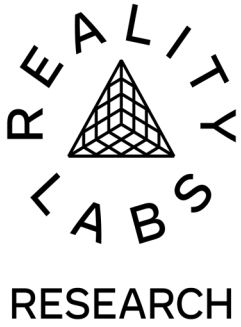
CVMP2023

The 20th ACM SIGGRAPH European Conference on Visual Media Production
30th November - 1st December 2023
BFI Southbank, London, UK

Programme

British Film Institute (BFI) Southbank
30th November & 1st December 2023
https://www.cvmp-conference.org/2023/

Conference Sponsors 2023



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

Message from the Chairs

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

Finally, we would like to thank everyone who submitted to CVMP this year, the invited speakers, the reviewers, our sponsors, and the organising committee for their hard work in bringing CVMP 2023 together!

Marco Volino and Armin Mustafa (Conference Chairs)
Peter Vangorp (Full Papers Chair)
Peter Eisert and Claudio Guarnera (Short Papers Chairs)
Oliver Grau and Abi Bowman (Industry Chairs)
Jeff Clifford and Peri Friend (Sponsorship Chairs)
Hansung Kim (Local Arrangements)
Da Chen and Moira Shooter (Public Relations)
Emily Ellis (Conference Secretary)

DAY 1 | Thursday 30th November 2023

Location: BFI Southbank

09:00 Registration opens | BFI Bar

09:20 Chairs' Welcome | Marco Volino, University of Surrey
Armin Mustafa, University of Surrey

09:30 SESSION 1 | Image-based AI

1. HDR Illumination Outpainting with a Two-Stage GAN Model
   Jack Hilliard, Adrian Hilton, Jean-Yves Guillemaut
2. One-shot Detail Retouching with Patch Space Neural Transformation Blending
   Fazilet Gokbudak, Cengiz Oztireli
3. A Compact and Semantic Latent Space for Disentangled and Controllable Image Editing
   Gwilherm Lesné, Yann Gousseau, Saïd Ladjal, Alasdair Newson
4. DECORAIT - DECentralized Opt-in/out Registry for AI Training
   Kar Balan, Alex Black, Andrew Gilbert, Simon Jenni, Andy Parsons, John Collomosse
5. Expression-aware video inpainting for HMD removal in XR applications
   Fatemeh Ghorbani Lohesara, Karen Eguiazarian, Sebastian Knor

11:10 Coffee Break | Foyer
Poster presenters put up posters

11:40 KEYNOTE 1 | Andrew Whitehurst, Industrial Light and Magic

12:40 SPOTLIGHT SESSION

13:00 POSTERS, DEMO AND LUNCH | BFI Bar

14:30 SESSION 2 | Pose and Motion

6. Optimising 2D Pose Representations: Improving Accuracy, Stability and Generalisability Within Unsupervised 2D-3D Human Pose Estimation
   Peter Hardy, Srinandan Dasmahapatra, Hansung Kim
7. BundleMoCap: Efficient, Robust and Smooth Motion Capture from Sparse Multiview Videos
   Georgios Albanis, Nikolaos Zioulis, Kostas Kolomvatsos

15:10 CVMP AWARDS | Jeff Clifford

15:30 Coffee Break | BFI Bar

16:00 KEYNOTE 2 | Jonathan Starck, Synthesia

17:00 NETWORKING RECEPTION | BFI Bar

19:00 Close

DAY 2 | Friday, 1st December 2023

Location: BFI Southbank

09:00 Registration opens | BFI Bar

09:30 SESSION 3 | Image Processing and Reconstruction

8. Redistributing the Precision and Content in 3D-LUT-based Inverse Tone-mapping for HDR/WCG Display
   Cheng Guo, Leidong Fan, Qian Zhang, Hanyuan Liu, Kanglin Liu, Xiuhua Jiang
9. A software test bed for sharing and evaluating color transfer algorithms for images and 3D objects
   Herbert Potechius, Gunasekaran Raja, Thomas Sikora, Sebastian Knorr
10. LFSphereNet: Real Time Spherical Light Field Reconstruction from a Single Omnidirectional Image
    Manu Gond, Emin Zerman, Sebastian Knorr, Mårten Sjöström
11. View-dependent Adaptive HLOD: real-time interactive rendering of multi-resolution models
    Rui Li

10:50 Coffee Break | BFI Bar

11:20 KEYNOTE 3 | Duygu Ceylan, Adobe Research

12:20 POSTERS, DEMO AND LUNCH | BFI Bar

14:00 INDUSTRY SESSION | Generative AI and XR

12. Applications in Media for Novel View Synthesis
    Graeme Phillipson, Hell Raymond-Hayling, Alia Sheikh
13. Making Gen AI work for us, rather than the other way around
    Will MacNeil
14. Overcoming Latency in Real-Time Virtual Production
    Dominic Brown
15. Beyond Reality - Towards Creative Applications for Mixed Reality
    Thu Nguyen-Phuoc

15:20 Coffee Break | BFI Bar

15:50 KEYNOTE 4 | Tupac Martir, Satore Studio

16:50 Prizes, Announcements and Closing | Marco Volino, University of Surrey
Armin Mustafa, University of Surrey

KEYNOTE 1 | Andrew Whitehurst
Industrial Light and Magic

The Magic of The Man in the Hat: Creating VFX for Indiana Jones and the Dial of Destiny

Thursday 30th November 2023

Andrew Whitehurst served as production Visual Effects Supervisor on Indiana Jones and the Dial of Destiny. The film features around 2350 visual effects shots, created over three years, using a wide range of technologies from traditional to state of the art. In this talk he will breakdown a handful of shots to explain the creative process from initial idea to completion, why certain techniques were selected for these shots, and why to succeed requires the use of the simplest of technologies, a pencil, as well as the most complex.

Andrew Whitehurst

Andrew Whitehurst is a senior visual effects supervisor at Industrial Light and Magic with 25 years of experience. He has worked on films as diverse as Ex Machina, for which he won an Academy Award, Paddington, and Skyfall. His most recent project was Indiana Jones and the Dial of Destiny.

KEYNOTE 2 | Jonathan Starck
Synthesia

Digital Humans in Generative Video
Thursday 30th November 2023

There has been an explosion of interest and adoption of Generative AI over the last few years, driven by the ability to generate human-level text, images, audio, and video. These techniques offer powerful tools to content creators to streamline the creation process, enhance creativity, and deliver personalised content at scale. The aim of this talk is to introduce the current trends and challenges in Generative AI, specifically for live-action content and visual storytelling. Synthesia focuses on democratising video creation for businesses, providing the ability to create photorealistic human avatars and enabling users to create professional videos using text as a simple and accessible interface. This brings particular challenges in creating controllable and life-like performances for digital humans in text-to-video.

Jonathan Starck

Jonathan Starck is CTO at Synthesia, a startup founded in 2017 and now a Generative AI Unicorn. Synthesia produces the world's #1 AI video generation platform that allows users to create professional live-action videos directly in the browser, removing the physical constraints of conventional production. No cameras, microphones, or studios. Just create, iterate and collaborate directly on final quality content using text as an interface. Prior to Synthesia, Jonathan was a Researcher on Digital Humans at the University of Surrey and then Head of Research at Foundry, introducing 3D computer tools to accelerate artist workflows in the Nuke Compositing system for Film VFX. At Synthesia he leads Research and Production and is responsible for generative humans - Synthesia "AI Avatars".

KEYNOTE 3 | Duygu Ceylan
Adobe Research

Towards Multi-Modal Generation
Friday 1st December 2023

We are witnessing an impressive pace of innovation happening in the generative AI space. 2D image domain is often at the frontier of this innovation followed by trends to extend the success to domains such as videos or 3D. While they seem as different domains, one can argue that these domains are in fact very much connected. In this talk, I will talk about some recent efforts that leverage the knowledge of foundational models trained for a particular domain to address tasks in other domains. I will also present thoughts around future opportunities that can leverage this tight connection to go towards universal generation models.

Duygu Ceylan

Duygu Ceylan is senior research scientist at Adobe Research. Prior to joining Adobe in 2014, Duygu obtained her PhD degree from EPFL where I worked with Prof. Mark Pauly. She graduated from Computer Engineering Department of METU in 2007 and got my Master's degree in CS from Bilkent University in 2009. Duygu received the Eurographics PhD Award in 2015 and the Eurographics Young Researcher Award in 2020. Her research interests include using machine learning techniques to infer and analyze 3D information from images and videos, focusing specifically on humans. She is excited to work at the intersection of computer vision and graphics where she explores new methods to bridge the gap between 2D & 3D.

KEYNOTE 4 | Tupac Martir
Satore Studio

Engines For Everything Except Games
Friday 1st December 2023

Although gaming engines are mainly used for creating games, over the past few years Satore has used them for anything but. Instead Satore has found that they are crucial elements of art, VR and immersive pieces. Tupac Martir will share how Satore uses gaming technologies and how they are applied within Satore's work and R&D, and how many of the techniques used are translatable to other aspects of entertainment and art. Tupac firmly believes that Technology is a character, with real-time softwares being the centre of importance at this time.

Tupac Martir

Tupac is the Creative Director and Founder of Satore Studio. He's a multimedia artist whose work spans the fields of technology, lighting, projection and video, sound design, music, and composition, as well as choreography and costumes. Vogue has described him as 'the visual designer and creative director behind some of the most important events in the world'. He has provided production design, visuals and lighting direction for the likes of Elton John, Beyoncé, Danny Boyle, the Coachella Music & Arts Festival and Serpentine Gallery.

Tupac is renowned within the fashion industry and has worked on ground-breaking shows for Alexander McQueen, Moschino, Alexander Wang, and Thomas Tait, among others. In 2019 Tupac Directed and produced "Cosmos Within Us", a performative reality piece that debuted at the Venice Film Festival and won the "Spirit of Raindance Award" at the Raindance London Film Festival. More recently Tupac Directed and produced "Unique" at the BFI London & Sonar, a performative reality piece that showed the future of performance with improvisation at the heart of it. Additionally, Tupac has led a team to become pioneers in Virtual Production. debuting "Haita" at Sónar+D, working with Es Devlin on an illuminated kinetic sculpture for Moet, and being part of TRANSMIXR, a 19 partner consortium across Europe that aims to create a range of human-centric tools for remote content production/ consumption in immersive technologies.

At Satore Studio, Tupac leads an international team of creative minds to create ground-breaking work, mixing art and technology. Tupac founded Satore Studio as a place to explore and create powerful art and multimedia experiences combining - lighting, projection, video, sound design, music, virtual reality and augmented reality.

INDUSTRY TALKS

Applications in Media for Novel View Synthesis

Graeme Phillipson, Hell Raymond-Hayling, Alia Sheikh (BBC)

Recent advances in novel view synthesis (NVS) have provided high quality photo realistic images from points of view outside the input image set. The generation times for NVS models have now decreased to the point where they can realistically be used for content production applications, such as the production of video assets. Additionally there has been the development of more accessible software which allows the models to be used by non-academics in practical industrial applications. For example to export any number of novel paths through a static scene as video. With the advent of models that can accommodate moving subjects it will soon be possible to more usefully apply assets created via novel view synthesis to common media applications. Here we will discuss how NVS models (in particular NVS models that can accommodate moving subjects) could be used in content creation, what possibilities they would enable and what challenges remain for the widespread adoption of these models in the content production landscape.

Making Gen AI work for us, rather than the other way around

Will MacNeil (The Mill)

Most Gen AI tools available right now present as black boxes: closed systems that give us a minimum of creative control. For those of us in the visual creative industry, used to bending software to our will to get exactly the look we want, this just isn't going to work. So how do we fit Gen AI into the processes and workflows we've built our careers around? Mill Creative Director Will MacNeil will show us how Mill artists are integrating text-to-image and large language models into their work, without sacrificing creative control.

The Mill is a world renowned studio, fuelled by award-winning VFX artists & Creative Technologists with decades of experience across the Advertising and Brand Experience Industries. We are founded on the passionate pursuit of excellence, fearlessly pioneering visual imagination for over 30 years, and embracing every evolution of our art with pinnacle craft. Making the impossible not just possible, but utterly believable to create brave new experiences for today's audiences. We work with diverse clients, including global corporations in sports and entertainment, as well as art institutions, museums, and charities. Our passion lies in creating captivating experiences that bridge the gap between creativity and technology.

Overcoming Latency in Real-Time Virtual Production

Dominic Brown (Disguise)

Immersive production experiences rely on a pipeline of tools for high quality content creation, and Virtual Production and Extended Reality require these pipelines to operate in real-time. One of the biggest challenges in VP/XR is latency, especially in moving shots when virtual content is rendered and streamed on the LED wall to the perspective of a moving camera. This talk will present Disguise Research work on a novel reprojection process that significantly reduces the visual error rates caused by content rendering latency. This allows for more creative freedom on camera motion, greater bandwidth for higher quality scene rendering, and opens up possibilities for remote real-time content rendering solutions.

Beyond Reality - Towards Creative Applications for Mixed Reality

Thu Nguyen-Phuoc (Reality Labs Research at Meta)

I will present our latest research progress in the transformation of lifelike 3D scenes and avatars into various artistic and creative styles. Imagine the multitude of possibilities with a wide range of 3D filters. You can effortlessly turn your living room into a van Gogh-inspired Impressionist painting, your office into a scene from the Matrix, or even morph your avatar into a spooky Halloween zombie. This research forms an integral part of our long-term effort to develop captivating visual effects for Mixed Reality (MR) applications. Our primary focus is on 3D content, ensuring faster processing and higher-quality outcomes that can be easily scaled up to accommodate millions of users.

HDR Illumination Outpainting with a Two-Stage GAN Model

Jack Hilliard, Adrian Hilton, Jean-Yves Guillemaut (CVSSP, University of Surrey)

In this paper we present a method for single-view illumination estimation of indoor scenes, using image-based lighting, that incorporates state-of-the-art outpainting methods. Recent advancements in illumination estimation have focused on improving the detail of the generated environment map so it can realistically light mirror reflective surfaces. These generated maps often include artefacts at the borders of the image where the panorama wraps around. In this work we make the key observation that inferring the panoramic HDR illumination of a scene from a limited field of view LDR input can be framed as an outpainting problem (whereby the original image must be expanded beyond its original borders). We incorporate two key techniques used in outpainting tasks: i) separating the generation into multiple networks (a diffuse lighting network and a high-frequency detail network) to reduce the amount to be learnt by a single network, ii) utilising an inside-out method of processing the input image to reduce the border artefacts. Further to incorporating these outpainting methods we also introduce circular padding before the network to help remove the border artefacts. Results show the proposed approach is able to relight diffuse, specular and mirror surfaces more accurately than existing methods in terms of the position of the light sources and pixelwise accuracy, whilst also reducing the artefacts produced at the borders of the panorama.

One-shot Detail Retouching with Patch Space Neural Transformation Blending

Fazilet Gokbudak, Cengiz Oztireli (University of Cambridge)

Photo retouching is a difficult task for novice users as it requires expert knowledge and advanced tools. Photographers often spend a great deal of time generating high-quality retouched photos with intricate details. In this paper, we introduce a one-shot learning based technique to automatically retouch details of an input image based on just a single pair of before and after example images. Our approach provides accurate and generalizable detail edit transfer to new images. We achieve these by proposing a new representation for image to image maps. Specifically, we propose neural field based transformation blending in the patch space for defining patch to patch transformations for each frequency band. This parametrization of the map with anchor transformations and associated weights, and spatio-spectral localized patches, allows us to capture details well while staying generalizable. We evaluate our technique both on known ground truth filters and artist retouching edits. Our method accurately transfers complex detail retouching edits.

A Compact and Semantic Latent Space for Disentangled and Controllable Image Editing

Gwilherm Lesné, Yann Gousseau, Saïd Ladjal, Alasdair Newson (LTCI Télécom Paris)

Recent advances in the field of generative models and in particular generative adversarial networks (GANs) have lead to substantial progress for controlled image editing. Despite their powerful ability to apply realistic modifications to an image, these methods often lack properties such as disentanglement (the capacity to edit attributes independently). In this paper, we propose an auto-encoder which re-organizes the latent space of StyleGAN, so that each attribute which we wish to edit corresponds to an axis of the new latent space, and furthermore that the latent axes are decorrelated, encouraging disentanglement. We work in a compressed version of the latent space, using Principal Component Analysis, meaning that the parameter complexity of our autoencoder is reduced, leading to short training times (

45 mins). Qualitative and quantitative results demonstrate the editing capabilities of our approach, with greater disentanglement than competing methods, while maintaining fidelity to the original image with respect to identity. Our autoencoder architecture is simple and straightforward, facilitating implementation.

DECORAIT - DECentralized Opt-in/out Registry for AI Training

Kar Balan, Alex Black, Andrew Gilbert (University of Surrey), Simon Jenni, Andy Parsons, John Collomosse (Adobe Inc.)

We present DECORAIT; a decentralized registry through which content creators may assert their right to opt in or out of AI training as well as receive reward for their contributions. Generative AI (GenAI) enables images to be synthesized using AI models trained on vast amounts of data scraped from public sources. Model and content creators who may wish to share their work openly without sanctioning its use for training are thus presented with a data governance challenge. Further, establishing the provenance of GenAI training data is important to creatives to ensure fair recognition and reward for their such use. We report a prototype of DECORAIT, which explores hierarchical clustering and a combination of on/off-chain storage to create a scalable decentralized registry to trace the provenance of GenAI training data in order to determine training consent and reward creatives who contribute that data. DECORAIT combines distributed ledger technology (DLT) with visual fingerprinting, leveraging the emerging C2PA (Coalition for Content Provenance and Authenticity) standard to create a secure, open registry through which creatives may express consent and data ownership for GenAI.

Expression-aware video inpainting for HMD removal in XR applications

Fatemeh Ghorbani Lohesara (Technische Universität Berlin), Karen Eguiazarian (Tampere University), Sebastian Knorr (Ernst Abbe University of Applied Sciences Jena)

Head-mounted displays (HMDs) serve as indispensable devices for observing extended reality (XR) environments and virtual content. However, HMDs present an obstacle to external recording techniques as they block the upper face of the user. This limitation significantly affects social XR applications, specifically teleconferencing, where facial features and eye gaze information play a vital role in creating an immersive user experience. In this study, we propose a new network for expression-aware video inpainting for HMD removal (EVI-HR-net) based on generative adversarial networks (GANs). Our model effectively fills in missing information with regard to facial landmarks and a single occlusion-free reference image of the user. The framework and its components ensure the preservation of the user's identity across frames using the reference frame. To further improve the level of realism of the inpainted output, we introduce a novel facial expression recognition (FER) loss function for emotion preservation. Our results demonstrate the remarkable capability of the proposed framework to remove HMDs from facial videos while maintaining the subject's facial expression and identity. Moreover, the outputs exhibit temporal consistency along the inpainted frames. This lightweight framework presents a practical approach for HMD occlusion removal, with the potential to enhance various collaborative XR applications without the need for additional hardware.

Optimising 2D Pose Representations: Improving Accuracy, Stability and Generalisability Within Unsupervised 2D-3D Human Pose Estimation

Peter Hardy, Srinandan Dasmahapatra, Hansung Kim (University of Southampton)

This paper investigated pose representation within the field of unsupervised 2D-3D human pose estimation (HPE). All current unsupervised 2D-3D HPE approaches provide the entire 2D kinematic skeleton to a model during training. We argue that this is suboptimal and disruptive as long-range correlations will be induced between independent 2D key points and predicted 3D coordinates during training. To this end, we conducted the following study. With a maximum architecture capacity of 6 residual blocks, we evaluated the performance of 7 models which each represented a 2D pose differently during the adversarial unsupervised 2D-3D HPE process. Additionally, we showed the correlations induced between 2D key points when a full pose is lifted, highlighting the unintuitive correlations learned. Our results show that the most optimal representation of a 2D pose during the lifting stage is that of two independent segments, the torso and legs, with no shared features between each lifting network. This approach decreased the average error by 20% on the Human3.6M dataset when compared to a model with a near identical parameter count trained on the entire 2D kinematic skeleton. Furthermore, due to the complex nature of adversarial learning, we showed how this representation can also improve convergence during training allowing for an optimum result to be obtained more often.

BundleMoCap: Efficient, Robust and Smooth Motion Capture from Sparse Multiview Videos

Georgios Albanis (University of Thessaly), Nikolaos Zioulis (Moverse), Kostas Kolomvatsos (University of Thessaly)

Capturing smooth motions from videos using markerless techniques typically involves complex processes such as temporal constraints, multiple stages with data-driven regression and optimization, and bundle solving over temporal windows. These processes can be inefficient and require tuning multiple objectives across stages. In contrast, BundleMoCap introduces a novel and efficient approach to this problem. It solves the motion capture task in a single stage, eliminating the need for temporal smoothness objectives while still delivering smooth motions. BundleMoCap outperforms the state-of-the-art without increasing complexity. The key concept behind BundleMoCap is manifold interpolation between latent keyframes. By relying on a local manifold smoothness assumption, we can efficiently solve a bundle of frames using a single code. Additionally, the method can be implemented as a sliding window optimization and requires only the first frame to be properly initialized, reducing the overall computational burden. BundleMoCap's strength lies in its ability to achieve high-quality motion capture results with simplicity and efficiency.

Redistributing the Precision and Content in 3D-LUT-based Inverse Tone-mapping for HDR/WCG Display

Cheng Guo (Communication University of China), Leidong Fan (Peking University), Qian Zhang, Hanyuan Liu (National Radio and Television Administration), Kanglin Liu, Xiuhua Jiang (Peng Cheng Laboratory)

ITM (inverse tone-mapping) converts SDR (standard dynamic range) footage to HDR/WCG (high dynamic range /wide color gamut) for media production. It happens not only when remastering legacy SDR footage in front-end content provider, but also adapting on-the-air SDR service on user-end HDR display. The latter requires more efficiency, thus the pre-calculated LUT (look-up table) has become a popular solution. Yet, conventional fixed LUT lacks adaptability, so we learn from research community and combine it with AI. Meanwhile, higher-bit-depth HDR/WCG requires larger LUT than SDR, so we consult traditional ITM for

an efficiency-performance trade-off: We use 3 smaller LUTs, each has a non-uniform packing (precision) respectively denser in dark, middle and bright luma range. In this case, their results will have less error only in their own range, so we use a contribution map to combine their best parts to final result. With the guidance of this map, the elements (content) of 3 LUTs will also be redistributed during training. We conduct ablation studies to verify method's effectiveness, and subjective and objective experiments to show its practicability.

A software test bed for sharing and evaluating color transfer algorithms for images and 3D objects

Herbert Potechius (Ernst Abbe University of Applied Sciences Jena, Technical University of Berlin), Thomas Sikora (Technical University of Berlin), Gunasekaran Raja (Anna University), Sebastian Knorr (Ernst Abbe University of Applied Sciences Jena, Technical University of Berlin)

Over the past decades, an overwhelming number of scientific contributions have been published related to the topic of color transfer, where the color statistic of an image is transferred to another image. Recently, this idea was further extended to 3D point clouds. Due to the fact that the results are normally evaluated subjectively, an objective comparison of multiple algorithms turns out to be difficult. Therefore, this paper introduces the ColorTransferLab, a web based test bed that offers a large set of state-of-the-art color transfer implementations. Furthermore, it allows users to integrate their implementations with the ultimate goal of providing a library of state-of-the-art algorithms for the scientific community. This test bed can manipulate both 2D images, 3D point clouds and textured triangle meshes, and it allows us to objectively evaluate and compare color transfer algorithms by providing a large set of objective metrics. As part of ColorTransferLab, we are introducing a comprehensive dataset of freely available images. This dataset comprises a diverse range of content with a wide array of color distributions, sizes, and color depths which helps in appropriately evaluating color transfer. Its comprehensive nature makes it invaluable for accurately evaluating color transfer methods.

LFSphereNet: Real Time Spherical Light Field Reconstruction from a Single Omnidirectional Image

Manu Gond, Emin Zerman (Mid Sweden University), Sebastian Knorr (Ernst Abbe University of Applied Sciences Jena), Mårten Sjöström (Mid Sweden University)

Recent developments in immersive imaging technologies have enabled improved telepresence applications. Being fully matured in the commercial sense, omnidirectional (360-degree) content provides full vision around the camera with three degrees of freedom (3DoF). Considering the applications in real-time immersive telepresence, this paper investigates how a single omnidirectional image (ODI) can be used to extend 3DoF to 6DoF. To achieve this, we propose a fully learning-based method for spherical light field reconstruction from a single omnidirectional image. The proposed LFSphereNet utilizes two different networks: The first network learns to reconstruct the light field in cubemap projection (CMP) format given the six cube faces of an omnidirectional image and the corresponding cube face positions as input. The cubemap format implies a linear re-projection, which is more appropriate for a neural network. The second network refines the reconstructed cubemaps in equirectangular projection (ERP) format by removing cubemap border artifacts. The network learns the geometric features implicitly for both translation and zooming when an appropriate cost function is employed. Furthermore, it runs with very low inference time, which enables real-time applications. We demonstrate that LFSphereNet outperforms state-of-the-art approaches in terms of quality and speed when tested on different synthetic and real world scenes. The proposed method represents a significant step towards achieving real-time immersive remote telepresence experiences.

View-dependent Adaptive HLOD: real-time interactive rendering of multi-resolution models

Rui Li (Sorbonne Université)

Real-time visualization of large-scale surface models is still a challenging problem. When using the Hierarchical Level of Details (HLOD), the main issues are popping between levels and/or cracks between level parts. We present a visualization scheme (both HLOD construction and real-time rendering), which avoids both of these issues. In the construction stage, the model is first partitioned (not cut) according to a Euclidean cubic grid, and the multi-resolution LOD is then built by merging and then simplifying neighboring elements of the partition in an octree-like fashion, fine-to-coarse. Some freedom applies to the simplification algorithm being used, but it must provide a child-parent relation between vertices of successive LODs. In the rendering stage, the octree-based hierarchy model is traversed coarse-to-fine to select the cube with the appropriate resolution based on the position of the viewpoint. Vertex interpolation between child and parent is used to achieve crack and popping-free rendering. We implemented and tested our method on a modest desktop PC without a discrete GPU, and could render scanned models of multiples tens of million triangles at optimal visual quality and interactive frame rate.

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

Light Field Video Compression Efficiency Through View Omission and Synthesis

Nusrat Mehajabin, Tala Bazzaza, Hamid Reza Tohidypour, Yixiao Wang, and Panos Nasiopoulos
http://www.dml.ubc.ca

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

Light field (LF) imaging, also known as plenoptic imaging, represents a groundbreaking technology continually evolving to deliver a human-like visual data, aspiring to faithfully replicate our visual environment and closely emulate human perception [2]. As opposed to traditional cameras that only capture the scene from a single viewpoint, a light field camera allows light to be captured from multiple viewpoints, preserving realistic vertical and horizontal parallax and effectively recording not only the intensity but also the direction of light rays. This results in ample data allowing post-scene depth of field, focal point, or resolution adjustments. Additionally, depth and distance data facilitate segmentation and object detection. Light field tech has diverse uses, including cinematography, augmented/virtual reality, and medical applications.

However, the substantial increase in captured data underscores the paramount importance of efficient compression techniques. Traditional compression standards are unsuitable for handling light field data. Naturally, an effective encoding method for managing this vast amount of data is a critical factor in enabling the technology and unlocking new market opportunities. State-of the-art LF compression methods center on organizing keyframes (I and P frames) and leveraging horizontal and vertical similarities within the hierarchical bi-directional (B) frames. Khoury et al. positioned the I-frame at the center and expanded the structure by placing the P-frames at the furthest cells horizontally and vertically, achieving 38% bitrate reduction compared to LF-MVC [3]. Mehajabin's et al. approach uses an SSIM based keyframe selection strategy to determine the correlation among the views being predicted and their references and choose accordingly the different type of frames [4]. This method is an extension of Multiview-HEVC and is shown to improve compression over [3] by 17%, making it the state-of-the-art for LF video compression at the time of writing this article.

In this paper, we investigate the potential for achieving greater compression efficiency with light field data by omitting certain views during transmission and then synthesizing them at the receiver end using a synthesis method tailored for LF [5]. Meanwhile, we compress the remaining views using the state-of-the-art LF coding approach presented in [4]. We evaluated the compression efficiency of the original arrangement that involves all the views with that of several subset view arrangements where some original views are dropped and then synthesized. Figure 1a shows the different view arrangements that we considered: (a) all views, (b) only columns, (c) only rows, (d) raster skip, and (e) extreme. All these view arrangements were compressed using Mehajabin's approach which is shown to effectively scale to any number of views. The missing LF views are synthesized using Wafa's et al.'s method, a deep recursive residual network designed to generate LF views from sparse input views. This state-of-the-art view synthesis approach is based on a GAN learning-based model that is trained using the spatial and angular information of the light field content.

We evaluated our approach on publicly available microlens based light field videos captured by the Raytrix LF camera [1]. The video sequences are 30fps with 2K resolution and duration of 10 seconds. We examined various compression quality levels by setting QP values to 25, 30, 35 and 40.

The results showcased in this article are based on the 'Chess-Pieces' light field video, with consistent outcomes observed across all other videos. Preliminary results showed that the two row view arrangements perform poorly at the reconstruction stage because of the significant number of

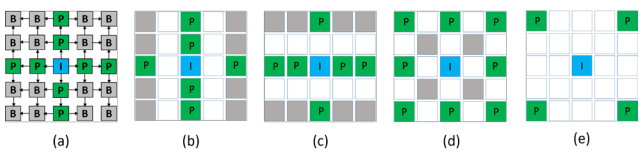
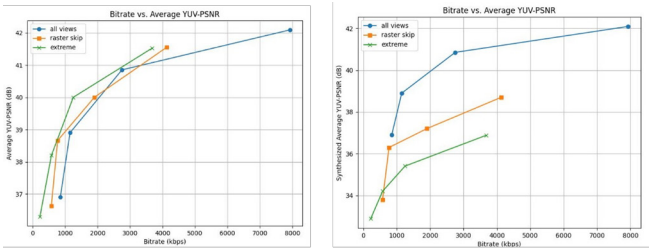


Figure 1: The different view arrangements considered in our tests: (a) all views, (b) only columns, (c) only rows, (d) raster skip, and (e) extreme.



(a) (b)

Figure 2: (a) bitrate and average PSNR for three compression scenarios: compressing all views, raster skip (half the views), and extreme view arrangements; (b) bitrate and average visual quality, quantified in terms of PSNR, encompassing all views, both transmitted and reconstructed.

adjacent views missing compared to the raster skip arrangement. Thus, we present results of the two representative arrangements, raster skip and extreme. Fig. 2a illustrates the relationship between bitrate and average PSNR for three compression scenarios: compressing all views, raster skip (half the views), and extreme view arrangements. We observe that, in general, the hypothesis of saving bandwidth by omitting views during transmission is true, however, there are important trade-offs to consider. Figure 2b displays the bitrate and average visual quality, quantified in terms of PSNR, encompassing all views, both transmitted and reconstructed. We observe that the performance of transmitting all the views is higher than that of omitting half or more views. The reason for this is that reconstructed quality of the missing views, although visually acceptable, does not match what would have been attained if the complete light field were transmitted.

These findings highlight the delicate balance between bandwidth efficiency and reconstruction quality in light field compression and transmission. While sending a subset of views can lead to cost-effective data transfer in bandwidth-constrained scenarios, it is crucial to acknowledge that this approach may not be suitable for applications demanding the highest possible image quality or where the omission of certain views could result in a loss of critical information.

Our findings also suggests that while compression algorithms for light field have made significant progress, there is still ample room for improvement in view synthesis algorithms. To advance the research further we can test different view synthesis algorithms to analyze the impact. we can also investigate which patterns of view omission strike the best balance between bandwidth efficiency and reconstructed view quality. This may involve omitting a specific subset of views based on their importance, spatial distribution, or other criteria.

[1] L. Guillo, X. Jiang, G. Lafruit, and C. Guillemot. Light field video dataset captured by a r8 raytrix camera (with disparity maps). In *HAL Id: hal-01804578*, 2018.

[2] L. Mignard-Debise I. Ihrke, J. Restrepo. Principles of light field imaging: Briefly revisiting 25 years of research. In *IEEE Signal Processing Magazine, vol. 33, no. 5, pp. 59 – 69, September*, 2016.

[3] J. Khoury, N. Mehajabin, M. T. Pourazad, P. Nasiopoulos, and V. C.M. Leung. An efficient three-dimensional prediction structure for coding light field video content using the mv-hevc standard. In *International Journal of Multimedia Intelligence and Security, vol 4, no. 1, pp. 47-64*, 2022.

[4] N. Mehajabin, M. T. Pourazad, and P. Nasiopoulos. An efficient pseudo-sequence-based light field video coding utilizing view similarities for prediction structure. In *IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 4, pp. 2356-2370, April*, 2022.

[5] A. Wafa and P. Nasiopoulos. Light field gan-based view synthesis using full 4d information. In *ACM SIGGRAPH European Conference on Visual Media Production (CVMP), London, UK, December*, 2022.

Human Pose Estimation Assisted Posture Reminder for E-Learning Video Production

Yixia Zhao
http://cnic.cas.cn/jgsz/kyywbm/xmtjsyyyfzb
Computer Network Information Center,
Chinese Academy of Sciences, Beijing, China

Zhenhua Feng
https://www.surrey.ac.uk/people/zhenhua-feng
School of Computer Science and Electronic Engineering,
University of Surrey, Guildford GU2 7XH, United Kingdom

In recent years, e-learning has seen considerable growth, particularly since the outbreak of COVID-19 in 2020. In the non-face-to-face e-learning scenario, pre-recorded video lectures are one of the most important learning materials [5]. During the video recording stage, the body language of a presenter is very important, which can affect the user's learning attention and efficiency [4]. However, most teachers do not have rich experience in recording a video presentation [3]. A general video recording monitor only displays the live streaming video to the presenter, but cannot provide clear and real-time instructions to adjust her/his body postures.

To address the above issue, the technique of Human Pose Estimation (HPE) has become a very useful tool that can capture the posture of a presenter using standard RGB cameras [2]. The aim of HPE is to predict the 2D or 3D key points of a person by giving a single image or a video sequence. This is helpful for understanding the pose of a presenter during the video recording stage and providing useful instructions. There are many existing human pose estimation tools. In this paper, we use MediaPipe[1], which is a graphics-oriented, cross-platform framework that facilitates comprehensive acceleration throughout the end-to-end process. MediaPipe is able to detect the presence of an individual within the current frame, estimate the coordinates of 33 key points of the human body, and precisely delineate the region of interest [1].

Based on MediaPipe, we apply HPE to develop a posture reminder system to assist a presenter during the lecture recording process. To be specific, we propose a method that includes three main adjustments for the postures of the head, arm and body of a presenter. We first obtain the key point information of more than 1000 Chinese Academy of Sciences Massive Open Online Courses (CASMOOC) videos using MediaPipe. Based on the analysis of these videos, we design three main pose adjustment principles for a teacher, including head pose, arm pose and body pose. The head pose includes the yaw, roll and pitch rotations of the head of the presenter. The arm pose consists of the angle between the upper arm and lower arm, and the angle between the upper arm and body. Last, we use the coordinates of all 33 key points as the body pose.

The overall pipeline of the proposed method is shown in Fig 1. We define the head pose from the 3D perspective. The threshold of each head angle (yaw, roll, and pitch) is set to 5°. If any of the three angles exceeds the threshold, the system is triggered and a visual reminder or audio reminder will be provided for the presenter.

At the same time, we take the angle between the upper arm and lower arm and the angle between the upper arm and body into consideration to solve the stiff and uncoordinated posture of the teacher's arms during the recording process. When the angle between the upper and lower arms is greater than 100° but less than 180°, and the angle between the upper and body is less than 30°, the teaching posture appears stiff and unnatural. In this case, a reminder will be sent to the presenter so she/he needs to adjust the arm posture.

The detection of the teacher's whole body posture mainly involves calculating the inclination angle of the body. It needs to adjust body gestures while the angle of the hip midpoint and shoulder midpoint along the centre of gravity line is more than 2°.

To not interrupt the recording of the presenter, we propose two kinds of reminders. The first one is screen reminder while the unsuitable gesture lasted more than 30 seconds and less than 120 seconds. The screen reminder is provided through methods of changing colours or using bold fonts. This signal can reduce the time for teachers to understand the reminder content. Under this circumstance, a teacher will automatically adjust her/his posture. If an inappropriate gesture lasts more than 120 seconds, the system will provide an audio reminder. In this situation, we give a strong reminder signal and we can better post-process the video recording by using the "drip" voice.

Tested on the CASMOOC video courseware datasets, the system achieved an accuracy of 81.25% in identifying improper poses. By deploying the system in CASMOOC lecture video production, the teachers are satisfied with the proposed real-time posture reminder system, verifying the effectiveness of the proposed system. Experiments demonstrate that the proposed system not only improves the quality of E-learning videos but also reduces the recording time. The posture reminder logs can enhance the efficiency of video editing by around 30% in CASMOOC production.

This research focuses on exploring the use of human posture estimation technology in lecture video recording and production. The conclusion is that teachers can receive timely feedback, adjust posture, and improve the efficiency and effectiveness of lecture video recording by using the proposed reminder system.

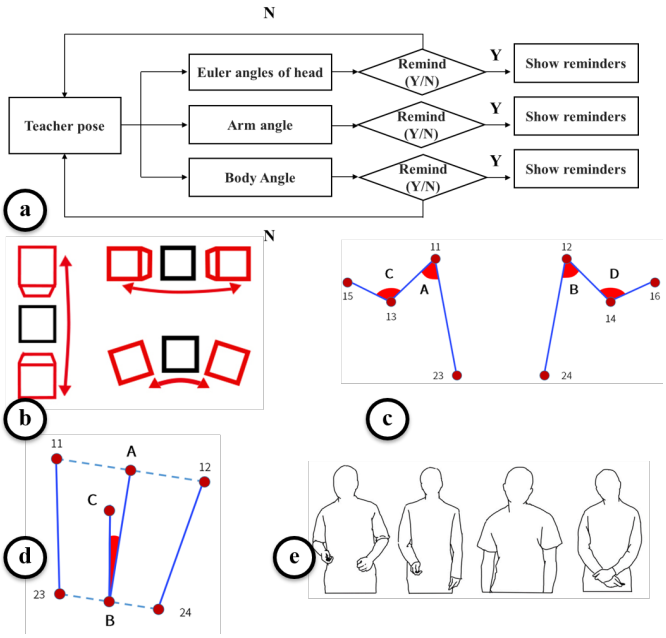


Figure 1: An overview of the proposed system: (a) flow diagram of the posture reminder system; (b) left pitch, top right yaw, and bottom right roll reminders; (c) key points and arm poses; (d) key points and body pose; (e) examples of postures.

[1] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[2] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer vision and image understanding*, 192:102897, 2020.

[3] Philip J Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 41–50, 2014.

[4] Shu-Sheng Liaw. Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the blackboard system. *Computers & education*, 51(2):864–873, 2008.

[5] Syed A Raza, Wasim Qazi, Komal Akram Khan, and Javeria Salam. Social isolation and acceptance of the learning management system (lms) in the time of covid-19 pandemic: an expansion of the utaut model. *Journal of Educational Computing Research*, 59(2):183–208, 2021.

[1] https://github.com/google/mediapipe

Practical Imaging and Analysis of Turbidity in Liquid

Kewei Li
kewei.li22@imperial.ac.uk
Department of Computing,
Imperial College London

Abhijeet Ghosh
ghosh@imperial.ac.uk
Department of Computing,
Imperial College London

Turbidity is the cloudiness phenomenon in the liquids, which can indicate water impurity. This phenomenon is caused by the scattering of the light. In volume rendering, the scattering properties of the volume liquids are described by scattering coefficient and phase function. Previous work on acquiring the scattering properties is measuring the scattering light in a tank filled with diluted liquid [2], or using Min/Max images and polarization to separate light components and then estimating coefficients [1]. In this paper, we propose a practical turbidity estimation pipeline building on the method of [1] to detect impurities dissolved in low concentration in water. Such water samples can look visually clear but can be harmful if consumed. Thus, its practical detection has application in health monitoring, particularly in developing countries.

Our estimation pipeline has three steps: image sampling, decomposition, and coefficient estimation. The acquisition setup in Figure 1 is very simple and can be applied in most developing countries. Ten HDR images are captured in the image sampling step: one image for the empty cup under the homogeneous illumination for reference, eight images for the liquid under the horizontal and vertical stripe illumination, and one cross-polarized image for the liquid under the homogeneous illumination.

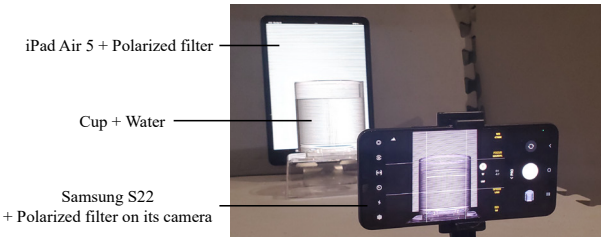


Figure 1: Acquisition setup, in which iPad is for illumination, Samsung S22 is for imaging, and the measured water is in a glass

In the decomposition step, we separate direct component, single and multiple scattering of the light. We first separate the direct and global component based on the Min/Max images. For the horizontal Min and Max images, we denote the images under the horizontal stripe illumination as I_{hi}, and the calculation of Min and Max images is written as

$$Min(x,y) = \min_{i=1}^{4} I_{hi}(x,y),\ Max(x,y) = \max_{i=1}^{4} I_{hi}(x,y) \tag{1}$$

The final Min and Max images are the average of the horizontal and vertical Min and Max images. The Min image captures the light scattering from the illuminated regions, and contains half of the global component, while the Max image captures the scattered light from other illuminated regions and the direct light, and contains the whole direct component and half of the global component. Therefore, the direct component = Max − Min, and the global component = 2 × Min. We then separate the single and multiple scattering in the global component by polarized imaging. We assume that the single scattering is polarized, while the multiple scattering is unpolarized and can be captured in the cross-polarized image. Therefore, the single scattering is calculated by subtracting the radiance of the cross-polarized image in the global component, and the multiple scattering is double the radiance of the cross-polarized image.

In the coefficient estimation step, we estimate the extinction coefficient, the scattering coefficient, and the scattering anisotropy of the liquid. We denote the radiance in the image under the homogeneous illumination and direct component as I_0 and I_d respectively, and choose the center of the cup as the region of interest (ROI). The extinction coefficient σ_t is estimated by inverting the volume rendering process:

$$\sigma_t = -\frac{1}{2r}\log\left[\frac{average\{I_d(x,y)\}}{average\{I_0(x,y)\}}\right] \tag{2}$$

in which the average is in the ROI, and r is the radiance of the cup. To estimate the extinction coefficient of the water with small turbidity, we use that of pure water as the reference, which may not be zero because of the noises of sampling. We assume the reference is absolutely pure. Given the extinction E_0 of the pure water and the extinction E of the measurement liquid, the normalized extinction coefficient σ_t' is

$$\sigma_t' = -\frac{1}{2r}\log\frac{E}{E_0} \implies \sigma_t' = \sigma_t - \sigma_{t0} \tag{3}$$

in which σ_{t0} is the extinction coefficient of the pure water.

We then estimate the scattering by a cross-section model which is similar to that in [1]. The iPad screen is modelled as an area light source emitting I_0 radiance, and the estimated single scattering I_s in ROI is

$$I_s = \beta \int_{x\in\text{iPad}} \int_{-r}^{r} I_0 \cos\theta \cdot e^{\sigma_t(d+r-z)} P(g,\cos\theta)\,dz\,dx \tag{4}$$

in which β is the scattering coefficient, $P(g,\cos\theta)$ is the H-G phase function, and g is the scattering anisotropy. We use the Nelder-Mead method to search the β and g making the Equation 4 to match the captured single scattering. The scattering coefficient (m^{-1}) has an almost positive relationship with the nephelometric turbidity unit (NTU) [3], and therefore we can estimate the turbidity from the scattering coefficient directly.

(a) Diluted dark soy sauce (b) Diluted milk (c) Diluted coffee

Figure 2: Photograph (left) and rendering (right) of different liquids

Figure 2 shows the comparisons of photographs and renderings generated by PBRT using the estimated coefficients, in which the rendered liquids are very similar to the real liquids. The renderings with estimated coefficients in an outdoor scene are shown in Figure 3. The result of the estimation shows that our method can handle turbidity of less than 3 NTU and even can classify the turbidity of less than 1 NTU for colorless liquid with high scattering. Since the radiance of light changes more with a longer transmitted distance, smaller turbidity can be measured by increasing the size of the cup. In conclusion, we propose a practical method to detect the low-concentration impurities in the water, which uses a simple acquisition setup and is more applicable in most regions, particularly in developing countries.

(a) Measurement concentration (b) Natural concentration

Figure 3: Rendering of the liquids in an outdoor scene. Left to right: dark soy sauce, semi-skimmed milk, and coffee. The concentration in (a) is measurement concentration, and that in (b) is natural concentration scaled from the measurement concentration.

[1] Jaewon Kim and Abhijeet Ghosh. Practical acquisition of translucent liquids using polarized transmission imaging. In *ACM SIGGRAPH 2017 Posters*, SIGGRAPH '17, 2017.

[2] Srinivasa G Narasimhan, Mohit Gupta, Craig Donner, Ravi Ramamoorthi, Shree K Nayar, and Henrik Wann Jensen. Acquiring scattering properties of participating media by dilution. In *ACM SIGGRAPH 2006 Papers*, pages 1003–1012. 2006.

[3] Gonzalo L Pérez, Ana Torremorell, José Bustingorry, Roberto Escaray, Patricia Pérez, Maria Diéguez, and Horacio Zagarese. Optical characteristics of shallow lakes from the pampa and patagonia regions of argentina. *Limnologica*, 40(1):30–39, 2010.

SDiT: Enhanced Stable Diffusion for Audio-to-Talking Face Generation

Fatemeh Nazarieh

Zhenhua Feng

School of Computer Science and Electrical Engineering,
University of Surrey,
Guildford GU2 7XH, UK

Audio-to-talking face generation is considered a subtask of cross-modality content generation. In this task, the generative model learns to generate realistic talking faces speaking the provided speech while maintaining the style of the reference image. Audio-to-talking face generation has many applications such as animation generation in the entertainment industry or generating talking virtual avatars to increase human-computer interaction. Audio and visual content are from different modalities, and bridging the gap between them is not a straightforward task. Important challenges faced in this task include audio-lip synchronization, high-quality image generation, and spatiotemporal consistency amongst generated video frames. To address the complexities involved in audio-to-talking face generation, we introduce SDiT, a novel approach that combines the state-of-the-art stable diffusion [3] technique with vision transformers (ViT [2]). It differs from previous works in utilizing ViT as the underlying neural backbone in stable diffusion. This enables SDiT to generate highly realistic and well-aligned talking faces based solely on speech input and a single reference frame of a particular identity. Further, our experiments demonstrate that incorporating vision transformers substantially enhances the speed of content generation.

Proposed Method: We specifically utilize ViT instead of the commonly used UNet to better learn the underlying relationship between the input speech and visual elements such as lip movements and facial expressions. During training, following the DiffTalk [4] approach, we employ a masked ground truth frame and face landmarks as control mechanisms to effectively learn audio-talking face synchronization. This is crucial since estimating head movements solely based on audio is a challenging task for the model. Before processing the image, we segment the input speech into overlapping windows of 500ms each and use the wav2vec [1] transformer for feature extraction. We find that these overlapping windows help capture context around each current frame, improving the generalization capabilities of the model.

Figure 1 illustrates the overall framework of the proposed model. The input reference frame ($x_{rf} \in \mathbb{R}^{(H,W)}$), ground truth frame ($x \in \mathbb{R}^{(H,W)}$) and masked ground truth frame ($x_m \in \mathbb{R}^{(H,W)}$) are concatenated channel-wise and processed using an image encoder $z_t = \varepsilon_\theta(x_t, t)$. By providing masked frames (x_m) during training, SDiT can learn the suitable head movements corresponding to each audio section. Obtained latent vector (z_t) contains processed visual features and is down-sampled by the factor of f ($x \in \mathbb{R}^{(H/f,W/f)}$) which makes the training process faster compared to using the initial resolution of the images. The determined vector (z_t) is utilized as the primary input of the diffusion model while audio and face landmarks are utilized as conditions (C). After adding noise to the latent vector (z_t) using a linear scheduler, the ViT predicts the amount of noise added at each diffusion time step t. The related objective function is presented based on Eq. 1:

$$L_{SDiT} = E_{(z,\varepsilon\sim N(0,1),C,t)}\left[\|\varepsilon - M(z_t, C, t)\|_2^2\right] \quad (1)$$

where M represents the denoising network with ViT, and t represents the diffusion time steps $t \in [1, 2, .., T]$. Further, since the input sequence and condition are from different modalities, the cross-attention mechanism is adopted for the vision transformer as follows:

$$Attention(Q, K, V) = softmax(\frac{(QK^T)}{\sqrt{d_q}}V \quad (2)$$

In Eq. 2, the Q carries visual information while K and V carry information related to the condition information. The attention mechanism enables the capturing of relationships and associations between sequence elements of different modalities. It should be noted that during inference, we only use the provided speech and reference frame for talking face generation.

Implementation details: The diffusion time step is set to 1000 for training and 200 for inference. The down-sampling factor is fixed to 8.

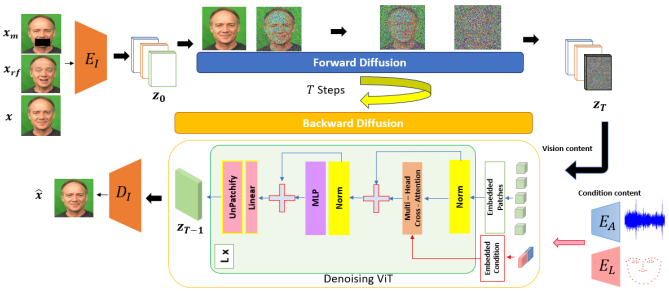


Figure 1: Overview of the SDiT architecture.

We train this model on 128x128 resolution videos, obtained from the CREMA-D dataset. We train this model using two Nvidia A100 GPUs with a batch size of 300. The training stage of SDiT finished in 10 days.

Experimental results: We evaluate our proposed audio-based talking face generation (SDiT) with two recent state-of-the-art models. While all methods performed competitively, SDiT demonstrated superior performance. Specifically, it outperformed Stable Diffusion and DiffusedHeads [5]) in terms of FID, SSIM and LPIPS (Table 1). Indicating better visual fidelity, audio-lip synchronization, and consistent facial structures.

Table 1: A comparison with SOTA methods in terms of image quality.

Model	FID↓	SSIM↑	PSNR↑	LPIPS↓
SDiT	11.676	0.688	23.698	0.082
StableDiffusion	12.459	0.674	23.991	0.094
DiffusedHeads	12.450	0.545	21.803	0.104

Further, we explored the effectiveness of adopting ViT with stable diffusion, in terms of generation speed. One of the key drawbacks of the most current generative models is the computational costs. Based on the determined results, SDiT presents a much faster inference speed, as shown in Table 2.

Table 2: A comparison with SOTA methods in terms of inference speed.

Model	Speed (FPS)	# Parameters
SDiT	227.16	110 M
StableDiffusion	60.55	1 M
DiffusedHeads	0.09	215 M

The speed-to-parameter ratio for SDiT provides a balanced profile, highlighting its applicability for deployment in real-world scenarios where both speed and complexity are key factors.

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. NIPS'20, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[3] Robin Rombach, Andreas Blattmann, et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[4] Shuai Shen, Wenliang Zhao, Zibin Meng, et al. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *CVPR*, 2023.

[5] Michał Stypułkowski, Konstantinos Vougioukas, et al. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023.

So you think you can dance? Controllable Multi-person Dance Generation

Emily Corby
www.linkedin.com/in/emily-corby-477ab5222
Edward Fish
https://ed-fish.github.io/
Andrew Gilbert
https://www.andrewjohngilbert.co.uk/

C-CATS:
Centre for Creative Arts and Technologies
University of Surrey
UK

The generation of synthetic images and videos has exploded in recent years, first with image and later with video, generally driven by diffusion models [4]. With the improvement from the era of GANs [3], to video2video style transfer [2] for transferring dance movements from a source video to a target individual, to fine-tuned pose controlled methods such as DisCo [7]. However, generally, these need more controllability or the ability to create multiple people, limiting real-world applications. Therefore, this work looks to remove these issues by adapting the previous work of DisCo [7] for the generation of videos that contain multiple backing dancers mimicking the dance style of a lead dancer. Disentangled Control for Referring Human Dance Generation in Real World (DisCo) is a model architecture that can generate dance videos and images with three properties: faithfulness (keeping the foreground and background consistent with the reference image whilst maintaining a precise pose), generalizability (can be used with unseen foreground, background and pose) and compositionality (can adapt to random compositions of seen and unseen foreground, background and pose using different video/image sources). We aim to build on this around multiple people and poses so it can be applied to real world dance scenarios to generate backing dancers in a video.

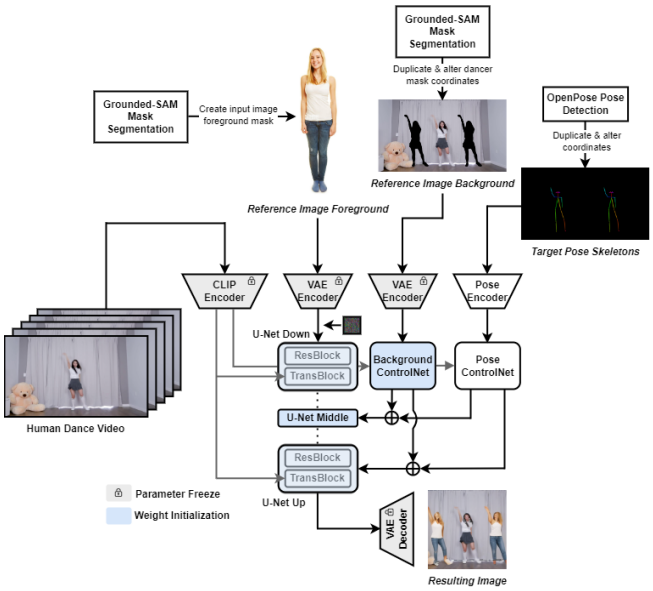


Figure 1: The model framework [7]

The model, as shown in Figure 1, is based on a pre-trained stable diffusion U-Net, with the addition of a ControlNet [6] for the background and pose parameters. For training, the foreground frame is split into reference foreground, reference background (both extracted with Grounded-SAM [5]), target pose and target input image. These are fed into frozen CLIP, OpenPose [1], and VAE encoders to create feature embeddings. The reference foreground and background are fed into the down block layer of the U-Net (alongside the noise image). The target pose and target image are fed through ControlNets to control the stable diffusion process further. The outputs of these two ControlNets are summed and fed into the middle and upper block layers of the U-Net.

Due to the disentangled control, the three aspects of the model, the foreground, background and pose, can be easily manipulated to befit a specific application design. Therefore, generating multiple backing dancers in a video at inference is possible. This can be done by duplicating the pose skeleton and altering the coordinates of each, so they stand on either side of the original dancer. The segmentation mask is modified so the dancer's mask is duplicated and moved to the same coordinates as the pose skeletons. This is so that when the reference background image is created, the dancer's shape is masked out in the desired position of the new backing dancers. These steps will need to be taken for each frame of the video. The input image can contain one person or two different people, and Figure 2 shows output examples.

Figure 2: Frames taken from the result of the video generation model when using one-person and two-person input images

Qualitatively, the dance moves are distinguishable, and the choreography recognisable, but the current model has a few limitations. Firstly, it is zero-shot, so the accuracy of the facial and limb details is lacking; however, this can be improved by applying DisCo's human-specific fine-tuning model to train DisCo on the specific images being used. Secondly, the model has been pre-trained on approximately 350 TikTok dance videos. However, none of the videos displayed the subject's legs, so when using zero-shot full body images, the model struggles to interpret them, and the legs are significantly less accurate in the generated video than the rest of the body; this also could be solved through improved diversity in the training dataset. Our model allows for the generation of automatic backing dancers in dance videos with faithful reproduction of the pose and dancer's appearance. However, this is a work in progress, and it would be possible to remove the limitations of the quality of the result through more diverse training data and make it even more suitable for real-world applications.

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017.

[2] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *ICCV*, 2019.

[3] Ian Goodfellow and et al. Generative adversarial networks. *Communications of the ACM*, 2020.

[4] Ajay Jain Jonathan Ho and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[5] Alexander Kirillov and et al. Segment anything. *arXiv*, 2023.

[6] Maneesh Agrawala Lvmin Zhang. Adding conditional control to text-to-image diffusion. *arXiv*, 2023.

[7] Tan Wang and et al. Disco: Disentangled control for referring human dance generation in real world. *arXiv*, 2023.

InclusivityXR – an online tool for detecting inclusivity issues in AR and VR

Andy Woods[1]
andy.woods@rhul.ac.uk
Umar Farooq[2]
f.umar@surrey.ac.uk
James Bennett[1]
james.bennett@rhul.ac.uk
Marco Volino[2]
m.volino@surrey.ac.uk

[1] StoryFutures,
Royal Holloway, University of London
[2] Centre of Vision, Speech and Signal Processing (CVSSP),
University of Surrey

1 Introduction

There is a huge push to bring virtual and augmented reality into the mainstream by companies such as Meta ($36 billion, 2019-2022 [4]) and Apple (14k AR apps in the App store, 2022), with evidence suggesting that AR is at the point of mass adoption (e.g. 63% of people in the UK regularly use AR [1]). At the same time, one in five people in the UK have disabilities [5] and it is estimated the 'Purple pound' is valued in the UK at over £212bn per year [3]. Appropriately, there is a drive to make immersive content accessible to all.

Whilst there are tools in development for detecting accessibility issues for web-based immersive experiences that follow web-based WCAG guidelines (e.g. `https://aria-at.w3.org/`), we are not aware of tools specifically for immersive experiences developed for outside the web-browser (although Unity[1] and Unreal[2] do provide some general advice and tools). One reason for this is that there are a great variety of ways to define elements in immersive experiences, and it is hard to create tools flexible enough to deal with this variation – for example, a button can be created in a plethora of ways, from being a 'stock' button provided by a platform, to being crafted via a GUI, to being generated in code; or indeed being a hybrid of several of these techniques.

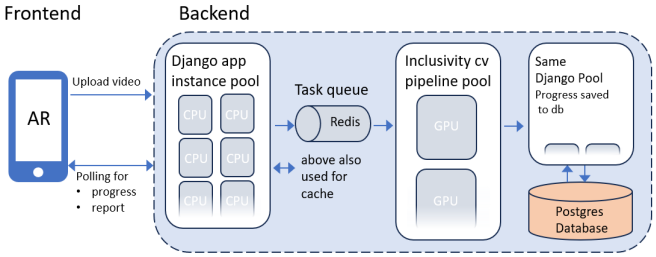


Figure 1: Server architecture. Note that multiple instances of the web-applications running at the same time is common policy, helping ensure reponsiveness during busy periods. During development, we hosted the above on single server, requiring about 10s to process / 1s realtime (8 x 2.1 GHz x86 intel Xeon Gold, 32GB RAM, no GPU)

2 InclusivityXR pipeline

InclusivityXR solves this problem by detecting issues of inclusivity at the level of pixels – whilst being agnostic to the code / schema that defines visual elements. It does this by means of a computer vision pipeline, through the following steps:

1. Developers upload a screen recording (video) of the app in action via a web application (Figure 1).

2. The video is processed (Figure 2) and an online report is generated, listing issues that have been detected.

Our approach to detecting inclusivity issues is as follows:

1. Identify key UI elements. We achieve this by building on the work by Chen, Jieshan, et al. [2] whose tool scans single still images and identifies UI elements using a range of traditional and deep-learning computer vision techniques. We extend their solution to work with video, in so doing augmenting a variety of time-based information to detected UI elements.

2. Apply a set of bespoke detectors each focusing on a given issue of inclusivity. So far we have implemented the following detectors: text too small; colour contrast inappropriate for regularly sighted individuals as well as individuals with 3 most common forms of colour-blindness; UI element too small; font overcrowding.

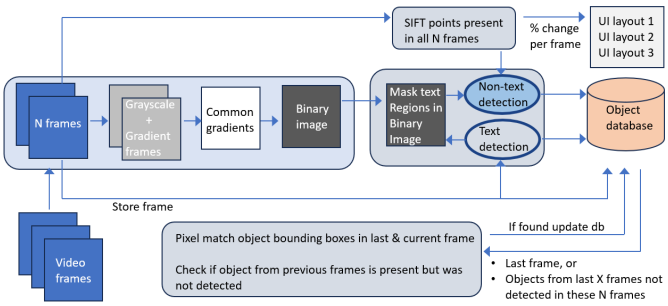


Figure 2: Computer vision pipeline

3 Future Work and Conclusion

We have in place the groundworks of a framework that can be readily extended with new bespoke detectors for other issues of inclusivity. A key next step is to identify which inclusivity issues the tool can have the most impact on for end-users and developers; to achieve this we plan to work with charities and individuals with different disabilities. A longer term goal is to enhance our pipeline via deep learning components, for which we will need more developers interested in AR/VR inclusivity to upload videos to the platform.

4 References

[1] J. Bennett, P. Dalton, O. Goriunova, C. Preece, L. Whittaker, I. Verhulst, and A. T. Woods. The story of immersive users. Technical report, StoryFutures, 2021. URL `www.storyfutures.com/resources/audience-insight-report`.

[2] J. Chen, M. Xie, Xing Z., C. Chen, X. Xu, Zhu L., and G. Li. Object detection for graphical user interface: Old fashioned or deep learning or a combination? In *28th ACM joint meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020.

[3] Department for Work and Pensions. High street could be boosted by £212 billion 'purple pound' by attracting disabled people and their families. Technical report, Gov.UK, 2014. URL `https://tinyurl.com/5n7khzrr`.

[4] ImmerseUK. Immersive economy report. Technical report, 2022. URL `www.immerseuk.org/wp-content/uploads/2022/10/Immersive-Economy-2022-final-13-Oct.pdf`.

[5] M. A. Vaughan. Family resources survey: Financial year 2019 to 2020. Technical report, Gov.UK, 2021. URL `https://tinyurl.com/ye2474s5`.

[1] `https://www.foundations.unity.com/fundamentals/accessibility`
[2] `https://docs.unrealengine.com/5.0/en-US/accessibility-in-unreal-engine/`

Towards Neural Representations of Heterogeneous Translucent Voxelised Media

Tom Gillooly[1]
thomas.b.gillooly@ntnu.no
Jon Y. Hardeberg[1], Abhijeet Ghosh[2], G. Claudio Guarnera[3]

[1] Norwegian University of Science and Technology
[2] Imperial College London
[3] University of York

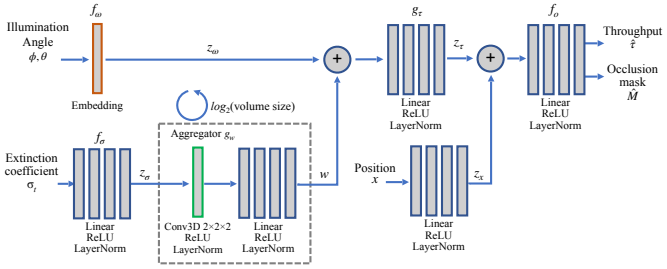
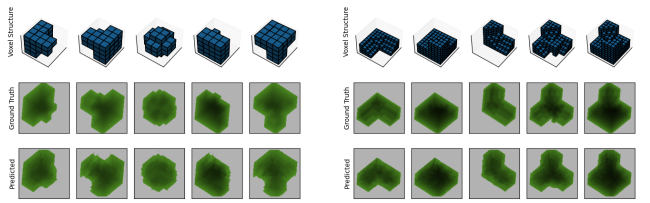


Figure 1: Full neural model. Vector concatenation is indicated by "+".



(a) Cube size 4, ρ_c=0.961, wMAPE=0.049 (b) Cube size 8, ρ_c=0.920, wMAPE=0.093

Figure 2: Neural renders of voxel structures at various scales.

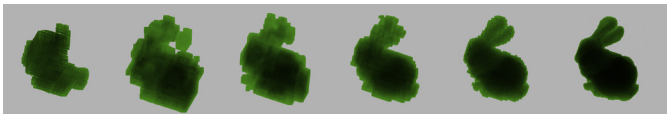


Figure 3: Neural rendering of a voxelized Stanford bunny. From left to right: full volume rendering with one feature vector, then progressively splitting the volume in half and using a feature vector per sub-volume.

When rendering homogeneous media, the log transmittance is the product of the extinction coefficient and the path length. Therefore, a reduction in path length can be offset by an increase in the extinction coefficient to maintain the same transmittance. We refer to this increased extinction coefficient as the equivalent optical parameter. However in heterogeneous structures, depending on the incident angle and position, the path may encounter varying optical parameters. Therefore, an equivalent optical parameter can no longer be expressed as a single scalar value; instead it must be represented as a function of material composition, incident angle, and incident position. Existing work on heterogeneous voxelised structures retains the full voxel grid in feature space [3], while other research on neural network-based rendering of translucent media relies on handcrafted volume representations [1]. In contrast, we learn a single vector representation by progressively aggregating latent optical parameter representations of a voxelised structure and train a neural rendering pipeline to convert these representations into throughput values.

Specifically, we individually encode raw optical parameters in a 3D structure $z_\sigma = f_\sigma(\sigma_t) \in \mathbb{R}^{N\times N\times N\times D}$, which are then aggregated with a learned function g_w, implemented as a 3D convolutional kernel:

$$w_n = g_w(w_{n-1}) \quad w_0 := z_\sigma \tag{1}$$

Once the volume has been fully aggregated, the resulting latent vector can be conditioned on an encoded position value to produce both a throughput prediction $\hat{\tau}$ and an occlusion prediction $\hat{M}$:

$$\hat{\tau} = f_\tau(z_\tau, z_x)_0 \quad \hat{M} = f_\tau(z_\tau, z_x)_1 \tag{2}$$

where $z_\tau = g_\tau(z_\omega, w_n)$ is the latent volume representation conditioned on the angle of illumination $z_\omega = f_\omega(\phi, \theta)$, and $z_x = f_x(x)$ is a vector representing the query location on the surface of the volume. The occlusion prediction indicates whether the incident illumination has passed cleanly through the volume without intersecting any voxels. Model accuracy is evaluated with the Concordance Correlation Coefficient ρ_c [2] and Weighted Mean Absolute Percentage Error (wMAPE), the latter given by the ratio $\sum_{i=1}^{n}|A_i - F_i| / \sum_{i=1}^{n}|A_i|$.

Our dataset consists of a set of voxel grids with varying occupancy, where an increase in voxel count is matched with a decrease in extinction coefficient to give overall identical throughput. Our motivation is to ensure that structures with equivalent properties are present in the dataset so that this regularity can be learned. Morphological operations are then applied, thus leading to structures with greater complexity.

To render the different configurations, we ray trace a single voxel and query the throughput by hit location on voxel layouts of increasing scale (see Fig 2 for some examples). The model performs well in predicting occlusion and transparency relative to the ground truth for cube sizes of 1, 2, and 4, each of which are present in the training data. On unseen larger volumes the wMAPE score begins to deteriorate, and the model appears to produce output for a lower frequency version of the input volume, resulting in correct overall shape but a loss of detail.

To render larger and more complex volumes, we divide the volume into sub-volumes, render them separately, and then composite them into a final volume, as shown in Fig 3. In the rightmost image, we split the full volume into $2 \times 2 \times 2$ sub-volumes, resulting in only 104 unique latent vectors that can be queried in parallel for throughput values, rather than processing the full 64^3 individual voxels.

Learning equivalence across different structures. We generate a set of optically equivalent structures and visualise the latent vector z_τ using t-SNE [4]. The output for an example configuration is shown in Fig 4. As the structures are equivalent per-angle, we expect that they cluster by illumination angle in high-dimensional space. As the figure shows (top row), this is indeed the case for the three volume sizes in the training dataset. While larger volumes are mapped to a different cluster (bottom row), each equivalent structure for these larger volumes does show similar clustering. Therefore, the latent space exhibits learned periodicity.

[1] Simon Kallweit, Thomas Müller, Brian Mcwilliams, Markus Gross, and Jan Novák. Deep scattering: Rendering atmospheric clouds with radiance-predicting neural networks. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017.

[2] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[3] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[4] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

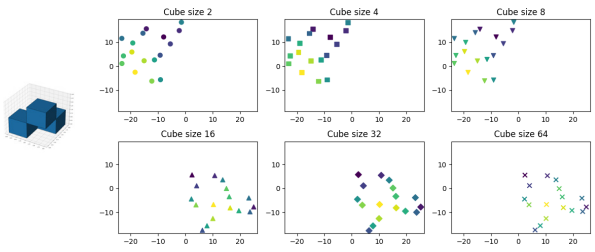


Figure 4: t-SNE plots of z_τ for optically equivalent structures. Different colours correspond to different conditioning illumination angles.

Refocus-NeRF: Focus-Distance-Aware Neural Radiance Fields Trained with Focus Bracket Photography

Yuki Yabumoto, Takuhiro Nishida, Takashi Ijiri Shibaura Institute of Technology

1 Introduction

Focus bracketing is a technique for quickly capturing a sequence of photographs by changing the focus distance. Focus stacking is a technique for synthesizing a single image from a sequence of photographs taken by focus bracketing, such that the image has a greater depth of field (DoF) and all subjects are entirely in focus. They are particularly useful for photographing small objects, such as flowers and insects, which require a macro lens with a very shallow DoF. Researchers have combined focus bracketing and focus stacking with photogrammetry to reconstruct three-dimensional (3D) shapes of small specimens [1, 4]. However, these photogrammetry-based methods have limitations in reconstructing transparent or specular objects.

Neural Radiance Fields (NeRF) [2] is a novel view synthesis method enabling reconstruction and rendering of 3D scenes from sparse 2D photographs. It represents a 3D scene with a multilayer perceptron (MLP) that takes a 3D location $\mathbf{x}$ and viewing direction $\mathbf{d}$ as inputs and outputs density and color at $\mathbf{x}$ viewed from $\mathbf{d}$. NeRF can reconstruct 3D scenes with transparent and specular objects in principle because it trains the MLP to render images with similar appearances to the input photographs. However, the original NeRF assumes input of deep DoF photographs and does not consider defocus blur effects. Wu et al. [5] added defocus blur effects to NeRF framework by blending the colors of neighboring rays. However, the method assumes that the radiance of each sampling ray is concentrated at a specific depth for screen space blending.

Our goal is to reconstruct a 3D scene from sequences of focus bracketing photographs, incorporating the inherent defocus blur effects. We present refocus-NeRF, which extends NeRF representation to receive the focus distance as well as the location and viewing direction as inputs. We train it with sequences of focus bracketing photographs taken from different viewpoints. Because we train the network with focus bracketing photographs, it can render images with defocus blur effects similar to actual photographs, even for scenes containing transparent or specular objects. Our method represents defocus blur as density and color in the 3D scene and dynamically modifies the scene according to the focus distance; it can render a scene using a standard camera ray sampling of the original NeRF without additional screen space blending.

2 Method

The refocus-NeRF model takes three inputs variables; location $\mathbf{x} \in \mathbb{R}^3$, viewing direction $\mathbf{d} = (\theta, \phi) \in \mathbb{R}^2$, and focus distance $f \in [0,1]$. The output of the model is the density $\sigma \in \mathbb{R}$ and RGB color $\mathbf{c} \in \mathbb{R}^3$ at the $\mathbf{x}$ viewed from $\mathbf{d}$. Because our method produces defocus blur effects by adjusting density and color, we input $\mathbf{x}$, $\mathbf{d}$, and f to the density MLP. We also apply multiresolution hash encoding to $\mathbf{x}$ and spherical harmonics encoding to $\mathbf{d}$ similarly to Instant-NGP [3].

We collect datasets to train the network as follows. We first perform focus bracketing from different viewpoints to obtain multiple sequences of photographs. We scale all photographs to align these in each sequence because the angles of view of photographs in a sequence slightly shift along with the change in focal distances (i.e., focus breathing). We define focus distance for a photograph in each sequence such that the photographs with the minimum and maximum focus distances have $f = 0.0$ and $f = 1.0$, respectively, while those in between have f values linearly interpolating 0.0 and 1.0. We next compute the focus stacking image for each bracketing sequence. Finally, we apply structure from motion to all focus stacking images to obtain their 3D poses.

We train the network using sequences of bracketing photographs after the alignment. We minimize the error between the observed photographs and the rendered images of refocus-NeRF from the corresponding viewpoints. Especially, when training with a photograph with focus distance f_i, we input f_i to the network, which enables us to obtain scene dynamically change according to focus distance. We use the Huber loss function for evaluating the error. We implemented our prototype system based on Instant-NGP [3].

Figure 1: We perform focus bracketing from multiple viewpoints and train a network using the obtained photographs (a). We synthesize from an unknown camera with different focus distances (b) and (c).

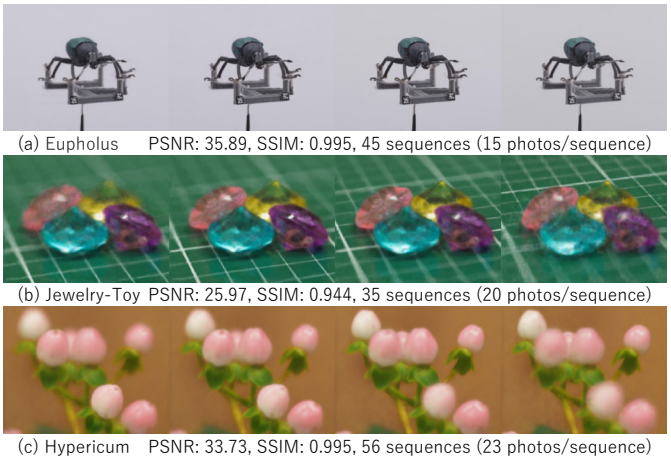


Figure 2: We render three scenes from an unknown camera pose with different f values. Each row shows the accuracy, the number of viewpoints, and the number of photos. We trained the network 40k steps (~15 min).

3 Results and Discussion

We reconstructed four different scenes to demonstrate the feasibility of the refocus-NeRF. Figures 1 and 2 present the rendering results from unknown camera poses with varying focus distances. Figure 2 summarizes the number of focus bracketing sequences (viewpoints) and photographs within each sequence. It also shows the accuracies (PSNR/SSIM) of the rendering results for viewpoints that were not used during training. Our method accurately rendered images containing out-of-focus blur. Our method dynamically modifies the scene using focus distance f; therefore, it can render images with different in-focus positions only with conventional camera ray sampling without screen space blending.

One limitation of our method is its memory cost; it requires more photographs than the traditional NeRF since multiple photographs exist at each viewpoint. In the future, we would like to extend our method to reconstruct similar scenes with fewer photographs. Another future work is to improve our approach to handle scenes beyond forward-facing scenes.

[1] T.-N. Doan and C. V. Nguyen. A low-cost digital 3d insect scanner. *Information Processing in Agriculture*, 2023.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 2021.

[3] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022.

[4] Y. Qiu, D. Inagaki, K. Kohiyama, H. Tanaka, and T. Ijiri. Focus stacking by multi-viewpoint focus bracketing. In *SIGGRAPH Asia 2019 Posters*, 2019.

[5] Z. Wu, X. Li, J. Peng, H. Lu, Z. Cao, and W. Zhong. Dof-nerf: Depth-of-field meets neural radiance fields. In *Proceedings of ACM MM '22*, 2022.

MV-SyDog: A Multi-View 3D Dog Pose Dataset for Advancing 3D Pose Estimation

Moira Shooter, Charles Malleson, Adrian Hilton
{m.shooter,charles.malleson,a.hilton}@surrey.ac.uk

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

Figure 1: Samples demonstrating the multi-view (a) and the different dog types aspect (b) from the MV-SyDog dataset.

Introduction: Computer-generated films featuring animals, like the Jungle Book (2016), and the Lion King (2019), often rely on 3D artists for animation. This process involves time-consuming and costly techniques such as key frame animation or motion capture (mocap). Animals and their unpredictable movements can pose additional challenges, making mocap difficult or impossible. While prior research has made significant progress in developing efficient and non-invasive techniques for human pose estimation using neural networks, the progress in the field of 3D animal pose estimation has been comparatively slower due to the scarcity of available animal pose datasets. Synthetic data offers advantages, such as the generation of extensive controlled datasets. However, models trained on synthetic data often demonstrate limited performance when evaluated on real-world datasets due to the domain gap.

Due to the lack of 3D animal pose datasets, the field of 3D animal pose estimation has primarily relied on available 2D pose datasets. While these approaches have shown promising results, the depth ambiguity still persists, particularly when observing subjects from different angles. The primary objective of this work is to address the issue of scarcity of datasets by generating and using synthetic data into our training process. We present preliminary 2D pose estimation results for dogs in both synthetic and real images to demonstrate the potential of our approach in capturing complex poses. Furthermore, these 2D results serve as a foundation and validation step towards our planned release of an extensive 3D dog pose dataset in the near future. Additionally, we show that our method effectively narrows the domain gap by leveraging features from DINOv2 [3]

Synthetic data generation and 2D pose estimation: We introduce MV-SyDog, a synthetic dataset featuring multi-view videos of dogs. We created this dataset using Unity3D and the Unity Perception package. It includes 1k videos, each lasting 2 seconds with 2D/3D pose ground truth and depth maps. We enhanced the dataset's ground truth by extending the Unity Perception package to capture 3D kinematic motion sequences for each video. For added scene diversity, we included five different dog models with textures, representing various dog breeds and sizes (Figure 1). To achieve realistic quadruped movements, we harnessed mode-adaptive neural networks [5] , enabling us to control animations by issuing commands such as walking, trotting, running, sitting and, lying down. Six virtual cameras were positioned in a circular arrangement around the subject to produce the multiple viewpoints. To introduce further variation, 420 different high dynamic range images (HDRIs) were used for realistic lighting and backgrounds. Randomness was introduced in the dog's appearance and pose, HDRIs, and view settings for each video. We included 3 dogs for training and 2 for testing. Additionally, for the test dataset certain HDRIs were excluded.

While our dataset is synthetic, it includes 3D labels and provides a greater variety of images compared to existing 3D animal pose datasets (as shown in Table 1). For instance, RGBD-Dog [2] was captured indoors with dogs wearing motion capture suits. We believe that due to the limited diversity in such datasets, networks trained on them exhibit poor generalisation capabilities when applied to in-the-wild images.

	Outdoor	Multi-view	# Species	# Keypoints	# Images
Ours	✔	✔	dogs	33	270K
RGBD-Dog [2]	✗	✔	dogs	41	136K
Animals3D [4]	✔	✗	40	26	3.4k (dogs: 1k)

Table 1: Comparison of MV-SyDog (Ours) with real 3D animal pose datasets mostly focused on dogs.

Our pose estimation network architecture exists of a feature extractor (DINOv2) and 1 head responsible for producing 2D joint heatmaps. This can be easily extended to predict 3D poses by for example adding location maps. We modify DINOv2 by unfreezing the last three layers and adding a dropout layer before the final layer.

Results: We use the percentage of correct keypoints (PCK ↑) and mean per joint position error (MPJPE ↓) to evaluate the network's performance. These metrics were normalised using the area of the segmentation map. Furthermore, we set the threshold of the PCK to 0.15. We achieve a PCK of 87.41 and MPJPE of 9.05 on the synthetic test dataset. While quantitatively the model performs poorly on real-world images (StanExt [1]) with a PCK of 36.18 and MPJPE of 25.97, qualitatively the model produces complete and plausible 2D poses by harnessing DINOv2 features (Figure 2), highlighting the superiority of DINOv2 over traditional backbones like ResNet. This is achieved without including real-world data samples into our model's training process. We believe the reason for not achieving higher qualitative results lies in the variations in keypoint semantics across datasets. This suggests that potential improvements in our results may be attainable through label refinement methods or the inclusion of real-world data.

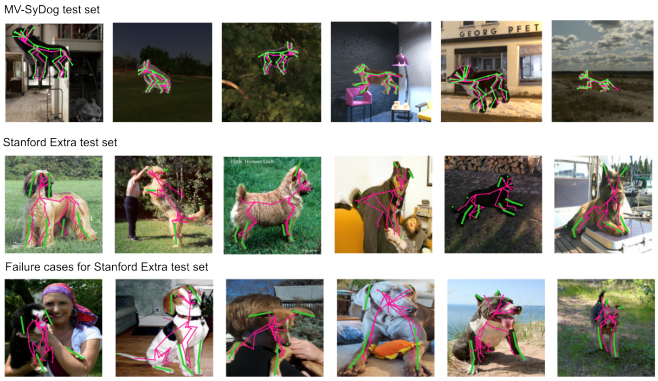


Figure 2: Qualitative results from network solely trained on the proposed synthetic data with ground truth (green) and predictions (pink). StanExt [1] contains only 24 labels whether the MV-SyDog contains 34 labels.

[1] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020.

[2] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[4] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, Wei Ji, Chen Wang, Xiaoding Yuan, Prakhar Kaushik, Guofeng Zhang, Jie Liu, Yushan Xie, Yawen Cui, Alan Yuille, and Adam Kortylewski. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9099–9109, October 2023.

[5] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Trans. Graph.*, 37(4), jul 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201366. URL https://doi.org/10.1145/3197517.3201366.

Artificial-intelligence Robotic Camera System for Live Soccer Broadcasts

Kota Takahashi
Toshie Misu
Kensuke Hisatomi
https://www.nhk.or.jp/strl/english/

Science & Technology Research Laboratories,
Japan Broadcasting Corporation

Program production for large-scale live sportscasting requires a large number of cameras and camera operators, but there is a shortage of skilled human resources. Therefore, the broadcasting industry requires automatic capturing techniques for sports broadcasting. Pixellot [1] and Veo Cam 2 [2] are automatic capturing cameras that are commercially available. Pixellot achieves automatic capturing by cropping a region of interest (RoI) in the panorama video of an entire field acquired from multiple cameras. Veo Cam 2 automatically generates a video including a ball by means of cloud artificial inteligence (AI) computing. We present an AI robotic camera system that automatically captures soccer games from the main (game-follow) camera position, which is roughly aligned with the half way line of the field. The system is capable of executing camerawork in accordance with the game situation by using a deep-neural-network model trained by the situation-dependent framing of skilled camera operators.

Figure 1 shows the configuration of the system, which consists of a sensor camera, a feature-extraction component, a framing component, a mechanical-control component, and a robotic camera. The sensor camera captures bird's eye images of the entire soccer field. To process the images, the feature-extraction component obtains a feature vector that consists of the positions, velocities, and head poses of the players and the position of the ball [3]. The framing component determines the capturing area with the framing AI trained on human-controlled camera operations. Finally, the mechanical-control component directs the camera motions depending on the capturing area on the basis of the feature map. This component has sliding mode control for high-speed coarse motion and proportional-derivative control for fine movements and stabilization. Each of these components executes processing independently and hands over the processed data in a best-effort manner. The feature-extraction component operates on a 0.1-s cycle. The framing component polls data (the feature map) from the feature-extraction component on an approximately 0.05-s cycle. The mechanical control component issues commands on an approximately 0.033-s cycle.

The framing AI refers to the feature map, state of play, and history of the capturing area over 2 s. The capturing area is described with three parameters: the in-plane coordinates [x,y] of the center of the RoI projected onto the ground plane and "half the width of the RoI r" measured on the ground plane. The framing AI determines the shooting area 1 s ahead using a predictive model trained with a feature map, because the end-to-end processing time from the imaging to the mechanical reaction is 1075 ms. As shown in Figure 2 (left), the camera movement by the 1-s prediction model best approximates the operator's maneuvering timing. The framing AI is modeled by a network with the following three stages: residual network (ResNet) [4], long short-term memory (LSTM) [5], and multilayer perceptron (MLP). LSTM suppresses temporal fluctuations of estimation resultes by taking into account the capturing areas in the previous 2 s, but this makes the framing AI have a slower response to changes in the match development. We therefore added another MLP network to only the training phase to short-cut the LSTM sub-network. This for-training-use-only network configures so-called ensemble learning, and its output elements are the same as the MLP. The network can respond quickly to changes occurring in the match, although the estimation results of the capturing area fluctuate from frame to frame because the time axis direction is not taken into account. The framing AIs neural network model is trained to minimize the loss function, which is the total value weighted in a trial-and-error manner against the mean squared error of two outputs. As shown in Figure 2 (right), the model based on ensemble learning responds more quickly to large changes than that without ensemble learning. The model with ensemble learning also yields more precise positioning.

Training data of the AI framing model is collected using two methods. The first one is for gatherings operators in-situ operation data in soccer stadiums. The data are measured using a pan-tilt head and a lens with rotary encoders installed on the camera. The other method is for obtain-

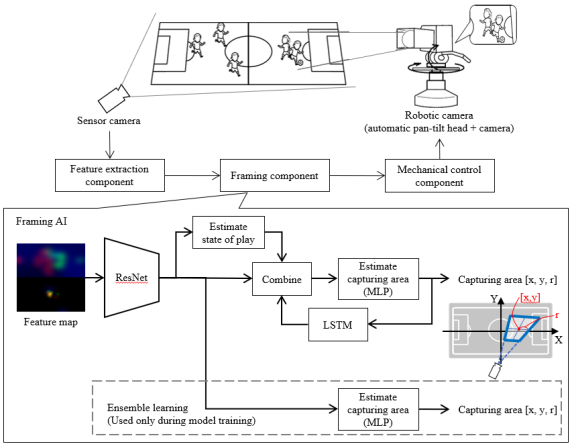


Figure 1: Configuration of AI robotic camera system

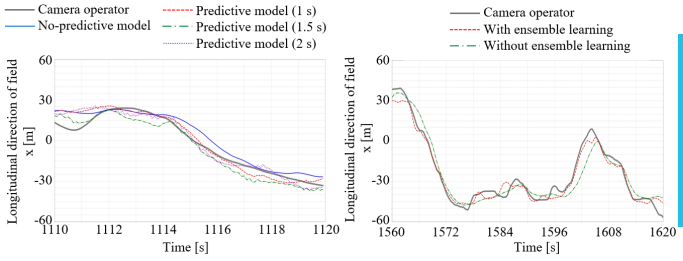


Figure 2: Effect of predictive models (left) and ensemble learning (right)

ings ex-situ data using a virtual-capturing system. The virtual-capturing system displays the entire soccer field on a large screen monitor, and the operator watches the monitor and manipulates a 3-degree-of-freedom measuring apparatus that measures the pan, tilt, and zoom operations. To simulate an actual capturing environment, another monitor (as a view finder) is installed on top of the apparatus to feed back the field-of-view of virtually captured video frames to the operator.

We conducted an automatic capturing experiment using our system with a trained predictive model incorporated into the framing component, and compared the video captured by the AI robotic camera with that captured by the human-operated camera. Our system could captured an area close to the human-operated camera. However, it was sometimes hard to see a camera image due to undesired tremors stemming from the fine-tuning of the angle and zoom in the system. A lag was also observed when the camera acceleration changed significantly. To solve this problem, we smoothed the capturing area in the time axis direction before processing in the mechanical control component, and the control parameters of the component were adjusted to values that emphasize responsiveness rather than suppressing fluctuations in the capturing area. Results of simulation, the timing of the motion and tracking speed were faster than at the time of the experiment after introducing the smoothing process. Fine adjustment of the focal length (related to the capturing range r of the capturing area) was also reduced.

[1] Automated Soccer Tracking Camera. https://www.pixellot.tv/sports/soccer

[2] Veo Cam 2. https://www.veo.co

[3] S. Yokozawa, M. Takahashi, H. Mitsumine, and T. Mishima. Head Pose Estimation for Football Videos by using Fixed Wide Field-of-View Camera. *ICPRS*, pages 83–88, 2017.

[4] K. Xe, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, pages 770–778, 2016.

[5] S. Hochreierr and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, pages 1735–1780, 1997.

Depth Reprojection for Mitigating Latency in XR Media Production

Dom Brown
dominic.brown@disguise.one
Sebastian Day
sebastian.day@disguise.one
Tom Whittock
tom.whittock@disguise.one

disguise Technologies Ltd
London, UK

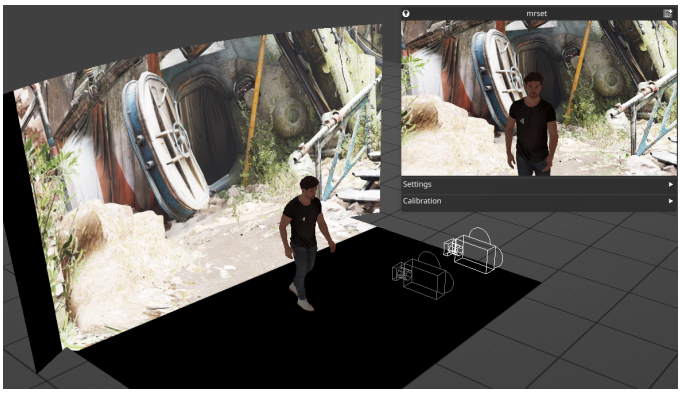


Figure 1: Virtual mock-up of latency in XR LED production. Content rendered at one camera location (grey) can be displayed at another (white).

Latency is a major issue in XR media production workflows involving LED volumes, where a the camera's location in physical space is used render a frame of the virtual content and projected onto the LED wall matching the camera's frustum (Figure 1). It can take up to 100s of milliseconds for the rendered image from a given perspective to be ready for display on the LED. This is a problem if the camera is moving, as when the rendered frame is displayed, its perspective is out of date. This results in noticeable perspective inaccuracies between the physical onset elements and the virtual content. The greater the round-trip latency, the greater this visual error.

Methods for reducing latency include post-render image warping techniques, which reproject content using homographic methods [1, 2, 3]. This works perfectly for rotational changes, but it does not address parallax inconsistencies with translational differences between perspectives.

We present Depth Reprojection, a technique that mitigates rendering latency by reprojecting content from the perspective it was rendered at to the perspective of the most up-to-date camera tracking using geometry generated on-the-fly from the content's depth map (Figure 2).

Each vertices' reprojected position v_r is determined by taking the original positions in the rendered content's image plane v_c (the x and y pixel locations in clip space and z depth value) and their original world space coordinates using the inverse of the content's View Projection matrix C^{-1}. The final clip space vertex position is determined by then applying the View Projection matrix of the most recent tracking reading R.

$$v_r = RC^{-1}v_c$$

This process is performed in real time using a graphics shader pipeline, with the geometry generation being performed by tessellation shaders that adaptively generate vertices based on the level of detail within the depth map.

The resulting effect is that the *global* error that is visible when planar reprojection is used is replaced with *local* areas of error around the transitions between depth planes, which are less noticeable to the human eye. SSIM metrics show that the Depth Reprojection method outperforms the standard homographic Planar method when reprojecting to account for translational differences between source and target perspectives (Figure 3).

The efficacy of this reprojection method in mitigating rendering latency for XR opens up promising opportunities: the ability to offload content rendering to the cloud and account for unpredictable network latency; spend more computation time rendering higher quality virtual content; and reduce the need for asset optimisation.

Figure 2: Top: Virtual content rendered in Unreal Engine. Bottom: Reprojection geometry generated in real time from the content's depth map.

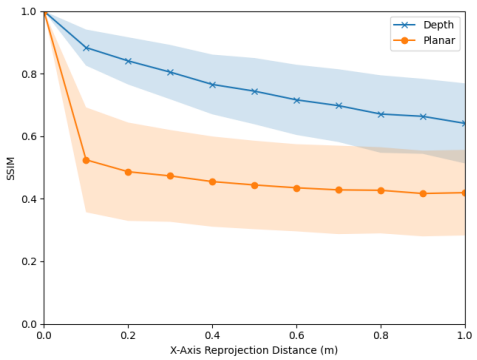


Figure 3: SSIM results from reprojecting at different horizontal offsets

[1] Ben Boudaoud, Pyarelal Knowles, Joohwan Kim, and Josef Spjut. Gaming at warp speed: Improving aiming with late warp. In *ACM SIGGRAPH 2021 Emerging Technologies*, SIGGRAPH '21, pages 1–4. Association for Computing Machinery, 2021. ISBN 978-1-4503-8364-6. doi: 10.1145/3450550.3465347. URL https://doi.org/10.1145/3450550.3465347.

[2] Joohwan Kim, Pyarelal Knowles, Josef Spjut, Ben Boudaoud, and Morgan Mcguire. Post-render warp with late input sampling improves aiming under high latency conditions. 3(2):1–18, 2020. ISSN 2577-6193. doi: 10.1145/3406187. URL https://dl.acm.org/doi/10.1145/3406187.

[3] OpenCV. OpenCV: Basic concepts of the homography explained with code, 2018. URL https://docs.opencv.org/4.x/d9/dab/tutorial_homography.html.

Real-Time Omnidirectional 3D Multi-Person Human Pose Estimation with Occlusion Handling

Pawel Knap[1] Peter Hardy[1] Alberto Tamajo[1] Hwasup Lim[2] Hansung Kim[1]
pmk1g20* p.t.d.hardy* at2n19* hslim@kist.re.kr h.kim*
*@soton.ac.uk

[1] University of Southampton
[2] Korea Institute of Science and Technology

Abstract

We present a multi-person 3D human pose estimation system that addresses a major limitation in existing models, namely the focus on single-person pose estimation. By using an off-the-shelf 2D detectors and 2D-3D lifting model, we first obtain the 3D poses of detected individuals in their own local coordinate system from video. We then use a radar sensing data and people-matching approach to localise the 3D poses within a global coordinate system accurately reconstructing the scene in real-time.

Introduction

Existing 3D human pose estimation (HPE) models primarily target single-person HPE, while our approach introduces a method for multi-person HPE. By using a 360° panoramic camera and mmWave radar sensors, our system effectively resolves depth and scale ambiguities. It uses a real-time occlusion-handling 2D-3D pose lifting algorithm [2] allowing for accurate performance capture both indoors and outdoors, all while remaining affordable and scalable. As evidence, we find that our system maintains a consistent time complexity irrespective of the number of detected individuals, achieving a frame rate of around 7-8 fps on a commercial-grade GPU. Our method revolves around transforming 2D body keypoints, detected by OpenPose[1], into predicted 3D keypoints in a global coordinate space obtained via radar sensing data. The system proceeds through several stages, which can be seen in figure 1.

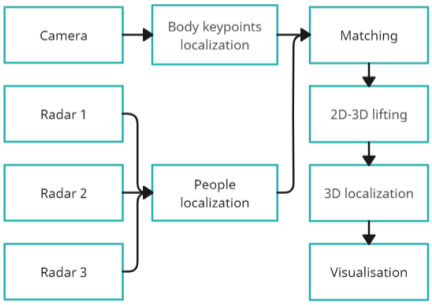


Figure 1: Overview of our approach, OpenPose extracts 2D keypoints from an image, while radars localize individuals. Subsequently, we match the individuals detected by both radar and camera systems. Next, the 2D keypoints are elevated into 3D, and their positions are adjusted within the global coordinate system using radar data.

System overview

Utilizing OpenPose [1], key body points' 2D Cartesian coordinates are extracted from the video frames, serving as inputs for the 2D-3D lifting model [2]. This algorithm employs a two-stage process, known as lift-then-fill, to address the problem of occlusion. Initially, it elevates the unobstructed 2D keypoints to form a partial 3D pose. Subsequently, an occlusion handling network completes missing joints caused by occlusions. The individuals detected by both radar and camera are matched using a binary search tree. This matching is based on the disparity between the average x-coordinate of 2D keypoints and the radar coordinates converted into the images x-coordinate space through a learned transformation. Finally, the detected pose is moved to its 3D coordinates by adding the radar-gathered positional data to 3D keypoints coordinates of identified individuals.

Results

Our system achieves a low average matching error of 4.63%, calculated as the disparity between a coordinate of correctly matched camera and radar individuals. The 2D-3D lifting algorithm [2] achieved competitive results with a PA-MPJPE of 37.2 and N-MPJPE of 61.7 on the GT 2D poses in the Human3.6M dataset, with qualitative results shown in Figure 2. Additionally, radar and camera calibrations reduced localisation errors by a few centimetres depending on the direction. We also conducted an ablation study which showed that the system performed consistently well with diverse poses, indoors and outdoors as shown in Figure 3. Furthermore, false positives were effectively filtered out. The system's primary constraint is the runtime of off-the-shelf 2D detectors, due to the large resolution present in 360° cameras. Additionally in our scenario, the radar detection range is approximately 3.5 meters.

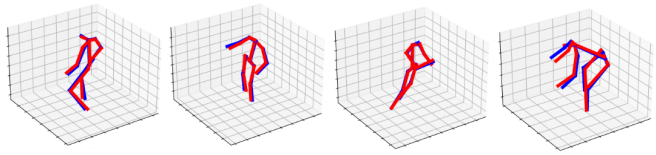


Figure 2: Qualitative pose reconstruction on the Human3.6M dataset. The GT 3D pose is in blue with our models predictions in red.

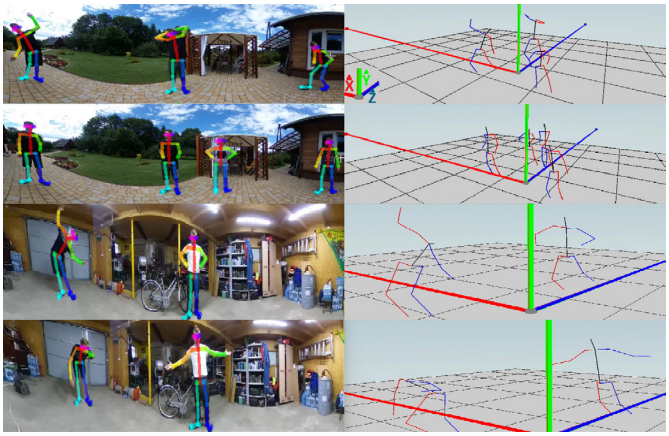


Figure 3: Actual poses captured by the camera, overlaid with OpenPose outputs, and reconstructed poses in the global 3D coordinate system.

Conclusion

Our real-time 3D multi-person detection system is robust, performs consistently regardless of the number of individuals, and theoretically can handle any number of detected people. The only limitations being, the speed of off-the-shelf 2D detectors and the range of the radar sensor. In our future work, we aim to develop a 2D detection approach that can process frames faster in high-resolution scenarios and extend the range of the radar which will enhance accuracy. Nonetheless, our contributions have paved the way for this system to be an affordable and dependable solution in the industry.

Acknowledgements

This work was partially supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction (EP/V03538X/1) and partially by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E32303).

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y.A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[2] Peter Hardy and Hansung Kim. Links - lifting independent keypoints - partial pose lifting for occlusion handling with improved accuracy in 2d-3d human pose estimation, 2023.

Using ML networks for turning 2D video into 3D Volumetric Video inside of Unreal Engine 5

Alex Grona
www.v3lox.com
Velox XR Limited

What is Velox XR

Velox XR is a cutting-edge extended reality technology solution that allows for real-time streaming into the Unreal Engine. Our platform streamlines the extended reality production process, making it faster, simpler, and more cost-effective for content creators, game developers, and virtual production studios.

Overview

The demo will show a workflow of creating interactive immersive experiences from live videos taken during the event.

Product

Our revolutionary 3D video format (VLX), currently patent-pending, enables us to reconstruct real-life events and environments from video footage captured on mobile devices with depth cameras. Using our fully operational workflow, we can effortlessly import any video footage with people into an Unreal Engine project, which enables us to composite it in real-time based on depth. This dynamic 3D mesh integration of individuals into any Unreal Engine 5 scene provides a stunningly realistic and fully animated experience, identical to its original filming, with the added benefit of eliminating the background.

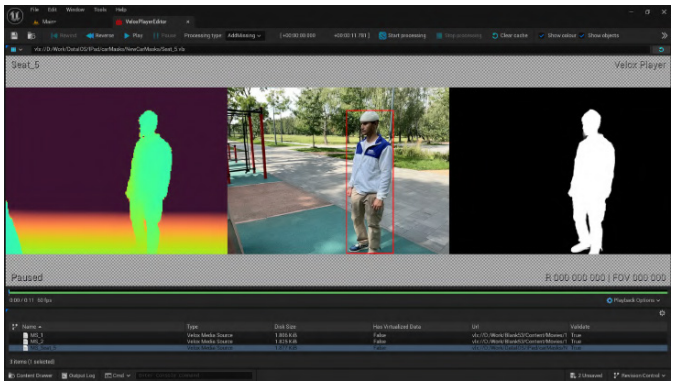


Figure 1: A view of the Velox Player Editor UI in Unreal Engine 5 showing three video channels:
1. Depth,
2. Video RGB with Region of Interest (ROI),
3. Mask

Velox XR will present two products at the show:

- An iOS [app] using Apple LiDAR for recording live footages in the VLX format
- Three UE5 plugins for creating volumetric videos
  - **VeloxPlayer**: Media framework extension to UE5 for playing VLX files
  - **VeloxPlayerPlus**: Extension to the free plugin for editing, saving and publishing VLX files
  - **VeloxNeuro**: ML inference engine.

Figure 2: Showcase video

Unique features:

- Mobile solution: no green screen, no special studio and no hardware calibration
- Real-time 3D reconstruction in UE5 on low spec hardware incl. Android (mobile, VR)
- Hardware agnostic: can support any depth camera available on the market
- !! Live streaming !! **NOT RELEASED YET** and can be introduced to the public at CVMP

Figure 3: Examples from presenting a demo @CVPR 2023

Machine Learning networks

The live video editing feature currently offers a number of ML models for **human tracking**, **segmentation** and **alpha matting mask generation** directly in the UE Editor**.**

Drawing from Motion Capture: Performance, 3D Drawing and Character Animation

Ricardo Megre
https://aimcreativestudios.com
AIM Creative Studios

This demo presentation aims to explore novel approaches of combining motion capture with drawing and 3D animation. As the art form of animation matures, possibilities of hybrid techniques become more feasible, even on smaller and independent projects, where crosses between traditional and digital media provide new opportunities for artistic expression.

The research was developed for a PhD thesis in the areas of 3D animation and motion capture, resulting in the initial development of a short-film titled Out-of-balance. The project is at the moment being developed at AIM Creative Studios.

Out-of-balance is an animated film that tells the story of Alice searching for herself and her place in the world. The story unfolds in a universe created from the drawn line. Alice lives trapped in her threads and the characters that cross her path are also built by lines.

Figure 1: 3D data representing the actors movements (left). Drawing of main character Alice, based on the actors performance (right).

The project focuses on alternative production methods that do not depend on mocap retargeting, and provide animators with greater options for experimentation and expressivity. As motion capture data is a great source for naturalistic movements, it was combined with interactive methods such as digital sculpting and 3D drawing.

The hybrid animation technique seeks to combine the expressiveness of drawing, based on the actors performance, with 3D animation.

The first step was the recording of a motion capture session where topics such as animation principles and drawing principles were used as motivation for the actors improvisation. Concepts such as “squash” and “stretch”, “twisting” and “rotation” were used to make the actors movements visible. Through drawing, the construction and rhythm of different motions was represented and studied.

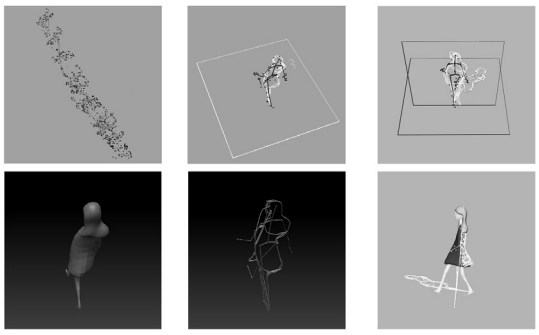


Figure 2: Process from mocap data to drawing on planes and finally on volumes.

The recorded data was used for the construction of the storyboards, informing on camera angles and performance of the characters. The ideas developed during this process did not necessarily match the actor’s original performance. Instead, many of the movements were deconstructed into key poses, re-timed and changed to better serve the story.

Secondly, 3D drawings were done on top of the mocap data and storyboards, using NURBS planes as canvases in Autodesk Maya. 3D lines and armatures were used in ZBrush as a reference to sculpt the volumes, which were later converted into NURBS planes for drawing the final lines in 3 dimensions.

Figure 3: Visual development and animation, combining 3D drawing with path-tracing rendering.

Having motion capture as a base reference helped significantly in order to find moments that could support the creation of the narrative, while offering the possibility to interact within the 3D space.

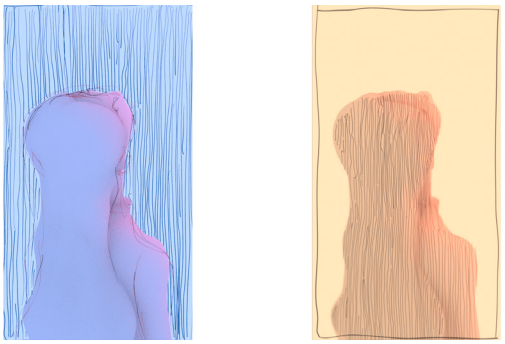


Figure 4: Using 3D lines to describe positive and negative spaces.

Lines were used to convey the positive and negative spaces of the composition, describing Alice’s inner and outer worlds. The lines were drawn vertically, although they are drawn on 3D surfaces. This expands the possibilities that two-dimensional media offer, as the camera can be moved through and around the lines, lights and depth of field can also be added. Deformers have been added to the lines, creating movement with a wavelike effect.

These methods are used for the visual development of the film, where they become relevant for the creation of a specific visual language, that can be used to articulate concrete ideas for storytelling in animation.

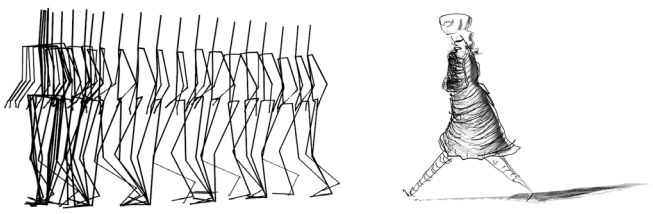


Figure 5: Selection of actor’s poses for the animation of Alice getting lost.

LightMoCap: Light-weight, Real-time & Scalable Markerless Motion Capture

Georgios Albanis[12]
giorgos@moverse.ai
Nikolaos Zioulis[1]
nick@moverse.ai
Anargyros Chatzitofis[1]
argyris@moverse.ai
Spyridon Thermos[1]
spiros@moverse.ai
Vladimiros Sterzentsenko[1]
vlad@moverse.ai
Kostas Kolomvatsos[2]
kostasks@uth.gr

[1] Moverse
moverse.ai
[2] Department of Informatics and Telecommunications, University of Thessaly

1 Introduction

Markerless motion capture (MoCap) technology emerges as an important technology for virtual and live productions. Conventional MoCap methods have long relied on encumbering full-body suits, which inadvertently constrain actor performance and limit the creative flexibility of producers. The real-time aspect is not only crucial for live/virtual productions, as pre-visualization is very useful even for post-production and can facilitate rapid prototyping and rehearsals. While modern data-driven monocular solutions can be efficient, they lack robustness and do not produce metric-scale outputs. Further, multi-view markerless MoCap solutions come with resource-intensive demands and despite that still suffer from prolonged processing times. In this demonstration, we will present *LightMoCap*, a multi-view markerless MoCap system that is highly scalable in terms of the camera-to-resources ratio.

2 Approach

LightMoCap relies on fitting an articulated parametric body $\mathcal{B}$ [2] to multi-view keypoint observations. The body function $(\mathbf{v},\mathbf{f}) = \mathcal{B}(\beta,\theta,\mathbf{T})$ jointly encodes shape, pose and global pose as a mesh surface comprising vertices v and faces $\mathbf{f}$, through a set of blendshape coefficients $\beta \in \mathbb{R}^{10}$, per joint $j := \{1,\ldots,J\}$, rotations $\theta \in \mathbb{SO}^{3\times J}$, and the transform $\mathbf{T} \in \mathbb{SE}^3$, respectively. A set of K landmarks $\ell \in \mathbb{R}^{3\times K}$ can be extracted using a linear function $\ell = \mathbf{R} * \mathbf{v}$, with $\mathbf{R}$ being a regressor matrix. Assigning proper landmarks to match the keypoints estimated by a 2D pose estimation model defines a set of correspondences upon which a minimization problem can be formulated.

To estimate the human pose and shape at a specific time instance t, an objective function with two λ-weighted components is used, a data term, $\mathcal{E}_{data}$, and a prior term, $\mathcal{E}_{prior}$:

$$\underset{\beta,\theta,\mathbf{T}}{\operatorname{argmin}}\, \lambda_{data}\mathcal{E}_{data} + \lambda_{prior}\mathcal{E}_{prior}. \tag{1}$$

The data term is an L2 image domain distance error between the projections of the landmarks on each viewpoint $p := \{1,\ldots,P\}$ via each viewpoint's projection function π_p and the corresponding estimated keypoints $\mathbf{k}^p$:

$$\mathcal{E}_{data} = \sum_{p}^{P}\sum_{k}^{K} ||\pi^p(\ell_k) - \mathbf{k}_k^p||_2^2. \tag{2}$$

The prior term is a data-driven joint angle distribution regularizer, relying on a per joint representation learning model [4] $\mathcal{M}_j$ that maps each joint's rotation to a Gaussian distributed latent variable $\mathbf{z}_j = \mathcal{M}_j(\theta_j)$:

$$\mathcal{E}_{prior} = \sum_{j}^{J} ||\mathcal{M}(\theta_j)||_2^2. \tag{3}$$

This is an important term as it regularizes the system's solution to valid poses even despite the lack of proper keypoint estimates to constrain the rotation of each joint.

3 Demonstration

We built a system using edge AI enabled devices, namely the Luxonis OAK-D Pro PoE [3]. This ensures the system's scalability in terms of the viewpoint-to-resources ratio. The edge devices are capable of neural inference, and thus, each of them hosts a light-weight keypoint estimation model [1]. Multiple devices can be connected on a single network interface and only need to transmit a light-weight payload, the estimated keypoints $\mathbf{k}^p$. Viewpoint metadata like the cameras' intrinsic parameters are available for each device, while a preparatory extrinsic calibration process also provides the viewpoints' spatial poses.

For each new subject entering the capturing volume a body calibration process initializes the system. Eq. (1) is optimized and a temporal average of the shape coefficients β initializes the subject's body shape and skeleton. Post initialization, a purely kinematic fit of the skeleton is performed using Eq. (1), fixing the shape coefficients β and only optimizing the pose related ones, θ and $\mathbf{T}$. This results in a light-weight process that can be optimized in real-time on a single CPU. The system runs at the rate of the keypoint estimators, which perform inference at 25Hz, using no GPU resources for neither fitting nor pose estimation. Nonetheless, future work can additionally exploit the availability of GPU compute to improve fitting performance.

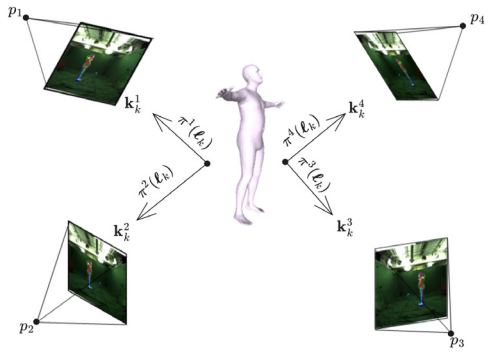


Figure 1: *LightMoCap* fits an articulated template mesh to 2D keypoint observations inferred on a set of edge AI devices, regularized by a joint angle prior term. The distribution of the processing allows for CPU-only solving of the body motion in real-time.

[1] Ivan Grishchenko, Valentin Bazarevsky, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, Richard Yee, Karthik Raveendran, Matsvei Zhdanovich, Matthias Grundmann, and Cristian Sminchisescu. BlazePose GHUM Holistic: Real-time 3d human landmarks and pose estimation. *arXiv preprint arXiv:2206.11678*, 2022.

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[3] Luxonis. OAK-D: Stereo camera with edge ai, 2020. URL `https://luxonis.com/`. Stereo Camera with Edge AI capabilities from Luxonis and OpenCV.

[4] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Demo: Audio-Driven Video Composition

Tim Rumpf
tim.a.rumpf@filmuniversitaet.de
Film University Babelsberg Konrad Wolf, Potsdam
Jakub Fiser
fiser@adobe.com
Adobe Research, London

Synchronizing video transitions and effects with a music beat is a common and powerful way to elevate the impact of multimedia content and foster an emotional connection with the viewer. However, achieving good synchronization manually can be a painstaking and error-prone process, often requiring frame-by-frame adjustments. Additionally, it becomes challenging to easily swap resources, such as exploring different audio tracks, while maintaining synchronization with the rest of the composition.

Professional editing software like Adobe Premiere Pro [1] offers fine-grained, frame-level control over the synchronization process, allowing skilled editors to meticulously align and time their visuals to match the music beat. However, this level of control demands expertise and deep domain knowledge, making it inaccessible to many content creators, especially novices.

On the opposite side of the spectrum, tools like Canva's Beat Sync [2] aim to simplify the process by offering an "auto-magical" approach. These tools typically work in a black-box fashion and automatically synchronize video transitions with music, and thus require minimal user intervention. While such approach is user-friendly, it often sacrifices creative control, leaving content creators with limited or no customization options.

Striking a balance between these two approaches is the challenge our demo system attempts to address. That is, we don't subject the user to laborious frame-level adjustments but still allow for meaningful higher-level editing operations such as changing the location of individual transitions while adhering to the beats or global adjustments like overall composition length and pacing.

Our demo is comprised of three key components:

1. Input Analysis
2. Effect Library
3. Matchmaking System

As input, our system ingests a single audio track, one or more videos, and optional set of short texts. We detect events in the audio, such as beat and onset times and peaks in multiple streams of audio analysis data, namely: amplitude, frequency spectrum, percussive and harmonic components, and pitch. For each video input, we detect cuts. While the audio is then used unchanged, each video sub-segment can be further trimmed to enable perfect transition timing.

Our effect library contains elements that are all authored ahead of time, and are used either at the transition times, or in between them. They can be divided in four categories:

- Video Filters: Directly manipulate the video's final appearance, altering pixels with effects like distortions, shifts in exposure or colors, glitching, blurring, or overlays like flares or shapes. Video filters are applied between transitions. Some of their parameters (e.g., brightness) are continuously driven by mapped data streams from the audio analysis, making them directly audio-reactive.
- Video Transitions and Cuts: Short-duration video filters with one or more animated parameters. Triggered at the transition times to be aligned with the audio events that initiate a cut in the video.
- Text Animations: Animations of text elements triggered at the transition times. Our system has a hierarchical representation of text elements from scene- down to glyph-level, which allows for authoring of rich and diverse set of animations. Additionally, each level of the animation hierarchy has its own masks that can be animated independently.
- Text Filters: Manipulate the final appearance of the text elements after their animations have been resolved. Driven by the continuous data streams from the audio analysis. Rendered on top of the video.

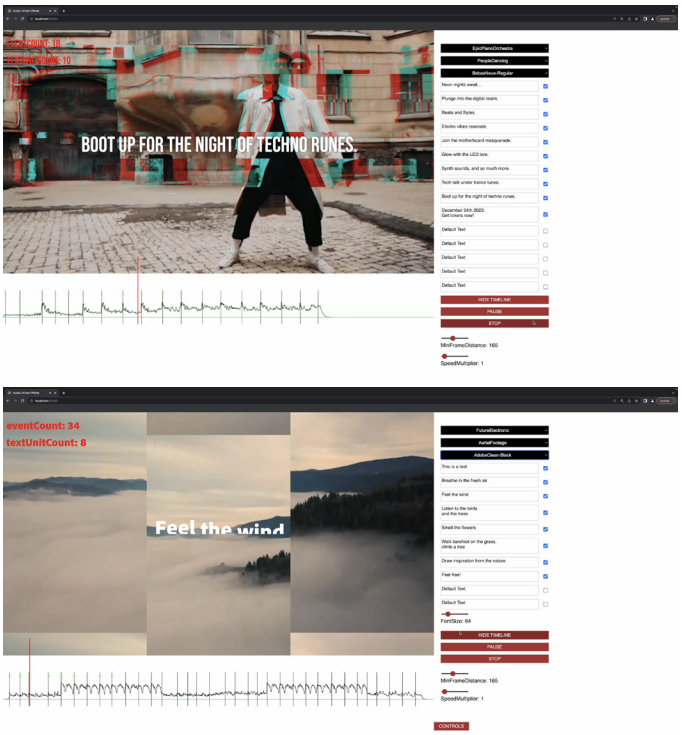


Figure 1: Screenshot of the demo interface. Above: Video playback with applied filter. Below: Active text animation and transition. The timeline gives feedback about the composition's overall pacing and the transition events' location. On the right-hand side, the user can tweak the template parameters and customize the animated text that is rendered on top of the video. Upon any parameter change, the affected parts of the composition are recomputed instantly. A particular choice of effects is baked into the template to maintain stylistic coherence.

In our matchmaking system, we apply and synchronize the effects with the user-given input. This process involves combining and mapping various effects within predefined templates and then applying and pairing them with the user-defined input. Our primary decision-making factor is the audio analysis. We score the detected audio events based on factors like peak significance, distribution across the audio track, emphasis on percussive elements and spectral edges, and preference for events synchronized with the beat. We also estimate the on-screen times of the input text when selecting the most suitable audio events for synchronization. To facilitate quick exploration of stylistically coherent results, we have constructed templates that package subsets of the effects from the effect library, ensuring that users can maintain a consistent and polished visual aesthetic.

The user experience of our system is depicted in Figure 1. The demo lets the user to customize one of the predefined templates and experiment with visual aspects of the composition and its pacing. The feedback is instant, and our system also allows the composition parameters to change even when the video is playing, which helps to iterate to the desired result quickly.

[1] Adobe. Premiere Pro. https://www.adobe.com/products/premiere.html, 2023. Version: 23.0.
[2] Canva. Beat Sync. https://www.canva.com/features/beat-sync/, 2023. Accessed: 2023-09-15.

TECHNICAL AWARDS

To celebrate the 20th anniversary of the CVMP conference, we are pleased to announce the inaugural CVMP Technical Awards in partnership with InnovateUK.

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

Innovate UK

NOTES

CHAIRS

Conference Chairs
Marco Volino, University of Surrey, UK
Armin Mustafa, University of Surrey, UK

Full Papers Chair
Peter Vangorp, Utrecht University, Netherlands

Short Papers & Demos Chair
Peter Eisert, Humboldt University, Germany
Claudio Guarnera, University of York, UK

Industry Chair
Oliver Grau, Intel, Germany
Abi Bowman, Disguise, UK

Sponsorship Chair
Jeff Clifford, Evastute/Milk VFX, UK
Peri Friend, Foundry, UK

Local Arrangements Chair
Hansung Kim, University of Southampton, UK

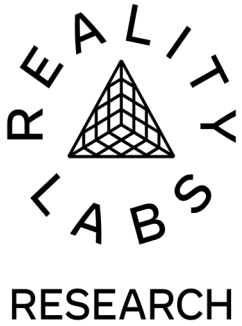
Public Relations Chair
Da Chen, University of Bath, UK
Moira Shooter, University of Surrey, UK

Conference Secretary
Emily Ellis, University of York, UK

Programme Committee
Kevin Matthe Caramancion, University of Wisconsin–Stout
Da Chen, University of Bath
Robert Dawes, BBC Research
Daljit Singh Dhillon, Clemson University
Peter Eisert, Humboldt University
Zhenhua Feng, University of Surrey
Elena Garces, Seddi
Joe Geigel, Rochester Institute of Technology
Andrew Gilbert, University of Surrey
Oliver James, DNEG
Hansung Kim, University of Southampton
Rafal Mantiuk, University of Cambridge
Kenny Mitchell, Edinburgh Napier University / Roblox
Marco Pesavento, University of Surrey
Christian Richardt, Meta Reality Labs Research
Nadejda Roubtsova, University of Bath
Moira Shooter, University of Surrey
Graham Thomas, BBC
Peter Vangorp, Utrecht University
Zhidong Xiao, Bournemouth University

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

Conference Sponsors 2023



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