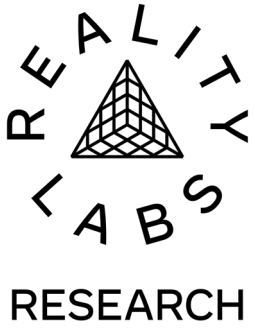
CVMP2024

The 21st ACM SIGGRAPH European Conference on Visual Media Production
18th - 19th November 2024
BFI Southbank, London, UK

Programme

British Film Institute (BFI) Southbank
18th - 19th November 2024
https://www.cvmp-conference.org/2024/

Conference Sponsors 2024



https://dl.acm.org/conference/cvmp

Message from the Chairs

We are pleased to introduce the programme for the twenty first ACM SIGGRAPH European Conference on Visual Media Production (CVMP). For 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 from academia and the creative industries, approximately 50:50.

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 nine full papers and twenty short papers and demos, 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 2024 together!

Armin Mustafa and Hansung Kim (Conference Chairs)
Claudio Guarnera (Full Papers Chair)
Peter Eisert and Peter Vangorp (Short Papers and Demos Chairs)
Oliver Grau and Sara Coppola (Industry Chairs)
Marco Volino (Sponsorship Chair)
Jeff Clifford (Awards Chair)
Changjae Oh (Local Arrangements)
Da Chen (Public Relations Chair)
Violeta Menendez (Social Chair)
Elizabeth James (Conference Secretary)

DAY 1 | Monday 18th November 2024

Location: BFI Southbank

09:00 Registration opens | Blue Room

09:20 Chairs' Welcome | Armin Mustafa, University of Surrey
Hansung Kim, University of Southampton

09:30 SESSION 1 | 3D Capture and Novel View Synthesis

1. High-Quality Facial Geometry from Sparse Heterogeneous Devices under Active Illuminationl
   Lewis Bridgeman, Gilles Rainer, Abhijeet Ghosh
2. Enhanced Illumination Adjustment in 3D Outdoor Reconstructions via Shadow Removal through Color Transfer
   Herbert Potechius, Selvam Essaky, Gunasekaran Raja, Thomas Sikora, Sebastian Knorr
3. RegSegField: Mask-Regularization and Hierarchical Segmentation for Novel View Synthesis from Sparse Inputs
   Kai Gu, Thomas Maugey, Knorr Sebastian, Christine Guillemot

10:30 Coffee Break | Foyer
Poster presenters put up posters

11:00 KEYNOTE 1 | Sarah Ellis, Royal Shakespeare Company

12:00 SPOTLIGHT SESSION

12:20 POSTERS, DEMO AND LUNCH | Blue Room

14:00 SESSION 2 | Image Enhancement and Computational Photography

4. Optimal OLAT Alignment for Image-Based Relighting with Color-Multiplexed OLAT Sequence
   Arvin Lin, Abhijeet Ghosh
5. Image-Based Material Editing Using Perceptual Attributes or Ground-Truth Parameters
   Victor Stenvers, Peter Vangorp
6. Low-light Video Enhancement with Conditional Diffusion Models and Wavelet Interscale Attentions
   Ruirui Lin, Qi Sun, Nantheera Anantrasirichai

15:00 CVMP AWARDS | Jeff Clifford

15:30 Coffee Break | Blue Room

16:00 KEYNOTE 2 | Ana Serrano, Universidad de Zaragoza

17:00 NETWORKING RECEPTION | Blue Room

19:00 Close

DAY 2 | Tuesday, 19th November 2024

Location: BFI Southbank

09:00 Registration opens | Blue Room

09:30 SESSION 3 | Human Motion, Privacy and Usability

7. Multi-Resolution Generative Modeling of Human Motion from Limited Data
   David E Moreno-Villamarin, Anna Hilsmann, Peter Eisert
8. PDFed: Privacy-Preserving and Decentralized Asynchronous Federated Learning for Diffusion Models
   Kar Balan, Andrew Gilbert, John Collomosse
9. Interacting from Afar: A Study of the Relationship Between Usability and Presence in Object Selection at a Distance
   Kalila Shapiro, Pedro Quijada Leyton, Lloyd Stemple, David Uberti

10:30 Coffee Break | Blue Room

11:00 KEYNOTE 3 | Shunsuke Saito, Reality Labs Research, Meta

12:00 SPOTLIGHT SESSION

12:15 POSTERS, DEMO AND LUNCH | Blue Room

13:45 INDUSTRY SESSION | Generative AI in Media Production

10. Building Safe and Fair Generative AI with Content Provenance
    John Collomose, Adobe/University of Surrey
11. AI In VFX: What works and what doesn't (yet)
    Graham Jack, BeloFX
12. Modern 3D scene understanding with Project Aria Glasses. A step closer towards always-on sensing with AI wearables
    Armen Avetisyan, Meta
13. Generative AI and Public Service Media at the BBC
    Graham Thomas, BBC R&D

15:25 Coffee Break | Blue Room

15:50 KEYNOTE 4 | Aljosa Smolic, Lucerne University of Applied Sciences and Arts

16:50 Prizes, Announcements and Closing | Armin Mustafa, University of Surrey
Hansung Kim, University of Southampton

KEYNOTE 1 | Sarah Ellis
Royal Shakespeare Company

Monday 18th November 2024

Sarah Ellis

Sarah Ellis is an award-winning producer currently working as Director of Digital Development for the Royal Shakespeare Company to explore new artistic initiatives and partnerships. The latest partnership for the RSC is the Audience of the Future Live Performance Demonstrator funded by Innovate UK - a consortium consisting of arts organisations, research partners and technology companies to explore the future of performances and real-time immersive experiences.As a spoken word producer, she has worked with the Old Vic Tunnels, Battersea Arts Centre, Birmingham REP, Contact, Improbable, Southbank Centre, Soho Theatre, and Shunt. She has been Head of Creative Programmes at the Albany Theatre and Programme Manager for Apples & Snakes.She is a regular speaker and commentator on digital arts practice, as well as an Industry Champion for the Creative Industries Policy and Evidence Centre, which helps inform academic research on the creative industries to lead to better policies for the sector. She has been appointed Chair of digital agency, The Space, established by Arts Council England and the BBC to help promote digital engagement across the arts.

KEYNOTE 2 | Ana Serrano
Universidad de Zaragoza

Understanding user behavior and attention in immersive environments
Monday 18th November 2024

Virtual reality (VR) is an exciting and rapidly growing medium that presents both challenges and opportunities. As VR techniques and applications continue to blossom, creating engaging experiences that exploit their potential becomes increasingly important. Understanding and being able to reliably predict human visual behavior is an essential factor in achieving this goal. This knowledge can be the key to designing more engaging storytelling experiences and developing efficient content-aware compression and rendering techniques that take into account users' attentional patterns and behavior. In this talk, we will explore approaches and challenges involved in modeling visual attention and gaze behavior in immersive 360º environments. By studying how users allocate their attention and direct their gaze, we can uncover valuable insights for creating immersive VR experiences.

Ana Serrano

Ana Serrano is an Associate Professor at Universidad de Zaragoza (Spain). Previously she was a Postdoctoral Research Fellow at the Max-Planck-Institute for Informatics. She received her PhD in Computer Science in 2019. Her doctoral thesis was recognized with one of the Eurographics 2020 PhD awards and she was recognized with the Eurographics Young Researcher Award in 2023 and the VGTC Significant New Researcher Award in 2024. Her research focuses on various areas of visual computing, including computational imaging, material appearance perception and editing, and virtual reality. She is particularly interested in using perceptually-driven approaches to improve user experiences and develop tools to assist content creation.

KEYNOTE 3 | Shunsuke Saito
Reality Labs Research, Meta

Foundations for 3D Digital Humans
Tuesday 19th November 2024

What constitutes the foundation for 3D digital hand avatars? In this talk, we aim to establish the essential components necessary for creating high-fidelity digital human models. We argue that relighting, animation/interaction, and in-the-wild generalization are crucial for bringing high-quality avatars to everyone. We will discuss several relightable appearance representations that achieve a photorealistic appearance under various lighting conditions. Furthermore, we will introduce techniques to effectively model animation and interaction priors. Finally, the talk will cover bridging the domain gap between high-quality studio data and large-scale in-the-wild data via a human-centric foundational model called Sapiens, which is key to enhancing robustness and diversity in avatar modeling algorithms. We will also explore how these foundations can complement and enhance each other.

Shunsuke Saito

Shunsuke Saito is a Research Scientist at Meta Reality Labs Research in Pittsburgh, where he leads the effort on next generation digital humans. He obtained his PhD degree at the University of Southern California. Prior to USC, he was a Visiting Researcher at University of Pennsylvania in 2014. He obtained his BE (2013), ME (2014) in Applied Physics at Waseda University. His research lies in the intersection of computer graphics, computer vision and machine learning, especially centered around digital human, 3D reconstruction, and performance capture. His work has been published in SIGGRAPH, SIGGRAPH Asia, NeurIPS, ECCV, ICCV and CVPR, three of which have been nominated for CVPR Best Paper Award (2019, 2021) and ECCV Best Paper Award (2024). His real-time volumetric teleportation work also won Best in Show award in SIGGRAPH 2020 Real-time Live!

KEYNOTE 4 | Aljosa Smolic
Lucerne University of Applied Sciences and Arts

AI-based Volumetric Content Creation for Immersive XR Experiences and Production Workflows

Tuesday 19th November 2024

Capture and 3D reconstruction of real-world objects, scenes, environments and people for creation of digital assets is an important task in many media productions. Classical methods of visual computing are known for instance as photogrammetry for static content or volumetric video for dynamic content. Recently, AI-based methods like Neural Radiance Fields (NeRF) and Gaussian Splatting disrupted the scientific field and created a lot of interest in the media production industry. However, integration of such content with standard production workflows is not straight-forward, due to the specific nature of the data. This talk will highlight examples of AI-based volumetric content creation and their application in media production workflows, as developed in different projects at HSLU

Aljosa Smolic

Aljosa Smolic is Professor in the Computer Science Department of the Lucerne University of Applied Sciences and Arts in Switzerland and Co-Head of the Immersive Realities Research Lab. Before he was Professor of Creative Technologies at Trinity College Dublin heading the research group V-SENSE, Senior Research Scientist and Group Leader at Disney Research Zurich, and Scientific Project Manager and Group Leader at Fraunhofer HHI. He is also a Co-Founder of the company Volograms, which commercializes volumetric video technology. Prof. Smolic's expertise is in the broad area of visual computing (covering image/video processing, computer vision, computer graphics) with a focus on immersive XR technologies. He published 250+ scientific papers and book chapters, holds 35+ patents and received several awards and recognitions for his research.

INDUSTRY TALKS

Building Safe and Fair Generative AI with Content Provenance

John Collomose, Adobe/University of Surrey

Provenance facts, such as who made an image and how, can provide valuable context for users to make trust decisions. In this talk I will contrast three provenance enhancing technologies: metadata, fingerprinting and watermarking, and discuss how their complementary strengths may be combined to provide robust trust signals to support stories told by real and generative images. I will outline Content Credentials (C2PA), an emerging standard for provenance metadata and how its rapid adoption promises to play an important role in fighting fake news and the spread of misinformation. Beyond authenticity, I will describe how provenance can also underpin new models for value creation in the age of Generative AI. In doing so I address risks arising with generative AI such as ensuring training consent, and the proper attribution of credit to creatives who contribute their work to train generative models. I show that provenance may be combined with distributed ledger technology (DLT) to develop novel solutions for recognizing and rewarding creative endeavour in the age of generative AI.

AI In VFX: What works and what doesn't (yet)

Graham Jack, BeloFX

This talk will dig into the practical realities of using AI based techniques in the production of visual effects for film and television. It will explore some areas where AI is already bringing huge productivity gains as well as looking at areas where there are still significant challenges to its use. The talk will discuss technical challenges such as resolution limitations and lack of temporal coherence; creative challenges including specificity, being able to work iteratively and the need to fit into established creative workflows and will touch on some difficult topics such as the legal and ethical implications and the effect that an increased use of AI could have on the artists working in our industry.

Modern 3D scene understanding with Project Aria Glasses. A step closer towards always-on sensing with AI wearables

Armen Avetisyan, Meta

Egocentric, multi-modal data on future augmented reality devices offers both challenges and opportunities for machine perception. These devices must be wearable all day, with a socially acceptable form factor, and operate on low-power compute resources. Project Aria serves as a prototype for this class of hardware and as a research platform for egocentric vision. Our research centers on developing models to construct spatial scene representations and object reconstructions from the sensing capabilities provided by these glasses. We present models that address the device's constraints and compute limitations, demonstrating impressive capabilities in real-world scenarios.

Generative AI and Public Service Media at the BBC

Graham Thomas, BBC R&D

The BBC is treading very carefully when considering applications of GenAI. Particular concerns include maintaining trust with audiences (not using GenAI to create something that could be interpreted as real, where this would be misleading, or that may contain errors), copyright protection (infringing the rights of other content producers), or generating content that may have inherent bias (e.g. summarisation of a programme that does not maintain editorial balance). Nevertheless, it is exploring this technology, both through evaluating commercial solutions and through developing its own. This talk will dig into the minefield where Public Service Media meets GenAI, and look at some of the areas that may be of particular interest to the CVMP audience, including content summarisation, free-viewpoint rendering, and combatting the rise of deep fake images through detection strategies and content provenance.

High-Quality Facial Geometry from Sparse Heterogeneous

Lewis Bridgeman (Lumirithmic Ltd), Gilles Rainer (Imperial College London), Abhijeet Ghosh (Lumirithmic Ltd, Imperial College London)

High-resolution facial geometry is essential for realistic digital avatars. Traditional reconstruction methods, such as multi-view stereo, often struggle with materials like skin, which exhibit complex light reflection, absorption, and scattering properties. Neural reconstruction methods have shown greater robustness to these view-dependent effects. However, positional-encoding-based implementations are typically slow, while faster hash-encoded methods may falter under sparse camera views. We present a geometry reconstruction method tailored for an active-illumination facial capture setup featuring sparse cameras with varying characteristics. Our technique builds upon hash-encoded neural surface reconstruction, which we enhance with additional active-illumination-based supervision and loss functions, allowing us to maintain high reconstruction speed and geometrical fidelity even with reduced camera coverage. We validate our approach through qualitative evaluations across diverse subjects, and quantitative evaluation using a synthetic dataset rendered with a virtual reproduction of our capture setup. Our results demonstrate that our method significantly outperforms previous neural reconstruction techniques on datasets with sparse camera configurations.

Enhanced Illumination Adjustment in 3D Outdoor Reconstructions via Shadow Removal through Color Transfer

Herbert Potechius (University of Applied Sciences), Selvam Essaky (Anna University), Gunasekaran Raja (Anna University), Thomas Sikora (Technical University of Berlin), Sebastian Knorr (University of Applied Sciences)

The introduction of 3D reconstruction technology has revolutionized the digitization of real-world objects, from small artifacts to large-scale structures like buildings. This technology enables the rapid creation of virtual representations of the real world with minimal design expertise, offering a novel way to experience reality. However, it presents challenges such as baked-in illumination, which complicates subsequent relighting and integration into digital environments. This paper introduces a shadow removal algorithm, SRCT, that uses simulated lighting and color transfer techniques to reduce the visible effects of self- and cast shadows in the texture maps of 3D models resulting from the 3D reconstruction process. The effectiveness of this approach is validated through a comparison with existing shadow removal techniques. This validation utilizes a newly introduced dataset, EDEN, which comprises 3D reconstructions of buildings derived from drone imagery for qualitative evaluation, along with an additional dataset, Sunlit3D, featuring 3D reconstructions of buildings under various simulated lighting conditions forquantitative analysis.

RegSegField: Mask-Regularization and Hierarchical Segmentation for Novel View Synthesis from Sparse Inputs

Kai Gu (Rennes University and Technische Universität Berlin), Thomas Maugey (Rennes University), Knorr Sebastian (Hochschule für Technik und Wirtschaft Berlin), Christine Guillemot (Rennes University)

Radiance Field (RF) representations and their latest variant, 3DGaussian Splatting (3D-GS), have revolutionized the field of 3D vision. Novel View Synthesis (NVS) from RF typically re-

quires dense inputs, and for 3D-GS in particular, a high-quality point cloud from a multi-view stereo model is usually necessary. Sparse input RFs are commonly regularized by various priors, such as smoothness, depth, and appearance. Meanwhile, 3D scene segmentation has also achieved significant results with the aid of RFs, and combining the field with different semantic and physical attributes has become a trend. To further tackle NVS and 3D segmentation problems under sparse-input conditions, we introduce RegSegField, a novel pipeline to utilize 2D segmentations to aid the reconstruction of objects and parts. This method introduces a novel mask-visibility loss by matching 2D segments across different views, thus defining the 3D regions for different objects. To further optimize the correspondence of 2D segments, we introduce a hierarchical feature field supervised by a contrastive learning method, allowing iterative updates of matched mask areas. To resolve the inconsistent segmentation across different views and refine the mask matching with the help of RF geometry, we also employed a multi-level hierarchy loss. With the help of the hierarchy loss, our method facilitates scene segmentation at discrete granularity levels, whereas other methods require sampling at different scales or determining similarity thresholds. Our experiments show that our regularization approach outperforms various depth-guided NeRF methods and even enables sparse reconstruction of 3D-GS with random initialization.

Optimal OLAT Alignment for Image-Based Relighting with Color-Multiplexed OLAT Sequence

Arvin Lin, Abhijeet Ghosh (Imperial College London)

We present two color-multiplexed illumination sequences for optimallyaligned one-light-at-a-time (OLAT) captures. We leverage colormultiplexing strategies to embed tracking frames within the OLAT photographs to correct for subject motion. Our method allows better motion estimation via optical flow than traditional methods, which interleaves tracking frames between OLATs. Comparison between rendered results and user study on comfortability both demonstrate that color-multiplexed sequences give better-aligned OLATs and are more comfortable for the subject during data capture. Our proposed sequences can replace traditional OLAT sequences for better data acquisition, which would benefit both light-stage rendering results and any state-of-the-art relighting methods that are trained on OLAT-generated data.

Image-Based Material Editing Using Perceptual Attributes or Ground-Truth Parameters

Victor Stenvers, Peter Vangorp (Utrecht University)

Image-based material editing neural networks have been trained on perceptual attributes because such attributes are human-friendly. But it seems that training such networks on non-perceptual material parameters has been neglected in comparison. It is interesting that collecting perceptual experiment data has been considered an acceptable additional effort until now. It would be much easier to generate a dataset with ground-truth material parameter attributes instead. Ground-truth parameters also avoid the noise that is inherent in perceptual experiment data. We show that existing neural networks can be trained on datasets with ground-truth material parameters and that they generate material edits of similar quality and that stay as close to the valid gamut of the trained material model as neural networks trained on perceptual material attributes. We expect that these results will encourage more study of the qualitative and quantitative differences between image-based material editing networks trained on material parameters and on perceptual attributes.

Low-light Video Enhancement with Conditional Diffusion Models and Wavelet Interscale Attentions

Ruirui Lin, Qi Sun, Nantheera Anantrasirichai (University of Bristol)

Videos captured in low-light conditions often su er from various distortions, such as noise, low contrast, color imbalance, and blur. Consequently, a post-processing work ow is necessary but typically time-consuming. Developing AI-based tools for videos also requires significantly more computational resources compared to those for images. This paper introduces a novel framework aimed at reducing memory usage and computational time by enhancing videos in the wavelet domain. The framework utilizes conditional diffusion models to enhance brightness and adjust colors in the lowpass subbands while employing interscale-attention mechanisms to enhance sharpness in the high-pass subbands. To ensure temporal consistency, we integrate feature alignment and fusion into the denoiser of the diffusion models. Additionally, we introduce adaptive brightness adjustment as a preprocessing module to reduce the workload of the learnable networks. Experimental results demonstrate that our proposed methods outperform existing low-light video enhancement techniques with competitive inference times compared to image-based methods.

Multi-Resolution Generative Modeling of Human Motion from Limited Data

David E Moreno-Villamarin, Anna Hilsmann, Peter Eisert (Fraunhofer Heinrich Herz Institute)

We present a generative model that learns to synthesize human motion from limited training sequences. Our framework provides conditional generation and blending across multiple temporal resolutions. The model adeptly captures human motion patterns by integrating skeletal convolution layers and a multi-scale architecture. Our model contains a set of generative and adversarial networks, along with embedding modules, each tailored for generating motions at specific frame rates while exerting control over their content and details. Notably, our approach also extends to the synthesis of co-speech gestures, demonstrating its ability to generate synchronized gestures from speech inputs, even with limited paired data. Through direct synthesis of SMPL pose parameters, our approach avoids test-time adjustments to fit human body meshes. Experimental results showcase our model's ability to achieve extensive coverage of training examples, while generating diverse motions, as indicated by local and global diversity metrics.

PDFed: Privacy-Preserving and Decentralized Asynchronous Federated Learning for Diffusion Models

Kar Balan, Andrew Gilbert (University of Surrey), John Collomosse (Adobe Research)

We present PDFed, a decentralized, aggregator-free, and asynchronous federated learning protocol for training image diffusion models using a public blockchain. In general, diffusion models are prone to memorization of training data, raising privacy and ethical concerns (e.g., regurgitation of private training data in generated images). Federated learning (FL) offers a partial solution via collaborative model training across distributed nodes that safeguard local data privacy. PDFed proposes a novel sample-based score that measures the novelty and quality of generated samples, incorporating these into a blockchain-based federated

learning protocol that we show reduces private data memorization in the collaboratively trained model. In addition, PDFed enables asynchronous collaboration among participants with varying hardware capabilities, facilitating broader participation. The protocol records the provenance of AI models, improving transparency and auditability, while also considering automated incentive and reward mechanisms for participants. PDFed aims to empower artists and creators by protecting the privacy of creative works and enabling decentralized, peer-to-peer collaboration. The protocol positively impacts the creative economy by opening up novel revenue streams and fostering innovative ways for artists to benefit from their contributions to the AI space.

Interacting from Afar: A Study of the Relationship Between Usability and Presence in Object Selection at a Distance

Kalila Shapiro (University College London), Pedro Quijada Leyton, Lloyd Stemple, David Uberti (Sony Interactive Entertainment)

Virtual Reality (VR) permits people to bend the rules of physics that exist in the real world, allowing for unique "magic" interactions. For example, if a user wants to interact with a far away object, they can summon it to themselves in the Virtual Environment (VE) as opposed to physically walking to it like they would need to outside of a headset. There is no set framework for how to facilitate this type of user interaction for the optimal user experience. We wanted to understand how people intuitively interact with objects at a distance, hypothesizing that the more usable someone found a system, the higher level of presence they would experience. We present a study that explores the relationship between presence and usability through questionnaires. We identify a positive correlation between presence and usability and propose a system that allows free-movement to encourage presence in VEs.

Full papers available from the ACM Digital Library:
https://dl.acm.org/conference/cvmp

TRACER: Modular Open-Source Framework for Real-Time XR Collaboration

Jonas Trottnow
https://animationsinstitut.de/en/studies/lecturers/jonas-trottnow
Simon Spielmann
https://animationsinstitut.de/en/simon-spielmann
Francesco Andreussi
https://www.linkedin.com/in/francesco-andreussi
Simon Haag
https://www.linkedin.com/in/simonhaag2094

Research & Development department at
Animationsinstitut of Filmakademie Baden-Württemberg

Data management in distributed XR productions has become a major topic ever since more and more digital elements were introduced into traditional workflows for filmmaking, experiences and presentations. Virtual productions, VR multi-player games and AR applications demand for open, software agnostic and future proof pipelines.

We introduce **TRACER** (Toolset for Realtime Animation, Collaboration & Extended Reality), a software agnostic communication infrastructure and toolset for plugging open-source tools into a production pipeline, establishing interoperability between open source and proprietary tools, targeting real-time collaboration and XR productions, with an operational layer for exchanging data objects and updates including animation and scene data, synchronization of scene updates of different client applications (Blender, UE, Unity ...), parameter harmonization between different engines/renderers, unified scene distribution and scene export which stores the current state of the scene.

TRACER foundation, a modular, open-source C# framework handling

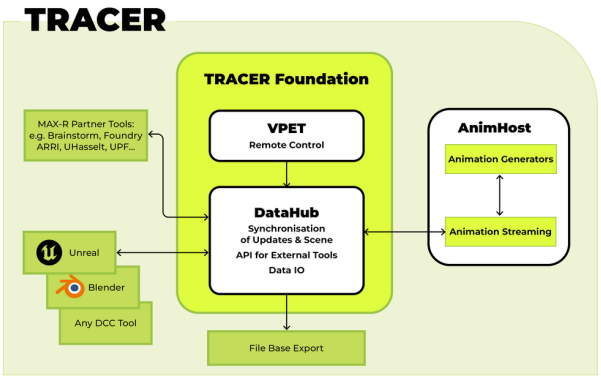


Figure 1: Overview of TRACER

scene operations, input methods, UI, animation and communication between various clients, is the core of the toolset. It is easily extensible by adding new modules. Compact and efficient protocols unify scene transfer and real-time updates across different DCC applications and game engines like Blender, Unity, Unreal Engine etc. All networking is based on netMQ / zeroMQ.

Along with the core and a basic set of modules a multitude of example implementations, integrations and extensions has been developed.

VPET [2], the Virtual Production Editing Tool is a tablet based app to visualize 3D scenes in AR and edit various aspects like object position, rotation, scale, light parameters, cameras and now even animations with an intuitive interface. Based on an initial implementation of VPET, the fundamental functionalities were made available for more general XR applications resulting in the TRACER foundation, through a re-factoring and reimplementation. VPET 2.0 now serves as a sample implementation in the TRACER foundation framework.

Along with VPET being a remote control for 3D scenes and beyond, a number of native scene providing plugins has been developed. Plugins for Blender, Unreal Engine, Unity 3D plugin for TRACER making it possible to distribute an authored scene directly to other TRACER clients like VPET, while changes to the scene are shared with all clients synchronously.

DataHub, another central component of the TRACER ecosystem. Being the connecting link between all instances, the DataHub has the ability to distribute and store incoming data and therefore act as a scene server and event recorder providing this data for potential post production steps. This makes it possible to store and load complete scenes, as well as their changes over time. Changes in the scene get redistributed to all connected clients.

With technologies like marker-less, video-based motion capturing and AI-generated character animation, pipelines for animated movies are transforming. Movie Productions utilising game engines for rendering demand for interactive and real-time animation directing capabilities that can be driven by artists and directors. While AI-based human character animation generators exist in research, their applied usage in animated movie production is sparse. The interactive nature combined with the demand for artist controllability asks for new user interfaces and pipelines.

AnimHost connects animation generators (such as AI deep neural networks trained on motion capturing databases, video-based, low-cost motion capturing and many more) to DCC applications, on-set tools like VPET or renderers in general. It is functionally independent of the animation receiving app and provides an intuitive interface to support new solvers, with a focus on real-time scenarios. Animations can be directed either in Blender or VPET for now. A simple walking path in form of a spline enriched with additional information like look-at rotations can be defined, based on which AnimHost generates a full-body bone animation for the given character. Just like DataHub, AnimHost is implemented as a Desktop application within the Qt framework. To evaluate this development the 3d short film production "Survivor"[1] has been initiated.

Examples include virtual productions, the origin of VPET and the concepts behind TRACER, while more recent implementations provide a variety of applications. For example, "Faith of the Minotaur" [1], a VR multi-player game, utilizes TRACER incl. DataHub as backend realizing collaborative player interactions with the game's world. Another application are "Digital Locations". Location Scouts can explore potential shooting locations digitally in VR/AR and in the browser, plan logistics or even use the digital representation in productions. TRACER provides a simple interface to load arbitrary scenes into VPET or other TRACER based applications through scanning a simple QR code[2].

The presented TRACER framework provides a comprehensive set of tools to address multiple challenges in modern XR production pipelines. The open-source nature and modularity makes it a valuable companion in research and production. All components can be obtained from our GitHub repositories[3].

TRACER is developed within the Max-R project which received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101070072.

[1] Andreas Dahn, Leszek Plichta, Simon Spielmann, Eduard Schäfer, and Justus Blönnigen. Fate of the minotaur: A scalable location based vr experience. In *ACM SIGGRAPH 2024 Immersive Pavilion*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705274. doi: 10.1145/3641521.3664406. URL https://doi.org/10.1145/3641521.3664406.

[2] Simon Spielmann, Volker Helzle, Andreas Schuster, Jonas Trottnow, Kai Götz, and Patricia Rohr. Vpet: virtual production editing tools. In *ACM SIGGRAPH 2018 Emerging Technologies*, SIGGRAPH '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358101. doi: 10.1145/3214907.3233760. URL https://doi.org/10.1145/3214907.3233760.

[1] https://research.animationsinstitut.de/survivor
[2] https://research.animationsinstitut.de/digitallocations
[3] https://github.com/FilmakademieRnd

Detection and Re-Identification in the case of Horse Racing

Will Binning
https://www.linkedin.com/in/will-binning-66a9b7221/
Sadegh Rahmani
https://www.linkedin.com/in/sadegh-rahmani/
Xu Dong
www.linkedin.com/in/xudong-442302166
Andrew Gilbert
https://andrewjohngilbert.github.io/

C-CATS,
Centre of Creative Arts and Technology
University of Surrey
UK

Despite its popularity and the substantial betting it attracts, horse racing has seen limited research in machine learning. Some studies have tackled related challenges, such as adapting multi-object tracking to the unique geometry of horse tracks [3] and tracking jockey caps during complex manoeuvres [2]. Building on this work, our research aims to create a helmet detector framework as a preliminary step for re-identification using a limited dataset. Specifically, we detected jockeys' helmets throughout a 205 seconds race with six disjointed outdoor cameras, addressing challenges like occlusion and varying illumination. Occlusion is a significant challenge in horse racing, often more pronounced than in other sports. Jockeys race in close groups, causing substantial overlap between jockeys and horses in the camera's view, complicating detection and segmentation. Additionally, motion blur, especially in the final stretch of the race, and the multi-camera broadcast capturing various angles—front, back, and sides—further complicate detection and consecutively re-identification (Re-ID). To address these issues, we focus on helmet identification rather than detecting the entire horses or jockeys. We believe helmets, with their simple shapes and consistent appearance even when rotated, offer a more reliable target for detection to make the Re-ID downstream task more achievable.

The Architecture: A summary of the architecture is presented in Figure 1. The system focuses on helmet detection and classification. A multi-class Convolutional Neural Network (ConvNet) based on ResNet-18 is trained on labelled helmet classes using a cross-entropy loss function. For helmet detection and segmentation, we employ Grounded-SAM [4] with the prompt "helmet" to accurately detect and segment jockeys' helmets, even capturing partially visible helmets in cases of mild occlusion. These segmented helmets are then cropped from the images to create our dataset. Based on the ResNet-18 architecture, the classification model processes these cropped helmet images and is trained using cross-entropy loss and the Adam optimizer. After training, we apply a confidence threshold to filter out false positives. Finally, we use the detected helmets to annotate the original images with colour-coded bounding boxes corresponding to each class.

Dataset: The racing broadcast company Racetech [1] provided the data and industry context that enabled the research presented in this paper. The dataset comprises a single outdoor competitive horse race, captured by six moving cameras, with a total duration of 205 seconds and 1026 frames sampled at 5 FPS. The race features 12 jockeys, and we developed a proof-of-concept labelled dataset focusing on 5 jockeys (or classes) across the six cameras. For semi-automated ground truth labelling, the cropped helmet images were sorted by primary colour, followed by a manual review to remove mis-classifications. The model was trained on 80% of the samples from camera 1, and tested on the remaining samples from camera 1, as well as on unseen data from cameras 2-6.

Results: The accuracy of the helmet detector across the 6 cameras is illustrated in Figure 2, highlighting the model's overall performance. Helmets with simple, solid-coloured designs, such as classes 1 and 2, achieved

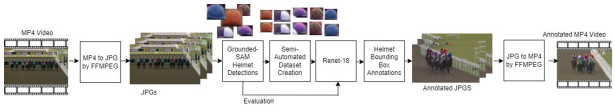


Figure 1: The model segments helmets from RGB images and utilizes a multi-class classifier to train the helmet detector. The final output is an annotated MP4 video, where helmet bounding boxes are colour-coded to represent the five distinct classes.

Figure 2: Confusion matrix for individual classes on each angle. Values are percentages of total frames that the class has been successfully annotated. After running through the model, these values were manually verified to remove false positives but do not account for frames where the class is not present.

higher accuracy. However, helmets with intricate designs, like those in classes 4 and 5, faced challenges, particularly in wide shots where lower resolution impacted detection. Additionally, cameras positioned around turns (such as cameras 4 and 5) were less successful due to issues with occlusion and blurring, which adversely affected both ResNet feature extraction and the initial helmet detection. However, there is a good performance in the classification of helmets despite only using examples from a single camera view to train the model. The camera view 1 helmet examples used to train this model ranged from 70 to 160 images per class, which, in conjunction with the single-angle training set and fairly simple model, shows that there is a prospect of success in this technique. For instance, the most successful class (class 1) had a mean 79.1% successful detection and classification during the race.

Conclusions: We initially selected five jockeys with distinct helmets to simplify dataset creation. However, expanding the dataset revealed challenges in manually classifying similar helmets. Future research will combine distinct helmet detections with jockey or horse tracking to improve classification, acknowledging that not all helmets in a race will be unique. Further exploration of the data requirements for effective helmet detection is also recommended.

[1] Racetech. `https://www.racetech.co.uk/`. Accessed: 2024-08-27.

[2] Mohammad Hedayati, Michael J Cree, and Jonathan B Scott. Tracking jockeys in a cluttered environment with group dynamics. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 67–73, 2019.

[3] Wing W.Y. Ng, Xuyu Liu, Xuli Yan, Xing Tian, Cankun Zhong, and Sam Kwong. Multi-object tracking for horse racing. *Information Sciences*, 638:118967, 2023.

[4] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

ATRA: An Adaptive Tick Rate Algorithm for Efficient Collaboration in Virtual Reality

William Naylor
W.Naylor@warwick.ac.uk
Alan Chalmers
Alan.Chalmers@warwick.ac.uk
Kurt Debattista
K.Debattista@warwick.ac.uk

Visualisation Group,
Warwick Manufacturing Group,
University of Warwick

Within remote collaborative applications, data must be sent periodically over a network to synchronise the states for each client in the virtual environment. The tick rate defines the frequency at which this is sent. At present, networked applications have only employed constant tick rates to provide game state updates to all clients [3], however this rigidity can introduce inefficiencies where the tick rate may be unnecessarily high in slow-paced scenarios, wasting network throughput for negligible perceptual benefit, or too low in high-intensity sequences, thus introducing visible jitter artefacts. This work seeks to present a dynamic tick rate framework to maximise quality of experience in collaborative virtual environments whilst minimising network throughput.

Intelligent networking algorithms present the possibility to reduce network traffic and computational demand [6], however, up to this point these algorithms have primarily focused on the network layer for optimal path selection to mitigate latency [2] and packet loss [1]. An application-led adaptive tick rate algorithm would enable content analysis techniques to dynamically adjust the tick rate to ensure the smallest possible load on the network whilst maintaining a consistently high quality of experience for all connected clients.

We present an Adaptive Tick Rate Algorithm (ATRA), which employs models of tick rate perception to analyse the content within a scene and the viewing media to accurately predict perceptually acceptable tick rates. The perceptual models utilised by ATRA were developed from our previous research surrounding tick rate psychometric functions [4] and just noticeable differences [5].

Algorithm

ATRA is a routine that runs periodically on the central server within a client-server architecture as it requires complete knowledge of the world state and it can operate for Virtual Reality clients as well as desktop clients. It is broken down into four stages as follows: Scene Velocity Analysis, Psychometric Function, Just Noticeable Difference, Rapid Change Filter.

The first stage, Scene Velocity Analysis, iterates over all clients and the set of all networked objects within the view frustum of that client to determine the maximum angular velocity visible to any client and the maximum result between all clients is passed to the next stage.

From there, Psychometric Function utilises the information regarding angular velocity and display format to calculate a tick rate that has a high probability of being perceived as acceptably smooth. The psychometric function for tick rates observes that higher tick rates are required for perceptual smoothness in VR than on a desktop and similarly, higher tick rates are required for high object velocities.

The Just Noticeable Difference phase then limits the tick rate change from the previous iteration to ensure the difference will be imperceptible. A strong interaction exists between the frame rate and the tick rate, therefore just noticeable differences should be calculated from the Frames per Update ratio, $FPU = frame\ rate / tick\ rate$.

Finally, the Rapid Change Filter analyses when the last major change in the tick rate was made to ensure multiple JND-scale drops in the tick rate do not occur within a short time span, as the multiple incremental changes could otherwise be cumulatively larger than a single JND and this would cause the change to become noticeable, which would negatively affect immersion and quality of experience in the program.

ATRA may run every tick to incentivise small continuous corrections to the tick rate each cycle and this may also reduce the number of large JND-scale changes presented to the observer. It does not necessarily need to be run every single update and may instead be run every second or two, as this would reduce the number of times the tick rate needs to be adjusted, however this would increase the chance of large changes to the tick rate each time the algorithm is run as the world-state would have changed more drastically between cycles.

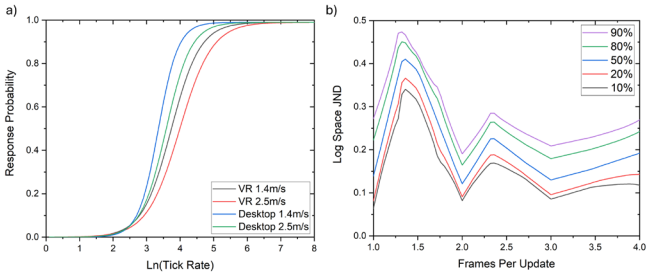


Figure 1: The perceptual models guiding the adaptive algorithm. a) Probability for a given tick rate to be perceived as smooth. b) Just noticeable difference thresholds for different detection probabilities

Evaluation

In order to test the implications of a dynamic tick rate, a collaborative Virtual Reality experiment was performed where participants were tasked to cooperate under different tick rate models, including ATRA. The experiment tasked participants to collaborate by driving heavy machinery to transport rubble around a virtual building site and the metrics of task performance, perceived quality and cybersickness were used to evaluate the impact of the tick rates.

Preliminary results suggest that ATRA is perceptually identical to higher tick rate models, such as a constant 90 Hz, whereas lower tick rates such as 22.5 Hz were rated significantly poorer. This is further supported by quality of experience questionnaires reporting 22.5 Hz to be more jittery than higher tick rates or ATRA. Cybersickness symptoms were also observed to increase over the course of the experiment, however tick rates did not affect the severity of symptoms. No significant effect on task performance was found.

This work demonstrates the viability for an adaptive tick rate algorithm to dynamically adjust the tick rate of a network, as ATRA may be employed to significantly reduce the network throughput compared to a constant tick rate model without jeopardising the perceptual quality of the networked virtual environment.

[1] B Chang. Analysis of adaptive cost functions for dynamic update policies for qos routing in hierarchical networks. *Information Sciences*, 151:1–26, 5 2003. doi: 10.1016/S0020-0255(02)00274-8.

[2] Niels Christensen, Mark Glavind, Stefan Schmid, and Jiří Srba. Latte: Improving the Latency of Transiently Consistent Network Update Schedules. *ACM SIGMETRICS Performance Evaluation Review*, 48 (3):14–26, 3 2021. doi: 10.1145/3453953.3453957.

[3] Steven W. K. Lee and Rocky K. C. Chang. Evaluation of lag-related configurations in first-person shooter games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, volume 1, pages 1–3. IEEE, 12 2015. doi: 10.1109/NetGames.2015.7382997.

[4] William Naylor, Kurt Debattista, and Alan Chalmers. Perception-based high quality distributed virtual reality. *Virtual Reality*, 27(3): 2529–2539, 9 2023. doi: 10.1007/s10055-023-00825-9.

[5] William Naylor, Alan Chalmers, and Kurt Debattista. Tick Rate - Frame Rate Cross Interaction And The Impact Upon Just Noticeable Differences In Virtual Reality. *Submitted to a journal, Awaiting reviewer feedback*, 2024.

[6] Leonisio Schepis *et al.* Adaptive Data Update for Cloud-based Internet of Things applications. In *Proceedings of the ACM MobiHoc Workshop on Pervasive Systems in the IoT Era*, pages 13–18, New York, NY, USA, 7 2019. ACM. doi: 10.1145/3331052.3332472.

High Fidelity 3D Head Reconstruction with 2D Gaussian Splatting

Anil Bas[1,2]
abas@bournemouth.ac.uk
Oleg Fryazinov[1]
ofryazinov@bournemouth.ac.uk
Xiaosong Yang[1]
xyang@bournemouth.ac.uk
Callum Rex Reid[2]
cal@visualskies.com

[1] National Centre for Computer Animation, Bournemouth University
[2] Visualskies Ltd

Implicit neural representations such as Neural Radiance Fields (NeRF) [8] have changed the landscape of novel view synthesis and scene reconstruction. This is extended to applications on virtual avatars as well, enabling highly realistic 3D head models. Despite progress, several challenges remain in 3D reconstruction using NeRF, such as handling occlusions and artefacts, building intricate geometries across diverse bodies and achieving cost-effective rendering.

Alternatively, 3D Gaussian Splatting (3DGS) [6] offers an explicit representation, capturing shape and appearance by optimising 3D Gaussians. These primitives, being differentiable volumetric representations, enable compact optimisation, resulting in efficient training times whilst providing exceptional visual quality. Traditional model-based methods are widely used for head reconstruction, however, they are limited in their ability to express intricate geometry and fine details. In the context of 3D head reconstruction, radiance fields could be adapted to capture complex nonlinear variations in facial appearance. For example, studies like NeRFace [2] and GaussianAvatars [9] explored this idea, demonstrating state-of-the-art rendering results. However, both rely on a morphable parametric face model, alongside other tools such as landmark detectors and face trackers, and are not primarily designed for extracting surface meshes.

Surface reconstruction using 3DGS presents additional challenges. As shown in SuGaR [3], 3D Gaussians do not directly align to the surface geometry, which complicates mesh extraction. Correspondingly, Dai et al. [1] state that the issue arises from the ellipsoid nature of Gaussians as well as ambiguity in the normal direction, and suggest flattening 3D Gaussians into 2D ellipses. Similarly, Huang et al. [4] propose the use of 2D Gaussian primitives (elliptical disks) as surface elements, which significantly improves the quality of the surface reconstruction.

In this work, we present 2D Gaussian Splatting for 3D head reconstruction, capable of capturing facial details, hair and mouth interior. Our approach builds on the similar representation from [1, 4]. This allows our method to efficiently optimise with multiple losses. We apply Poisson reconstruction and perform further refinements to extract the final mesh.

Our approach follows the same structure as 3DGS [6]; 3D Gaussians are represented by a centre point $\boldsymbol{\mu} \in \mathbb{R}^3$, opacity $a \in [0,1]$, colour (as spherical harmonics coefficients) $\mathbf{c} \in \mathbb{R}^3$ and covariance matrix $\sum \in \mathbb{R}^{3x3}$:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad \text{and} \quad \Sigma = \mathbf{RSS}^T\mathbf{R}^T \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{3x3}$ and $\mathbf{S} \in \mathbb{R}^{3x3}$ are the rotation matrix and scaling matrix, respectively. The covariance matrix is parameterised by a rotation quaternion $\mathbf{q} \in \mathbb{R}^4$ and a scaling vector $\mathbf{s} \in \mathbb{R}^3$. Borrowing the Gaussian flattening approach from [1], we set z-scale to zero as $\mathbf{s} = [s_x, s_y, 0]$ and define Diag[.] as diagonal matrix where the updated covariance is:

$$\Sigma = \mathbf{R}\,\mathrm{Diag}\left[s_x^2, s_y^2, 0\right]\mathbf{R}^T. \tag{2}$$

For rendering, we follow volumetric alpha blending, where the colour $\mathbf{C}$ of a pixel is computed by blending N ordered points:

$$\mathbf{C} = \sum_{i \in N} \mathbf{c}_i a'_i \prod_{j=1}^{i-1} (1 - a'_j) \tag{3}$$

where $\mathbf{c}_i$ is colour (view dependent appearance) of each point (modeled via spherical harmonics) and a' denoting blending weight.

We optimise 2D Gaussians with a combination of photometric loss $\mathcal{L}_{\text{rgb}}$, normal loss $\mathcal{L}_{\text{n}}$ and mask loss $\mathcal{L}_{\text{m}}$. We found additional regularisation with precomputed normal maps and background masks effective for improving reconstruction quality. Our simplified final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{n}}\mathcal{L}_{\text{n}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} \text{ and } \mathcal{L}_{\text{rgb}} = (1-\lambda_{\text{rgb}})\mathcal{L}_1 + \lambda_{\text{rgb}}\mathcal{L}_{\text{D-SSIM}} \tag{4}$$

where $\mathcal{L}_{\text{rgb}}$ is from 3DGS [6] and λ indicates respective weights.

Figure 1: High-quality mesh reconstruction from a limited set of images. Note the quality of the hair and facial details, including the mouth interior.

We applied screened Poisson surface reconstruction [5] to resampled point cloud, alongside rendered multi-view depth and normal maps. Here, it is also possible to use truncated signed distance fusion for mesh extraction. Following [1], we perform volumetric pruning to the outlying points.

Fig 1 shows surface reconstruction results as rendered images. We preprocess 16 images, similar to GaussianAvatars [9], for two subjects from the NeRSemble dataset [7]. For each subject, we provide background masking, normal maps and known camera parameters as well.

This work addresses the surface alignment problem by treating 3D Gaussians (ellipsoids) as 2D disks (ellipses). Unlike previous approaches in NeRF and GS-based 3D avatar reconstruction, our method eliminates the need for morphable models (e.g. FLAME or BFM), templates, trackers or annotations; directly guiding Gaussians with input data. Building on prior research, we demonstrate how utilising 2D Gaussian Splatting for 3D head reconstruction enables the effective extraction of high-fidelity, textured meshes.

The results incorporated in this work have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 900025.

[1] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using Gaussian surfels. In *Proc. SIGGRAPH*, 2024.

[2] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Proc. CVPR*, 2021.

[3] Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *Proc. ECCV*, 2024.

[4] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian splatting for geometrically accurate radiance fields. In *Proc. SIGGRAPH*, 2024.

[5] Michael Kazhdan and Hugues Hoppe. Screened Poisson surface reconstruction. *ACM Trans. Graph.*, 2013.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.

[7] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 2023.

[8] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 2021.

[9] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3D Gaussians. In *Proc. CVPR*, 2024.

A Novel Motion Control Workflow for Capturing and Reproducing Realistic Lens Flares

Vincent Maurer
vincentmaurerlt@gmail.com

Technical Directing,
Filmakademie Baden-Württemberg, Animationsinstitut

Lens flares are a common optical artifact in photography and filmmaking caused by reflections and scattering within a camera system. While they can degrade image quality, lens flares are sometimes harnessed artistically to enhance visual appeal. In contemporary film and animation productions, like "Dune" or "The Lego Movie"[4], the characteristics of vintage lenses serve as creative tools. Existing techniques predominantly employ moving 2D sprites, computationally intensive ray tracing or hybrid techniques[2]. However, these methods often can't capture the full complexity of real-world lens flares. My involvement with the final year 3d animated film project "The Deep Above" [1] led me to recreate a specific lens flare of an anamorphic lens. Traditional compositing plugins did not achieve the results the director envisioned. This prompted me to propose a novel workflow for capturing and reproducing them.

Several factors like lens design, incoming light angle, aperture, focus, size and colour of the light source contribute to the appearance of lens flares. To limit the amount of variables to capture, the research focused on three factors: angle of incoming light, lens model and aperture.

The flares are captured by moving a small point of light across the field of view of a camera, creating a regular grid of point flares. A focused white LED light source with an adjustable aperture was placed in a blacked out room. The light was pointed directly at the camera, which was mounted to an industrial motion control robot. [2] A script for the MIMIC for Maya Plugin [3], a tool to create animations for motion control robots, was developed. It rotates the camera around its nodal point in a line-by-line serpentine pattern. The dataset has a density of 128 rows and 72 columns normalized to the field-of-view of the lens, combined with 1.75x overscan to record light sources that are not visible in the frustum, but still produce flares. Each lens was captured at four different apertures from wide open (usually about f1.4) to four stops closed, while adjusting ISO and exposure time to compensate for the loss of light. For further processing the footage was denoised, the exposure values were visually matched and the shots exported as EXR sequences in ACEScg colour space. The dataset includes a variety of lenses - vintage and current, spherical, anamorphic, and zoom - to capture diverse flare scenarios.

Initially, a compact Nuke gizmo selected appropriate lens flare images and employed a dissolve operation to blend the four closest images. This resulted in high-quality still imagery, but lacked performance and exhibited blending artifacts(Figure 1.1). Inspired by previous work by [3], the lens flare images were stored in a sprite sheet. Screen space UVs were used to crop out the four adjacent tiles around the selected position. The images were then blended by distance.

Based on this method, a quick prototype for the Unreal Engine (Figure 1.8) was developed. It showed more complex flare characteristics than the default Unreal implementation (Figure 1.7), which is implemented as a post-process filter, but only creates one flare, limiting flexibility. To create animatable point light flares with fewer artifacts, multiple interpolation techniques for generating smoother in-between frames were evaluated. Using Nuke's Kronos node, I interpolated two samples of each line in the grid along the x-axis and used the outputs to interpolate the y-axis. This led to sharper images and smoother interpolation. The challenges of this optical flow based technique are morphing artifacts due to the overlapping moving transparent layers in a lens flare.

In addition to exploring purely image-based techniques, I tested an implementation of the machine learning based RIFE [1]network in Nuke. This network, built for retiming, estimates optical flow and interpolates images similar to the Kronos node. It results in sharper images, but showed similar morphing artifacts. Using retiming tools to interpolate in a 2D grid of images has limitations due to their one-dimensional nature. To address this, I trained Nuke's machine learning framework, Copycat, to interpolate between four images concurrently. Although the results showed less clarity then those produced by traditional retiming networks 1.6), this approach presents an intriguing avenue for further research.

[1]The Deep Above Making Of: https://www.therookies.co/entries/31298
[2]Light: Godox S30, Robot: Kuka KR-16)
[3]MIMIC Plugin for Maya: https://www.mimicformaya.com/

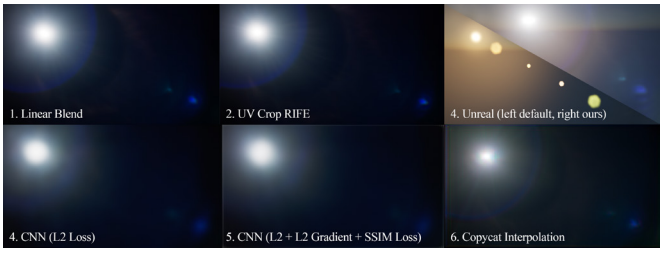


Figure 1: Comparison between different explored lens flare techniques

Instead of relying solely on interpolation, I also explored custom models for inferring the images. The input consists of a lens flare image and the corresponding three-dimensional vector: x and y positions in normalized device coordinates (NDC) and an aperture value normalized from f1 to f22 down to the range 0-1. While training a conditional generative adversarial network (cGAN) on the dataset, I encountered multiple issues. These included mode collapse, instability, and non-convergence.

Subsequently, I reverted to training an inverse classification convolutional neural network (CNN) due to its good fit for the problem and relatively simple implementation. The architecture is composed of 6 upsampling convolution layers with batch normalization and LeakyReLU. This is followed by a fully connected layer to the final resolution of 512x256. The resulting images lost high frequency detail (Figure 1.1) compared to the ground truth and appeared cloudy (Figure 1.4). This might be connected to an insufficient L2 loss function, but could also be a result of misalignment in the dataset due to our capture method. To enhance detail preservation, I combined L2 loss, L2 gradient loss, and structural similarity index (SSIM) with equal influence. This exhibited sharper edges, but still failed to resolve some small details (Figure 1.5). After training, the tested networks were exported as TorchScript and loaded in as a gizmo in Nuke's Cattery Format [4] for further comparison and inference.

The innovative capture method and its resulting lens flare tools have demonstrated successful applications in various VFX and CG productions at Filmakademie. Notably, the anamorphic flares and glows have piqued the interest of our compositors due to their ease of use and instant realistic appearance, requiring minimal manual adjustments. Looking ahead, I envision developing an affordable, compact motion control setup to capture flare data and want to explore more advanced machine learning techniques. A portion of the dataset will also be published for further research [5]. Ultimately, a deeper understanding of lens and camera characteristics, coupled with more accurate data, will undoubtedly enhance the image quality of visual effects and animated films.

The author would like to thank the staff at Doublecheese Film that supported us with their Motion Control System, Benjamin Völker (Zeiss) and Johanna Barbier (The Foundry), as well as the Filmakademie team: Niklas Wolff, Andreas Blind, Max Pollmann, Liliane Maurer, Fynn Aurich, Alexander Kreische, Jonas Trottnow, Simon Spielmann and Volker Helzle.

[1] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation, 2022. URL https://arxiv.org/abs/2011.06294.

[2] M. Hullin, E. Eisemann, H.-P. Seidel, and S. Lee. Physically-based real-time lens flare rendering. *ACM Trans. Graph.*, 30(4), jul 2011. doi: 10.1145/2010324.1965003.

[3] S. Jo, Y. Jeong, and S. Lee. Real-time nonlinear lens-flare rendering method based on look-up table. *Journal of KIISE*, 44:253–260, 03 2017. doi: 10.5626/JOK.2017.44.3.253.

[4] E. Pekkarinen and M. Balzer. Physically based lens flare rendering in "the lego movie 2". In *Proceedings of the 2019 Digital Production Symposium*. ACM, 2019. doi: 10.1145/3329715.3338881.

[4]Nuke Cattery: https://community.foundry.com/cattery
[5]Lens Flare Dataset: https://vincent-maurer.github.io/lens-flare-capture/

Efficient Audio-Visual Fusion for Video Classification using Attend-Fusion

Mahrukh Awan, Asmar Nadeem, Armin Mustafa
{mahrukh.awan,armin.mustafa,asmar.nadeem}@surrey.ac.uk

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

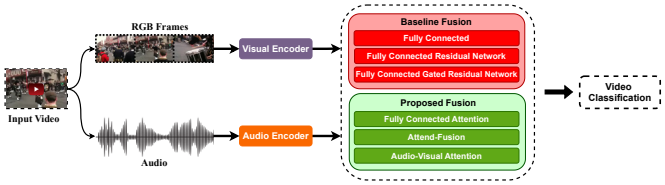


Figure 1: Overview of our proposed audio-visual video classification framework on YouTube-8M dataset [1], illustrating different fusion mechanisms of audio and visual modalities.

We present Attend-Fusion, a novel and efficient approach for audio-visual fusion in video classification tasks. Our method addresses the challenge of exploiting both audio and visual modalities while maintaining a compact model architecture. Through extensive experiments on the YouTube-8M dataset [1], we demonstrate that our Attend-Fusion achieves competitive performance with significantly reduced model complexity compared to larger baseline models.

The YouTube-8M dataset comprises millions of YouTube videos, each annotated with labels from a diverse vocabulary of 4,716 visual entities. Our approach focuses on effectively fusing audio and visual modalities to improve classification accuracy on this challenging multi-label dataset.

We compare our Attend-Fusion model with several baseline approaches, including Fully Connected (FC) networks. The FC Late Fusion model emerged as the best-performing baseline. Our Attend-Fusion model incorporates self-attention mechanisms [2], defined as:

$$\mathbf{X}_{att} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are query, key, and value matrices derived from the input features, and d is the feature dimension.

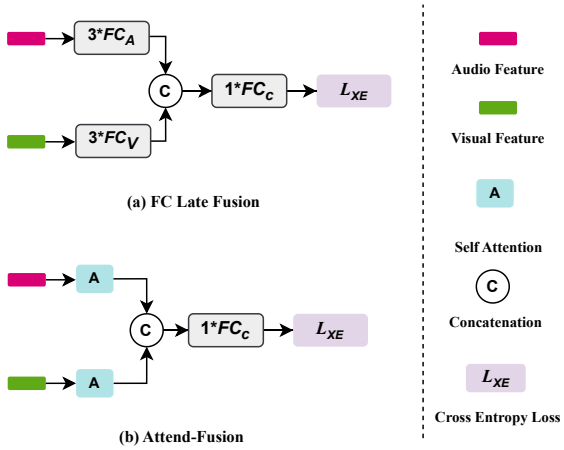


Figure 2: Comparison of Fully-Connected (FC) Late Fusion (baseline) and Attend-Fusion architectures.

Metric	FC Late Fusion	Attend-Fusion
GAP (%)	80.87	80.55
F1 Score (%)	75.96	75.64
Parameters (M)	341	72

Table 1: Comparison of FC Late Fusion and Attend-Fusion.

Table 1 compares the performance and efficiency of FC Late Fusion and Attend-Fusion. Our model achieves comparable performance with significantly fewer parameters, demonstrating its efficiency.

The Attend-Fusion architecture processes audio and visual features separately through attention networks, which consist of fully connected layers and self-attention mechanisms. The attended features are then fused using a late fusion strategy, allowing the model to learn both modality-specific and cross-modal representations.

We conducted comprehensive ablation studies to investigate the contributions of different components. Removing the attention mechanism or using only a single modality led to significant drops in performance, confirming the value of our approach. The cross-entropy loss function used for multi-label classification is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) + (1-y_{i,c})\log(1-\hat{y}_{i,c}) \tag{2}$$

where N is the number of samples, C is the number of classes, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability.

Figure 3: Qualitative results comparing the top-3 predictions of Attend-Fusion, FC Late Fusion (SOTA baseline), and the ground truth labels on representative examples from the YouTube-8M dataset.

Qualitative analysis of Attend-Fusion's predictions on representative examples from the YouTube-8M dataset showcases its ability to accurately classify videos across various domains, including sports, music, gaming, and more. The model demonstrates superior performance in capturing fine-grained details and maintaining coherence in its predictions, highlighting the effectiveness of the attention mechanism in integrating audio and visual information.

Attend-Fusion's compact design and high performance make it well-suited for real-world applications where computational resources are limited, such as mobile devices or edge computing scenarios. Our work contributes to the ongoing efforts to develop more sustainable and deployable AI solutions for video understanding tasks, opening new possibilities for efficient video classification across various domains.

[1] Abu-El-Haija, S. et al. (2016). YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv preprint arXiv:1609.08675.

[2] Vaswani, A. et al. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems, 5998-6008.

2D or not 2D: Live broadcasting into a game engine using a pseudo-3D approach

Graham Thomas
https://www.bbc.co.uk/rd/author/?id=author-people-g-a-thomas
Fiona Rivera
https://www.bbc.co.uk/rd/author/?id=author-people-fiona-rivera

BBC Research & Development,
Media City UK, Salford

Figure 1: Capture of a performance, with central and left-hand cameras visible (left); camera feed from left-hand camera embedded in virtual night club (right).

Game-like environments that offer live multiplayer capability are becoming a major form of entertainment. A large community, estimated to be 3.2M in the UK, also spend time in these environments for social & experiential reasons, rather than gaming. The ability to easily deliver a live event such as a concert to users of such platforms is thus of interest to broadcasters.

Volumetric capture is one approach that could be taken to capture and stream performers. However, this can be challenging, particularly when trying to capture a large stage at an event such as a music festival where there may be multiple performers, effects such as smoke and lasers, and lighting that might form an integral part of the show.

A simpler approach that may suffice in situations where a fully-free viewpoint is not required, is to use one or more conventional 2D video streams (sometimes referred to as video billboards), or so-called 2.5D video (where depth data may be used to help extend the range of viewpoints) [1]. Such approaches may be particularly suitable when the range of viewpoints is naturally limited, such as for an audience viewing a stage from the front. The lack of true 3D may also be less apparent in browser-based VR use cases (which are our current focus), where the user has no stereoscopic depth cues, rather than applications using a VR headset. There is also anecdotal evidence that flat images placed within a 3D environment can trick the brain into seeing it as true 3D, as illustrated by the successful use of Pepper's Ghost illusions for so-called holographic performers on stage [2].

Adding depth data to images, or simply using an alpha mask to define the foreground area, can help make a performer captured in a single video stream look better-integrated into a 3D environment. However, often the stage background or lighting/haze effects form an important part of a live performance and would also make depth/alpha estimation difficult, so for our initial work we have focused on what can be achieved with one or more conventional video streams without alpha or depth.

This Short Paper (which is a more focused version of a broader paper [3]) presents an experiment to evaluate how well one or more 2D images can be used to give a feeling of 3D. We built a 3D "nightclub-style" environment in Unreal Engine that included a replica of the real-world stage on which we had recorded a number of musical performances, as shown in Figure 1. We tested user responses to the following conditions:

C1: Single centre camera view only. This provides seamless viewing, but skewed perspectives to either side of the stage.

C2: Multi-cam view with 3 cameras. This provides more accurate viewing perspectives from the sides, and thus potentially more sense of depth, but at the cost of a visual glitch when the video stream switches to the feed closest to the avatar's position.

C3: Multi-cam view as in C2, but with flashing stage light distractions when the video stream switches between cameras.

Consistent viewing angles and duration for each condition were provided by the point-of-view from a player avatar following a specified path around the virtual venue. We used a total of 9 video clips of the experience (3 conditions x 3 artists) recorded in the game engine. Participants were instructed to view each video and state the degree to which they agreed that "the appearance of the performance was satisfactory from each angle shown".

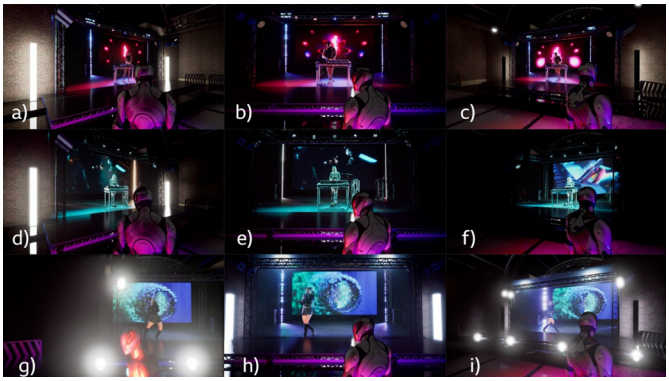


Figure 2: Example views from the study. The top row (a), (b), (c) show single camera views from either side of stage and centre. The middle row (d), (e), (f) show multi-cam views. The bottom row (g), (h), (i), show the multi-cam views when the distraction lights are triggered (g), (i) as the player avatar point of view changes when moving to the side.

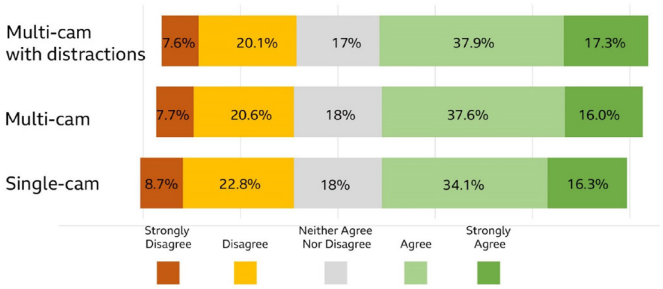


Figure 3: Overall study results, showing percentage of respondents' agreement/disagreement that the appearance of the performance was satisfactory from each angle shown

Figure 3 shows that our 75 respondents had a small preference for C3 over C1 (about 5% more participants agreeing or strongly agreeing that the appearance was satisfactory), with C2 being about 3% above C1. We deem these small differences worthy of further exploration in an interactive version of the study, where participants can freely explore the space and their own viewing angles. The qualitative feedback also indicates that the relationship between the lighting in the virtual space and in the real-world performance is worth further exploration.

These preliminary findings suggest that simple video-based representations can satisfy the majority of the audience (about 55% of the audience for C3), where the range of viewing angles is constrained. This also has design implications for virtual venues, suggesting for example that that a wedge-shaped audience space with the stage at the narrow end might be worth considering to constrain the position of users. Our continuing work will test this hypothesis further.

This work was carried out in the MAX-R project, co-funded by Innovate UK and the Horizon Europe Research & Innovation Programme under Grant Agreement No. 101070072. Thanks to artists Badliana, KDYN, TWST, and Production Park, UK for help with the test shoot.

[1] O. Grau, M. Price, and G. Thomas. Use of 3-d techniques for virtual production. *BBC R&D White Paper WHP033*, 2002.

[2] K. Grow. Live after death: Inside music's booming new hologram touring industry. *Rolling Stone Magazine*, September 2019.

[3] G. Thomas, F. Rivera, L. Kelso, B. Weir, P. Rich, and O. Moolan-Ferose. Live music in virtual immersive spaces. *Proceedings of IBC*, September 2024.

Overcoming Spatial Limitations in Optical Motion Capture Using Nested Volumes

Paulo Scatena
https://www.technicalwhat.com/

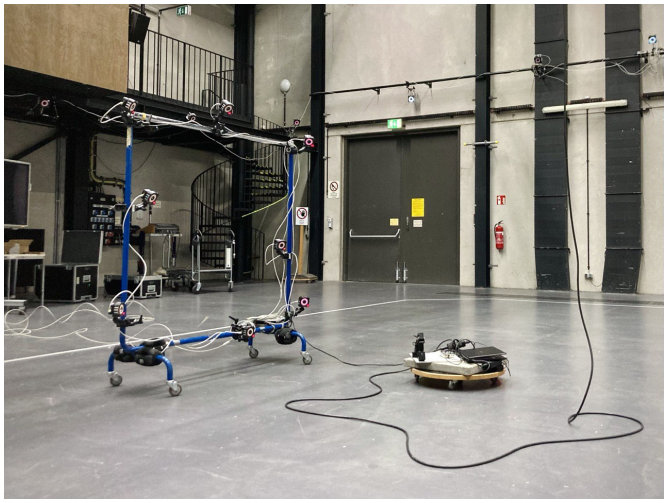


Figure 1: Motion capture rig nested within larger capture volume

Abstract: This paper presents a semi-dynamic motion capture setup that incorporates a mobile camera rig nested within a larger static volume. This approach is shown to mitigate occlusion issues and extend capture coverage by allowing on-the-fly adjustments to the effective capture volume. Although this setup introduces minor errors, it proves to be a valid tool for increasing visibility in complex or constrained environments.

1 Introduction

Optical motion capture is subject to the problem of occlusion: things get in the way. Relying on stationary camera arrangements does not make it any less challenging. Inspired by targeted visual effects techniques in cinema, such as those used in "The Irishman", this paper proposes a mobile "mocap universe" that can be focused on key performance areas, much like a cameraman framing a lead actor, providing increased capture resolution where performance is most critical.

This setup provides a practical solution for capturing data in dynamic or occlusion-heavy environments, which are increasingly common in modern VFX and virtual production pipelines.

2 Method

For a practical test, a mobile rig was equipped with eight OptiTrack S250e motion capture cameras (Figure 1), while 24 Prime 41 cameras covered the tracking volume (an area of approximately 100m²). Having cameras on wheels gives us a de facto dynamic capture universe, but lacks the solidity of static tracking. A naive approach is proposed here: the rig itself was tracked so that its coordinate system would be placed within the global one.

The hypothesis was that, under certain conditions, this nested tracking system could reduce the loss of information compared to static tracking alone. For the fixed universe, tracking a large object with large markers is easy, so it has no problem tracking the rig on wheels. This rig can then be positioned to give these cameras an optimal view of the target. They "illuminate" the area, providing local resolution, while the static volume provides global coverage.

For this study, a small rigid body consisting of six passive markers - our 'hero asset' - was attached to a motorised device programmed to perform a steady motion. This was placed in a wheeled platform that the mocap rig would follow closely.

The entire array was initialised as a cohesive volume so that the correct affinity between the coordinate systems could be defined. Subsequent recordings could then be post-processed in relation to specific cameras to export the tracking corresponding to their universes. Data were recorded under different occlusion conditions and movement patterns.

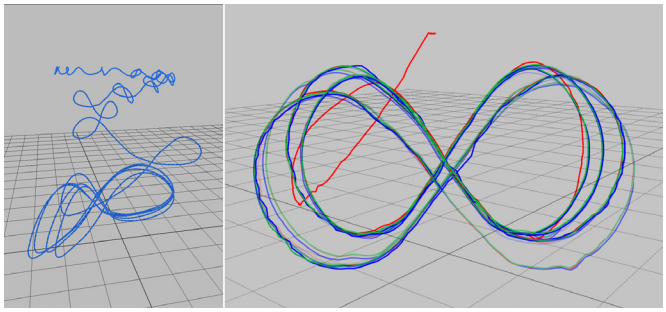


Figure 2: Trajectory in global space (left) and local space (right) in three sets: control (green), extreme (red) and mobile (blue)

3 Results and Discussion

Illustrative results are shown in Figure 2. In this scenario, the hero asset followed a pattern 30cm wide, while the combined array moved around 5 meters across the room. Data from three sets are compared: static cameras with good visibility (control), static cameras with reduced visibility due to set elements (extreme), and the moving cameras (nested tracking). The data is presented in two different coordinate spaces so that spatial deviations are clear.

The nested tracking introduced a positional error of about one centimetre, with greater inaccuracies during movement. However, it was able to track the hero asset at all times, where occlusion might otherwise have compromised the data. These results suggest that, with further refinement, nested tracking could provide a practical solution for extending mocap volumes in constrained environments, dynamically expanding an otherwise static motion capture volume.

Immediate gains in reliability and ease of use would be achieved by using active or semi-active markers (off-the-shelf infrared LEDs can appear to the cameras as very reliable "passive" markers). The simplicity of the study presented here should be emphasised. It was set up and recorded in spare studio time, over the course of two days. This means that there is room for improvement in all directions. The work presented here is shared in the belief that further investigation may be justifiable, resources permitting.

While the proposed setup may be unconventional, it may be refreshing to consider optimising tracking systems for the talent being tracked, rather than for empty space. As tracking systems become increasingly self-calibrating, fully dynamic setups may not be far off in the future. Perhaps modular motion capture volumes are an early step.

4 Conclusion

This brief study shows that a nested tracking system can in effect mitigate tracking losses due to occlusion, while allowing for some dynamic adaptation of an optical motion capture volume.

While this approach introduces some complexity and noise, the potential to improve capture in constrained or occlusion-prone environment makes it a promising tool, particularly for entertainment applications where intent of action may be more valuable than spatial precision, and motion capture can perhaps be instructed to follow our point of interest.

5 Acknowledgements

The author would like to thank the Animationsinstitut at the Filmakademie Baden-Württemberg, for access to their space and hardware, and Charley Henley for triggering the reflection on the matter.

A Preliminary Study on Real-Time Compositing for 2D Cel-Animation Production

Takeshi Okuya[12]
https://orcid.org/0009-0003-1226-2753
Brian Pascente[13]
pascenby@rose-hulman.edu

[1] ufotable Inc.
[2] Waseda University
[3] Rose-Hulman Institute of Technology

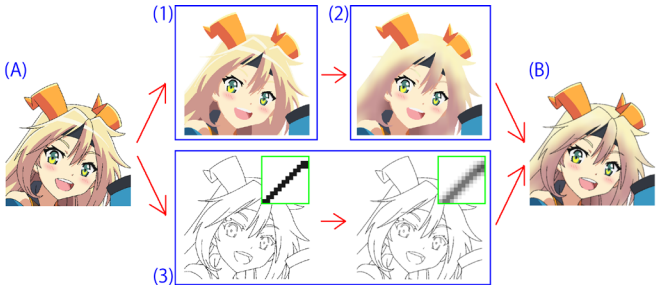


Figure 1:[1]Example of character processing, including hair gradation and line smoothing. (A) Input source. (1) Contour line removal. (2) Specified color area blur. (3) Contour line extraction blur. (B) Composite result.

1 Introduction

In cel-animation production, images are created by artists specializing in line drawing, painting, and background processing, and the character and background images are composited. The compositing process, which used to be done during film shooting, is now managed by compositing software. In addition, during this process, the character image is enhanced, and various effects are applied. An example of character processing is shown in Figure 1, where the hair gains a natural appearance with shading by gradation, and the contour lines are smoothed without jagged edges. These processes are typically based on classical filter algorithms rather than neural networks, allowing composite artists to make fine adjustments.

However, real-time rendering has not yet been achieved in the compositing process due to the computationally heavy character processing. This results in a waiting time for artists between setting parameters and previewing, which reduces work efficiency. While the filters employed are combinations of general algorithms and methods for acceleration are well-known [2][3], to the best of our knowledge, acceleration techniques that are specifically optimized for both modern hardware and the unique features of animation character images have not been thoroughly explored.

In this study, our goal is to achieve real-time rendering of the compositing process. By implementing acceleration techniques and measuring the speed of each step, we investigated the challenges that need to be addressed to achieve real-time performance.

2 Implementation

To evaluate the speed of each image processing algorithm shown in Figure 1, we implemented several standard methods and acceleration techniques on both CPU and GPU. One acceleration method involved separating the filter in the XY direction [2][3]. Additionally, since the character image consists of a limited color palette, with each color occupying a small area relative to the total pixel count, processing was further accelerated by pre-listing the pixels to be processed. The following sections detail the implementation of each algorithm.

Contour Line Removal: The color of the pixel designated as the contour line is replaced with the surrounding color. In the first implementation, replacement was performed on all pixels in a single pass. If replacement failed due to surrounding pixels being lines, a count was incremented, and the process was repeated until the count reached zero. In the next implementation, line pixels were added to a list, and replacement was conducted using this list, continuing until the list was empty.

Specified Color Area Blur: Multiple colors to be blurred are specified, affecting only those areas. In the first implementation, blur processing was applied in both vertical and horizontal directions in a single pass, with blur strength adjusted based on the specified color. In the next implementation, to reduce pixel samplings and multiplication operations, the X-direction blur was applied in the first pass, followed by the Y-direction blur in the second pass. Additionally, by adding pixels to be processed to a list during the first scan and executing the blur in both directions using this list, the number of pixel samplings was minimized.

Contour Line Extraction Blur: The color of the contour line is specified, enabling blur processing only for that color. In the first implementation, a mask was created to isolate the contour line, followed by blur processing on this image. In the next implementation, line extraction and X-direction blur were performed in the first pass, with Y-direction blur applied in the subsequent process. Additionally, we implemented processing using pixel lists and applied blurs in both the X and Y directions.

3 Results and Discussion

The implemented algorithms were executed on both CPU and GPU, and their performance was measured. The test environment consisted of a Ryzen7 7700X and a GeForce RTX 3060, with implementation on Windows 11 Pro 64-bit, Visual Studio 2022, OpenCV 4.10.0, and Unity 6000.0.19f1. Each process was tested on 86 images[1] with a resolution of 1684x1000 pixels, and the average processing time per image was calculated. The CPU ran in a single thread using SIMD instructions, while the GPU utilized Unitys Visual Compositor [1] with fragment and compute shaders. GPU processing time was measured in FPS for each process combination, and execution times were derived from the reciprocal of the FPS and their differences. The results are shown in Tables 1-3.

Table 1: Processing time of each contour line removal process [ms]

	CPU	GPU
Sample all pixels every time	17.8	4.88
Create line pixel list first	10.5	4.23

Table 2: Processing time of each specified color area blur process [ms]

	CPU	GPU
1 pass (5 px)	96.8	0.27
2 pass (5 px)	36.7	0.27
Pixel list and 2 pass (5 px)	33.6	0.44
1 pass (50 px)	7170	4.88
2 pass (50 px)	145	0.63
Pixel list and 2 pass (50 px)	155	0.44

Table 3: Speed results for contour extraction blur [ms]

	CPU	GPU
1 pass (5 px)	530	0.20
2 pass (5 px)	135	0.28
Pixel list and 2 pass (5 px)	119	1.28

The results demonstrate that pixel listing, which accounts for the unique features of animation character images, was particularly effective in replacement processing. However, for blur processing, performance improvements were minimal or even slightly slower in some cases. This is attributed to the significant reduction in iterations for replacement processing, while the overhead of pixel listing proved relatively high for blur processing. Moreover, GPU implementation exhibited greater acceleration for blur processing compared to the CPU, although the increase in GPU command overhead appeared to limit acceleration in replacement processing. Encouragingly, substantial acceleration was observed on the GPU through pixel listing, especially when using larger filter sizes (50 px). Overall, this study concludes that pixel listing is effective for acceleration under specific conditions. Additionally, the GPU implementation achieved 10.5 ms (94.8 FPS) from input to displaying all filters (color area blur of 50 px), demonstrating real-time rendering for character processing on the GPU. Future challenges include verifying multi-threaded performance on the CPU, further optimizing GPU implementations, and exploring other image processing algorithms.

[1] Visual compositor overview. https://docs.unity3d.com/Packages/com.unity.visual-compositor@0.30. (Accessed on 09/20/2022).
[2] Tomas Akenine-Mller, Eric Haines, and Naty Hoffman. *Real-Time Rendering, Fourth Edition*, chapter 12 Image-Space Effects, pages 513–543. A. K. Peters, Ltd., 2018.
[3] Richard Szeliski. *Computer Vision: Algorithms and Applications, Second Edition*, chapter 3 Image processing, pages 107–189. Springer, 2022.

[1]Test images are publicly available(a) and provided under Unity-Chan License Terms(b).
(a) https://unity-chan.com/download/releaseNote.php?id=UnityChanAnimationMaterialData
(b) https://unity-chan.com/contents/guideline_en/

Synthesis of Realistic Tongue Movements for Speech Animation Using Gated Recurrent Units

Jake Mwangi-Powell[1], Robert Kosk[2], Tianxiang Yang[2],
Cathair Kerrigan[2], Marco Volino[1]

[1]University of Surrey, [2]Humain Studios

1 Introduction

Speech animation has long focused on lip-sync and facial movements, neglecting the tongue's role in articulation, which is critical for achieving realistic character animations [2]. This project addressed this gap by proposing a machine learning (ML)-based approach to synthesize realistic tongue movements from speech audio inputs, resulting in more lifelike digital speech. Previous methods primarily focused on lips and facial movements. ML, particularly Recurrent Neural Networks (RNNs) and LSTM networks, have shown potential in speech animation [3]. However, these techniques have not been fully utilized for tongue synthesis. Transformers have shown success in natural language tasks but are computationally intensive. This project explored these models to find the most efficient and accurate architecture for tongue movement synthesis.

This paper presents a method for synthesizing realistic tongue movements from speech inputs using machine learning, specifically Gated Recurrent Units (GRUs). By leveraging a dataset of magnetic Articulography (EMA), various neural architectures such as Long Short-Term Memory (LSTM) networks, transformers, and GRUs were evaluated. After comprehensive hyper-parameter optimization, the GRU model achieved the best performance with a Mean Squared Error (MSE) loss of 1.473. The proposed approach improves speech animation by dynamically generating tongue movements that enhance the realism of digital character animations.

2 Methodology

Pre-existing tongue position data that was captured using Electromagnetic Articulography (EMA), which tracks tongue and lip positions in 3D space [1] was utilized. This dataset, combined with synchronised speech audio, was pre-processed to ensure uniform length for batch processing. A custom forward kinematics (FK) rig of the tongue was created for the project, as shown in Figure 1, which allowed precise control over key parts of the tongue. Joints placement overlaps with positioning of EMA sensor coils, allowing for seamless transfer from sensors data onto tongue mesh animation. Several neural architectures were evaluated, including RNNs, LSTMs and transformers. The GRU model, with bidirectionality enabled, was selected for its simplicity, efficiency, and ability to capture sequential data dependencies. Hyperparameter tuning identified an optimal learning rate of 0.001, 256 hidden dimensions, 4 layers, and a dropout rate of 0.3. The Adam optimizer was used due to its superior performance over alternatives like RMSprop and SGD. Cross-validation was employed to ensure the model generalized well across unseen data, with training, validation, and test sets split to measure robustness.

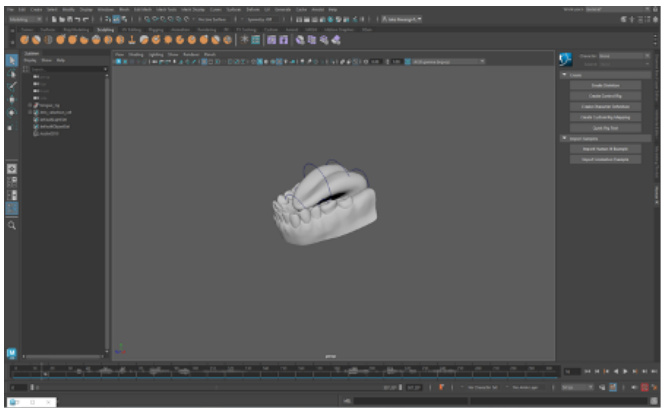


Figure 1: Tongue rig rendered in Autodesk Maya

3 Results

The GRU model achieved an MSE loss of 1.473, outperforming LSTM and transformer models in both accuracy and computational efficiency. A batch size of 8 and early stopping with a patience value of 5 further optimized performance. Models without bidirectionality performed worse, demonstrating the importance of capturing context from both past and future frames in speech data. Evaluation on unseen data confirmed the model's ability to generalize well, with accurate predictions of tongue movements closely aligned with speech inputs, as shown in Figure 2. Objective evaluations using MSE, combined with subjective assessments of the generated animations, indicated a high level of realism in the synthesized tongue movements. Figure 2 shows that the model's predictions closely follow the actual values throughout most of the time steps, with minimal deviations. Despite some fluctuations, particularly in the earlier time steps, the overall alignment between predictions and actual values suggests the model captures the underlying pattern effectively.

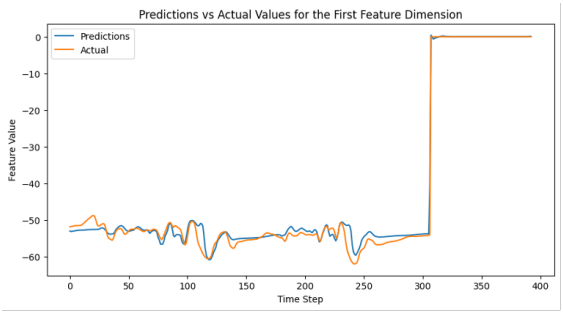


Figure 2: X-coordinate position for feature 1 at each frame

4 Conclusion

This work successfully developed a GRU-based approach for synthesizing realistic tongue movements from speech inputs. The proposed method enhances the realism of speech animations in digital characters, with applications in fields such as gaming, film, and VR. By demonstrating the effectiveness of ML in handling complex articulatory movements, this project opens new avenues for more immersive and lifelike digital animations. Future work could focus on expanding the dataset to include a wider range of speakers and speech variations. Integrating this tongue animation system with facial animation frameworks would create a more cohesive speech animation system. Additionally, optimizing the model for real-time performance would enable its use in interactive applications like virtual reality (VR) or live animation environments.

References

[1] Salvador Medina, Denis Tome, Carsten Stoll, Mark Tiede, Kevin Munhall, Alex Hauptmann, and Iain Matthews. Speech driven tongue animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2022.

[2] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01344.

[3] Pengcheng Zhu, Lei Xie, and Yunlin Chen. Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings. In *Interspeech 2015*, pages 2192–2196, 2015.

Efficient Multi-Scale 3D Gaussian Splatting

Umar Farooq, Jean-Yves Guillemaut, Adrian Hilton, Marco Volino
{m.farooq,j.guillemaut,a.hilton,m.volino}@surrey.ac.uk

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

Figure 1: **Top:** RGB rendering. **Middle:** Gaussian Visualization. **Bottom:** Size, SSIM, cumulative training time and global iteration. Each column indicates one level of detail used in our coarse-to-fine optimization and the last column shows 3DGS for comparison.

Introduction: 3D Gaussian Splatting (3DGS) [1] is an explicit point-based novel view synthesis technique that achieves high visual quality and fast training times. However, 3DGS suffers from high memory and storage usage [3, 4] limiting its applicability across various device form factors. In this context, we introduce a novel and efficient 3DGS coarse-to-fine optimization strategy. Our method reduces the memory overhead of 3DGS by initiating the training process with significant over-reconstruction, which serves as an effective regularizer, and progressively refines the scene representation. Our approach also produces stand-alone scene representations at each level-of-detail in the progressive refining process, enabling variable storage, transmission, and rendering based on the downstream requirements.

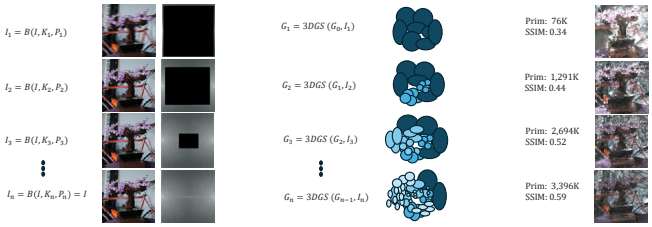


Figure 2: Our frequency modulation and progressive Gaussian levels-of-details method diagram.

Method: We employs a progressive frequency control strategy with five distinct image and scene quality levels I_n and G_n respectively in range $n = (1-5)$ as depicted in our method diagram shown in Figure 2. We start by applying the blur with a large kernel size to the training images, reducing their detail and noise. This initial reduction in detail allows the model to focus on learning the broader, more significant structures of the scene. As the training progresses, the level of blur is gradually reduced by reducing the size of the blur filter which reintroduces higher frequency details in a controlled manner.

The optimization starts from the initial 3D points obtained from SfM on the training images I and this initial sparse representation acts as our G_0 at $n = 0$. Then for each subsequent level in our coarse-to-fine optimization process we use the last set of Guassians G_{n-1} as the starting point and the 3DGS[1] optimization loss is used to optimize G_{n-1} using the filtered images I in range $n = (1-5)$. Where the images I_n are obtained from our frequency modulation function. For each level the images get progressively sharper and more high frequency content is allowed to remain in the image. For the last level at $n = 5$ used we directly pass the orignal training images to the 3DGS optimizer.

Results: Our method reduces the number of primitives required by 62%, lowers GPU memory usage by 40% and reduces optimization time by 20% as shown in Figure 3. Our method successfully reconstructs subtle details like grass and shrubs all the while using fewer Gaussian primitives. The proposed method is combined with an off-the-shelf compression method [3] to obtain further compression and to showcase the general nature of our contribution.

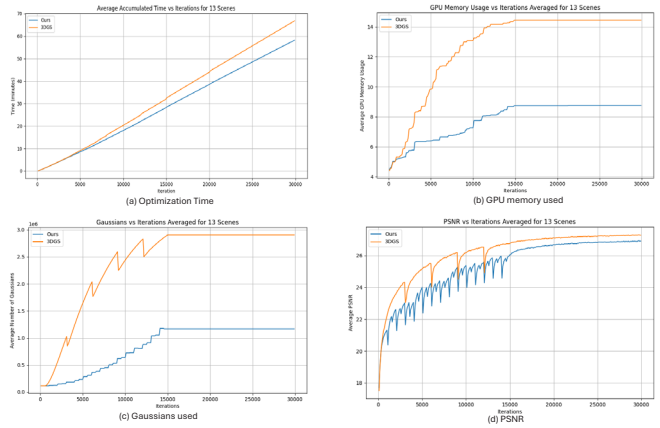


Figure 3: Shows optimization time, GPU memory usage, number of Gaussian primitives and PSNR for our method compared to 3DGS[1].

Table 1: Quantitative results for our method on commonly used benchmark datasets.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	Size (MB) ↓
Mip-NeRF360				
Plenoxels	0.626	23.080	0.463	2,100.0
INGP-Base	0.671	25.300	0.371	13.0
INGP-Big	0.699	25.590	0.331	48.0
Mip-NeRF 360	0.792	27.690	0.237	8.6
3DGS	0.815	27.210	0.214	734.0
Reduced-3DGS[4]	0.809	27.100	0.226	29.0
Compact-3DGS[2]	0.797	27.030	0.247	29.1
Compress-3DGS[3]	0.801	26.981	0.238	28.8
Ours-Progressive Only	0.796	27.111	0.259	21.8
Ours-Full	0.797	26.777	0.256	15.87
Tanks & Temples				
Plenoxels	0.719	21.080	0.379	2,300.0
INGP-Base	0.723	21.720	0.330	13.0
INGP-Big	0.745	21.920	0.305	48.0
Mip-NeRF 360	0.759	22.220	0.257	8.6
3DGS	0.841	23.140	0.183	411.0
Reduced-3DGS[4]	0.840	23.570	0.188	14.0
Compact-3DGS[2]	0.831	23.320	0.202	20.9
Compress-3DGS[3]	0.832	23.324	0.194	17.3
Ours-Progressive Only	0.834	23.465	0.203	13.5
Ours-Full	0.819	23.061	0.224	9.93
Deep Blending				
Plenoxels	0.795	23.060	0.510	2,700.0
INGP-Base	0.797	23.620	0.423	13.0
INGP-Big	0.817	24.960	0.390	48.0
Mip-NeRF 360	0.901	29.400	0.245	8.6
3DGS	0.903	29.410	0.243	676.0
Reduced-3DGS[4]	0.902	29.630	0.249	18.0
Compact-3DGS[2]	0.900	29.730	0.258	23.8
Compress-3DGS[3]	0.898	29.381	0.253	25.3
Ours-Progressive Only	0.900	29.540	0.260	19.4
Ours-Full	0.898	29.350	0.268	12.02

Applications: Our approach enables efficient and usable optimization of 3DGS up to the highest resolution level-of-detail before running out of GPU memory. A scene can be rendered at a different level-of-detail depending on device hardware and user requirements. We also enable downstream applications where parts of a scene can be rendered in variable quality which can further decrease memory footprint and increase render speed.

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023.

[2] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. *arXiv preprint arXiv:2311.13681*, 2023.

[3] Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. *arXiv preprint arXiv:2401.02436*, 2023.

[4] Panagiotis Papantonakis, Georgios Kopanas, Bernhard Kerbl, Alexandre Lanvin, and George Drettakis. Reducing the memory footprint of 3d gaussian splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1):1–17, 2024.

SINGLE-IMAGE COHERENT RECONSTRUCTION OF OBJECTS AND HUMAN

Sarthak Batra, Adrian Hilton, Armin Mustafa
s.batra@surrey.ac.uk

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

1 Abstract

Existing methods for reconstructing objects and humans from a monocular image suffer from severe mesh collisions and performance limitations for interacting occluding objects. This paper introduces a method to obtain a globally consistent 3D reconstruction of interacting objects and people from a single image. Our contributions include: 1) an optimization framework with collision loss to handle human-object and human-human interactions; and 2) a novel technique to robustly estimate 6 DOF poses, for heavily occluded objects, exploiting image inpainting. Notably, our proposed method operates effectively on images from real-world scenarios, without necessitating scene or object-level 3D supervision (Fig. 1).

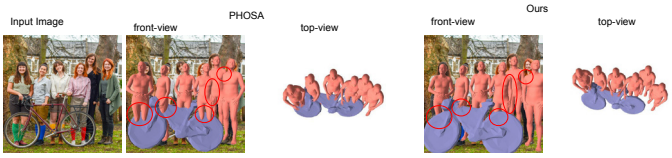


Figure 1: Comparison of the proposed method (right) reconstruction with PHOSA(middle).

2 Methodology

Existing methods for human and object reconstruction are either limited to single objects and humans or give limited performance for complex images with multiple people and objects. These methods estimate the 3D poses of humans and objects independently and do not take into account the human-human interactions [4] and even if they do they generally follow a supervised approach. This leads to large collisions between the meshes with incoherent reconstructions. Our contributions are:

- A method for generating a cohesive scene reconstruction from a single image by capturing interactions among humans and between humans and objects within the scene, all without relying on any explicit 3D supervision.
- A collision loss in an optimization framework to robustly estimate 6 DOF poses of multiple people and objects in crowded images.
- An inpainting-based method to improve the segmentation mask of heavily occluded objects that greatly boosts the precision of 6 DOF object position estimations.

The proposed method (Fig.2) reconstructs interacting humans and objects in a 3D scene from a single RGB image. It begins by detecting humans and objects, followed by SMPL-based per-person reconstruction, which initially results in incorrect spatial arrangements due to collisions. Human 3D locations and poses are refined using an optimization process with a collision loss to prevent overlaps. To accurately estimate the 3D object pose (6-DoF), a differentiable renderer aligns 3D object meshes with 2D segmentation masks [1]. Occluded object masks are corrected via image inpainting, unlike PHOSA [4], which uses an occluded object mask. Finally, joint optimization integrates human-human and human-object interactions for a coherent and realistic 3D reconstruction.

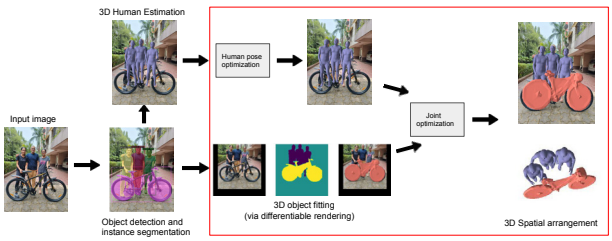


Figure 2: Overview of the proposed method

3 Results

We perform both quantitative and qualitative assessments of the performance of our technique on the COCO-2017 dataset on images that include interactions of humans and objects against PHOSA [4], ROMP[2], and BEV[3].

Methods	E_{H-col}	$E_{H-depth}$	E_{HO-col}	$E_{HO-depth}$
PHOSA	79.42	86.68	78.21	68.84
ROMP	63.51	74.27	-	-
BEV	35.25	56.17	-	-
Ours	**16.46**	**48.37**	**26.65**	**33.77**

Table 1: Quantitative evaluation with PHOSA [4], ROMP[2], and BEV[3]. BEV and ROMP only reconstruct humans. Equations of each evaluation parameter are given in the supplementary.

To perform quantitative evaluation a random sample of images from the COCO-2017 test set was selected, and scene reconstructions were performed. The method was compared with PHOSA, ROMP, and BEV by analyzing mesh collisions for human-human (E_{H-col}) and human-object (E_{HO-col}) interactions, as well as incorrect depth ordering for human-human ($E_{H-depth}$) and human-object ($E_{HO-depth}$) interactions. The results, averaged across images, show that the proposed approach outperforms state-of-the-art techniques in multi-human and multi-human-object reconstruction, leading to more coherent and realistic results with fewer ambiguities as seen in fig.3 and fig.4.

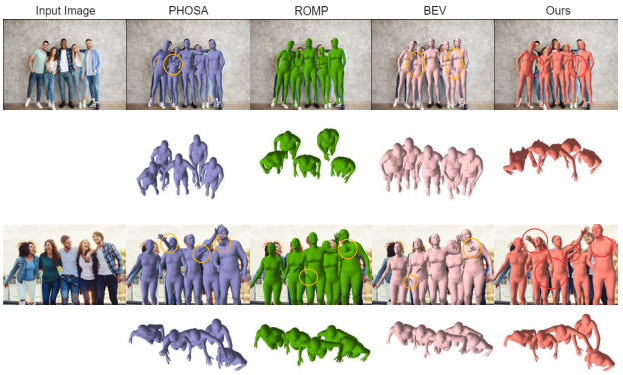


Figure 3: Qualitative comparison between our method with other multi-human reconstruction methods

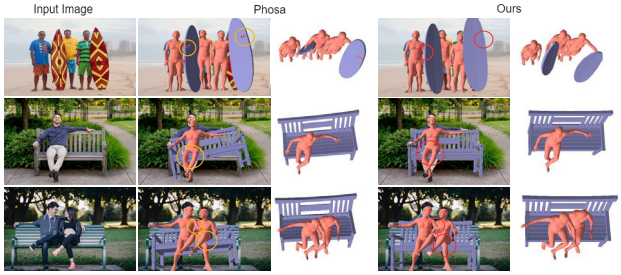


Figure 4: Qualitative comparison between our method and PHOSA

[1] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

[2] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021.

[3] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022.

[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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020.

A New Dataset for Multi-Camera Frame Synthesis

Conall Daly
dalyc21@tcd.ie
Anil Kokaram
anil.kokaram@tcd.ie

Sigmedia Group,
Department of Electronic and Electrical Engineering,
Trinity College Dublin, Ireland

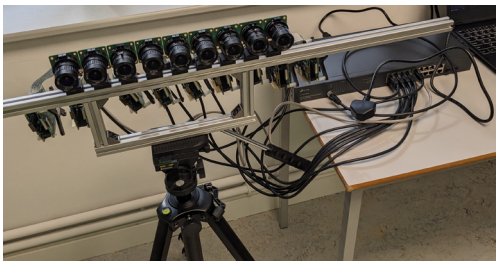


Figure 1: Rig used for frame capture, communication between controller PC and each of the 9 cameras is facilitated via a network switch.

Most popular view interpolation datasets focus on the task of temporal interpolation for scenes taken with a single hand held camera[6] [1]. Most multi-view datasets are limited to dual camera setups for depth estimation tasks and other stereoscopic effects. Or multi-view datasets for scene reconstruction which are either single camera with no frame ordering and/or not densely sampled enough spatially. This limits the ability to compare classical/deep learning based algorithms [1, 4, 5] against newer dynamic radiance field rendering algorithms [2]. Inspired by time-slice scene acquisition [3] we have developed a multi-view dataset that overcomes these issues. This dataset is available via the following link [2].

Camera Rig Details: Figure 1 shows our multi-view camera rig used to capture the dataset. There are 9 Raspberry Pi's each connected to a Raspberry Pi High Quality Camera with a 12.3 Megapixel ($\approx$ 4k) Sony IMX477 sensor. A controller PC interfaces over a network switch with all Raspberry Pi's to provide gen-lock for synchronised capture and picture transfer to a centralised storage.

The mechanical design of the rig required some thought in order to allow the cameras to be spaced closely. Custom mounts to support the Raspberry Pi's were designed in a CAD package and manufactured using a 3D printer. The board which the camera sensor is attached to is the primary limitation on getting the cameras as close as possible. This board has a width of 38 mm[3] and so this is approximately the distance between the centres of the camera sensors. The camera rig therefore spans a total distance of 34.2 cm.

Dataset Description: We employ 10 objects for our scenes and these are shown in Figure 2. They are shot using a two-point LED lighting setup, against a green screen backdrop. We record the objects at three different distances from the camera *close* (0.75 m), *medium* (1 m), and *far* (1.25) to create varying relative motion caused by parallax.

Post-Processing Pipeline: Each camera records an image with a slightly different colour space. Furthermore, the orientation of the camera plane is often not what was intended. When played back as a sequence, these views show very poor colour smoothness and heavy judder. We therefore implement a number of post-processing steps to correct these distortions using a pipeline in Nuke. First, we remove geometric distortion from each frame. The distortion is estimated by capturing a chequerboard grid with each camera and using grid detection and Nuke's LensDistortion node to estimate parameters for the distortion model. We then perform a colour correction step. It is not necessary to estimate some "exact" colour reproduction of the scene, only to smooth out the differences between cameras as much as possible. Hence we then balance the colour across all cameras using a Calibrite ColorChecker Classic chart paired with the CalibrateMacbeth Blinkscript node developed by Jedediah Smith. We then stabilise the frames using the Tracker node in Nuke. As a final step, we crop our scene to 1080p resolution and then down-sample this to create a second version with a resolution of 720p.

[1] https://media.xiph.org

[2] https://drive.google.com/drive/folders/1J7QdGFcYw_AAAO6U9TNgBvTJwNCTYUo-?usp=sharing

[3] https://datasheets.raspberrypi.com/hq-camera/hq-camera-cs-mechanical-drawing.pdf

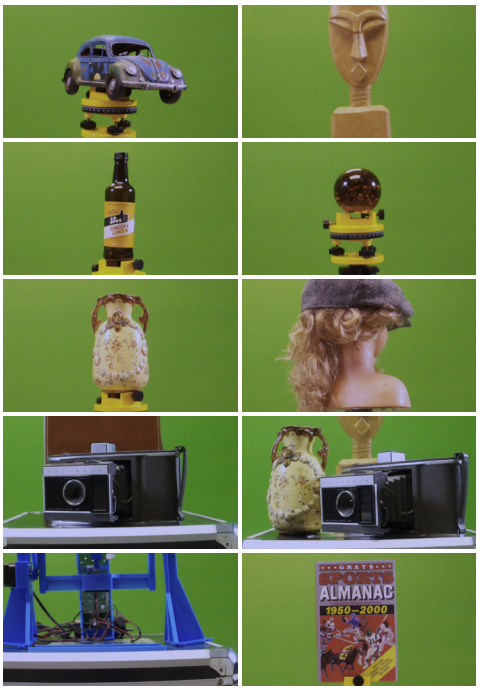


Figure 2: 1080p images of real dataset objects at a medium distance, captured using our multi-view rig.

Future Work: In this work we only consider static scenes as a first iteration of this dataset and intend to add dynamic scenes in the next iteration. Along with this we aim to perform a thorough statistical analysis and comparison of frame interpolation and view synthesis algorithms against one another.

Acknowledgements: The research conducted in this publication was funded by the Irish Research Council under grant number EPSPG/2023/1515 and OvercastHQ.

[1] Duolikun Danier, Fan Zhang, and David Bull. ST-MFNet: A spatio-temporal multi-flow network for frame interpolation. In *CVPR*, 2022.

[2] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023.

[3] Markus Kettern, David C. Schneider, Benjamin Prestele, Frederik Zilly, and Peter Eisert. Automatic acquisition of time-slice image sequences. In *CVMP*, 2010.

[4] Anil Kokaram, Davinder Singh, and Simon Robinson. A Bayesian View of Frame Interpolation and a Comparison with Existing Motion Picture Effects Tools. In *ICIP*, 2020.

[5] Simon Niklaus, Long Mai, and Feng Liu. Video Frame Interpolation via Adaptive Separable Convolution. In *ICCV*, 2017.

[6] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017.

MAX-R Demo: Digital Location

Jonas Trottnow
https://animationsinstitut.de/en/studies/lecturers/jonas-trottnow
Simon Spielmann
https://animationsinstitut.de/en/simon-spielmann
Volker Helzle
https://animationsinstitut.de/en/studies/lecturers/prof-volker-helzle
Alexandru-Sebastian Tufis-Schwartz
https://www.linkedin.com/in/alexandru-sebastian-tufis-schwartz-019457131/

Research & Development department at
Animationsinstitut of Filmakademie Baden-Württemberg

Digital Locations are representations of potential film locations existing in reality. They are digital twins created through 3D scanning or similar approaches. Currently most potential film locations are offered to location scouts as images and textual descriptions. This often makes it difficult to decide for a location and plan a production at the location. Usually travelling to the location is needed to define places for equipment (e.g. trucks & generators), plan shots (e.g. perspectives, lighting etc.) making the process of location scouting tedious, time consuming and in some cases CO2 emissive. With digital locations this can be simplified as many of the decisions and plans can be laid out utilising the digital twin of the location, given that software tools offer intuitive and feature rich possibilities to work with it. The technology developed within Max-R offers great potential to enhance this. In addition this approach offers improved sustainability supporting the Green Deal by minimising the need for travel.

Within the Max-R project TRACER has been extended to meet the requirements of a Digital Location pipeline. TRACER is a software agnostic communication infrastructure and toolset for plugging open-source tools into a production pipeline, establishing interoperability between open source and proprietary tools, targeting real-time collaboration and XR productions, with an operational layer for exchanging data objects and updates, synchronization of scene updates of different client applications (Blender, UE, Unity, VPET ...). VPET, an exemplary implementation of an XR tool using TRACER foundation as base, has been advanced for Digital Locations. VPET is a tablet-based, collaborative tool that allow real-time on-set light, asset and animation editing via an intuitive interface. It provides functionality to edit assets and synchronize changes between different VPET or TRACER clients. The requirements and possibilities for Digital Locations have been developed together with the Film Commission Stuttgart, a local authority maintaining locations in Baden-Württemberg (Germany) and the film company Third Picture.

Figure 1: Digital Location demo website with embedded wgpuEngine viewer and QR code for VPET

Within the Digital Location demo a 3D digital location can be displayed

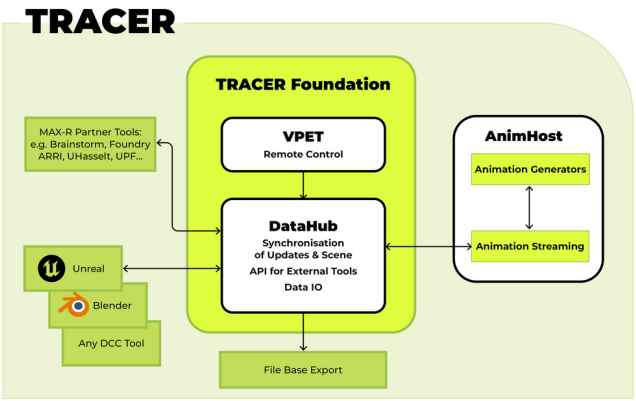


Figure 2: Overview of TRACER

interactively in a browser using wgpuEngine developed by UPF within Max-R. Camera perspectives can be changed and the location can be virtually inspected directly on a website. The scene can then be loaded directly from the website into VPET by scanning a QR code. The scene can then be explored and edited in AR on a tablet. DataHub, being part of TRACER, synchronises VPET clients working on the same Digital Location simultaneously. Editable assets can be additional set elements, props and production equipment or moveable objects on location. This makes it possible for location scouts, producers, the set dressing department, production designers, grip, ... to plan scenes, set builds and logistics remotely with the intuitive tablet based interface VPET offers, without the requirement of knowing how to operate 3D software. Lighting department and directors of photography (DoP) can plan lighting setups as well as camera placements and movements digitally with the actual location.

All components can be obtained from our GitHub repositories[1]. TRACER

Figure 3: VPET with digital location in AR

is developed within the Max-R project which received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101070072.

[1] https://github.com/FilmakademieRnd

MAX-R Demo: AnimHost

Jonas Trottnow
https://animationsinstitut.de/en/studies/lecturers/jonas-trottnow
Simon Spielmann
https://animationsinstitut.de/en/simon-spielmann
Volker Helzle
https://animationsinstitut.de/en/studies/lecturers/prof-volker-helzle
Francesco Andreussi
https://www.linkedin.com/in/francesco-andreussi
Simon Haag
https://www.linkedin.com/in/simonhaag2094/

Research & Development department at
Animationsinstitut of Filmakademie Baden-Württemberg

With technologies like marker-less, video-based motion capturing and AI-generated character animation, pipelines for animated movies are transforming. Movie productions utilising game engines for rendering demand for interactive and real-time animation directing capabilities that can be driven by artists and directors. While e.g. AI-based human character animation generators exist in research, their applied usage in animated movie production is sparse. Integrating such solutions in industry standard DCC applications, game engines etc. is a lengthy process, especially since the interactive nature and the demand for artist controllability asks for new user interfaces and pipelines. AnimHost is addressing these challenges.

AnimHost is part of the TRACER ecosystem. TRACER is a software agnostic communication infrastructure and toolset for plugging open-source tools into production pipelines, establishing interoperability between open-source and proprietary tools, targeting real-time collaboration and XR productions, with an operational layer for exchanging and synchronising scene data and updates between different client applications (Blender, UE, Unity ...).

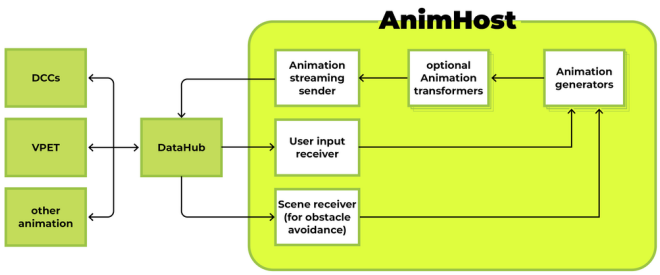


Figure 1: Overview of AnimHost

AnimHost connects animation generators (such as AI deep neural networks trained on motion capturing databases, video-based, low-cost motion capturing, ...) to DCC applications, on-set tools or renderers in general. It is functionally independent of the animation receiving app and provides an intuitive interface to support new solvers, with a focus on real-time scenarios. AnimHost addresses compatibility issues between generators and end users by offering a defined API and data structure for sending human bone animations into arbitrary receivers. Generators are integrated into AnimHost as PlugIns. A simple node graph editor is used as a user interface to set up the animation pipeline within AnimHost. Generators get connected to the sending node, while needed transformations or eventually re-targeting can be added in the middle. Being part of TRACER, automatic generated or captured animations can be sent to multiple devices simultaneously through DataHub. Interactive character control through a spline based walking path can be authored e.g. with Blender, which is also able to receive the streamed animations. Thereby it is possible to generate animations automatically and permit interactive directing/blocking by non-professionals and experts alike. The development is engine-agnostic, allow high-level interaction with a character in real-time, and be adaptable to any future DCC application.

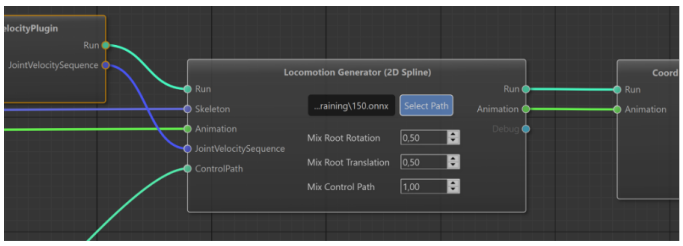


Figure 2: Screenshot of node based AnimHost UI

The demo will showcase the pipeline provided by AnimHost and the TRACER framework based on assets from the testproduction "Survivor". Walking paths for a character can be defined as a spline by artists in standard DCC tools like Blender. AnimHost then immediately generates bone animations locally that are then applied in the original DCC application. The animated character animation can then be refined with a post-processing pipeline using classical animation tools like control rigs, inverse kinematics and forward kinematics. Being an open-source framework, TRACER, which includes AnimHost, also provides rich possibilities for future advancements through the modular plugin interfaces.

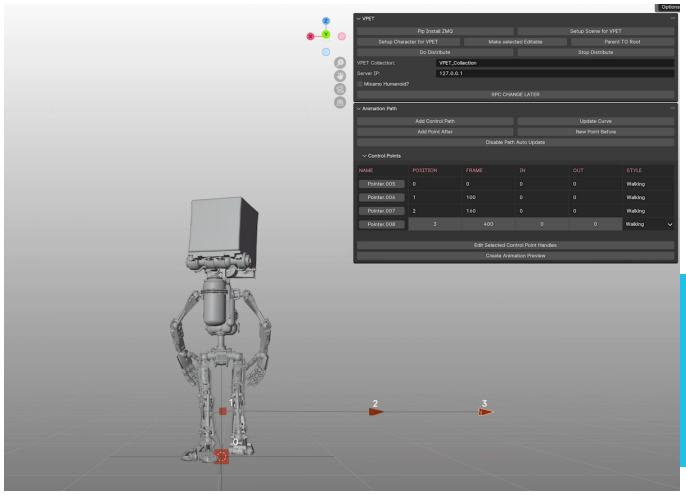


Figure 3: Animation authoring interface as plugin in Blender

All components can be obtained from our GitHub repositories[1].

Figure 4: Screenshot of the animatic of the 'Survivor' testproduction

TRACER and AnimHost are developed within the Max-R project which received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101070072.

[1] https://github.com/FilmakademieRnd

Demonstrating Causal-Temporal Narrative Video Captioning using NarrativeBridge

Asmar Nadeem[1,2]

Mahrukh Awan[1,2]

Armin Mustafa[1,2]

[1]SAIReco Ltd., United Kingdom

[2]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom

contact@saireco.org

We present a live demonstration of NarrativeBridge [3], a novel framework that enhances video captioning by incorporating Causal-Temporal Narrative (CTN) understanding. NarrativeBridge addresses a critical limitation in existing video captioning models and benchmarks: the lack of coherent representations of causal and temporal relationships in video content. Our approach consists of two key components:

1. **CTN Captions Benchmark:** A novel dataset of causal-temporal narrative captions generated via few-shot prompting, encoding cause-effect relationships in videos.

2. **Cause-Effect Network (CEN):** An architecture with dedicated cause and effect encoders for learning and generating contextually rich video descriptions.

NarrativeBridge significantly outperforms existing state-of-the-art methods in articulating the causal and temporal aspects of video content, demonstrating improvements of 17.88 and 17.44 CIDEr points on the MSVD and MSR-VTT datasets, respectively.

1 Demo Overview

Attendees will interact with two main components:

1. **CTN Caption Generator:** Visitors can input video clips and observe the real-time generation of CTN captions using our approach, as illustrated in Figure 1. This figure illustrates the CTN caption generation process, which forms the basis of our NarrativeBridge system. Attendees can observe this process during the demo.

2. **CEN Caption Showcase:** A display of captions generated by our CEN model (Figure 2) trained on our novel CTN benchmark, highlighting the model's ability to capture complex causal-temporal narratives in various video scenarios (Figure 3). Figure 2 shows the CEN architecture, which enables effective learning and generation of captions with causal-temporal narratives. Figure 3 illustrates NarrativeBridge's performance across diverse video content.

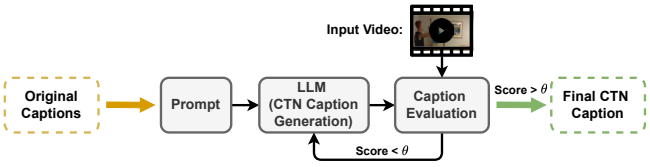


Figure 1: CTN caption generation pipeline.

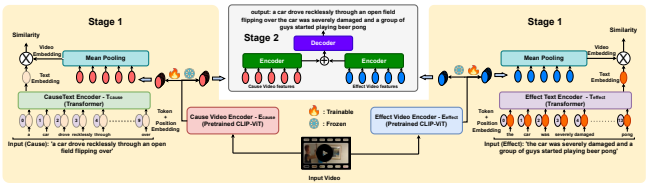


Figure 2: The two-stage Cause-Effect Network (CEN) architecture.

2 Interactive Features

Our demo station offers:

- A touchscreen interface for uploading or selecting sample videos
- Live generation and display of captions from our CEN model trained on the CTN benchmark
- Side-by-side comparison of CEN-generated captions with baseline methods [1, 2, 4] and ground truth CTN captions
- Interactive exploration of CEN performance across various video types and scenarios

3 Demonstration Scenarios

Visitors can explore NarrativeBridge through various scenarios, for example, as shown in Figure 3, all featuring captions generated by our CEN model:

1. **Video Game Analysis:** Generate CTN captions for game footage, capturing complex action sequences and their outcomes.

2. **Sports Highlight Understanding:** Analyze sports clips, demonstrating the system's ability to capture cause-effect relationships in dynamic scenes.

CTN caption GT: 'a player performs a fatality move in mortal kombat another character is killed in the game'

CEN (OURS): 'a video game character performs a fatality move in mortal kombat video game killing another' **SEM-POS:** 'a player is playing a game' **AKGNN:** 'freddy krueger is playing a video game and freddy krueger in a video game featuring freddy krueger and freddy krueger and freddy krueger in a video game' **GIT:** 'a boy is playing a video game'

(a)

CTN caption GT: 'a soccer player kicked the ball with precision the ball successfully went into the goal'

CEN (OURS) : 'a soccer player kicked the ball with precision the ball went into the goal scoring a point' **SEM-POS:** 'a player is kicking a ball' **AKGNN:** 'the soccer player scored a goal and the team scored a goal' **GIT:** 'soccer players are playing soccer'

(b)

Figure 3: Qualitative examples of NarrativeBridge performance across different scenarios, showing ground truth CTN captions and our CEN model's generated captions.

4 Technical Setup

Our demo runs on a high-performance GPU workstation, featuring:

- Real-time video processing and CEN-based caption generation
- A display for showcasing results, including side-by-side comparisons of CEN-generated captions with baselines and ground truth
- Touchscreen tablet for attendee interaction and scenario selection

We invite CVMP attendees to experience NarrativeBridge and explore its potential in advancing video understanding and captioning technologies across a wide range of applications in visual media production.

[1] Willy Fitra Hendria, Vania Velda, Bahy Helmi Hartoyo Putra, Fikriansyah Adzaka, and Cheol Jeong. Action knowledge for video captioning with graph neural networks. *Journal of King Saud University-Computer and Information Sciences*, 35(4):50–62, 2023.

[2] Asmar Nadeem, Adrian Hilton, Robert Dawes, Graham Thomas, and Armin Mustafa. Sem-pos: Grammatically and semantically correct video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2023.

[3] Asmar Nadeem, Faegheh Sardari, Robert Dawes, Syed Sameed Husain, Adrian Hilton, and Armin Mustafa. Narrativebridge: Enhancing video captioning with causal-temporal narrative. *arXiv preprint arXiv:2406.06499*, 2024.

[4] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Advancements in Networked Pipelines for Virtual Production

Philip Coulam-Jones
philip.coulam-jones@disguise.one
Kenneth Leung
kenneth.leung@disuise.one
Josh McNamee
josh.mcnamee@disguise.one
Disguise Technologies

Augmented and extended reality media is a growing area of media production which requires the development of new tools and new pipelines, as well as the extension and adaptation of existing ones. Since September 2022, Disguise has been involved in a Europe-wide consortium project called MAX-R (Mixed, Augmented and eXtended Reality Media Pipeline) dedicated to building new production pipelines for a variety of hybrid media styles, including XR, VR and MILEs. We have developed a number of cutting-edge prototypes that will enable the next generation of hybrid productions. We are aiming to put together a demonstration session that allows us to highlight some of these prototypes, with an opportunity to exhibit them to interested experts and answer any questions that they may have about these emerging technologies.

Our focus for a CVMP presentation will be demonstrating two projects: RSConnect, a tool for exporting audiovisual and metadata out of the Designer instance running on set, and Remote Reimaging, our new workflow for updating and controlling rendering servers in use on a production. As well, we will be available to discuss our other MAX-R related work including our new Depth Reprojection algorithm for VP[1], along with updates to our system to incorporate state-of-the-art OpenColourIO colour management options and broaden API control of on-stage virtual content.

1 RenderStream Connect

RenderStream Connect (RSConnect) is a prototype library we have developed for next-generation data interchange on virtual production sets or extended reality (XR) stages. The design is intended to allow our networked show coordination software, Designer, to output to an ecosystem of arbitrary modern software in a manner that compliments our existing use of HTTP APIs. RSConnect incorporates network discovery, capability advertisement and data format negotiation to allow the export of audio-visual data and metadata from Designer to a wider range of endpoints than has previously been possible. For MAX-R we have been considering various use cases including an on-set dedicated shot archiving device, a confidence-monitoring platform, and functionality for pushing audio and video directly from an XR stage into the metaverse.

RSConnect makes use of open standards for communication and serialisation, specifically mDNS/DNSSD and gRPC. For audio-visual interchange we targeting using RTP, RTMP and WebRTC. By utilising these methods our aim is for RSConnect to form a generic protocol that can be used independently of the Designer software.

A case study for the RSConnect system has already been developed in collaboration with MAX-R partners at the BBC and Improbable Worlds, which involves streaming a live performer on an LED stage recorded by the BBC, with visuals and capture coordinated using Disguise's Designer system, into Improbable's metaverse platform via RSConnect. This would form an engaging, visually interesting and interactive demo.

The set-up will consist of one camera pointed at a performer on the demonstration stage in our London office, connected to a computer running an instance of our Designer software and the Improbable metaverse platform. We will then live stream the capture of the presenter into the metaverse environment without the need for OBS, allowing for the control of both stream and stage set-up from within the same application. Improbable's platform will be available to attendees on their mobile devices, accessed via a simple sharing link in the form of a QR code, as well as displayed on-screen for the audience to view.

2 Remote Reimaging

OS management and system configuration on an enterprise level typically uses cloud-based solutions like Windows Deployment Services[2] and Orel Cloud Image Management Service[3]. These solutions allow centralised and remote control of large server systems without direct intervention. In the virtual production industry however, these technologies have not yet been made available for managing hardware which may have specific physical dependencies. Operators of those systems have traditionally relied upon USB drives to deploy OS images onto individual servers.

Figure 1: Capture of live performers on an LED stage being composited into Improbable's metaverse platform via RenderStream Connect

With our new reimaging workflow, we have made steps to move towards emulating the cloud-based solution by utilising the local network as the OS deployment method. The process is managed centrally through a webapp, which communicates with servers with new HTTP API endpoints. Via this interface the user can also extensively customise server configuration for their specific project.

Our demonstration will consist of a pre-recorded video walkthrough simulating a scenario where a customer is renting a set of rendering RX servers hosted in the cloud, centrally managed by Disguise. After they finish using the servers, Disguise operators use the webapp to remotely clean the servers' Media drives to protect the customer's data. The webapp then automatically transitions to re-image the servers to a different Operating System. When the re-image process is complete, the operator and applies the machine configuration for the next customer. The customer then starts up a remote rendering workload on another server, using the reconfigured RX servers as render nodes. This demonstrates the reconfiguration of servers without manual intervention, captured on video to streamline the process and highlight the key steps to attendees.

3 Conclusion

Demoing RSConnect and Remote Reimaging in this setup will allow for a great space for interested industry parties to learn more about our work on the MAX-R project over the course of the past two years. We hope that CVMP attendees find the session engaging and that it generates a wealth of ideas for how to implement these prototypes in their own projects once they've made their full release.

[1] S. Day D. Brown and T. Whittock. Depth reprojection for mitigating latency in xr media production (short paper). In *CVMP*, 2024.
[2] Microsoft. Windows deployment services. https://learn.microsoft.com/en-us/previous-versions/windows/it-pro/windows-server-2012-r2-and-2012/hh831764(v=ws.11).
[3] Orel. Image management service. https://www.orelcloud.com/image-management-service/.

EnVisualAIzer – Explorable AI-Generated Environments for All Filmmakers in Virtual Production

Pauline Leininger
p.leininger@hff-muc.de

University of Television and Film Munich
Ludwig-Maximilian-University Munich

The evolution of artificial intelligence (AI) is transforming virtual film production (VP), providing filmmakers with tools for immersive scene creation and previsualization. AI-generated environments lower barriers for small teams by enabling rapid, cost-effective scene design without physical sets. Technologies like generative models, Neural Radiance Fields (NeRF) [3], and 3D Gaussian Splatting (GS) [2] are increasingly promising in film production, though their use is often limited to visual effects (VFX) experts. *EnVisualAIzer* addresses this by making AI-generated environments accessible to filmmakers from all departments.

EnVisualAIzer Prototype

EnVisualAIzer aims to integrate AI-generated environments into all phases of virtual production, from ideation to final production, enabling playful interaction and experimentation. Built with Unreal Engine 5, it is designed to familiarize filmmakers with the possibilities of AI-generated environments in a way that is easy to use, even for those without technical expertise. The prototype features an evaluation ground for various environment types, including Stable Diffusion Depth Mesh [4], 360° Panoramic Images (Blockade Labs 3D [1]), and Gaussian Splats (Volinga AI [5]), enabling assessment of their effectiveness in VP workflows.

Key Features

- **Dynamic Scene Exploration:** Users can freely navigate and interact with AI-generated environments, adjusting camera angles, placing virtual actors, and exploring scene compositions in real-time.
- **Object and Actor Placement:** With drag-and-drop functionality, users can place characters and objects in the scene, adjusting their positions for previsualization or scene planning.
- **Real-Time Annotations and Sketching:** Filmmakers can annotate environments or add sketches directly within the tool to communicate ideas or make quick notes for later reference.
- **Previsualization and Camera Controls:** The tool includes a Previs Camera Mode, which enables customizable camera settings, allowing users to capture stills or plan camera movements within the environment.
- **On-the-Fly Image2Image Generation:** Integrated with Stable Diffusion API, *EnVisualAIzer* allows users to generate or modify scene images directly, enabling dynamic adjustments and quick visual changes.

Figure 1: Overview of the main views of the *EnVisualAIzer* prototype: (1) Sample project selection, (2) the explorable scene with note and character placement, (3) the camera previs mode, and (4) a detailed view of a saved scene showcasing an Image2Image adjustment.

Figure 2: Overview of the six demo environments: three AI environment types (1) Depth Mesh, (2) Blockade 360° Panoramic Images and (3) Volinga Gaussian Splat, in an *indoor* and *outdoor* setting each.

Evaluation and Outlook

A user study with 15 film professionals from various departments showed that *EnVisualAIzer* is effective in making AI-generated environments accessible, particularly for previsualization and scene ideation. Gaussian Splatting was favored for high-fidelity tasks, while Depth Mesh and Blockade 3D proved valuable for quick concept development. The tool's accessibility allowed non-technical users to engage in tasks like storyboarding and collaborative scene design. Based on this feedback, future work will focus on adding real-time customization, dynamic scenes, and AR/VR integration to enhance collaboration and environment fidelity.

[1] Blockade Labs. Blockade labs | 360° panorama generator. Online, 2024. URL https://www.blockadelabs.com. Accessed: September 5, 2024.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering, 2023. URL https://arxiv.org/abs/2308.04079v1.

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020. URL http://arxiv.org/abs/2003.08934. arXiv:2003.08934 [cs].

[4] Thygate. Stable diffusion webui depth maps extension. GitHub Repository, 2023. URL https://github.com/thygate/stable-diffusion-webui-depthmap-script. Available: https://github.com/thygate/stable-diffusion-webui-depthmap-script Accessed: September 5, 2024.

[5] Volinga AI Technologies. Homepage of volinga ai. Online, 2024. URL https://volinga.ai. Accessed: September 5, 2024.

The Changing Forest: managing mass personalisation via audio-visual media flexibilities

Craig Cieciura

Maggie Kosek

Elettra Bargiacchi

Philip J. B. Jackson

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

https://www.surrey.ac.uk/centre-vision-speech-signal-processing

This demo presents *The Changing Forest*[1], an interactively personalisable pre-authored audio-visual experience. produced by combining inter-compatible adaptations that reflect a diverse set of personalisation types. We describe the overall approach to the experience's production, as well as the workflows employed by the design team to achieve the goals of type of personalisation. The result is presented as a short piece of audio-visual content with a coherent narrative that can switch across the combinations of personalisation settings. We reflect on the benefits and potential of this approach for large-scale personalisation while maintaining significant editorial control with favourable scalability.

When thinking about the kind of adaptation needed to make content suitable for a particular individual, it is not enough to offer the choice of one kind of accessibility or another, or one kind of user choice alone. In reality, each user may have a complex set of needs and desires to be addressed. Accordingly, the option to improve or adapt material in one sense or another should not preclude all other choices. In this spirit, the present demonstration develops a diverse set of personalisation types each of which may be combined with other choices, thus offering a combinatorial explosion of media experiences by the selection of a few limited choices. In this study, researchers explored readily available mechanisms available to design teams during postproduction in order to create alternative versions suitable for inter-compatibility.

The content selected was based on a previous asset base of object-based media developed to showcase the capabilities of spatial audio and virtual reality. The original content source is an audio-lead narrative of a children's fantasy set in a forest in Autumn [2], called The Turning Forest[2]. In this new variation, which we call The Changing Forest, a shortened cut of that story was created as a default or standard version, considering a standard landscape 16:9 visual aspect ratio, a stereo audio playback, using the English-speaking narrator to make a 71-second vignette. As with the original VR experience, for this piece, we first explored the auditory modality, developing range of different audio mixes across different types of personalisation [1]. Here we also developed alternative visual presentations of the content based on 2D layers containing the original components but again combining them in different ways to achieve significant variations. This design of the set of personalisation types was founded on an analysis of user needs across the existing UK population, suggesting that certain sectors of the population could be better served by these different types of intervention. In particular, the media was adapted in terms of: duration, language, style, clarity, and the user's reproduction or playout system. A snapshot showing how the visual content changed across these alternative versions is illustrated in Figure 1.

The resulting combinations of alternative variations along these different dimensions of personalisation lead to a large number of possible renders, much larger than may have been produced practically if done individually. However by developing these alternative versions along orthogonal axes, the user's selections could be combined to produce the integrated alternative version that best meets their needs and desires (see Figure 2). We believe the potential of this production approach is that a media production organisation can effectively provide a wide range of versions to address the needs and preferences of different audience segments, including intersectional demographics, achieving greater audience coverage without a vast increase in the production costs.

This interactive audio-visual media experience demonstrates a potentially viable approach to the production of flexible media for mass personalisation by exploiting the compatibility of a diverse set of flexibilities. This short piece exemplifies the user value in each of the personalisation adaptations that it offers, meanwhile indicating ways in which the versatility of that personalisation may be allowed to grow incrementally, towards individually personalised experiences for large-scale regional, national or international audiences. In future work, we wish to evaluate the selection of personalisation types, quantify the gains in value with users, and prioritise the defined options. Trials with other pieces and genres of content may be envisaged to test this concept of incompatible flexibilities with such materials for different programme formats. Finally, we would like to explore the benefits of applying each type of personalisation to additional media types, for example graphic overlays, captions and other automatically generated content.

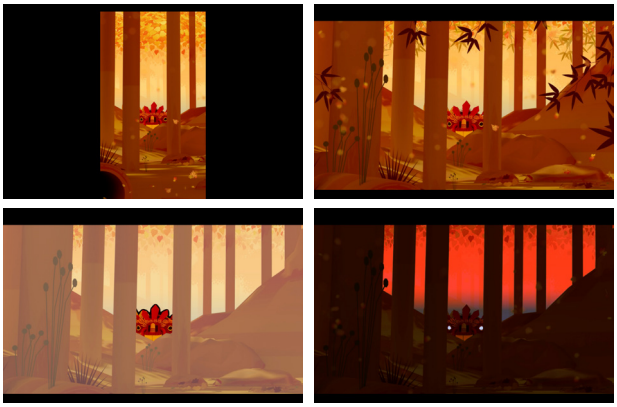


Figure 1: Visual comparison of alternative versions at 0:26 (clockwise from top left): for 9:16 mobile phone screen, regionalised for Asia, clarity and teenage horror versions.

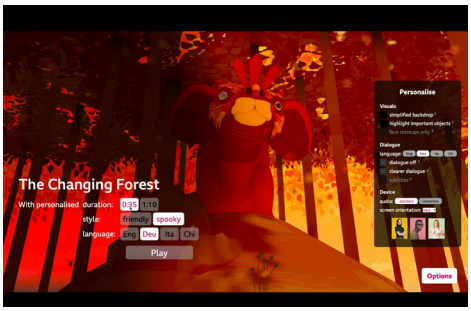


Figure 2: Interactive interface offering the user options to personalise The Changing Forest according to a diverse set of options, relating to physical and sensory accessibility needs, as well as more hedonic preferences.

[1] C. Cieciura, E. Bargiacchi, and P. J. B. Jackson. Authoring inter-compatible flexible audio for mass personalization. In *Audio Engineering Society*, New York, NY, Oct. 2024.

[2] James Woodcock, Chris Pike, Frank Melchior, Philip Coleman, Andreas Franck, and Adrian Hilton. Presenting the S3A Object-Based Audio Drama Dataset. In *The 140th Audio Engineering Society Convention*, May 2016.

[1]This work was supported by UKRI Pulse funding to the EPSRC-BBC Prosperity Partnership 'AI4ME: Future personalised object-based media experiences delivered at scale anywhere' (EP/V038087/1). Thanks to our AI4ME colleagues, especially Blaise Galinier, Maxine Glancy and Joannar Risznica for help in development of the design for the user interface.

[2]https://www.bbc.co.uk/taster/pilots/turning-forest

AI-Driven Real-Time Visualization of Music

Jenny Huang
j.huang@hff-muc.de
University of Television and Film Munich
LMU Munich

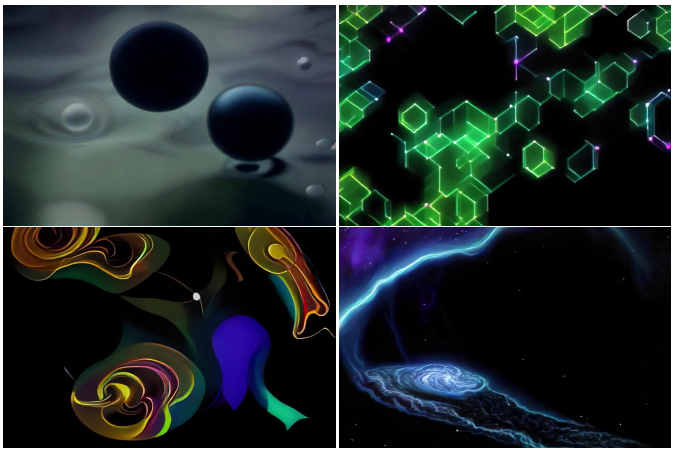


Figure 1: Music Visualizations Generated With Our Prototype

Music visualizations are visual representations or interpretations of music that often dynamically respond to audio. They can create more immersive and captivating music experiences, while also making them more accessible to individuals with hearing loss.

However, most existing real-time music visualization systems tend to fall into two distinct categories. The first category focuses heavily on the informative aspect, offering data-driven visualizations that typically rely on symbolic representations like MIDI data (e.g. Hiraga et al. [7], Fonteles et al. [5], Pouris and Fels [9]). While informative, these visuals may lack appeal for non-musicians and can be seen as boring, as shown by feedback from deaf and hard-of-hearing (DHH) individuals [6]. The second category prioritizes aesthetics, producing visually pleasing but largely preprogrammed patterns. This approach, seen in visualizers integrated into media players, can become repetitive over time and fail to capture more complex musical aspects such as mood or emotion.

To address the shortcomings of both approaches, we developed an advanced system that combines several AI technologies to create dynamic, meaningful visualizations in real-time (examples see Figure 1). Our system generates visuals that are not only audio-reactive but also reflect the emotional and aesthetic dimensions of the music, delivering a richer, multisensory music experience.

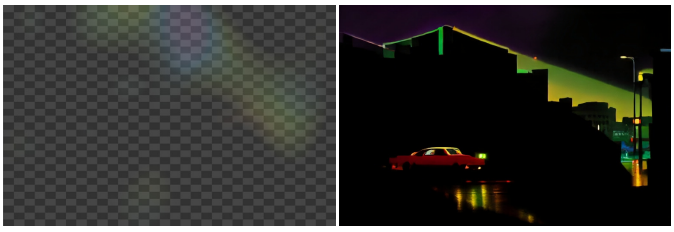


Figure 2: Audio-Responsive Noise Animation (left) and Generated Image (right) using the generated prompt "A retro-futuristic cityscape bathed under an orange sunset-sky, twisting vinyl records as building roofs, neon pink and sapphire streetlights dancing along viridian cobblestone pathways. Silver chrome retro cars hover over magenta-coated roads; Reflections of teal and lavender in shop windows, Quiet pulse of golden light from rooftop lounges, An opulent, yet approachable urban vista of smooth funk and vibrant"

Building on a comprehensive review of related literature and preliminary interviews with DHH individuals, we established the objectives for our prototype and ultimately designed a music visualization system that integrates Music Information Retrieval (MIR), Large Language Models (LLMs), and Image Generation Models.

To develop the core architecture of the prototype, we utilized TouchDesigner [1], a node-based visual programming language known for its versatility in interactive media creation.

For image generation, we employed the diffusion pipeline StreamDiffusion [8] alongside the "sd-turbo" model [4] by StableDiffusion. This configuration allowed us to achieve 5-6 frames per seconds at a resolution of 768x512 pixels on a consumer-grade gaming laptop, with even higher framerates on more powerful hardware.

The image generation process in our system requires two key inputs: a reference image and a text prompt.
The reference image is an animated noise texture, with its parameters dynamically controlled by low-level audio features, such as frequency amplitudes, extracted from the audio signal. This ensures that the visuals respond in real time to changes in the audio.
The text input, on the other hand, is repeatedly generated based on high-level musical information analyzed over time. This includes details such as the genre, mood, and instrumentation. For these MIR tasks, we used pre-trained models from Essentia [2]. The classified musical information is then processed by an LLM — in this case, GPT-4-turbo via the OpenAI Chat Completions API [3] — which is instructed to interpret the given musical data and create a visual elaboration that serves as the prompt for the image model.

Figure 2 shows an audio-responsive noise reference image alongside the final visualization produced from it, using the generated text prompt.

Our real-time music visualization system can enhance music events, such as concerts, festivals, and nightclubs, and paves the way for future research focused on developing immersive and inclusive music experiences.

[1] Touchdesigner by derivative. https://derivative.ca/. (Accessed on 05/10/2024).

[2] Essentia website. https://essentia.upf.edu/. (Accessed on 05/23/2024).

[3] Text generation - openai api. https://platform.openai.com/docs/guides/text-generation. (Accessed on 05/10/2024).

[4] stabilityai/sd-turbo · hugging face. https://huggingface.co/stabilityai/sd-turbo. (Accessed on 05/31/2024).

[5] Joyce Horn Fonteles, Maria Andréia Formico Rodrigues, and Victor Emanuel Dias Basso. Creating and evaluating a particle system for music visualization. *Journal of Visual Languages & Computing*, 24 (6):472–482, December 2013. doi: 10.1016/j.jvlc.2013.10.002.

[6] David W. Fourney and Deborah I. Fels. Creating access to music through visualization. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, pages 939–944. IEEE, September 2009. doi: 10.1109/TIC-STH.2009.5444364.

[7] Rumi Hiraga, Reiko Mizaki, and Issei Fujishiro. Performance visualization: a new challenge to music through visualization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 239–242. ACM, December 2002. doi: 10.1145/641007.641054.

[8] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. StreamDiffusion: A pipeline-level solution for real-time interactive generation. URL http://arxiv.org/abs/2312.12491.

[9] Michael Pouris and Deborah I. Fels. Creating an Entertaining and Informative Music Visualization. In *Computers Helping People with Special Needs*, volume 7382, pages 451–458. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-31522-0_68. Series Title: Lecture Notes in Computer Science.

TECHNICAL AWARDS

Here are the CVMP Technical Awards:

Research Impact Award

Awarded to individual(s) who have performed key research that has later been taken by other third parties and used effectively in media production or product.
Examples include:

- An author of a paper that has subsequently been used by a third party at a media company to develop a set of tools or effect used on production media.
- This award is for the subsequent impact a research/paper has achieved.

Collaboration Award

Collaboration award represents joint effort on a dedicated project between individuals jointly across both academia and industry.
Examples include:

- Where an academic has been based at a company to create a tool used directly on production(s).
- Grant based collaborations between academia and industry for media.

Implementation Award

Awarded to individual(s) who have taken academic research and then have pioneered in a timely manner a tool that implements that research in a novel way.
Examples include:

- Taking research and applying it in an original way to solve a problem that was not originally foreseen as an application.
- Creating a tool that applies underlying research that provides a level of artistic control beyond the current state-of-the-art effective for media production

NOTES

NOTES

CHAIRS

Conference Chairs
Armin Mustafa, University of Surrey, UK
Hansung Kim, University of Southampton, UK

Full Papers Chair
Claudio Guarnera, University of York, UK

Short Papers & Demos Chair
Peter Eisert, Humboldt University, Germany
Peter Vangorp, Utrecht University, Netherlands

Industry Chair
Oliver Grau, Intel, Germany
Sara Coppola, UK

Sponsorship Chair
Marco Volino, University of Surrey, UK

Awards Chair
Jeff Clifford, Evastute/Milk VFX, UK

Local Arrangements Chair
Changjae Oh, Queen Mary University London, UK

Public Relations Chair
Da Chen, University of Bath, UK

Social Chair
Violeta Menendez, University of Surrey, UK

Conference Secretary
Elizabeth James, University of Surrey, UK

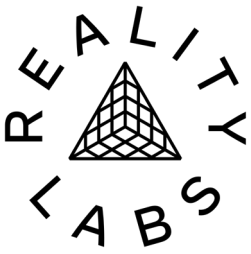
Programme Committee

Da Chen, University of Bath
Peter Eisert, Humboldt University
Dar'ya Guarnera, University of York
Hansung Kim, University of Southampton
Gilles Rainer, Imperial College London
Kenny Mitchell, Edinburgh Napier University
Christian Richardt, Meta Reality Labs Research
Zhidong Xiao, Bournemouth University
Peter Vangorp, Utrecht University
Daljit Singh Dhillon, Clemson University
Andrew Gilbert, University of Surrey
Craig Kaplan, University of Waterloo
Rafal Mantiuk, University of Cambridge
Nadejda Roubtsova, University of Bath
Graham Thomas, BBC
Moira Shooter, University of Surrey
Erik Reinhard, Technicolor

Steering Committee

Neill Campbell, University of Bath
John Collomosse, University of Surrey
Oliver Grau, Intel
Anil Kokaram, Trinity College Dublin
Jeff Clifford, Evastute/Milk VFX
Abhijeet Ghosh, Imperial College London
Peter Hall, University of Bath
Will Smith, University of York

Conference Sponsors 2024



RESEARCH

People-Centred AI
UNIVERSITY OF SURREY

CAMERA
Centre for the Analysis of Motion,
Entertainment Research and Applications

ACMSIGGRAPH

Published by ACM

https://dl.acm.org/conference/cvmp