing:
Prior to participating in the study, you will be asked to provide your name and a phone number at which you can be contacted to be used for contact tracing. This information will be written down and kept in a locked filing cabinet in the lab, only accessible to Dr. Alan Kingstone, his graduate students, and research assistants.

Confidentiality
Your identity will be kept strictly confidential. All documents will be identified only by code number and kept in a locked filing cabinet. You will not be identified by name in any reports of the completed study. Data will be stored on UBC’s Microsoft OneDrive and will also be identified only by code number. The OneDrive account will be password protected so that only the principle investigator, Dr. Alan Kingstone, his graduate students, and research assistants will have access to it. Note, the results of this study will be used to write a scientific report, and may be part of student thesis research.

Remuneration
Depending on the experiment, you will participate for between 45 to 60 minutes. You will receive ½ course credit per ½ hour, or $5 per ½ hour cash for your participation. Your credit will automatically be submitted approximately 24 hours after the end of the session.

Contact for information about the study
This study is being conducted by Dr. Alan Kingstone, the principal investigator, with Kevin Roberts, Nicola Anderson, Natalie Brown, Leilani Forby, Oliver Jacobs, Anita Schmalor, Veronica Dudarev, Farid Pazhoohi, Jake Gerlofs, Brandon Fors, and Grayson Mullen as the co-investigators for the study. Please call any one of them at [contact information]. Dr. Kingstone may be reached at [contact information].

Contact for concerns about the rights of research subjects
If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [contact information] or if long distance e-mail [contact information] or call toll free [contact information].

Consent
Your participation in this study is entirely voluntary and you may refuse to participate or withdraw from the study at any time without jeopardy to your class standing. You may also withdraw from the experiment at any time during or after your participation and request that your data be deleted. Please feel free to ask the experimenter any additional questions you may have about the study. By signing below, you are indicating that you are at least 18 years old, have read this consent form, and agree to participate in this research study.

If you would like a copy of this form for study H10-00527 for your records, please notify the researcher, and provide them with your email address.

Subject Signature __________________________ Date __________________________

Printed Name of the Subject __________________________
A.1.2 Study Protocol

A quick breakdown of the study steps, links to the meeting point, route, and survey are provided to the researcher conducting the study as a gentle reminder of key milestones that are necessary for a successful run of the study (see Figure A.1).

PIETVIA: Data Collection Study Instructions

Protocol: [1 hour]
1. Meet the researchers in front of the ICICS Building
2. Sign the consent form
3. Wear the Pupil Labs glasses, attach the glasses to the motorola phone
4. Gaze Calibration
   a. Hold the phone in your hands while the researchers perform the eye calibration
      i. Hold your head still and move your eyes to follow the tracker with the white outer circle, look at the target where there is a plus sign
      ii. Move your head and follow the tracker with your eyes
4. Gaze Calibration
5. You will be shown the route on a device so that you are aware of where to walk
6. We will do this walk 3 times,
   a. The first time, the researcher will accompany you (b/c may be interchanged)
   b. The next time, you will do this walk alone
   c. The next time, you will be required to hold the phone in your hand, occasionally look at the video to make sure that it is still recording
7. Between each walk we will start a new calibration and recording
8. If you find that you are lost, we ask that you stop your walk, and look at this route on your personal phone.
   You can scan this for the route:

9. We will have a short debrief to talk about the purpose of this study:
   https://ubc.ca1.qualtrics.com/jfe/form/SV_6sSljU50XG4JM8u
10. Get the debriefing form from the researcher

Figure A.1: Study protocol. The study protocol serves as a quick checklist highlighting all the key steps of the study.

A.1.3 Demographic Questionnaire

The online questionnaire can be found here: https://ubc.ca1.qualtrics.com/jfe/form/SV_6sSljU50XG4JM8u. This questionnaire was shared with participants via the
following QR code.

![QR Code]

**Figure A.2: Questionnaire QR code.** The QR code used to share the questionnaire with participants. This allows the participants to perform the questionnaire on their own device with ease.

The contents of this questionnaire are documented in Figure A.3.

### A.1.4 Debriefing Form

Prior to beginning the experiment, the participants should have limited knowledge of its purpose such that they are not biased by the hypothesis or what they perceive to be the researchers’ preferred outcomes. As such, it is important to share the full study details with participants after the study has been completed. The debriefing form presented in Figure A.4 is provided to the participants at the very end of the study. The researcher must provide adequate time for the participant to read through the debriefing form and ask any follow up questions.
Figure A.3: Demographic questionnaire. This questionnaire is administered to participants after their participation in the PIETVA data collection study.
When people cross the street, they make a series of movements that are indicative of their attention, intention, and comprehension of the roadside environment. Research has shown that there are distinct patterns of variation across different classes of pedestrians. Factors such as age, gender, geographical context, crossing type, time of day, presence of a companion, and much more may contribute to individual crossing behaviour. Often times, autonomous vehicle algorithms consider pedestrians as a uniform entity when predicting their future trajectory. Observational studies in VR environments revealed the tendency for older pedestrians to be fixated on their feet when crossing the road and making overly cautious or overly risky movements due to poor judgement of vehicle speeds (Zito et al., 2015). Higher compliance to laws when crossing was also attributed more closely to older pedestrians than younger pedestrians in a simulation environment (Yang et al., 2006). Lower scan times prior to crossing and wait times prior to crossing at signalized intersections were often exhibited by males when compared to those of females (Tiwari et al., 2007). Although these general findings have been observed in a variety of geographical contexts, these variables have not been attributed in a real life North American context across multiple types of crossings. While more work is needed to define these patterns with respect to crossing type, the general consensus is that patterns of behaviour can be attributed to specific pedestrian characteristics.

In the present work, we ask whether these idiosyncratic tendencies extend to how people orient the whole attention (including eye, head, and body movements) when interacting with distinct crossing types in real life.

In this experiment, you were asked to walk a route which contained 2 zebra crossings, 2 stop sign crossings, and 2 traffic signal crossings. You were asked to walk this route while accompanied by a companion, by yourself, and with a mobile device in your hand. Afterward, you were asked a few questions about your age, gender identity, and the geographical context in which you grew up.

We hypothesize that people with similar age, gender, and places of origin will display similar patterns in the way they orient their body, head, and eyes when crossing at different intersections when crossing alone. We hypothesize that all pedestrians will exhibit similar patterns in the way they orient their body, head, and eyes when crossing with a companion and with a mobile device in hand.

The independent variables was your age, gender, and the norms associated with the geographical context within which you were raised. The dependent variables we measured included scanpath similarity with individuals with similar demographics, average duration of fixations, crossing speed, average scan times before crossing, and potential objects in the scene that were the subject of fixation.

For further information about the subject of this experiment, please read:


Figure A.4: Study debriefing. The debriefing form outlines a brief description of the prior art, the objective of the study, including the hypothesis and study variables, the nature of the experiment, and relevant contact information.