CVMP 2022

The 19th ACM SIGGRAPH European Conference on Visual Media Production
1–2 December 2022, London

Programme

ODEON Covent Garden, London
1st & 2nd December 2022
https://www.cvmp-conference.org/2022/

Conference Sponsors 2022

ACM ISBN: 978-1-4503-9939-5

Message from the Chairs

We are pleased to introduce the programme for the nineteenth ACM SIGGRAPH European Conference on Visual Media Production (CVMP). For almost two decades, CVMP has built a reputation as the prime venue for researchers to meet with practitioners in the Creative Industries: film, broadcast and games. The conference brings together expertise in computer vision, computer graphics, video processing, machine learning, games, XR, animation and physical simulation. It provides a forum for presentation of the latest research and application advances, combined with keynotes and invited talks on state-of-the-art industry practice. CVMP regularly attracts attendees approximately 50:50 from academia and the creative industries.

CVMP has a traditionally strong technical papers programme but this year has seen an increase in the number of submitted papers and we are delighted to present ten full papers and fourteen short papers, from both academia and industry. Full papers were subject to double-blind peer review by our international programme committee, and short papers by jury from our paper and programme chairs. Special care was taken to ensure peer-review was handled by non-conflicted reviewers. This makes for what we believe is a great papers line-up for oral and poster presentations at CVMP, and is a strong indicator of the quality of research in our area. We are also continuing with spotlight presentations for short papers which proved to be very popular in previous years.

Finally, we would like to thank everyone who submitted to CVMP this year, the invited speakers, the reviewers, our sponsors, and the organising committee for their hard work in bringing CVMP 2022 together!

Rafał Mantiuk and Marco Volino (Conference Chairs)
Armin Mustafa (Full Papers Chair)
Yulia Gryaditskaya (Short Papers Chair)
Valentin Deschaintre (Industry Chairs)
Jeff Clifford (Sponsorship Chair)
Giuseppe Claudio Guarnera (Local Arrangements)
Peter Vangorp (Public Relations)
Emily Ellis (Conference Secretary)

DAY 1 | Thursday 1st December 2022
Location: ODEON Covent Garden

09:00 **Registration opens |** Foyer

09:30 **Chairs' Welcome |** **Marco Volino**, University of Surrey
Rafał Mantiuk, University of Cambridge

09:40 SESSION 1 **| A Lovely View**

1. **Neural apparent BRDF fields for multiview photometric stereo**
   Meghna Asthana, William A. P. Smith, Patrik Huber (University of York)
2. **A wide-baseline multiview system for indoor scene capture**
   Théo Barrios (University of Reims Champagne-Ardenne), Cédric Niquin (XD Productions), Stéphanie Prévost (University of Reims Champagne-Ardenne), Philippe Souchet (XD Productions), Céline Loscos (University of Reims Champagne-Ardenne)
3. **Light field GAN-based view synthesis using full 4D information**
   Abrar Wafa, Panos Nasiopoulos (University of British Columbia)
4. **From fibers to pixels: SEDDI's approach for scalable textile digitization** (industry talk)
   Elena Garces (SEDDI)

11:00 **Coffee Break |** Foyer
Poster presenters put up posters

11:30 KEYNOTE 1 **| Erik Reinhard**, InterDigital

12:30 SPOTLIGHT SESSION

12:50 POSTERS, DEMO AND LUNCH **|** Foyer

14:30 SESSION 2 **| The Main and th' Aerial Blue**

5. **Generating real-time detailed ground visualisations from sparse aerial point clouds** (industry talk)
   Aiden Murray, Scarlet Mitchell, Alexander Bradley (Cobra Simulation), Eddie Waite, Caleb Ross, Joanna Jamrozy, Kenny Mitchell (Edinburgh Napier University, Cobra Simulation)
6. **Semantic segmentation for multi-contour estimation in maritime scenes**
   Alastair J. Finlinson, Sotiris Moschoyiannis (University of Surrey)
7. **Tragic talkers: a shakespearean sound-and light-field dataset for audio-visual machine learning research**
   Davide Berghi, Marco Volino, Philip J. B. Jackson (University of Surrey)

15:30 **Coffee Break |** Foyer

16:00 KEYNOTE 2 **| Ben Cowell-Thomas**, DNEG

17:00 NETWORKING RECEPTION **|** Foyer

19:00 **Close**

DAY 2 | Friday, 2nd December 2022
Location: ODEON Covent Garden

09:00 **Registration opens |** Foyer

09:30 SPECIAL SESSION **| Delivering Stories through Light and Color**

8. **The colour of horror**
   Lesley Istead (Carleton University), Andreea Pocol, Sherman Siu, William Chen, Alex Zdanowicz, Alex Rowaan, Craig Kaplan (University of Waterloo)
9. **Model-based deep portrait relighting**
   Frederik David Schreiber, Anna Hilsmann, Peter Eisert (Fraunhofer Heinrich Hertz Institute)
10. **LED Virtual production - what the tutorials don't tell you!** (industry talk)
    Peter Kirkup (Disguise)

10:50 **Coffee Break |** Foyer

11:20 KEYNOTE 3 **| Bernd Bickel**, IST Austria

12:20 POSTERS, DEMO AND LUNCH **|** Foyer

14:00 SESSION 4 **| Touch-up Touch Down**

11. **Assessing advances in real noise image denoiser**
    Clément Bled, François Pitié (Trinity College Dublin)
12. **U-attention to textures: hierarchical hourglass vision transformer for universal texture synthesis**
    Shouchang Guo (University of Michigan), Valentin Deschaintre (Adobe Research), Douglas C. Noll (University of Michigan), Arthur Roullier (Adobe)
13. **Distilling style from image pairs for global forward and inverse tone mapping**
    Aamir Mustafa, Param Hanji, Rafał Mantiuk (University of Cambridge)
14. **Capturing, inferring and applying body contact for human-scene interaction understanding** (industry talk)
    Chun-Hao Paul Huang (Adobe)

15:20 **Coffee Break |** Foyer

15:50 KEYNOTE 4 **| Belen Masia**, Universidad de Zaragoza

16:50 **Prizes, Announcements and Closing |** **Marco Volino**, University of Surrey
Rafał Mantiuk, University of Cambridge

17:00 **Close**

KEYNOTE 1 | Erik Reinhard
InterDigital

Lights, Camera, Climate Action!
Thursday 1st December 2022, 11:30

From production to transmission and final consumption in cinemas and at home: movies, series and other (live) programmes all cost non-trivial amounts of energy to produce, process, transmit and display. Advances in framerate, colour gamut, dynamic range and pixel resolution all have a further energetic impact, as does the introduction of new media, and the move from traditional broadcast to streaming technologies. The aim of this presentation is to create awareness of the magnitude of the problem, to show what efforts are underway (production, certification, standardization, image processing, displays), and where opportunities lie to do better.

Erik Reinhard

Erik Reinhard is Distinguished Scientist at InterDigital, a research, innovation and licensing company in wireless and video communication. He participates in sustainability programs at ITU-R, DVB, and SMPTE. He currently uses his expertise in color science, image/video processing and psychophysics to reduce the environmental impact of video communication and display.

KEYNOTE 2 | Ben Cowell-Thomas
DNEG

A New Era for the VFX Industry? – How DNEG Utilised Gaming Technology for The Matrix Resurrections

Thursday 1st December 2022, 16:00

In this session Ben Cowell-Thomas, DFX Supervisor at DNEG, will be breaking down The Matrix Resurrections' dojo fight sequence. He'll specifically focus on how DNEG utilised Unreal Engine for rendering, why this convergence of gaming and VFX technology is so exciting, and the benefits it offers to filmmakers. He'll also look forward to the near future to discuss the technology innovations that will positively impact the VFX and film industry, and why they should be embraced.

Ben Cowell-Thomas

Ben Cowell-Thomas is a DFX Supervisor at DNEG with over 20 years of VFX industry experience. After graduating from Bournemouth University, he started his career in education heading up the Animation and VFX course at Temasek Polytechnic Singapore. He spent three years in Sydney, before returning to the UK and working as Head of Studio / VFX Supervisor at Nexus Studios in London. Switching his focus from TV to feature film, Ben joined DNEG in 2016. His impressive filmography includes 'Star Trek Beyond', 'Justice League', 'Deadpool 2', 'Avengers: Endgame', 'Fast & Furious 9', and 'The Matrix Resurrections'.

KEYNOTE 3 | Bernd Bickel
IST Austria

Reimagining Design and Fabrication with Computational and Data-Driven Methods

Friday 2nd December 2022, 11:20

Advanced fabrication techniques have grown in sophistication over the last decade, vastly extending the scope of structures and materials that can be manufactured. While providing new opportunities for personalized fabrication, product design, engineering, architecture, art, and science, the potential impact of these techniques is tightly coupled with the availability of efficient computational methods for design.

In this talk, I will describe how techniques from visual computing enable the transformation of design workflows. I will first introduce a generic optimization approach based on the extended finite element method, which allows for the optimization of a wide range of design objectives directly on parameterized 3D CAD models. Leveraging optimization-based design and a tailored data-driven model of the materials' responses, I will then introduce novel approaches for interactive shape exploration and demonstrate its applicability to designing cold-bent glass facades and deployable structures. Furthermore, I will show how insights from geometry can be used to derive an intuitive and rigorous characterization of the design space of planar elastic rods, which can be shaped into intriguing curved elements.

Finally, I will reflect on the successes and the challenges of algorithms and artificial intelligence as tools for the future of design and discuss research opportunities in this area.

Bernd Bickel

Bernd Bickel is a Professor at IST Austria, heading the Computer Graphics and Digital Fabrication group, and a senior research scientist at Google. He is a computer scientist interested in computer graphics and its overlap into robotics, computer vision, machine learning, material science, and digital fabrication. His main objective is to push the boundaries of how digital content can be efficiently created, simulated, and reproduced.
Bernd obtained a PhD in Computer Science from ETH Zurich in 2010. From 2011-2012, Bernd was a visiting professor at TU Berlin, and in 2012 he became a research scientist and research group leader at Disney Research. In early 2015 he joined IST Austria. He received the ETH Medal for Outstanding Doctoral Thesis in 2011, the Eurographics Best PhD Award in 2012, the Microsoft Visual Computing Award in 2015, an ERC Starting Grant in 2016, the ACM SIGGRAPH Significant New Researcher Award in 2017, and a technical achievement award from the Academy of Motion Picture Arts and Sciences in 2019.

KEYNOTE 4 | Belen Masia
Universidad de Zaragoza

Modeling Attention and Gaze Behavior in Immersive Environments
Friday 2nd December 2022, 15:50

Creating engaging and compelling experiences in Virtual Reality is a challenging task: large bandwidth, computation and memory requirements are limiting factors; on top of that, there is the added difficulty of designing content for users who have control over the point of view. We argue that understanding user behavior and attention in immersive environments can help address these challenges. In this talk, we explore approaches to modeling visual attention and gaze behavior in 360º environments. Applications range from compression to realistic avatar simulation or scene content design, as well as furthering our understanding of human perception, and in particular how we selectively process the sensory information we receive.

Belen Masia

Belen Masia is an Associate Professor in the Computer Science Department at Universidad de Zaragoza, and a member of the Graphics and Imaging Lab. Before, she was a postdoctoral researcher at the Max Planck Institute for Informatics. Her research focuses on the areas of appearance modeling, applied perception and virtual reality. She is a Eurographics Junior Fellow, and the recipient of a Eurographics Young Researcher Award in 2017, a Eurographics PhD Award in 2015, an award to the top ten innovators below 35 in Spain from MIT Technology Review in 2014, and an NVIDIA Graduate Fellowship in 2012. She is also a co-founder of DIVE Medical, a startup devoted to enabling an automatic, fast, and accurate exploration of the visual function, even in non-verbal patients.

INDUSTRY TALKS

From fibers to pixels: SEDDI's approach for scalable textile digitization

Elena Garces (SEDDI)

The fashion industry is facing an unprecedented challenge to digitalize its processes, starting with textiles. However, acquiring these digital copies is typically a cumbersome and slow process that requires expensive machines and several manual steps, creating roadblocks for scalability, repeatability, and consistency. In this context, casual capture systems for optical digitization provide a promising path for scalability. These systems leverage low-cost devices (such as smartphones), one or more different illuminations, and learning-based priors to estimate the material's diffuse and specular reflection lobes. However, to train a machine learning-based solution, data is needed, which is not easy to obtain for textile materials. In this talk, I will discuss our end-to-end approach to providing a scalable solution to digitize textiles, including the design of a dual-scale optical gonioreflectometer (capable of seeing yarns at the fiber level) for dataset creation and our deep learning-based solution that only requires a single image as input.

Elena is a Technology Manager and Co-Founder at SEDDI, where she leads the AI and Optics team. She is also a Juan de la Cierva Fellow researcher at Universidad Rey Juan Carlos (Madrid, Spain) since 2022. In SEDDI, she has led the research and development of a textile digitization solution and contributed to several projects towards accelerating garments and avatars simulation using deep learning techniques. Previously, she was a postdoctoral researcher at Technicolor R&D (France) and a PhD student at the University of Zaragoza (Spain), where she specialized in inverse problems of appearance capture and style understanding.

LED virtual production - what the tutorials don't tell you!

Peter Kirkup (Disguise)

Delivers the film industry soars in growth, a progressing new technology space is accelerating the use of virtual production. Join disguise Solutions Director Peter Kirkup, as he talks through how you can take in-camera visual efforts to the next level. You'll learn all about what should be considered when exploring LED Production, all the components needed for a stage, and challenges you can prepare for.

Peter heads up Global Technical Solutions, working closely with our users to understand their projects and building solutions tailored to their specific needs. Peter has been fascinated with the world of entertainment technology from the age of 7, when he first operated lighting for a school concert. He since went on to study Stage Management at university and has held roles at Zero 88 lighting and LumenRadio in Sweden, before joining disguise in 2015.

Capturing, inferring and applying body contact for human-scene interaction understanding

Chun-Hao Paul Huang *(Adobe)*

Human motion capture has long been studied in computer vision and computer graphics. On one hand, mature maker-based solutions exist for industrial applications, such as Vicon; on the other hand, marker-less approaches also show great potentials, enabling assorted VR/AR applications. Despite rapid progress, reconstructing human motions and 3D scenes, from one or multiple RGB images, are still commonly treated as separate tasks. Motion capture is usually performed in a controlled environment, whereas scene reconstruction often considers only the static content, excluding humans. To capture both humans and 3D scenes as well as the interaction between them, one key component is the "contact" between bodies and scene objects. In this talk, I will present our research on how body-contact can be captured and estimated from RGB images, and how we apply body-contact to reconstruct more complex scenes which humans interact with.

Chun-Hao Paul Huang is a research scientist at Adobe London Lab. Before joining Adobe, he was a postdoctoral researcher at Max Planck Institute for Intelligent Systems, Tübingen, working closely with Dr. Michael Black (2019-2022). He obtained his Ph.D. with summa cum laude at Technische Universität München (TUM) under the supervision of PD. Dr. Slobodan Ilic (2016). He has interned at Microsoft Research Redmond, Disney Research Zurich and collaborate closely with Dr. Edmond Boyer at INRIA Grenoble. Chun-Hao's primary research focus is on human-related visual perception, such as 3D body reconstruction, human-scene interaction, marker-less motion capture, and 3D-guided human image generation. His work has been shortlisted in best-paper candidates (CVPR 2021, CVPR 2022) and awarded as best-paper runner up (3DV 2013).

Generating real-time detailed ground visualisations from sparse aerial point clouds

Aiden Murray, Scarlet Mitchell, Alexander Bradley (Cobra Simulation), Eddie Waite, Caleb Ross, Joanna Jamrozy, Kenny Mitchell (Edinburgh Napier University, Cobra Simulation)

We present an informed kind of atomic rendering primitive, which forms a local adjacency aware classified particle basis decoupled from texture and topology. Suited to visual synthesis of detailed landscapes inferred from sparse unevenly distributed point clouds. They enable real-time, flexible, interpretive fragment shader rendering spanning their contained bounds. Results are applicable to digital twins for simulation and training, and entertainment sectors.

FULL PAPERS | Abstracts

Neural apparent BRDF fields for multiview photometric stereo

Meghna Asthana, William A. P. Smith, Patrik Huber (University of York)

We propose to tackle the multiview photometric stereo problem using an extension of Neural Radiance Fields (NeRFs), conditioned on light source direction. The geometric part of our neural representation predicts surface normal direction, allowing us to reason about local surface reflectance. The appearance part of our neural representation is decomposed into a neural bidirectional reflectance function (BRDF), learnt as part of the fitting process, and a shadow prediction network (conditioned on light source direction) allowing us to model the apparent BRDF. This balance of learnt components with inductive biases based on physical image formation models allows us to extrapolate far from the light source and viewer directions observed during training. We demonstrate our approach on a multiview photometric stereo benchmark and show that competitive performance can be obtained with the neural density representation of a NeRF.

A wide-baseline multiview system for indoor scene capture

Théo Barrios (University of Reims Champagne-Ardenne), Cédric Niquin (XD Productions), Stéphanie Prévost (University of Reims Champagne-Ardenne),
Philippe Souchet (XD Productions), Céline Loscos (University of Reims Champagne-Ardenne)

We present a complete multiview acquisition system, a camera array allowing depth reconstruction based on disparity. We built a new wide-baseline camera grid supported by an interactive camera controller purposely built for indoor large scene capture. It is composed of 16 cameras aligned as a 4x4 grid, synchronized, and characterized. The design of the camera system manages storage and real-time capture viewing.
We also propose a DNN-based approach to estimate the floating-point disparity values, which is adapted to wide-baseline configurations while providing high precision, even for sharp and concave objects. The ultimate result is a dense 3D point cloud which offers versatile possibilities of viewing.

Light Field GAN-based View Synthesis using full 4D information

Abrar Wafa, Panos Nasiopoulos (University of British Columbia)

Light Field (LF) technology offers a truly immersive experience having the potential to revolutionize entertainment, training, education, virtual and augmented reality, gaming, autonomous driving, and digital health. However, one of the main issues when working with LF is the amount of data needed to create a mesmerizing experience with realistic disparity, smooth motion parallax between views. In this paper, we introduce a learning based LF angular super-resolution approach for efficient view synthesis of novel virtual images. This is achieved by taking four corner views and then generating up to five in-between views. Our generative adversarial network approach uses LF spatial and angular information to ensure smooth disparity between the generated and original views. We consider plenoptic, synthetic LF content and camera array implementations which support different baseline settings. Experimental results show that our proposed method outperforms state-of-the-art light field view synthesis techniques, offering novel generated views with high visual quality.

Semantic Segmentation for Multi-Contour Estimation in Maritime Scenes

Alastair J. Finlinson, Sotiris Moschoyiannis (University of Surrey)

In the maritime environment, navigation and localisation are primarily driven by systems such as GPS. However, in a scenario where GPS is not available, e.g., it is jammed or the satellite connection was lost, navigators can use visual methods derived from surrounding land masses and other permanent features in the perceptual range. To enable autonomous navigation, specifically localisation, a vessel must determine its position by extracting and matching the contours of its surrounding environment with an elevation model. The contours of interest are the true horizon line, visible horizon line and shoreline. Extracting these contours is commonly approached using computational methods such as edge detection or pixel clustering techniques that are not robust and build on weak priors. To this end, we propose the first learning-based framework that explores the fusion of inertial data into an encoder-decoder model and extracts the contours. In addition, extensive data augmentation methods are used to extend the MaSTr1325 dataset, introducing further robustness to the common environmental challenges faced by the sensors of unmanned surface vessels. We form a small curated dataset containing 300 images - composed of six component segmentation masks and three further masks describing the true horizon and visible horizon contours and the shoreline, for evaluation. We experimented extensively with popular segmentation models such as UNet, SegNet, DeepLabV3+ and TransUNet with various backbones for a quantitative comparison. The results show that, within a small margin of ten pixels and in a high-resolution image, our system detects three key contours, namely shorelines, true and visible horizon contours, used in navigation with an accuracy of 63.79%, 68.94% and 89.75%, respectively.

Tragic Talkers: A Shakespearean Sound- and Light-Field Dataset for Audio-Visual Machine Learning Research

Davide Berghi, Marco Volino, Philip J. B. Jackson (University of Surrey)

3D audio-visual production aims to deliver immersive and interactive experiences to the consumer. Yet, faithfully reproducing real-world 3D scenes remains a challenging task. This is partly due to the lack of available datasets enabling audio-visual research in this direction. In most of the existing multi-view datasets, the accompanying audio is neglected. Similarly, datasets for spatial audio research primarily offer unimodal content, and when visual data is included, the quality is far from meeting the standard production needs. We present "Tragic Talkers", an audio-visual dataset consisting of excerpts from the "Romeo and Juliet" drama captured with microphone arrays and multiple co-located cameras for light-field video. Tragic Talkers provides ideal content for object-based media (OBM) production. It is designed to cover various conventional talking scenarios, such as monologues, two-people conversations, and interactions with considerable movement and occlusion, yielding 30 sequences captured from a total of 22 different points of view and two 16-element microphone arrays. Additionally, we provide voice activity labels, 2D face bounding boxes for each camera view, 2D pose detection keypoints, 3D tracking data of the mouth of the actors, and dialogue transcriptions. We believe the community will benefit from this dataset as it can assist multidisciplinary research. Possible uses of the dataset are discussed.

The Colour of Horror

Lesley Istead (Carleton University), Andreea Pocol, Sherman Siu, William Chen, Alex Zdanowicz, Alex Rowaan, Craig Kaplan (University of Waterloo)

In this paper, we present a simple method to produce a colour palette for film trailers. Our method uses k-means clustering with a saturation-based weighting to extract the dominant colours from the frames of the trailer. We use our method to generate the palettes of 29 thousand film trailers from 1960 to 2019. We aggregate these palettes by era, genre, and director by re-applying our clustering method and we note various trends in the use of colour over time and between genres. We also show that our generated palettes reflect changes in mood and theme across films in a series, and we demonstrate the palettes of notable directors.

Model-Based Deep Portrait Relighting

Frederik David Schreiber, Anna Hilsmann, Peter Eisert (Fraunhofer Heinrich Hertz Institute)

Like most computer vision problems the relighting of portrait face images is more and more being entirely formulated as a deep learning problem. However, data-driven approaches need a detailed and exhaustive database to work on and the creation of ground truth data is tedious and oftentimes technically complex. At the same time, networks get bigger and deeper. Knowledge about the problem statement, scene structure, and physical laws are often neglected. In this paper, we propose to encompass prior knowledge for relighting directly in the network learning process, adding model-based building blocks to the training. Thereby, we improve the learning speed and effectiveness of the network, thus performing better even with a restricted dataset. We demonstrate through an ablation study that the proposed model-based building blocks improve the network's training and enhance the generated images compared with the naive approach.

Assessing Advances in Real Noise Image Denoiser

Clément Bled, François Pitié (Trinity College Dublin)

Recently image denoiser networks have made a number of advances to go beyond additive Gaussian white noise and deal with real noise, such as produced by digital cameras. We note that some of the performance gains reported in the state of the art could potentially be explained by an increase in network sizes. In this paper we propose to revisit some of these advances, including the synthetic noise generator and noise maps proposed in CBDNet, and re-assess them using a simple DnCNN baseline network, and thus attempt at measuring how much gains can be attributed to using more modern architectures.

U-Attention to Textures: Hierarchical Hourglass Vision Transformer for Universal Texture Synthesis

Shouchang Guo (University of Michigan), Valentin Deschaintre (Adobe Research), Douglas C. Noll (University of Michigan), Arthur Roullier (Adobe)

We present a novel U-Attention vision Transformer for universal texture synthesis. We exploit the natural long-range dependencies enabled by the attention mechanism to allow our approach to synthesize diverse textures while preserving their structures in a single inference. We propose a hierarchical hourglass backbone that attends to the global structure and performs patch mapping at varying scales in a coarse-to-fine-to-coarse stream. Completed by skip connection and convolution designs that propagate and fuse information at different scales, our hierarchical U-Attention architecture unifies attention to features from macro structures to micro details, and progressively refines synthesis results at successive stages. Our method achieves stronger 2 times synthesis than previous work on both stochastic and structured textures while generalizing to unseen textures without fine-tuning. Ablation studies demonstrate the effectiveness of each component of our architecture.

Distilling Style from Image Pairs for Global Forward and Inverse Tone Mapping

Aamir Mustafa, Param Hanji, Rafał Mantiuk (University of Cambridge)

Many image enhancement or editing operations, such as forward and inverse tone mapping or color grading, do not have a unique solution, but instead a range of solutions, each representing a different style. Despite this, existing learning-based methods attempt to learn a unique mapping, disregarding this style. In this work, we show that information about the style can be distilled from collections of image pairs and encoded into a 2- or 3-dimensional vector. This gives us not only an efficient representation but also an interpretable latent space for editing the image style. We represent the global color mapping between a pair of images as a custom normalizing flow, conditioned on a polynomial basis of the pixel color. We show that such a network is more effective than PCA or VAE at encoding image style in low-dimensional space and lets us obtain an accuracy close to 40 dB, which is about 7-10 dB improvement over the state-of-the-art methods.

Full papers available from the ACM Digital Library
https://dl.acm.org/doi/proceedings/10.1145/3565516

SHORT PAPERS

Rendering for Hyper-Realistic Displays: A Benchmark Study

Akshay Jindal
https://www.cl.cam.ac.uk/~aj577
Rafał Mantiuk
https://www.cl.cam.ac.uk/~rkm38
Department of Computer Science and Technology,
University of Cambridge

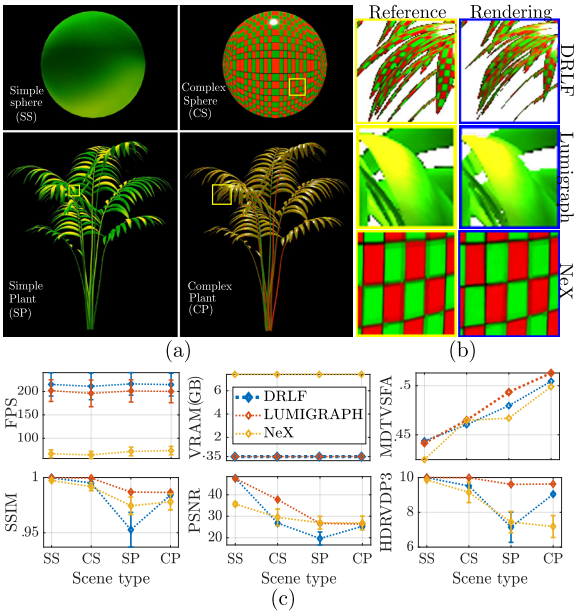


Figure 1: 3D scenes used for comparing different IBR methods (a), sample artefacts caused by three rendering methods (b), and performance and qualitative comparison of the three rendering methods (c).

Introduction

Hyper-realistic displays [5] are an emerging class of computational displays that strive to reproduce reality by delivering a large subset of perceptual realism cues such as high spatio-temporal resolution, high dynamic range (HDR), binocular disparity, motion parallax, accommodation, and colour. The rendering methods for these displays are required to process high-resolution HDR images in real-time to generate novel views from arbitrary viewpoints while delivering correct depth cues. These requirements make image-based rendering (IBR) methods a natural choice for such displays due to their relatively low computational cost and high-quality results. Each method comes with its own set of trade-offs; thus, it is necessary to understand how their artefacts translate to perceived quality. In this work, we compare three real-time techniques from different ends of the IBR spectrum: dynamically reparameterised lightfield (DRLF) [2], lumigraph implemented as a mesh with view-dependent textures [5], and NeX, a neural multi-plane images method [4]. We present performance benchmarks for the three algorithms, rate their visual artefacts using objective quality metrics, and show how the existing quality metrics are insufficient to evaluate the quality of IBR techniques.

Method

We compare the three methods on 4 forward-facing synthetic scenes (2 geometries $\times$ 2 materials) generated using Unity3D (Figure 1a). The geometries used were a sphere mesh (2 800 vertices and 960 triangles) and a plant mesh (254 244 vertices and 84 748 triangles). The meshes were mapped to two materials: a Lambertian material with a low-frequency gradient image as its diffuse component and a specular material (Phong shading) with a high-frequency checkerboard as its diffuse component.

We rendered 20 images of 2160$\times$1440 resolution and 16-bit per channel with a baseline of 100 mm for all 4 scenes. Even numbered images were fed to the rendering methods for novel view synthesis, and the odd images were used for validation. The three methods were implemented in MATLAB and OpenGL and were used to generate new views from the perspective of validation images to facilitate objective quality evaluation. The computational performance of each method was measured over 1000 frames on an i7-8700 CPU @ 3.20GHz, 32GB RAM and an RTX 2080Ti GPU with V-Sync disabled. In addition to the validation poses, 60 Hz videos were rendered and are available on the project web page[1].

Results

The DRLF method is computationally the simplest method with the highest average FPS and lowest VRAM requirements (Figure 1c). The lumigraph method is a close second in terms of performance and memory requirements. The memory required by its mesh representation is negligible compared to HDR images. There is also no noticeable effect of the number of triangles between sphere and plant scene, indicating that mesh size is not a bottleneck on modern GPUs. NeX is the slowest and most-memory intensive method (20$\times$ more memory). We found one of the biggest bottlenecks in NeX's performance to be the texture-lookup operation on MPI LUTs. An implementation that makes the better use of the GPU memory architecture might improve performance.

The artefacts induced by the three methods are also quite different (Figure 1b). DRLF works well on the simple sphere but blurs the regions away from the focal plane in other scenes. This is visible as blurred textures or missing thin edges. Lumigraph renders sharp textures in all scenes but maps incorrect textures near thin edges (disocclusions). Since lumigraph uses the rasterization of a mesh, the geometry edges can also suffer from aliasing artefacts. In both methods, these artefacts become more prominent in videos and appear as distracting flickering near thin edges or juddery specular highlight. NeX works well on both thin edges and specular highlights but gravely affects the texture appearance of checkerboard material. Also, NeX induces unnatural halos around the object when the rendering viewpoint is far from MPI's reference pose.

To quantify the quality of each method, we compared their renderings for the validation camera views against the validation images using 3 popular objective quality metrics: PSNR, SSIM, and HDR-VDP-3. Since PSNR and SSIM were designed for SDR images, the HDR images were first encoded using PU21 transform [1]. HDR-VDP-3 assumed a linear RGB colour space and 45 ppd resolution. The plots in (Figure 1c) show the quality results averaged over 10 validation images, and the error bars show the standard deviation. We also run a blind video quality assessment metric MDTVSFA [3] on the videos recorded for each method. Overall, the lumigraph rendering method is consistently rated as the highest quality in all 4 scenes. However, there is not much consensus on the rating and ranking of the methods across 4 metrics. Lumigraph and DRLF have similar quality (low std. deviation) across all 10 validation images (except DRLF for SP), but NeX has a higher variance due to artefacts in images far from the reference pose. Since the metrics we used are some of the most popularly used metrics in IBR evaluation, it is crucial to understand which of these metrics best relate to human ratings. Note that none of these metrics were designed for IBR artefacts, and most of them do not account for HDR or temporal artefacts such as judder and flicker.

[1] M. Azimi and R. K. Mantiuk. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*. IEEE, 2021.

[2] A. Isaksen, L. McMillan, and S. J. Gortler. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[3] D. Li, T. Jiang, and M. Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 2021.

[4] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proc. of Computer Vision and Pattern Recognition*, 2021.

[5] F. Zhong, A. Jindal, A.O. Yöntem, P. Hanji, S. J. Watt, and R. K. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Trans. Graph.*, 2021.

[1] https://www.cl.cam.ac.uk/research/rainbow/projects/hdrmfs/rendering_methods/

Super-resolution 3D human digitization from a single low-resolution image

Marco Pesavento, Marco Volino, Adrian Hilton
{m.pesavento,m.volino,a.hilton}@surrey.ac.uk

Centre of Vision, Speech and Signal Processing (CVSSP),
University of Surrey

1 Introduction

Although consumer cameras are nowadays able to capture high-resolution (HR) images, there are several scenarios where the image of a single person is not at the full camera image resolution but at a relatively low-resolution (LR) sub-image. Reconstruction of high quality shape from LR images of people is therefore important for example for images of multiple people, scenes requiring a large capture volume or when people are distant from the camera (Fig. 3). Since the LR image contains little detail information, state-of-the-art approaches cannot represent fine details in the reconstructed shape. To tackle this problem, we introduce a new framework that generates Super-Resolution Shape (SuRS) via a high-detail implicit function which learns the mapping from a low-resolution shape to its high resolution counterpart. We apply SuRS to the task of 3D human digitization to estimate high-detail 3D human shape from a single LR RGB image (256×256) without assisting the training with auxiliary data such as normal maps or parametric models. Our approach learns the missing information of S_{LR} in order to estimate fine shape details from a LR image even if these details are not clearly visible in the input image. An extended version of this short paper will appear in ECCV22 [2].

2 Method

An implicit function defines a surface as a level set of a function f, e.g. $f(X) = 0$ where X is a set of 3D points in $\Re^3$. To represent a 3D surface S, this function $f(X)$ is modelled with a Multi-Layer Perceptron (MLP) that classifies the 3D points as either 'inside' or 'outside' the 3D surface S. A HR surface S_{HR} can be represented as a 0.5 level-set of a continuous 3D occupancy field:

$$f^{gt}_{HR}(x_{HR}) = \begin{cases} 1, \text{if } x_{HR} \text{ is inside } S_{HR} \\ 0, \text{otherwise} \end{cases} \quad (1)$$

where x_{HR} is a point in the 3D space around the HR surface S_{HR}. To map the LR surface to the HR surface, we adapt the implicit function representation to the 3D super-resolution shape. We define a new ground truth for the high-detail implicit function of the estimated super-resolution shape S_{SR}:

$$f^{gt}_{SR}(x_{LR}) = \begin{cases} 1, \text{if } x_{LR} \text{ is inside } S_{HR} \\ 0, \text{otherwise} \end{cases} \quad (2)$$

where x_{LR} are the 3D points sampled from the space of the LR surface S_{LR}. In contrast to the common implicit function, f^{gt}_{SR} is created by classifying the 3D points x_{LR} with respect to the HR surface instead of the LR one: some x_{LR} points are labelled as 'outside' even if they are 'inside' S_{LR}, and viceversa.

We apply SuRS to the task of 3D human digitization from a single LR RGB image. We modify the pixel-aligned implicit function representation first introduced by Saito et. al. [3] for this task by adapting it to the reconstruction of super-resolution shape from a low-resolution image I_{LR}:

$$f_{SR}(\phi(p_{LR}, I_{LR}), z(x_{LR}), \hat{s}_{MR}) = \hat{s}_{SR}, \hat{s}_{SR} \in \mathbb{R} \quad (3)$$

where x_{LR} are the 3D points sampled from the space of the LR surface S_{LR}, $p_{LR} = \pi(x_{LR})$, $\hat{s}_{MR}$ is a mid-resolution estimation of the shape computed with $f^{gt}_{HR}(x_{HR})$ from a LR input image and $\hat{s}_{SR}$ is a super-resolution estimation of the shape computed with $f^{gt}_{SR}(x_{LR})$. SuRS is modelled via a neural network architecture composed of 3 modules and trained end-to-end (Fig. 1).

• **Image features extractor**: We design a novel U-net architecture to extract both high and low resolution features. The former embeds the fine detail of the input image while the latter maintains holistic reasoning.

• **Mid-resolutiono MLP**: a classic pixel aligned implicit function $f_{MR}(\phi(p, I_{LR}), z(x_{HR}))$ is modelled by this first MLP and its estimation $\hat{s}_{MR}$ represents a mid resolution shape because the fine details are not embedded in the low resolution input image and cannot be reproduced.

• **Super-resolution multi-layer perceptron (SR-MLP):** the final estimation $\hat{s}_{SR}$ is obtained by a second MLP: $\hat{s}_{MR}$ is concatenated with the feature embedding and processed by SR-MLP in order to facilitate the learning of the map. Compared to MR-MLP, this one infers the difference between S_{LR} and S_{HR} to the final estimation, representing fine details on the super resolution shape even if they are not represented in the low resolution input image.

The super-resolved estimation is obtained and the marching cube algorithm is applied to reconstruct the super-resolution shape.

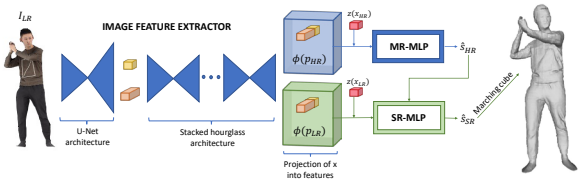
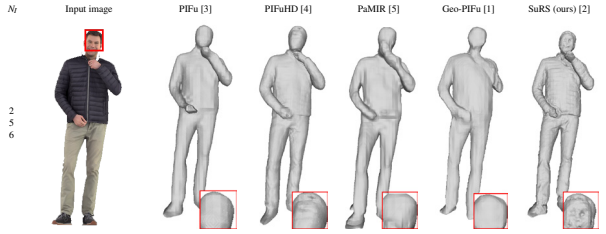


Figure 1: Overview of the approach. Given a low-resolution input image f a human, SuRS outputs a super-resolution shape by learning an high-quality implicit representation.

3 Results

We evaluate our method by comparing with state-of-the-art approaches for 3D human digitization from single image from synthetic data (Fig. 2) as well as from real data (Fig. 3). PIFuHD [4], PaMIR [5] and Geo-PIFu [1] use auxiliary data to facilitate the reconstruction while SuRS [2] and PIFu [3] uses only RGB images both during training and inference.



	PIFu [3]	PIFuHD [3]	PaMIR [5]	Geo-PIFu [1]	SuRS [2]
CD	1.519	1.062	1.644	1.853	**1.057**
Normal	0.1198	0.1257	0.1410	0.1658	**0.1227**
P2S	1.581	**1.037**	1.740	1.652	1.247

Figure 2: Qualitative (top) and quantitative (bottom) comparisons with state-of-the-art approaches for 3D human digitization from single image.

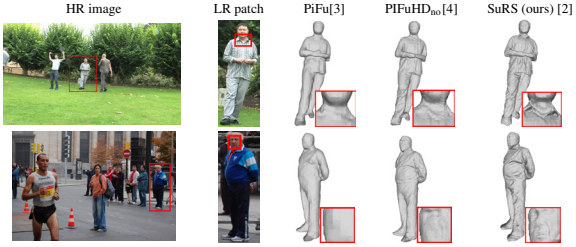


Figure 3: Real data comparisons with approaches that only use RGB input.

[1] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*, 2020.

[2] Marco Pesavento, Marco Volino, and Adrian Hilton. Super-resolution 3d human shape from a single low-resolution image. *arXiv preprint arXiv:2208.10738*, 2022.

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

[4] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.

[5] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Reconstructing Multiple People and Objects from a Single Image

Sarthak Batra, Simon Hadfield, and Armin Mustafa

The significant growth in gaming, self-driving cars, AR/VR, and film making, has led to an increased demand for reconstructing objects in 3D in general dynamic scenes from a single image. However, the majority of the existing methods reconstruct either only rigid or non-rigid objects in the scene with the incorrect spatial arrangement. In this paper, we simultaneously reconstruct both rigid and non-rigid objects in 3D with the correct relative spatial arrangement in a globally consistent 3D scene from a single RGB image. The contributions of the paper are as follows: (a) A unified approach to reconstruct rigid and non-rigid objects in a dynamic scene. The models of people and objects in the scene are animatable in contrast to existing methods.; and (b) Create a spatially coherent reconstruction with correct relative location of humans and objects in the scene. A qualitative and quantitative evaluation is performed on KITTI and WAYMO datasets against state-of-the-art method in multi-object reconstruction [4] demonstrating improved performance.

1 Methodology

In this paper, we present a method that can simultaneously recover the 3D shape and spatial arrangement of humans and objects in complex outdoor street scenes from a single RGB image, as shown in Figure 1. We present a method that takes a single RGB image as input and outputs the 6 D.O.F poses for humans and objects in a common 3D coordinate system along with their 3D models. We begin by first detecting humans and objects using an object detector [1], which gives a bounding box around each object. People in the scene are reconstructed using state-of-the-art human reconstruction methods [2] and predefined meshes are used for non-rigid objects such as cars, bicycles, trucks etc. The meshes of both people and objects in the scene are fully animatable models and have tracking information over time.

Bounding boxes are extracted for these objects followed by estimation on translation and orientation of each object to obtain a spatially coherent reconstruction of the entire scene For rigid objects, we crop the objects from the image using their corresponding bounding boxes. Then we pass those cropped images as input to our model which returns the dimension and local yaw for the objects and using these we determine the locations and place the objects accordingly in the scene. For humans convolution layers are used to extract features of objects along with fully connected layers to obtain 3D object pose (translation and orientation). We use the idea described in [3] that the perspective projection of a 3D bounding box should fit tightly within its 2D detection window. We assume that the 2D object detector has been trained to produce boxes that correspond to the bounding box of the projected 3D box. Our method uses 3D to 2D bounding box projection constraints to recover translation and a multi-bin orientation loss and a deep CNN network is used, in which the input is the image of an individual object and the output are the dimensions (height, width and length) along with the allocentric orientation for that object. Using the dimensions and allocentric orientation we determine the translation of the objects.

2 Experiments

We implemented our method using PyTorch. For training the model, we use Stochastic Gradient Descent as an optimizer, with learning rate and momentum set to 0.0001 and 0.9 respectively. We use batch size 8, and run the training for 50 epochs. Other hyperparameters like γ and w are set to 0.6 and 0.4 respectively. Each cropped window is resized to a height and width of 224 pixels.

The results are demonstrated on the KITTI and WAYMO datasets by reconstructing the spatial arrangement of multiple people and objects present in an uncontrolled urban street environment in Figure 2. We mainly focus on car, cycles and humans which are most likely objects to be seen on urban streets. We perform quantitative comparison with PHOSA [4] following the average Orientation Score(OS) and average L2 loss as performance metrics across all test images for KITTI and WAYMO dataset. The results in Table 1 demonstrate that the proposed method out performs PHOSA in terms orientation estimation and dimension estimation across all classes(person, cycle and car).

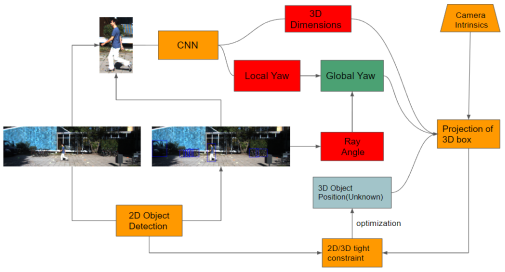


Figure 1: The pipeline to infer 3D position from 2d bounding box

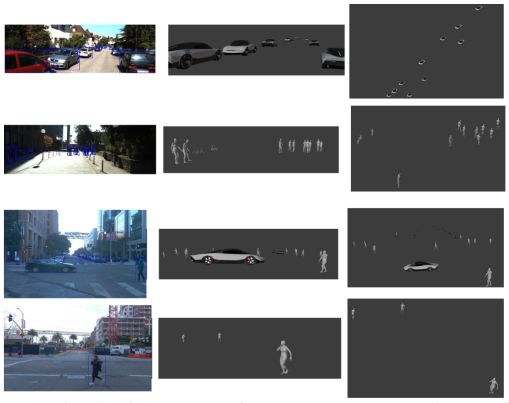


Figure 2: **Qualitative results of our method on test images from KITTI(top 3 rows) and Waymo (bottom 2 rows) datasets. The order from left to right is input image, front view and top view of the reconstruction respectively.**

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[2] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2021.

[3] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *CVPR*, 2017.

[4] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

Method	Dataset	Class	Evaluation Metrics	
			OS	Dim.Acc
Our Method	Kitti	Person	0.9828	0.9513
		Car	0.9723	0.9614
		Cycle	0.9816	0.9225
Phosa	Kitti	Person	0.6381	0.9213
		Car	0.5737	0.9312
		Cycle	0.5349	0.8927
Our Method	Waymo	Pedestrian	0.7823	0.8764
		Vehicle	0.7720	0.8345
		Cyclist	0.7657	0.8137
Phosa	Waymo	Pedestrian	0.5825	0.8296
		Vehicle	0.5149	0.7949
		Cyclist	0.4817	0.7729

Table 1: Quantitative comparison with PHOSA

RealMonoDepth: Self-Supervised Monocular Depth Estimation for General Scenes

Mertalp Ocal, Armin Mustafa, Adrian Hilton
{m.ocal,armin.mustafa,a.hilton}@surrey.ac.uk

Centre of Vision, Speech and Signal Processing (CVSSP),
University of Surrey

Learning monocular depth across diverse scenes is a challenging problem, due to large changes in depth range. Indoor scenes have a depth range of $< 10m$ and outdoor scenes are commonly 100s of meters, Fig. 1. Monocular depth estimation should estimate depth across scenes with a wide variation in the depth range. Existing self-supervised methods can be trained only on datasets with similar depth ranges [4, 9], limiting the number of images that can be used for training. As a result, they demonstrate poor generalisation performance and can only perform specific tasks, such as depth estimation in outdoor driving scenes with a fixed stereo baseline. We introduce RealMonoDepth a self-supervised monocular depth estimation approach which learns to estimate the real scene depth for a diverse range of indoor and outdoor scenes. A novel loss function with respect to the true scene depth based on relative depth scaling and warping is proposed. This allows self-supervised training of a single network with multiple datasets for scenes with diverse depth ranges from both stereo pair and in the wild moving camera datasets.

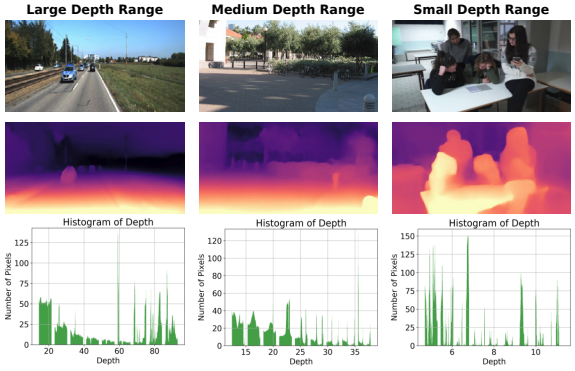


Figure 1: Depth estimation for scenes with different depth ranges: left column - outdoor $\approx 500m$; middle - outdoor $\approx 25m$; and right - indoor ($\approx 10m$). **Top Row:** Input image. **Middle Row:** Depth from proposed method. **Bottom Row:** Depth histogram.

Method: Given two images of a scene from different viewpoints (I_1, I_2), the depth network (Relative DepthNet) predicts the corresponding per-pixel relative depth maps ($D_{1,Rel}$, $D_{2,Rel}$) using shared weights. The relative depth maps are transformed to real depth (D_1, D_2) using the scale transform module. The self-supervised loss is then computed using the warped real depth estimates (D_1, D_2) and warped images (I_2, I_1). Using estimated calibration and camera poses, real depth values allow us to reconstruct the input images (I'_1, I'_2) and depth maps (D'_1, D'_2). This information is interpolated to compute photometric loss (L_{ph}) and geometric consistency (L_{gc}) loss, that supervises the depth network. SSIM and smoothness losses are used to regulate the depth estimation. The loss function for the proposed method is:

$$L = \sum_s \lambda_{ph} L^s_{ph} + \lambda_{gc} L^s_{gc} + \lambda_{ssim} L^s_{ssim} + \lambda^s_s L^s_{smooth} \quad (1)$$

where s indexes over different image scales and λ_{ph}, λ_{gc}, λ_{ssim} and λ_{smooth} are the weighting terms. To learn to estimate depth from images across diverse scenes with varying depth ranges, we normalise depth across the images and datasets using a non-linear scale transform and train the network to estimate relative depth. Given the relative depth map prediction as input, our scale transform module outputs the real depth map. This is formulated as:

$$D_k = \mu_k e^{(D_{k,Rel})} \text{ for } k = 1, 2 \quad (2)$$

where μ_k is the median depth value for two images I_1 and I_2, D_1 and D_2 are the real scene depth maps. During training, camera calibration is required to estimate the median depth value for the scale transform. For datasets with unknown calibration, an off-the-shelf SFM method i.e. COLMAP is used to solve for camera calibration and sparse correspondences between views. If the camera calibration is known (e.g. KITTI dataset), it is used to compute sparse correspondences between views. The sparse correspondences are triangulated in 3D exploiting camera pose to get a sparse reconstruction of the scenes. The sparse 3D points are projected on each view to obtain sparse depth maps for each viewpoint. Depth values are then sorted and the median depth value is estimated. Median depth enables prediction of real depth maps which are used together with the input images to estimate the loss in Equation 1.

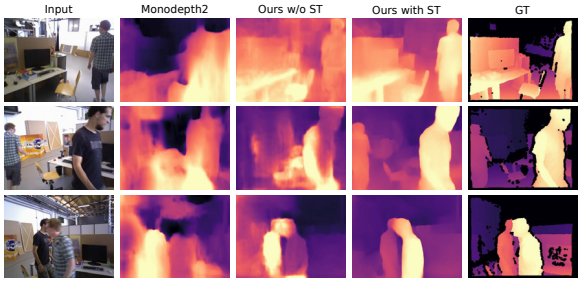


Figure 2: Qualitative comparisons of proposed method on TUM dataset.

Results : In Table 1 and Fig. 2, we demonstrate that the proposed self-supervised loss function using real depth dramatically improves generalisation performance when trained on both moving camera (Mannequin Challenge (MC) mostly indoor) and stereo (KITTI outdoor) datasets jointly. These datasets contain both indoor (1–10m) and outdoor (1–1000m) scenes with a wide variation in depth range. We test the same trained model on several benchmark datasets which the network has not seen during training: Make3D (outdoor buildings), NYUDv2 test split (indoor) and dynamic subset of TUM-RGBD (humans in indoor environments).

Table 1: Quantitative results on Make3D, NYUDv2 and TUM Dynamic Objects RGBD datasets for different methods trained on various scenes.

		Make3D			NYUDv2			TUM		
	Method	Training	Abs Rel	RMS	Training	Abs Rel	RMS	Training	Abs Rel	RMS
Supervised	Laina [5]	Make3D	**0.176**	4.45	NYU	0.129	0.583	NYU	0.223	0.947
	Chen [1]	DIW	0.550	7.25	-	-	-	NYU+DIW	0.262	1.004
	Li [6]	Megadepth	0.402	6.23	-	-	-	-	-	-
	Fu [2]	-	-	-	-	-	-	NYU	**0.194**	0.925
Self-supervised	Monodepth [3]	KITTI	0.525	9.88	-	-	-	-	-	-
	Monodepth2 [4]	KITTI	0.322	7.42	KITTI	0.342	1.183	KITTI	0.427	1.616
	Ours	KITTI	0.300	**7.09**	KITTI	0.287	0.956	KITTI	0.356	1.406
	Ours	MC	0.335	7.82	MC	0.200	0.701	MC	**0.182**	1.229
	Ours	MC+KITTI	**0.297**	7.38	MC + KITTI	**0.196**	**0.691**	MC+KITTI	0.188	**0.985**
Unsupervised	Zhou [9]	KITTI	0.651	8.39	-	-	-	-	-	-
	TrainFlow [8]	-	-	-	NYU	**0.189**	**0.686**	-	-	-
	DDVO [7]	KITTI	**0.387**	**8.09**	-	-	-	-	-	-

[1] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in neural information processing systems*, pages 730–738, 2016.

[2] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.

[5] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[6] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

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

[8] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020.

[9] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

A Deep Learning Based Tone Mapping Operator for 4K HDR Images Focusing on Color Accuracy and Artistic Intent

Junbin Zhang, Yixiao Wang, Hamid Reza Tohidypour, Mahsa T. Pourazad, Panos Nasiopoulos
{zjbthomas,yixiaow,htohidyp,pourazad,panosn}@ece.ubc.ca

Department of Electrical and Computer Engineering,
University of British Columbia,
Vancouver, Canada

High dynamic range (HDR) has arguably been established as the preferred image and video format for content providers. As standard dynamic range (SDR) displays still dominate the market, there is a need for finding efficient ways to convert HDR content to the SDR format, a process known as tone mapping. Recently, many tone mapping operators (TMOs) have been proposed that are based on deep learning approaches. However, the biggest challenge in training such deep learning networks is lack of truthful SDR and HDR datasets that would lead to highly accurate TMOs. In this paper, we introduce a new high-quality 4K HDR-SDR dataset of image pairs, covering a wide range of brightness levels and colors. We propose a TMO that is based on generative adversarial network (GAN).

Dataset Construction. To create a dataset that covers a wide range of brightness and colors and a huge range of scenes, we extracted around 2,000 unique 4K HDR images from 14 representative high-quality 4K HDR videos from the 4K Media depository [5]. We categorized these HDR images in terms of brightness based on the expected value of the luminance in the light domain.

To find the tone-mapped SDR images with the best visual quality for the corresponding HDR images, we conducted a comprehensive subjective test. We tone-mapped a large number of representative HDR images using multiple state-of-the-art TMOs and asked the subjects to pick the tone-mapped image with the best visual quality. The results showed that Mantiuk's [2] TMO yielded the best bright tone-mapped images, while for medium scenes the Ploumis's [3] TMO performed better. Finally, for dark scenes, once more Mantiuk's and Ploumis's TMOs outperformed the rest with both having their strengths and weaknesses. Hence, we included both SDR images in our database (for each such HDR input, one of the two SDR images is randomly selected as the target during training).

Network Design. We modified the ESRGAN network architecture [4], to match our tone mapping requirements. Only the generator of ESRGAN is changed. We removed the up-sampling convolutional layers in the original ESRGAN (see dotted lines, Fig. 1) so that the resolution of the output matches that of the input. In addition, we added a tangent hyperbolic (Tanh) activation function at the end of the network. Our experiments showed that use of the Tanh activation function allows our network to learn to avoid clipping and eventually move all the pixel values within the [-1, 1] range. The rest of the generator, the discriminator, and the objective functions remain the same.

Evaluation results. We compared our TMO with Mantiuk's and Ploumis's TMOs.

1) Subjective test. The best reference image for the generated SDR is the original HDR, as this comparison allows to determine if the mapping process preserved the original artistic intent. We, thus, conducted a subjective test to compare the resultant tone-mapped images with the original HDR.

Our testing dataset consisted of 51 representative frames from 17 4K HDR sequences provided by [6], which include a wide variety of bright, median, and dark scenes. Participants were shown the resultant tone-mapped images side-by-side with their original HDR version. We observed that the subjects ranked the images generated by our TMO to have the highest visual fidelity to the original HDR content. It is also worth noting that its performance remained steady across the various scenes.

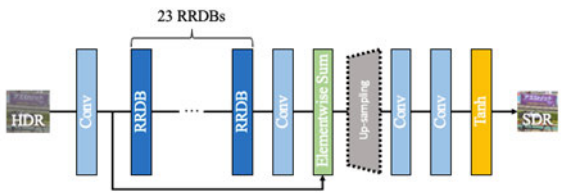


Figure 1: Our modified generator.

Figure 2: Comparison of results between our proposed method, Mantiuk's and Ploumis's TMOs.

2) Qualitative visualization. We visualized a representative frame from one 4K HDR video [7] in Fig. 2. We observe that the visual quality of our network is superior to that of the other TMOs, preserving global and local information. That is due to the fact that the convolution operators and the layers allow it to learn features ranging from almost the global level to the pixel level.

3) Objective metric. We also evaluated the performance of our approach using the Naturalness Image Quality Evaluator (NIQE) [1], a no-reference metric that measures the naturalness of the generated SDR images. Our method was shown to achieve the best (lowest) NIQE among all three TMOs.

Conclusion. Our TMO achieves high perceptual quality, maintaining the artistic intent and providing better color representation compared to existing state-of-the-art TMOs.

[1] A. Mittal et al. Making a 'completely blind' image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2012.

[2] R. Mantiuk et al. Display adaptive tone mapping. *ACM Trans. Graphics*, 27:1–10, 2008.

[3] S. Ploumis et al. Perception-based histogram equalization for tone mapping applications. In *DMIAF*, 2016.

[4] X. Wang et al. ESRGAN: enhanced super-resolution generative adversarial networks. https://arxiv.org/abs/1809.00219, 2018.

[5] 4K Media. 4K Media. https://4kmedia.org/, 2022.

[6] The Institute of Image Information and Television Engineer. Ultra-high definition/wide-color-gamut HDR standard test sequences. https://www.ite.or.jp/content/test-materials/uhdtv_hdr/, 2019.

[7] Jacob + Katie Schwarz. Morrocco 8K HDR 60FPS (FUHD). https://www.youtube.com/watch?v=hVvEISFw9w0&t=195s, 2018.

Volumetric and User-Centric Rendering Techniques for Lens Flare and Film Grain in Virtual Reality Environments

Johann Wentzel
http://johannwentzel.ca
Lesley Istead
https://carleton.ca/hci/people/lesley-istead/

School of Computer Science,
University of Waterloo
School of Information Technology,
Carleton University

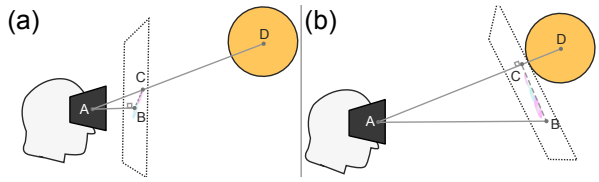


Figure 1: Two techniques for rendering lens flare in VR: (a) *Headset Flare*, which mounts the rendering plane in front of the user's head; and (b) *Directional Light Flare*, which mounts the rendering plane in front of the sun. The angle between the user's gaze direction (*AB*) and the direction toward the sun (*AD*) determine the spread and angle of the flare's rings, expanding outward from plane intersection point *C* along *CB*.

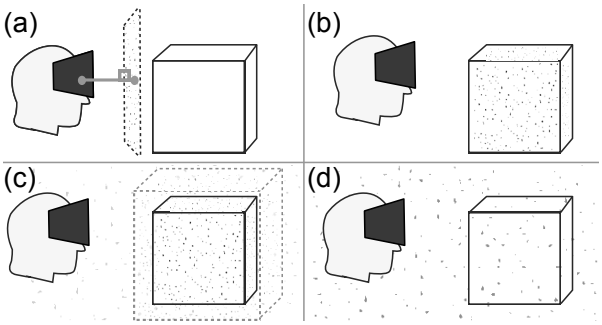


Figure 2: Four techniques for rendering film grain in VR: (a) *Headset Grain*; (b) *On-Surface Grain*; (c) *Object Cloud*; and (d) *Free Cloud*.

While several techniques can recreate lens flare and film grain effects in both monoscopic [1] and stereoscopic [2] formats, little work has explored applying these effects in real-time immersive environments like virtual reality (VR). Developing visual effects for VR involves several distinct challenges compared to other formats, the most prominent of which is user comfort. If a user moves in space, how should an effect react in order to remain visible, while ensuring that the user can fuse their left- and right-eye views? To address this, we present several methods for producing lens flare and film grain effects in VR. Lens flare is rendered in two different styles, taking as input the headset's forward direction and vector toward a light source. The primary difference between these styles is the position of the rendering plane, either mounted to the user's headset (*Headset Flare)* or mounted on the light source itself (*Directional Light Flare*). Film grain is rendered using two techniques: shader techniques which render grains as a 2D texture (*Headset Grain* and *On-Surface Grain*); and volumetric techniques which render grains as 3D particles within the environment (*Object Cloud* and *Free Cloud*).

We simulate lens flare with two artifacts rendered in the environment. The first is a series of semi-transparent rings with chromatic scattering similar to those found in analog lens flare. Ring positioning is based on three components (Figure 1): the vector of the user's headset gaze direction (*AB*), the vector between the user's headset and the sun (*AD*), and the intersections of those vectors with a rendering plane. Ring artifacts are distributed on the rendering plane the along the vector between these intersection points (*B* and *C*), with the size of $\angle CAB$ determining their spread. From the user's perspective, this results in ring artifacts splayed from the light source toward the center of their camera view, increasing in spread as the user moves their gaze away from the light. These rings are rendered in world space such that the user's eyes can fuse the two lens flare images. The second effect, a static glare artifact accompanied by a volumetric blur, is rendered at point *C* to appear to the user as surrounding the sun. Both effects fade in and out upon $\angle CAB$ reaching a set size, which in our implementation was 15 degrees.

We implemented lens flare in two styles, the difference between them being the position and anchoring of the rendering plane. *Headset Flare* (Figure 1a) displays these artifacts on a plane about 5 cm in front of the user's headset position, anchored to and centered with the user's headset view. From the user's perspective, flare effects would appear near their headset, close to their eyes. *Directional Light Flare* (Figure 1b) positions this plane near the sun, distant from the user. From the user's perspective, flare effects would appear distant, spread across the sky.

We simulate film grain using two rendering techniques. The *Shader* technique uses Perlin noise to procedurally generate grains of varying shape, brightness, and transparency. This shader's intensity and number of grains varies based on the distance from the camera to the object to which this shader is applied. This results in a flat texture containing small, semi-transparent grains simulating film grain. The *Volumetric* rendering technique creates small 3D grain objects, randomly scaled between 0.5 and 3 cm in all 3 dimensions, with random brightness and transparency. When applied to a volume, this technique creates a bounding volume V_b of a user-configurable scale larger than the original volume V_o (in our implementation 110% of the V_o size), and generates grain objects at random points inside the volume $V_b - V_o$. Grain positions are chosen by randomly choosing two vertices in the volume $V_b - V_o$ then linearly interpolating between them by a random amount. Grains are repositioned every 20 ms as recommended by Templin et al. [2]. Each grain is rendered in world space, meaning the user's eyes can focus on individual grains.

We implemented the two rendering techniques in four different styles. The first two use *Shader* rendering, applying the Perlin shader to either the user's view or objects in the scene. *Headset Grain* (Figure 2a) renders the grain texture on a plane covering the user's entire field of view, anchored to the user's head about 5 cm away. Users perceive film grain particles along a single "sheet" in their vision. *On-Surface Grain* (Figure 2b) uses the Perlin shader to render grain as a flat texture, which is then added to the surfaces of objects in the scene.

The last two styles use *Volumetric* rendering, creating grains as individual 3D particles in the scene. *Object Cloud* (Figure 2c), inspired by previous work on perceptually-motivated stereoscopic film grain [2], renders grains at varying density depending on the camera's distance from objects in the scene. Grains are randomly distributed in space where there is no object in view, but become more densely concentrated around objects in the scene depending on their depth from the camera. Infinite-depth grain is achieved by creating one large surrounding grain volume for particles to appear randomly, while objects in view dynamically add or remove particles from their grain volume V_b depending on their Euclidean distance to the headset. *Free Cloud* (Figure 2d) disperses grains randomly within a large predetermined cylindrical volume surrounding the user. Because the volumetric grain technique is applied to the user instead of an object, we set the size of V_o to be 0 such that grains can appear at any point within the user's surrounding cylindrical volume.

An 8-participant pilot study of these lens flare and film grain techniques showed that participants preferred volumetric film grain effects and distant lens flare effects in VR environments, but context and the surrounding scene can affect the strength of this preference.

[1] Matthias Hullin, Elmar Eisemann, Hans-Peter Seidel, and Sungkil Lee. Physically-based real-time lens flare rendering. *ACM Trans. Graph.*, 30(4), 2011.

[2] Krzysztof Templin, Piotr Didyk, Karol Myszkowski, and Hans-Peter Seidel. Perceptually-motivated stereoscopic film grain. *Comput. Graph. Forum*, 33(7):349–358, October 2014.

Learning Texture Transformer Network for Light Field Super-Resolution

Javeria Shabbir — College of Computing, Georgia Institute of Technology

M. Zeshan Alam — Department of Computer Science, Brandon University

M. Umair Mukati — GN Jabra A/S

Contrary to the traditional camera, the light field (LF) camera captures light rays approaching its surface, preserving the angular information of the incident light rays on the camera's sensor. Light field acquisition can be done in a variety of ways, including micro-lens arrays (MLAs) [1], coded masks [2], and camera array [5]. Among these different implementations, MLA based LF cameras offer a cost-effective approach. However, there lies a spatio-angular tradeoff in this design, since a single sensor is shared to capture both spatial and angular information.

To overcome this tradeoff, recently, some learning-based methods have been proposed [3, 6, 7]. In [7], a residual convolutional network is utilized to achieve high spatial resolution. Whereas, in [3] each sub-aperture image is super resolved individually and then to maintain parallax structure, a regularization network was appended. In [6], a texture transformer network is proposed that transfers high-quality texture from the reference image for target image generation. Motivated by this design, we propose to utilize a high-quality reference image to improve the resolution of all the views of the light field.

Our target is to improve the spatial resolution of light-field images by four times. We propose a modular technique comprising three different modules to achieve this task. The first module, named All-In-Focus High-Quality Reference Generator (AHQRG), generates a high-resolution image of the central view of the input low-resolution LF. The second module is the Texture Transformer Network for Image Super-Resolution (TTSR), which takes in two inputs: the output of the AHQRG module, which is treated as a reference image, and the low-resolution perspective image of the light field. TTSR is used to generate high-resolution views of the LF sequentially. However, since each view of the LF is super-resolved independently, the entire LF may not follow the regularity in the LF structure. This knowledge is taken as prior to further improve the LF's spatial resolution in the third module called LF refinement (LFREFINE).

All-In-Focus High Quality Reference Generator: We adopted spatial and angular interleaved convolution from [3] and modified this network such that it takes 7×7 views of the LF as input and results in a single view at the output. AHQRG module shown in Figure 1, has 3×3 convolutional layers followed by a set of four interleaved filters that alternates between the angular and spatial representation of LFs. Three 3d convolution layers are introduced after the interleaved filters to extract information from the spatial and angular coordinates simultaneously. Finally, the resulting residual is added to the $4\times$ bicubically upsampled central view producing a high-quality central view.

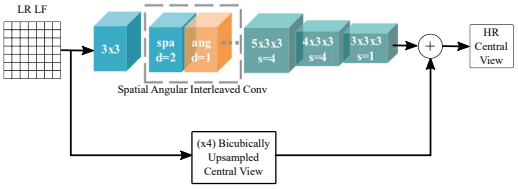


Figure 1: Block diagram of All-In-Focus High Quality Reference Generator (AHQRG).

Texture Transformer Network: TTSR module sequentially takes one low-resolution LF view and a high-resolution reference image produced by AHQRG module and outputs a corresponding high-resolution view of the LF. TTSR extracts feature from the reference image using a texture transformer to super-resolve low-resolution image. Bicubically down-sampled and up-sampled reference image serves as key for texture transformer, bicubically up-scaled low-resolution image as query and reference image as value. Texture transformer successfully transfers high-resolution features from reference to low-resolution image. The performance is further enhanced as the features at different scales produced by multiple texture transformers are combined to create a high-resolution image.

Light-field Refinement: Since each view of the LF is super-resolved using TTSR independently, the resulting LF may not follow the LF structure. We take this problem as an opportunity to further enhance the quality of the LF by imposing the LF constraint. Similar research is done in [3] to remove the artifacts during the view synthesis process. We utilize the LF refinement step (known as LF blending module) to improve the overall quality. Furthermore, EPI loss function is incorporated to enforce the LF prior.

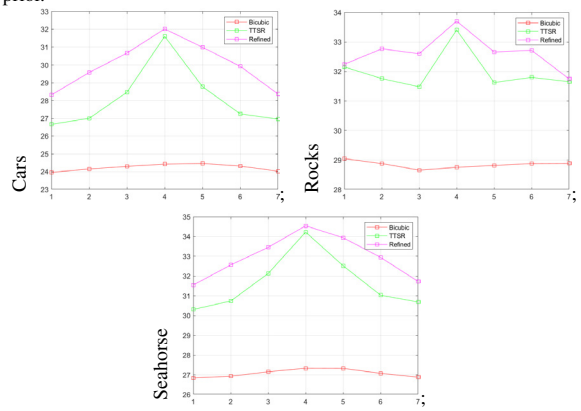


Figure 2: Comparison of the proposed method with bicubically resized view, and outputs from TTSR three different LFs from [4] dataset

Performance Evaluation: We quantitatively compare the performance of the proposed method with the bicubically resized LF and the output of sequentially super-resolved LF views from TTSR. We utilize PSNR as an evaluation metric. For simplicity of representation, we plotted the PSNR of only the diagonal views of the LF images in Figure 2. A significant gain in PSNR is evident for the proposed method as compared to the bicubically resized views. Though TTSR can generate a high-quality central view, as soon as the view deviates from the central location, the quality starts dropping. We believe this may be due to incorrect texture placement. On the other hand, LFREFINE considerably improves performance by applying LF prior. This module improves the quality of the views away from the central location as well as the central view.

[1] M. Z. Alam and B. K. Gunturk. Hybrid light field imaging for improved spatial resolution and depth range. *Machine Vision and Applications*, 29:11–22, 2018.

[2] M. Z. Alam and B. K. Gunturk. Deconvolution based light field extraction from a single image capture. In *IEEE Intl. Conf. on Image Processing*, pages 420–424, 2018.

[3] J. Jin, J. Hou, H. Yuan, and S. Kwong. Learning light field angular super-resolution via a geometry-aware network. In *AAAI Conf. on Artificial Intelligence*, volume 34, pages 11141–11148, 2020.

[4] N. K. Kalantari, T. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. on Graphics*, 2016.

[5] B. Wilburn, N. Joshi, V. Vaish, E. V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. on Graphics*, pages 765–776, 2005.

[6] F Yang, H Yang, J Fu, H Lu, and B Guo. Learning texture transformer network for image super-resolution. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[7] S. Zhang, Y. Lin, and H. Sheng. Residual networks for light field image super-resolution. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 11046–11055, 2019.

Towards Production-Ready Machine Learning For Animation Cleanups

Arun Leander Boudodimos — Film University Babelsberg KONRAD WOLF

Lena Gieseke — Film University Babelsberg KONRAD WOLF

Machine learning algorithms have been applied to various problems regarding free-hand sketches. In their survey, Xu et al. [2] identify sketch simplification as one major category of investigation, and Yan et al. [3] benchmark existing cleanup algorithms. These algorithms refer to the *cleanup* step in hand-drawn 2d animation, in which *rough* drawings are simplified by retracing the lines of the original. Pursuing a production-ready automatization of this step, based on real-world artistic processes, in the following we 1.) focus on synthesizing training data with a common production setup and 2.) propose for future work to investigate the integration of further artistic input. As there is no single clean solution to a rough drawing, artists receive additional information, such as a character sheet and textual descriptions, guiding the cleanup work. We envision to translate this approach by processing existing reference images of the target style as input, for example, to a cGAN model. Such references could be broken down into parameters describing the cleanup style. As a first step, we present work-in-progress to produce cleanups from roughs with a cGAN. Our goal is to develop a pipeline based on data that could realistically be generated and controlled in a production studio and that could in future work be extended to include further input such as character sheets.

Synthesizing Data

The quality and accessibility of machine learning approaches highly depend on the training data [1] but finding suitable existing data is challenging. However, synthesizing visual data is the everyday work of an animation studio. Also, in animation productions the rough animation may get significantly altered during the cleanup, e.g., shapes and poses might be artistically adjusted or completely removed and additional drawings might be added in between cleaned sequences. Both, the rough and the clean data might undergo many changes in a real-world production context. Hence, we propose a more flexible generation of synthetic training data with a 3d software, namely *Blender*. Blender includes the non-photorealistic rendering tool *Freestyle*, which renders lines based on 3d geometry (Fig. 1). Samples of the generated training data consists of 3d solids and abstract renderings. Such an approach of synthesizing training data with common 3d software, is not only more familiar for an animation studio, but synthesizing the data also prevents overfitting. However, as of now, it also appears to not encompass the desired domain (see Sec. Results).

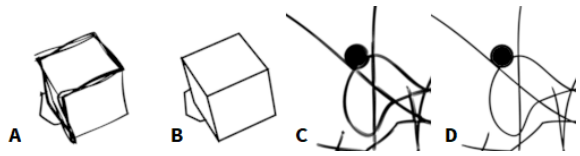


Figure 1: Examples of the synthetic training data generated with the non-photorealistic rendering tool *Freestyle* in Blender. A: 3d solids (rough), B: 3d solids (clean), C: abstract lines (rough), D: abstract lines (clean).

Architecture of the cGAN

For the image-to-image translation of rough to clean images, we use a modified Wasserstein GAN setup (Fig. 2) with grayscale raster images, with one channel for the rough and the other one for the clean data. The generator is a U-net, which receives the rough channel of the samples for input. Xception-like blocks are used between pooling (downsampling) and upscaling. The loss function is a combination of the generator's modified Wasserstein-loss as well as a direct comparison of the generator output and our synthesized target image. The critic is trained on the rough reference (Fig. 1A) concatenated with the clean target (Fig. 1B) or the generator's output. For this pipeline, we used a training set of 5000 128x128 samples and trained the model on mini-batches.

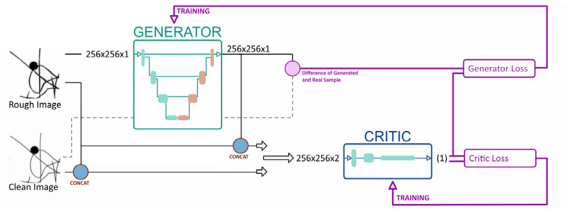


Figure 2: The architecture for training the cGAN.

Results

Visually, many requirements of a cleanup are met in our results (Fig. 3). The loss of the training data quickly drops and the difference between generated samples is indistinguishable despite the shallowness of the generator. The lines in the result have an even thickness, and repeatedly traced strokes are either separated into two distinct lines or combined into a single line. However, although the body parts of the bear character are clearly separated, not all lines are correct. The line between the ears of the bear is wrongly split, instead of being merged as a single line. Also, the knob-like shape of the nose is cleaned into an empty circle. Even though the synthesized data contains images with and without knobs (see Fig. 1C and 1D), the GAN has no means to assess if the nose is supposed to be filled or not. Hence, bear noses are randomly solved to filled or empty shapes. For consistent results, the training data requires information specific to this character, e.g., "the knob is the nose of a bear". The data needs to fulfill the requirements of the bear domain, which is difficult to synthesize.

Figure 3: The cGAN produces a cleanup (right) from the input image (left). The resulting lines have a consistent thickness and structure, but some strokes were not merged correctly, e.g., the top of the head and the nose has not been filled.

Next Steps

Possible improvements include adding more depth to the generator and critic. The generator would need to process global information from the image as a whole in order to determine how various parts of the image are translated. Furthermore, we would like to investigate how to supplement the input with data that is commonly available in animation productions, such as a series of reference images, from which the features of the subject can be inferred.

[1] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Real-Time Data-Driven Interactive Rough Sketch Inking. *ACM Trans. Graph.*, 37(4), 2018.

[2] Peng Xu, Timothy M. Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2022.

[3] Chuan Yan, David Vanderhaeghe, and Yotam Gingold. A benchmark for rough sketch cleanup. *ACM Trans. Graph.*, 39(6), 2020.

Distance in CLIP embedding space as a perceptual loss for fine-grained visual tasks

Gianluca Berardi *+
https://www.unibo.it/sitoweb/gianluca.berardi3/en
Yulia Gryaditskaya +
https://yulia.gryaditskaya.com/

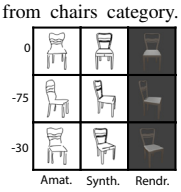
* Department of Computer Science and Engineering, University of Bologna
+ Centre for Vision, Speech and Signal Processing, University of Surrey

In this work, we evaluate the distance in CLIP [2] embedding space as a perceptual loss on 3 image domains: RGB renderings {rendering}, Non-Photorealistic Renderings (NPRs) {synthetic} and freehand amateur sketches [1] {amateur}. The high diversity and abstraction of amateur sketches represent an extremely challenging test scenario. At the same time, it is one of the image domain that could benefit the most from a strong perceptual loss, and is of particular interest in this study.

1 Methodology

We run two experiments. First, we study the in-variance of CLIP embeddings of 3D shapes represented with their multi-view projections in the three aforementioned domains. Second, we investigate the similarity of CLIP embeddings of individual viewpoints of the same object between the three considered domains.

In this work, we only consider 3D shapes from chairs category. In both experiments, we use three different views for every object, with the camera azimuth angles set to $0°$, $-30°$ and $-75°$. The inset on the right shows example viewpoints for one of the considered 3D shapes. We use the third layer of the pre-trained ResNet101 CLIP image encoder as our CLIP embedding space, following [3].



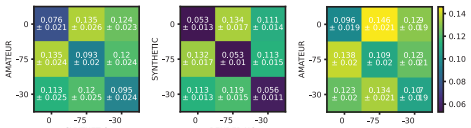
Experiment 1 We consider 10 objects, and compute the distance between two 3D shapes A and B as follows:

$$Dist(A^i, B^j) = 1/V \sum_{k=1}^{V} \left(CLIP(A^i_k) - CLIP(B^j_k) \right)^2 \quad (1)$$

where $V = 3$ is the number of views, and subscripts i and j denote one of three image domains.

For this experiment, we plot pairwise distances between shapes, when their views come from one of the three image domains. We can see that, in all three configurations, comparing the same object between different domains results in the lowest average distance (darker color) in most of the cases. This shows general robustness of the CLIP model across different domains. However, failure cases are possible, in particular, when amateur sketches are concerned: Sometimes the loss in CLIP embedding space can not confidently distinguish two shapes in two different domains.

Experiment 2 When comparing individual views k and h in domains i and j, the same object is considered for both domains, averaging over different objects:



Experiment 2

$$Dist(k^i, h^j)_{k,h \in \overline{1,3}} = 1/N \sum_{A=\mathrm{obj}_1}^{\mathrm{obj}_N} \left(CLIP(A^i_k) - CLIP(A^j_h) \right)^2 \quad (2)$$

where $N = 100$ is the number of objects. From the respective figure, we observe that for all configurations the average lowest values are obtained when comparing the same view. This shows that, in general, CLIP is able to match the same view of an object between different domains. However, when comparing amateur sketch views with views from other domains, we observe that the confidence intervals can overlap. This means, that occasionally an incorrect viewpoint in a different domain can be selected.

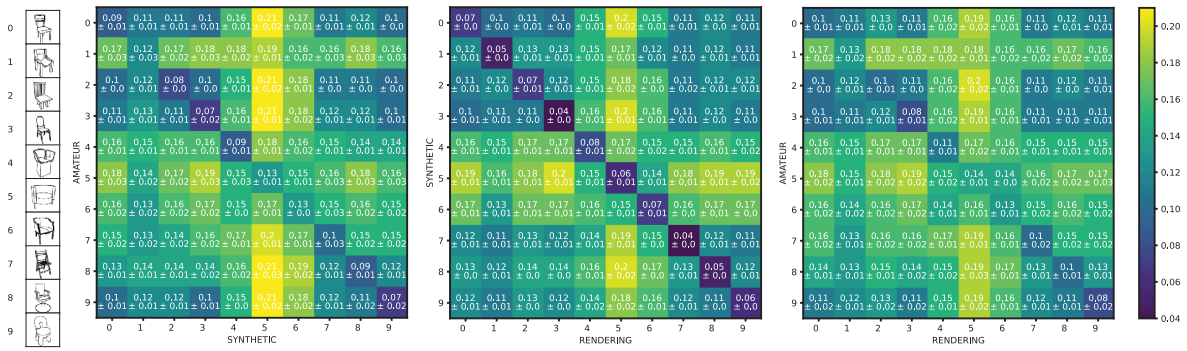
2 Conclusion

Overall, our analysis shows that the distance in CLIP embedding space is a promising perceptual loss. In both experiments, the performance is strong when comparing renderings and synthetic sketches: (1) the mean distance is low for the corresponding objects or individual views, and (2) even when considering standard deviations, there is no overlap with the distances to other objects, or views. Amateur sketches introduce a performance drop in both experiments. This is expected due to abstract nature and diversity of these sketches. However, we should also consider that amateur sketches labeled with a certain view are not always correct, because for humans is hard to draw objects from an exact point of view. This contributes to the lower performance on amateur sketches.

[1] Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Toward fine-grained sketch-based 3d shape retrieval. *TIP*, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[3] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *SIGGRAPH*, 2022.



Experiment 1

An Introduction to the Virtual Autonomous Cinematography Environment (VACE)

Will Kerr
wk307@bath.ac.uk
Wenbin Li
https://researchportal.bath.ac.uk/en/persons/wenbin-li
Tom S. F. Haines
https://researchportal.bath.ac.uk/en/persons/tom-fincham-haines

Department of Computer Science,
University of Bath,
BA2 7AY, United Kingdom

Figure 1: Example VACE usage: camera pose imitation based on face position. L: input, R: output. Input images courtesy [5]

1 Background

Virtual filming environments allow cinematographers to visualise, test and plan camera shots prior to costly real-world filming production. This creative workflow is predominantly manual, requires a lot of time and expertise, and benefits from expert-level cinematography training. These constraints can be reduced through automation and integrating prior knowledge into camera pose recommendation engines. This concept has received attention in previous work and by which this paper aims to develop further. The main contribution of this paper has been to develop an integrated environment allowing vision-based analysis and control of cameras within 3D virtual settings, named the Virtual Autonomous Cinematography Environment (VACE).

Commercial solutions and previous research work have shown progress in this area, covering the 3D environment, manual control of camera trajectories, and autonomous virtual cinematography research [2, 3, 4], However the open integration of camera pose and vision analysis tools has yet to receive much attention.

VACE offers a flexible environment where 3D scenes, actors, props, lighting and cameras can be positioned by manual and modular automated processes. This then provides a platform for further research into camera pose solving using a variety of novel techniques, in order to assist the cinematography planning process.

2 System Development

The VACE system was developed with the following aims: 1) to require as input a 3D scene, actor, and a goal, 2) to provide an optimization environment between 3D environment manipulation and virtual camera image analysis according to a desired goal, 3) to output proposed camera pose recommendations as per the goal.

Figure 2: VACE System Components L to R: 3D environment, I/O, Control

The basic building blocks for such a system are: a 3D environment which can be manipulated in a deterministic manner, some control program which is tasked to control the 3D environment, and a communication pathway that allows bi-directional interaction between the 3D environment and control program.

Considering previous work and available applications, Unreal Engine 4 (UE) was selected as the appropriate 3D environment — being well supported, allowing multiple virtual camera and timeline control, be externally controlled through API, and render virtual camera images to be made available to external programs. The communication mechanisms eventually selected was the UE API which includes HTTP and websocket control of nearly every parameter available within the engine, and a Spout [1] camera renderer plugin which allows shared GPU memory access to camera images from UE4 to external programs. Multiple camera feeds are produced simultaneously to allow parallel processing. Python was selected as the control program. See Figure 2.

3 Evaluation

The first task put upon VACE was to solve camera poses in order to imitate human face qualities from various reference images extracted from films [5]. This required optimisation, as such a Particle Swarm Optimization approach was taken as this provided good coverage within the allowable search bounds, generally reducing swarm variance after 50 iterations, and providing believable results for the final camera pose image.

A custom loss function with a combination of face width, position and angle measurements was created giving rise to 6 individual feature scores, combined in a weighted L1 fashion. Face detection was implemented with a yolov5 face detector and angle by an fsa-net model.

431 images from [5] were used as stimulus, tasking the system with imitating the face position and angle with a new scene / actor. Runtime defined 100 iterations and 48 particles by the GlobalBestPSO algorithm supplied by pyswarms python library. Examples shown in Figure 1.

4 Conclusion and Future work

VACE has been presented as a tool for further research into autonomous cinematography, and its use described in a basic application of camera pose solving for face qualities.

Further work is required in: 1) additional compositional element analysis (contrast ratio, key, blur etc), 2) temporally based recommendations (i.e. video), 3) integrating more generalised reference material so the system solves from an entire film or directors' back-catalogue, so *style* can be replicated, rather than individual images, and 4) a user-study to qualitatively measure how such a tool improves the cinematographic planning process.

[1] L. Jarvis. Spout - a video frame sharing system for Microsoft Windows. URL `https://spout.zeal.co/`.

[2] H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics*, 39(4), 2020.

[3] R. Ronfard. Film Directing for Computer Games and Animation. In *Computer Graphics Forum*, volume 40, pages 713–730. Wiley Online Library, Wiley Online Library, 2021.

[4] C. Sanokho, C. Desoche, B. Merabti, T. Y. Li, and M. Christie. Camera Motion Graphs. *SCA 2014 - Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 177–188, 2014.

[5] M. Savardi, A. Signoroni, P. Migliorati, and S. Benini. Shot scale analysis in movies by convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing*, pages 2620–2624, 2018.

Automatic Real Time Camera View Switching in Multi-camera Setup for Ice Hockey Games: An Object Detection Approach

Hamid Reza Tohidypour, Yixiao Wang, Mahsa T. Pourazad, Panos Nasiopoulos, Gurpreet Heir, Derinsola Ibikunle, Anthony Li, Fawaz Ahmed Saleem, and Zhao-bang Luo
http://www.dml.ubc.ca

Department of Electrical and Computer Engineering, University of British Columbia, Canada

Switching camera views while broadcasting ice hockey has a significant impact on the viewer's quality of experience. In professional coverage, this process involves expensive specialized equipment and highly skilled individuals such as camera operators and a director responsible for supervising and deciding the overall operation. Unfortunately, such an expense is prohibitive when it comes to broadcasting amateur community or school sports. In this case, despite the fact that more than one camera may be used, real-time coverage involves only a main view, without offering the option of watching another view that may better cover crucial moments during the game.

As a result, this monotonous coverage of regional games may potentially hinder the viewership and thus be detrimental in the progress of school and amateur sports. Thus, there is a need for a cost-effective, fully automated camera view switching system, which analyzes the importance of the scene covered by each camera and then switches the view in a manner that is pleasant to the viewer. Figure 1 shows an example of a multi-camera setup for hockey arenas.

In this paper, we propose an innovative camera switching method which is based on deep learning, namely the YOLOv4 architecture [3], and a temporal tracking scheme to automatically pick the most important view to be broadcasted. Here, in order to show the validity of our approach, we chose to train our network for ice hockey, as the network needs to be retrained for each sport, using a dataset that corresponds to the specific game. Our model receives video feeds from all the cameras around an ice hockey arena and detects the puck, net, goalie, players, and referees in real time with very good precision. A novel camera switching algorithm that weights the objects detected by each camera view according to their importance and uses the predicted confidence values for the different objects and temporal tracking to choose which camera view to be broadcasted. Our proposed method is play-centered unlike the player centered work presented in [5].

We based our method on the state-of-the-art object recognition approach named YOLov4 architecture. The main reason for this choice is that YOLOv4 reported to give promising results in detecting small objects with very fast inference time [3]. This is important since one of the main drawbacks of the Faster-RCNN based camera switching method proposed in [4], was the low accuracy of detecting the puck. In order to train our YOLOv4 model, we generate a comprehensive dataset for our application by collecting a large number of videos from local amateur ice hockey games. These videos were captured by cameras mounted on the side views and goalie views. In addition, we also collected a large number of professional ice hockey videos from YouTube [2]. From all the videos we selected 5000 well representative frames for the training-validation phase, avoiding subsequent, similar frames and preferring frames that included the puck. Almost half of these frames were selected from the videos captured from the side view, and the remaining from the goalie view. The selected frames were labeled according to our objects of interest (players, goalie, puck, net, and referee), while the audience were excluded. 80% of the training-validation dataset was randomly selected as the training dataset and the remaining 20% was considered for the validation phase.

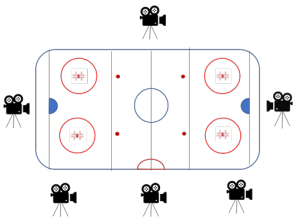


Figure 1: An example of a multi-camera setup for ice hockey arenas.

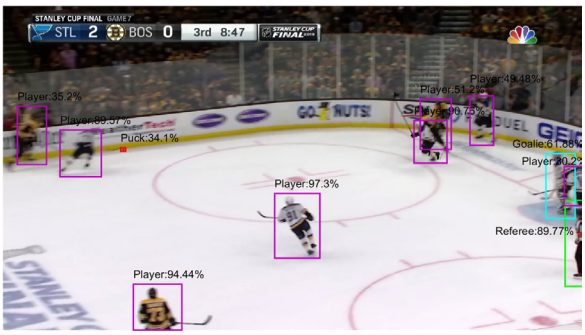


Figure 2: Example frame from the test set that shows the players, puck, goalie, and referees, which were detected correctly by our model.

The training-validation frames were utilized to train ourYOLOv4 using a state-of-the-art advanced research computing network [1]. Our trained model achieved very high accuracy on the validation set.

In order to evaluate the performance of our trained model, we examined it on the test videos with unseen frames. We used our deep learning model to detect the objects of interest. Figure 2 shows the predicted objects and the probability values assigned to the bounding boxes for anl example test image. Our camera switching algorithm considers the position and confidence level of detection of all the objects, as each one has different roles to play in determining the best camera view for the current moment of the game. It is important to note that designing our algorithm to be biased towards the importance of objects to the fans, will allow our solution to be focused on the action. Driven by professional game coverage, we assume that the most important object/event in hockey broadcasting involves the puck. Following the above observation and the outcome of many trials asking subjects to validate the validity of our switching scheme, we assigned a weight to the confidence values predicted for each object type according to its importance. More precisely, the confidence of each detected object in the current camera view is weighted according to its object type and the weighted values are summed up to calculate the score for the current camera view. Preliminary results show our camera switching method outperformed the state-of-the-art described in [4].

[1] *Compute Canada state-of-the-art advanced research computing network*. Available from: https://www.computecanada.ca.

[2] *2019 IIHF Ice Hockey World Championship*. Available from: IHF Worlds 2021, YouTube, https://www.youtube.com/c/IIHFWorlds/videos.

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao. Yolov4: Optimal speed and accuracy of object detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[4] H. R. Tohidypour, Y. Wang, M. Gholami, M. Kalia, K. Wen, L. Li, M. T. Pourazad, and P. Nasiopoulos. A deep learning based approach for camera switching in amateur ice hockey game broadcasting. In *5th International Conference on Signal Processing and Information Communications (ICSPIC 2022), Accepted*, 2022.

[5] L. Wu. *Multi-view hockey tracking with trajectory smoothing and camera selection*. Thesis, University of British Columbia. Retrieved from https://open.library.ubc.ca, 2008.

High-Quality Variable Speed Video for Stationary Cycling

Charles Malleson
charles.malleson@surrey.ac.uk

Centre for Vision, Speech and Signal Processing,
University of Surrey

We propose an approach for efficiently producing high-quality video of outdoor trails for playback on a stationary cycling rig. Footage is captured in a single pass walking along a trail using hand-held, consumer-grade equipment. Intermediate processing is performed on the footage to produce output video sequences for a range of target cycling speeds. The approach yields virtually artifact-free video with smooth, judder-free motion and realistic motion blur over a range of cycling speeds.

The video is played back on a stationary cycling rig (Fig. 1, right) at a speed varied according to the pedalling speed of the user. Because of the large size, high resolution and close proximity of the display to the user, high image quality and stable motion are essential for quality of user experience. Video content of off-road trails available on the web is typically recorded using action cameras (e.g. GoPro) mounted to the bicycle/rider. While the convenience and low cost of such capture setups is attractive, the video quality is often subpar. On the other hand, high-end video cameras with 4D stabilization (e.g. [2]) are beyond the reach of casual content creators. The proposed approach uses a consumer-grade mirrorless hybrid camera (Panasonic GH6) with a custom light-weight stabilization rig for cost-effective capture of high-quality UHD footage (Fig. 1, left). This hand-held rig allows for capture of trails where bicycles or other vehicles may not be permitted. More-over, through judicious choice of capture settings and processing steps, only one pass of the trail need be recorded (at walking speed) for playback at a range of cycling speeds, while exhibiting stable, judder-free motion and consistent, realistic motion blur.

Stabilization Based on the principle 'prevention is better than cure', we aim to have the video captured as steadily as possible. We thereby avoid the need for any in-camera or post-processing software-based video stabilization, which could degrade image quality. The lens and in-body optical image stabilization systems onboard the camera are enabled, and the camera is mounted on a commodity hand-held 3-axis gimbal (DJI Ronin SC [3]). These are effective in reducing most of the camera shake, however there remains an undesirable 'bobbing up and down' motion in the video due to the walking of the operator. To mitigate this, we propose to augment the standard gimbal, with its 3 rotational degrees of freedom, by mounting it in a custom active vertical stabilization rig.

The vertical stabilization rig uses an RC servo to move the base of the gimbal according to the detected vertical motion. An accelerometer mounted on the actuated platform provides vertical acceleration measurements to a microcontroller (on a Raspberry Pi Pico), which controls the servo using a basic feedback control loop. This reduces the vertical motion of the camera. Field trials using tracked fiducial markers indicate that the system reduces vertical oscillation at a typical walking speed of 4 km/h by approximately 56% (from 25 mm down to 11 mm peak-to-peak), thus producing in more convincing simulated cycling motion than the 3-axis gimbal alone. We expect that further reduction in the vertical motion could be achieved with the same hardware by refinement and further tuning of the feedback control software on the microcontroller.

Retiming We aim to match the effective shutter angle of the output video to the de facto standard [4] of 180° at all playback speeds. A naive retiming of the input footage to an arbitrary speed by sampling from the nearest input frame would result in the effective shutter angle decreasing (and the motion appearing choppier) with increasing playback speed. In general, naively blending multiple consecutive video frames to simulate motion blur leads to artifacts. It is possible to retime video arbitrarily while simulating convincing 180° motion blur by using optical flow-based warping and blending of input frames (for instance using Adobe After Effects [1]). Optical flow is, however generally computationally expensive and is prone to fail in the vicinity of fine structures such as plant stems, potentially leading to objectionable image artifacts [5].

In our approach, the video is captured at 59.94 fps with the shutter speed fixed at $\frac{1}{60}$ s, yielding an input shutter angle close to 360°. Output frames are then synthesised with an effective continuous 180° exposure for output at even multiples of the capture speed (or close to 180° for odd multiples), by blending (averaging) input frames (Fig. 2). Note that the blending needs to be performed in linear (gamma expanded) space,

Figure 1: Hand-held video capture rig (left) and stationary cycling rig (right). A commodity 3-axis gimbal is augmented with a custom vertical stabilization system. The capture vantage point and field of view are matched to the 55" UHD display of the playback rig.

otherwise the blended pixel values would be biased towards darker samples, resulting in artifacts. By capturing at walking speed, integer speedup factors from 2-8x can be used to achieve playback at a suitable range of cycling speeds without judder. These speed retimed sequences are pre-generated off-line and cued up at runtime for playback according to the user's current pedalling speed (and virtual gear).

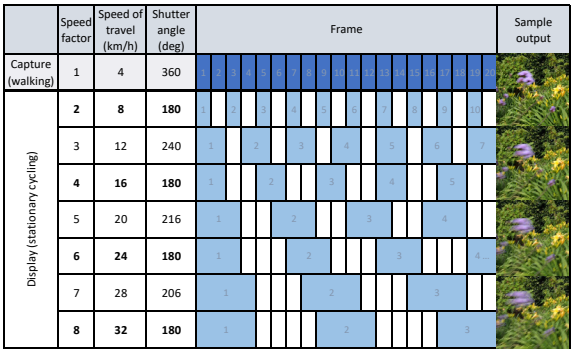


Figure 2: Input video sequence (walking at 4km/h), retimed to various target cycling speeds from 8-32km/h, by blending the appropriate input frames to produce judder-free output video with realistic motion blur.

Acknowledgement: *This work was partially supported by the Leverhulme Trust Early Career Fellowship scheme.*

[1] Adobe After Effects. `https://www.adobe.com/uk/products/aftereffects`. Accessed: 2022-08-31.

[2] DJI Ronin 4D. `https://www.dji.com/uk/ronin-4d`, . Accessed: 2022-08-31.

[3] DJI Ronin SC. `https://www.dji.com/uk/ronin-sc`, . Accessed: 2022-08-31.

[4] Ianik Beitzel, Aaron Kuder, and Jan Frohlich. The effect of synthetic shutter on judder perception — an HFR and HDR data set and user study. *SMPTE Motion Imaging Journal*, 129:42–50, 01 2020.

[5] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong. Optical flow and scene flow estimation: A survey. *Pattern Recognition*, 114:107861, 2021.

Enabling Virtual Theatre: Two Technical Approaches for Live Distributed Performance

Joe Geigel
http://www.cs.rit.edu/~jmg

Department of Computer Science,
Rochester Institute of Technology

1 Introduction

We define **Virtual Theatre** as shared, live performance with participants contributing from different physical locales. The challenges of virtual theatre involve not only integrating technical components to enable such an experience, but doing so while maintaining aspects of theatrical performance (e.g. liveness, social presence, and perspective [1]) that make it a particularly unique art form. In this work, we present two technical frameworks for enabling virtual theatre utilized on recent live productions.

2 Video Based Delivery

Inspired by the Interplay [3], a telematic theatre experience created and streamed over Internet2, we use a video based approach for our presentation of *Canadian Wiggler*, performed in September of 2020 as part of the Rochester 2020 Virtual Fringe Festival. (Figure 1).

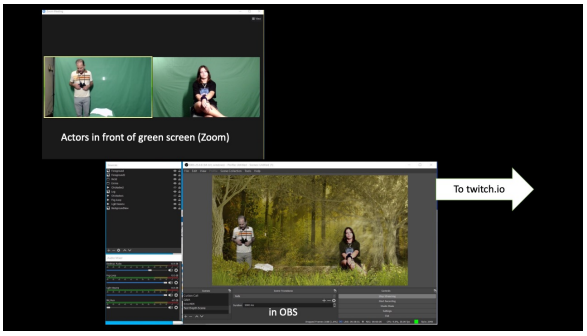


Figure 1: *The Canadian Wiggler* - 2020

Like the Interplay, the "stage" is created using realtime video mixing with layers originating from participants in different locations. We use OBS, an open source video mixing and presentation tool, to perform this real-time mixing with the result being streamed to a live video streaming service (we utilized `twitch.io`).

Actors performed in front of green screens and used Zoom to stream their performance. These feeds were captured, chroma-keyed and placed in the proper position on the composite video. Actors used a separate facetime connection using their phones to gauge their position on the virtual stage with respect to other actors.

The entire presentation is managed by the stage / screen manager who synchronized the placement, timing, and appearance of video layers which themselves containing animation and moving video elements.

Audience members watching the performance maintained a sense of shared social presence by using the live chat feature provided by the streaming platform.

The show was performed over a three day span with audience numbers ranging from 20-50 from all around the country. The show was conceived, created, rehearsed, and presented completely remotely with the participants involved never having to meet physically.

3 Theatre in the MetaVerse

Given the immersive nature of virtual reality, there is growing interest is using VR as a for experiencing live theatre [2]. In Fall 2021, we presented *Been Set Free* (Figure 2), a live dance performance realized completely in a virtual space with audience member immersed using VR headsets.

The main stage was implemented using Unreal Engine (UE) 4.26 with the performance implemented as a multiplayer game allowing all participants simultaneous access to the performance space. UE supports the creation of multiplayer games allowing us to utilize the already existing networking infrastructure provided by the engine.

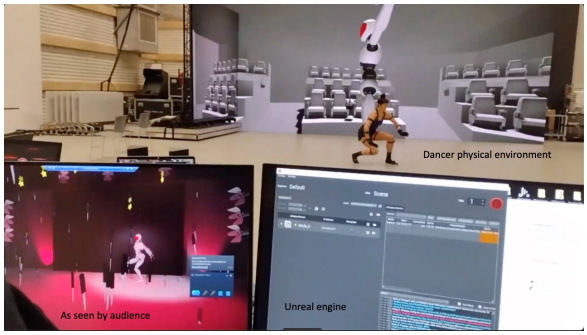


Figure 2: *Been Set Free* - November 2021

Each audience member was a player in the shared simulation and was represented in the virtual space by a futuristic head which reflects the head motions and gaze of the audience member (Figure 3).

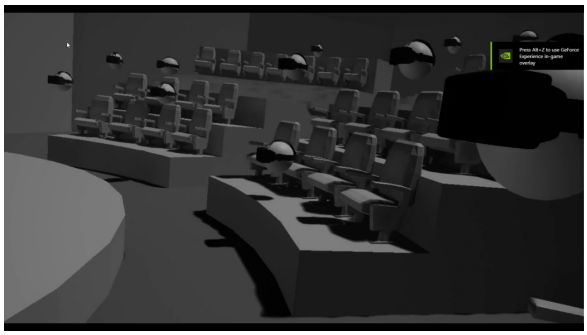


Figure 3: In world audience for *Been Set Free*

The dancing avatar's movement was defined using motion capture in conjunction with UE's LiveLink plugin. The dancer's physical environment was presented on a large video wall to immerse the dancer on the virtual stage and provide a social connection between the actor and audience members.

4 Conclusion

Both frameworks broaden the access of theatrical experiences to those not physically able to attend due to distance or other reasons. This new kind of distributed theatre also opens up new opportunities both technically and creatively as theatrical pieces are written specifically for these platforms.

[1] Joe Geigel. Creating a theatrical experience on a virtual stage. In *International Conference on Advances in Computer Entertainment*, pages 713–725. Springer, 2017.

[2] Linjia He, Hongsong Li, Tong Xue, Deyuan Sun, Shoulun Zhu, and Gangyi Ding. Am i in the theater? usability study of live performance based virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, New York, NY, USA, 2018. ACM.

[3] Charles Nichols, W. Scott Deal, Timothy J. Rogers, Jimmy Miklavcic, Beth Miklavcic, Many Ayromlou, Robert Wachtel, Paul Mercer, Joe Humbert, and Rob King. Musical performance over internet2 using the accessgrid. In *Proceedings of the 2006 International Computer Music Conference, ICMC 2006, New Orleans, Louisiana, USA, November 6-11, 2006*. Michigan Publishing, 2006.

TONGA - a project and equipment management platform for educational film productions

Dr. Michael Witt
https://www.filmuniversitaet.de/portrait/person/michael-witt
Stefan Beckers
https://www.filmuniversitaet.de/portrait/person/stephan-beckers

TONGA Project - IT-Service Department,
Filmuniversity Babelsberg KONRAD WOLF
https://tonga.filmuniversitaet.de/blog
https://www.filmuniversitaet.de

Project management software is an established tool in many different industries to handle the coordination of complex processes with different actors. However film production and especially the teaching of film production require special capabilities. These involve the handling of actions taken and assets created by project members as well as their supervision, revision and approval by supervisors during the production process.

Additional requirements are added in the university context by the need to rent out equipment to students for their film projects. In contrast to standard rental systems the aforementioned capabilities are required here too. Lecturers must supervise and approve requested equipment and should be able to engage with the students about choices made. Additionally a permission based system that depends on the students course of study, study progress and certificates acquired needs to be employed to meet the universities demands.

Tonga is an integrated project management system with a lightweight rental system. The software was developed tailored for universities. As of today, many different systems are often used to organise equipment rental, student project progress (like film productions) and financial management. All these systems often form stand-alone data islands which cause delays in organisational processes and incoherent data. *Tonga* was designed to integrate all required systems into one platform that encompasses the requirements of all stakeholders involved. Its name reflects this as the Kingdom of Tonga archipelago consists of 171 islands that form a single country.

Film production projects and other projects in the university context are highly individual. The software therefore supports different types of projects which are highly customisable. Each project is backed by a preconfigured workflow. A workflow consists of phases which again consist of assets. This allows lecturers to structure complex projects into logical units and to guide students through the process of realisation of their project. However this structure does not necessarily have to be a strict tree. Assets that span multiple phases (e.g. a film script which is evolved during the project process) can be placed into all of them without the need for users to copy any data or to upload things twice.

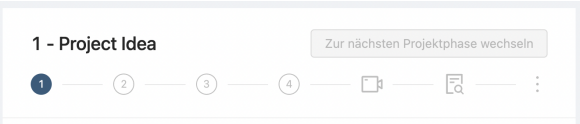


Figure 1: Project progress in *Tonga* is displayed by phases to emphasize the previous, current and upcoming steps required to successfully realise the project. Quick access to equipment rental and search functions are also present.

Tonga supports different types of assets that are created or utilised for film projects. Assets can be simple documents (which are automatically versioned), production plans, collections of contracts, film distribution information, links, annotated texts and lists of persons. Person list assets allow the project owner to assign people to different positions in the project. These positions can further be linked to contract assets to keep track of legal requirements. To encompass supervision and revision tasks each asset, as well as each phase, can be configured to require approval. Approvals can be granted by entities such as specific users, a group of users, as well as persons that assume a certain role in the project (e.g. production manager or post-production supervisor) or supervises a student for a specific position. With each approval documents and comments can be added to communicate more information within the project.

Finally, each asset supports comments for a discussion among project members and to share related information. Furthermore *Tonga* possesses a sophisticated notification system that collects messages generated by the application on different occasions. These messages are presented in a centralised message board with quick access links to the relevant software component like the project detail page, approval screen or equipment rental information.

To support the individual nature of film productions *Tonga* creates a *project workflow* (PW) for each project based on the selected preconfigured workflow. This PW can be modified while the project is running and allows authorised users to add or remove phases, assets, approval configurations and other related data. All these PW adjustments only affect the assigned project and the originating workflow remains unchanged and can be reused as is.

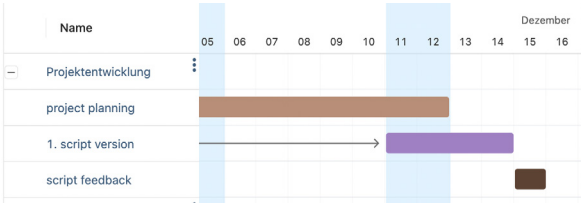


Figure 2: Production plan asset with GANTT style visualisation of tasks, their time slot as well as dependencies.

Lecturers and technical staff can access all supervised projects as well as projects where approvals are waiting to be reviewed. These projects, pending approvals as well as a history of all approval decisions are provided in a clear management view to easily identify open tasks and access projects that require attention. To meet the requirements for flexible substitutions between users (e.g. in case of sickness, planned leave etc.) each person can configure a list of people that are allowed to perform approvals on their behalf. Once configured, the person that is being substituted or the substituting person can activate the substitution. From this point on, all notifications are also send to the substituting user and projects and approvals become accessible.

The unimpeded progression of projects managed by *Tonga* was a major software design goal. The sophisticated substitution system, the fine-grained configuration of approving entities per project as well as special approving permissions for production managers realise this goal.

Connected to the project management component but also usable as a stand alone feature is the equipment rental system. It supports approvals of equipment requests before students and project members can rent them out. Approval entities can be configured based on articles available in the rental system or based on projects that are associated with a request. This on one hand allows e.g. for a department-based approval system where lecturers from the sound department approve sound equipment. On the other hand this also supports a project based approach, where a responsible production supervisor approves equipment that should be rented for this project.

Once approved, equipment rental is managed as known from common rental systems. Users can always access their rental history and display articles waiting for return. They also get notified if the return date approaches. A mechanism to automatically inform a person about overdue equipment with a subsequent blockage from further rental requests is also part of the *Tonga* software.

Finally transparency of responsibilities and traceability are important aspects of *Tonga*. Therefore uploaded documents are always versioned to keep previous versions accessible. Furthermore, required approvals as well as entities who are permissioned to perform these are always visible to project members. Once an approval is made the date, time and approving user are displayed.

These features make *Tonga* a powerful tool for universities to manage film projects of students. This is illustrated by the Filmuniversity Babelsberg KONRAD WOLF as well as the HFF Munich. Both institutions develop *Tonga* in a cooperative effort and are currently in the process of deployment into production usage.

3D VR Sketch-Based 3D Shape Retrieval Demo

Ling Luo
https://rowl1ng.com
SketchX, CVSSP,
University of Surrey

Yulia Gryaditskaya
https://yulia.gryaditskaya.com
SketchX, CVSSP, Surrey Institute for People Centred AI,
University of Surrey

Tao Xiang
http://personal.ee.surrey.ac.uk/Personal/T.Xiang
SketchX, CVSSP,
University of Surrey

Yi-Zhe Song
http://personal.ee.surrey.ac.uk/Personal/Y.Song
SketchX, CVSSP,
University of Surrey

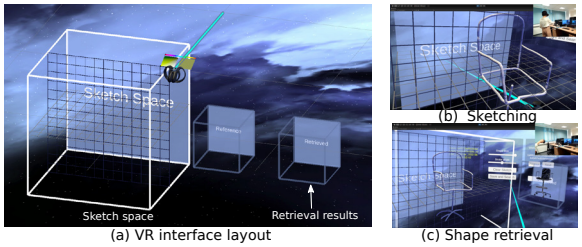


Figure 1: Retrieval demo interface. (a) The interface includes a space for sketching and a space to display the retrieved 3D shapes. We display two orthogonal grids to facilitate sketching in 3D. (b) Users are expected to create a quick 3D sketch using only a set of sparse lines, while wearing a Virtual Reality (VR) headset and operating a VR controller. We refer to such sketches as 3D VR sketches. (c) We feed a complete sketch as a query into our retrieval algorithm and display 3D shapes selected from a collection of 3D shapes, ranked by predicted similarity.

1 Motivation

Growing free online 3D shapes collections propel research on 3D retrieval. Target applications include interior design, the creation of virtual worlds, or online shopping. However, the best choice of query modality remains an open question. Text queries are habitual search queries, but finding the right description quickly becomes a frustrating experience when one is interested in a specific form. We refer to the tasks where one is interested in retrieving a particular instance as *fine-grained*. Fine-grained 3D shape retrieval from a 2D sketch has been shown to be a very challenging task. Therefore, we consider a different query modality: *3D VR sketches* – 3D sketches created while wearing a Virtual Reality (VR) headset and operating a VR controller. Our demo interface is designed so that end users can freely retrieve a 3D model by drawing it in the air in a VR environment. To facilitate wide applicability, we favor the most convenient sketching scenario where the sketch consists of sparse lines, and users are not expected to have any advanced sketching skills, prior training, or to create a time-consuming accurate drawing. This demo demonstrates the first 3D form retrieval system that uses such 3D VR sketches as a query modality and is based on our recent works [1, 2, 3].

2 Setup

There are a number of VR painting and design applications that allow users to sketch and draw in 3D (such as Tilt Brush by Google, Quill by Facebook or Gravity Sketch), however we choose to implement our own interface which is (1) tailored to quick retrieval applications by providing the minimal interface and supporting the most basic sketching, undo and save functions; (2) allows for interactions with our learning-based algorithms.

As a result, users can familiarize themselves with the interface quickly and directly start exploring the retrieval functionality. An example usage-tutorial consists of the following steps: (1) get familiar with hand controllers and the functions of each button and trigger: sketch, undo last stroke, delete all strokes, menu click button; (2) practice grabbing and rotating reference space and sketching space; (3) practice adjusting line width; and (4) practice drawing random lines.

We implemented our custom 3D sketching environment based on Oculus Rift platform and Unity engine. The demo's code is available online: https://github.com/Rowl1ng/SketchyVR-Retrieval.

Figure 2: Example VR sketches and retrieval results obtained with our model proposed in [3]. It can be observed that the top retrieval results exhibit strong structural similarity to the input sketch. The green square highlights the ground-truth 3D shape among the ranked retrieval results.

The demo interface is demonstrated in Figure 1. Once the user completes the sketch and clicks on the search button, the sketch is passed to our retrieval model, which processes the query and performs the retrieval on the server. The most relevant results are then displayed in the VR interface, as shown in Figure 1(c).

3 Retrieval model

We used an interface[1], similar to this demo, to collect the first large-scale 3D VR-sketches. We collected 1,497 sketches for a subset of 1,005 3D chairs in the ShapeNetCore dataset. Our dataset is available online https://cvssp.org/data/VRChairSketch/. We used this dataset to train and test our deep-learning-based models.

During training and inference, both sketches and shapes are first normalized to fit a unit bounding box, and then are converted to point clouds. During training the same encoder is used for sketch and shape point clouds. The training is driven by a triplet loss which considers the ground truth 3D shapes for a given sketch as a positive example, while all other shapes are considered as negative examples. More details can be found in [2, 3].

In [3], the accuracy of the retrieval results is improved by taking structural shape similarity into consideration during training, and proposing a new triplet loss formulation. Our demo implements this model. Some retrieval examples are shown in Figure 2.

4 Conclusion

This demo showcases the first practical fine-grained 3D shape retrieval system that uses a quick and sparse 3D sketch as a query. We hope our work and its demo will inspire further exploration of interactions using 3D sketches in VR scenarios (e.g. sketch-based 3D editing).

[1] Ling Luo, Yulia Gryaditskaya, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Towards 3D VR-Sketch to 3D Shape Retrieval. In *2020 International Conference on 3D Vision (3DV)*, pages 81–90. IEEE, 2020.

[2] Ling Luo, Yulia Gryaditskaya, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Fine-grained vr sketching: Dataset and insights. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021.

[3] Ling Luo, Yulia Gryaditskaya, Tao Xiang, and Yi-Zhe Song. Structure-Aware 3D VR Sketch to 3D Shape Retrieval. In *International Conference on 3D Vision (3DV)*. IEEE, 2022.

[1] https://github.com/Rowl1ng/SketchyVR

Lumirithmic Demo: Desktop-based High-Quality Facial Capture

Yiming Lin[1]
Jayanth Kannan[1]
Ekin Oztur[1 2]
Luca Filipi[1]
Gaurav Chawla[1]
Abhijeet Ghosh[1 2]

[1] Lumirithmic
[2] Imperial College London

(a) Tablet-based Setup (b) Rendering

Figure 1: Our novel desktop-based setup (a) for high-quality facial capture (b). Our setup consists of a set of portable mobile devices (iPads and iPhones) for static facial capture.

Realistically rendered human faces have wide-ranging applications in computer graphics, entertainment, advertising, and virtual presence in AR and VR, and in the envisioned *metaverse*. Realistic modeling of facial shape and appearance has been revolutionized with the development of acquisition techniques for high-quality 3D facial capture. However, realistic facial appearance capture typically requires use of custom designed and complex apparatus such as the Lightstage [1, 2].

We present a novel desktop-based system (see Fig. 1) for high-quality facial capture including geometry and facial appearance (as presented in [3]). The proposed acquisition system is highly practical, consisting purely of commodity components, enabling a significant reduction in cost, along with increased portability and scalability compared to alternative capturing systems. This is a demo at CVMP for on-site facial capture.

1 Desktop-based Capture System

Our setup consists purely of a set of mobile devices – eight tablets and five smartphones, that are mounted on a desk as shown in Fig. 1 (a). The tablets (iPad Air 4th generation) are arranged in two rows (latitudes) of four devices and oriented longitudinally so as to cover a significant zone of the frontal hemisphere around a subject's face. The screens of the tablets are oriented towards the subject and provide controlled piece-wise continuous illumination. We also mount five smartphones (iPhone 12 Pro) in the setup along the equatorial plane and employ their high-resolution back cameras (zoom lens) for acquiring facial reflectance and photometric normals. The devices are all controlled in synchronization during capture process where one device acts as the master and wirelessly communicates capture command and timings to other devices.

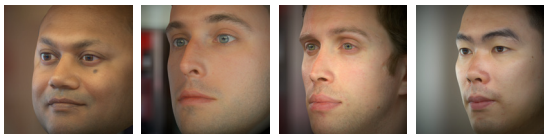
We employ horizontally and vertically aligned binary illumination patterns over the hemispherical zone of illumination of the proposed capture setups for acquiring albedo and photometric normals with diffuse-specular separation. This reduces the number of measurements to only four photographs under form-factor modulated horizontal and vertical binary patterns (see Fig. 2). We further reduce the measurements to just *two photographs* using spectral multiplexing of these patterns into R, G, and B color channels of display illumination. More technical details can be found in [3]

Figure 2: Horizontal and vertical binary patterns and their complements employed with form-factor modulation.

(a) Diff. Alb. (b) Spec. Alb. (c) Diff. Norm. (d) Spec. Norm.

(e) Photo (f) Render. (2D) (g) Geom. (h) Render. (3D)

Figure 3: Various subjects acquired with our tablet-based setup using our 2-shot method.



(a) Render. 1 (b) Render. 2 (c) Render. 3 (d) Render. 4

Figure 4: 3D renderings for different subjects being captured with our novel setup.

2 Results and Conclusion

We deconstruct a full set of our captured results with 2D and 3D renderings in figure 3. Our novel two-shot facial capturing setup coupled with our algorithms yields high-quality BRDF and normal results which generate high-fidelity renderings as shown in figure 4. Our setup is highly practical and scalable, consisting entirely of commodity components, and can make high-quality static facial capture widely accessible.

[1] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM TOG*, 30(6), 2011.

[2] Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. Diffuse-specular separation using binary spherical gradient illumination. In *EGSR (EI&I)*, pages 1–10, 2018.

[3] Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. Practical and scalable desktop-based high-quality facial capture. ECCV, 2022.

NOTES

NOTES

NOTES

CHAIRS

Conference Chairs
Marco Volino, University of Surrey
Rafał Mantiuk, University of Cambridge

Full Papers Chair
Armin Mustafa, University of Surrey

Short Papers & Demos Chair
Yulia Gryaditskaya, University of Surrey

Industry Chair
Valentin Deschaintre, Adobe Research

Sponsorship Chair
Jeff Clifford, Wavecrest

Local Arrangements Chair
Giuseppe Claudio Guarnera, University of York

Public Relations Chair
Peter Vangorp, Utrecht University

Conference Secretary
Emily Ellis, University of York

Programme Committee
Akin Caliskan, University of Surrey
Kevin Matthe Caramancion, University at Albany
Dan Casas, Universidad Rey Juan Carlos
Robert Dawes, BBC Research
Daljit Singh Dhillon, Clemson University
Peter Eisert, Fraunhofer Heinrich Hertz Institute
Zhenhua Feng, University of Surrey
Andrew Gilbert, University of Surrey
Tom Haines, University of Bath
Oliver James, DNEG
Hansung Kim, University of Southampton
George Alex Koulieris, Durham University
Marco Pesavento, University of Surrey
Stanislav Pidhorskyi, Meta Reality Labs Research
Erik Reinhard, InterDigital
Christian Richardt, Meta Reality Labs Research
Zhidong Xiao, Bournemouth University

Steering Committee
Neill Campbell, University of Bath
Jeff Clifford, Wavecrest
John Collomosse, University of Surrey
Abhijeet Ghosh, Imperial College London
Oliver Grau, Intel
Peter Hall, University of Bath
Volker Helzle, Filmakademie
Anil Kokaram, Trinity College Dublin
Will Smith, University of York

Conference Sponsors 2022

ACM ISBN: 978-1-4503-9939-5