OULU 2023
C 891

ACTA UNIVERSITATIS OULUENSIS

Phong Nguyen

NEURAL SCENE REPRESENTATIONS FOR LEARNING-BASED VIEW SYNTHESIS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

ACTA UNIVERSITATIS OULUENSIS
C Technica 891

PHONG NGUYEN

NEURAL SCENE REPRESENTATIONS FOR LEARNING-BASED VIEW SYNTHESIS

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in the OP auditorium (L10), Linnanmaa, on 25 August 2023, at 12 noon

UNIVERSITY OF OULU, OULU 2023

Acta Univ. Oul. C 891, 2023

Supervised by
Professor Janne Heikkilä

Reviewed by
Doctor Martin Danelljan
Professor Tobias Ritschel

Opponent
Professor Serge Belongie

ISBN 978-952-62-3739-8 (Paperback)
ISBN 978-952-62-3740-4 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

PUNAMUSTA
TAMPERE 2023

Nguyen, Phong, Neural scene representations for learning-based view synthesis.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
Acta Univ. Oul. C 891, 2023
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

This thesis introduces learning-based novel view synthesis approaches using different neural scene representations. Traditional representations, such as voxels or point clouds, are often computationally expensive and challenging to work with. Neural scene representations, on the other hand, can be more compact and efficient, allowing faster processing and better performance. Additionally, neural scene representations can be learned end-to-end from data, enabling them to be adapted to specific tasks and domains.

Conventional structure-from-motion, structure-from-depth, and multi-view geometry techniques prescribe how the 3D structure of the environment is represented. This thesis introduces architectures that learn this representational space, allowing it to express concisely the presence of textures, parts, objects, lights, and scenes using a single vector. In addition, the methods can account for the uncertainty of understanding the scene's content in the face of severe occlusions and partial observations.

Large-scale novel view synthesis aims to generate photo-realistic images of arbitrary targets in the 3D space. Recent research has produced target views by interpolating in ray or pixel space and they often suffer from artifacts arising from occlusions or inaccurate geometry. This work proposes novel, efficient frameworks that represent 3D scenes as multiple-depth planes. The trained model can render color and depth images of the novel views. The proposed architectures are compact and produce plausible results on unseen data without fine-tuning or test-time optimization.

Human capture and rendering is the process of capturing the appearance and motion of a human and generating a realistic 3D representation of that person. Existing methods tackle this problem using expensive multi-view capture setups. This thesis focuses on the issue of predicting novel views of an unseen dynamic human using a single viewpoint. Instead of representing the input as a point cloud, this work presents an efficient sphere-based view synthesis network that produces higher-quality results than multi-view approaches. Despite being trained solely on synthetic data, the work also shows great generalization performance on real images.

Keywords: human synthesis, neural radiance fields, novel view synthesis, plane sweep volumes, sphere-based rendering, vector-based representation

Nguyen, Phong, Neuraaliset näkymien esitystavat oppimispohjaiseen näkymäsynteesiin.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta
Acta Univ. Oul. C 891, 2023
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Tämä väitöskirja esittelee lähestymistapoja oppimispohjaiseen uuden näkymän synteesiin käyttäen erilaisia neuraalisia näkymän esitystapoja. Perinteiset esitystavat, kuten vokselit tai pistepilvet, ovat usein laskennallisesti kalliita ja haastavia käsitellä. Neuraaliset näkymän esitystavat voivat toisaalta olla kompaktimpia ja tehokkaampia, mikä mahdollistaa nopeamman käsittelyn ja paremman suorituskyvyn. Lisäksi neuraaliset näkymän esitystavat voidaan oppia päästä päähän datasta, jolloin ne voidaan mukauttaa tiettyihin tehtäviin ja alueisiin.

Perinteiset rakenne-liikkeestä-, rakenne-syvyydestä- ja moninäkymägeometriatekniikat määräävät, miten ympäristön 3D-rakenne esitetään. Tämä väitöskirja esittelee arkkitehtuurit, jotka oppivat tämän esitystapa-avaruuden mahdollistaen sen, että tekstuurien, osien, esineiden, valojen ja näkymien olemassaolo voidaan tiiviisti ilmaista yhdellä vektorilla. Lisäksi menetelmät voivat ottaa huomioon näkymän sisällön ymmärtämiseen liittyvän epävarmuuden vaikeiden okkluusioiden ja osittaisten havaintojen yhteydessä.

Laajamittainen uuden näkymän synteesi pyrkii luomaan fotorealistisia kuvia mielivaltaisista kohteista 3D-avaruudessa. Aiemmat tutkimukset ovat tuottaneet kohdenäkymiä interpoloimalla säde- tai pikseliavaruudessa, ja ne kärsivät usein okkluusioista tai epätarkasta geometriasta johtuvista artefakteista. Tässä työssä ehdotetaan uusia, tehokkaita viitekehyksiä, jotka esittävät 3D-näkymiä monisyvyystasoina. Koulutettu malli osaa renderöidä väri- ja syvyyskuvia uusista näkymistä. Ehdotetut arkkitehtuurit ovat kompakteja ja tuottavat uskottavia tuloksia ennalta näkemättömään dataan perustuen ilman hienosäätöä tai testausajan optimointia.

Ihmisen kapturointi ja renderöinti on prosessi, jossa tallennetaan ihmisen ulkonäkö ja liike sekä luodaan realistinen 3D-esitys kyseisestä henkilöstä. Nykyiset menetelmät ratkaisevat tämän ongelman käyttämällä kalliita usean näkymän kuvankaappausasetelmia. Tämä väitöskirja keskittyy ongelmaan, jossa ennustetaan uusia näkymiä ennalta näkemättömästä dynaamisesta ihmisestä yhden kuvakulman avulla. Sen sijaan, että esitettäisiin syöte pistepilvenä, tämä työ esittelee tehokkaan pallopohjaisen näkymän synteesiverkon, joka tuottaa laadukkaampia tuloksia kuin monen näkymän lähestymistavat. Huolimatta siitä, että verkko on koulutettu pelkästään synteettisellä datalla, työ osoittaa myös erinomaista suorituskyvyn yleistyvyyttä todellisilla kuvilla.

Asiasanat: ihmisen synteesi, neuraaliset radianssikentät, pallopohjainen renderöinti, tasonpyyhkäisyvolyymi, uuden näkymän synteesi, vektoripohjainen esitystapa

To my family and friends

Acknowledgements

The research presented in this thesis was conducted in the Center for Machine Vision and Signal Analysis (CMVS) at the University of Oulu between 2018 and 2023. I am grateful to my supervisors, Professor Janne Heikkilä and Professor Esa Rahtu. Their guidance, feedback, and encouragement have been vital for completing this thesis. I am grateful for the freedom they gave me to explore my research ideas throughout my studies. At the same time, they knew when to re-direct me back to the appropriate path.

I want to thank my other co-authors Professor Jiri Matas from the Czech Technical University, and my fellow Ph.D. friends, Animesh Karnewar and Lam Huynh, for all the exciting discussions and ideas. Their constructive feedback helped me significantly improve the publications and my research. I also sincerely appreciate my colleagues, Nikolaos Sarafianos, Christoph Lassner, and Tony Tung, for their instructions during my first industrial internship at Meta Research Reality Labs. I also appreciate Professor Sanja Fidler, Sameh Khamis, Francis Williams, Zan Gojcic, and Or Litany at NVIDIA AI Research for mentoring me in the final stage of my Ph.D. I want to express my deepest gratitude to all my colleagues at CMVS, NVIDIA, and Meta, with whom I have worked.

I acknowledge the follow-up group members Professor Olli Silvén and Professor Susanna Pirttikangas for their feedback on my studies and research. I thank pre-examiners Postdoctoral Researcher Martin Danelljan from ETH Zurich and Professor Tobias Ritschel from University College London for their valuable feedback. While living in Finland, it has been my absolute pleasure to know so many remarkable people, and this would not be complete without mentioning some of you: Khanh. N, Nhat. V, Hoang. N, Le. N, Nhan. N, Tung. P, Janne.M, Snehal. B, Manuel. L, Zhou.S, Constantino. A, Praneeth.S, Silvia. Z, Miguel. L, Peng. W, Haoyu. C. I also want to thank Professor Nguyen Duc Toan for his kindness and altruism so I could come to South Korea and start chasing my dream.

My deepest gratitude goes to my family, especially my mother, for giving me love and life. Thank you for always being by my side, caring for me, and teaching me mathematics and various life lessons. Because of your endless efforts and encouragement, this journey is possible for me. Last but not least, I also would like to thank my girlfriend for her caring support and loving moments that gave me the strength to finish this thesis.

During the most challenging moments of my life, you showed me the light at the end of the tunnel.

Oulu, January 2023
Phong Nguyen Ha

List of abbreviations

x	*A 3D point in space*
d	*Viewing direction of a ray*
a	*The radius of a sphere*
f	*A mapping function*
ϕ	*The learnable weights of a neural network*
γ	*Positional encoding*
c	*Color of a 3D point*
σ	*Density of a 3D point*
I	*RGB image*
v	*Camera pose*
r	*The vector-based scene representation*
$\mathcal{N}$	*The Normal distribution*
z	*Latent code*
K	*Camera intrinsic matrix*
R	*3D rotation matrix*
t	*3D translation vector*
H	*Homography matrix*
V	*A volume of features*
F	*2D feature maps*
P	*Plane sweep feature volume*
D	*Depth maps*
Enc	*Encoder*
Dec	*Decoder*
Gen	*Generator network*
Dis	*Discriminator network*

NVS	*Novel view synthesis*
2D	*Two-dimensional*
3D	*Three-dimensional*
4D	*Four-dimensional*
5D	*Five-dimensional*

RGB	*Red, green and blue*
RGB-D	*Red, green and blue, depth*
RGB-α	*Red, green and blue, alpha*
IBR	*Image-based rendering*
SfM	*Structure from motion*
MVS	*Multi-view stereo*
MPI	*Multiple plane image*
LiDAR	*Light detection and ranging*
MLP	*Multi-layer perceptron*
ReLU	*Rectified linear unit*
NeRF	*Neural radiance fields*
GAN	*Generative adversarial network*
KL	*Kullback–Leibler divergence*
ELBO	*Evidence lower bound*
GQN	*Generative Query Network*
GAQN	*Generative Adversarial Query Network*
LSGAN	*Least Squares Generative Adversarial Network*
T-GQN	*Transformer-based Generative Query Network*
HVS	*Human view synthesis*
PSV	*Plane Sweep Volume*
ConvLSTM	*Convolution Long Short Term Memory*
SPADE	*Spatially-adaptive denormalization*
HRF	*Holistic radiance fields*
GPU	*Graphic processing units*
GT	*Ground truth*
VR	*Virtual reality*
AR	*Augmented reality*
FFC	*Fast Fourier convolution*
VGG	*Visual Geometry Group*
HD	*High defenition*

List of original publications

This dissertation is based on the following articles which are referred to in the text by their Roman numerals (I–V):

I Nguyen-Ha, P., Huynh, L., Rahtu, E. & Heikkilä, J. (2019, June). Predicting Novel Views Using Generative Adversarial Query Network. Scandinavian Conference on Image Analysis (SCIA). Norrköping, Sweden.
https://doi.org/10.1007/978-3-030-20205-7_2

II Nguyen-Ha, P., Huynh, L., Rahtu, E. & Heikkilä, J. (2020, November). Sequential View Synthesis with Transformer. Asian Conference in Computer Vision (ACCV). Kyoto, Japan.
https://doi.org/10.1007/978-3-030-69538-5_42

III Nguyen-Ha, P., Animesh, K., Huynh, L., Rahtu, E., Jiri, M., & Heikkilä, J. (2021, December). RGBD-Net: Predicting Color and Depth images for Novel Views Synthesis. International Conference on 3D Vision (3DV). Surrey, UK.
https://doi.org/10.1109/3DV53792.2021.00117

IV Nguyen-Ha, P., Huynh, L., Rahtu, E., Jiri, M., & Heikkilä, J. (2022, December). HRF-Net: Holistic Radiance Fields from Sparse Inputs. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). The paper is currently under a major revision.
https://arxiv.org/abs/2208.04717

V Nguyen-Ha, P., Nikolaos, S., Christoph, L., Heikkilä, J. & Tony, T. (2022, October). Free-Viewpoint RGB-D Human Performance Capture and Rendering. European Conference in Computer Vision (ECCV). Tel Aviv, Isarel.
https://doi.org/10.1007/978-3-031-19787-1_27

The author of this dissertation had the main responsibility of preparing articles I-V. This includes the implementation of the algorithms, experiments, and writing. The ideas presented in the articles were devised in group discussions with the co-authors, during which they provided valuable suggestions and feedback.

Contents

1 Introduction

1.1 Background and motivation

View synthesis generates a new view of a scene from one or more existing pictures of that scene. This can be used to create a novel view of a scene from a different perspective or to fill in missing regions in a current view. Novel view synthesis can refer specifically to the generation of a view that does not correspond to any real-world observation of the scene but rather represents a novel or fictional perspective. This can be used for a variety of purposes, such as virtual reality [1], special movie effects [2], and image-based rendering [3]. Many different techniques can be used for novel view synthesis, including 2D image warping, 3D reconstruction, and machine learning-based approaches.

The history of view synthesis can be traced back to the earliest days of computer graphics when researchers first began exploring ways to generate synthetic images from geometric models. One of the first notable efforts in this area was the work of Ivan Sutherland and his students at the University of Utah in the 1960s, who developed early 3D graphics systems that could generate synthetic views of simple geometric scenes [4]. In the decades that followed, there have been many significant developments in the field of view synthesis, including the introduction of more sophisticated geometric modeling techniques, the development of image-based rendering approaches that do not rely on explicit geometric models, and the application of machine learning techniques to view synthesis.

In recent years, there has been a particular focus on using deep learning to enable more realistic and flexible view synthesis, with many researchers exploring the use of generative models [5, 6] and other deep learning approaches to synthesize new views. There are several motivations for using neural scene representations for view synthesis. One motivation is to improve the realism of synthesized views. Traditional view synthesis methods often produce unrealistic results because they need to fully capture the complexity of scenes [7]. By using neural scene representations, which are trained on either synthetic or real-world data, it is possible to generate more realistic synthesized views [8]. Overall, the use of neural scene representations for view synthesis represents a promising direction for improving the realism and usefulness of novel views.

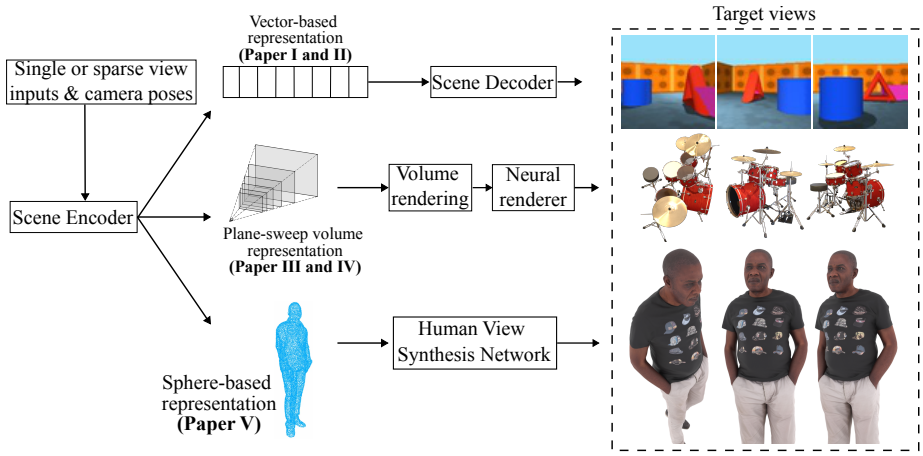

Fig. 1. An overview of different neural scene representations that have been used in this thesis for the topic of novel view synthesis using a set of sparse observations.

1.2 Scope of the thesis

This thesis utilizes several neural scene representations and deep learning techniques for novel view synthesis as can be seen in the Figure 1. The approaches in Paper I and Paper II propose an implicit vector-based representation of a synthetic 3D scene and then use a variational decoder to generate new views. Paper I presents two novel adversarial losses to improve the performance and help the learning process become much more stable. Paper II introduces a sequential synthesizing scheme to speed up the quality of the rendered novel images and fasten the training process. Paper III and Paper IV address the problem of novel view synthesis in large-scale real-world 3D scenes. Both papers described learning to construct memory efficient plane-sweep volumes and infer color and depth images of the unknown views. Paper IV proposes a generalized view synthesis approach utilizing both neural radiance fields and convolution neural networks. Finally, Paper V explores generating new views of an unseen dynamic human using a single RGBD image via sphere-based rendering. All methods presented above show plausible performance on unseen data without any scene-specfic fine-tuning or test-time optimization.

1.3 Contributions

The main contributions of the thesis are listed below.

– A general learning framework for novel view synthesis which introduces an additional discriminator network encourages the model to predict more accurate novel views. Moreover, the combination of the least square loss and the feature matching loss helps stabilize both the generator and discriminator training processes. (Paper I)
– A sequential novel-view synthesis approach renders more accurate distant novel views. Inspired by Transformers [9] architectures, the method can put more attention on the views that are useful to render novel views. Moreover, the proposed method requires less time to reach convergence than previous approaches. (Paper II)
– A generalized novel view synthesis method is presented that attains high-quality synthesis results for both seen and unseen data. Adaptive depth scaling that enables producing photorealistic novel views with and without per-scene optimization. A spatial-temporal module is proposed to produce a smooth sequence of rendered novel views along a continuous camera path. [1] (Paper III)
– An efficient sparse view synthesis network that employs a coarse radiance field predictor and a neural renderer to holistically predict novel images approximately two orders of magnitude faster than prior arts. The proposed scene-specific model requires only 10-15 minutes of fine-tuning to achieve high-quality results. IFuthermore, the proposed method does not require additional depth supervision. (Paper IV)
– A robust sphere-based synthesis network that generalizes to multiple identities without per-human optimization. Another refinement module is presented to enhance the self-occluded regions of the initial estimated novel views. This introduces a novel yet simple approach to establishing dense surface correspondences for the clothed human body. [2] (Paper V)

1.4 Overview of original articles

Paper I introduces a general learning framework for novel view synthesis that encodes input views into a latent representation used to generate a new view through a recurrent variational decoder. The proposed method improves the view synthesis by adding a least-square adversarial loss and a feature-matching loss. The experiments demonstrate that the trained model can produce high-quality results and faster convergence than the conventional approach. This paper received the Best Paper Award in the SCIA 2019 conference.

[1] https://github.com/phongnhhn92/RGBDNet
[2] https://www.phongnhhn.info/HVS_Net/index.html

Paper II addresses the problem of novel view synthesis sequentially. From the query pose and a set of input poses, an ordered set of observations that leads to the target view is created. The problem of single novel view synthesis is then reformulated as a sequential view prediction task. A Transformer-based architecture is then proposed to extract multiple scene representations. Experimental results show good performance on various challenging synthetic datasets and demonstrate that the trained model not only gives consistent predictions but also does not require any retraining for finetuning.

Paper III presents a cascaded architecture for novel view synthesis, called the RGBD-Net, which consists of two core components: a hierarchical depth regression network and a depth-aware generator network. The former predicts the target views' depth maps using adaptive depth scaling. At the same time, the latter leverages the predicted depths and renders spatially and temporally consistent target images. Moreover, the method can be optionally trained without depth supervision while retaining high-quality rendering. Thanks to the depth regression network, the extracted dense 3D point clouds are more accurate than those produced by some state-of-the-art multiview stereo methods.

Paper IV proposes a neural radiance fields-based view synthesis pipeline that renders the entire view in a single forward pass during training and testing. Instead of rendering the novel views directly at the original scale, the method infers lower-resolution radiance fields. Then it uses an up-scale neural renderer to obtain the final estimates. Although the generalized model can render plausible results on both unseen data, the trained model can further be optimized using more images to achieve state-of-the-art results.

Paper V takes as input a single, sparse RGB-D image of the upper body of a human and a target camera pose, and generates a high-resolution rendering from the target viewpoint. To account for the sparseness of the input depth, a sphere-based neural renderer is utilized to create a denser, warped image compared to simply performing geometry warping from one view to the other. Combined with an encoder-decoder architecture and trained end-to-end processing, this method can synthesize novel views of unseen individuals and inpaint areas that are not visible from the main input view.

1.5 Outline of the thesis

This thesis consists of an overview and an appendix containing the original articles described in the previous section. The remainder of the thesis is organized as follows. Chapter 2 presents a brief background to novel view synthesis and the recent learning-based methods that utilize neural scene representations. Chapter 3 introduces the

vector-based scene representations that render novel views of small-scale synthetic scenes. Chapter 4 tackles real-world and large-scale novel view synthesis by relying on plane sweep volume representations. Chapter 5 presents a generalizing dynamic human-view synthesis approach based on sphere-based rendering. Chapter 6 presents the conclusions.

2 Novel view synthesis

The fundamental goal of novel view synthesis is to render an image from a novel viewpoint, given a sparse set of reference images and corresponding camera poses. As opposed to traditional computer graphics, where the scene is constructed from hand-crafted 3D models, the goal of view synthesis algorithms is to use images captured from the real world as a medium for rendering scenes. The idea of using images as a rendering medium has remained a challenging and long-standing topic in academia due to a number of challenges. Some of the key challenges include camera pose estimation, inferring the geometric structure of the scene, modeling view-dependent lighting, and gracefully handling missing information. In addition to synthesizing plausible views, the practicality of the algorithm imposes additional challenges, such as achieving a high rendering speed and a low memory footprint.

2.1 Traditional view synthesis

The ways that novel view synthesis has been approached historically can be categorized into two approaches: light field rendering and image-based rendering. Light field rendering methods are based on directly interpolating between densely sampled images. Furthermore, image-based rendering (IBR) takes advantage of geometrical information derived from multi-view stereo to synthesize views.

2.1.1 Light-field rendering

Early work in novel view synthesis [10] defines the novel view synthesis problem as approximating a light field representation, which is a continuous 5-dimensional function that returns incoming radiance originating from the scene for a given position $x \in \mathbb{R}^3$ and view direction $d \in \mathbb{R}^2$.

The light field captures the flow of light in the scene, making it possible to render novel views by sampling it from desired locations. Given a collection of source views and camera poses that densely cover the parameter space, the light field can be discretized into a regular grid and interpolated between samples to synthesize novel views [11, 12]. This discretization can be further reduced to a 4-dimensional grid by noting that radiance remains constant along a ray in free space. This approach was termed the lumigraph,

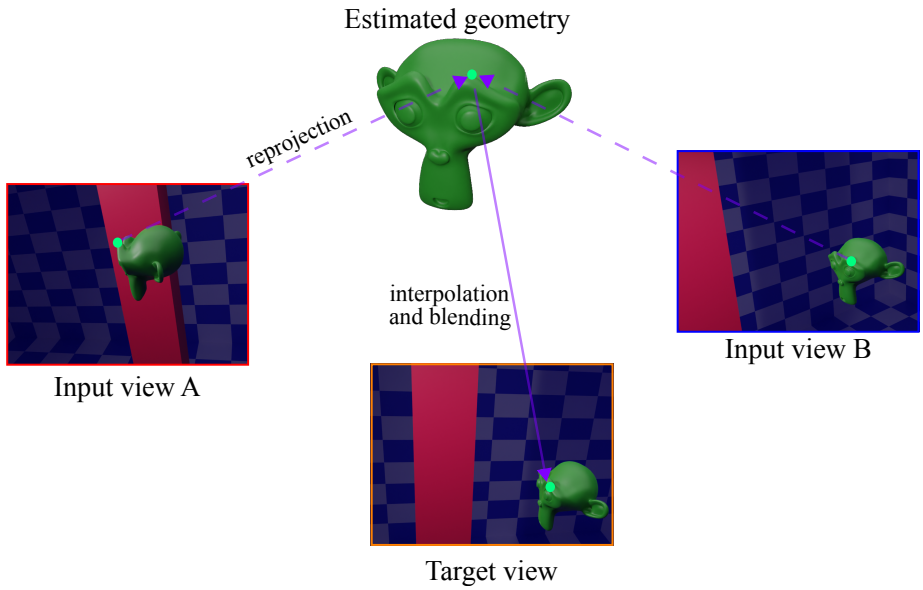

Fig. 2. An overview of image-based rendering. The scene geometry is first estimated using structure from motion or multi-view stereo. Each pixel of the target view is rendered by blending re-projected pixels from near-by input views.

in which the light field of an enclosed scene is discretized using bounding planes. However, this approach works only with a sufficiently dense collection of images. A more recent work [13] quantified this requirement by showing that light field rendering has a fundamental minimum sampling rate.

2.1.2 Image-based rendering

Image-based rendering refers to novel view synthesis techniques that utilize a geometrical estimate of the scene and nearby views to synthesize novel views. The development of structure from motion (SfM) and multi-view stereo (MVS) algorithms introduced a reliable way to estimate the camera poses and geometry of the scene from a set of images, which became a foundational technique for image-based rendering. The goal of SfM is to find the parameters of each camera and the 3D coordinates of shared points in the scene. The traditional way of solving this problem involves an image correspondence search and an incremental reconstruction procedure, explained in more detail [14]. The MVS methods build upon this technique by associating each pixel of the source views

with a globally consistent depth value [15]. The result is a dense point cloud that can be further refined into a textured mesh using e.g., Delaunay triangulation.

Traditional IBR algorithms utilize MVS to estimate the surface geometry of the scene and use it to reproject nearby images into a novel viewpoint [16]. As illustrated in Fig. 2, the reprojection process involves projecting the pixels of nearby pictures to the surface geometry and back to the target view. The reprojected pixels can overlap and end up in sub-pixel locations, which is handled by aligning the pixels with the target view utilizing, e.g., inverse bilinear interpolation and then using a blending operation to form the final image. The blending operation varies between authors; for example, Chaurasia et al. [17] specify the weights by camera orientation and the reliability of depth information at each pixel. Later studies improve the geometry estimation with various modifications to MVS, such as per-view meshes [18] and modeling depth uncertainty [7].

2.1.3 Limitations of traditional view synthesis

While the two discussed approaches can synthesize plausible views, they are limited in various aspects and have room for improvement. The approach of light field interpolation is only feasible when working with a sufficiently dense collection of images. Given the required sampling rate and the high memory cost of discretizing the parameter space, light field rendering only lends itself to large-scale scenes.

Image-based rendering is more scalable and can be integrated into existing graphics engines; however, there are several issues. Firstly, it is fundamentally limited by the accuracy of the reconstructed surface geometry. Scenes involving detailed geometry and view-dependent materials are incredibly challenging for SfM and MVS to reconstruct, producing noisy or spurious geometry and missing regions. The blending process is a crude approximation of how light can scatter from a surface, especially in scenarios with challenging light interactions such as reflection and refraction.

In summary, synthesizing views from scenes involving sparse views, detailed geometry, or view-dependent materials remains challenging for traditional view synthesis. Recent studies address the above mentioned issues by exploring several neural scene representations.

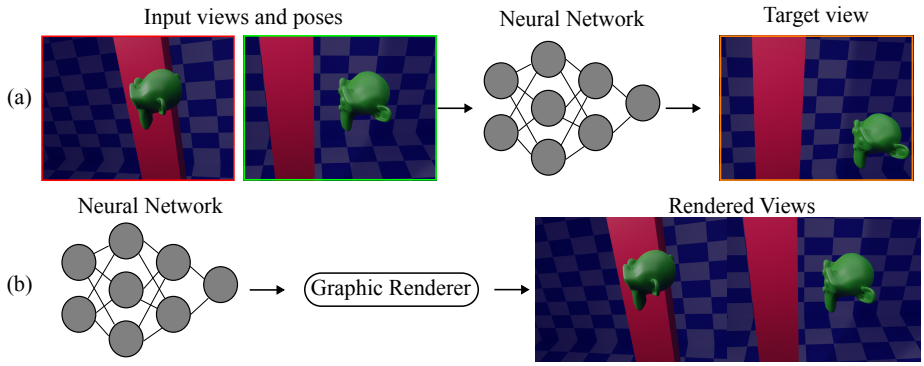

Fig. 3. An illustration of 2D (a) and 3D (b) learning-based view synthesis.

2.2 Learning-based view synthesis

Traditional computer graphics allows us to generate high-quality controllable imagery of a scene; all physical parameters of the scene, for example, camera parameters, illumination, and materials of the objects, need to be provided as inputs. To generate controllable imagery of a real-world scene, we need to estimate these physical properties from existing observations such as images and videos. This estimation task is considered inverse rendering and is challenging, especially when the goal is photo-realistic synthesis. In contrast, learning-based view synthesis is a rapidly emerging field that allows the compact representation of scenes, and rendering can be learned from existing observations using neural networks. Like classical computer graphics, the goal of neural rendering is to generate photo-realistic imagery in a controllable way, for example generating novel views, re-lightning, or scene decomposition.

As seen in Figure 3, there are two lines of research on learning-based view synthesis in 2D and 3D. The former approaches focus on training a neural network to render the target image from some nearby input views and their poses. The trained 2D neural rendering models can be generalized to new scenes without finetuning. In contrast, 3D neural rendering overfits a neural network to a single scene and uses a computer graphics engine to render the images. Unlike 2D neural rendering, the optimized network does not learn how to render, but it learns to represent the scene in 3D, and that scene is then rendered according to the physics of the image formation. Essentially, a neural rendering pipeline learns to generate and represent a scene from real-world imagery, an unordered set of images, or structured, multi-view images or videos. It does so by mimicking the physical process of a camera that captures a scene. A fundamental property of

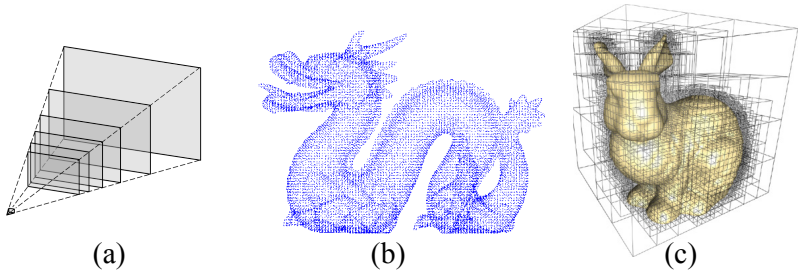

Fig. 4. An overview of some popular explicit scene representations for novel view synthesis: (a) multiple-plane images, (b) point-cloud and (c) voxel-grid (sparse octree).

learning-based view synthesis is the disentanglement of the camera capturing process (i.e., the projection and image formation) and 3D scene representation during training. This disentanglement has several advantages and leads especially to a high level of 3D consistency during the synthesis of images (e.g., for novel view synthesis). Modern neural rendering methods disentangle the projection and other physical attributes of the 3D scene by relying on different types of neural scene representation, and we will discuss them in the following section.

2.2.1 Explicit representations

Recent deep view synthesis methods learn to deal with 3D scenes using explicit 3D representations such as multiple plane images, voxel-grids and point clouds, as can be seen in Figure 4.

Multiple plane images. A significant number of studies [19, 20, 21] on view synthesis represent the 3D using Multiple Plane Images (MPIs). Each MPI includes multiple RGB-α planes, where each plane is related to a certain depth. The target view is generated by using alpha composition. Given a camera ray that intersects each image at plane depth, the projected color is formed by compositing each intersected color in a front-to-back fashion. A deep convolutional neural network is presented to predict MPIs that reconstruct the target views for the stereo magnification task [19].

Later work considerably improves the quality of synthesized images in the light-field setups [22]. They propose a novel network with a regularized gradient descent method to refine the generated images gradually. Local Light Field Fusion (LLFF) [23] introduces a practical high-fidelity view synthesis model that blends neighboring MPIs to the target view. Instead of using a set of simple RGB-α planes, the later work [24] models

view-dependent effects by parameterizing each pixel as a linear combination of basis functions learned from a neural network. This approach achieves the best overall scores across all significant metrics on frontal-facing scenes with faster rendering time than previous studies [19, 23]. In addition to multi-plane images, other researchers have considered alternative image formats such as spherical [25] and cylindrical [26] image representations to synthesize views from outward facing 360° scenes. This approach is essentially similar except for a different warping scheme and finds use in a stereo setup of 360° cameras. The critical advantage of multi-plane images is that they can be readily stored in image formats and integrated into graphics engines for real-time rendering without the need to evaluate a neural network. However, this representation could be more extensive regarding translational movement in the scene.

Point-based representation. Point cloud is a lightweight 3D representation that closely matches the raw data that many sensors (i.e. LiDARs, depth cameras) provide, and hence is a natural fit for applying 3D learning. A drawback of the point-based representation is that there might be holes between points after projection to the screen space. NPBG and its variant [27, 28] train a neural network to learn feature vectors that describe 3D points in a scene. These learned features are then projected onto the target view and fed to a rendering network to produce the final novel image. SynSin [29] lifts per-pixel features from a source image onto a 3D pointcloud that can be explicitly projected to the target view using a U-Net model. A more recent work [30] optimizes all of the scene's parameters such as the camera model, camera pose, point position, point color, environment map, rendering network weights, vignetting, camera response function, per image exposure, and per image white balance using a differentiable U-Net renderer. However, these point-based methods often suffer from temporal instabilities between generated novel views of a smooth camera path.

Sphere-based representation. Recent work on neural sphere-based rendering [31] introduces a fast, general purpose, sphere-based, differentiable renderer. Instead of treating the entire 3D scene as a set of 3D points, each sphere is parameterized by its position in space and its radius. Each sphere has an assigned opacity and can have an arbitrary vector as a payload, such as a color or a general latent feature vector. This makes it easy to handle point cloud data from 3D sensors directly, allows for the optimization of the scene representation without problems of changing topology, and is more efficient for rendering than recent approaches based on volumetric grids or fully-connected networks [8]. Moreover, the sphere-based representation eliminates the

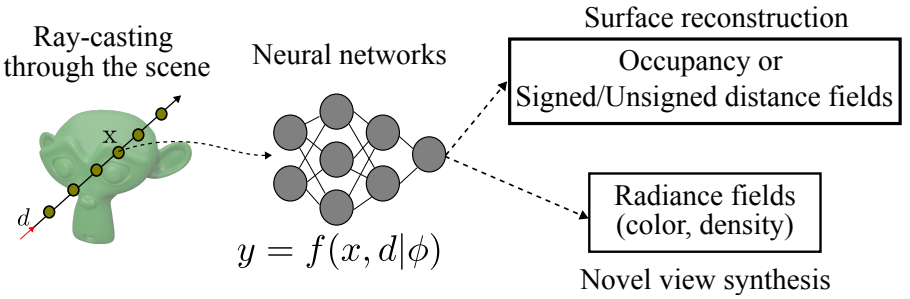

Fig. 5. An overview of implicit scene representations. A learned neural network f (parameterized by ϕ) is optimized to encoded a 3D point $x \in \mathbb{R}^3$ and its view direction $d \in \mathbb{R}^3$ to either binary occupancy and signed/unsigned distance fields for surface reconstruction or radiance fields for novel view synthesis.

need for acceleration structures, such as a k-d tree or octree, thus it can naturally support dynamic scenes.

Grid-based representation. Grid-based representations are similar to MPI representation but are based on a dense uniform grid of voxels. This representation has been used as the basis for neural rendering techniques to model object appearance. An early work [32] learns a persistent 3D feature volume for view synthesis and employs learned ray-marching on the task of object-centric reconstruction. Neural Volumes [33] is an approach for learning dynamic volumetric representations of multi-view data. The method consists of an encoder-decoder network that transforms input images into a 3D volume representation and a differentiable ray-marching operation that enables end-to-end training. However, the main limitation of these early approaches is the required cubic memory footprint.

Recent research addresses the above issue by utilizing sparse voxel octrees [34] to store and render voxels in a more efficient manner [35, 36]. While sparse voxel octrees are more challenging to work with than cartesian grids, they take significantly less memory to store and are faster to render. Another interesting approach is to learn a tri-plane representation [37] for novel view synthesis and generative modeling, and the features of each 3D point are obtained via tri-linear sampling. A tiny neural network then decodes those features to predict the 3D scene properties.

2.2.2 Coordinate-based representations

Conventional computer graphics algorithms and techniques developed assume meshes or point clouds as 3D scene representations for rendering and editing. Early work on 2D neural rendering [38] implicitly encodes the entire 3D scene into a latent vector and then uses a generator network to render novel views of the target viewpoints. Although this approach performs well on unseen synthetic data, it is not trivial to generalize to large-scale scenes. Recently, a large number of works have tried to represent and render a 3D scene using a coordinate-based mapping function such as a neural network. These recent studies show that the mapping function $f(x,d|\phi)$ between the spatial 3D coordinate x, view direction d, and the scene property y can serve as an implicit scene representation parameterized by a single neural network ϕ as can be seen in the Fig. 5. Such a new paradigm has drawn significant attention: one can train the neural network defined in a continuous and differentiable function space to recover fine-grained details at the scene scale with efficient memory consumption, which offers excellent benefits over the alternatives.

Surface reconstruction. Different types of implicit functions such as binary occupancy [39], signed, and unsigned distance functions [40, 41] are famous for this task. A common choice of network architecture for these representations is multi-layer perceptrons with ReLU activation. While ReLU is computationally convenient, it is found that using ReLU for this purpose causes the network to struggle with modeling high-frequency content. Fourier Features [42] tackles this issue by mapping the input coordinates x to a frequency domain $\gamma(x)$ as a pre-processing step. This mapping, known as a positional encoding, is found to increase the performance of multi-layer perceptron (MLP) when expressing low-dimensional functions. Alternatively, a later work [43] showed that using a sine activation function in place of ReLU achieves similar results.

Novel view synthesis. A recently emerging view synthesis method called Neural Radiance Fields (NeRF) [8] represents a continuous 3D scene in an MLP network f that can be queried using classical volume rendering. Essentially, a continuous scene is parameterized as a 5D vector-valued function which consists of a 3D location x and 2D viewing direction d. For each sampled 3D point along the ray, NeRF predicts its emitted color $c = (r,g,b) \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}^1$. The classical volume rendering technique [44] is utilized to accumulate those colors and densities and render 2D images. However, NeRF has to be evaluated at a large number of sample points along each

camera ray. This makes generating a full image with NeRF extremely slow. Despite the high quality of the synthesized novel images, NeRF also requires per-scene training.

Later, a vast amount of research has been dedicated to improving and extending neural radiance fields in various directions. For a complete list of neural field papers, there are several surveys [45, 46] that are useful to read. Recent volumetric aproaches [47, 48, 49] address the generalization issue of NeRF by incorporating a latent vector extracted from reference views. The feature of each sampled 3D point x is obtained via tri-linear interpolation. An MLP network encodes those features to get the radiance fields, including colors and densities. A volume rendering of NeRF is employed to obtain the novel views. Although these methods show generalization on selected testing scenes, they inherit the original method's slow rendering property.

Several studies address the slow rendering time since evaluating a multi-layer perceptron at each point in the ray is rather time-consuming. A recent work [50] introduces a student-teacher distillation scheme to effectively sample importance points along the rays. The first sampling network predicts suitable sample locations using a single evaluation per view ray, while the second shading network adaptively shades only the most significant samples per ray. Combining both networks shows a real-time rendering speed of 26 frames-per-second at a resolution of 1008×756 and a significantly small model size of 4.1 MB.

Considering the problem of novel view synthesis from only a set of 2D images, recent methods [51, 52] remove the requirement of known or pre-computed camera parameters, including both intrinsic and 6DoF poses. BARF [52] tackles the above problem by introducing a theoretical connection between classical image alignment to joint registration and reconstruction with NeRF. A coarse-to-fine registration is necessary for joint registration and reconstruction with coordinate-based scene representations. However, both of these methods can only optimize poses from scratch for wide-baseline 360-degree captures. GNeRF [53] achieves this by training a set of cycle-consistent networks (a generative NeRF and a pose classifier) that map from pose to image patches and back to pose again, optimizing until the classified pose of real patches matches that of sampled patches. They alternate this GAN training phase with a standard NeRF optimization phase until the result converges.

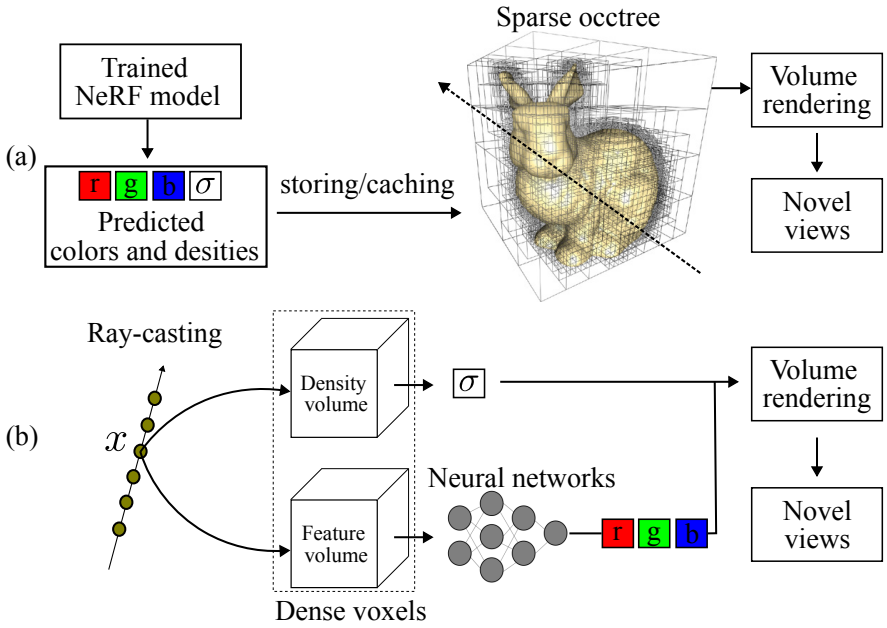

Fig. 6. An overview of hybrid scene representations for novel view synthesis: (a) storing/caching predicted radiance fields of a trained NeRF model to a sparse octtree and (b) optimizing both density, feature volumes and a tiny neural network to estimate radiance fields. Novel views are rendered via the volume rendering step of NeRF.

2.2.3 Hybrid representations

As seen in Section 2.2.2, a major drawback of coordinate-based representation is slow novel views rendering. It is therefore essential to improve the rendering speed without losing the high-fidelity output. Recent studies on neural rendering address this issue by combining fast-access explicit data structures such as sparse octree or dense voxel grids to render novel views efficiently (see Fig. 6).

The former approach [35, 54, 55, 56] utilizes a trained NeRF model to obtain a dense radiance field of the entire 3D scene. This method modifies how view-dependent colors are predicted to facilitate faster rendering and smaller memory requirements for the cached representations. PlenOctrees [54] queries the MLP to produce a sparse voxel octree of volume density, spherical harmonic coefficients and further finetunes this octree representation using a rendering loss to improve its output image quality. In contrast, the latter approach DirectVoxGO [57], proposes optimizing two explicit density and feature grids along with a small MLP network to obtain the radiance fields. Both methods above

significantly improve the rendering speed compared to NeRF, but the training times still take almost an hour per scene. Recent work called Instant-NGP [58] enables the training of a NeRF in a few seconds by exploiting a multi-resolution hash encoding instead of an explicit grid structure. Moreover, the proposed hash encoding is also generalized to several vision tasks such as gigapixel image approximation [59], surface reconstruction via signed distance functions [40], and neural radiance caching [60].

The above methods can synthesize high-quality images but require high-resolution feature grids to achieve good quality. This makes them less practical for graphics systems operating within tight memory, storage, and bandwidth budgets. Beyond compactness, it is also desirable for shape representation to dynamically adapt to the spatially varying complexity of the data, the available bandwidth, and desired level of detail. A recent work [61] on compression-aware neural fields proposes a vector-quantized auto-decoder that replaces bulky feature vectors with indices into a learned codebook. Experimental results show that the method can reduce the storage required by two orders of magnitude with relatively little visual quality loss without entropy encoding. Another interesting development [62] is to represent the 3D scene as a 4D tensor and factorize the tensor into multiple compact low-rank tensor components for efficient scene modeling. The trained model outperforms previous state-of-the-art methods and has a small model size ($<$ 4MB).

3 Vector-based neural scene representation

This chapter presents neural rendering approaches that compress the entire 3D scene as a vector. Conventional structure-from-motion, structure-from-depth, and multi-view stereo methods [63, 64] often represent the 3D scenes explicitly using point-clouds or meshes from sparse observations. Although these methods show good view synthesis results on much source data, they cannot recover the desired target views with a limited number of input images due to the ambiguity of 3D environments. Moreover, estimating the entire 3D scene structure may be more challenging than synthesizing novel images from new viewpoints.

The Generative Query Network (GQN) [38] tackles this problem by proposing an encoder-decoder network to predict the entire 3D scene given a few sparse observations (see Fig. 7). The encoder network first encodes input views into a single scene vector r. Since the encoder does not know which target view to render, it must find an efficient way of describing the actual layout of the scene as accurately as possible. It captures the essential elements, such as object positions, colors, and the room layout, in a concise, distributed representation. The encoder learns about specific objects, features, relationships, and environmental regularities during training. Instead of decoding the novel view from the encoded scene vector, the decoder is trained to produce a latent vector that matches the statistic of the ground-truth novel view. However, training GQN requires much synthetic data with perfect ground truths. Moreover, this method is known for its large memory consumption, and the predicted novel views are sometimes blurry. The proposed Generative Adversarial Query Network (GAQN) from Paper I and the Transformer-based Generative Query Network (T-GQN) from the Paper II improve the rendering quality of GQN by introducing two adversarial training losses and a sequential rendering scheme.

3.1 Generative query networks

In this section, we provide the reader with a brief background to GQN. Given the observations that include N source images $\{I_n\}_{i=1}^{N}$ and their corresponding camera poses $\{v_n\}_{i=1}^{N}$, GQN solves the view synthesis problem by using a encoder-decoder neural network to predict the target image I_q at an arbitrary query pose v_q.

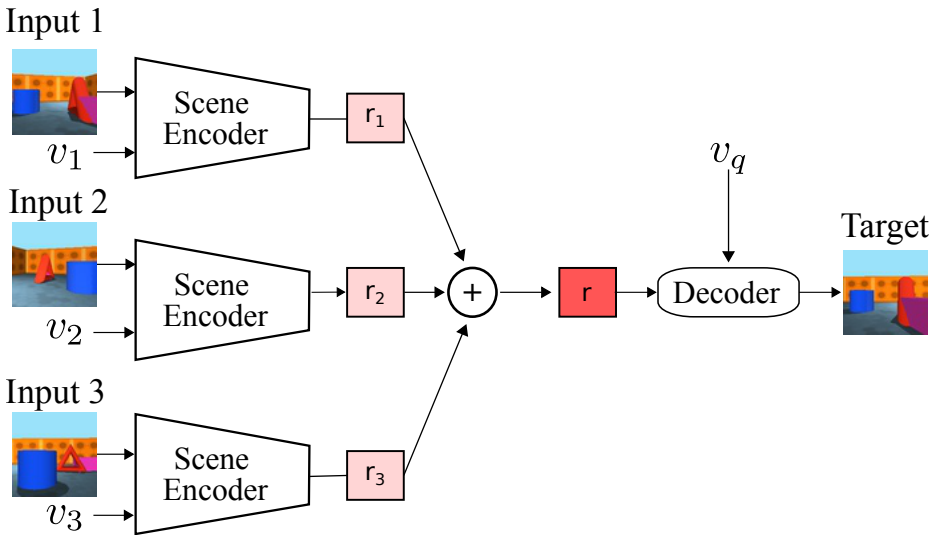

Fig. 7. An overview of a Generative Query Network (GQN). Adapted by permission, Paper I © 2019 Springer Nature.

First, the encoder is a feed-forward neural network that takes N observations as input and produces a single implicit scene representation $r = \sum_{n=1}^{N} r_n$ by performing a element-wise sum of N encoded scene representation r_n. The decoder then takes r and v_q as an input and predicts the new view I'_q from that viewpoint. The decoder network is a conditional latent variable model [65] which includes M pairs of Generation and Inference convolutional LSTM networks. At each generation step, the hidden state of the Generation and Inference LSTM core is utilized to approximate the prior π and posterior distribution q. Since the target view I_q is fed into the Inference sub-network, minimizing the Kullback-Leibler (KL) distance between π and q would help the Generation sub-network to produce an accurate result. Both the encoder and decoder networks are trained jointly to minimize the ELBO loss $\mathscr{L}_{GQN}$ function as follows:

$$\mathscr{L}_{GQN} = \left[-\ln \mathscr{N}(I_q|I'_q) + \sum_{m=1}^{M} KL\Big[\mathscr{N}(q_m)||\mathscr{N}(\pi_m)\Big]\right]. \tag{1}$$

3.2 Generative adversarial networks

Before introducing the proposed Generative Adversarial Query Network of Paper I, a brief introduction of Generative Adversarial Networks (GANs) is given. This method consists of two competing architectures referred as a generator (*Gen*) and a discriminator

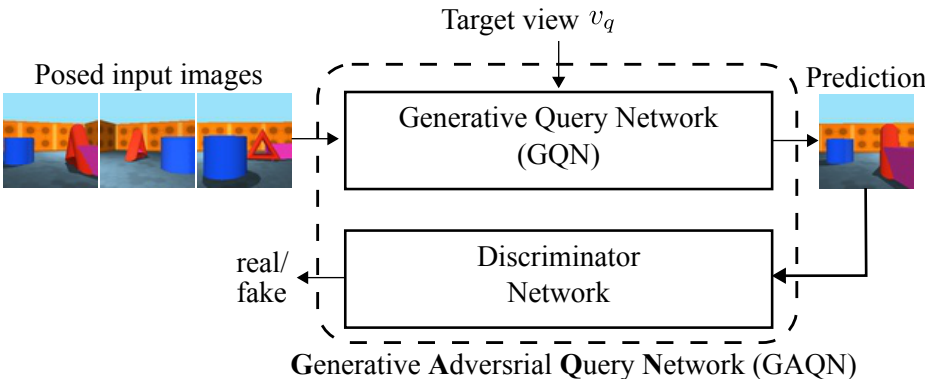

Fig. 8. Illustration of a Generative Adversarial Query Network (GAQN). The proposed network has an additional Discriminator network to further enhance the quality of the predicted novel views. Adapted by permission, Paper I © 2019 Springer Nature.

(*Dis*). The generator *Gen* maps a given latent representation z (possibly a vector with random values) into a novel image $x' = Gen(z)$ that is then passed to the discriminator network *Dis*. The discriminator aims to determine if the given sample is produced by the generator, or if it is a real image taken from the training set. Denoting the real training samples as x, the conventional generator loss L_G^{GAN} and discriminator loss L_D^{GAN} are defined as:

$$L_G^{GAN} = -\mathbb{E}_{z\sim P_z}[\log Dis(Gen(z))], \tag{2}$$

$$L_D^{GAN} = -\mathbb{E}_{x\sim P_x}[\log Dis(x)] - \mathbb{E}_{z\sim P_z}[\log(1 - Dis(Gen(z))]. \tag{3}$$

Both networks are trained simultaneously in an alternating fashion. In the ideal case, the procedure guides the generator to produce images that are indistinguishable from the training image distribution. However, in practice the training procedure is challenging due to various problems such as mode collapses [66].

3.3 Generative adversarial query networks

The proposed Generative Adversarial Query Network (GAQN) of the Paper I builds on top of the GQN architecture by introducing two GAN training losses. As illustrated in Fig. 7 and 8, the proposed GAQN consists of three components: an encoder network *Enc*, a decoder network *Dec*, and a discriminator network *D*. The GAQN utilizes the same *Enc* and *Dec* architecture as standard GQN to generate a novel view $I'_q = Dec(Enc(x,p),p_q)$. However, inspired by the aforementioned GAN methods, we include an additional

discriminator network D to distinguish between the generated fake images from the GQN and the ground truth view in the training data. In this way, the discriminator is trained to enhance the image fidelity of the GQN model.

3.3.1 Applying least-square adversarial loss

Numerous studies [67, 68] have shown that the training of the GAN may be unstable due to the vanishing gradients problem caused by the binary cross entropy loss of the vanilla GAN approach (see Equation 2 and 3). In this work, we avoid the problem by adopting the least-square loss function from the previously proposed Least Squares Generative Adversarial Networks (LSGANs) [69]. The idea of LSGANs has proven effective since it tries to pull the fake samples closer to the decision boundary of the least-square loss function. Based on the distance between the sampled data and the decision boundary, LSGANs manage to generate better gradients to update its generator. Furthermore, LSGANs also exhibit less mode-seeking behavior, stabilizing the training process. Equation (4) and (5) provide the generator and discriminator loss of LSGANs that we use to train the GQN decoder and our proposed discriminator as follows:

$$\mathscr{L}_G^{LSGAN} = \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}}\Big[(Dis(I'_q)-1)^2\Big], \tag{4}$$

$$\mathscr{L}_D^{LSGAN} = \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}}\Big[(Dis(I'_q))^2\Big] + \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\Big[(Dis(I_q)-1)^2\Big]. \tag{5}$$

3.3.2 Applying feature-matching loss

Inspired by the recent studies [68] on improving the stability of the GAN training, we add an extra feature matching loss to train the generator network. The main idea of this feature matching loss is to use the discriminator network as a feature extractor and guide the generator to generate data that matches the feature statistics of the real data. There are several approaches to exploiting the feature matching loss in training the generator.

Specifically, the common GAN generator loss is being replaced by a mean feature matching loss. It has been argued that this mean feature matching loss helps prevent the gradient vanishing problem during the training. In our research, we have already adopted the least square loss to address the above problem, but there is no guarantee that the problem is completely solved. Therefore, we train our GAQN generator network using a unified loss function as the combination of LSGANs generator loss $\mathscr{L}_G^{LSGAN}$ and

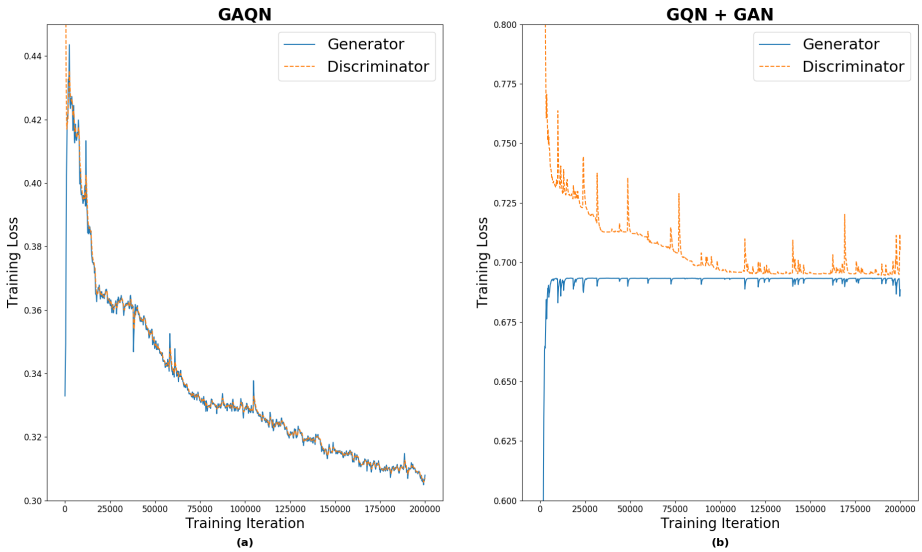

Fig. 9. Comparison of generator and discriminator training loss between our proposed GAQN (a) and GQN + GAN (b). Both the generator and discriminator of GQN+GAN are suffering from mode collapsing and vanishing gradients. Using the least-square loss and the mean feature loss, our GAQN achieves a stable learning process. Reprinted, with permission, from Paper I © 2019 Springer Nature.

mean feature matching loss $\mathscr{L}_{FM}$. Let $f_D()$ be the mean of the output feature maps from the 3^{rd} layer of the discriminator network, the mean feature matching loss is define as follows:

$$\mathscr{L}_{FM} = ||\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} f_D(x_{gt}) - \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}} f_D(x')||_2^2. \tag{6}$$

3.3.3 Discussion

Paper I applies both the least-square GANs and the discriminator feature matching losses and has shown significantly better novel views than those produced by the original GQN method. In addition, these two new losses also contribute to faster and more stable training. In Figure 9, there is a clear difference in the training loss landscape between GAQN and GQN+GAN. The training procedure of GQN + GAN is highly unstable due to mode collapsing and diminishing gradients. In contrast, the GAQN model eliminates both problems by using the least-squares loss and the discriminator mean feature loss.

Figure 10 illustrates the predicted novel view using the plain GQN architecture and the proposed GAQN architecture. Although the obtained GQN model successfully

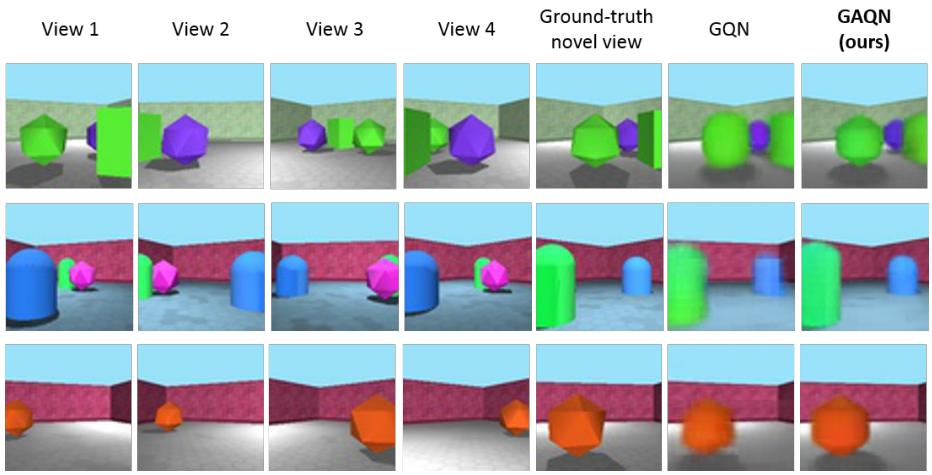

Fig. 10. Comparison between generated a novel view using the GQN and the GAQN architectures. Reprinted, with permission, from Paper I © 2019 Springer Nature.

predicts the correct location and color of three objects in the 3D scene, their edges are blurry. Meanwhile, GAQN produces sharper object edges, especially to the middle green icosahedron.

3.4 Transformers-based generative query network

Despite having better results than GQN, the proposed GAQN inherits the long training convergence and low-quality novel views at the distant target pose. In this section, a Transformer-based Generative Query Network (T-GQN) is introduced to infer the underlying 3D scene structure and faithfully produce the target view even at a distant query pose. GQN and GAQN use a simple summation function to represent the entire 3D scene as a single implicit representation. Although they successfully render the target view, a large amount of data is required for training, which takes a long time to converge. Moreover, both methods focus on synthesizing only a single target image. They are inefficient when rendering target images for distant query poses subject to strong geometric transformations and severe occlusions.

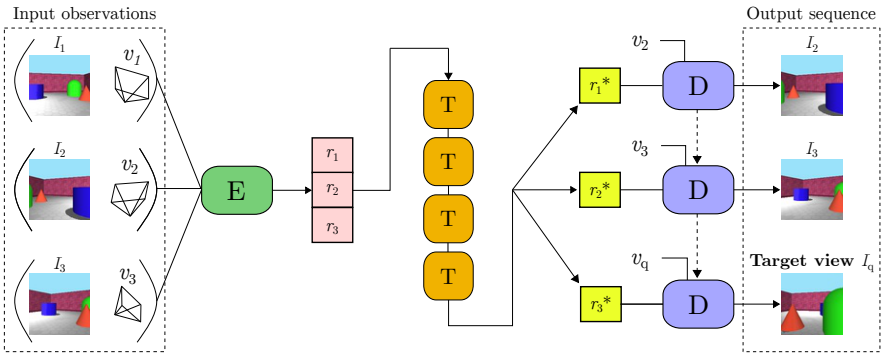

Fig. 11. Illustration of the Transformer-based Generative Query Network (T-GQN). Reprinted, with permission, from Paper II © 2020 Springer Nature.

3.4.1 *Sequential view synthesis*

The problem of single view synthesis can be redefined as a problem of sequential view synthesis by predicting a sequence of N novel views $S_{out} = \{I'_2, ..., I'_N, I'_q\}$ from a sequence of observations $S_{in} = \{(I_1, v_1), ..., (I_N, v_N)\}$. Since we have N different target views, having N other scene representations would be beneficial. To have multiple implicit scene representations, a Transformer Encoder [9] is utilized to learn the dependencies between input observations at each rendering step. As seen in Figure 11, a set of scene representations r is fed as input to the Transformer Encoder and produces another set of enhanced scene representations r^*_n that are trained to exploit the multi-view dependencies.

As shown in Fig. 12 (a), GQN predicts the target view I_q in a single rendering step. If the query pose v_q is distant from all input poses, then the target view might look completely different from all input views. In this case, it is non-trivial for GQN to generate a plausible target view, and it might take a long training time to reach the convergence. If the model is able to predict an input view I'_n for $n > 1$ based on previous input data $\{(I_1, v_1), \ldots, (I_{n-1}, v_{n-1})\}$ then it also renders the target view I_q at the query pose v_q provided that the camera poses $\{v_1, \ldots, v_N, v_q\}$ have been organized as a sequence where the adjacent poses are the closest ones. The proposed T-GQN model is trained using multiple rendering steps to achieve such ability. Each rendering step is identical to the GQN except that different sets of input observations and query poses are fed as input to the network. Fig. 12 (b) illustrates these sets of input observations at each rendering step with boxes of different colors.

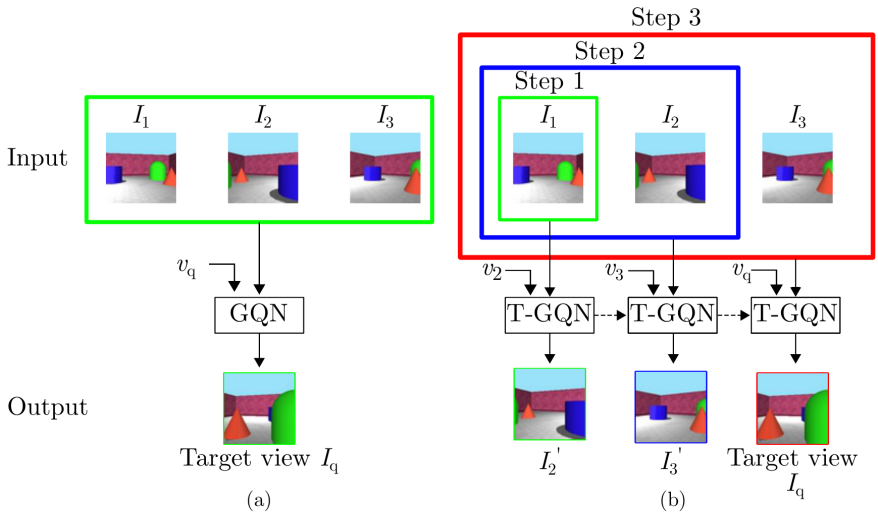

Fig. 12. An illustration of the single view synthesis (a) compared to our proposed sequential view synthesis (b). Reprinted, with permission, from Paper II © 2022 Springer Nature.

3.4.2 Discussion

In novel view synthesis, the input images are often randomly captured, and they would make rendering a distant target view much more challenging than those predicted by nearby photos. Paper II tackles this issue using the sequential rendering scheme and multiple scene representations. The proposed Transformer architecture's attention score further demonstrates its effectiveness in both cases when the camera movement is restricted or not. In both cases, the model is trained to put the higher score on views that are the most relevant to render at each rendering step (see Figure 13).

Finally, the effectiveness of the multi-step scene representation is presented in Figure 14. In this case, the T-GQN model of Paper II significantly outperforms both GQN and a variant SeqGQN, which uses a single representation r as an input to the decoder network. Although both GQN and SeqGQN use the exact aggregated scene representation r, SeqGQN can produce better and more accurate target views than the baseline. This result demonstrates that sequentially approaching the neural rendering leads to more precise view synthesis.

However, the proposed method manages to hallucinate the novel views given a few sparse observations. Experimental results show that both GAQN and T-GQN struggle to render photo-realistic images compared to those estimated using multi-view

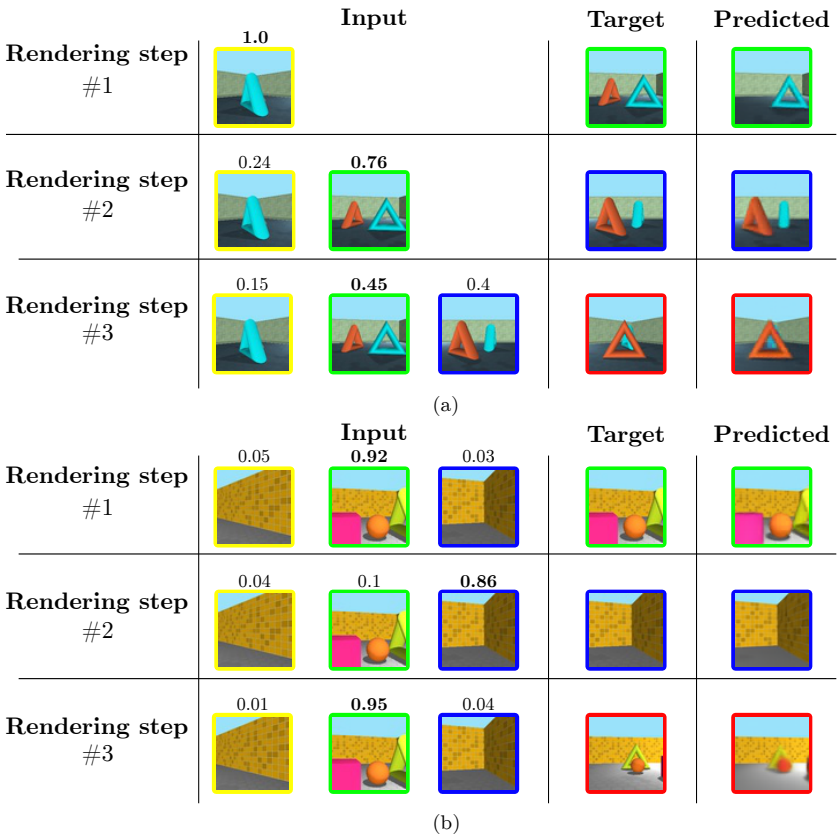

Fig. 13. Visualization of multi-view attention scores at each rendering step produced by our methods when (a) the camera movement is restricted and (b) the camera is free to move. Note that the order of the input sequence does not affect the learned multi-view attention scores. Reprinted, with permission, from Paper II © 2020 Springer Nature.

data. Future research should address the problem of lifting a single image to a 3D object and demonstrate the ability to generate a plausible 3D object with 360° views that correspond well to the given reference image. This would further shed light on a promising direction of easing the workflows for creative artists and designers.

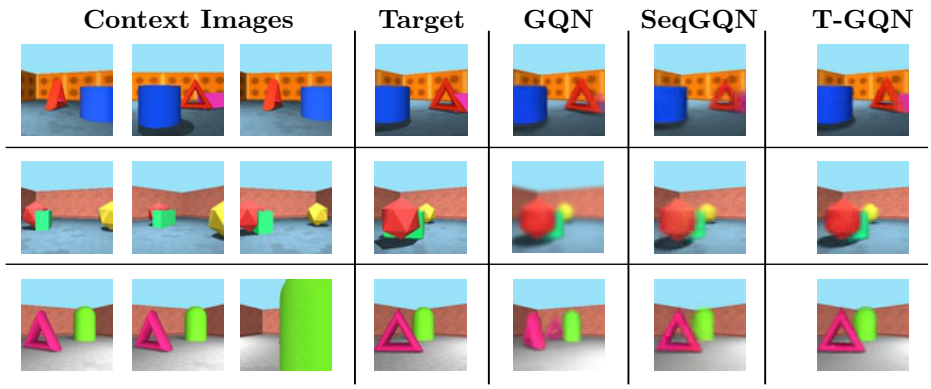

Fig. 14. Example of generated novel views compared between our proposed T-GQN and variants on the RRC dataset. Reprinted, with permission, from Paper II © 2020 Springer Nature.

4 Plane sweep volume representation

The previous chapter introduced two novel view synthesis networks by compressing the entire 3D scene as a single vector. Although those networks can hallucinate the full synthetic scenes given a few sparse observations, it is non-trivial to generalize to real large-scale images with higher complexities. This chapter focuses on the explicit Plane Sweep Volume (PSV) representation which serve as the medium to synthesize new views. Both Paper II and Paper IV utilize this representation and achieve state-of-the-art results on the benchmark real and synthetic datasets.

Early studies on view synthesis with deep learning already use PSV [70] for novel view synthesis. Each input image is projected onto successive virtual planes of the target camera to form a PSV. Kalantari et al. [71] calculate the mean and standard deviation per plane of the PSV to estimate the disparity map and render the target view. The inputs to the recent MPI-based methods [23, 22] are also PSVs. However, those PSVs are constructed on a fixed range of depth values. In contrast, the proposed RGBD-Net of Paper III builds multi-scale PSVs which use adaptively sampled depth planes. In Paper IV, the PSVs can also be used to efficiently infer the neural radiance fields of the entire novel views.

4.1 Predicting color and depth images for novel views synthesis

This section first describes the differentiable homography warping step to obtain a PSV of the novel view. The volume is later utilized by the deep neural network of Paper III called RGBD-Net (see Fig. 15), which comprises two modules. The former network is a hierarchical depth regression network (Section 4.1.2) that estimates the multi-scale depth map of the novel view, and the latter depth-aware refinement network (Section 4.1.3) enhances the warped images to produce the final target image.

4.1.1 Constructing plane sweep volume via homography warping

Both Paper III and Paper IV propose to generate cost volumes for the multi-view images by adopting the traditional plane stereo [72]. The basic idea of constructing a PSV is to back-project the input image onto successive virtual planes in the camera frustum of the

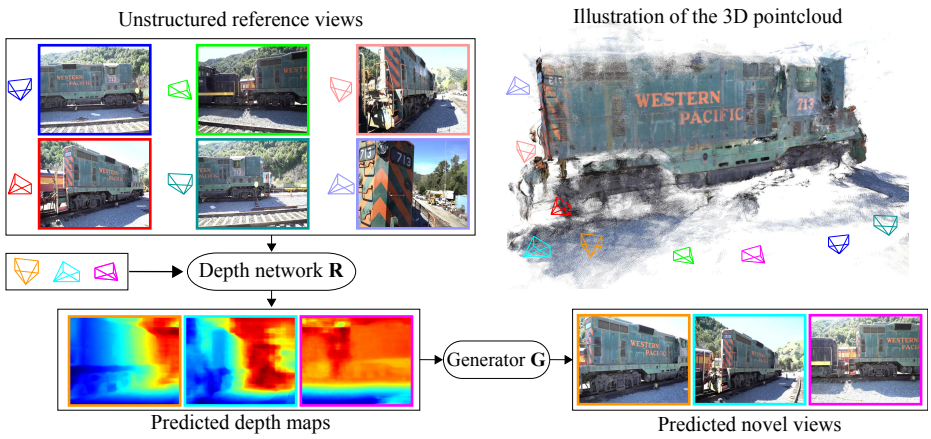

Fig. 15. The overview architecture of the proposed RGBD-Net. Reprinted, with permission, from Paper III © 2021 IEEE.

novel view and measure the photo consistency among the warped volumes for each pixel.

For simplicity, I_n and $v_n = \{K_n, R_n, t_n\}$ are denoted as the n^{th} input image and its camera pose, I_q and $v_q = \{K_q, R_q, t_q\}$ are denoted as the target image and its camera pose and K, R, t are the camera intrinsics, rotation matrix and translation vector of each viewpoint. A set of M virtual planes perpendicular to the z-axis of the novel view are first uniformly sampled in the inverse-depth space. All input images I_n are warped into those planes to form N volume $\{V_n\}_{n=1}^{N}$. The coordinate mapping from the warp feature map $V_n(d_m)$ to I_n at depth d_m is determined by the plane transformation $\hat{x} \sim H_n(d_m) \cdot x$, where $\sim$ denotes the projective equality and $H_n(d_m)$ is the homography between the input and target poses at the depth d. Let n_q be the principle axis of the novel view, the homography is expressed by a 3×3 matrix as follows:

$$H_n(d_m) = K_q \cdot R_q \cdot (I - \frac{(t_q - t_n) \cdot n_q^T}{d}) \cdot R_n^T \cdot K_n^T. \tag{7}$$

Both Paper III and Paper IV use bilinear interpolation to sample deep extracted features rather than using raw RGB pixels of the input views. As the core step to bridge the 2D feature extractor and the 3D rendering decoder, this warping operation is implemented differentiably, enabling end-to-end training of novel view synthesis.

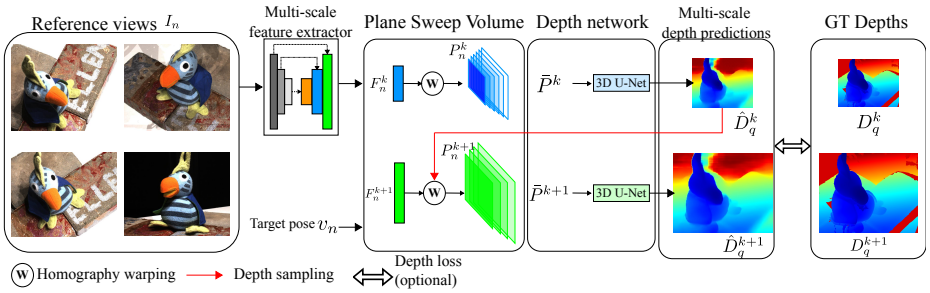

Fig. 16. Illustration of the depth regression network of the RGBD-Net. Reprinted, with permission, from Paper III © 2021 IEEE.

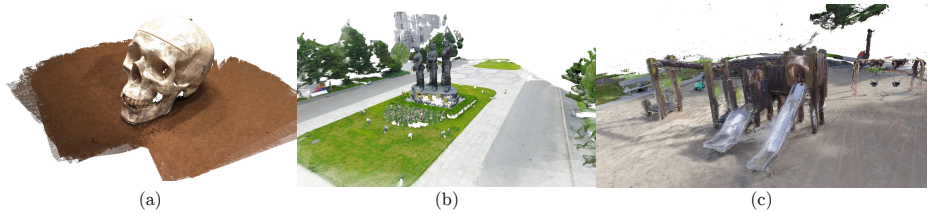

Fig. 17. Predicted pointcloud of the RGBD-Net on various scale dataset: (a) DTU datasets [76], (b) Blended MVS dataset [77], (c) Tanks & Temples dataset [78]. Adapted by permission, Paper III © 2021 IEEE.

4.1.2 Depth regression network

In this section, the pipeline of the depth regression network to estimate the depth map of the target view as illustrated in Figure 16 using a set of unstructured input images and their poses. Each input view is I_n first fed to the Feature Pyramid Network [73] to extract K multi-scale features F_n^k. We then apply homography warping (Section 4.1.1) to each input feature map to construct a PSV P_n^k of the target view with a set of M_k hypothesis depth planes. At each scale, a mean PSV volume $\bar{P}^k$ between all input views is fed to a 3D Convolution U-Net to estimate a coarse novel depth map $\hat{D}_q^k$. Inspired by the recent work on multi-view stereo [74], we estimate the depth map of the novel views in a coarse-to-fine manner. A depth plane resampling technique is proposed from the MVS literature [75] to efficiently sample near-surface depth planes using the predicted coarse novel depth map prediction. This also helps reduce the GPU memory required to process high-resolution feature maps. As can be seen in Figure 16, the method can estimate high-resolution depth maps $\hat{D}_q^{k+1}$ using the coarse estimates.

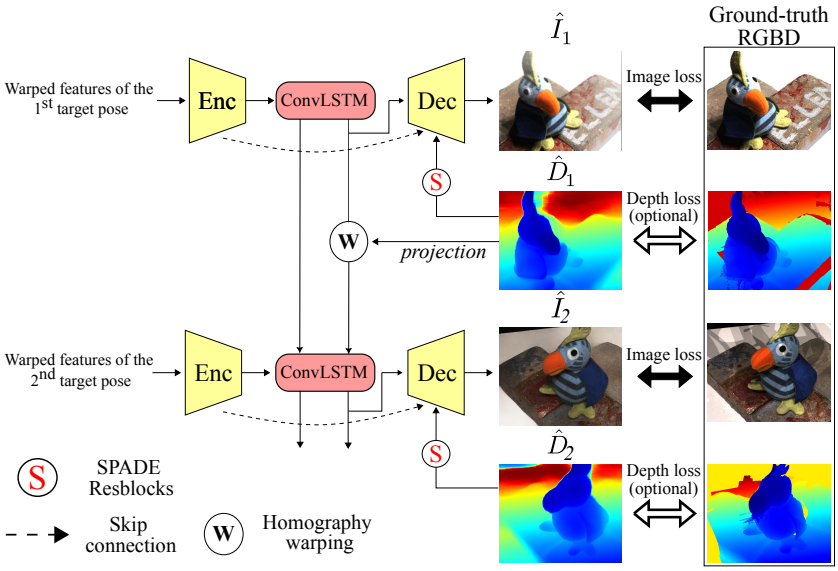

Fig. 18. Illustration of the depth-aware refinement network of the RGBD-Net. Reprinted, with permission, from Paper III © 2021 IEEE.

The proposed coarse-to-fine depth estimation network is first trained using the object-centric DTU dataset [76] from scratch, and it obtains state-of-the-art results on the testing set of the same source data. However, we observe that the model needs help to generalize to the large-scale scenes from the BlendedMVS [77] and Tanks & Temples [78] datasets. Therefore, we propose an adaptive depth scaling method that normalizes the GT depths of those large-scale datasets to have a similar range to the DTU dataset during training. During the testing time, we then scaled the depth predictions back to the original range. As seen in Figure 17, this technique allows the model to generalize and produce high-quality 3D scene reconstruction on unseen data.

4.1.3 Depth-aware refinement network

A depth-aware refinement network is proposed using the predicted depths of the regression network above to produce the novel views $\hat{I}_q$ since the RGBD-Net uses a set of sparse reference views to synthesize a novel view. Generating videos along smooth camera paths is therefore potentially subject to temporally inconsistent predictions and flickering artifacts due to the independent rendering at each new viewpoint. Inspired

by the sequential view synthesis of Paper II, this refinement network (see Figure 18) also consists of multiple rendering steps that produce Q sampled novel views. Each rendering step of the RGBD-Net consists of feature fusion and view synthesis.

The former step leverages the regressed depth maps to warp multi-scale input features to the novel views via bilinear interpolation. Experimental results show that naively concatenating or summing those warped features often leads to sub-optimal and blurry view synthesis results. Applying a Transformer-based architecture to learn a unified representation leads to a bigger model. An inverse depth-based fusion method is therefore introduced to add different weights to those warped features. As the depth network gets better at predicting depth maps, so does the feature fusion method.

The latter view synthesis step uses the fusion features and the positional encoding of the viewing direction as input to a U-Net with a shared ConvLSTM in the bottleneck. This allows our model to perform long-range dependencies between different target views and achieves spatio-temporally consistent novel images. Instead of directly using the hidden state of the previous rendering step, the predicted depth maps are leveraged to warp the hidden state to the current step. This also ensures that the predicted novel image has similar sharp edges as the depth map produced by the depth regression network. Moreover, the trained RGBD-Net model also shows outstanding generalization performance on the unseen dataset, as seen in Figure 19. Although both FVS [79] and NeRF++ [80] are trained on the Tanks and Temples [78] dataset, the proposed RGBD-Net of Paper III still outperforms both of them both qualitatively and quantitatively.

4.2 Cascaded and generalizable neural radiance fields

In the previous work of Paper III, the method relies on a ConvLSTM network to render Q targets and enforce the view consistency between those rendered images. There is a drawback: it would take a long training time to reach the convergence and achieve view-consistency novel images. Recent studies [47, 48, 81] on generalized novel view synthesis alleviates this problem by leveraging the continuous and implicit NeRF representation. Instead of directly predicting the depth maps from the PSVs, these methods learn neural radiance fields that extract consistent depths and color images via volume rendering [44]. Moreover, they also train a generalized model using multiple scenes and then perform 15- to 1-hour test-time optimization to achieve photo-realistic renderings on a single scene.

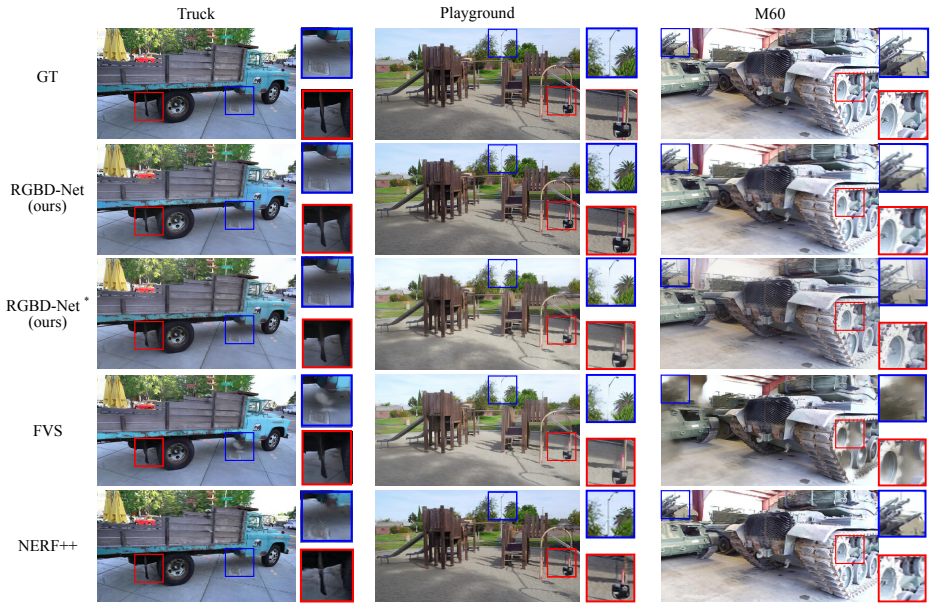

Fig. 19. Example of the generated novel views by the RGBD-Net and state-of-the-art methods for three scenes from the Tanks and Temples dataset. Reprinted, with permission, from Paper III © 2021 IEEE.

4.2.1 Coarse radiance fields predictor

Despite having state-of-the-art results, the NeRF-based generalized methods above inherit the slow rendering property of the original NeRF method. MVSNeRF [81] first infers a low-resolution PSV from a few sparse observations and then decodes such volume into high-resolution novel views. The decoding step requires independently querying each pixel's radiance fields using a learned Multi-Layer Perception (MLP) network. Therefore, rendering the entire high-resolution images is very expensive, and it takes 15 seconds to obtain a single image. Paper IV addresses this issue by introducing a Cascade and Generalizable Neural Radiance Fields (CG-NeRF) method that significantly speeds up the rendering and also achieves state-of-the-art results on both the seen and unseen datasets as can be seen in the Figure 20.

Similar to the RGBD-Net of Paper III, the first step is to estimate the depth map of the target view. However, instead of regressing multi-scale depth maps, a coarse radiance fields predictor is utilized to infer the novel view's depths and deep features in the lower resolution. Instead of using the costly 3D U-Net of the RGBD-Net, a

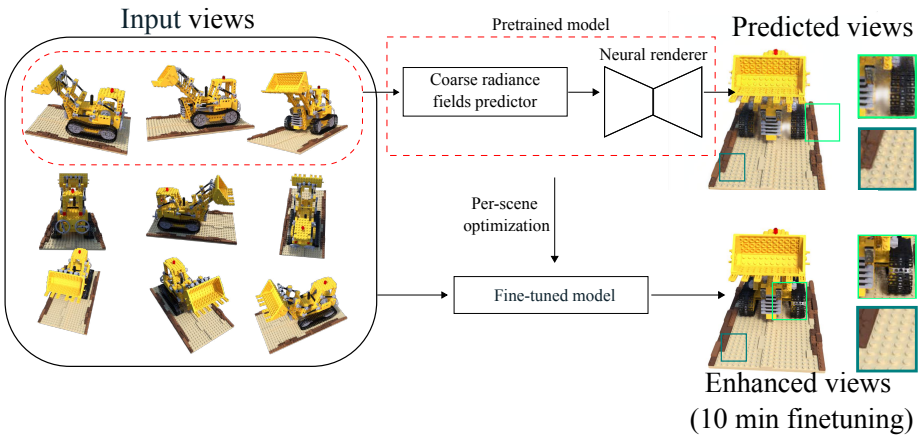

Fig. 20. An overview of pretrained and scene-specific optimization of the Cascade and Generalizable Neural Radiance Fields (CG-NeRF). Adapted by permission, Paper IV © 2021 IEEE.

memory efficient Transformer-based MobileViT block [82] is utilized to regularize this cost volume. This network learns the long-range dependencies via the multi-head attention [9] between the non-overlapping patches of multiple input PSVs. The output of the MobileViT is a unified coarse volume $\hat{V}$ that can be used to infer the density σ and radiance features f of the query point x and its viewing direction d. By accumulating σ and f of every 3D point along the rays, the entire low-resolution depths D and features F are obtained from a single forward GPU pass.

4.2.2 Up-scaling neural renderer

The method presented in Paper IV feeds the coarse estimation of the scene geometry and appearance of the above network to an up-scaling auto-encoder network and renders the higher-resolution target images. However, it is challenging to generate such images using coarse features. Paper IV also leverages the depth plane resampling technique and feature fusion of Paper III to obtain the near-depth features. Instead of progressively predicting depth maps from coarse to fine, the coarse depth map is bilinearly up-sampled to match the original resolution. Both coarse and fine features are then fed to the auto-encoder network. All modules of the CG-NeRF model are trained by photometric losses between the high-resolution predicted and GT novel views. An additional coarse

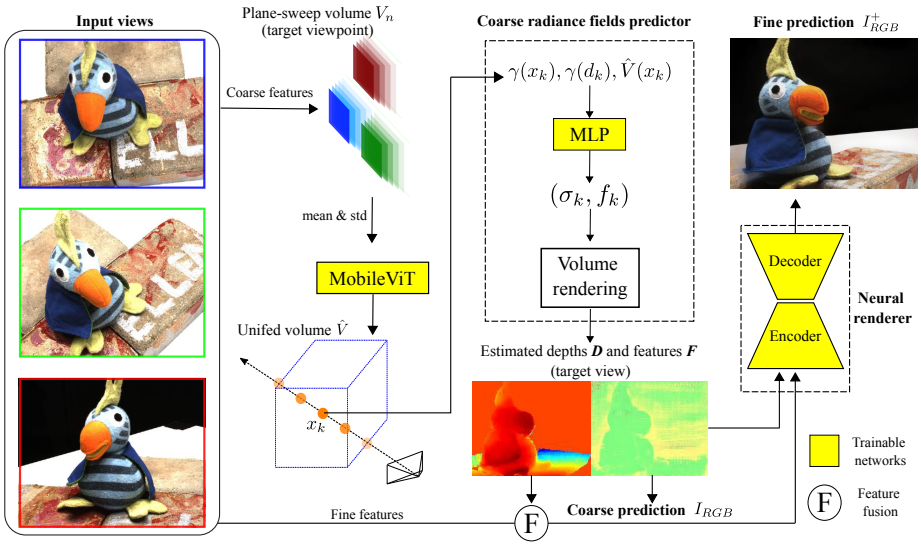

Fig. 21. The proposed CG-NeRF model comprises several parts: 1) a memory-efficient MobileViT architecture that fuse multiple low-resolution plane-sweep volumes of the target viewpoint into a single unified volume $\hat{V}$, 2) a coarse radiance fields predictor that estimates target depth and features in low resolution, and 3) an auto-encoder network to render novel views at the original resolution. Reprinted, with permission, from Paper IV © 2022 IEEE.

RGB loss and an adversarial loss [83] are added to further supervise the coarse radiance fields predictor and enforce the view consistency.

4.3 Discussion

Paper III and Paper IV introduced novel approaches towards real-time and generalizable novel view synthesis using a set of sparse observations. They show several advantages over prior arts, including 1) a high accuracy 3D scene reconstruction is obtained as a side product of the view synthesis process and 2) a good performance base model that can be further finetuned for a short amount of time to achieve state-of-the-art results on a specific testing scene.

Figure 22 illustrates the qualitative comparisons between the generated novel views of the generalized or scene-specific fine-tuning CG-NeRF with those produced by other methods. Although the model is trained from scratch using the DTU dataset [76], the generalized version of CG-NeRF still shows competitive results to a 15 minutes fine-tuning MVSNeRF [81] model. Optimizing for a quarter of an hour, the finetuned

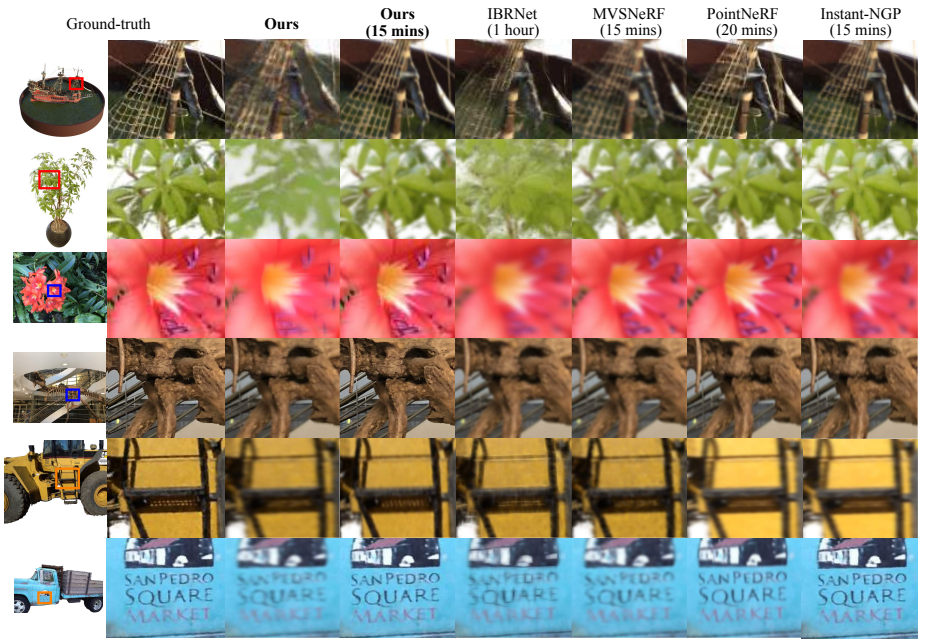

Fig. 22. Qualitative comparisons between CG-NeRF and state-of-the-art methods on three different unseen datasets. Reprinted, with permission, from Paper IV © 2022 IEEE.

CG-NeRF model produces cleaner and more photo-realistic novel views than all baseline methods on unseen synthetic and real datasets. In addition, both the RGBD-Net and the CG-NeRF are very efficient at rendering novel views thanks to the coarse-to-fine rendering strategy.

There is still a gap when employing the trained models on resource constraint devices as the current approaches first aim to have high rendering speed on the commercialized GPUs. Combining the current techniques with recent work [61] on compressing and streaming neural radiance fields would reduce the model size and the memory consumption significantly. Another interesting approach is to expand the current approaches on unbounded large-scale scenes [84] captured from drones or satellite.

5 Sphere-based dynamic human rendering

Previous chapters in this thesis have mainly focused on learning-based view synthesis on static and large-scale scenes. However, it is non-trivial to apply these methods to living people who are constantly moving daily. Capturing and faithfully rendering photorealistic humans from novel views has been a fundamental problem of 3D computer vision [85]. This chapter focuses on the problem of free-viewpoint human capture and rendering, which refers to capturing the appearance and motion of a human performer from a single viewpoint, and then synthesizing a new view of the performer from any arbitrary viewpoint. This allows for creating virtual reality (VR) or augmented reality (AR) experiences in which the viewer can freely change their perspective and see the performer from any angle as if they were physically present in the same space. Moreover, free-viewpoint human capture and rendering have many potential applications, including film and television production and virtual try-on applications for clothing and accessories.

5.1 Multi-view human performance capture and rendering

This section briefly introduces conventional free-viewpoint human rendering using multiple input views. There are several existing studies on capturing and rendering the appearance and motion of a human performer for free-viewpoint rendering. Recent human-specific neural rendering approaches [86, 87] use multiple cameras to capture the performer from different viewpoints. The captured images are then used to create a 3D model of the performer, which can be manipulated to synthesize new views from any desired viewpoint. However, such methods can be prohibitively expensive to run and cannot generalize to unseen humans but instead create a dedicated model for each human that they need to render.

There is another line of research that tries to generate dynamic humans using a generalized human view synthesis network as seen in the Figure 23 and they train the model either by single or multiple view captured data. Martin-Brualla et al. [88] use multiple depth sensors or structured lights to capture the 3D geometry of the performer directly. However, their capture setup produces dense geometry, which makes this a comparatively easy task: the target views stay consistent with the input views. A recent approach [89] uses a frontal input view and a large number of calibration images to

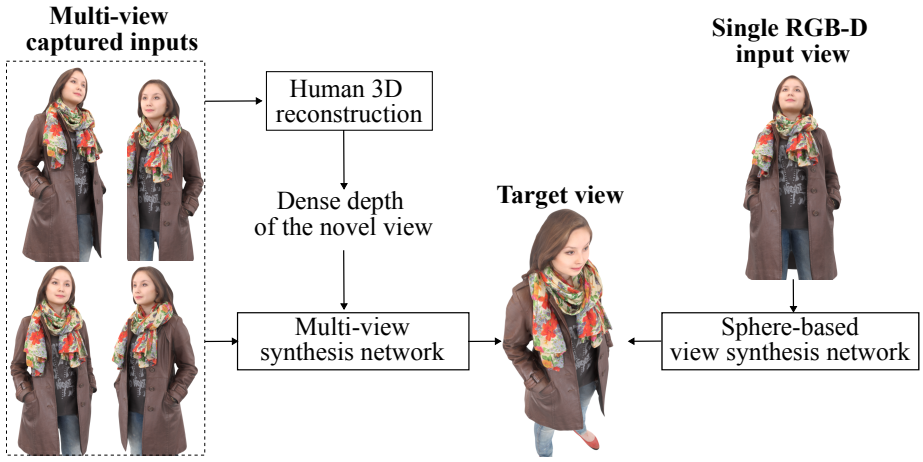

Fig. 23. Comparison of single and multi-view synthesis approaches on the human performance capture and rendering.

extrapolate novel views. This method relies on a keypoint estimator to warp the selected calibrated image to the target pose, leading to unrealistic results for hands, occluded limbs, or large body shapes. Despite having excellent results, those multi-view synthesis approaches require an expensive multi-view capture rig which is not trivial to setup. Paper V tries to tackle this problem of using a single view RGB-D image, and the depth image can be very sparse. Moreover, the input camera is fixed on the ground, looking up to the human. This setup is chosen for practical purposes but also for increasing the geometric distance between input and target poses.

5.2 Single view human performance capture and rendering

This section presents the main pipeline of the Human View Synthesis Network (HVS-Net) from Paper V that generates high-fidelity rendered images of clothed humans using just a single RGBD image from the frontal viewpoint. There are many challenges: 1) generalization to new subjects at test-time as opposed to models trained per subject, 2) the ability to handle dynamic behavior of humans in unseen poses as opposed to animating humans using the same poses seen at training, 3) the ability to handle occlusions (either from objects or self-occlusion), 4) capturing facial expressions and 5) the generation of high-fidelity images in a live setup given a single-stream, sparse RGB-D input (similar to a low-cost, off-the-shelf depth camera).

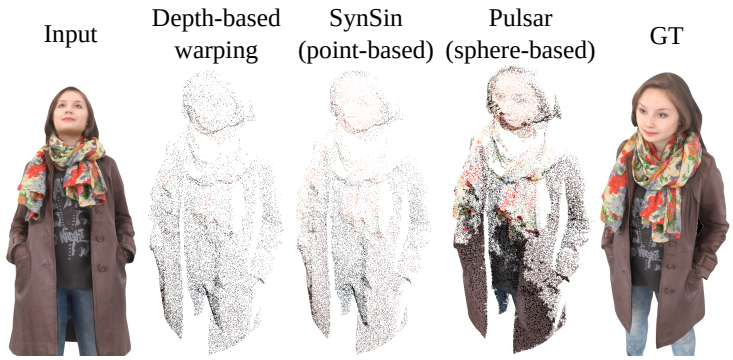

Fig. 24. Comparison of sparse 3D point cloud transformations from the input view to the novel view. The novel image warped by sphere-based rendering is significantly denser. Reprinted, with permission, from Paper V © 2022 Springer Nature.

5.2.1 Sphere-based neural rendering

Given the depth of every pixel from the original viewpoint as well as the camera parameters, these points can naturally be projected into a novel view. This makes the use of depth-based warping or a differentiable point- or sphere-renderer [31] a natural choice for the first step in the development of the view synthesis model. The better this renderer can transform the initial information into the novel view, the better this projection step is automatically correct (except for sensor noise) and not subject to training errors.

Fig. 24 shows a comparison between warped novel views from a single RGBD image using three different methods: depth-based warping [90], point-based rendering [91] and sphere-based rendering [31]. Depth-based warping [90] represents the RGD-D input as a set of pixel-sized 3D points, and thus, the correctly projected pixels in the novel view are very sensitive to the density of the input view. The widely-used differentiable point-based renderer [91] introduces a global radius-per-point parameter that produces a somewhat denser image. Since it uses the same radius for all points, this comes with a trade-off: if the selected radius is too large, details in dense regions of the input image are lost; if the chosen radius is too small, the resulting images get sparser in sparse regions. The recently introduced sphere-based renderer not only provides the option to use a per-sphere radius parameter, but also provides gradients for these radii, enabling us to set them dynamically. As depicted in Figure 24, this allows us to produce denser images than the other methods. More information about efficient differentiable rendering can be found in the Pulsar sphere-based neural rendering paper [31].

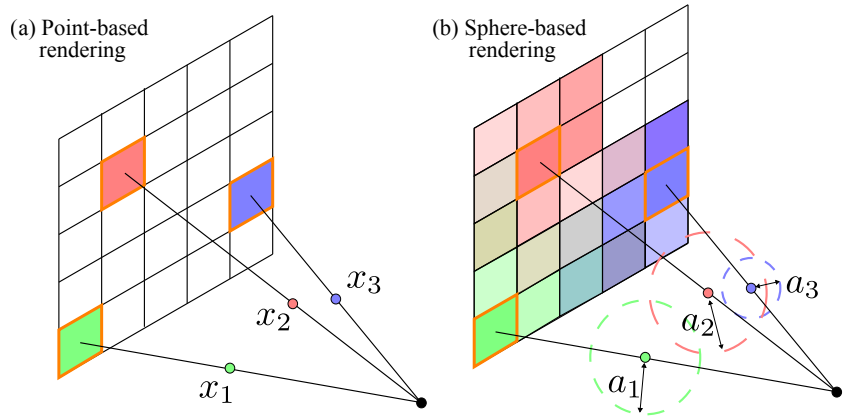

Fig. 25. Visualization of the *rendered features* between (a) point and (b) sphere-based rendering methods. Reprinted, with permission, from Paper V © 2022 Springer Nature.

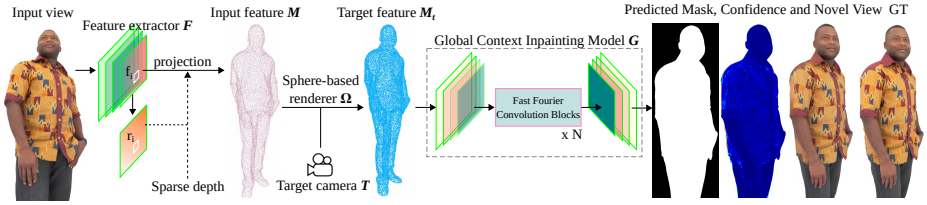

Fig. 26. Sphere-based view synthesis network architecture. Reprinted, with permission, from Paper V © 2022 Springer Nature.

5.2.2 Sphere-based view synthesis network

Instead of naively warping RGB pixels, the sphere-based renderer warps the extracted deep features of the input image to the novel view. In addition, those extracted features, along with positional encoding of the pixel position and its viewing direction, are fed to a shallow convolution layer, followed by a sigmoid activation to estimate the radius-per-sphere. In Fig. 25, we show the visualization of rendered feature maps from a set of sparse points using point and sphere-based renderers. In the case of point-based rendering [91], each 3D point p_i can render a single pixel. A large amount of pixels can-not be rendered because there is no ray connecting those pixels with valid 3D points. In contrast, the sphere-based neural renderer [31] renders a pixel by blending the colors of any intersecting spheres with the given ray. Since we estimate radius a_i of each sphere

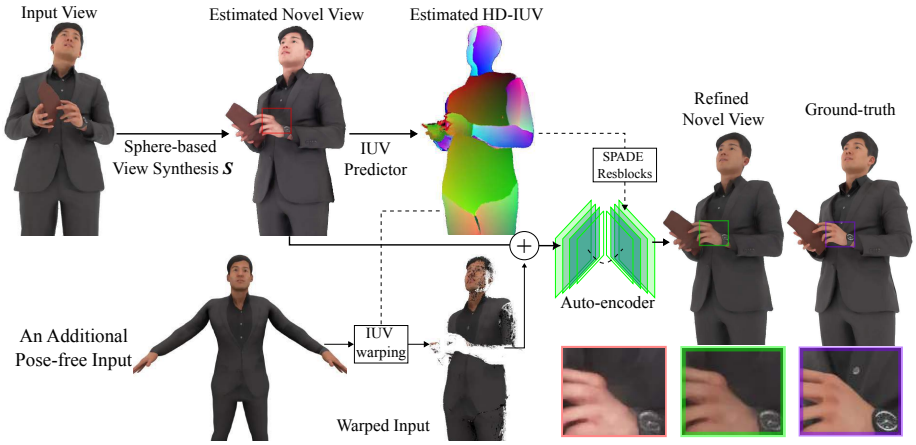

Fig. 27. An overview of the IUV-based image refinement network. Reprinted, with permission, from Paper V © 2022 Springer Nature.

(dashed circle) using a shallow network, this allows us to render pixels that do not have a valid 3D coordinates.

Although the warped features using sphere-based rendering are denser than other methods, many pixels are still left to be filled. This remains a challenging problem since having several "gaps" in the re-projected feature images cannot be avoided. As seen in the Figure 26, Paper V presents an efficient encoder-decoder-based inpainting model to produce the final renders. The encoding bottleneck significantly increases the model's receptive field size, allowing it to fill in more of the missing information correctly. Additionally, the method employs a series of Fast Fourier Convolutions (FFC) [92] to account for the image-wide receptive field. The model can hallucinate missing pixels much more accurately compared to regular convolution layers [93]. The output of the encoder-decoder structure is a foreground mask, a confidence mask and an novel view RGB images. Both masks are then applied to the predicted RGB to obtain the initial prediction. Paper V trains the above network using several photometric losses such as L1 image loss, VGG perceptual loss and adversarial loss.

5.2.3 Occlusion-aware rendering

The sphere-based view synthesis network predicts plausible novel views with high quality. As seen in Figure 27, if the person is holding an object such as a wallet or if their hands are obstructing large parts of their torso, then the warped transformation will

result in missing points in this region. This leads to low-fidelity texture estimates for those occluded regions when performing novel view synthesis with a target camera that is not close to the input view. Hence, to further enhance the quality of the novel views, two additional modules are introduced: 1) an HD-IUV predictor to predict dense correspondences between an RGB image (render of a human) and the 3D surface of a human body template, and 2) a refinement module to warp an additional occlusion-free input (a selfie in a practical application) to the target camera and enhance the initial estimated novel view to tackle the self-occlusion issue.

There are several studies [94, 95] on predicting the dense correspondences of a human from a single image. Still, the estimated IUV predictions cover only the naked body instead of the clothed human, and they are inaccurate as they are trained based on sparse and noisy human annotations. Instead, Paper V introduces an IUV predictor from scratch using synthetic RenderPeople scans. This dataset contains 3D models that can be processed to obtain accurate ground-truth correspondences. The output of the HD-IUV predictor is then utilized to warp all visible pixels to the human in the target camera and obtain a partially warped image. The partially warped image is finally fed to another refinement U-Net model to refine the initial estimated novel view. This module addresses two key details: 1) it learns to be robust to artifacts that originate either from the occluded regions of the initially synthesized novel view as well as texture artifacts that might appear because we rely on HD-IUV dense correspondences for warping and 2) it is capable of synthesizing crisper results in the occluded regions as it relies on both the initially synthesized image as well as the warped image to the target view based on HD-IUV.

5.3 Discussion

Training the HVS-Net of Paper V requires multi-view ground-truth images and their human-dense correspondences. It is non-trivial to acquire such data from real human captures. Therefore, the network is trained solely using synthetic data. Although the trained model has not seen any real images, it can still synthesize photo-realistic novel views of unseen humans, as illustrated in Figure 28.

Another impressive feat of the HVS-Net is that the method does not rely on dense input depth maps to synthesize high-quality novel views. Figure 29 shows novel view synthesis results using different levels of sparsity of the input depth maps. The method can maintain the quality of view synthesis despite substantial reductions in point

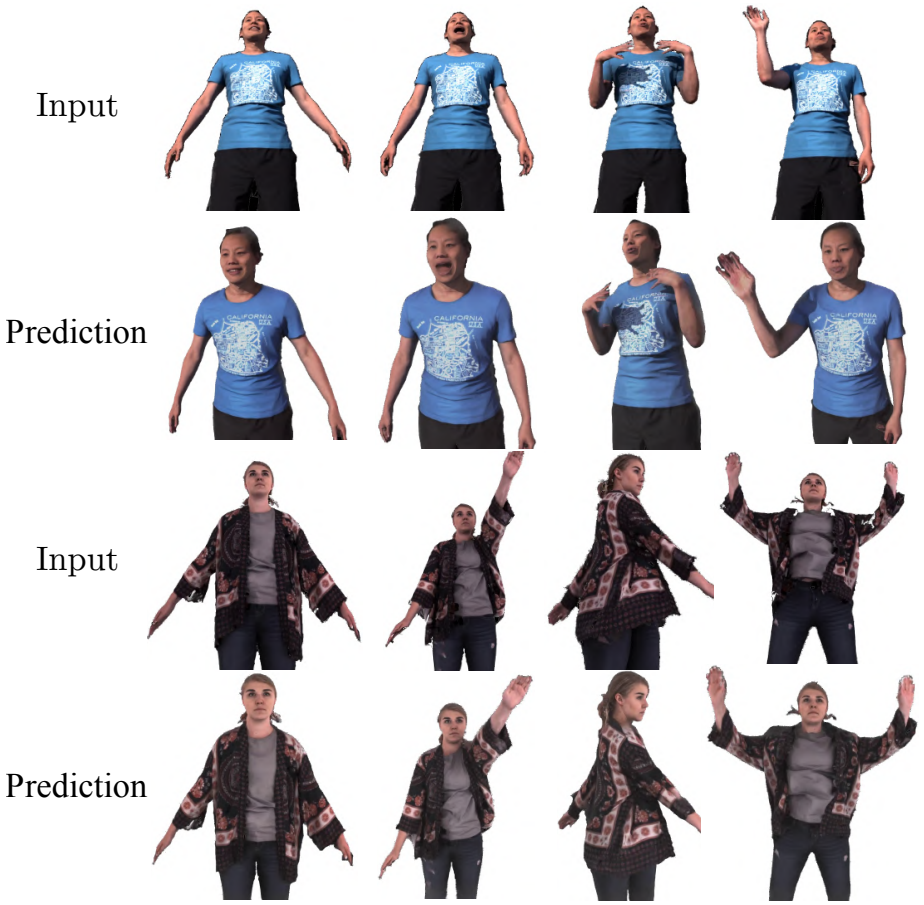

Fig. 28. The novel views generated by HVS-Net on the real-world examples without any finetuning. Adapted by permission, Paper V © 2022 Springer Nature.

cloud density. This highlights the importance of the proposed sphere-based rendering component and the occlusion-aware rendering.

For AR/VR applications, a prime target for a method like the one proposed, runtime performance, is critical. At test time, the HVS-Net generates 1K resolution images at 21FPS using a single NVIDIA V100 GPU. This speed can be further increased with more efficient data loaders and an optimized implementation that uses the NVIDIA TensorRT engine. Finally, different depth sparsity levels do not significantly affect the average runtime of the HVS-Net, which is a plus compared to prior studies [88, 89].

(a) 5% foreground points (b) 10% foreground points (c) 25% foreground points

Fig. 29. A comparison between generated novel views on different levels of depth sparsity. Reprinted, with permission, from Paper V © 2022 Springer Nature.

Although the HVS-Net is superior in generating dynamic unseen humans, the method requires a set of sparse 3D points as input to the pipeline. Applying this method on single-frontal camera smartphones is therefore more complex. Future research on this topic should address the problem of generating novel views of an unseen dynamic human using the captured smartphone data.

6 Summary and conclusion

This thesis has presented novel learning-based view synthesis approaches using different neural scene representations. This includes the simplest form of a single vector, as seen in Paper I and Paper II. Although both methods show excellent performance on unseen data and can hallucinate unseen regions, extending them to real-world data is not straightforward due to the lack of training data. This motivates the use of much more complicated scene representations, such as multiple plane images or plane sweep volumes (Paper III). However, training such a model requires depth supervision to obtain high-quality and consistent novel views. Combining such a method with neural radiance fields addresses this issue and achieves state-of-the-art results on static synthetic and real datasets (Paper IV). Chapter 5 discusses generating dynamic humans for the AR/VR application (Paper V).

High-performance novel view synthesis from a set of sparse observations using deep learning resort sto massive training and large network architectures. Paper I proposed a method that leverages the adversarial training scheme to improve the visual quality and convergence speed. Moreover, a feature-matching loss function is presented for stabilizing the training procedure. The experiments demonstrate that Paper I can produce high-quality results and faster convergence compared to the conventional approach [38].

Rendering far-away target views is a challenging problem of novel view synthesis. Paper II tackles this issue by introducing an attention-based model that leverages the long-range dependencies learning of Transformer architectures [9]. Instead of directly predicting the target view in a single pass, a multiple-step rendering strategy is proposed that sequentially renders novel views in an ordered set based on the geometric distance between input and target poses. The learned model can attend to the most critical view closest to the target at each rendering step. Moreover, it performs well on various challenging datasets and gives consistent predictions without any retraining for finetuning.

View synthesis methods struggle to render novel views on unseen data and have slow rendering speeds. Paper III tackles these issues by introducing a view synthesis network that estimates both color and depth images of the novel views in a coarse-to-fine manner. A memory-efficient multiple-plane volume is constructed to extract multi-scale depth maps, and they are then utilized to obtain near-surface features. A spatio-temporal

consistency synthesis module is then used to obtain high quality and jitterless novel views. The network can be trained end-to-end with or without depth supervision and perform well on large-scale real data.

Previous research on view synthesis often uses depth supervision for training, and requires a long training time to reach convergence. Paper IV presents a view synthesis approach that leverages the coordinate-based scene representation of radiance fields [8] that relies only on photometric loss to train. However, this method takes more than 30 seconds to render a single image. An cascade and efficient radiance fields predictor is therefore present first to infer coarse estimates of the novel views. An up-scaling neural renderer is then applied to predict the color image at the original resolution. Further optimizing the trained model on multiple scenes in 15 minutes enables state-of-the-art results on a specific testing scene.

Rendering a dynamic human has been an important problem in computer vision; a way of rendering an unseen human using just a single image has yet to be discovered. Paper V proposes a novel view synthesis framework that generates realistic renders from unseen views of any human captured from a single-view and sparse RGB-D sensor, similar to a low-cost depth camera, and without actor-specific models. The proposed architecture creates dense feature maps in novel views obtained by sphere-based neural rendering and creates complete renders using a global context inpainting model. Additionally, an enhancer network leverages the overall fidelity, even in occluded areas from the original view, producing crisp renders with fine details. Experimental results show that high-quality novel views of synthetic and real human actors are generated given a single-stream, sparse RGB-D input. The trained model also generalizes to unseen identities and new poses and faithfully reconstructs facial expressions. Moreover, this method outperforms prior view synthesis methods and is robust to different levels of depth sparsity.

References

[1] R. Szeliski and H.-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 251–258. [Online]. Available: https://doi.org/10.1145/258734.258861

[2] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, 2004, pp. II–II.

[3] S.-H. Sun, M. Huh, Y.-H. Liao, N. Zhang, and J. J. Lim, "Multi-view to novel view: Synthesizing novel views with self-learned confidence," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 162–178.

[4] I. E. Sutherland, "A head-mounted three dimensional display," in *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, ser. AFIPS '68 (Fall, part I). New York, NY, USA: Association for Computing Machinery, 1968, p. 757–764. [Online]. Available: https://doi.org/10.1145/1476589.1476686

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[7] E. Penner and L. Zhang, "Soft 3d reconstruction for view synthesis," *ACM Trans. Graph.*, vol. 36, no. 6, nov 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130855

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 405–421.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[10] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '95. New York, NY, USA: Association for Computing Machinery, 1995, p. 39–46. [Online]. Available: https://doi.org/10.1145/218380.218398

[11] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '96. New

York, NY, USA: Association for Computing Machinery, 1996, p. 31–42. [Online]. Available: https://doi.org/10.1145/237170.237199

[12] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 43–54. [Online]. Available: https://doi.org/10.1145/237170.237200

[13] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '00. USA: ACM Press/Addison-Wesley Publishing Co., 2000, p. 307–318. [Online]. Available: https://doi.org/10.1145/344779.344932

[14] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[15] M. Goesele, B. Curless, and S. Seitz, "Multi-view stereo revisited," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2402–2409.

[16] M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klowsky, D. Steedly, and R. Szeliski, "Ambient point clouds for view interpolation," *ACM Trans. Graph.*, vol. 29, no. 4, jul 2010. [Online]. Available: https://doi.org/10.1145/1778765.1778832

[17] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM Trans. Graph.*, vol. 32, no. 3, jul 2013. [Online]. Available: https://doi.org/10.1145/2487228.2487238

[18] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow, "Scalable inside-out image-based rendering," *ACM Trans. Graph.*, vol. 35, no. 6, nov 2016. [Online]. Available: https://doi.org/10.1145/2980179.2982420

[19] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," in *SIGGRAPH*, 2018.

[20] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, "Pushing the boundaries of view extrapolation with multiplane images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 175–184.

[21] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "Deepview: View synthesis with learned gradient descent," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2367–2376.

[23] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, 2019.

[24] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn, "Nex: Real-time view synthesis with neural basis expansion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin, "Matryodshka: Real-time 6dof video view synthesis using multi-sphere images," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 441–459.

[26] K.-E. Lin, Z. Xu, B. Mildenhall, P. P. Srinivasan, Y. Hold-Geoffroy, S. DiVerdi, Q. Sun, K. Sunkavalli, and R. Ramamoorthi, "Deep multi depth panoramas for view synthesis," in

Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 328–344.

[27] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky, "Neural point-based graphics," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 696–712.

[28] R. Rakhimov, A.-T. Ardelean, V. Lempitsky, and E. Burnaev, "Npbg++: Accelerating neural point-based graphics," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 948–15 958.

[29] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, "Synsin: End-to-end view synthesis from a single image," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7465–7475.

[30] D. Rückert, L. Franke, and M. Stamminger, "Adop: Approximate differentiable one-pixel point rendering," *ACM Trans. Graph.*, vol. 41, no. 4, jul 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530122

[31] C. Lassner and M. Zollhofer, "Pulsar: Efficient sphere-based neural rendering," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 1440–1449. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00149

[32] V. Sitzmann, M. Zollhofer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," in *Advances in Neural Information Processing Systems*, 2019.

[33] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 65:1–65:14, Jul. 2019.

[34] S. Laine and T. Karras, "Efficient sparse voxel octrees," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 8, pp. 1048–1059, 2011.

[35] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," *NeurIPS*, 2020.

[36] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5491–5500.

[37] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *CVPR*, 2022.

[38] S. M. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu, and D. Hassabis, "Neural scene representation and rendering," *Science*, vol. 360, no. 6394, pp. 1204–1210, 2018. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aar6170

[39] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4455–4465.

[40] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.

[41] J. Chibane, A. Mir, and G. Pons-Moll, "Neural unsigned distance fields for implicit function learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020.

[42] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *NeurIPS*, 2020.

[43] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[44] M. Levoy, "Efficient ray tracing of volume data," *ACM Trans. Graph.*, vol. 9, no. 3, p. 245–261, jul 1990. [Online]. Available: `https://doi.org/10.1145/78964.78965`

[45] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik, "Advances in neural rendering," *Computer Graphics Forum*, vol. 41, no. 2, pp. 703–735, 2022. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14507`

[46] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *Computer Graphics Forum*, 2022.

[47] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4576–4585.

[48] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "Ibrnet: Learning multi-view image-based rendering," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4688–4697.

[49] M. Mihajlovic, A. Bansal, M. Zollhoefer, S. Tang, and S. Saito, "KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints," in *European conference on computer vision*, 2022.

[50] A. Kurz, T. Neff, Z. Lv, M. Zollhöfer, and M. Steinberger, "Adanerf: Adaptive sampling for real-time rendering of neural radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022.

[51] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "NeRF−−: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[52] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5721–5731.

[53] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, and J. Yu, "Gnerf: Gan-based neural radiance field without posed camera," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6331–6341.

[54] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5732–5741.

[55] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 326–14 335.

[56] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec, "Baking neural radiance fields for real-time view synthesis," *ICCV*, 2021.

[57] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5449–5459.

[58] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[59] J. N. P. Martel, D. B. Lindell, C. Z. Lin, E. R. Chan, M. Monteiro, and G. Wetzstein, "Acorn: Adaptive coordinate networks for neural scene representation," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459785

[60] T. Müller, F. Rousselle, J. Novák, and A. Keller, "Real-time neural radiance caching for path tracing," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459812

[61] T. Takikawa, A. Evans, J. Tremblay, T. Müller, M. McGuire, A. Jacobson, and S. Fidler, "Variable bitrate neural fields," in *ACM SIGGRAPH 2022 Conference Proceedings*, ser. SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3528233.3530727

[62] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022.

[63] Y. Zhang, W. Xu, Y. Tong, and K. Zhou, "Online structure analysis for real-time indoor scene reconstruction," *ACM Trans. Graph.*, vol. 34, no. 5, pp. 159:1–159:13, Nov. 2015. [Online]. Available: http://doi.acm.org/10.1145/2768821

[64] D. Jimenez Rezende, S. M. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, "Unsupervised learning of 3d structure from images," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 4996–5004.

[65] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1462–1471. [Online]. Available: http://proceedings.mlr.press/v37/gregor15.html

[66] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=BydrOIcle

[67] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 214–223. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[68] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[69] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[70] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 358–363.

[71] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, Nov. 2016. [Online]. Available: `https://doi.org/10.1145/2980179.2980251`

[72] R. Collins, "A space-sweep approach to true multi-image matching," in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 358–363.

[73] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[74] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2495–2504.

[75] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *European Conference on Computer Vision (ECCV)*, 2018.

[76] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vision*, vol. 120, no. 2, p. 153–168, Nov. 2016. [Online]. Available: `https://doi.org/10.1007/s11263-016-0902-9`

[77] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[78] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[79] G. Riegler and V. Koltun, "Free view synthesis," in *European Conference on Computer Vision*, 2020.

[80] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv:2010.07492*, 2020.

[81] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," *arXiv preprint arXiv:2103.15595*, 2021.

[82] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations*, 2022. [Online]. Available: `https://openreview.net/forum?id=vh-0sUt8HlG`

[83] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *arXiv*, 2021.

[84] Y. Xiangli, L. Xu, X. Pan, N. Zhao, A. Rao, C. Theobalt, B. Dai, and D. Lin, "Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering," in *The European Conference on Computer Vision (ECCV)*, 2022.

[85] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, nov 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[86] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021.

[87] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose," in *Advances in Neural Information Processing Systems*, 2021.

[88] R. Martin-Brualla, R. Pandey, S. Yang, P. Pidlypenskyi, J. Taylor, J. Valentin, S. Khamis, P. Davidson, A. Tkach, P. Lincoln, A. Kowdle, C. Rhemann, D. B. Goldman, C. Keskin, S. Seitz, S. Izadi, and S. Fanello, "Lookingood: Enhancing performance capture with real-time neural re-rendering," *ACM Trans. Graph.*, vol. 37, no. 6, dec 2018. [Online]. Available: https://doi.org/10.1145/3272127.3275099

[89] R. Pandey, A. Tkach, S. Yang, P. Pidlypenskyi, J. Taylor, R. Martin-Brualla, A. Tagliasacchi, G. Papandreou, P. Davidson, C. Keskin, S. Izadi, and S. Fanello, "Volumetric capture of humans with a single rgbd camera via semi-parametric learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2019, pp. 9701–9710. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00994

[90] H.-A. Le, T. Mensink, P. Das, and T. Gevers, "Novel view synthesis from a single image via point cloud transformation," in *BMVC*, 2020.

[91] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, "Synsin: End-to-end view synthesis from a single image," in *CVPR*, 2020.

[92] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," in *NeurIPS*, 2020.

[93] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *WACV*, 2022.

[94] A. Ianina, N. Sarafianos, Y. Xu, I. Rocco, and T. Tung, "BodyMap: Learning full-body dense correspondence map," in *CVPR*, 2022.

[95] N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer," in *ECCV*, 2018.

Original publications

I Nguyen-Ha, P., Huynh, L., Rahtu, E. & Heikkilä, J. (2019, June). Predicting Novel Views Using Generative Adversarial Query Network. Scandinavian Conference on Image Analysis (SCIA). Norrköping, Sweden.
https://doi.org/10.1007/978-3-030-20205-7_2

II Nguyen-Ha, P., Huynh, L., Rahtu, E. & Heikkilä, J. (2020, November). Sequential View Synthesis with Transformer. Asian Conference in Computer Vision (ACCV). Kyoto, Japan.
https://doi.org/10.1007/978-3-030-69538-5_42

III Nguyen-Ha, P., Animesh, K., Huynh, L., Rahtu, E., Jiri, M., & Heikkilä, J. (2021, December). RGBD-Net: Predicting Color and Depth images for Novel Views Synthesis. International Conference on 3D Vision (3DV). Surrey, UK.
https://doi.org/10.1109/3DV53792.2021.00117

IV Nguyen-Ha, P., Huynh, L., Rahtu, E., Jiri, M., & Heikkilä, J. (2022, December). HRF-Net: Holistic Radiance Fields from Sparse Inputs. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). The paper is currently under a major revision.
https://arxiv.org/abs/2208.04717

V Nguyen-Ha, P., Nikolaos, S., Christoph, L., Heikkilä, J. & Tony, T. (2022, October). Free-Viewpoint RGB-D Human Performance Capture and Rendering. European Conference in Computer Vision (ECCV). Tel Aviv, Isarel.
https://doi.org/10.48550/arxiv.2112.13889

874. Laukkanen, Johanna (2023) Alkali-activation of industrial aluminosilicate sidestreams : application for mine water treatment and sediment remediation

875. Ismail, Mostafa (2023) Advanced imaging of lignocellulosic and cellulose materials

876. Abdelrahim, Ahmed (2023) Towards lower CO_2 emissions in iron and steelmaking : hydrogen-assisted reduction and cement-free briquettes

877. Zhou, Jin (2023) Conductive composites of engineered nanomaterials with excellent gas sensing properties

878. Ali, Farooq (2023) A framework for analyzing, developing, and managing stakeholder network relationships in collaborative hospital construction projects

879. Rönkkö, Pasi (2023) Circular economy as an enabler of improved resilience and material availability in supply chains

880. Isteri, Visa (2023) Alternative ye'elimite (CSA) cement clinkers from industrial byproducts

881. Aprialdi, Dwinata (2023) Forest hydrological studies to improve water management in marine lowlands in Palembang, Indonesia

882. Koskela, Aki (2023) Utilisation of lignin-based biocarbon in pyrometallurgical applications

883. Jääskä, Elina (2023) Game-based learning methods in project management higher education

884. Adediran, Adeolu Idowu (2023) Alkali activation of iron-rich fayalite slag : fresh, hardened and durability properties

885. Mansoori, Solmaz (2023) Improving data utilization in construction : on the path towards industrialization

886. Hannila, Esa (2023) Quality assessment and reliability of additively manufactured hybrid structural electronics

887. Alorwu, Andy (2023) User perceptions of personal data in healthcare : ethics, reuse, and valuation

888. Nyameke, Emmanuel (2023) Project identity formation and maintenance in a temporary organization environment

889. Koivikko, Niina (2023) Silica-titania supported vanadia catalysts in the utilization of industrial sulfur-contaminated gaseous methanol streams

890. Ye, Liang (2023) Automated multi-modal recognition of school violence

Book orders:
Virtual book store
https://verkkokauppa.omapumu.com/fi/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-3739-8 (Paperback)
ISBN 978-952-62-3740-4 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)

UNIVERSITY OF OULU