Day 10: Part 1

Speakers: Prof Christian Richardt, University of Bath

Title: Towards Reconstructing + Editing the Visual World

Space Life Capture

4 Papers will be discussed in the sessions

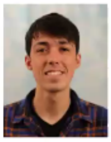
Roadmap

#	Project	Venue
1	MatryODShka	ECCV 2020 (oral)
2	Deep Video Portraits	SIGGRAPH 2018
3	HoloGAN	ICCV 2019
4	BlockGAN	NeurIPS 2020

ECCV'20 ONLINE 23-28 AUGUST 2020

MatryODShka: Real-time 6DoF Video View Synthesis using Multi-Sphere Images

Benjamin Attal [1, 2] Selena Ling [1] Aaron Gokaslan [1] Christian Richardt [3] James Tompkin [1]





[1] BROWN [2] Carnegie Mellon University [3] UNIVERSITY OF BATH

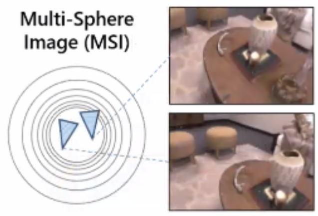
Multi-plane and multi-sphere images

- Multi-plane images
  - Perspective view synthesis
  - RGB + alpha layers
  - Inferred from plane sweep volumes

Disney's multi-plane camera

Multi-plane and multi-sphere images

- Multi-plane images
  - Perspective view synthesis
  - RGB + alpha layers
  - Inferred from plane sweep volumes
- Multi-sphere images
  - 360° view synthesis
  - RGB + alpha layers
  - Inferred from sphere sweep volumes

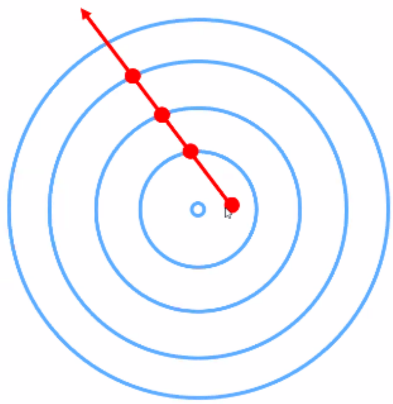


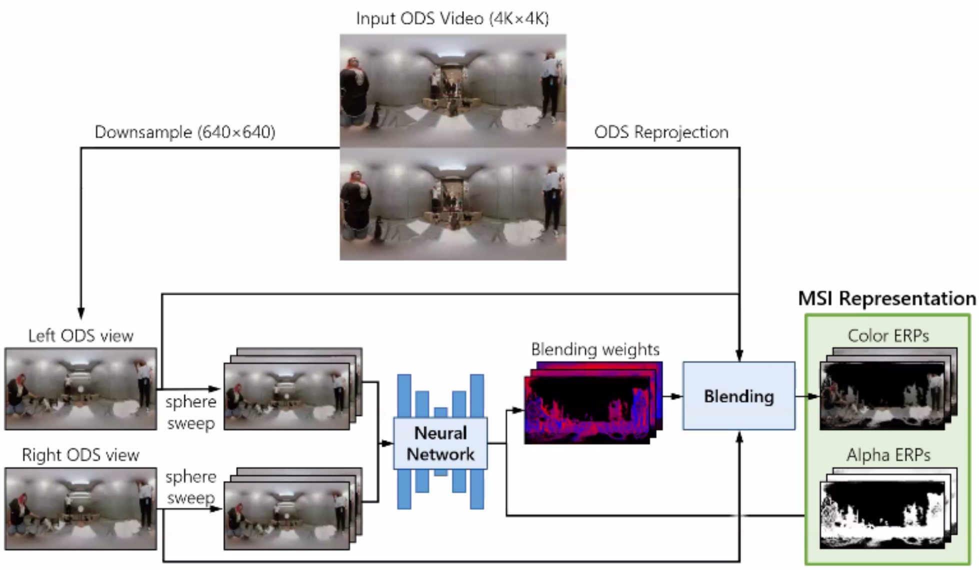
Matryoshka + ODS = MatryODShka

Multi-sphere image rendering

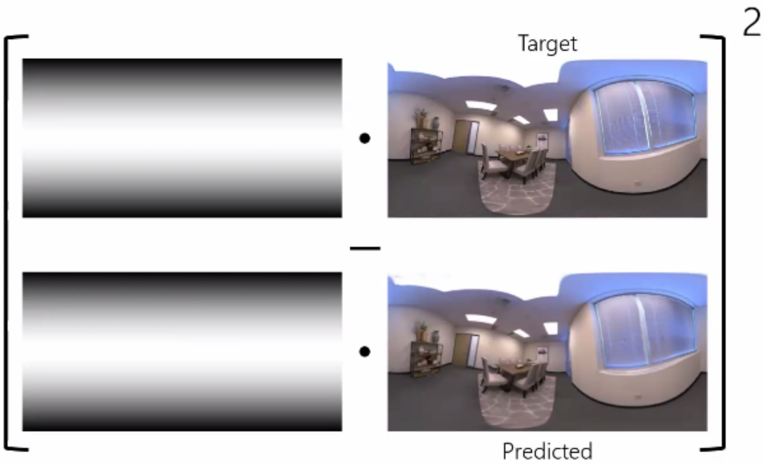
1. Intersect ray with each layer of MSI
2. Over-composite colors $\mathbf{c}_i$ and alphas α_i of intersection points:

$$\mathbf{c} = \sum_{i=1}^{N} \mathbf{c}_i \cdot \underbrace{\alpha_i \cdot \prod_{j=1}^{i-1} (1 - \alpha_j)}_{\text{Net opacity of layer } i}$$

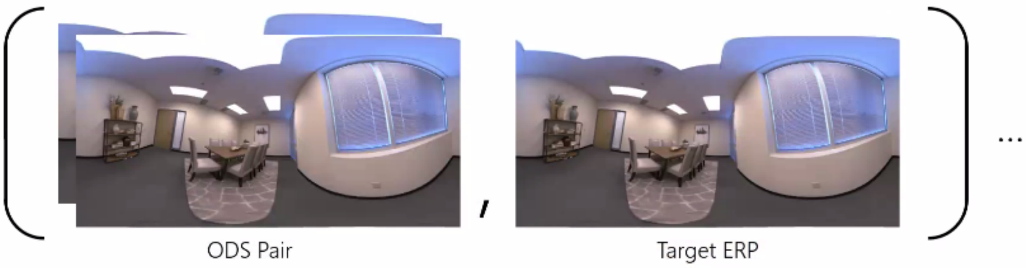




Spherically adapted training loss



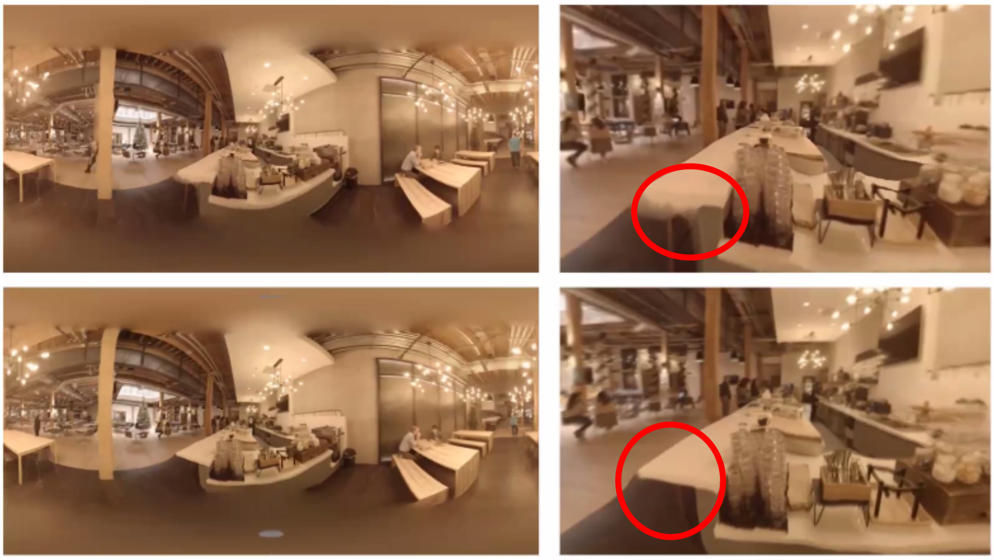
Training data



Straub et al. (2019). The Replica dataset: A digital replica of indoor spaces

Panoramic view synthesis

	Equirectangular View	Perspective View
Serrano et al. [IEEE VR 2019]		
MatryODShka [Attal et al. 2020]		



Concurrent work on multi-sphere images

Live view synthesis

Immersive Light Field Video
[Broxton et al., SIGGRAPH 2020]

- 46 camera input
- Offline (25 CPU hours per frame)
- High-resolution inference

MatryODShka (Ours)
[Attal et al., ECCV 2020]

- ODS input
- Online (30 Hz)
- Low-resolution inference

Grand Canyon

Live view synthesis

Live view synthesis

Deep Video Portraits

Hyeongwoo Kim[1] Pablo Garrido[2] Ayush Tewari[1] Weipeng Xu[1]
Justus Thies[3] Matthias Nießner[3] Patrick Pérez[2]
Christian Richardt[4] Michael Zollhöfer[5] Christian Theobalt[1]

[1] MPI Informatics [2] Technicolor [3] TU Munich [4] University of Bath [5] Stanford University

GENERATIONS SIGGRAPH2018

GENERATIONS / VANCOUVER SIGGRAPH2018

Contribution

Head Pose along with expression Captured

Original video | Pose editing | Expression editing

- Editing of head pose, rotation, face expression and eye gaze
- Combination of model-based face capture and CNN

Video: courtesy of UK government (Open Government Licence)

Related Work

Model-based face capture and reenactment

Garrido et al., ToG 2016

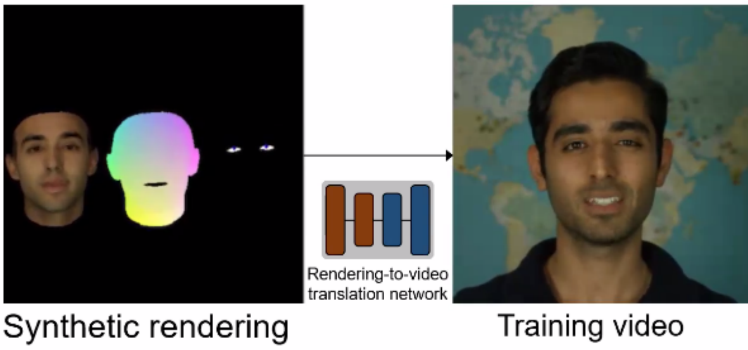
Kemelmacher-Shlizerman et al., ECCV 2010
Shi et al., ToG 2014
Suwajanakorn et al., ICCV 2015
Thies et al., CVPR 2016
Averbuch-Elor et al., ToG 2017
Thies et al., SIGGRAPH 2018

CNN-based methods

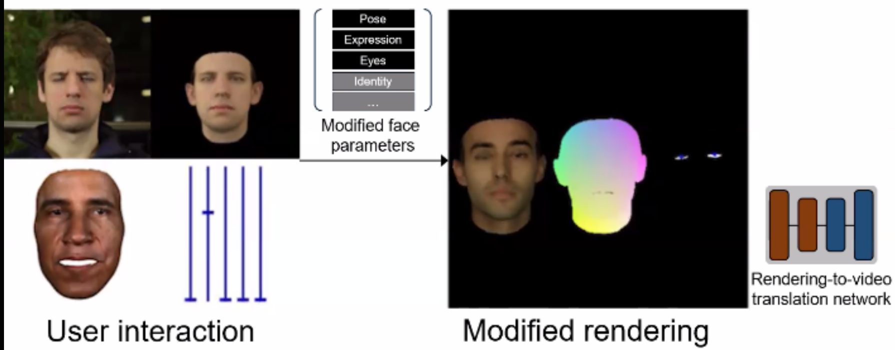
Karras et al., ICLR 2018

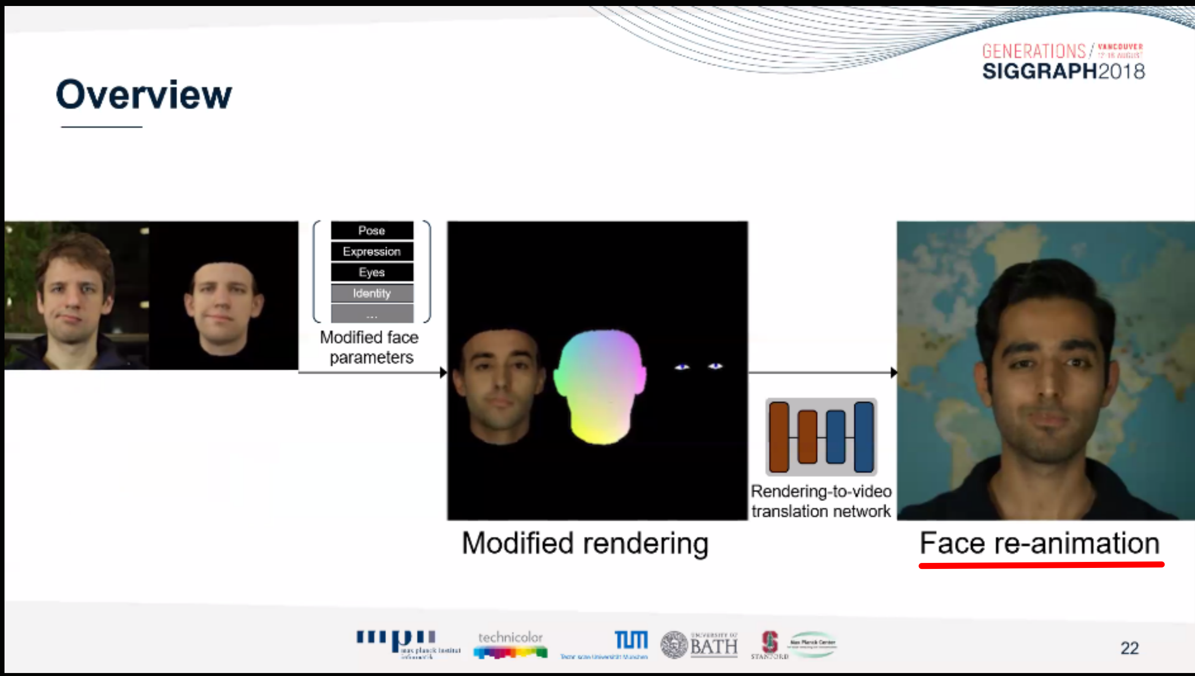
Goodfellow et al., NIPS 2014
Isola et al., CVPR 2017
Chen and Koltun, ICCV 2017
Tewari et al., ICCV 2017
Olszewski et al., ICCV 2018
Wang et al., CVPR 2018

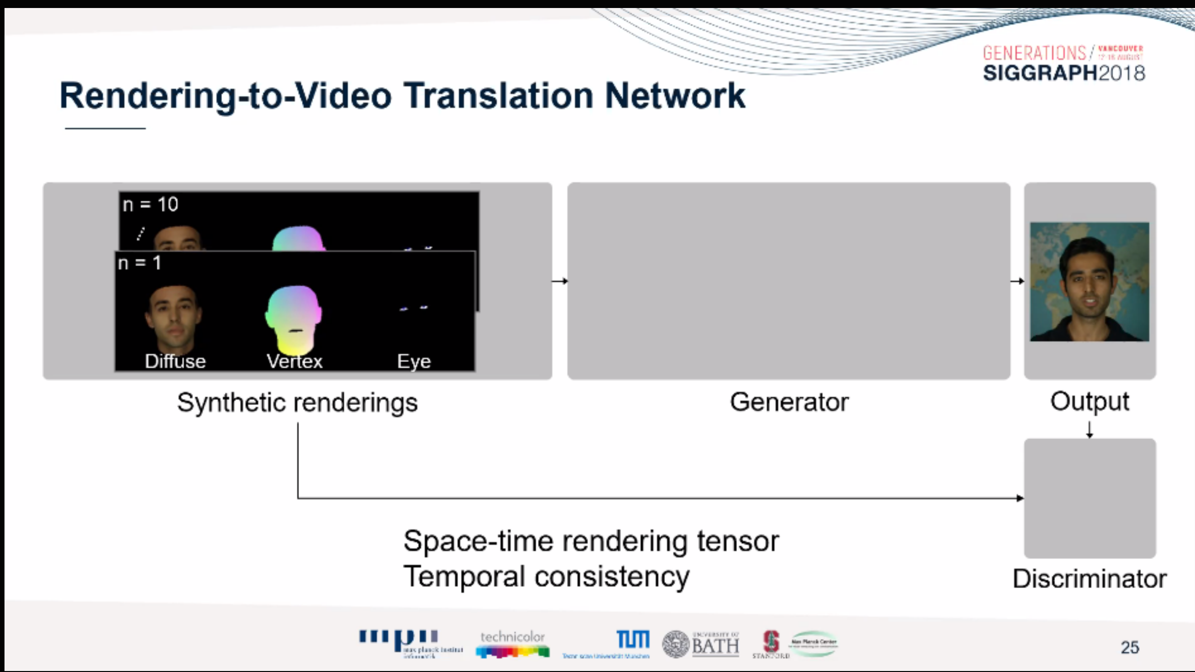
Overview

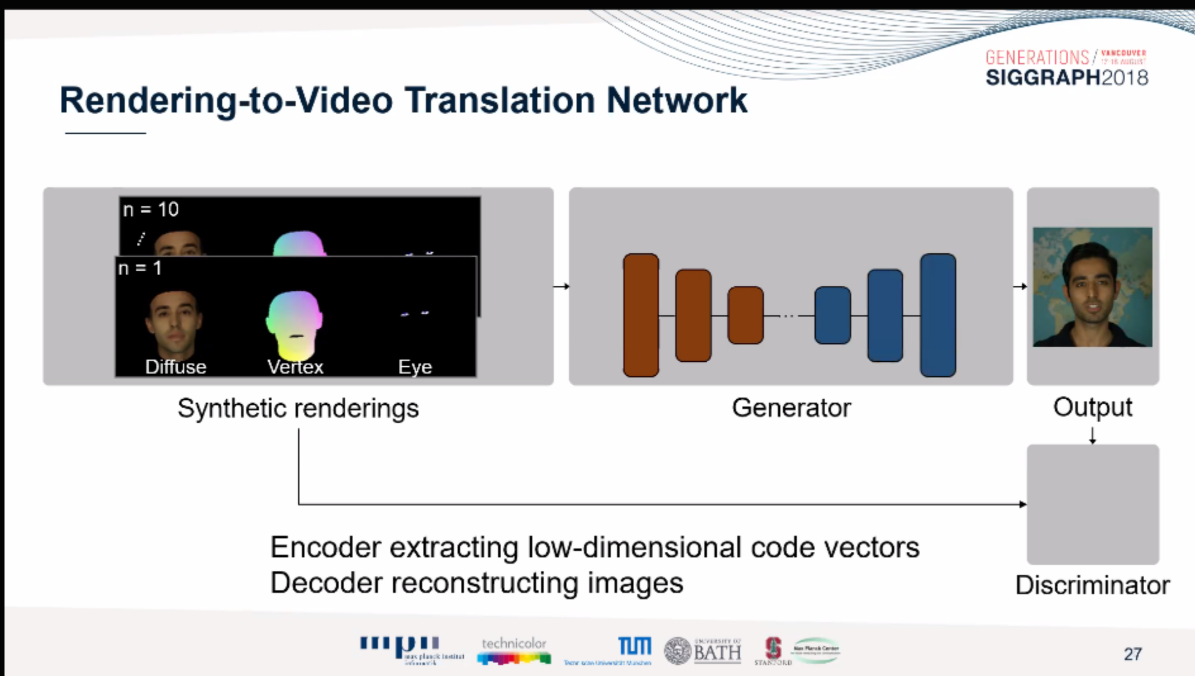


Overview



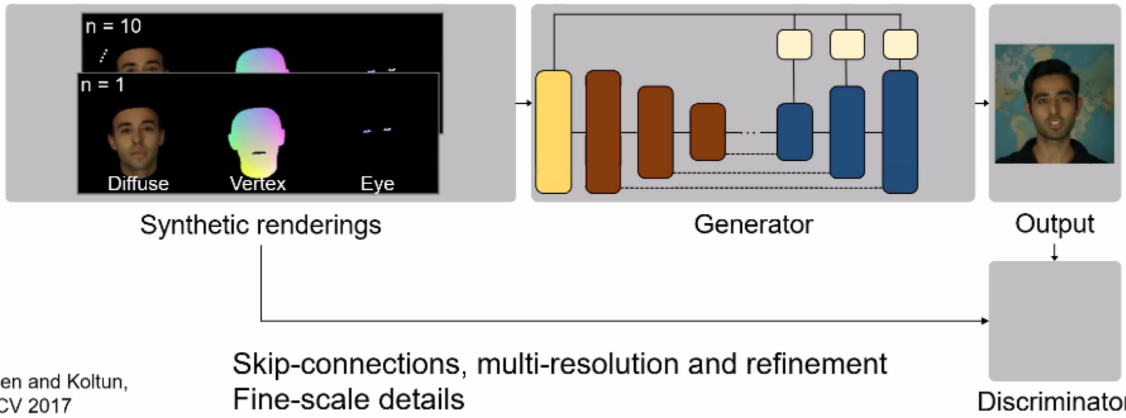

GENERATIONS / VANCOUVER
SIGGRAPH2018
Overview
Pose
Expression
Eyes
Identity
...
Modified face parameters
Modified rendering
Rendering-to-video translation network
Face re-animation
technicolor
BATH
22


GENERATIONS / VANCOUVER
SIGGRAPH2018
Rendering-to-Video Translation Network
n = 10
n = 1
Diffuse
Vertex
Eye
Synthetic renderings
Generator
Output
Space-time rendering tensor
Temporal consistency
Discriminator
technicolor
BATH
25


GENERATIONS / VANCOUVER
SIGGRAPH2018
Rendering-to-Video Translation Network
n = 10
n = 1
Diffuse
Vertex
Eye
Synthetic renderings
Generator
Output
Encoder extracting low-dimensional code vectors
Decoder reconstructing images
Discriminator
technicolor
BATH
27

GENERATIONS / VANCOUVER
SIGGRAPH2018

Rendering-to-Video Translation Network

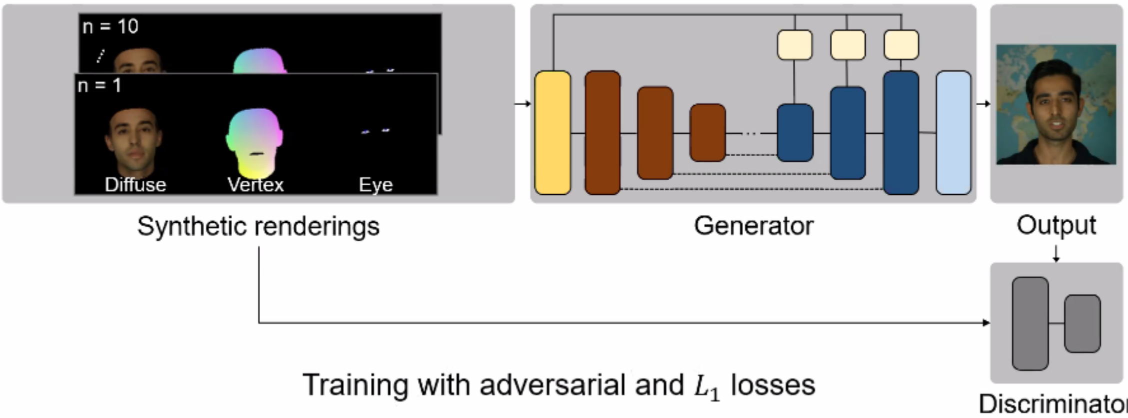


Skip-connections, multi-resolution and refinement
Fine-scale details

Chen and Koltun, ICCV 2017

GENERATIONS / VANCOUVER
SIGGRAPH2018

Rendering-to-Video Translation Network



Training with adversarial and L_1 losses

GENERATIONS / VANCOUVER
SIGGRAPH2018

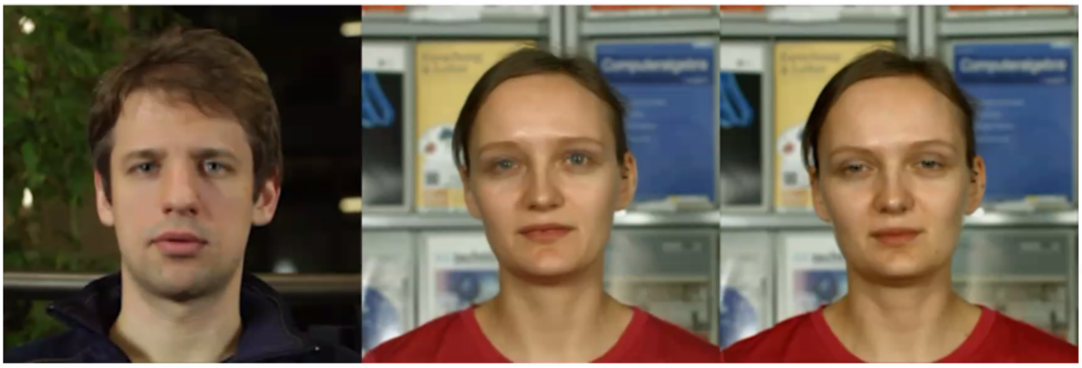
Result: Facial Reenactment

Full reenactment of head pose, head rotation, face expression and eye gaze

Source Target

Result: Facial Reenactment

Full reenactment of head pose, head rotation, face expression and eye gaze



Source | Target | Face2Face (Thies et al., 2016)

Result: Facial Reenactment

Source | Target | Result

Result: Visual Dubbing

Visual discomfort due to the discrepancy between video and audio tracks

Dubbing actor video | Original video

Result: Visual Dubbing

Modification of mouth motion to match audio tracks

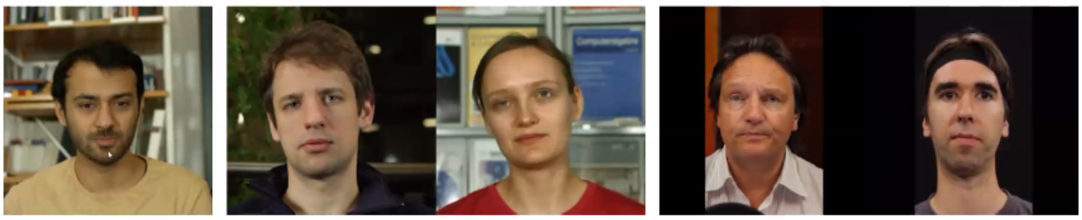
Dubbing actor video | Dubbed video | Garrido et al., 2015

Result: Interactive Editing

Pose | Expression | Shape

Approximately 9 fps

Summary

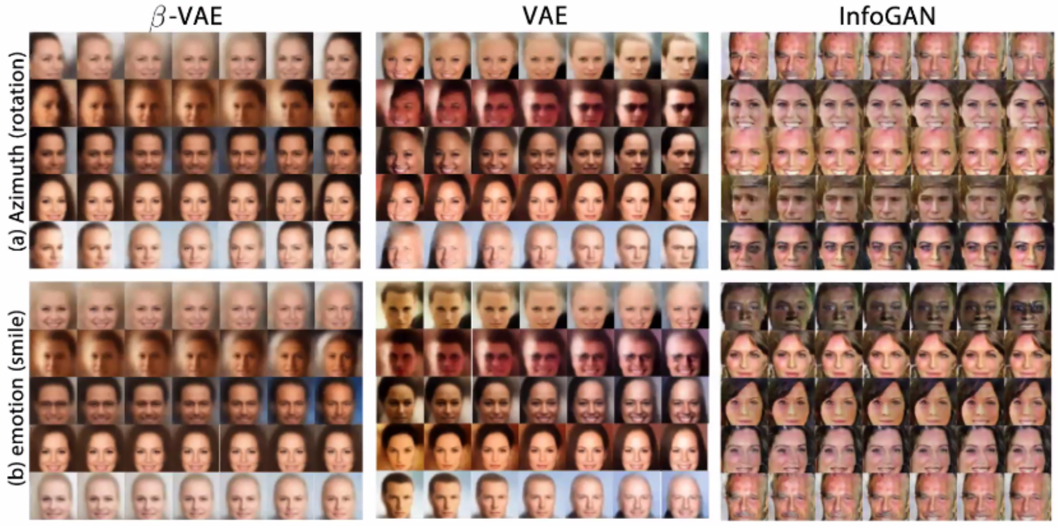


Future work:

- Pushing toward higher quality and resolution
- Video authentication and forensics

*missed slide (to be updated)

Disentangled representation learning



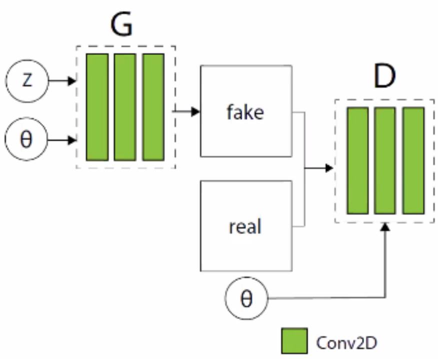
[Higgins et al., beta-VAE, 2014]

[Chen et al., InfoGAN, 2015]

HoloGAN

Multi objects

Conditional GANs

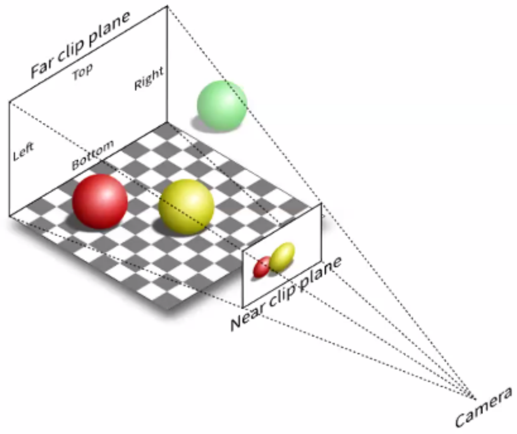


Computer
Graphics →
Basics

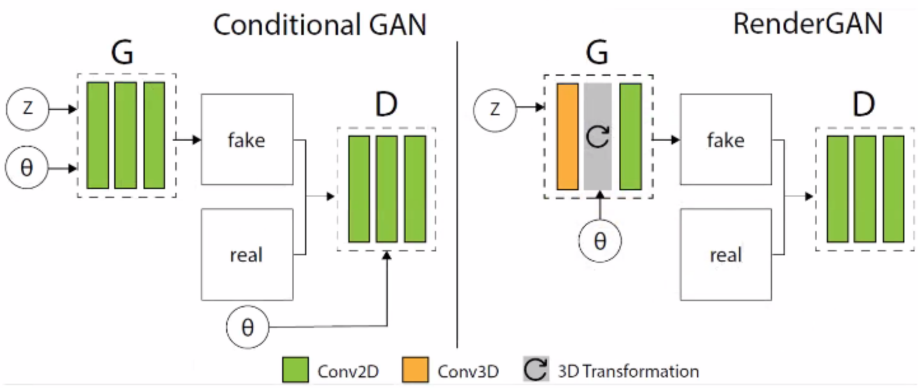
Back to the basics ...

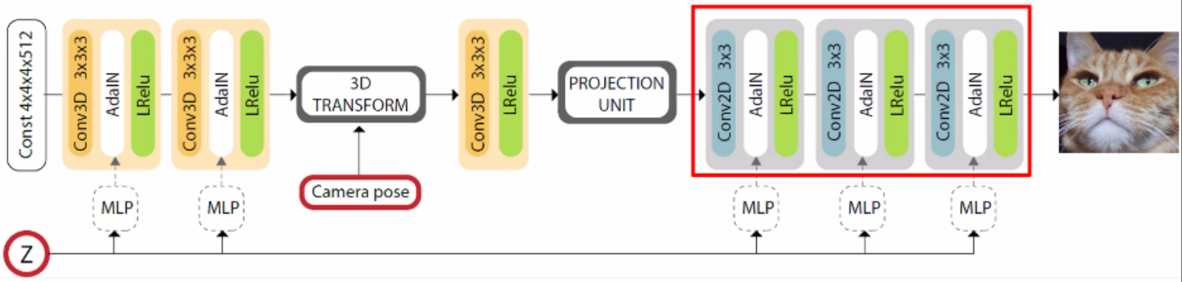
Rendering comprises two main steps:

- Compute visibility
- Compute shading



GAN architecture comparison





$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(T(z))))]$$

Datasets

Basel

CelebA

Cats

Chairs

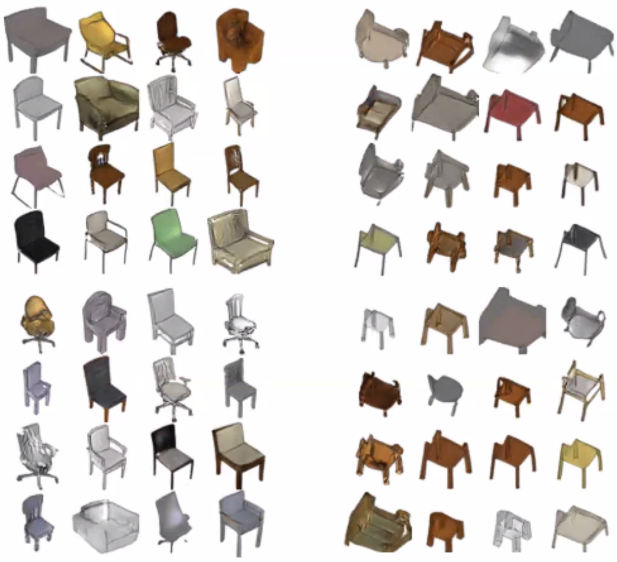
Cars

Cats

CelebA

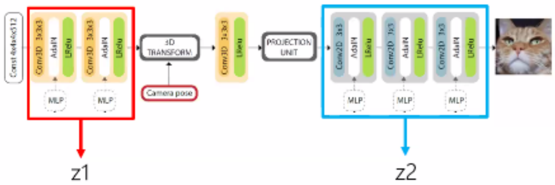
Basel Face Model

Chairs



Cars

Style mixing



Separating shape + appearance

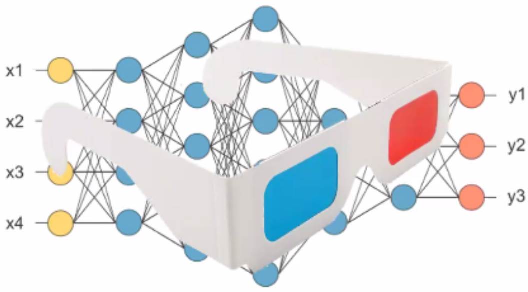
Separating shape + appearance

2021-08-13 Christian Richardt – Towards reconstructing and editing the visual world 59

The next frontier — complex scenes

HoloGAN conclusion

- Adding inductive bias about the 3D world to a neural network
  - Better image quality
  - Better 3D understanding
  - Completely <u>unsupervised</u>



BlockGAN

Learning 3D object-aware scene representations from unlabelled images

NeurIPS 2020

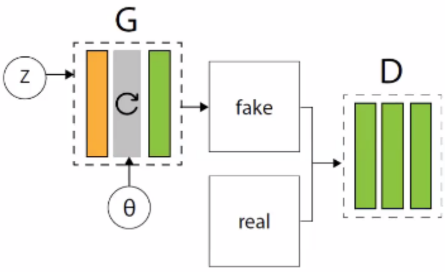
Thu Nguyen-Phuoc Christian Richardt Long Mai Yong-Liang Yang Niloy Mitra

Current work

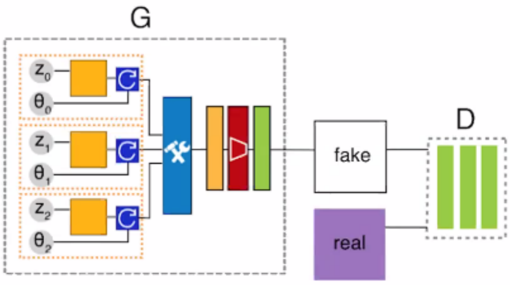
Scene Representation Networks
[Sitzmann et al. 2019]

HoloGAN
[Nguyen-Phuoc et al. 2019]

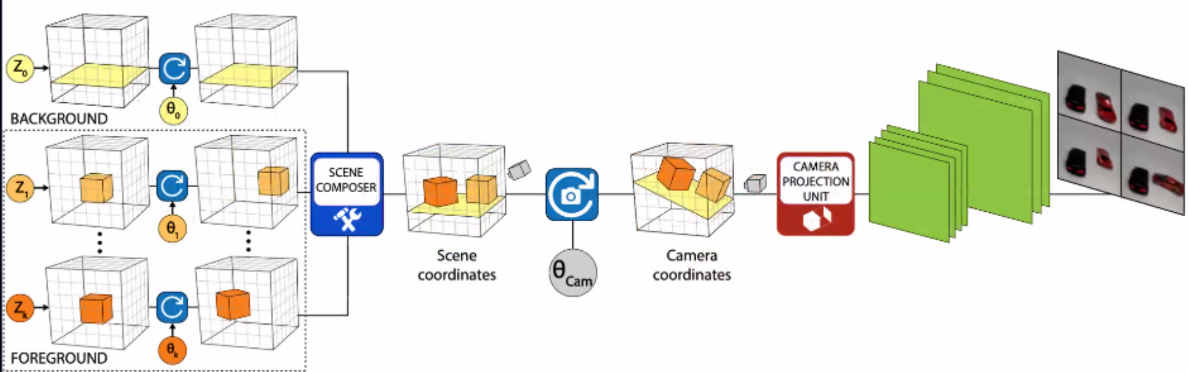
HoloGAN



BlockGAN



BlockGAN architecture



Datasets

Chair 1

Car 2

CLEVR 4

Cars (64×64)

Rotation

* all datasets rotated front & back

Translation

* up & down translati-on

Changing background appearance

light & shadow

Changing object #1 appearance

Rotation

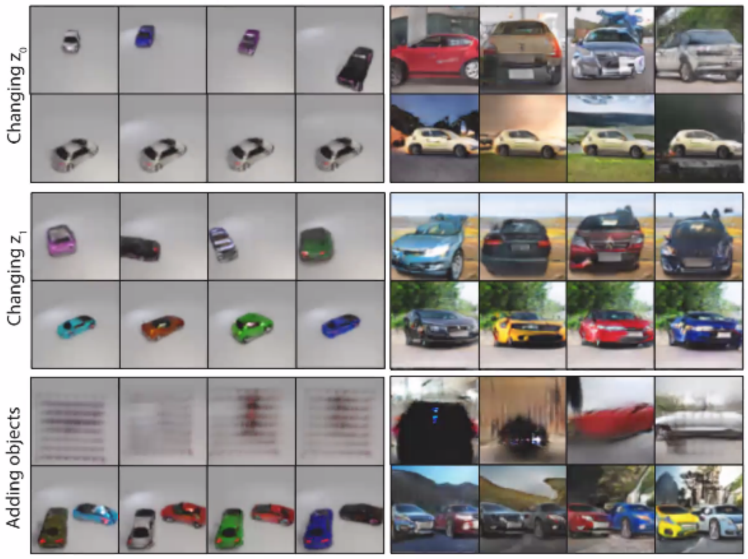
Translation

Change background

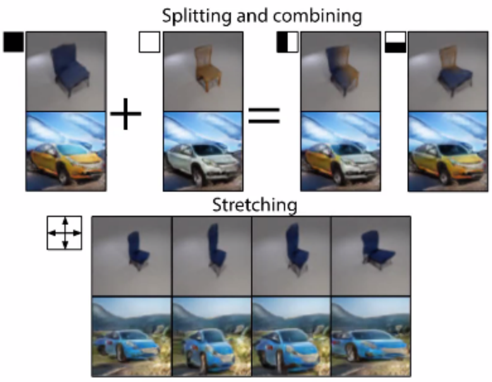
Change foreground

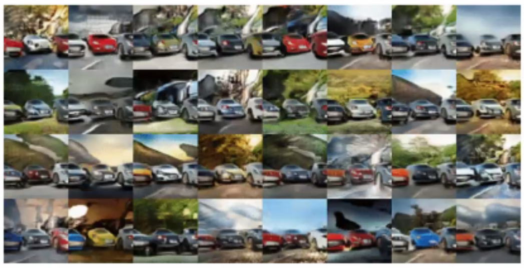
Generalisation

Comparison with 2D methods

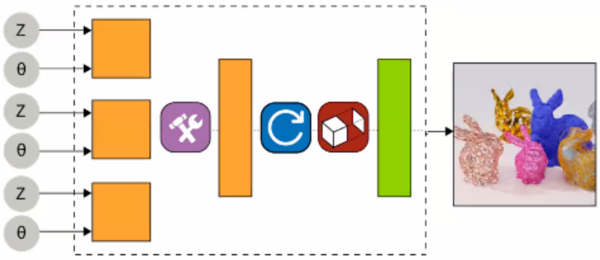


Generalisation





BlockGAN



- BlockGAN offers control over pose and appearance of individual object in the scene
- Deep voxel representations offer intuitive object spatial manipulation and composition

Takeaways

- Traditional visual computing approaches build on explicit models
- Explicit models have limits:
  - sometimes difficult to model the real world, e.g. realistic faces
- Machine learning is reinvigorating visual computing:
  - can learn to model relationships from sufficient data (& expressive network)
  - no explicit model necessary, but often a black box …
  - deep learning in particular has shown amazing results
- Benefits to applying domain knowledge to structure networks:
  - helps learning + we know what's going on!

Questions?

Engineering and Physical Sciences Research Council

Horizon 2020
MSCA FIRE

Christian Richardt

Towards reconstructing + editing the visual world

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

UNIVERSITY OF BATH

richardt.name
c_richardt