CVMP 2021

The 18th ACM SIGGRAPH European Conference on Visual Media Production
6–7 December 2021, London

Programme

The British Library, London
6th & 7th December 2021
https://www.cvmp-conference.org/2021/

Conference Sponsors 2021

ACM ISBN: 978-1-4503-9094-1
Copyright © 2021 by the Association for Computing Machinery, Inc

Message from the Chairs

We are pleased to introduce the programme for the 18th ACM SIGGRAPH European Conference on Visual Media Production (CVMP). For almost two decades, CVMP has built a reputation as the prime venue for researchers to meet with practitioners in the Creative Industries, including film, broadcast, games, immersive and beyond.

This year, we are cautiously returning to an in-person conference in London with an exciting programme that continues our signature mixture of keynotes, peer-reviewed papers, invited talks and special sessions, with speakers from across academia and industry. Our focus this year is on *digital and virtual humans* and the impact they have across the Creative Industries.

CVMP has firmly established itself as a leading venue for anyone interested in visual computing and its applications. This year, we are delighted to host an exciting array of keynote speakers from both academia and industry. Cengiz Öztireli (University of Cambridge) will focus on techniques at the intersection of computer graphics, vision, and machine learning for capturing and replicating the visual world in 3D. Darren Cosker (Microsoft) will be taking us on a journey into the metaverse and discuss what will be necessary to create compelling immersive experiences in mixed reality. Siyu Tang (ETH Zurich) will present her research on learning to bring 3D scenes alive by plausibly populating them with realistic synthetic 3D humans. Tobias Ritschel (University College London) completes the line-up by taking a fresh look at image synthesis for future near-eye VR displays using a perceptually inspired approach.

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 seven full papers and nine 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 a 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.

A small conference like CVMP in an expensive city like London would not be viable without the generous support of our sponsors. We sincerely thank our gold (YouTube, Autodesk, NVIDIA), silver (Adobe, Foundry, Vicon) and bronze sponsors (Activision, IO Industries, CAMERA, CVSSP) without whom this conference would not be possible. Finally, we would like to thank everyone who submitted their work to CVMP this year, the invited speakers, the reviewers and the organising committee for their hard work in putting together CVMP 2021!

Rafał Mantiuk and Christian Richardt (Conference Chairs)
Marco Volino (Full Papers Chair)
Armin Mustafa (Short Papers Chair)
Duygu Ceylan and Ilke Demir (Industry Chairs)
Bernhard Egger (Sponsorship Chair)
Peter Vangorp (Public Relations)
Alex King (Conference Secretary)

DAY 1 | Monday 6th December 2021

Location: The British Library, Knowledge Centre

09:00 **Registration opens with Coffee |** Knowledge Foyer

09:30 **Chairs' Welcome | Rafał Mantiuk**, University of Cambridge
Christian Richardt, University of Bath

09:40 SESSION 1 **| Bring on the colours!**

1. **Semantic-driven colorization**
   Man M. Ho (Hosei University), Lu Zhang (INSA Rennes),
   Alexander Raake (TU Ilmenau), Jinjia Zhou (Hosei University)
2. **Arnold 7 update** (industry talk)
   Frederic Servant (Autodesk)
3. **Photometric stereo with area lights for Lambertian surfaces**
   Jiangbin Gan and Thorsten Thormählen (University of Marburg)
4. **Material acquisition and editing** (industry talk)
   Valentin Deschaintre (Adobe)

11:00 **Break with Coffee |** Knowledge Foyer

11:30 KEYNOTE 1 **| Cengiz Öztireli**, University of Cambridge

12:30 SPOTLIGHT SESSION

12:40 LUNCH, POSTER AND DEMO SESSION **|** Knowledge Foyer

14:30 INDUSTRY SPECIAL SESSION **| Digital humans**

5. **Volumetric video at the intersection of visual effects and virtual production**
   Juraj Tomori, Charles Dupont, George Ash and Mike Pelton (dimension)
6. **High-performance multi-camera systems for volumetric capture and 4D face/body scanning**
   Andrew Searle (IO Industries Inc)
7. **The creation of 3D human datasets for CV research**
   Lukas Lamprecht (Renderpeople)

15:30 POSTER AND DEMO SESSION **|** Knowledge Foyer with Coffee

16:00 KEYNOTE 2 **| Darren Cosker**, Microsoft

17:00 NETWORKING RECEPTION **|** Knowledge Foyer

18:30 **Close**

DAY 2 | Tuesday, 7th December 2021

Location: The British Library, Knowledge Centre

09:00 **Registration opens with Coffee |** Knowledge Foyer

09:30 SESSION 2 **| And Action!**

8. **Automatic camera control and directing with an ultra-high-definition collaborative recording system**
   Bram Vanherle, Tim Vervoort, Nick Michiels and Philippe Bekaert (Hasselt University)
9. **Contact-rich simulation in NVIDIA Omniverse** (industry talk)
   Kier Storey and Michelle Lu (NVIDIA)
10. **FacialFilmroll: High-resolution multi-shot video editing**
    Bharath Bhushan Damodaran, Emmanuel Jolly (InterDigital R&D), Gilles Puy (In his own name), Philippe-Henri Gosselin, Cédric Thébault, Junghyun Ahn (InterDigital), Tim Christensen, Paul Ghezzo (In his own name), Pierre Hellier (InterDigital)
11. **Foundry and machine learning** (industry talk)
    Ben Kent (Foundry)

10:50 **Break with Coffee |** Knowledge Foyer

11:20 KEYNOTE 3 **| Siyu Tang**, ETH Zurich

12:20 LUNCH, POSTER AND DEMO SESSION **|** Knowledge Foyer

14:00 SESSION 3 **| Gimme the data!**

12. **Depth estimation from a single omnidirectional image using domain adaptation**
    Yihong Wu, Yuwen Heng, Mahesan Niranjan and Hansung Kim (University of Southampton)
13. **VPN: Video provenance network for robust content attribution**
    Alexander Black, Tu Bui (University of Surrey), Simon Jenni (Adobe Research), Viswanathan (Vishy) Swaminathan (Adobe), John Collomosse (Adobe Research)
14. **High-fidelity procedural data synthesis for validation and training of perception function** (industry talk)
    Oliver Grau and Korbinian Hagn (Intel)
15. **Speech-driven conversational agents using conditional flow-VAEs**
    Sarah Taylor, Jonathan Windle, David Greenwood (University of East Anglia), Iain Matthews (Carnegie Mellon University)

15:20 POSTER AND DEMO SESSION **|** Knowledge Foyer with Coffee

15:50 KEYNOTE 4 **| Tobias Ritschel**, University College London

16:50 **Announcements and Prizes | Rafał Mantiuk**, University of Cambridge
Christian Richardt, University of Bath

17:00 **Close**

KEYNOTE 1 | Cengiz Öztireli
University of Cambridge

3D Digital Reality – Modeling for Perception
Monday 6th December 2021, 11:30

Creating digital models of reality is one of the grand challenges of computer science. In this talk, I will summarize some of our efforts towards achieving this goal to allow machines to perceive the world as well as and beyond humans. The focus will be on capturing and replicating the visual world and techniques at the intersection of computer graphics, vision, and machine learning to solve several fundamental problems and their practical applications.

Cengiz Öztireli

Cengiz Öztireli is an Associate Professor at the University of Cambridge and a Senior Researcher at Google. He previously worked as a Research Scientist at Disney Research, and as a Senior Research Associate at ETH Zürich. He obtained his M.S. and Ph.D. degrees in computer science from ETH (jointly funded by the Swiss National Science Foundation) and completed a double major in computer and electronics engineering at Koç University (summa cum laude, valedictorian). He has been honored with several awards including the Eurographics Best Ph.D. Thesis Award, Fulbright Science and Technology Award, and the UKRI Future Leaders Fellowship.

KEYNOTE 2 | Darren Cosker
Microsoft

Creating Presence in Mixed Reality and the Metaverse
Monday 6th December 2021, 16:00

Imagine being able to have a conversation with someone who is hundreds of miles away, but it feels like they are actually there. Technology which can achieve this would change society — bringing distant family and friends closer together, transforming the way we work and reducing carbon footprints. However, creating compelling interactive experiences involving other people in mixed reality and the metaverse is a challenging task combining expertise in computer vision, graphics, AI and engineering. In this talk, I will examine some of the technologies required to make this a reality and the challenges ahead.

Darren Cosker

Darren Cosker is a Scientist at Microsoft's Mixed Reality and AI laboratory (Cambridge), and holds a part-time full Professor position at the University of Bath. He was previously the founding Director of CAMERA (2015–2021) — a multi-disciplinary research centre based dedicated to understanding and modelling human motion and appearance. At Microsoft, Darren is helping realise the vision of 'presence' in mixed reality and the metaverse through products such as Microsoft Mesh. Prior to joining Microsoft, Darren held personal research fellowships from the Royal Society (2012–2016) and the Royal Academy of Engineering (2007–2012).

KEYNOTE 3 | Siyu Tang
ETH Zurich

Learning to capture and synthesise 3D humans in 3D scene
Tuesday 7th December 2021, 11:20

In recent years, many high-quality datasets of 3D indoor scenes have emerged such as Replica and Gibson, which employ 3D scanning and reconstruction technologies to create digital 3D environments. Also, virtual robotic agents exist inside of 3D environments such as the Habitat simulator. These are used to develop scene understanding methods from embodied views, thus providing platforms for indoor robot navigation, AR/VR and many other applications. Despite this progress, a significant limitation of these environments is that they do not contain people. The reason such worlds contain no people is that there are no fully automated tools to synthesise realistic people interacting with 3D scenes naturally, and manually doing this requires significant artist effort. In this talk, I will present our previous and ongoing research about capture and synthesis of realistic people interacting realistically with 3D scenes and objects.

Siyu Tang

Siyu Tang is an assistant professor at ETH Zürich in the Department of Computer Science since January 2020. She received an early career research grant to start her own research group at the Max Planck Institute for Intelligent Systems in November 2017. She was a postdoctoral researcher in the same institute, advised by Dr. Michael Black. She finished her PhD at the Max Planck Institute for Informatics and Saarland University in 2017, under the supervision of Professor Bernt Schiele. Before that, she received her Master's degree in Media Informatics at RWTH Aachen University, advised by Prof. Bastian Leibe and her Bachelor degree in Computer Science at Zhejiang University, China. She has received several awards for her research, including the Best Paper Award at BMVC 2012 and 3DV 2020, Best Paper Award Candidates at CVPR 2021, an ELLIS PhD Award and a DAGM-MVTec Dissertation Award.

KEYNOTE 4 | Tobias Ritschel
University College London

Perceptually-inspired VR Image Synthesis
Tuesday 7th December 2021, 15:50

Images shown on future near-eye displays will be perceived differently. In this talk I will argue that, hence, all image synthesis itself will need to change. I will mostly discuss means to reduce bandwidth and/or latency. This can be achieved by rendering images that are perceived like other images directly (Ventral Metamers), by changing how the deepest internals of graphics hardware work (Perceptual Rasterization) or by switching to a domain different from pixels (Laplacian).

Tobias Ritschel

Professor Tobias Ritschel has received his PhD from Saarland University (MPI) in 2009. He was a post-doctoral researcher at Telecom ParisTech / CNRS 2009–10 and a Senior Researcher at MPI 2010–15. Tobias was appointed Senior Lecturer at University College London in 2015, where he was named Full Professor of Computer Graphics in 2019. His work has received the Eurographics Dissertation (2010) and Young Researcher Award (2014). His interests include Image Synthesis and Human Visual Perception, now frequently including applied AI.

INDUSTRY SPECIAL SESSION ON DIGITAL HUMANS

Volumetric video at the intersection of visual effects and virtual production

Juraj Tomori, Charles Dupont, George Ash and Mike Pelton (dimension)

Juraj Tomori, software developer at dimension, has experience with volumetric video, facial performance capture and processing and deep domain knowledge of visual effects. Positioned at the intersection of research and development, he is eager to get the latest innovations into production.

High-performance multi-camera systems for volumetric capture and 4D face/body scanning

Andrew Searle (IO Industries Inc)

IO Industries Inc. develops market-leading specialty video cameras, designed with the needs of VR/AR/XR content generators in mind. Whether it's a pair of cameras in a 3D stereoscopic rig, a handful of cameras for a 360° VR configuration, or an array of over 100+ cameras set up for volumetric video capture, IO Industries cameras have the features and flexibility it takes to make these configurations happen.

The creation of 3D human datasets for CV research

Lukas Lamprecht (Renderpeople)

Huge datasets of scanned 3D humans are a necessity to solve many current persistent and human-related problems within the CV industry, e.g. when it comes to fields like pose estimation or human digitization. Supervised learning with synthetic ground truth human 3D data has been proven to be a highly effective approach to develop machine learning models, but it goes hand in hand with the issue that compared to 2D imagery it's difficult to get large datasets of annotated hyper-realistic and accurate 3D scanned human data. Since 2013 Renderpeople is one of the world leaders in the production, development, and distribution of scanned 3D People models as stock footage. In his talk, their CEO Lukas Lamprecht will share some insights about Renderpeople's challenges as well as approaches and ideas about the further creation of human datasets for CV research.

Lukas Lamprecht is the CEO at Renderpeople, Germany since 2019. Before that he was a 3D artist at Renderpeople from 2015 to 2019. He was a lecturer and supervisor for 3D & VFX at SAE Institute, Germany between 2013 and 2015. He obtained his Bachelor of Arts (Hons.) degree in Digital Film Making from Middlesex University, London, UK in 2015.

Volumetric video at the intersection of visual effects and virtual production

Juraj Tomori, Charles Dupont, George Ash, Mike Pelton
{juraj,charles,george,mike}@dimensionstudio.co
dimension,
London

1 Introduction

Dimension operates at the intersection of volumetric video production, visual effects, and virtual production. Evolution in the media and entertainment industries is fuelling a growing demand for the rapid creation of realistic content, both for visual effects and for virtual production. Creating digital humans has long been an especially challenging task - an active area of research since the inception of computer graphics. Dimension is meeting the need for the scalable production of realistic virtual humans by bringing its extensive experience with volumetric video capture and virtual avatar creation to bear. In this talk we describe our approach to increasing the quality and editability of volumetric video assets, overcoming some inherent limitations of the medium, and increasing its applicability in the worlds of media and entertainment.

2 Volumetric video

Volumetric video, or "free viewpoint video", is a scalable and realistic way of capturing and representing human performances, and typically acquired with multi-camera stages. Capture reconstruction results in a lightweight, and believable, animated asset that can be viewed from an arbitrary point of view within the navigation range. Volumetric video is a broad term that includes multiple approaches to achieve the same goal - having the ability to freely change the viewpoint from which the video can be played back [5]. These approaches can be broadly divided into three categories:

- Model-based approaches, for example virtualized reality [3]
- Image-based approaches, for example light-fields [4]
- Hybrid approaches, for example that of [6]

Each approach is making a viewpoint range, compression, visual fidelity, cost, acquisition and reconstruction difficulty tradeoffs [5]. Recent research includes [1]'s model-based approach which has been commercialized and provides state-of-the-art results and [2]'s model-based approach utilizing the light stage in addition to multi-camera stage, being able to infer photometric normals and reflectance properties. Dimension is utilizing [1]'s approach with our own advances described in the talk.

3 Avatars

The use of avatars is familiar ground, and the preferred approach for visual effects, virtual production, games, and the media and entertainment industry in general. Producing an avatar typically involves multiple steps and skills: designing the concept art, 3D sculpting, modelling, rigging, shading, animation, and adding character effects, which together provide very tight control over the final result, but are laborious and time consuming to achieve.

4 Limitations

While volumetric video capture typically provides more realism than a bottom-up avatar-based approach, editing the content is more challenging than that of an avatar-based approach. On the other hand, volumetric video production typically replaces tedious work for many artists with extensive computation and reduced human input, making it a more scalable approach. The time to produce a segment of volumetric video is typically an order of magnitude shorter than that for an avatar-based workflow. There are of course challenges with volumetric video reconstruction - amongst them:

1. A lack of physically based textures, which [1] doesn't infer, but are required for seamless integration into the virtual environment,
2. A lack of motion vectors, which are needed for the deformation motion blur that features in a typical visual effects workflow,
3. The need for animation edits that fit with established animation pipelines,
4. The need for extreme visual fidelity in close-up effects shots,
5. Reconstruction difficulties arising from fast-moving or thin, transparent, translucent objects, and
6. The need for artist-friendly control over the materials and textures.

5 Results

In this talk we describe the steps we are taking towards addressing these challenges, and the benefits for visual effects and virtual production projects. Our progress is enabling new applications for volumetric video, including the delivery of high quality assets for digital doubles in visual effects close-ups, and lighter assets for animatable CG characters in medium and long shots, and in virtual production scenarios.

[1] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766945. URL https://doi.org/10.1145/2766945.

[2] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356571. URL https://doi.org/10.1145/3355089.3356571.

[3] T. Kanade, P. Rander, and P.J. Narayanan. Virtualized reality: constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1): 34–47, 1997. doi: 10.1109/93.580394.

[4] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, page 31–42, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237199. URL https://doi.org/10.1145/237170.237199.

[5] Aljoscha Smolic. 3d video and free viewpoint video—from capture to display. *Pattern Recognition*, 44(9):1958–1968, 2011. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.09.005. URL https://www.sciencedirect.com/science/article/pii/S0031320310004450. Computer Analysis of Images and Patterns.

[6] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, page 600–608, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 9781450378239. doi: 10.1145/1186562.1015766. URL https://doi.org/10.1145/1186562.1015766.

INDUSTRY TALKS

Arnold 7 update

Frederic Servant (Autodesk)

Deliver beautiful results faster with amplified rendering performance, interactivity, and reliability! In the latest Arnold release, performance is significantly upgraded, scalability on GPU is greatly improved, and fullframe imagers are now updated during rendering. Intel's Open Image Denoise, a fast, AI accelerated, high-quality denoiser is now integrated. This release also adds production-ready USD enhancements for specific procedurals and shapes in Hydra, deep AOVs, light linking, and more.

Material acquisition and editing

Valentin Deschaintre (Adobe Research)

Materials are an essential part of virtual environments. Such environments are now ubiquitous across industries, from entertainment to architecture and medicine. Yet, the creation of virtual materials remains challenging and requires multiple hours for trained artists. In this talk I discuss our recent work on making material acquisition, creation and editing more accessible and what I believe to be next interesting steps in this direction. The work discussed in this talk will include: few images material capture, material re-sampling and procedural representations.

Valentin is a Research Scientist at Adobe Research in the London lab, working on virtual material creation and editing. He previously worked in the Realistic Graphics and Imaging group of Imperial College London hosted by Abhijeet Ghosh. He obtained his PhD from Inria, in the GraphDeco research group, under the supervision of Adrien Bousseau and George Drettakis. During his PhD, he spent 2 months under the supervision of Frédo Durand, at MIT CSAIL. His research covers material and shape (appearance) acquisition, creation, editing and representation, leveraging deep learning methods.

Contact-rich simulation in NVIDIA Omniverse

Kier Storey and Michelle Lu (NVIDIA)

Kier is a distinguished engineer and architect on the PhysX team with over 15 years of experience writing high-performance physics simulations. After completing his Ph.D. in computer science, Kier worked as a physics programmer on a number of AAA games before joining the NVIDIA PhysX team. During his time at NVIDIA, he has worked on a wide range of simulations including rigid bodies, robotics, soft bodies, clothing, fluid, and particles using both high-performance multi-core CPUs and GPUs that have been used in a wide range of gaming and simulation platforms.

Michelle Lu is Director of Simulation Technology in the PhysX team. She has been working on physics simulation for over 15 years, including rigid body dynamics, robotics simulation, deformable simulation, clothing, and fluid dynamics.

Foundry and Machine Learning

Ben Kent (Foundry)

Foundry's Nuke is often seen as ubiquitous in the VFX industry. Part of that involves staying up-to-date and providing artists with the latest tech. Nuke 13.0 was the first implementation of machine learning—a goal that Foundry's A.I. Research team was created for. Ben Kent, Research Engineering Manager and A.I. Team Lead, will walk through the founding pillars of A.I. at Foundry, from where it started to where it is now. The session will cover the extensive work from the Foundry Research team, how machine learning tool CopyCat was developed and implemented in Nuke 13, and what the future holds for Foundry and AI.

Ben is the Research Engineering Manager at Foundry as well as a screenwriter/director. Ben won an Academy Sci-Tech Award for his work on Foundry's Furnace tools and now leads the A.I. research team. As a filmmaker, his feature debut, comedy horror Killer Weekend was released around the world in 2019.

INDUSTRY TALKS

High-fidelity procedual data synthesis for validation and training of perception function

Oliver Grau
Oliver.Grau@intel.com
Korbinian Hagn
Korbinian.Hagn@intel.com
Intel Deutschland GmbH,
Neubiberg, Germany

Figure 1: Example scene with randomly selected and place objects.

Figure 2: Realistic sensor effect simulation, (left) standard Blender tone-mapped output, (right) the sensor simulation output.

This contribution summarizes aspects of application of a highly realistic procedural scene and sensor simulation for data synthesis for uses in validation and training of AI-based perception functions for automated driving. Our approach allows production of highly varied scenes by probabilistic combination of 3D assets from a database. Further, we developed a sensor model, that can be adjusted to real camera systems. We are able to show that this combination can decrease the domain gap between real and our synthetic data set for training of deep neural networks (DNNs) for object detection and semantic segmentation of urban traffic scenes.

Computer generated imagery (CGI) is increasingly popular for training and validation of DNNs as synthetic data can avoid privacy issues found with recordings of members of the public and can automatically produce ground truth data at higher quality and reliability than costly manually labeled data. Moreover, simulations allow synthesis of rare scene constellations helping validation of products targeting safety critical applications, specifically automated driving. We are now able to synthesize virtual scenes in a visual quality that is hard to distinguish from real photographs for human observers. On the other hand, we have emerging complex technical and particular autonomous systems sensing the real world and aiming at resembling some perceptional tasks formerly only feasible by humans. Because of the complexity of reality and the related increasing complexity of the technical tasks, validation, that makes sure these systems work as intended and are safe are increasingly important.

Our approach for validation of perception systems is to use a probabilistic grammar system [1, 4] to generate 3D scenarios which include a catalogue of different object classes, and places them relative to each other to cover the complexity of the input domain. In addition we vary scene parameters, like time-of-the-day and object positions[3].

Recently, specifically in the domain of driving scenarios, game engines have been adopted, like `CARLA` [2] based on a commercial game engine [1]. In contrast to these game engine-based simulators we use physical-based rendering techniques. These techniques provide a more realistic lighting computation than game engines with typically fixed rendering quality of 8bit per RGB color channel. For our experimental work, we use the physical-based open source `Blender Cycles` renderer[2] in high dynamic range (HDR) resolution, followed by our realistic sensor simulation.

Model trained on	mIoU ↑
Cityscapes (baseline)	81.56
Synthetic images w/o artifacts	40.37
Synthetic images w/ random artifacts	45.76
Synthetic images w/ Cityscapes optimized artifacts	**47.63**
GTAV	39.08
Synscapes	59.33

Table 1: Performance results of `DeeplabV3+` models trained on different datasets evaluated on the Cityscapes dataset.

The effect of sensor and lens effects on perception performance has not been studied a lot. We implemented a sensor model which expects images in linear RGB space and floating point resolution. We then simulate a camera error model by applying *sensor noise*, as added Gaussian noise and an automatic, histogram-based exposure control (linear tone-mapping), followed by non-linear *Gamma correction*. Further, we simulate the following lens artifacts *chromatic aberration*, and *blur*. Fig. 2 shows a comparison of the standard tone-mapped 8bit RGB output of Blender (left) with our sensor simulation adapted to match the camera characteristic of Cityscape images.

By application of sensor lens artifacts on a synthetic training dataset one can achieve higher performance on the real-world validation dataset and successfully reduce the domain gap between synthetic and real-world domain. In table 1 we trained `DeeplabV3+` models on synthetic datasets such as Synscapes, GTAV and our synthetic dataset, with and without realistically modeled sensor artifacts, and evaluated these models on the Cityscapes dataset.

While the Synscapes [4] trained model achieves highest cross domain performance, beating the GTAV and our synthetic images, we can still show that by application of a realistic sensor simulation we are able to increase the cross-domain performance by over 7% mIoU. Synscapes applies, similar to our optimized artifact simulation, a sensor simulation that should closely resemble the sensor artifacts of the Citycsapes dataset[3] but they additionally have modeled their dataset to look as much as the Citycsapes dataset in terms of objects and scenes as possible, explaining the high cross-domain performance.

Next steps in our research is to disentangle the effects on scene complexity, geo-localization and rendering fidelity on the domain gap.

Acknowledgement: The work presented in this paper was partially funded by the BMWi project KI-Absicherung.

[1] Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Learning to generate synthetic datasets. In *ECCV*, 2020.

[2] Alexey Dosovitskiy et al. CARLA: an open urban driving simulator. *CoRR*, abs/1711.03938, 2017.

[3] Qutub Syed Sha, Oliver Grau, and Korbinian Hagn. Dnn analysis through synthetic data variation. In *Computer Science in Cars Symposium*, CSCS '20, 2020.

[4] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018.

[1] Unreal Engine 4, https://www.unrealengine.com

[2] https://www.blender.org/

[3] https://www.cityscapes-dataset.com

Semantic-driven colorization

Man M. Ho (Hosei University), Lu Zhang (INSA Rennes), Alexander Raake (TU Ilmenau), Jinjia Zhou (Hosei University)

Recent colorization works implicitly predict the semantic information while learning to colorize black-and-white images. Consequently, the generated color is easier to be overflowed, and the semantic faults are invisible. According to human experience in colorization, our brains first detect and recognize the objects in the photo, then imagine their plausible colors based on many similar objects we have seen in real life, and finally colorize them, as described in Figure 1. In this study, we simulate that human-like action to let our network first learn to understand the photo, then colorize it. Thus, our work can provide plausible colors at a semantic level. Plus, the semantic information predicted from a well-trained model becomes understandable and able to be modified. Additionally, we also prove that Instance Normalization is also a missing ingredient for image colorization, then re-design the inference flow of U-Net to have two streams of data, providing an appropriate way of normalizing the features extracted from the black-and-white image. As a result, our network can provide plausible colors competitive to the typical colorization works for specific objects. Our interactive application is available at https://github.com/minhmanho/semantic-driven_colorization.

Automatic camera control and directing with an ultra-high-definition collaborative recording system

Bram Vanherle, Tim Vervoort, Nick Michiels, Philippe Bekaert (Hasselt University)

Capturing an event from multiple camera angles can give a viewer the most complete and interesting picture of that event. To be suitable for broadcasting, a human director needs to decide what to show at each point in time. This can become cumbersome with an increasing number of camera angles. The introduction of omnidirectional or wide-angle cameras has allowed for events to be captured more completely, making it even more difficult for the director to pick a good shot. In this paper, a system is presented that, given multiple ultra-high resolution video streams of an event, can generate a visually pleasing sequence of shots that manages to follow the relevant action of an event. Due to the algorithm being general purpose, it can be applied to most scenarios that feature humans. The proposed method allows for online processing when real-time broadcasting is required, as well as offline processing when the quality of the camera operation is the priority. Object detection is used to detect humans and other objects of interest in the input streams. Detected persons of interest, along with a set of rules based on cinematic conventions, are used to determine which video stream to show and what part of that stream is virtually framed. The user can provide a number of settings that determine how these rules are interpreted. The system is able to handle input from different wide-angle video streams by removing lens distortions. Using a user study it is shown, for a number of different scenarios, that the proposed automated director is able to capture an event with aesthetically pleasing video compositions and human-like shot switching behavior.

Depth estimation from a single omnidirectional image using domain adaptation

Yihong Wu, Yuwen Heng, Mahesan Niranjan, Hansung Kim (University of Southampton)

Omnidirectional cameras are becoming popular in various applications owing to their ability to capture the full surrounding scene in real-time. However, depth estimation for an omnidirectional scene is more difficult than normal perspective images due to its different system properties and distortions. It is hard to use normal depth estimation methods such as stereo matching or RGB-D sensing. A deep-learning-based single-shot depth estimation approach can be a good solution, but it requires a large labelled dataset for training. The 3D60 dataset, the largest omnidirectional dataset with depth labels, is not applicable for general scene depth estimation because it covers very limited scenes. In order to overcome this limitation, we propose a depth estimation architecture for a single omnidirectional image using domain adaptation. The proposed architecture gets labelled source domain and unlabelled target domain data together as its input and estimated depth information of the target domain using the Generative Adversarial Networks (GAN) based method. The proposed architecture shows >10% higher accuracy in depth estimation than traditional encoder-decoder models with a limited labelled dataset.

VPN: Video provenance network for robust content attribution

Alexander Black, Tu Bui (University of Surrey), Simon Jenni (Adobe Research), Viswanathan (Vishy) Swaminathan (Adobe), John Collomosse (Adobe Research)

We present VPN – a content attribution method for recovering provenance information from videos shared online. Platforms, and users, often transform video into different quality, codecs, sizes, shapes, etc. or slightly edit its content such as adding text or emoji, as they are redistributed online. We learn a robust search embedding for matching such video, invariant to these transformations, using full-length or truncated video queries. Once matched against a trusted database of video clips, associated information on the provenance of the clip is presented to the user. We use an inverted index to match temporal chunks of video using late-fusion to combine both visual and audio features. In both cases, features are extracted via a deep neural network trained using contrastive learning on a dataset of original and augmented video clips. We demonstrate high accuracy recall over a corpus of 100,000 videos.

FacialFilmroll: High-resolution multi-shot video editing

Bharath Bhushan Damodaran, Emmanuel Jolly (InterDigital R&D France), Gilles Puy (In his own name), Philippe-Henri Gosselin, Cédric Thébault, Junghyun Ahn (InterDigital), Tim Christensen, Paul Ghezzo (In his own name), Pierre Hellier (InterDigital/Technicolor)

We present FacialFilmroll, a solution for spatially and temporally consistent editing of faces in one or multiple shots. We build upon unwrap mosaic [Rav-Acha et al. 2008] by specializing it to faces. We leverage recent techniques to fit a 3D face model on monocular videos to (i) improve the quality of the mosaic for edition and (ii) permit the automatic transfer of edits from one shot to other shots of the same actor. We explain how FacialFilmroll is integrated in post-production facility. Finally, we present video editing results using FacialFilmroll on high resolution videos.

Speech-driven conversational agents using conditional flow-VAEs

Sarah Taylor, Jonathan Windle, David Greenwood (University of East Anglia), Iain Matthews (Carnegie Mellon University)

Automatic control of conversational agents has applications from animation, through human-computer interaction, to robotics. In interactive communication, an agent must move to express its own discourse, and also react naturally to incoming speech. In this paper we propose a Flow Variational Autoencoder (Flow-VAE) deep learning architecture for transforming conversational speech to body gesture, during both speaking and listening. The model uses a normalising flow to perform variational inference in an autoencoder framework and is a more expressive distribution than the Gaussian approximation of conventional variational autoencoders. Our model is non-deterministic, so can produce variations of plausible gestures for the same speech. Our evaluation demonstrates that our approach produces expressive body motion that is close to the ground truth using a fraction of the trainable parameters compared with previous state of the art.

Photometric stereo with area lights for Lambertian surfaces

Jiangbin Gan and Thorsten Thormählen (University of Marburg)

This paper presents a photometric stereo technique that uses area lights for normal recovery and 3D geometry reconstruction of mid-sized objects. The object is illuminated in succession by several off-the-shelf LED area lights and images are captured by at least two DSLR cameras. Compared to point light sources, area lights have the advantage of producing high illuminance, resulting in low image noise and fast shutter speed, which is important if the captured object is not completely static during the acquisition of the images, e.g., when capturing a human face. Area lights are standard photo equipment which makes them cheaper, easier to obtain, and install than specialized many-lights hardware. The normal map of the object is recovered by our photometric stereo approach that uses ray tracing techniques to simulate the light transport in the scene. Furthermore, our approach takes the effects of occlusion and interreflections into account. The normal map is iteratively optimized which in turn is utilized to update the depth information of the object. Our synthetic and real-world experiments show that area lights are applicable for photometric stereo at the cost of an increased computational effort.

Full papers available from the ACM Digital Library
https://dl.acm.org/doi/proceedings/10.1145/3485441

SHORT PAPERS

One-shot SVBRDF Estimation Including Anisotropic material

Nozomu Terada, Ikuko Shimizu

Department of Computer and Information Sciences,
Tokyo University of Agriculture and Technology

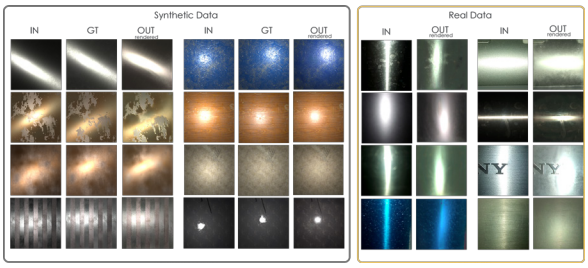


Figure 1: Rendering results using our network output for synthetic data and real data

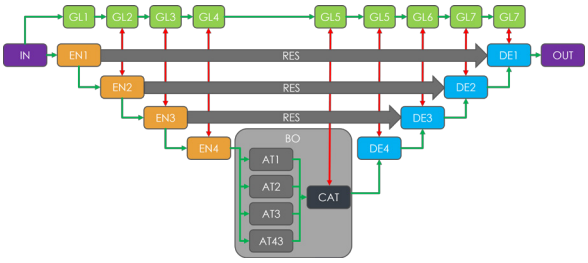


Figure 2: Network Structure

For rendering 3D object realistically in CG, Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF) is essential. Therefore, SVBRDF estimation methods have been studied for many years. In recent years, various methods for estimating SVBRDF employing deep learning have been proposed[2, 3, 6].

Methods based on deep learning require huge dataset. Though there are many SVBRDF datasets, they are limited to support only relatively simple isotropic materials. However, the optical phenomenon of anisotropy, where highlights are elongated as shown in Fig.1, is particularly common in metals. This is because industrial products are often processed in an artistic way to exhibit anisotropy. However, existing methods ignore the anisotropic property. To deal with the metal objects in CG, an easy and more unconstrained SVBRDF data estimation method is required.

This paper proposes a method for estimating SVBRDF including anisotropy from one image based on deep learning. Our contributions are as follows: (1) new network structure for estimating SVBRDF with four important characteristics, (2) the loss function for training our network by combination of three loss functions, and (3) a new SVBRDF dataset including anisotoripic materials for training.

The network structure of our method is shown in the Fig. 2. Basically, there are 4-layers encoder / decoder and skip connections. Our four important characteristics are ResPath for skip connection, Global Feature subnetwork, Bottom Block, and Partial Convolution.

For the skip connection, we use ResPath [5]. This is because ResPath is known to have effect of suppressing the gap between the encoder and the decoder.

The node “GL” in Fig. 2 is a Global Feature subnetwork proposed by Deschaintre et al. [2], which showed good performance in isotropic SVBRDF estimation. This global feature network is known to have the effect of improving the ability to integrate information from distant locations in an image, which improves the ability to assign similar values to similar materials in the image. By introducing the global feature network, the properties of uniform regions become more uniform.

The structure of the bottom block of our network, marked with “BO” in Fig. 2, is inspired by the method called Atrous Spatial Pyramid Pooling (ASPP) proposed in [1]. Because a wider field of view by downsampling makes the low-level features too abstract compared to the original input image, we use the Atrous convolution to get a wider field of view instead of downsampling.

In the SVBRDF estimation task, artifacts are known to occur when the input image have the strong and distinct highlights such as saturation. In the image of anisotropic materials, the strong and distinct highlights are often observed, and moreover, saturation tends to occur over a wider range than the image of standard materials. To overcome this problem, we utilize the convolution in the network with Partial Convolution [7]. In the Partial Convolution, the saturated regions are not convoluted by using a mask for the saturated regions of the input image.

Next, we explain about the loss function for training our network. We introduce a combination of three loss functions. The first one is the rendering loss function, which compares rendered image by estimated SVBRDF data to rendered image by the target SVBRDF data, which is common in SVBRDF estimation tasks. The second one is a loss function that compares the rendered image by estimated SVBRDF data whose anisotropy data is replaced by the target anisotropy data to rendered image by target SVBRDF data. Because it is difficult to estimate the anisotropic direction, this loss function facilitates the estimation. The third one uses the Perceptual loss proposed in [4]. This loss function makes it easier for the network to understand the global structure.

In addition, we estimate the height map instead of the normal map. This is because the normal map can be calculated using the height map and the height map estimation is easier than the normal map estimation.

Before training the network, we first created a dataset for training because there is no other SVBRDF datasets including anisotoripic materials. Our dataset is composed of data based on the Disney Principled SVBRDF model, which has intuitive parameters and is compatible with various software. For isotropic materials, data in Substance Share [1] were edited and used in our dataset. For anisotropic materials, we newly created data.

The results obtained by our method for the synthetic data and real data are shown in Fig. 1. For the synthetic date, although some of the results are unstable, we confirmed that we can estimate the SVBRDF data even when we input complex anisotropic data such as dirt, noise, and separations. Note that the highlights of the input image were removed fairly cleanly and the results are very similar to the ground truth data. For the real data, it is not possible to estimate properly under all conditions. However, we confirmed that satisfactory data was obtained from various inputs.

[1] L. C. Chen, G. Papandreou., I. Kokkinos., K. Murphy., A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRF. IEEE Trans. on Pattern Analysis and Machine Intelligence PP(99). June 2016.

[2] V. Deschaintre.., M. Aittala.., F. Durand.., G. Drettakis., and A. Bousseau. Single-Image SVBRDF Capture with a Rendering-Aware Deep Network. ACM Trans. Graph. Vol. 37, No. 4, Article 128, August 2018.

[3] D. Gao., X. Li., Y. Dong., P. Peers., K. Xu., and X. Tong. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. ACM Trans. on Graph. Vol. 38, Issue 4, Article134. July 2019.

[4] L. A. Gatys., A. S. Ecker., M. Bethge. A neural algorithm of artistic style. cs.CV, Sep 2015.

[5] N. Ibtehaz. and M. S. Rahman. MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. cs.CV, Feb 2019.

[6] Z. Li., K. Sunkavalli.., and M. Chandraker. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. cs.CV, Apr 2018.

[7] G. Liu., F. A. Reda., K. J. Shih., T. C. Wang., A. Tao., B. Catanzaro. Image Inpainting for Irregular Holes Using Partial Convolutions. ECCV 2018, pp 89–105.

[1] https://share.substance3d.com/

AURealnessGAN - An Architecture that Enables Manipulation of FACS Action Units in Face Image Generation

Koyo Ishihara
s200918y@st.go.tuat.ac.jp
Tokyo University of Agriculture and Technology

Ikuko Shimizu
ikuko@cc.tuat.ac.jp
Tokyo University of Agriculture and Technology

Akio Sashima
sashima-akio@aist.go.jp
National Institute of Advanced Industrial Science and Technology

Koichi Kurumatani
k.kurumatani@asule.org
National Institute of Advanced Industrial Science and Technology

We propose a novel GAN architecture called AURealnessGAN that can generates facial images corresponding to desired facial expressions by specifying the intensity in the form of basic facial actions (Action Units: AU) defined in the Facial Action Coding System (FACS)[1]. From the psychological point of view, FACS AU can be used to formally represent emotions and to manipulate the facial images corresponding to desired emotions. Our basic idea is to construct a GAN architecture that generates facial images corresponding to a specified emotion represented by FACS AU values, by introducing the values to a generator part of GAN, i.e., multiple layers of deconvolutional networks. In FACS, 44 basic AUs are defined such as Cheek Raiser (AU6), Lip Corner Puller (AU12), and so on.

The structure of AURealnessGAN is shown in Fig.1. It allows the input of AU intensity in addition to latent variables. This network separates the layer for learning individual features from the one of AU. In addition, the structure allows manipulation of AU intensity from a low-resolution image state by mixing the layer for learning features related to AU and one of individual features during deconvolution process.

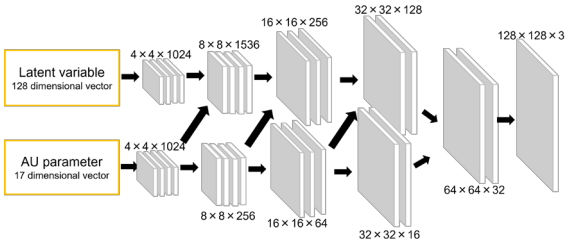


Figure 1: Generator of AURealnessGAN

We use RealnessGAN[4] as the base model. The Generator of AURealnessGAN is shown in Fig.1. The discriminator is a modified version of RealnessGAN that allows input of AU intensities in addition to images. For the learning, facial expressions with metadata for each AU intensity is required. Because it is difficult to create data by annotating AU intensity on face images with qualification, we utilize OpenFace[2], a network that can measure 17 AU intensities, to create a new dataset in which each AU intensity is annotated as metadata on the image. Some images in the dataset are hidden by hair or hands. It is impossible to measure the AU of that hidden part. The images with AU value that cannot be measured by OpenFace are annotated with -1 to indicate that it cannot be measured. If AU can be measured, AU intensity is normalized from 0 to 1 and annotated to the image to create a dataset.

We used CelebA[3], which is a public dataset for the image dataset. Some of the CelebA images do not measure all of the 17 AUs. After removing such images, we obtained 193,718 images of CelebA in which one of the AUs can be measured by OpenFace.

The generated images by AURealnessGAN are shown in Fig.2. We measured the AUs from these original images and input the AUs to the AURealnessGAN to generate face images with the same expressions as the original images. Compared to the case of neutral, the generated images has a change in facial expression of the original image while retaining the identity other than facial expression. In the generated images of sadness, black noise appears on the forehead in all images. This is probably due to the bias in the number of expressions in the dataset.

The dataset CelebA is designed for training of facial attribute recognition, face recognition, and face detection. Each expression is not equally included in the dataset. The intensity of AU6 (Cheek Raiser), which tends to be stronger when smiling, was measured from 102,446 images out of 193,718 datasets. The intensity was almost evenly distributed from weak to strong. On the other hand, AU4 (Brow Lowerer), whose intensity tends to be strong when the subject is sad, could be measured from only 58,652 images. In addition, most of the images were weak in intensity. This indicates that a lack of data on the face images with sad expressions may have resulted in not enough learning and noise in the generated images.

We proposed an AURealnessGAN that can manipulate the Action Unit. To solve the noise problem, different data set and/or improvement of base model is necessary.

Figure 2: Example of generated images(Left: Original image (from top to bottom: neutral, happiness, surprised, angry, sadness), The 2nd to 4th from left: generated images based on AU of original image)

[1] Paul Ekman, Wallace V. Friesen, O'Sullivan Maureen, Chan Anthony, Diacoyanni-Tarlatzis Irene, Heider Karl, Krause Rainer, LeCompte William Ayhan, Pitcairn Tom, Ricci-Bitti Pio E, Scherer Klaus, Tomita Masatoshi, Tzavaras Athanase. "Universals and cultural differences in the judgments of facial expressions of emotion". Journal of Personality and Social Psychology, 1987.

[2] OpenFace 2.0: Facial Behavior Analysis Toolkit, Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, IEEE International Conference on Automatic Face and Gesture Recognition, 2018.

[3] Deep Learning Face Attributes in the Wild, Liu, Ziwei and Luo, Ping and Wang, Xiaogang and Tang, Xiaoou, Proceedings of International Conference on Computer Vision (ICCV), 2015.

[4] Real or Not Real, that is the Question, Xiangli, Yuanbo and Deng, Yubin and Dai, Bo and Loy, Chen Change and Lin, Dahua, International Conference on Learning Representations, 2020.

[5] Development and validation of a facial expression database based on the dimensional and categorical model of emotions, Fujimura Tomomi. and Umemura Hiroyuki, Cognition and Emotion, 2018.

Learning semantic object segmentation for video post-production

Flavien Jourdren, Emmanuel Jolly, Claire-Hélène Demarty, InterDigital R&I, Rennes, France
Frédéric Lefebvre, Pierre Hellier
firstname.lastname@InterDigital.com

1 Introduction

Video postproduction pipeline will increasingly benefit from artificial intelligence tools. For instance, the automatic extraction of specific objects helps the postproduction workflow. In particular, booms mics removal could be accelerated, and color chart detection could end up in a more efficient color pipeline. For now, the segmentation of these objects is usually done via rotoscoping and consequently necessitates huge manual work. Semantic segmentation has made huge progress since the use of convolutional networks. Existing and publicly available frameworks such as Detectron2 [1] and PointRend [1] already allow to perform high quality detection and segmentation of 80 different generic classes. However, the performance of these frameworks is very much bound to the quantity and quality of training data. Unfortunately, fetching relevant video footage and manually extracting the objects (e.g., boom mics and color charts) is out of reach. To alleviate this problem, we propose in this paper a lightweight training strategy: training data is generated synthetically by inserting in an existing dataset the desired objects, along with data augmentation. A pretrained network is used and fine-tuned using this new dataset. Despite its simplicity, we show in this paper that the system can achieve good performances for an automatic video postproduction pipeline.

2 Method

Generation of specific training data PointRend was trained on the COCO dataset (`https://cocodataset.org/#home`), which contains 80 generic object categories. We propose to extend the COCO dataset with two new post production classes: boom mics and color charts. A very small set of manually segmented images of boom mics (24 instances) and color charts (450 instances) were used, thus requiring an acceptable level of manual work. We first create synthetic insertions of these objects in already annotated images from the COCO dataset. To reduce color and illumination discrepancy between inserted objects and the background images, the insertion is performed with the OpenCV seamless blending technique (`https://opencv.org/`). In case of overlapping between COCO annotated objects and the new classes, the new objects are placed in the foreground while the corresponding masks of the occluded COCO objects are modified accordingly. From 24 boom mics samples, 60% (14) were used for training, and 20% were used for validation and test respectively. In total, 1500 (respectively 400) images were generated for training (resp. validation). Each dataset was equally split into positive (containing the object) and negative (without the object) samples. For the first half of our data, some post production objects are randomly picked from example sets (with some horizontal or vertical flip), then inserted at a random location in a background image containing a person to guarantee a semantically consistent context. For the second half, we just pick random background images that we do not modify to have some negative examples (images without synthetic insertions). In a second step, data augmentation was used as implemented in Detectron2: ResizeShortestEdge, horizontal RandomFlip (prob: 20%), vertical RandomFlip (prob: 20%), RandomBrightness (prob: 20%, range: 0.75-1.25), RandomContrast (prob: 20%, range: 0.75-1.25), RandomSaturation (prob: 20%, range: 0.75-1.25) and RandomRotation (prob: 20%, range: -30-+30). Data augmentation increases the diversity of the final dataset and helps the generalisation of the model.

Network architecture and training We use transfer learning to learn our post production related classes by retraining only the model's prediction components (region proposal network and ROI heads). Transfer learning exploits the low-level extraction of features already computed on a large class of natural objects. Hence, the detection of new objects can be done with a limited number of training data. Transfer learning enables to both benefit from the good generalization of the already trained lower layers and to reduce the processing time by learning only the last parameters of the network. While common object segmentation methods generally predict labels on low-resolution crops (e.g., 28x28), PointRend extracts objects from 224x224 images, which is highly desirable for professional contents with resolutions 4K or higher. Hence, PointRend was selected as our base framework. The network backbone remains frozen, while the predicting parts are initialized with the original weights from PointRend Detectron2 and further retrained to take into account our new classes, i.e., boom mics and color charts. Both Resnet50 and Resnet101 were tested as network backbone. We trained for 3500 iterations with a batch size of 5. Training lasted one hour in average for the two added classes using one tesla 100 GPU, thus demonstrating that the process can be easily repeated for additional classes.

[1] `https://github.com/facebookresearch/detectron2`

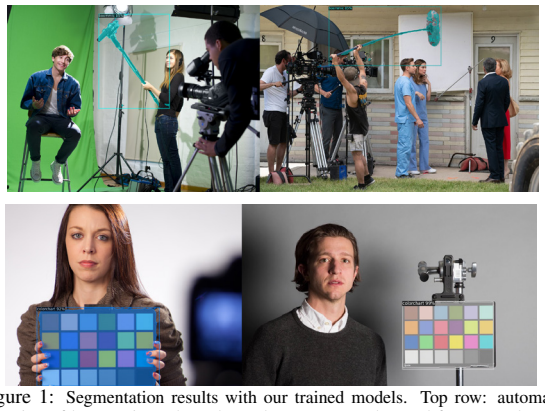


Figure 1: Segmentation results with our trained models. Top row: automatic detection of boom mics, where the garbage mate can be used for removal. Bottom row: automatic detection of color charts. An automatic color neutralization/homogenization/transfer tool could benefit from the detected chart.

3 Results

We present in figure 1 visual results of detection obtained on real images not used during training. Although the detection contours do not exactly match the geometric contours of the objects, we believe this paves the way for an automatic usage in production. We also computed the average precision (AP) on the test set. For the ResNet50 (resp. ResNet101) backbone, the average precision was 12.95 (resp. 16.35) for boom mics and 94.43 (resp. 93.99) for color charts. The AP values for boom mics are due to the inaccurate nature of the objects' contour. In addition, our network tends to fail detecting the whole handle of boom mics, leading to unwanted holes and certainly contributing to the low AP values. However, if used for removal with a dilatation of the obtained mask, we believe this average precision is acceptable for post-production routine. The precision obtained for color charts is better, which can be explained by the lower visual variability of the object. In conclusion, we have presented a framework to automatically extract specific post-production classes such as boom mics and color charts, while resorting to very low manual resources. Leveraging the transfer learning concept, we show that the synthetic generation of data leads to a production usable detection. Further work will focus on improving the plausibility of the data generation process by adding additional constraints (for example, constraining the boom mic position according to other objects in the scene).

4 References

[1] A. Kirillov *et al.* PointRend: Image segmentation as rendering. CVPR, 2020.

A Deep Learning Based Approach for Camera Switching in Amateur Ice Hockey Game Broadcasting

Hamid Reza Tohidypour, Yixiao Wang, Mohsen Gholami, Megha Kalia, Kexin Wen, Lawrence Li, Panos Nasiopoulos
http://www.dml.ubc.ca
Department of Electrical and Computer Engineering, University of British Columbia, Canada

Mahsa T. Pourazad
https://www.telus.com
TELUS Communications Inc., Canada, University of British Columbia, Canada

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

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

To this end, in this paper, we propose a solution that is based on deep learning, namely the Faster-RCNN architecture [6], to optimize view switching in regional ice hockey games. The main reason for this choice is that Faster-RCNN is proven to be more accurate and much faster compared to its predecessors [3, 4, 5], making it an ideal approach for real-time object detection of the ice hockey fields [6]. Our deep learning-based object recognition network receives video feeds from the two primary camera views of ice hockey that include the side view that shows the arena (please see left image in figure 1) and gives a wide view of the field, and the goalie views (please see right image in figure 1) that show a closer view of the nets. Then, it detects the players, net, and the puck in real time with very good precision and based on the predicted confidence values for the different objects, our algorithm decides which camera view should be broadcasted. As a result, the proposed method is play-centered unlike the only other existing work that is player centered [7].

In order to train our Faster-RCNN model, we generate a comprehensive dataset for our application by downloading several hockey videos of the resolution of 1920 x 1080 from YouTube [2]. From those videos, 1000 representative frames were selected for the training-validation phase, skipping redundant frames and considering only frames with significantly different content to avoid overfitting, preferably including the puck and of high visual quality – avoided blurry, fast moving puck frames. The selected frames were labeled according to our objects of interest (the players, net, and the puck), while the referees and audience were excluded. The training-validation frames were utilized to train our Faster-RCNN using a state-of-the-art advanced research computing network [1] and achieved the mean average precision of 78.9% for the validation images.

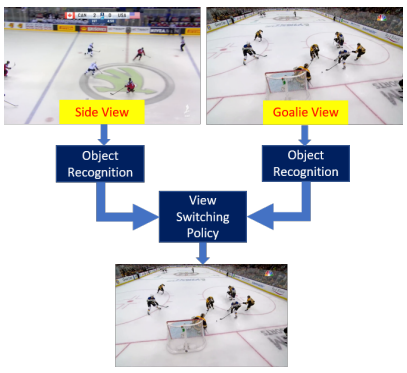


Figure 1: Our proposed scheme for automatic camera view switching.

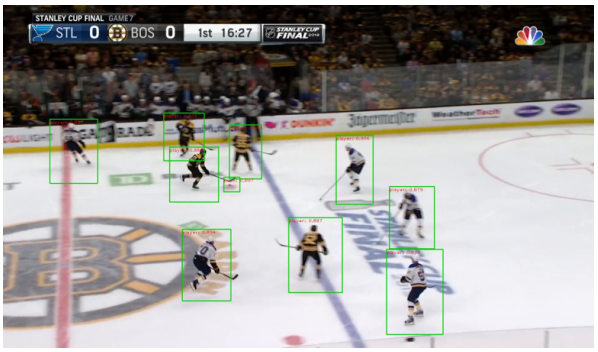


Figure 2: Example frame from the test set that shows players and the puck, which were detected correclty by our model.

In order to evaluate the performance of our trained model, we examined it on the test videos with unseen frames. We used our deep learning model to detect the objects of interest. Figure 2 shows the predicted objects and the probability values assigned to the bounding boxes for anl example test image. Our camera switching algorithm considers the position and confidence level of detection of all the objects, as each one has different roles to play in determining the best camera view for the current moment of the game. It is important to note that designing our algorithm to be biased towards the importance of objects to the fans, will allow our solution to be focused on the action. Driven by professional game coverage, we assume that the most important object/event in hockey broadcasting involves the puck. Following the above observation and the outcome of many trials asking subjects to validate the validity of our switching scheme, we assigned a weight to the confidence values predicted for each object type according to its importance. More precisely, the confidence of each detected object in the current camera view is weighted according to its object type and the weighted values are summed up to calculate the score for the current camera view. Our results show an accuracy of 75% for our camera switching method in real-time. Considering the fact that only 1000 frames were used for the training and validation phases, our camera switching approach achieved a great performance.

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

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

[3] R. Girshick. Fast r-cnn. In *Proc. IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE conference on computer vision and pattern recognition (ICCV)*, pages 580–587, 2014.

[5] K. He, X.Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn. towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

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

A Step Towards Automating the Synthesis of a Scene Script

Américo Pereira[12]
americo.j.pereira@inesctec.pt
Ricardo Carvalho[2]

Pedro Carvalho[13]

Luís Côrte-Real[12]

[1] INESC TEC,
Porto, Portugal

[2] Faculty of Engineering,
University of Porto, Portugal

[3] School of Engineering of the Polytechnic Institute of Porto,
Portugal

Generating 3D content is a task mostly done by hand. It requires specific knowledge not only on how to use the tools for the task, but also on the fundamentals of a 3D environment. Recent works such as [3] and [2] explore the idea of using natural language as input to a network that generates 3D shapes. However, their usage is not intuitive for new content developers and the content generated is directly related to the data used for training. In this work, we show that automatic generation of content can be achieved, from a scene script, by leveraging existing tools, so that non experts can easily engage in 3D content generation without requiring vast amounts of time in exploring and learning how to use specific tools.

In this article we explore the possibilities of easing the process of generating 3D content for non experts. To this end, we propose an architecture designed to provide a flexible automatic generation of 3D content based on a textual summary of a scene. Our main goal is to explore existing technologies and leverage them, showing that the task of manually creating and refining 3D content can be relieved by automating part of the creation process. The current proposal has two main targets: (1) entry level content creators, so that the creation of 3D virtual worlds that can be modified and used for further refinement are readily available without spending vast amounts of time; (2) augment the flexibility and automation of scene synthesis, particularly in scenarios where detail is not a main requirement. In our proposal we use the concept of a scene script, which is a file where the objects, lights, materials and so on are detailed. This way, by simply changing the file, variations of the scene can be generated. A high level description of the proposed framework is depicts in Figure 1.

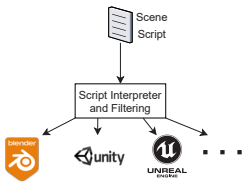


Figure 1: High Level Description of the Visualization Framework.

By providing an input scene script, the synthesizer starts by processing and extracting all the relevant information from the file and structures according to the target 3D generation platform. It then maps the described data with a direct connection with the chosen rendering API, forwarding the 3D content generation. By taking this approach, the content generated is guaranteed to be editable by the 3D generation platform, enabling more advanced users to start with a fast creation of 3D content that can be edited to add more complex details. Given that the goal of this work is to provide means to facilitate the creation of content by non experts and lessen the initial burden of creating 3D worlds for more advanced users, we decided to demonstrate our proposal by focusing on engines or rendering software that: provide means to automatically generate content; enable manual edition of the outcome.

To demonstrate that our proposal can indeed be used for automatic creation of 3D content based on a script of a scene, we prepared a proof of concept that takes a bottom-up approach. For this we have chosen Blender [1] as it allows to define, create and manipulate arbitrary geometry, materials and animations automatically, while also providing a detailed API for interconnections. In this proof of concept we show that Blender API calls can be used to directly map processed descriptions of a scene into actual 3D content. Due to the way the Blender API is structured, we separate the 3D environment creation and manipulation into different parts, smoothing the transition between the script and the scene. This integration with the Blender API is illustrated in Figure 2, where the different components of the description are forwarded to the corresponding 3D definitions.

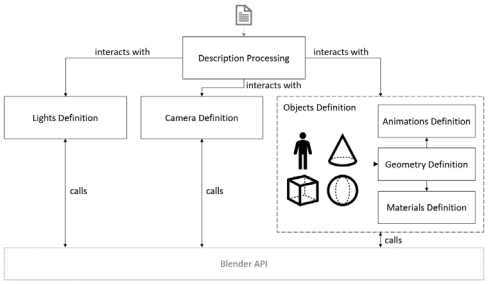


Figure 2: Virtual Scene Synthesis Process using Blender as Rendering Engine.

To better assess the ability of our proposal to automatically generate believable 3D environments from a script of a scene we asked 9 volunteers, with varying levels of mastery in 3D design and in different tools, to compare their work with our proposal. For this evaluation we defined a more complex scenario in a scene script that was given to the framework and to the volunteers. The volunteers were given 1h to complete the scene using their preferred tools. As we expected in this proof of concept, the visual aspect of the scene generated by hand shows more details. However, the volunteers commented that for an automatically generated scene, it produced a good starting point for further visual improvements.

Overall the volunteers commended the proposal mostly by the substantial reduction in time that it could accomplish, as the automatic generation process took only 10s to generate the scene given the script using an Intel i5 2600 CPU. Furthermore, the automatically generated results were considered to have a good enough quality as to allow users to understand the virtual scene. When looking into more detail, a lack of precise details such as facial features or more improved clothing, was identified. As the intend of this work was to show the possibility of automatically create content by using existing tools, it was within our expectations that the proof of concept did not convey precise details.

As future work we intent to explore topics such as human parametric models, cloth and texture generation; and explore with more detail the possibility of expressing data extracted from a scene in a structured and hierarchically coherent way, so that automatic generation of content can be made even more accessible and detailed.

This work was funded by Fundação para a Ciência e Tecnologia (FCT) with PhD Grant SFRH/BD/146400/2019.

[1] Blender Foundation. Blender, Mar 2021. URL `https://www.blender.org/`. Last accessed 20 January 2021.

[2] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*, pages 100–116. Springer, 2018.

[3] Ang Li, Jin Sun, Joe Yue-Hei Ng, Ruichi Yu, Vlad I Morariu, and Larry S Davis. Generating holistic 3d scene abstractions for text-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–201, 2017.

Look-Up-Table Mystified

Jurgen Stauder
Patrick Morvan
Angelo Mazzante
Anita Orhand
InterDigital R&D France
www.interdigital.com
John Frith
Moving Picture Company
www.moving-picture.com

Look-up-Tables (LUT) are used in a huge variety of applications. However, a LUT has inherently a limited size and applying the LUT to a signal instead of applying the underlying mathematical function always involves LUT application errors. Often, a LUT is accompanied by a pre-LUT allowing for non-regular sampling and error shaping. Reducing and shaping LUT application errors is essential for keeping errors under acceptable levels. In this paper, we present two new approaches that help into this direction.

The first new approach addresses a problem inherent in the use of a pre-LUT. In fact, a nonlinear pre-function applied as a pre-LUT allowing to focus LUT precision to a part of the range of the input signal enhancing hereby the precision for this part of the range. In order to not change the overall functionality, the inverse of the non-linear pre-function needs to be integrated into the calculation of the LUT entries. This necessary step causes a degradation of precision in a micro-scale in between two LUT entries.

The first new approach is a specific linearization of the pre-LUT in order to reduce overall LUT interpolation error. First, the pre-LUT is applied to the input signal. Then, the LUT is applied defined on regular grid. This regular grid is linked by the pre-LUT to a non-regular grid in the input signal domain. The approach is to define the pre-LUT linearly on this non-regular grid. Additionally, the inverse of the linearized pre-LUT is used for the calculation of the LUT entries.

We applied this approach to gamut mapping in transmission of high dynamic range (HDR) video using SL-HDR scheme (HDR single layer) [1] that is SDR backwards compatible. Figure 1 illustrates a sample result showing that the proposed approach considerably enhances the LUT precision.

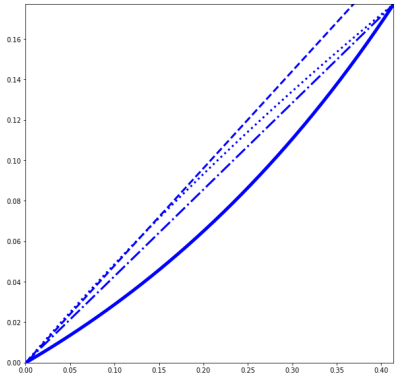


Figure 1: Application of underlying mathematical function (cont. line) compared to application of LUT with either linearized pre-LUT (dashed dotted line), classical pre-LUT (dotted line) or without (dashed line)

In visual tests using critical images such as shown in Figure 2 we found that ΔE CIE 1976 is reduced from 0.68 to 0.32 in overall range and from 0.196 to 0.076 in darks.

Figure 2 : Critical HDR frame involving smooth color slopes in the sky after transmission using the partially linearized pre-LUT (left of cross) and using classical pre-LUT (right)

The second new approach addresses a problem of preserving negative coordinates from camera noise in cinematographic post-production. When ingesting camera images into ACEScg color space using a LUT for gamut mapping, these noise values are required to be not affected by LUT interpolation errors.

The second new approach is the creation of a transparent pathway by a specific linear section in the pre-LUT corresponding to a specific linear section in the LUT. These sections are linked to the underlying mathematical function representing the gamut mapping. In fact, the pre-LUT is perfectly linear within the range of the noise and maps the noise into an intermediate range of the pre-mapped signal. The main LUT then maps the pre-mapped signal being perfectly linear within this intermediate range.

Figure 3 (top) shows a sample result of preserving noise using the proposed transparent pathway.

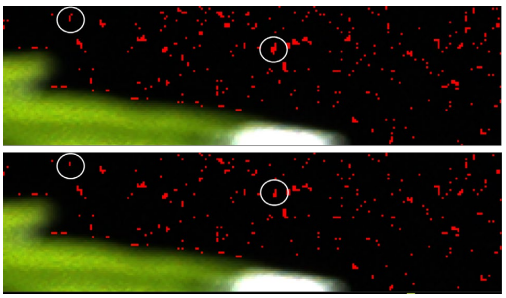


Figure 3: A camera image with marked negative noise (red pixels) before (top) and after ingesting and exporting from/to ACEScg using a 65x65x65 LUT (bottom). Only few noise (white circle) is lost since affected by the transition between pathway and gamut mapping function

[1] High-Performance Single Layer High Dynamic Range (HDR), System for use in Consumer Electronics devices, Part 1: Directly Standard Dynamic Range (SDR), Compatible HDR System (SL-HDR1), ETSI TS 103 433-1 V1.2.1 (2017-08)

Spatio-temporal algorithm for 3D sequences noise reduction

Ljubomir Jovanov
http://telin.ugent.be/~ljj
Danilo Babin
http://telin.ugent.be/~dbabin
Wilfried Philips
http://telin.ugent.be/~philips

imec-IPI,
Ghent University

In the past decade we are witnessing a rapid advance of cameras capable to simultaneously capture 3D information and color textures. Despite this progress the quality of the captured image/3D sequence is not sufficient to perform a complete textured 3D reconstruction of objects and large scenes at desired level of quality. One of the main reasons for this are different types of noise inherent to current depth sensors, their low spatial resolution and often poor image quality of color cameras used. In this paper we aim at simultaneous reduction of noise present in 3D measurements and color texture.

The proposed method operates on multiple frames in order to exploit spatio-temporal redundancy in both modalities. Moreover, we propose the use of color information to search for correspondences in previous frames used for temporal noise reduction. The first step in the proposed method is a dense motion estimation using the DeepFlow algorithm [3], where the algorithm extracts the most similar regions to the current one, from a buffer containing previous frames. Noise estimates are utilized to calculate reliabilities of the motion estimates and interpolation/filtering weights inside the temporal filtering module. Residual noise and interpolation artefacts are filtered inside the spatial filtering module.

Motion estimation has a crucial importance for temporal denoising. Accurate motion estimation enables filtering noisy pixels using their correspondences from previous frames. In the case of inaccurate motion estimation, temporal denoising produces motion blur artefacts around the edges of moving objects. To avoid this, motion estimation algorithm should be able to track large motions, since the artefacts most often occur in such situations. Another important factor is providing a dense motion estimates instead of block-wise motion vectors, since such motion field enables more precision and avoids block artefacts.

This algorithm relies on SIFT descriptors to find the best matching regions in previous frames. In order to increase robustness and handle large motions, the algorithm operates on multiresolution pyramids. Unlike most optical flow algorithms, DeepFlow operates from a fine level, and proceeds towards coarser levels, built by aggregating responses of smaller patches. The second component of the DeepFlow algorithm is variational optical flow, which relies on the deep matching framework. Additionally, the data term from deep matching is normalized to reduce the influence of regions with high values of spatial derivatives. Weights at different scales are different in order to reduce the influence of matching terms at finer scales. Finally optical flow optimizes the following cost function:

$$E(w) = \int_{\Omega} (E_D + \alpha E_S + \beta E_M) dx \tag{1}$$

where E_D is the data term, E_S smoothness term and E_M a matching term.

In order to perform joint motion estimation of depth and color we have replaced the blue channel with normalized depth values. An example of optical flow calculated on an "Orbit" sequence is shown in Figure 1b. Based on the estimated motion vectors we first perform motion compensated temporal filtering of color and depth. The proposed temporal filtering is performed on all noisy pixels $s^C(k,t)$ as follows:

$$\hat{s}_T^C(k) = \sum_{t=T-\frac{w}{2}}^{T+\frac{w}{2}} \sum_{h \in H} \alpha(t,h) s_t^C(h), \tag{2}$$

where $\hat{s}_T^C(k)$ is the temporally filtered version of depth and color at the location k of the current frame. Furthermore $s_t^C(h)$ contains values from the frame F_t at the location h. The amount of filtering is controlled through the weighting factors $\alpha(t,h)$ which depend on reliability of the motion estimation.

After temporal filtering, a certain amount of noise remains in the depth sequence. To remove it we rely on the method from [2]. This method starts from the assumption that image priors must not necessary be learned from data, and that a large portion of image statistics can be deduced from the structure of ConvNets generator. This algorithm relies on untrained ConvNets and fits a generator network to a single degraded image. Random initialized network weights are then fitted to a degraded image, conform to a task dependent observation model. This way the only information needed to perform restoration task is a degraded image and the structure of the network for the reconstruction. The network used in this method follows encoder-decoder architecture, with a small number of hyper parameters. LeakyReLU is used as a non-linear function.

We evaluate the proposed method using objective quality measures, by testing the performance using a well known “Interview” and “Orbit” sequence as a groundtruth. This sequence, acquired using a camera presented in [1], is often using in literature for benchmarking of various depth restoration, view interpolation and depth compression methods.

In the first experiment we have added artificial signal dependent noise, in accordance with the sensor characteristics as shown in Figure 1c. For an “Interview” sequence the average PSNR of was 20.47dB before the restoration. After denoising using the proposed method, PSNR averaged over the whole sequence was 30.41dB which is significantly improved compared to the noisy sequence. In the case of “Orbit” sequence, the PSNR was 20.1dB and 31.41dB after the restoration.

Figure 1: (a) Texture image associated to the scene (b) Estimated motion vectors (c) Noisy depth map (d) Denoised depth map

[1] G. J. Iddan and G. Yahav. G.: 3d imaging in the studio (and elsewhere. *Proceedings of SPIE*, 4298:48–55, 2001.

[2] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 128 (7):1867–1888, JUL 2020. doi: {10.1007/s11263-020-01303-4}.

[3] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *2013 IEEE International Conference on Computer Vision*, pages 1385–1392, 2013. doi: 10.1109/ICCV.2013.175.

Image Super-Resolution via Hierarchical Attention-Based Multi-References Sampling

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

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

1 Introduction

To overcome the limitations of single image super-resolution (SR) approaches that produce blurry super-resolved images, recent research has introduced the sub-problem of reference image super-resolution (RefSR). Given a low resolution (LR) input image and a similar high resolution (HR) reference image, RefSR approaches estimate a SR image. Reference super-resolution with a single reference image has been demonstrated to improve performances over general SR methods achieving large up-scaling with reduced visual artefacts. We generalise reference super-resolution to use multiple reference images giving a pool of image features and propose a novel attention-based sampling approach to learn the perceptual similarity between reference features and the LR input. As shown in Figure 2, given N_M reference images, our approach produces a $4\times$ SR image which is perceptually plausible and has a similar level of detail to the ground-truth HR image. An extended version of this short paper will appear in ICCV [2].

2 Method

The problem of multiple-reference super-resolution can be stated as follows: given a LR input I_{LR} and a set of HR reference images $\{I^m_{ref}\}_{m=1}^{N_M}$, estimate a spatially coherent SR output I_{SR} with the structure of I_{LR} and the appearance detail resolution of the multiple-reference images. Figure 1 presents an overview of the proposed approach to achieve multiple-reference super-resolution, which comprises the following stages.

Feature Extraction: to reduce GPU memory consumption with multiple reference images, the LR input I_{LR} and HR reference images $\{I^m_{ref}\}_{m=1}^{N_M}$ are divided into N_I and N_R sub-parts, respectively. Image features are extracted from these parts using a pre-trained VGG-19 network. The input vector is further divided into subvectors to focus the learning attention on input features while computing similarity maps with reference features.

Hierarchical Attention-based Similarity: the objective of this stage is to map the features of the LR input to the most similar features of the HR reference images. The output is a feature vector that contains the values of these most similar reference features. A hierarchical approach of similarity mapping is performed over $l = N_L$ levels. For every level l of the hierarchy, a similarity map between LR input subvectors and reference features is computed:

$$s^l_k = \phi^c(I_{LR}) * \frac{P_k(O^{l-1,r,m}_{ref})}{||P_k(O^{l-1,r,m}_{ref})||} \tag{1}$$

$k = c$ if $l = 1$, $k = r$ or $k = m$ otherwise. P is the patch derived from the application of the patch-match approach: patches of the reference features $O^{l-1,r,m}_{ref}$ are convoluted with subvectors $\phi^c(I_{LR})$ of the LR input to compute the similarity. When the similarity map s^l_k is evaluated, a vector O^l_{ref} containing the most similar features of O^{l-1}_{ref} is created by applying either one of two distinct approaches:

1. **Input attention mapping** ($l = 1$): in the first level a feature vector is created by maximising over every subvector of the input:

$$O^{1,r,m}_{ref}(x,y) = P_{k^*}(\phi^r(I^m_{ref}))(x,y) \tag{2}$$
$$k^* = \underset{k=c}{\operatorname{argmax}}\, s^1_k(x,y)$$

   $O^{1,r,m}_{ref}(x,y)$ is the (x,y) value of the k^* patch $P(\phi^r(I^m_{ref}))$ whose s^1 is the highest among all the similarity values $s^1_k(x,y)$ for each subvector of the LR input feature vector.
2. **Reference attention mapping** ($l > 1$): for subsequent levels of the hierarchy, a feature vector is created by maximising a new similarity s^l_k map over the feature vector created in the previous level.

$$O^{l,k}_{ref}(x,y) = O^{l-1,k^*}_{ref}(x,y) \tag{3}$$
$$k^* = \underset{k}{\operatorname{argmax}}\, s^l_k(x,y)$$

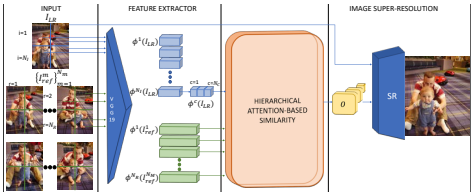


Figure 1: Overview of the approach. Given a LR input image and N_M reference images, the approach produces an HR reconstruction of the LR input image exploiting the references.

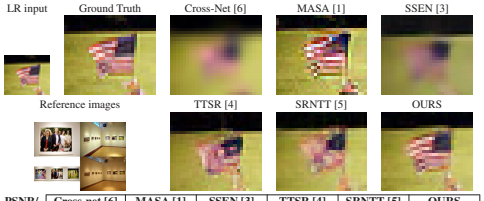
$k = r$ or $k = m$ depending on which level is processed. The value of $O^{l,k}_{ref}$ in the (x,y) position is the value of $O^{l-1,k}_{ref}$ with the highest s^l among all the $s^l_k(x,y)$ of $O^{l-1,k}_{ref}$.

The final output, obtained when the similarity mapping is performed for all the levels of the hierarchy, is a feature vector which contains the features of the references that are most similar to the features of the LR input.

Image Super-resolution: given the feature similarity mapping O, a generative adversarial network super-resolves the LR input to obtain the SR output which maintains the spatial coherence of the input with the HR appearance detail of the reference images. We modified the architecture of the generator of [5] by eliminating the batch normalization layers since they reduce the accuracy for dense pixel value predictions.

3 Results

We evaluate our method by comparing with state-of-the-art single RefSR approaches. Figure 2 shows the superiority of our approach.
We also confirmed (Figure 3) that increasing the number of reference images will lead to an improvement of the performance of the approach.



PSNR/ SSIM	Cross-net [6]	MASA [1]	SSEN [3]	TTSR [4]	SRNTT [5]	OURS
	26.00/.7576	24.84/.7311	22.71/.7169	25.59/.7645	26.42/.7738	**27.49/.8145**

Figure 2: Qualitative (top) and quantitative (bottom) comparisons with RefSR approaches.

PSNR/ SSIM	Ref. 1	Ref. 2	Ref. 4
	26.77/.7882	27.30/.8087	**27.49/.8145**

Figure 3: Qualitative (top) and quantitative (bottom) results of using different numbers of reference images.

[1] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. *arXiv preprint arXiv:2106.02299*, 2021.

[2] Marco Pesavento, Marco Volino, and Adrian Hilton. Attention-based multi-reference learning for image super-resolution. *arXiv preprint arXiv:2108.13697*, 2021.

[3] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020.

[4] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[5] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[6] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2018.

Human Point Cloud Generation using Deep Learning

Ryan Spick and James Walker
{ryan.spick,james.walker}@york.ac.uk
University of York,
York, Uk
Tim Bradley and Nigel John Williams
{nigel.williams,tim.bradley}@sony.com
Sony Interactive Entertainment,
London, UK

1 Introduction

Generative deep learning has been applied to a multitude of areas across many domains, each of these areas providing a different type of data from text, images, videos, and music [1, 3, 7, 8]. These examples use a variety of different network architectures, but the goal of each is to learn or exploit the underlying distribution of its training data. In this paper, a novel method of generating accurate pose and animation of human point cloud data using generative deep learning methods is presented, which uses dense correspondence based data, in which all points within the point cloud align with every other point in corresponding data points.

1.1 Point Clouds

Point clouds are sets of coordinates representing some multi-dimensional data, typically in a three-dimensional Cartesian coordinate frame, representing objects, surfaces, or shapes. In these cases, each point is represented with an x, y, and z component determining the geometric coordinate of each point in the cloud. These data points are usually the result of a type of 3D scanning such as LiDAR.

As of late, PointNet [5] type architectures, which facilitate the optimal consumption of point clouds directly by a neural network, have received a great deal of attention, though these approaches disregard potentially deep correspondences between points. PointNet++ [6] attempts to define weak correspondence through sampling overlapping regions/clusters, but this structure excludes any one to one point correspondence. If more information is known about the structure and layout of the point sets, then it is possible to derive well-defined correspondence that isn't specifically within the euclidean or geodesic space. PointNet architectures and their derivatives have been steadily applied in areas such as object detection [4], amongst many others.

2 Methodology

This work utilises the MPI-Faust data set [2] as the input data for the experiments and deep learning models in this paper. Each model file contains exactly 6890 points, where every point corresponds with every other point in the data-set. The data set consists of 10 unique models, of varying body structure and shape. There are 10 poses mirrored across every model. This dense formatting of the data allows for uniform learning of complex floating-point data through standard generative approaches, such as a Convolutional neural network or a static fully connected network. The MLP network is simply a fully connected network where every node in subsequent layers is connected to the previous layer's nodes. For the generator, the number of nodes increases with a factor of 2 each in each layer. The final output of the network increases to the size of the point cloud data, 6890, which is the size required for the input to the discriminator to match the real samples. The discriminator is a reverse of the generator, leading down to a node size of 1 with a Sigmoid activation, signaling if a sample passed in is determined real, or fake. A 1D convolutional network was also tested, but proved to have difficulties learning the symmetry of the data across different body shapes. We believe this was due to the inherent nature of the convolutions, having a lack of connection between those points that are close in space, but not in the data set.

2.1 Dot Based Loss Function

The idea of using the dot product is to add stability in the early stages of training. Because the data is in dense correspondence, a pre-computed dot product across the training data can be used to determine how well the generated samples conform to the original data distribution. A sample of the dot product of all of the training data was taken, where each dot product was taken for every point in each data point. Initially the calculation would take a point and its neighbouring point - where $D = Dot(N, N+1)$. Indexing all of the points proved far too inefficient to use during training, so random jittering was employed. This was changed to a stochastic approach where now - $D = Dot(N, N+step)$ with the step being a random value between roughly 0.8% and 1.2% of the data set size, the dot calculation was performed until N >= data length. This helped with performance without drastically reducing the quality of convergence to the pose shape.

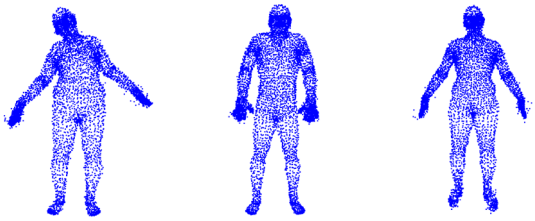


Figure 1: Generated examples from the MLP GAN, each model shows a different type of pose and body shape. Conditions were later added to control the type of pose and body shape through prior labelling.

3 Conclusion

This paper outlined a new method of loss calculation for ordered point cloud data, together with a deep parameter exploration of two neural network architectures has resulted in a robust method of generating new human pose and human poses animation from existing point cloud data. The idea of taking a prior dot product calculation and incorporating it into a weighted binary cross-entropy loss function provided a large stability increase when training the generator of the network. Subsequently improving the visual fidelity of human pose outputs and early training convergence.

[1] Adrián Barahona-Rıos and Sandra Pauletto. Synthesising knocking sound effects using conditional wavegan. In *17th Sound and Music Computing Conference, Online*, 2020.

[2] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. June 2014.

[3] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. Texture synthesis with spatial generative adversarial networks. *arXiv preprint arXiv:1611.08207*, 2016.

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[6] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[7] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.

[8] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

DEMO

Demo: Video Provenance Network for Robust Content Attribution

Alexander Black[1]
alex.black@surrey.ac.uk
Tu Bui[1]
t.v.bui@surrey.ac.uk
Simon Jenni[2]
jenni@adobe.com
Viswanathan Swaminathan[2]
vishy@adobe.com
John Collomosse[1 2]
collomos@adobe.com

[1] CVSSP, University of Surrey
[2] Adobe Research

Video is a powerful medium for storytelling. Yet, the ease with which digital video may be synthesized, manipulated and shared (*e.g.* via social media) presents a growing societal threat via the amplification of misinformation and fake news [2].

Videos often undergo various transformations during online distribution; changes in format, resolution, size, padding, effect enhancement *etc.* We present a system for matching partial video queries robust to such transformations. We build an inverse index of robust audio-visual features trained using contrastive learning and a rich set of augmentations representative of transformations typically applied to video 'in the wild' during online content distribution. We demonstrate our matching technique within the context of a system for tracing the provenance of video assets, using a corpus of 100,000 videos from the VGGSound dataset [1]. We demonstrate that our system is able to match fragments of video (*i.e.* partial or truncated videos) to determine not only the complete source video but also the time offset at which that fragment exist.

To learn a robust video fingerprinting model, we propose a self supervised network capable of encoding both visual and audio streams in a video. We leverage contrastive learning and a rich set of data augmentations for videos to train our model. To enable partial video matching, we follow a 'divide and conquer' approach where video is split into chunks and each chunk is indexed and search-able within an inverted index.

We train a CNN model to project a video frame to a compact embedding space. We employ the ResNet50 architecture [3], replacing the final classifier layer with a 256-D fully connected (fc) layer that serves as the embedding.

Similar to text search systems, we construct an inverted index that supports video retrieval at chunk-level. We sample 1M random descriptors and build a dictionary with codebook size K using KMeans. Given a database video, we break it into chunks where each chunk is represented as a bag of codewords. The K codewords are used as entries to our inverted index, listing all chunks in the database (a mapping between a chunk and video ID is also stored).

An advantage of our chunking and inverted index method is that it enables retrieval even if the query is only a segment of a database video. As a by product, our method also supports localization of a video segment by searching for the closest chunk in the database.

The user experience of the demo [1] is depicted in Figures 1 and 2. . The demo allows to create a custom query video, by selecting a temporal fragment from one of the VGGSound validation set videos and applying an augmentation to it. Generated query is used to search within 100,000 videos from the VGGSound dataset. The green/red text box above each of retrieval results indicates correct/incorrect retrieval results. The heatmap bar shows the edit distance between the query sequence of codewords and a same-length segment of the candidate video in sliding window fashion, which could be used to represent the confidence in localization of the query within the candidate video.

[1] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. VGG-sound: A large scale audio-visual dataset. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[2] S. Gregory. Ticks or it didn't happen. Technical report, Witness.org, 2019.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[1] Video recording of the demo is available at `https://youtu.be/c4Qv9IqD4J4`

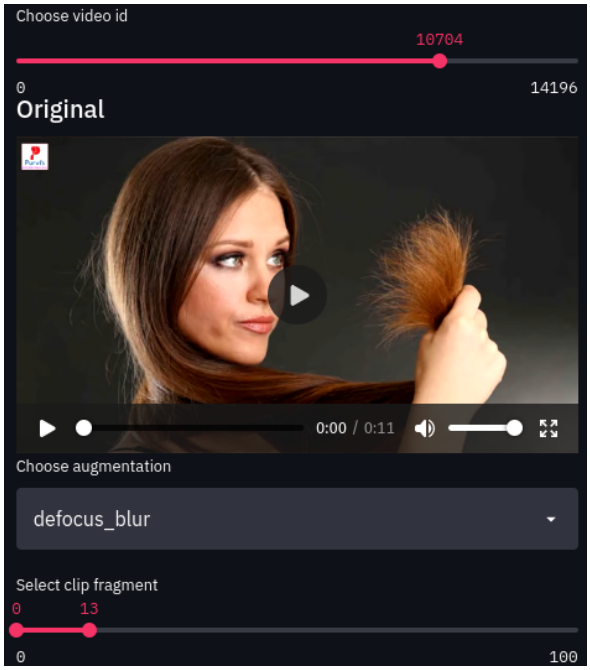


Figure 1: The user can generate a custom query based on any one of the 14,196 videos from VGGSound validation set. The user is prompted to select an augmentation type as well as start and end points of a clip fragment.

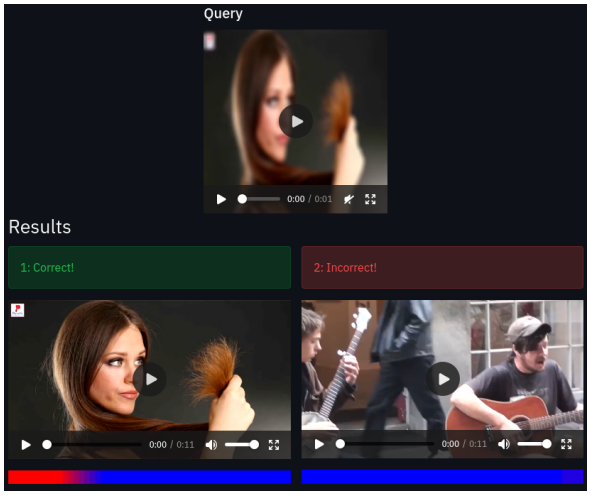


Figure 2: After the query is generated, it is presented to the user, along with top k results. The heatmap bars represent the confidence in localization of the query within the candidate video.

NOTES

NOTES

NOTES

CHAIRS

Conference Chairs
Rafał Mantiuk, University of Cambridge
Christian Richardt, University of Bath

Full Papers Chair
Marco Volino, University of Surrey

Short Papers & Demos Chair
Armin Mustafa, University of Surrey

Industry Chairs
Duygu Ceylan, Adobe Research
Ilke Demir, Intel Corporation

Sponsorship Chair
Bernhard Egger, University of Erlangen-Nürnberg

Public Relations Chair
Peter Vangorp, Edge Hill University

Conference Secretary
Alex King, University of York

Programme Committee
Akin Caliskan, University of Surrey
Dan Casas, Universidad Rey Juan Carlos
Robert Dawes, BBC Research
Daljit Singh Dhillon, Clemson University
Peter Eisert, Fraunhofer Heinrich Hertz Institute
Andrew Gilbert, University of Surrey
Tom Fincham Haines, University of Bath
Oliver James, Double Negative
Hansung Kim, University of Southampton
Koki Nagano, NVIDIA
Alexandros Neophytou, Microsoft
Marco Pesavento, University of Surrey
Erik Reinhard, Technicolor
Nadejda Roubtsova, University of Bath
William Smith, University of York
Kartic Subr, University of Edinburgh
Graham Thomas, BBC
Zhidong Xiao, Bournemouth University

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

Conference Sponsors 2021

ACMSIGGRAPH

ACM ISBN: 978-1-4503-9094-1
Copyright © 2021 by the Association for Computing Machinery, Inc