Trinity College Dublin

The University of Dublin

V-SENSE

Perception and Quality of Immersive Media

Professor Aljosa Smolic

SFI Research Professor of Creative Technologies

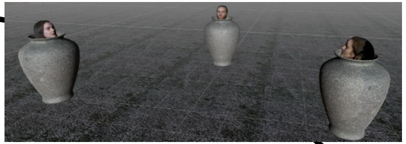
Trinity College Dublin
The University of Dublin
V-SENSE

Science Foundation Ireland sfi For what's next

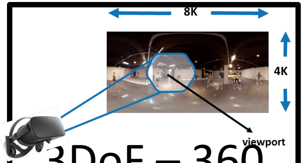
- **Extending Visual Sensation through Image-Based Visual Computing**
- **Visual computing at the intersection of**
    - Computer graphics
    - Computer vision
    - Media technology
- **Algorithms on pixels from capture to display**
- **Immersive visualization, XR (VR, AR), light fields**

V-SENSE Research Areas

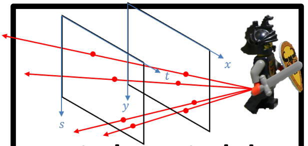
Creative
Experiments and Demonstrations



Visual Effects & Animation



3DoF – 360 VR Video

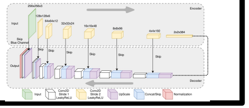


Light Field Imaging

6DoF – AR/VR & Free Viewpoint Video

Coding/Streaming

Deep Learning for Visual Computing



Quality, Visual Attention

Outline

- **Introduction to Visual Attention and Saliency**
- **Omnidirectional Video – 3DoF**
- **Volumetric Video – 6DoF**
- **Light Fields**
- **Summary**

Trinity College Dublin
The University of Dublin

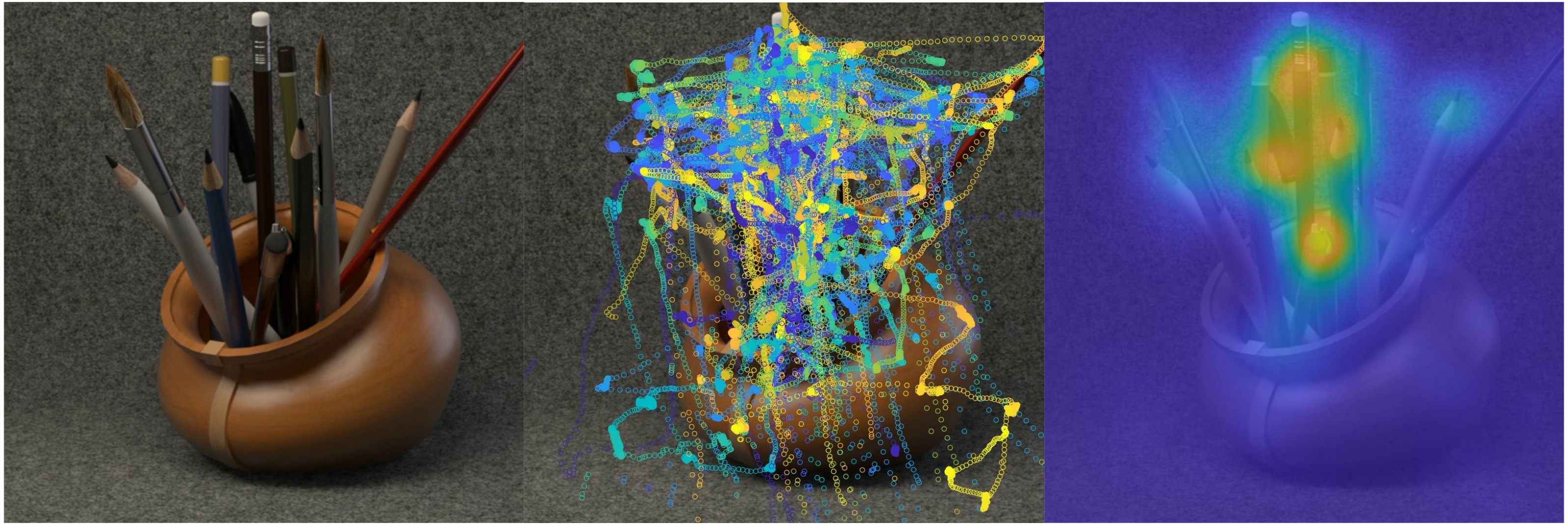
V-SENSE

Introduction

Professor Aljosa Smolic
SFI Research Professor of Creative Technologies

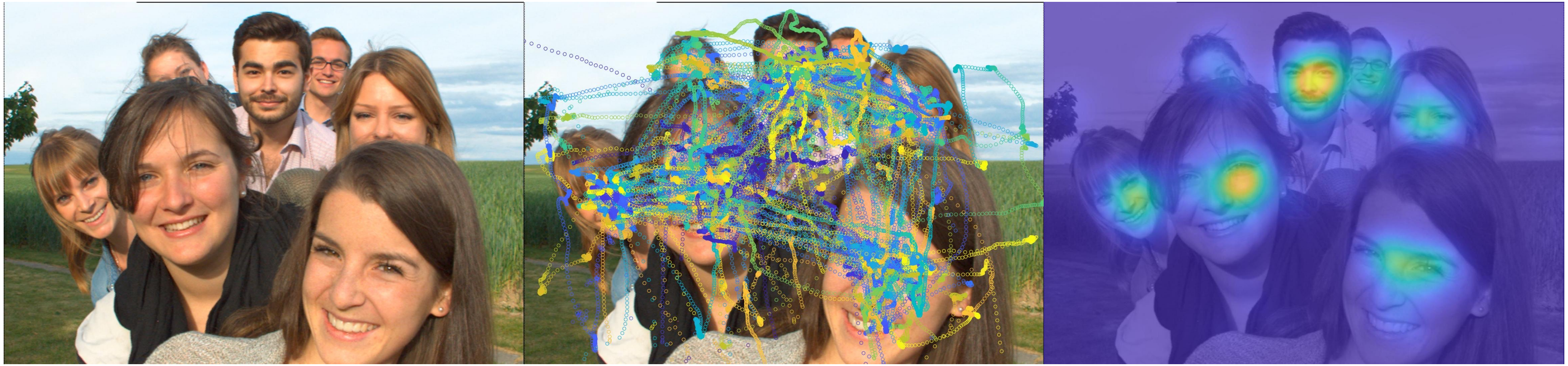
Visual Attention

Where people look when viewing a visual scene.



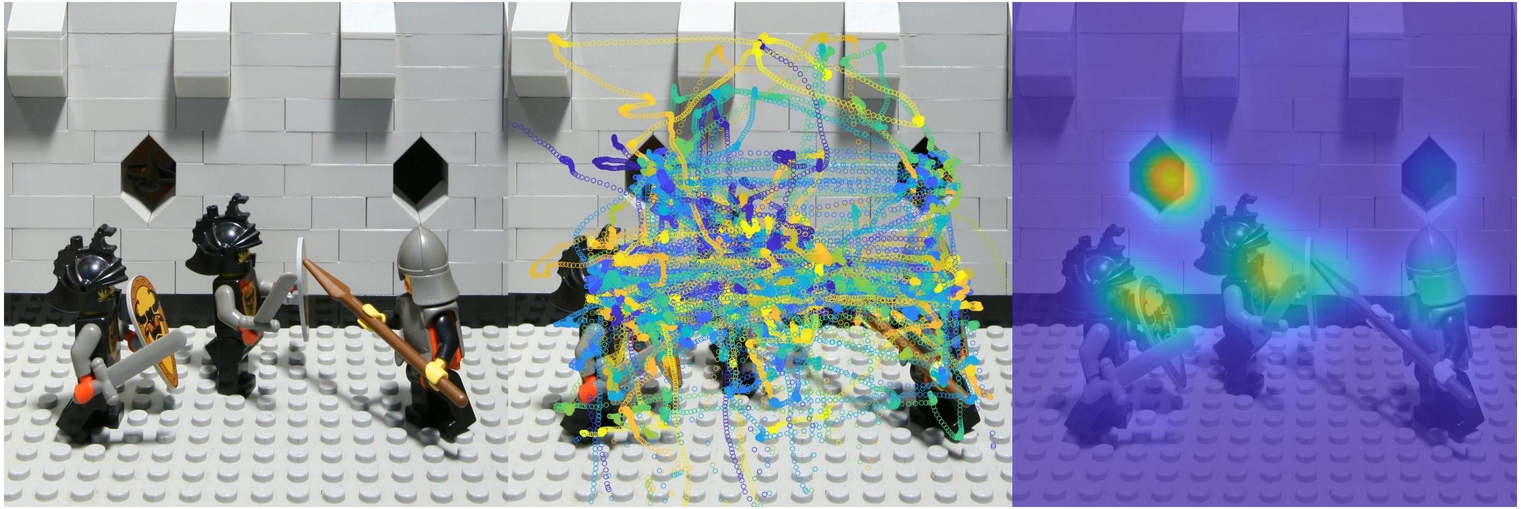
Visual Attention

Where people look when viewing a visual scene.



Visual Attention

Where people look when viewing a visual scene.

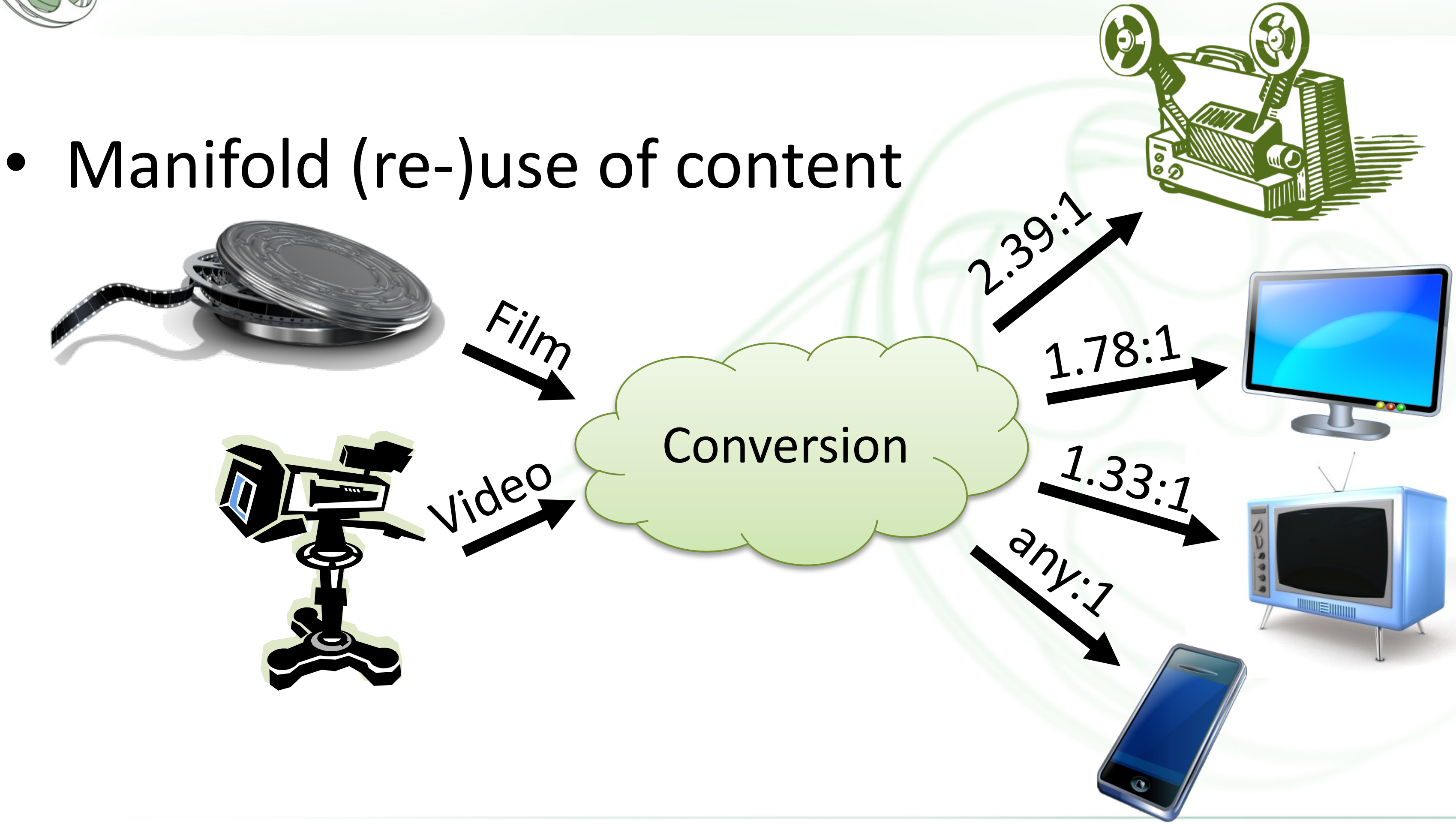


Purpose Visual Attention and Saliency

- **Understanding human perception**
- **Analysing/evaluating properties of the content**
- **Assigning resources to important parts of the content**
  - Coding/compression, streaming
  - Rendering
- **Quality assessment**
- **Optimizing algorithms driven by perceptual priority**

Current Video Distribution

- Manifold (re-)use of content



Film

Video

Conversion

2.39:1

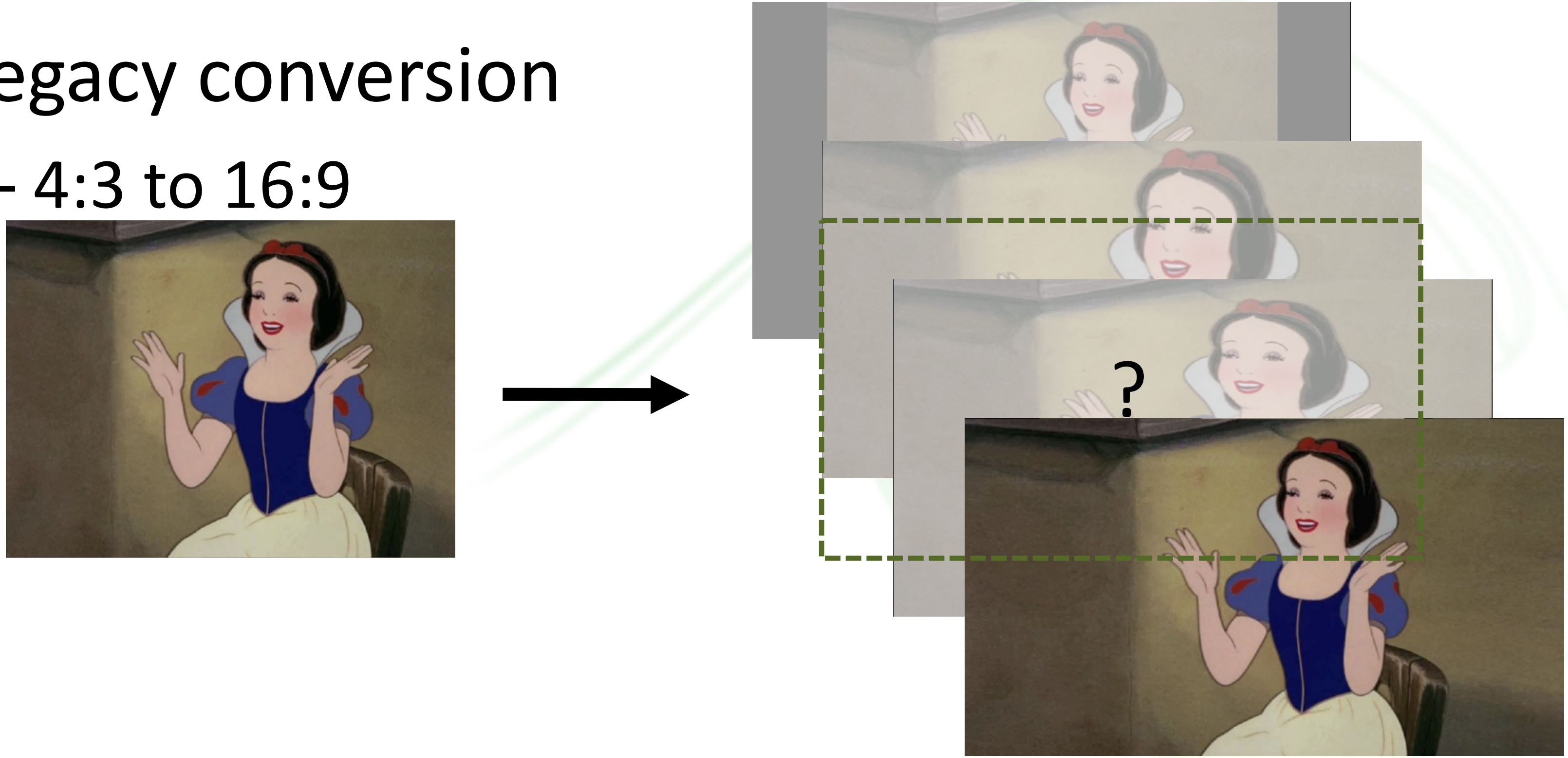
1.78:1

1.33:1

any:1

Linear Aspect Ratio Conversions

- Legacy conversion
  - 4:3 to 16:9

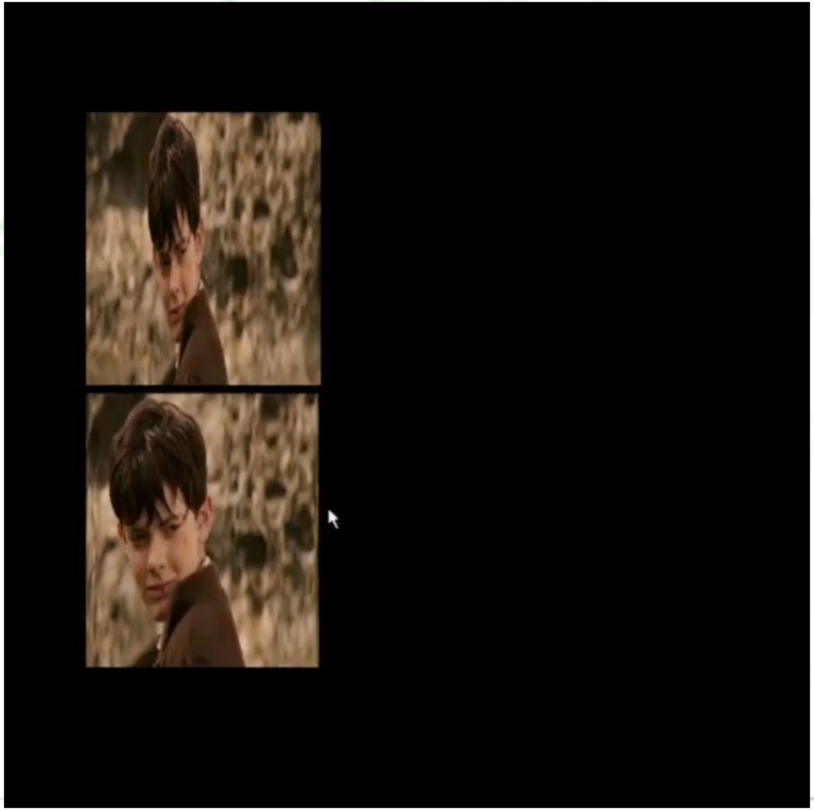


Intelligent Video Retargeting

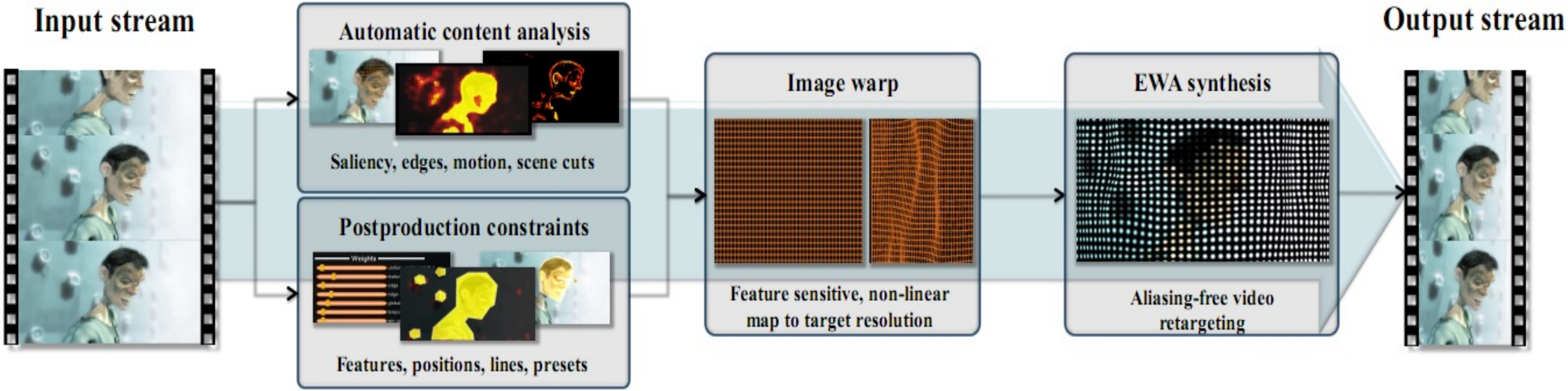
- Principle: scale visually less important regions

Stretching:

Our solution:



Video Retargeting


Input stream
Automatic content analysis
Saliency, edges, motion, scene cuts
Postproduction constraints
Features, positions, lines, presets
Image warp
Feature sensitive, non-linear map to target resolution
EWA synthesis
Aliasing-free video retargeting
Output stream

A System for Retargeting of Streaming Video

Philipp Krähenbühl, Manuel Lang, Alexander Hornung, Markus Gross

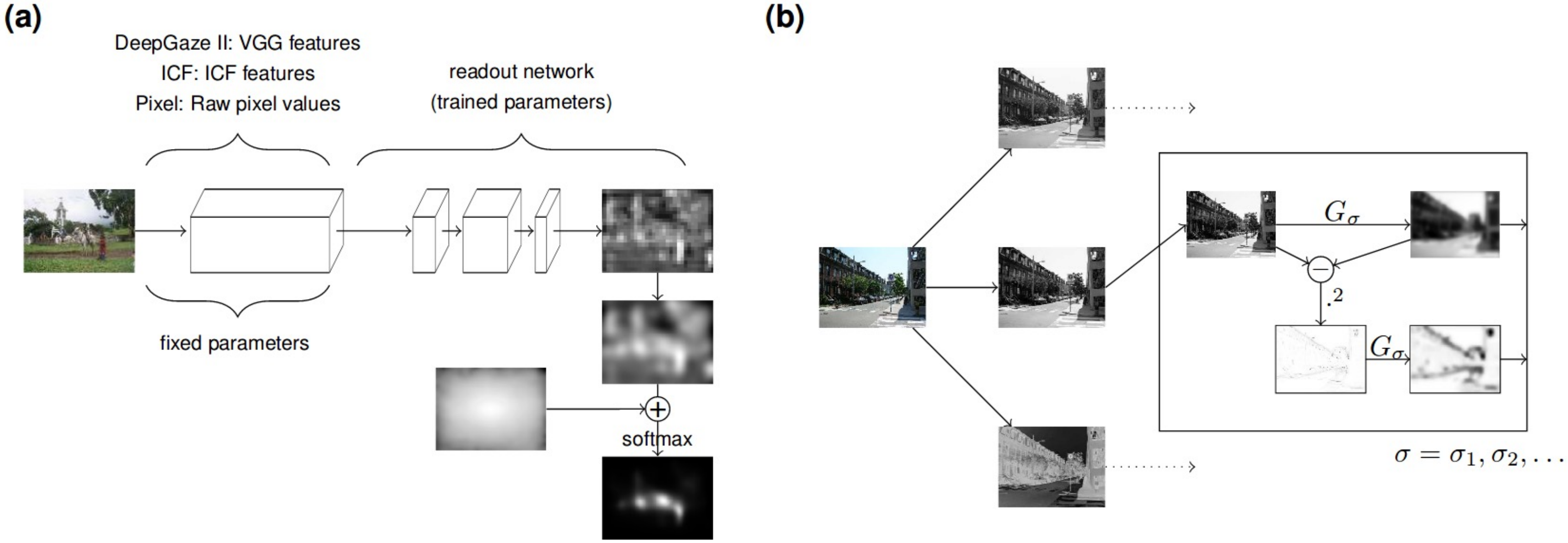
SIGGRAPH ASIA 2009

Vintage Saliency Estimation

- **Computational modelling of human visual perception**
- **Detectors of important visual features including:**
  - Faces, humans
  - Text
  - Colour, texture, edges
  - For video: motion
  - Etc.
- **Handcrafted algorithms validated through comparison to ground truth eye tracking data**

State-of-the-Art Saliency Estimation

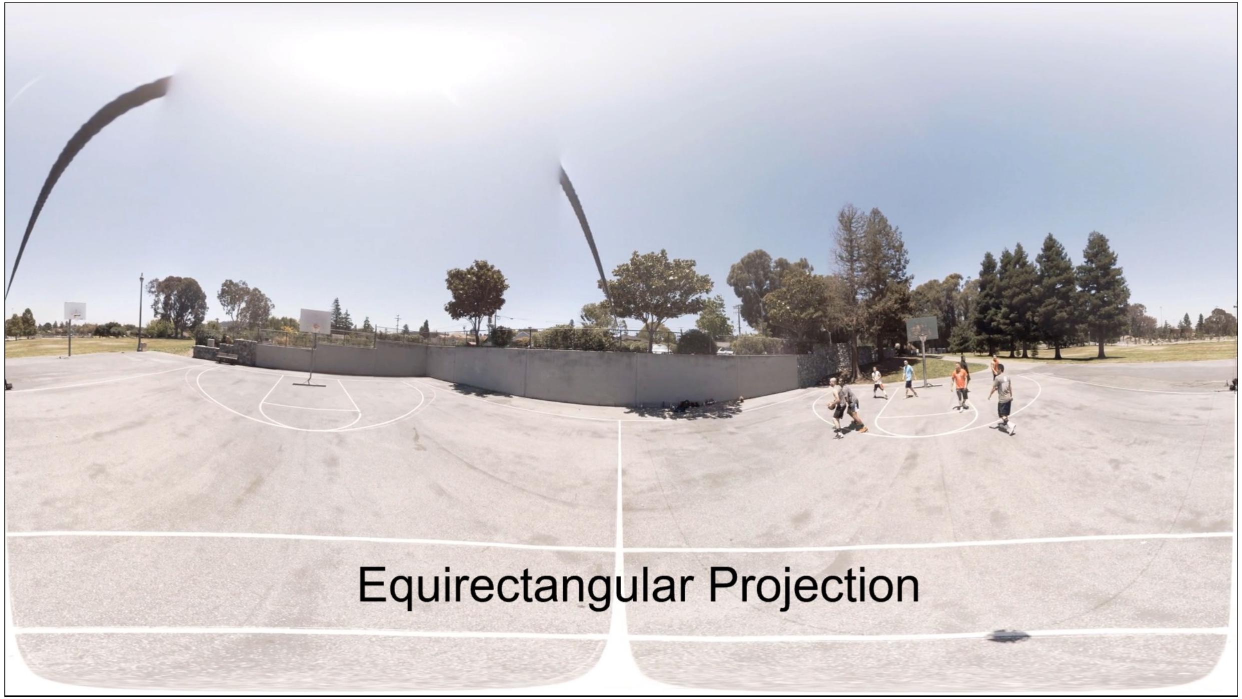
- **Deep learning**



M. Kümmerer, T. S. Wallis, and M. Bethge, "**DeepGaze II: Reading fixations from deep features trained on object recognition"** arXiv preprintarXiv:1610.01563, 2016.

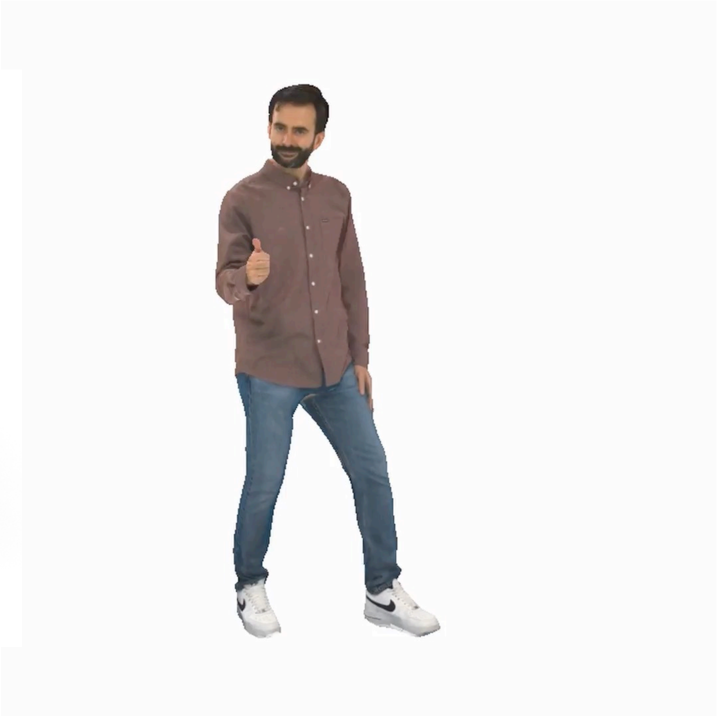
ODV – 3DoF Interaction

Viewing characteristics: free look around in 3DoF



VV – 6DoF Interaction

Viewing characteristics: free look around in 6DoF



eXtended Reality (XR) Content in 6DoF

Augmented and virtual reality experiences at V-SENSE

Augmented Reality

Virtual Reality

LFs – 6DoF Interaction and Refocusing

Viewing characteristics: limited look around in 6DoF and refocusing

Perception of Immersive Media

- **User interaction poses novel challenges for understanding of visual attention and saliency of immersive media**
- **Modelling of user behaviour becomes important**
- **Saliency models have to incorporate user interaction and content properties**

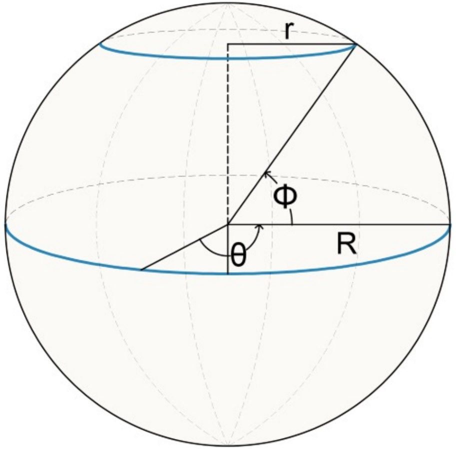
Trinity College Dublin
The University of Dublin

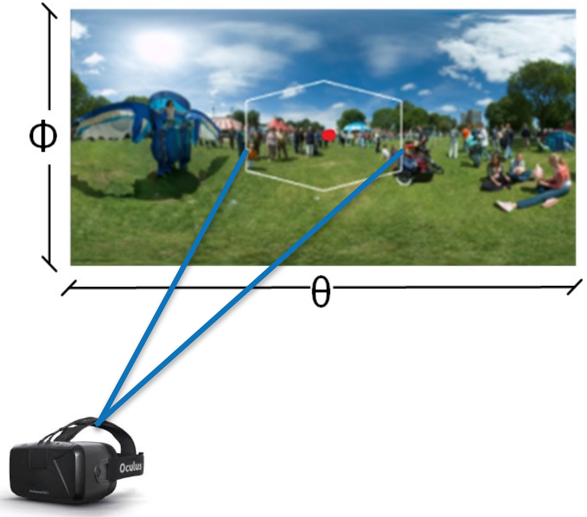
V-SENSE

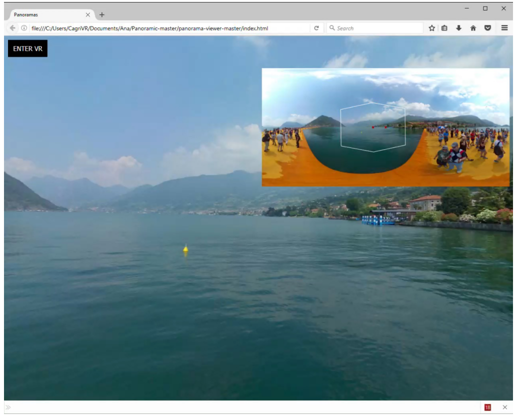
Omnidirectional Video – 3DoF

Professor Aljosa Smolic
SFI Research Professor of Creative Technologies

Omnidirectional Images (ODIs) in VR






Panoramas
file:///C:/Users/CagriVR/Documents/Ana/Panoramic-master/panorama-viewer-master/index.html
Search
ENTER VR

- Spherical captured images
- ODIs are stored in a planar representation e.g., **equirectangular**, cylindrical, cubic
- Projected back into a 3D geometry for rendering

Visual Attention

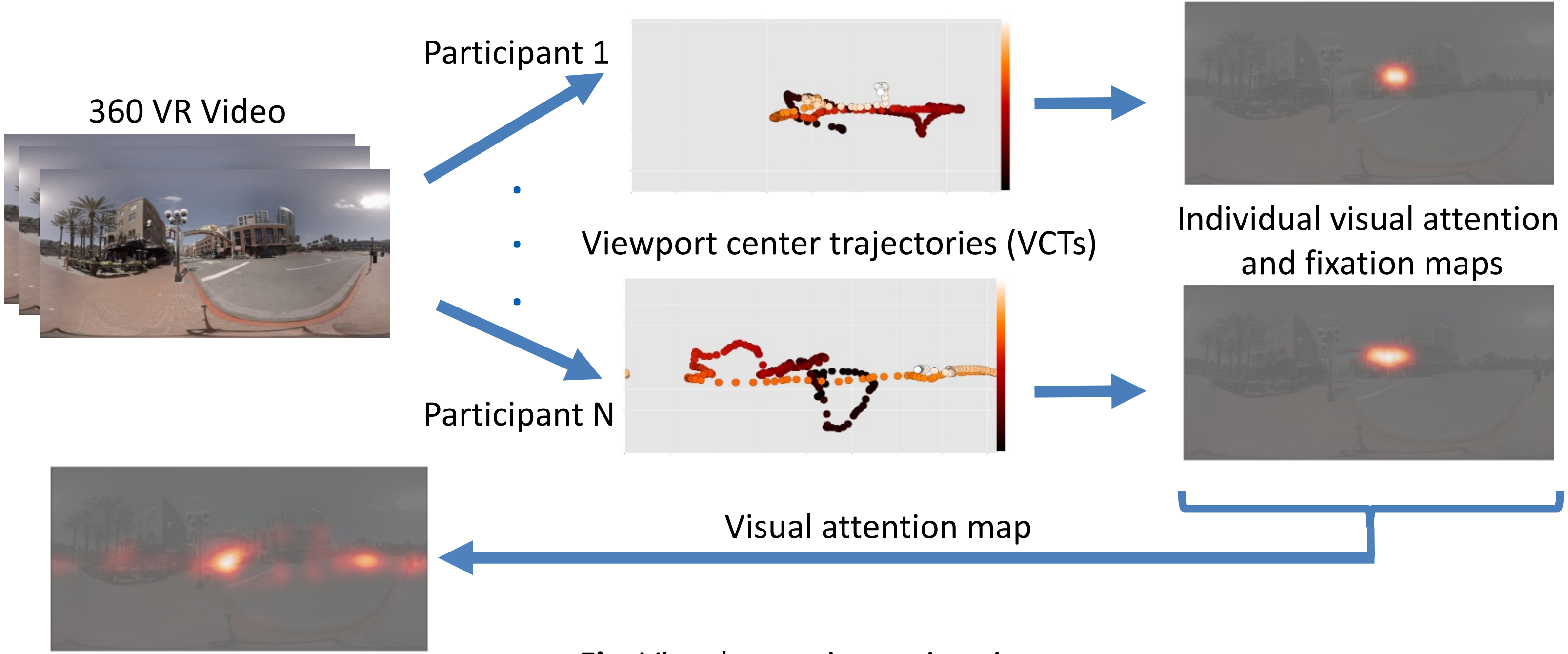

360 VR Video
Participant 1
Viewport center trajectories (VCTs)
Participant N
Individual visual attention and fixation maps
Visual attention map

Fig. Visual attention estimation.

- Abreu, Ana De; Ozcinar, Cagri; Smolic, Aljosa; “Look around you: saliency maps for omnidirectional images in VR applications,” **9th IEEE International Conference on Quality of Multimedia Experience (QoMEX), 2017.**
- Ozcinar, Cagri; Smolic, Aljosa; “Visual Attention in Omnidirectional Video for Virtual Reality Applications,” **10th IEEE International Conference on Quality of Multimedia Experience (QoMEX), 2018**.

Visual Attention

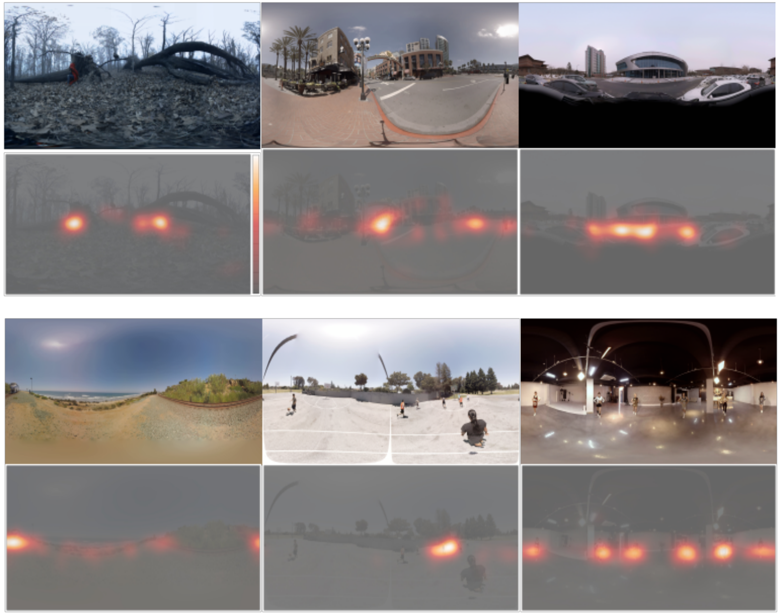
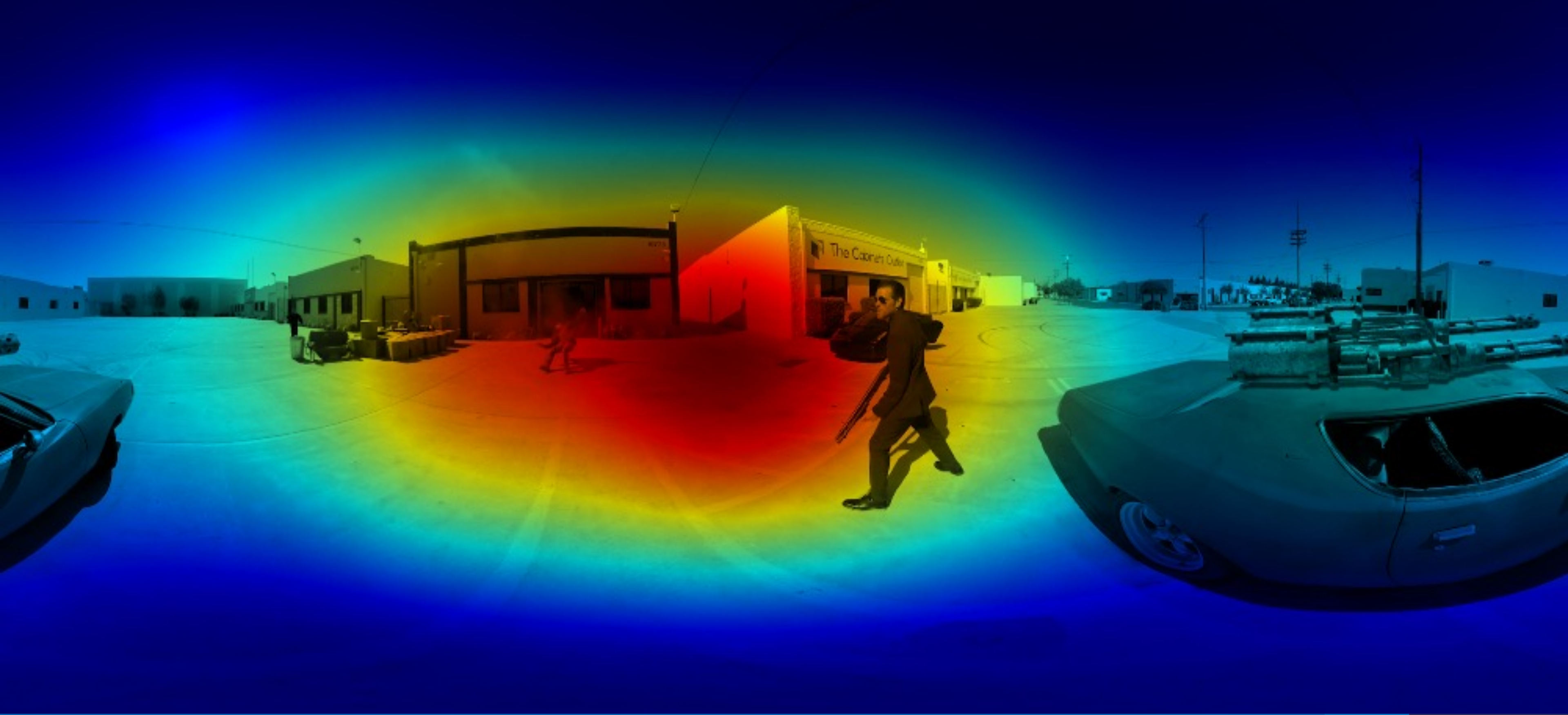


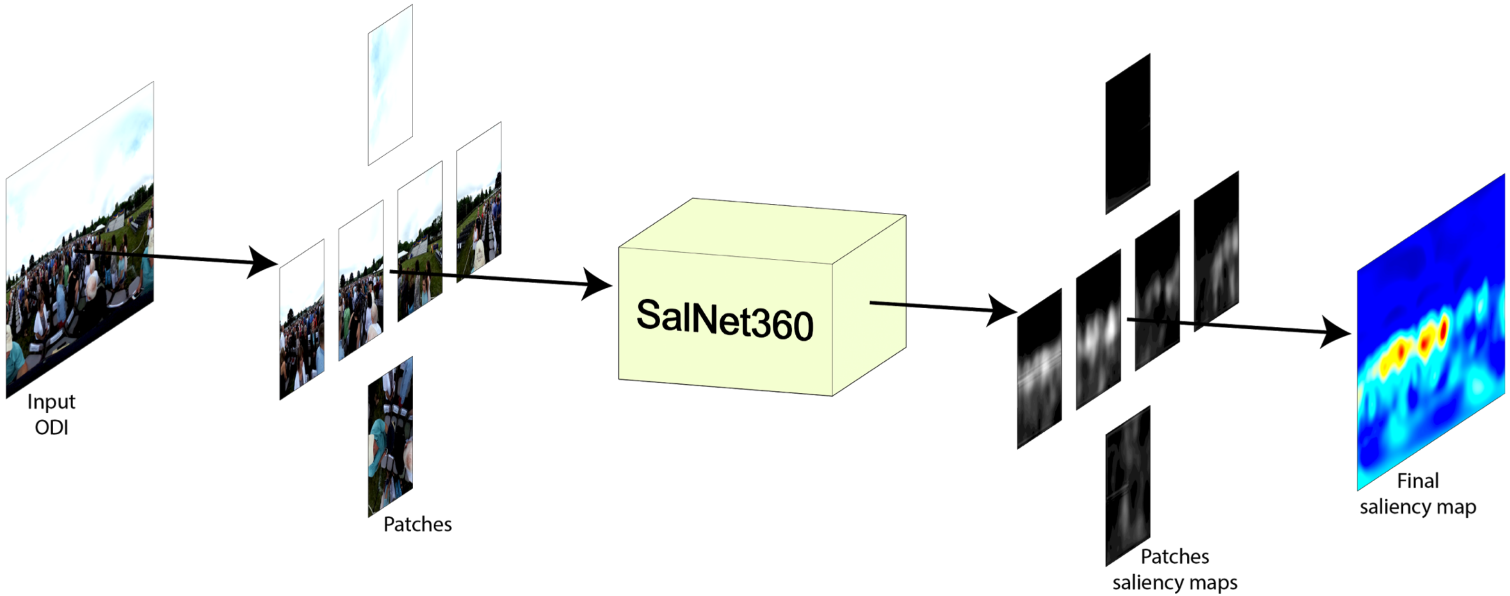
Fig. A sample thumbnail frame with its estimated visual attention for each ODV.

- Abreu, Ana De; Ozcinar, Cagri; Smolic, Aljosa; “Look around you: saliency maps for omnidirectional images in VR applications,” **9th IEEE International Conference on Quality of Multimedia Experience (QoMEX), 2017.**
- Ozcinar, Cagri; Smolic, Aljosa; “Visual Attention in Omnidirectional Video for Virtual Reality Applications,” **10th IEEE International Conference on Quality of Multimedia Experience (QoMEX), 2018**.

Saliency Map



SalNet360


Input
ODI
Patches
SalNet360
Patches
saliency maps
Final
saliency map

Modelling of Visual Attention

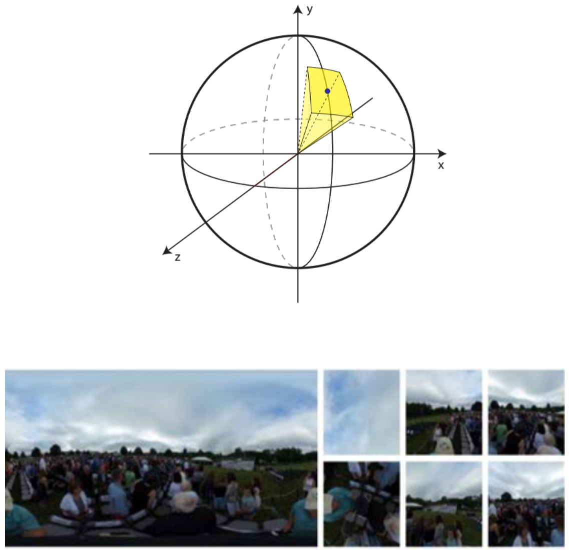


Fig. Sliding frustum used to create multiple patches.

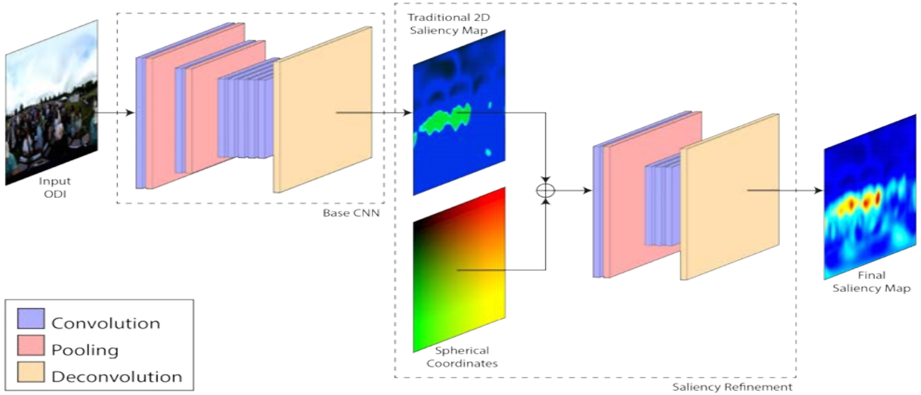


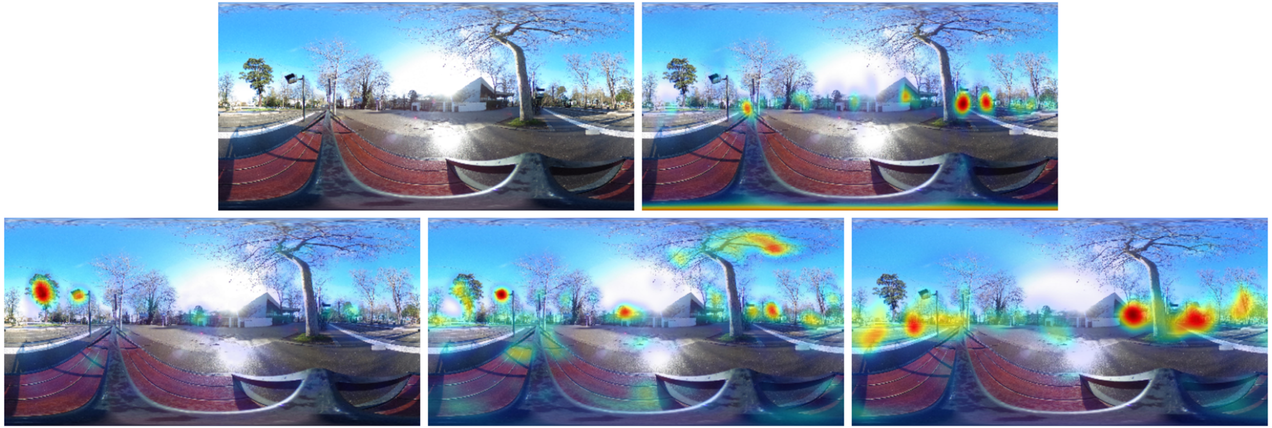
Fig. Network architecture of the SalNet360.

- Monroy, Rafael; Lutz, Sebastian; Chalasani, Tejo; Smolic, Aljosa; “SalNet360: Saliency Maps for omni-directional images with CNN,” **Signal Processing: Image Communication, 2018.**

Results

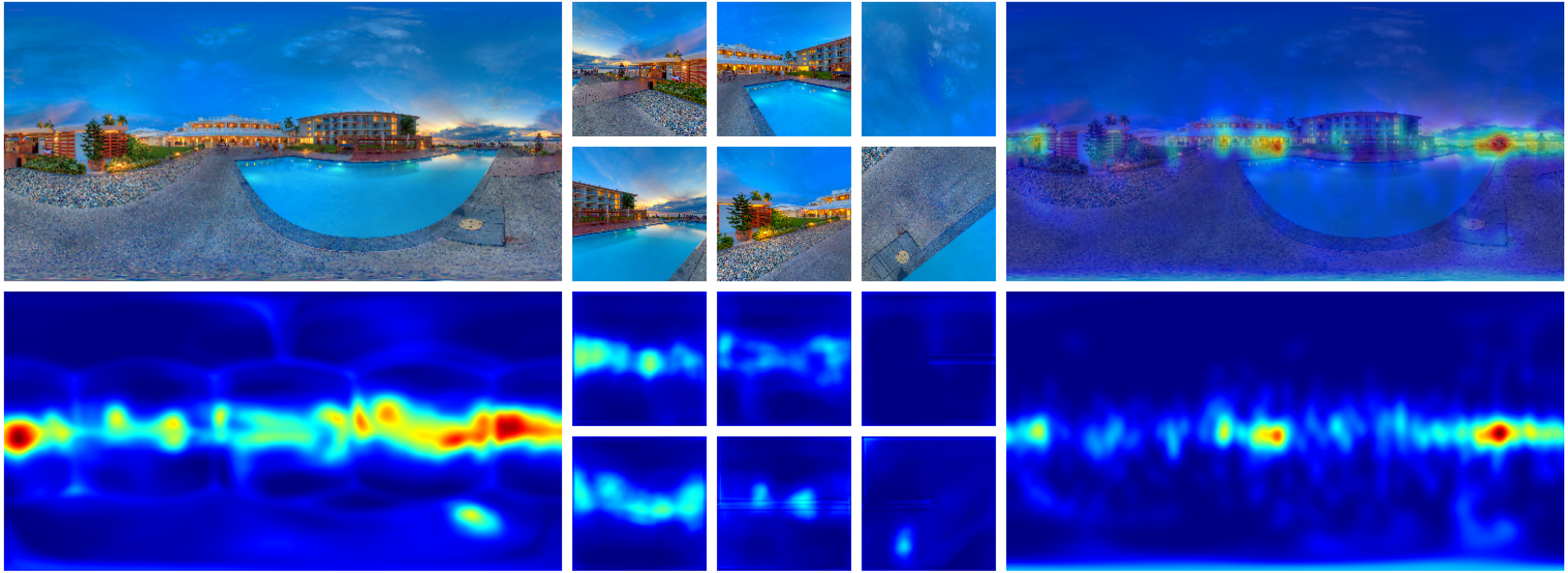
- Experiments with three scenarios
  - Base CNN, Base CNN + Patches, Base CNN + Patches + Sph. Coordinates

Top: Image and Ground Truth



Bottom: Base CNN, Base CNN + Patches, Base CNN + Patches + Sph. Coordinates

Results



Trinity College Dublin

The University of Dublin

V-SENSE

Adaptive video streaming for VR video

Viewport-aware streaming

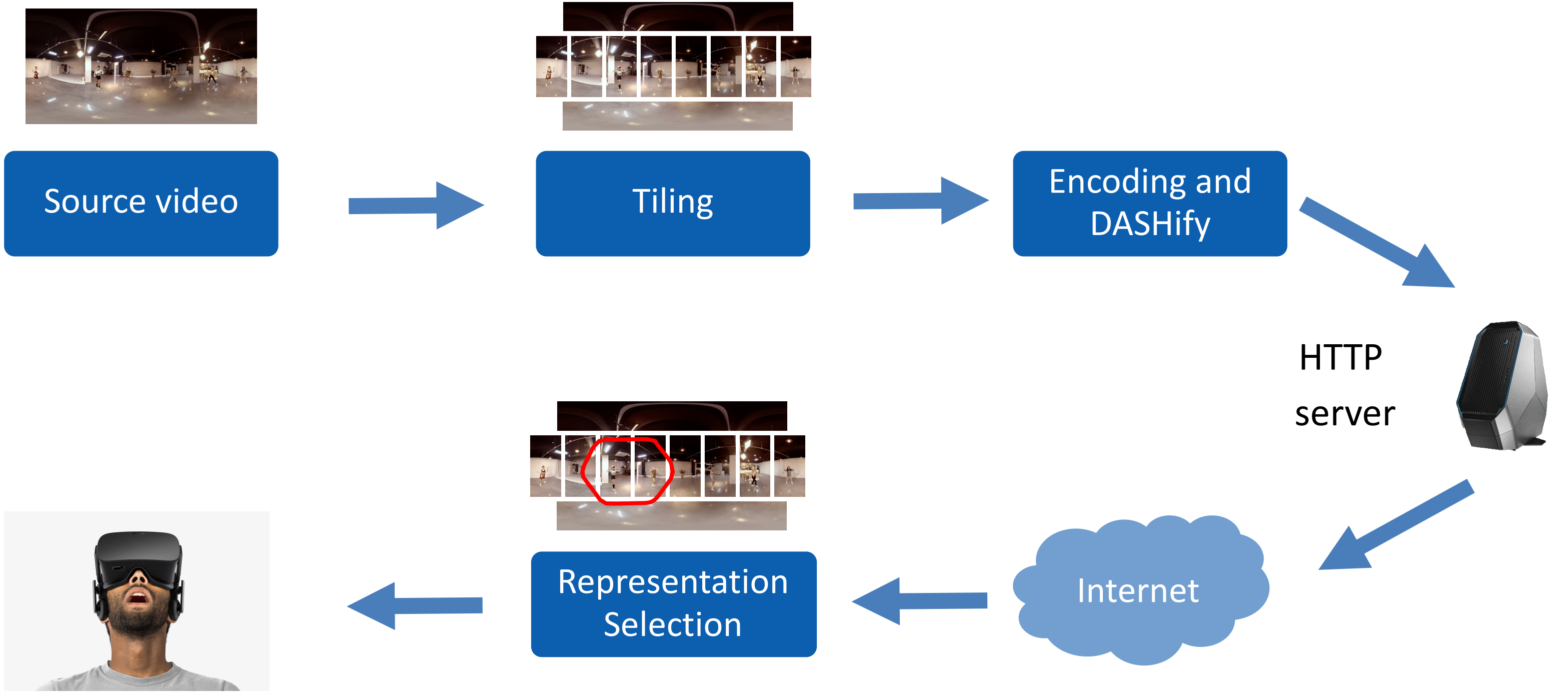
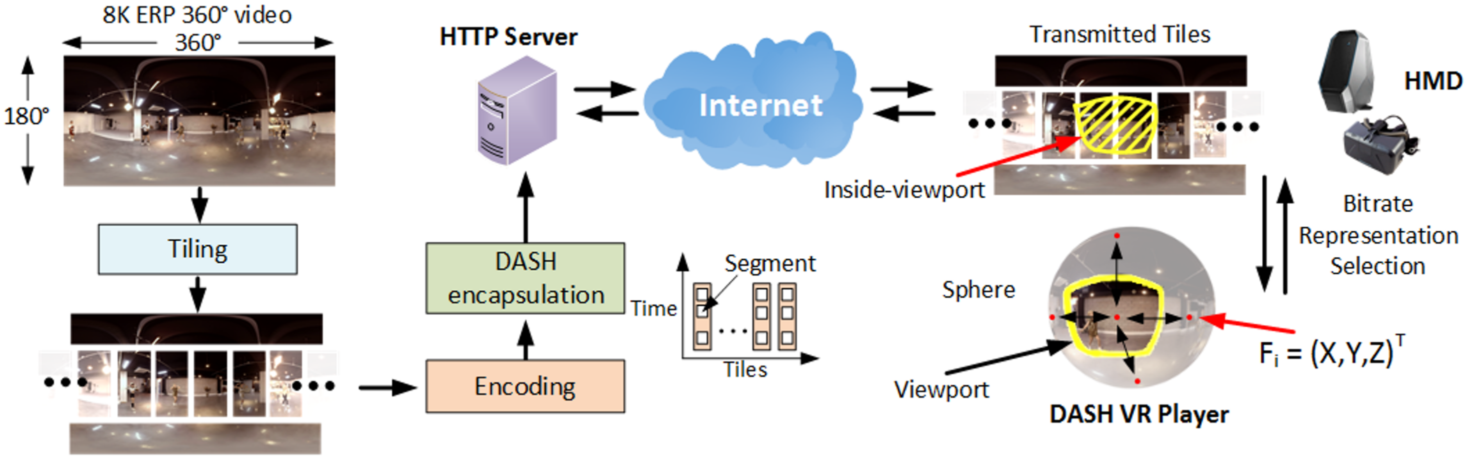


Fig: Proposed viewport-aware adaptive streaming using tiles method.

- Ozcinar, Cagri; Abreu, Ana De; Smolic, Aljosa; “Viewport-aware adaptive 360° video streaming using tiles for virtual reality,” **IEEE International Conference on Image Processing (ICIP), 2017.**

Viewport-aware streaming using tiles

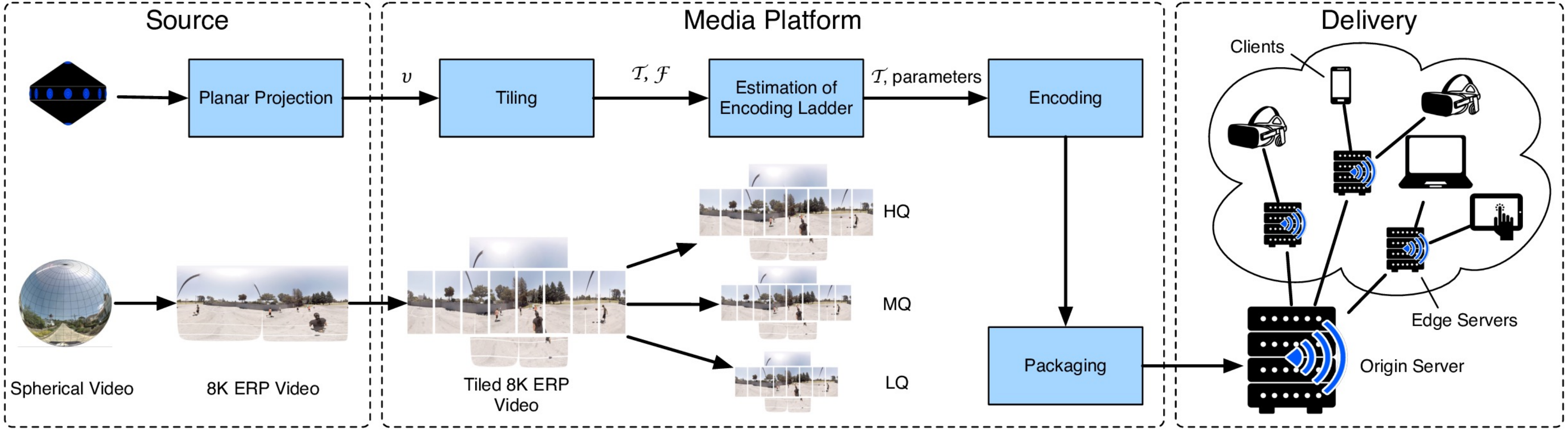


An end-to-end streaming system implementation that contains tiling, an extension of MPD, and DASH bitrate level selection in a viewport-aware way.

The proposed DASH player efficiently distributes the available bandwidth to tiles and requests the best bitrate representation for each tile in a viewport-aware manner.

C. Ozcinar, A. De Abreu, A. Smolic, “Viewport-aware adaptive 360° video streaming using tiles for virtual reality”, 2017 IEEE International Conference on Image Processing, Shanghai, China

Optimal encoding ladders in Adaptive Streaming



Cost-optimal encoding ladders is estimated in order to reduce storage capacity utilization and computational costs of CDN.
The method targets both the provider’s and client’s perspectives and introduces a technique for content-aware encoding ladder estimation of 360-degree video in adaptive streaming systems.

- Ozcinar, Cagri; Abreu, Ana De; Knorr, Sebastian; Smolic, Aljosa; “Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems,” **19th IEEE International Symposium on Multimedia (ISM), 2017.**

Visual Attention-Driven Dynamic Tiling

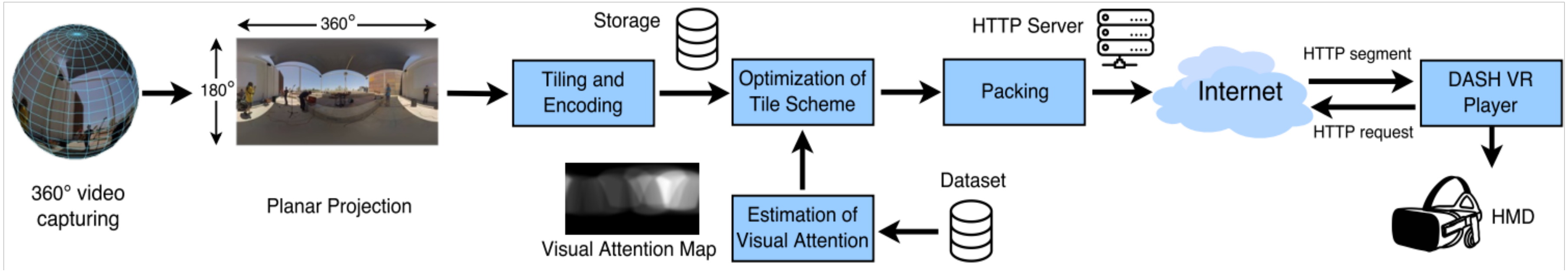


Fig: Schematic diagram of the proposed adaptive 360 VR video streaming system.

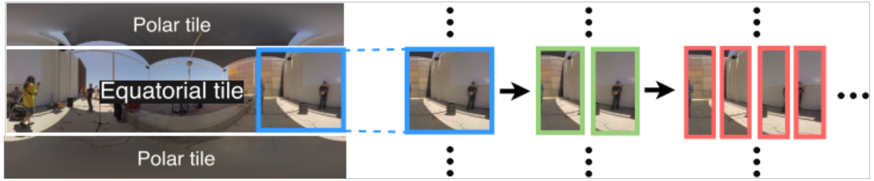


Fig: The used tiling scheme with its different structure.

- Ozcinar, Cagri; Cabrera, Julian; Smolic, Aljosa; "Omnidirectional Video Streaming Using Visual Attention-Driven Dynamic Tiling for VR," **IEEE International Conference on Visual Communications and Image Processing (VCIP), 2018**.
- Ozcinar, Cagri; Cabrera, Julian; Smolic, Aljosa; "Viewport-aware omnidirectional video streaming using visual attention and dynamic tiles," **7th European Workshop on Visual Information Processing (EUVIP), 2018**.

Trinity College Dublin
The University of Dublin

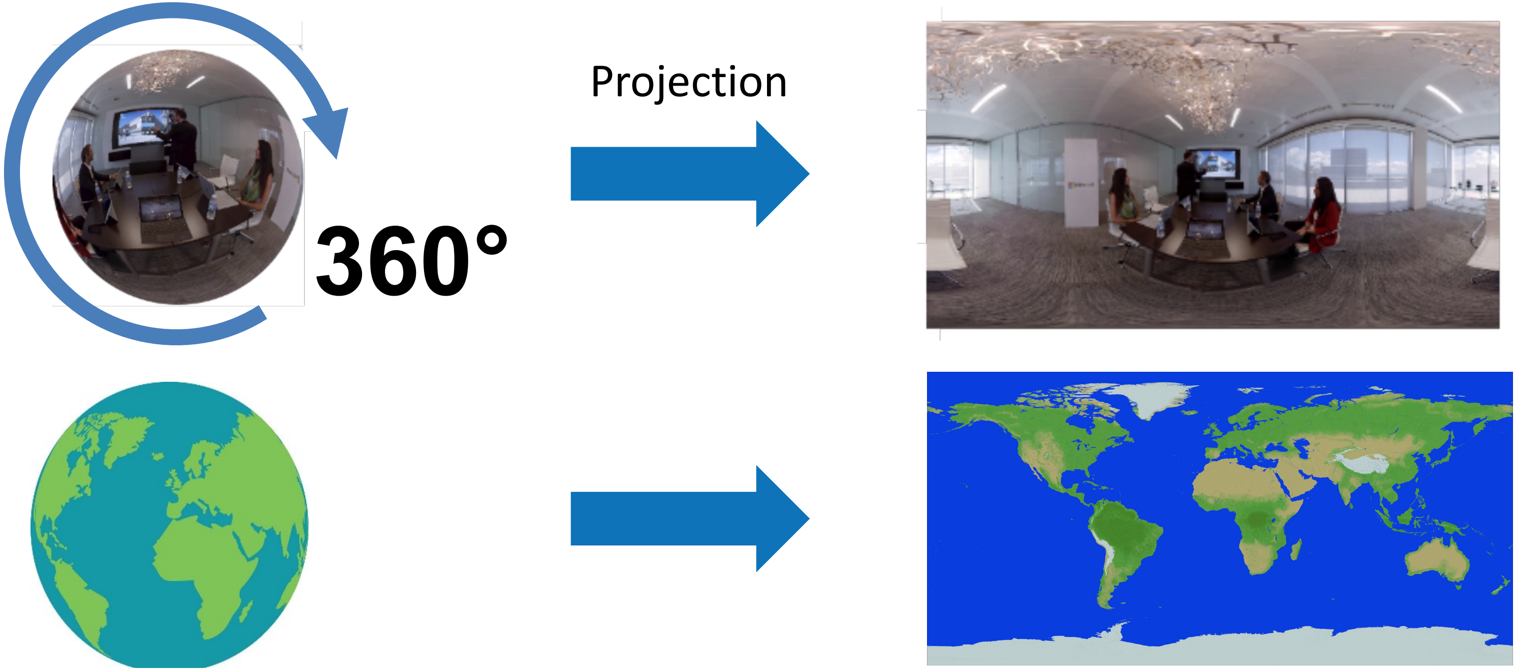
V-SENSE

VI-VA-METRIC: Omnidirectional Video Quality Assessment based on Voronoi Patches and Visual Attention

Simone Croci, Emin Zerman, and Aljosa Smolic

Unique Aspects of ODV

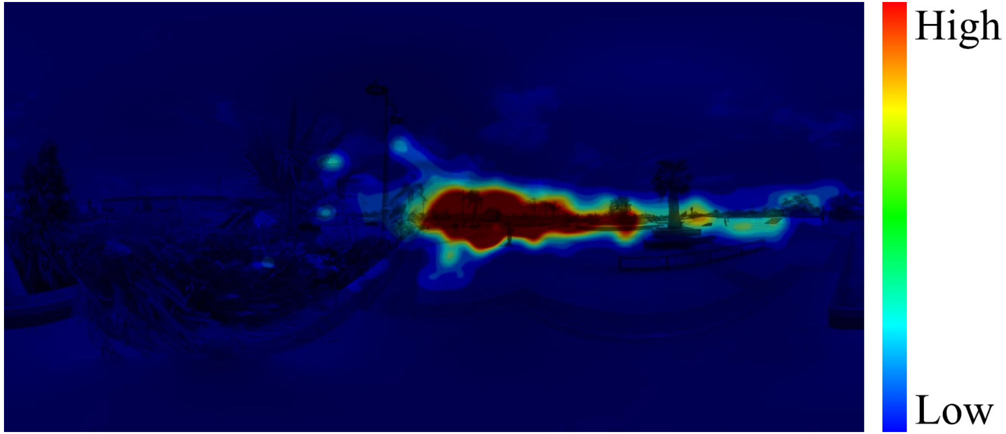
1. **Spherical nature but stored in planar representations**



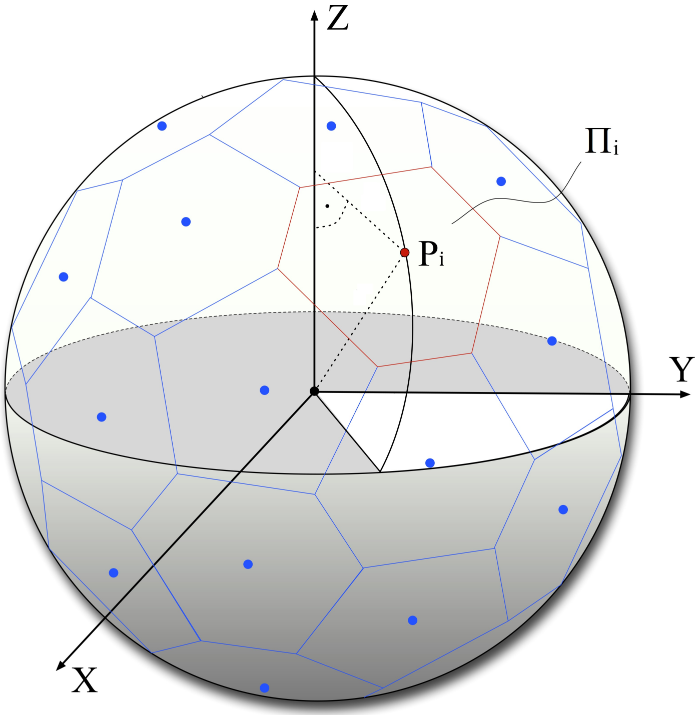
Unique Aspects of ODV

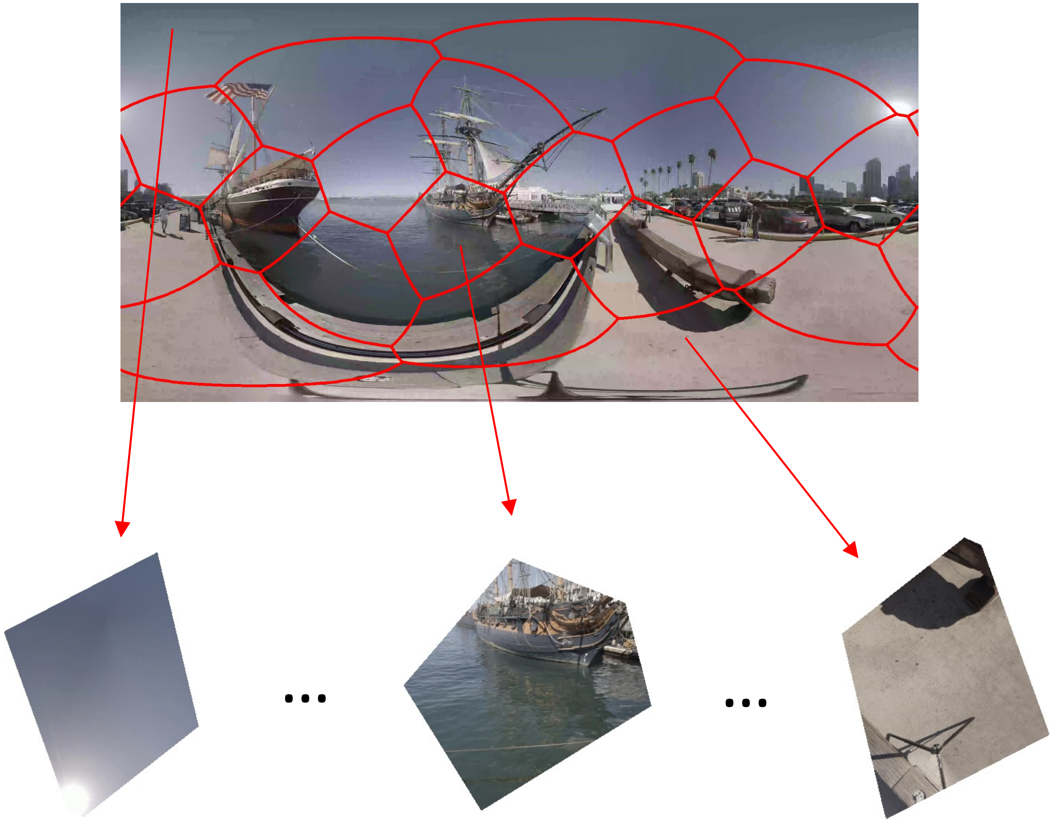
2. **Viewing characteristics: free look around, only viewport**

Visual Attention

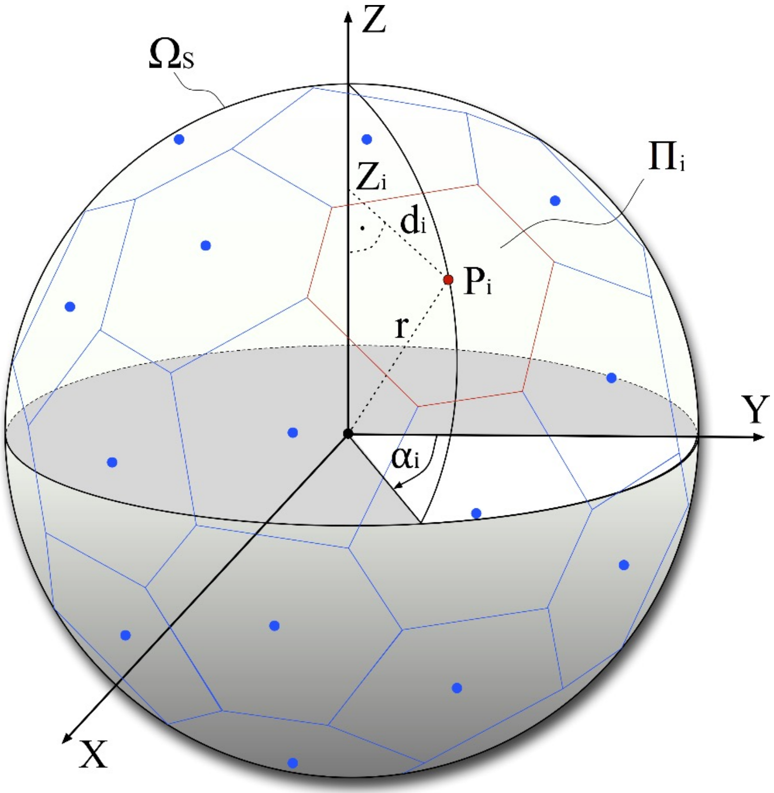


Voronoi Patch Extraction





Voronoi Patch Extraction



1) *N* evenly distributed points $\mathbf{P}_i = (X_i, Y_i, Z_i)$ with $i = 0 \dots N-1$

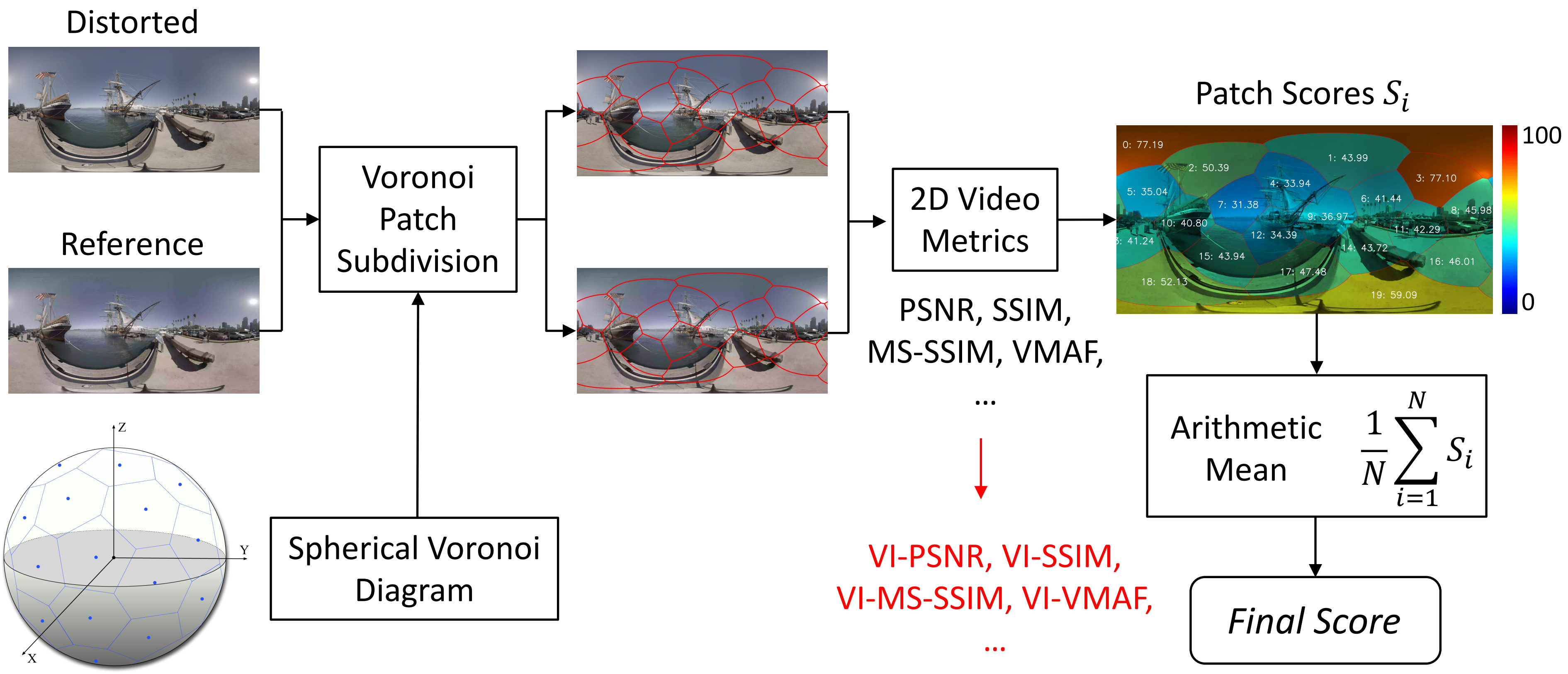
$$\begin{aligned}
\alpha_i &= i\pi \cdot \left(3 - \sqrt{5}\right) \\
Z_i &= \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2i}{N-1}\right) \\
d_i &= \sqrt{1 - Z_i^2} \\
X_i &= d_i \cdot cos(\alpha_i) \\
Y_i &= d_i \cdot sin(\alpha_i)
\end{aligned}$$

2) Spherical Voronoi Diagram => spherical patch Π_i

3) Planar patch Π'_i corresponding to the spherical patch Π_i

4) Pixels of planar patch Π'_i by sampling ODV in ERP

Voronoi-based Metrics



Distorted

Reference

Voronoi Patch Subdivision

Spherical Voronoi Diagram

2D Video Metrics

PSNR, SSIM, MS-SSIM, VMAF, …

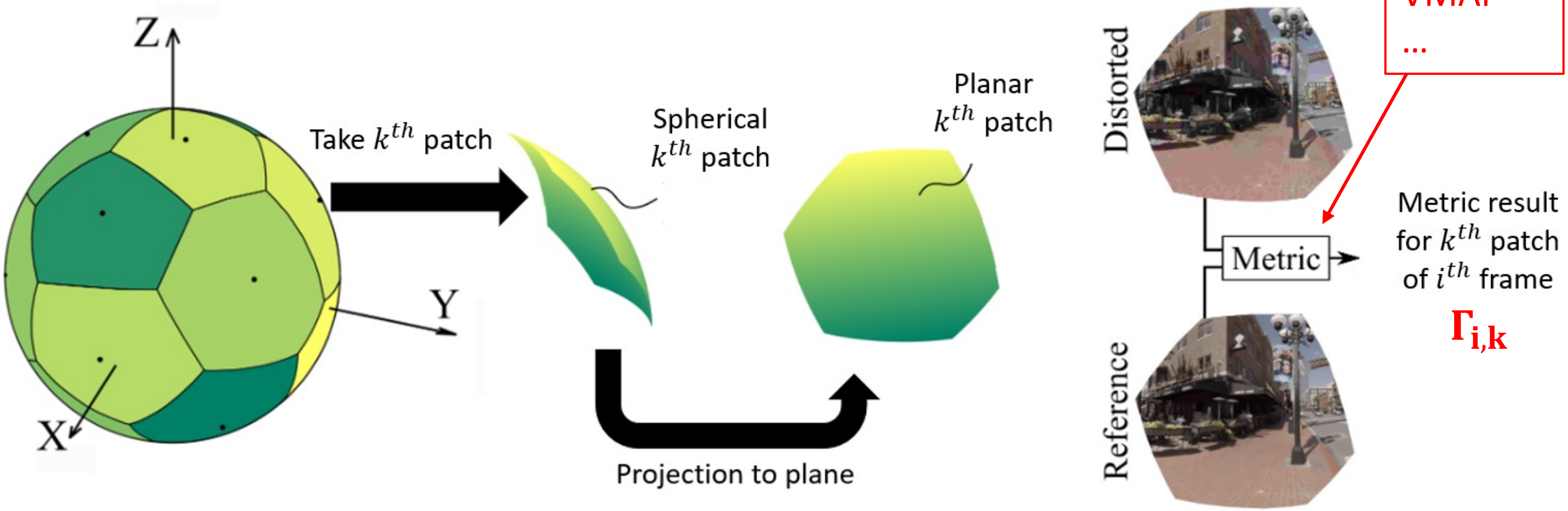
VI-PSNR, VI-SSIM, VI-MS-SSIM, VI-VMAF, …

Patch Scores S_i

Arithmetic Mean $\frac{1}{N}\sum_{i=1}^{N} S_i$

Final Score

VI-VA-METRIC Framework



VI-VA-METRIC Framework

Score of frame *i*:

$$T_i = \frac{\sum_{k=1}^{M} \Gamma_{i,k}}{M}$$

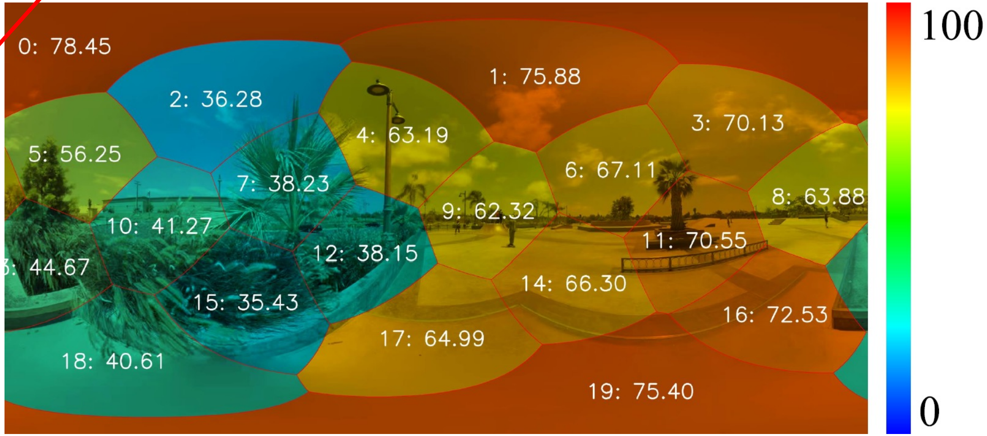
$$T_i' = \frac{\sum_{k=1}^{M} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} \nu_{i,k}}$$

$\Gamma_{i,k}$ Patch score

$\nu_{i,k}$ Visual attention weight

Score of patch *k* of frame *i:*

$\Gamma_{i,k}$



VI-VA-METRIC Framework

Score of frame *i*:

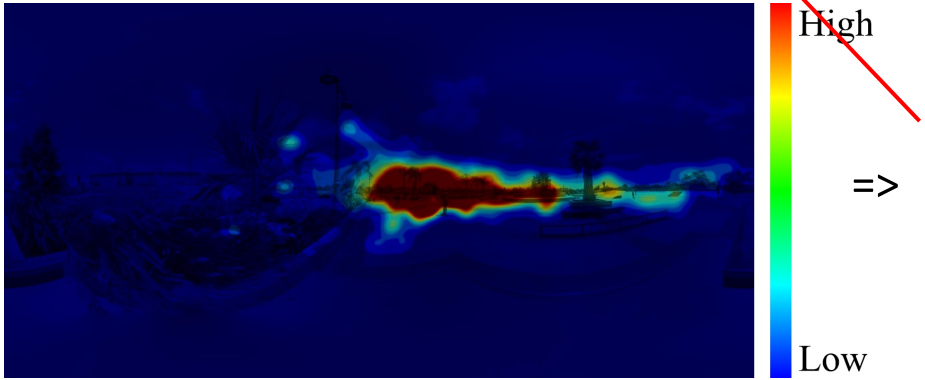
$$T_i = \frac{\sum_{k=1}^{M} \Gamma_{i,k}}{M}$$

$$T_i' = \frac{\sum_{k=1}^{M} v_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} v_{i,k}}$$

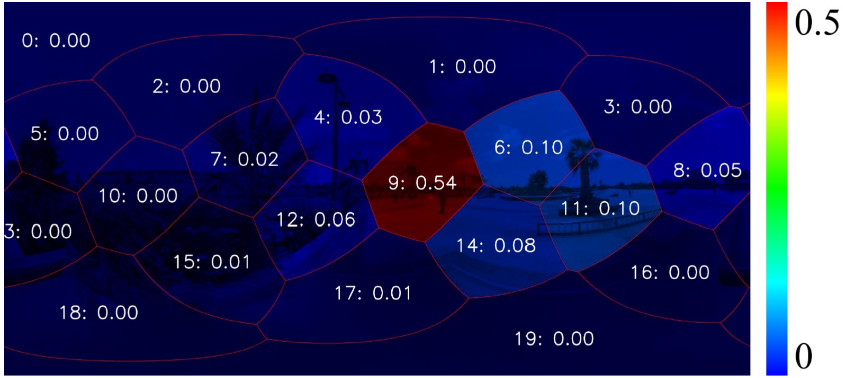
$\Gamma_{i,k}$ Patch score

$v_{i,k}$ Visual attention weight

Visual attention weight of patch *k* of frame *i:*



=> $v_{i,k}$



VI-VA-METRIC Framework

Final score from temporal pooling of frame scores

$$\text{VI-METRIC} = P(T_1, T_2, \dots, T_N)$$

$$\text{VI-VA-METRIC} = P(T'_1, T'_2, \dots, T'_N)$$

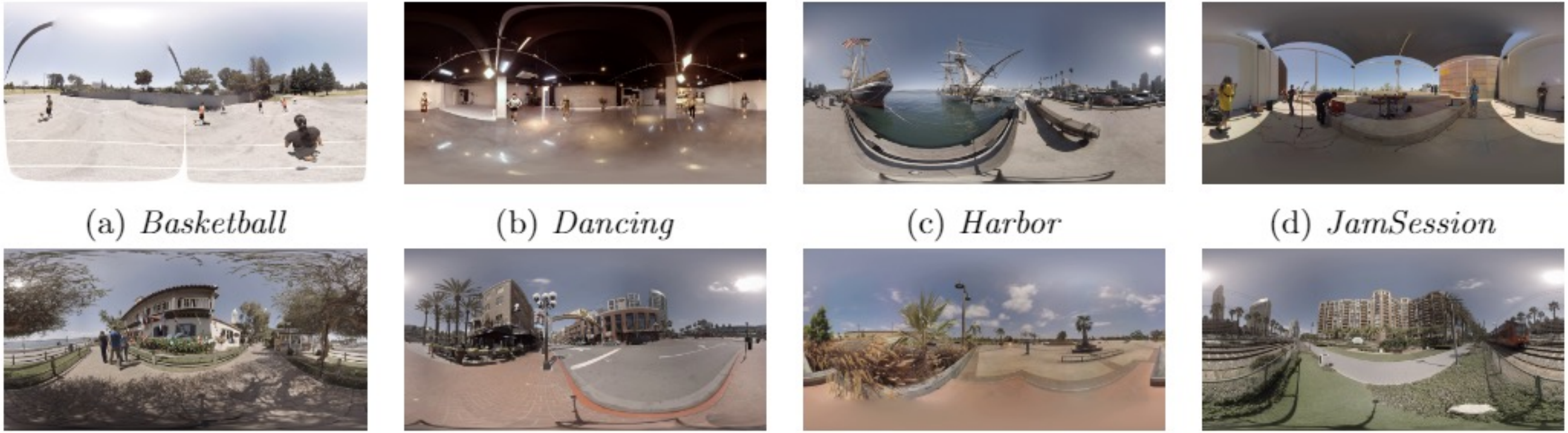
P: arithmetic mean, harmonic mean, min, median, p-th percentile, …

ODV Dataset and Subjective Experiments

- **Goal: metric evaluation**
- **ODV Dataset**
  - 8 reference and 120 distorted ODVs
  - Scaling and compression distortions
- **Subjective Experiments**
  - Subjective scores (DMOS) and visual attention data

ODV Dataset

- 8K x 4K ERP
- YUV420p
- 10 sec.



(a) *Basketball*

(b) *Dancing*

(c) *Harbor*

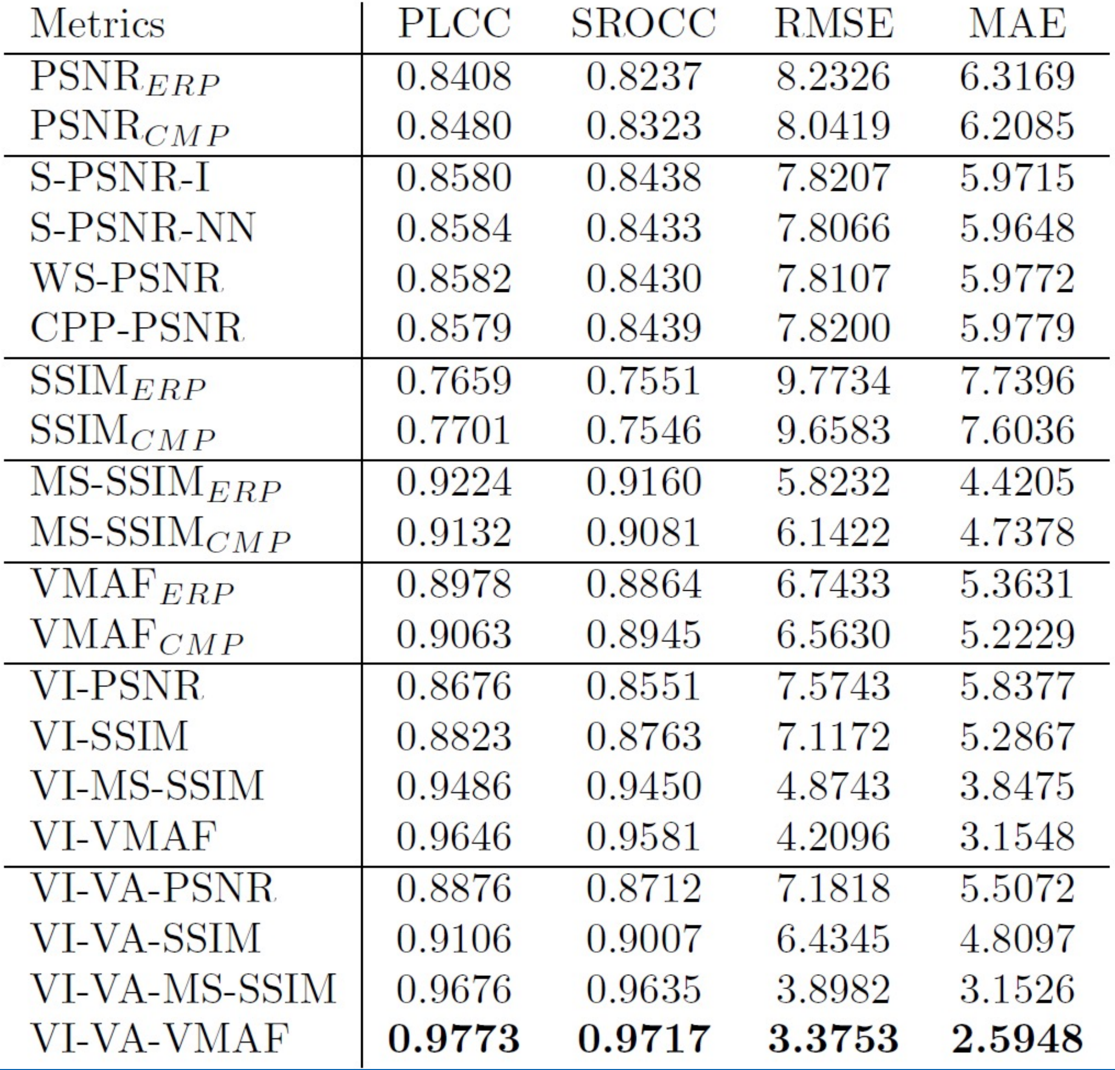
(d) *JamSession*

(e) *KiteFlite*

(f) *Gaslamp*

(g) *SkateboardTrick*

(h) *Trolley*



Metrics	PLCC	SROCC	RMSE	MAE
PSNR$_{ERP}$	0.8408	0.8237	8.2326	6.3169
PSNR$_{CMP}$	0.8480	0.8323	8.0419	6.2085
S-PSNR-I	0.8580	0.8438	7.8207	5.9715
S-PSNR-NN	0.8584	0.8433	7.8066	5.9648
WS-PSNR	0.8582	0.8430	7.8107	5.9772
CPP-PSNR	0.8579	0.8439	7.8200	5.9779
SSIM$_{ERP}$	0.7659	0.7551	9.7734	7.7396
SSIM$_{CMP}$	0.7701	0.7546	9.6583	7.6036
MS-SSIM$_{ERP}$	0.9224	0.9160	5.8232	4.4205
MS-SSIM$_{CMP}$	0.9132	0.9081	6.1422	4.7378
VMAF$_{ERP}$	0.8978	0.8864	6.7433	5.3631
VMAF$_{CMP}$	0.9063	0.8945	6.5630	5.2229
VI-PSNR	0.8676	0.8551	7.5743	5.8377
VI-SSIM	0.8823	0.8763	7.1172	5.2867
VI-MS-SSIM	0.9486	0.9450	4.8743	3.8475
VI-VMAF	0.9646	0.9581	4.2096	3.1548
VI-VA-PSNR	0.8876	0.8712	7.1818	5.5072
VI-VA-SSIM	0.9106	0.9007	6.4345	4.8097
VI-VA-MS-SSIM	0.9676	0.9635	3.8982	3.1526
VI-VA-VMAF	**0.9773**	**0.9717**	**3.3753**	**2.5948**

Findings

- **VI-METRICs better than original metrics**
  - Low projection distortion of Voronoi patches
- **VI-VA-METRICs better than VI-METRICs**
  - Visual attention is important
- **Best: VI-VA-VMAF**

Croci, Simone; Ozcinar, Cagri; Zerman, Emin; Knorr, Sebastian; Cabrera, Julian; Smolic, Aljosa
Visual Attention-Aware Quality Estimation Framework for Omnidirectional Video using Spherical Voronoi Diagram Journal Article
In: Springer Quality and User Experience, 2020.

ISO/IEC JTC 1/SC 29/AG 5 N00013
Draft Overview of Quality Metrics and Methodologies for Immersive Visual Media (v2)

AUDIO-VISUAL PERCEPTION OF OMNIDIRECTIONAL VIDEO FOR VIRTUAL REALITY APPLICATIONS

[1]**Fang-Yi Chao**, [2]Cagri Ozcinar, [2]Chen Wang, [2]Emin Zerman, [1]Lu Zhang,
[1]Wassim Hamidouche, [1]Olivier Deforges, [2]Aljosa Smolic

[1]Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
[2]V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland

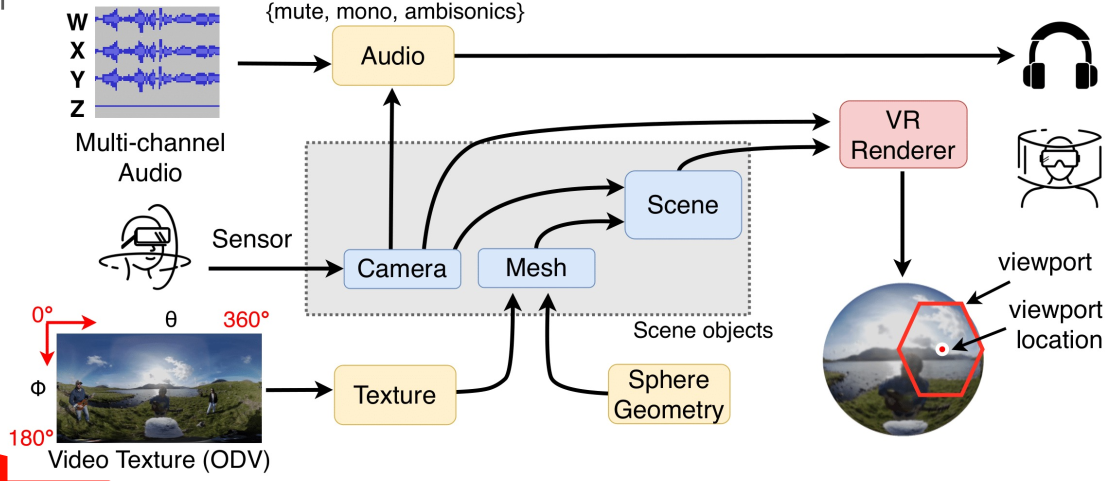
Presentation Date: 10/07/2020

Open source: https://v-sense.scss.tcd.ie/research/360audiovisualperception/

SUBJECTIVE EXPERIMENTS

- **Testbed: Records participants’ viewport center tragectories (VCTs) in mute, mono, ambisonics**

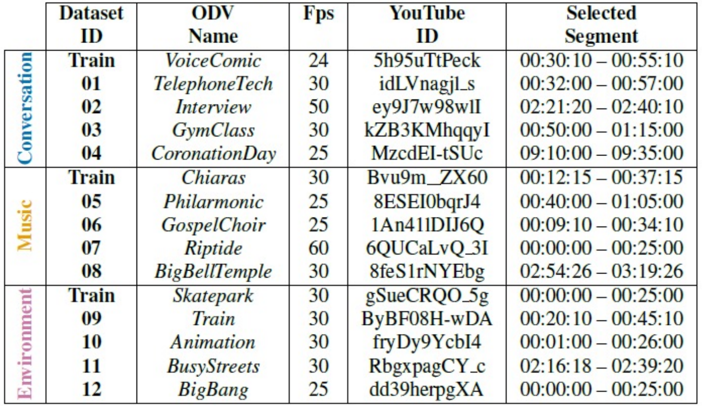
  1. Materials: Monoscopic ODV, First order ambisonics (4 channel: WXYZ)
  2. Equipements: Oculus Rift, Bose QuietComfort noise-canceling headphones
  3. Softwares: three JavaScript libraries:
     a) (1) three.js, (2) WebXR: enable the creation of fully immersive ODV experiences in a web browser
     b) (3) JSAmbisonics: nonindividual head-related transfer functions based on spatially oriented format for acoustics
  4. Each



SUBJECTIVE EXPERIMENTS

Materials: 3 training ODVs and 12 test ODVs in 3 audio modalities

Table 1: Description of the ODVs in our dataset.



	Dataset ID	ODV Name	Fps	YouTube ID	Selected Segment
Conversation	**Train**	*VoiceComic*	24	5h95uTtPeck	00:30:10 – 00:55:10
	01	*TelephoneTech*	30	idLVnagjl_s	00:32:00 – 00:57:00
	02	*Interview*	50	ey9J7w98wlI	02:21:20 – 02:40:10
	03	*GymClass*	30	kZB3KMhqqyI	00:50:00 – 01:15:00
	04	*CoronationDay*	25	MzcdEI-tSUc	09:10:00 – 09:35:00
Music	**Train**	*Chiaras*	30	Bvu9m_ZX60	00:12:15 – 00:37:15
	05	*Philarmonic*	25	8ESEI0bqrJ4	00:40:00 – 01:05:00
	06	*GospelChoir*	25	1An41lDIJ6Q	00:09:10 – 00:34:10
	07	*Riptide*	60	6QUCaLvQ_3I	00:00:00 – 00:25:00
	08	*BigBellTemple*	30	8feS1rNYEbg	02:54:26 – 03:19:26
Environment	**Train**	*Skatepark*	30	gSueCRQO_5g	00:00:00 – 00:25:00
	09	*Train*	30	ByBF08H-wDA	00:20:10 – 00:45:10
	10	*Animation*	30	fryDy9YcbI4	00:01:00 – 00:26:00
	11	*BusyStreets*	30	RbgxpagCY_c	02:16:18 – 02:39:20
	12	*BigBang*	25	dd39herpgXA	00:00:00 – 00:25:00

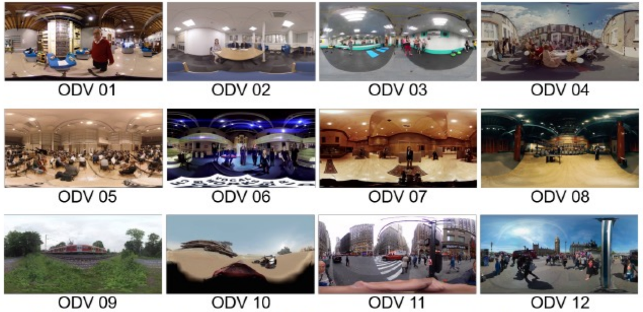


Fig. 2: Examples for each ODV used in subjective experiments. Rows from top to bottom respectively belong to category: Conversation; Music; Environment.

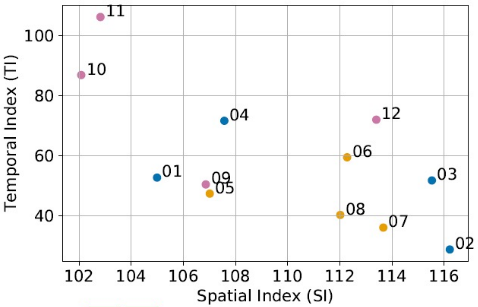


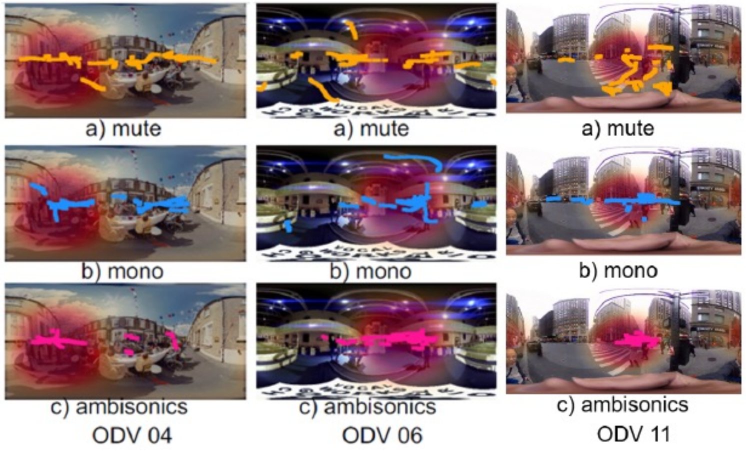
Fig. 3: SI and TI [19, 20] for each ODV used in subjective experiments. Each color visualizes each category: Conversation; Music; Environment.

Participants:

- **45 participants were recruited in this subjective experiment.**
- **Each ODV with each modality was viewed by 15 participants, and each participant viewed each ODV only once.**
- These participants were aged between 21 and 40 years old with an average of 27.3 years old
- 16 of them were female. 8 of them were familiar with VR, and the others were naive viewers.
- All were screened and reported normal or corrected-to-normal visual and audio acuity,
- 24 participants wore glasses during the experiment.

USER BEHAVIOR ANALYSIS

- **Analysis1: Do audio source locations attract attention of users?**
  - Normalized Scanpath Saliency (NSS) of fixations falling in sound source area
  - A higher NSS score indicates more fixations are attracted to areas of audio source locations with audio energy map (AEM)
  - negative NSS scores indicate most fixations are not corresponding to areas of audio source locations.


a) mute a) mute a) mute

b) mono b) mono b) mono

c) ambisonics c) ambisonics c) ambisonics

ODV 04 ODV 06 ODV 11

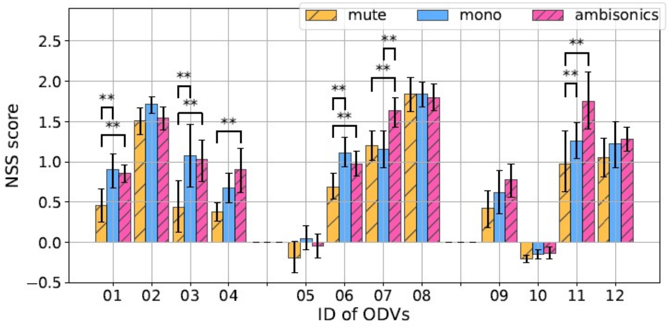
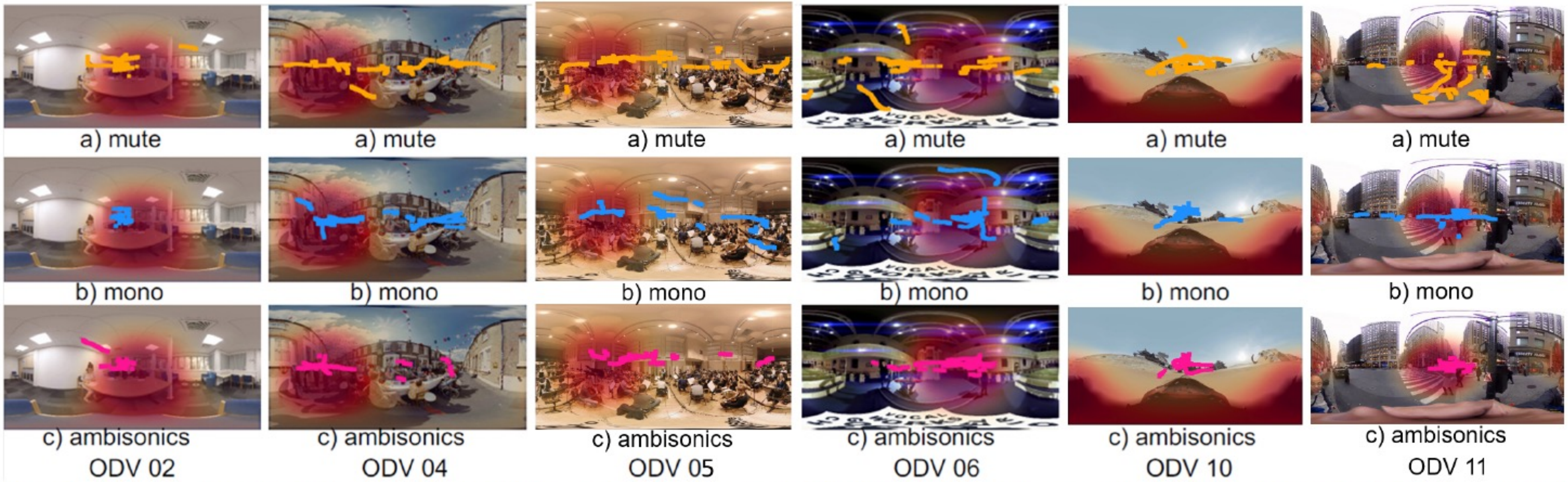


Fig. 4: Mean and 95% confidence interval of Normalized Scanpath Saliency (NSS) of fixations falling in sound sources areas under three audio modalities. ** marks statistically significant difference (SSD) between two modalities.

- Fig. 4 shows that users may tend to follow audio stimuli (especially human voice) in categories conversation and music while they tend to look around in general regardless of the background sound in category environment.

USER BEHAVIOR ANALYSIS

- **Analysis3: Does sound affect observers' navigation?**
  - Fig.6 shows fixation distributions of all observers overlaying on AEM and ODV frames in a second



a) mute
b) mono
c) ambisonics
ODV 02

a) mute
b) mono
c) ambisonics
ODV 04

a) mute
b) mono
c) ambisonics
ODV 05

a) mute
b) mono
c) ambisonics
ODV 06

a) mute
b) mono
c) ambisonics
ODV 10

a) mute
b) mono
c) ambisonics
ODV 11

Fig. 6: A sample thumbnail frame with its AEM and fixations for each ODV, where the red represents AEM and the orange, blue, and pink denotes fixations recorded under none, mono, and ambisonics modality, respectively. A frame for each ODV ID from left to right: *02*, *04*, *06*, *08*, *09*, and *10*.

- In most of the cases as shown in Fig. 6, the distribution of fixations for the ODVs with ambisonics modality is more concentrated.

TOWARDS AUDIO-VISUAL SALIENCY PREDICTION FOR OMNIDIRECTIONAL VIDEO WITH SPATIAL AUDIO

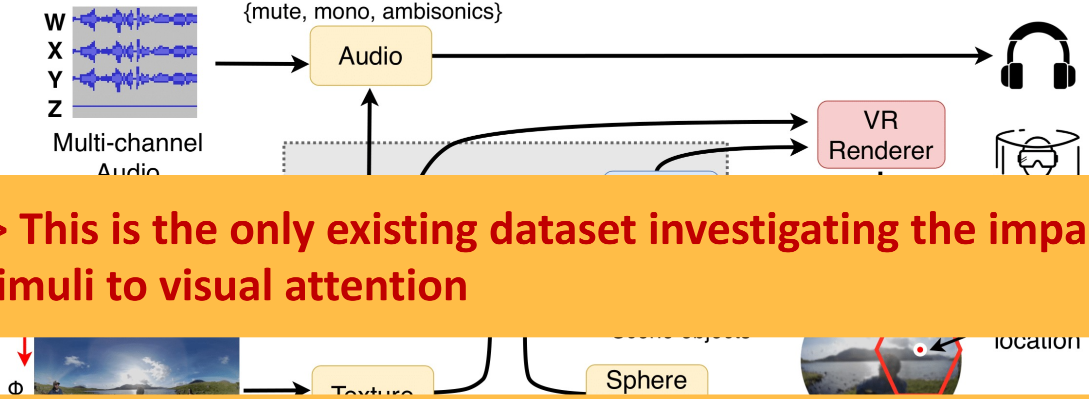
[1]Fang-Yi Chao, [2]Cagri Ozcinar, [1]Lu Zhang, [1]Wassim Hamidouche, [1]Olivier Deforges, [2]Aljosa Smolic

[1]Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
[2]V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland

Presented by Fang-Yi CHAO, Date: 03/12/2020 on VCIP

Code available: https://github.com/FannyChao/AVS360-audiovisual-saliency-360

RELATED WORKS



=> This is the only existing dataset investigating the impact of audio stimuli to visual attention

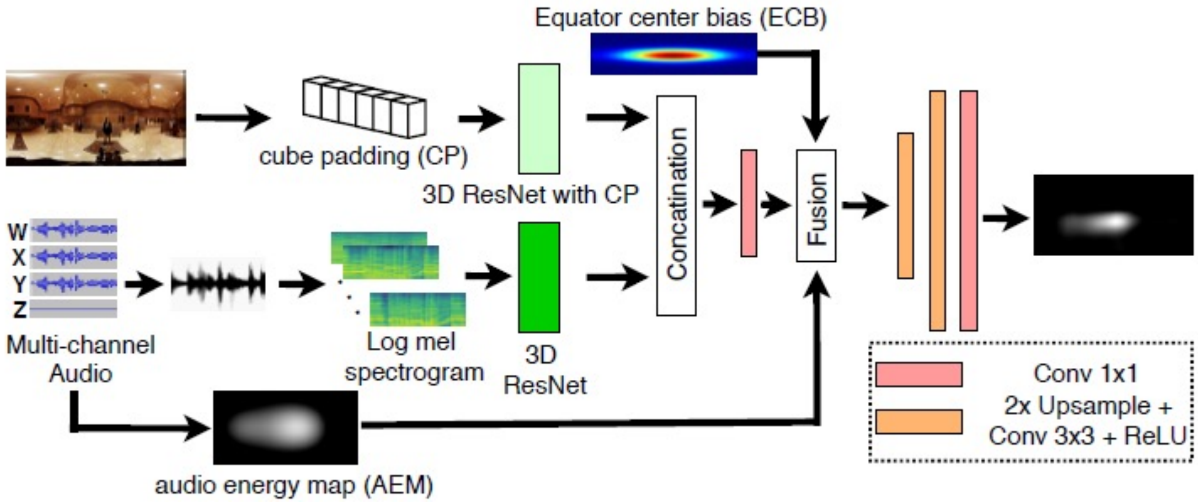
=> We propose the first audio-visual saliency model AVS360 to consider the impact of spatial audio

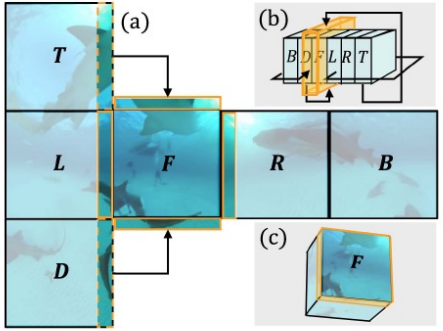
- Chao et al. [1] proposed **360AV-HM** dataset recording visual attention in mute, mono, ambisonic modalities
- They show that different audio-visual contents (i.e., conversation, music, and environment) of ODVs and different audio modalities (i.e., mute, mono, and ambisonics) have a different interactive effect on human visual saliency.

[1] F. Chao et al., "Audio-visual perception of omnidirectional video for virtual reality applications," In 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), July 2020.

AVS360

- **Network architecture**



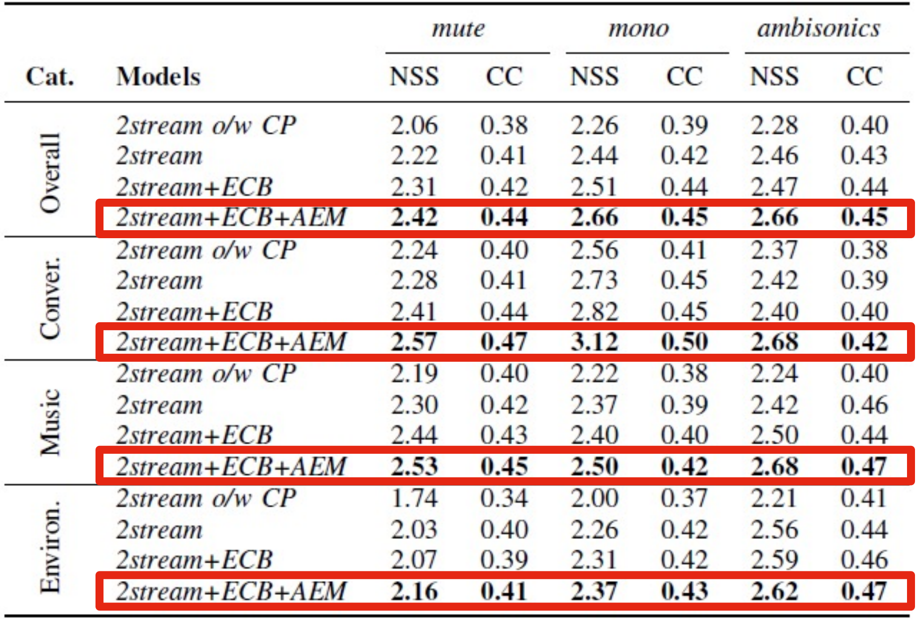


Cube padding [2]

[2] H.-T. Cheng et al., “Cube padding for weakly-supervised saliency prediction in 360 videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

RESULTS

- Contribution of each component in our model in all categories

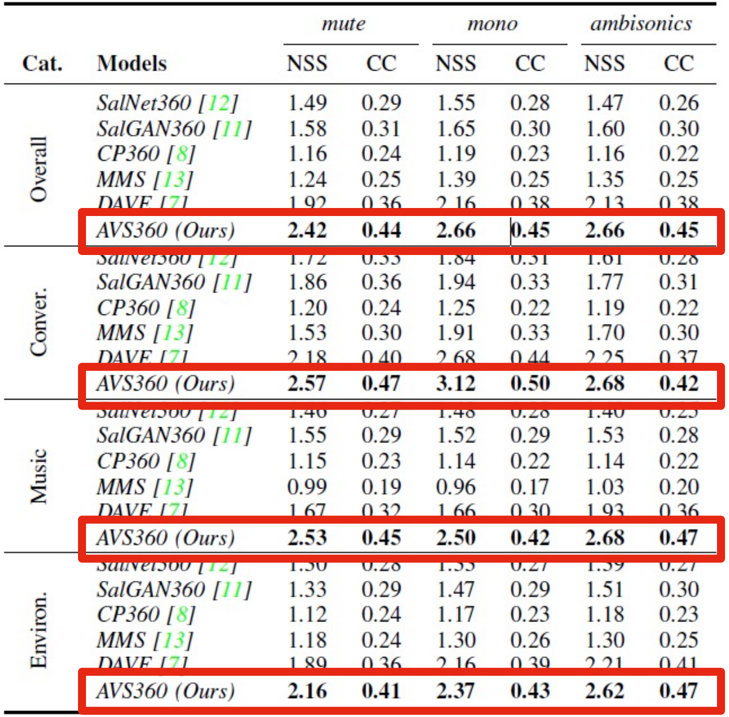


		mute		*mono*		*ambisonics*	
Cat.	**Models**	NSS	CC	NSS	CC	NSS	CC
Overall	*2stream o/w CP*	2.06	0.38	2.26	0.39	2.28	0.40
	2stream	2.22	0.41	2.44	0.42	2.46	0.43
	2stream+ECB	2.31	0.42	2.51	0.44	2.47	0.44
	2stream+ECB+AEM	**2.42**	**0.44**	**2.66**	**0.45**	**2.66**	**0.45**
Conver.	*2stream o/w CP*	2.24	0.40	2.56	0.41	2.37	0.38
	2stream	2.28	0.41	2.73	0.45	2.42	0.39
	2stream+ECB	2.41	0.44	2.82	0.45	2.40	0.40
	2stream+ECB+AEM	**2.57**	**0.47**	**3.12**	**0.50**	**2.68**	**0.42**
Music	*2stream o/w CP*	2.19	0.40	2.22	0.38	2.24	0.40
	2stream	2.30	0.42	2.37	0.39	2.42	0.46
	2stream+ECB	2.44	0.43	2.40	0.40	2.50	0.44
	2stream+ECB+AEM	**2.53**	**0.45**	**2.50**	**0.42**	**2.68**	**0.47**
Environ.	*2stream o/w CP*	1.74	0.34	2.00	0.37	2.21	0.41
	2stream	2.03	0.40	2.26	0.42	2.56	0.44
	2stream+ECB	2.07	0.39	2.31	0.42	2.59	0.46
	2stream+ECB+AEM	**2.16**	**0.41**	**2.37**	**0.43**	**2.62**	**0.47**

Mean values for saliency prediction accuracy of each component in our AVS360 model evaluated with the 360AV-HM dataset. (best in bold in each content category and audio modality)).

RESULTS

- Comparison to the state of the arts



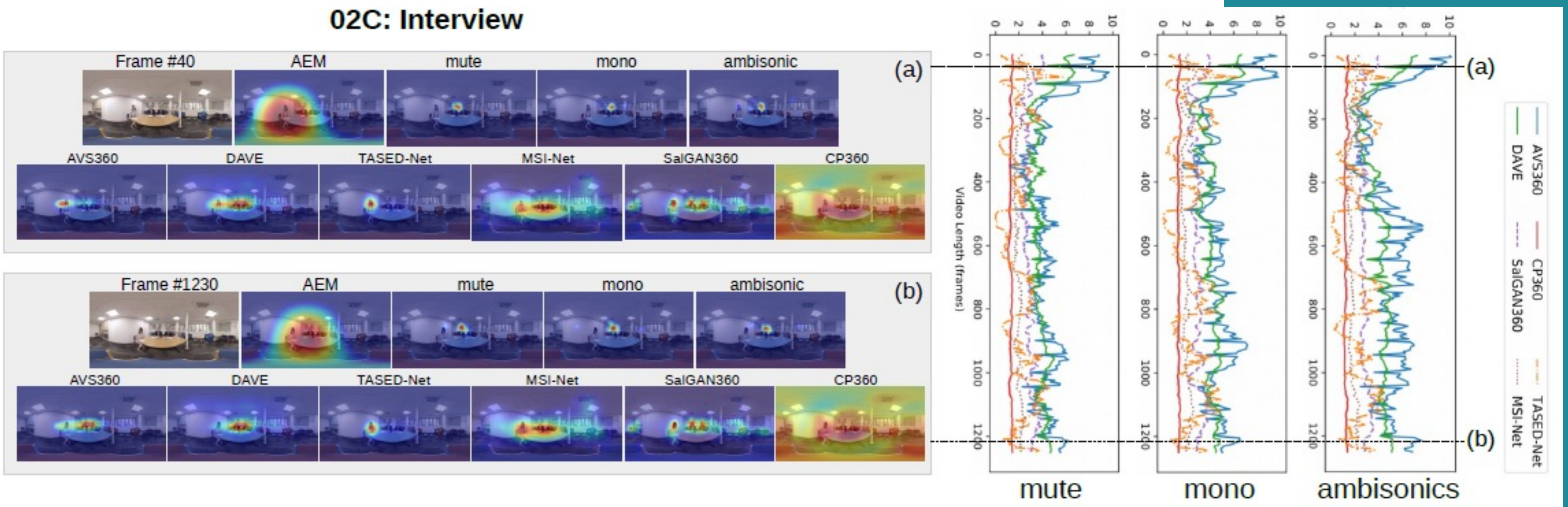
		mute		*mono*		*ambisonics*	
Cat.	**Models**	NSS	CC	NSS	CC	NSS	CC
Overall	*SalNet360 [12]*	1.49	0.29	1.55	0.28	1.47	0.26
	SalGAN360 [11]	1.58	0.31	1.65	0.30	1.60	0.30
	CP360 [8]	1.16	0.24	1.19	0.23	1.16	0.22
	MMS [13]	1.24	0.25	1.39	0.25	1.35	0.25
	DAVE [7]	1.92	0.36	2.16	0.38	2.13	0.38
	AVS360 (Ours)	**2.42**	**0.44**	**2.66**	**0.45**	**2.66**	**0.45**
Conver.	*SalNet360 [12]*	1.72	0.33	1.84	0.31	1.61	0.28
	SalGAN360 [11]	1.86	0.36	1.94	0.33	1.77	0.31
	CP360 [8]	1.20	0.24	1.25	0.22	1.19	0.22
	MMS [13]	1.53	0.30	1.91	0.33	1.70	0.30
	DAVE [7]	2.18	0.40	2.68	0.44	2.25	0.37
	AVS360 (Ours)	**2.57**	**0.47**	**3.12**	**0.50**	**2.68**	**0.42**
Music	*SalNet360 [12]*	1.46	0.27	1.48	0.28	1.46	0.25
	SalGAN360 [11]	1.55	0.29	1.52	0.29	1.53	0.28
	CP360 [8]	1.15	0.23	1.14	0.22	1.14	0.22
	MMS [13]	0.99	0.19	0.96	0.17	1.03	0.20
	DAVE [7]	1.67	0.32	1.66	0.30	1.93	0.36
	AVS360 (Ours)	**2.53**	**0.45**	**2.50**	**0.42**	**2.68**	**0.47**
Environ.	*SalNet360 [12]*	1.30	0.28	1.35	0.27	1.39	0.27
	SalGAN360 [11]	1.33	0.29	1.47	0.29	1.51	0.30
	CP360 [8]	1.12	0.24	1.17	0.23	1.18	0.23
	MMS [13]	1.18	0.24	1.30	0.26	1.30	0.25
	DAVE [7]	1.89	0.36	2.16	0.39	2.21	0.41
	AVS360 (Ours)	**2.16**	**0.41**	**2.37**	**0.43**	**2.62**	**0.47**

Mean values for saliency prediction accuracy of the state-of-the-art models evaluated with the dataset 360AV-HM (best in bold in each audio modality and content category).

EXAMPLES

Category Conversion:

02C: Interview

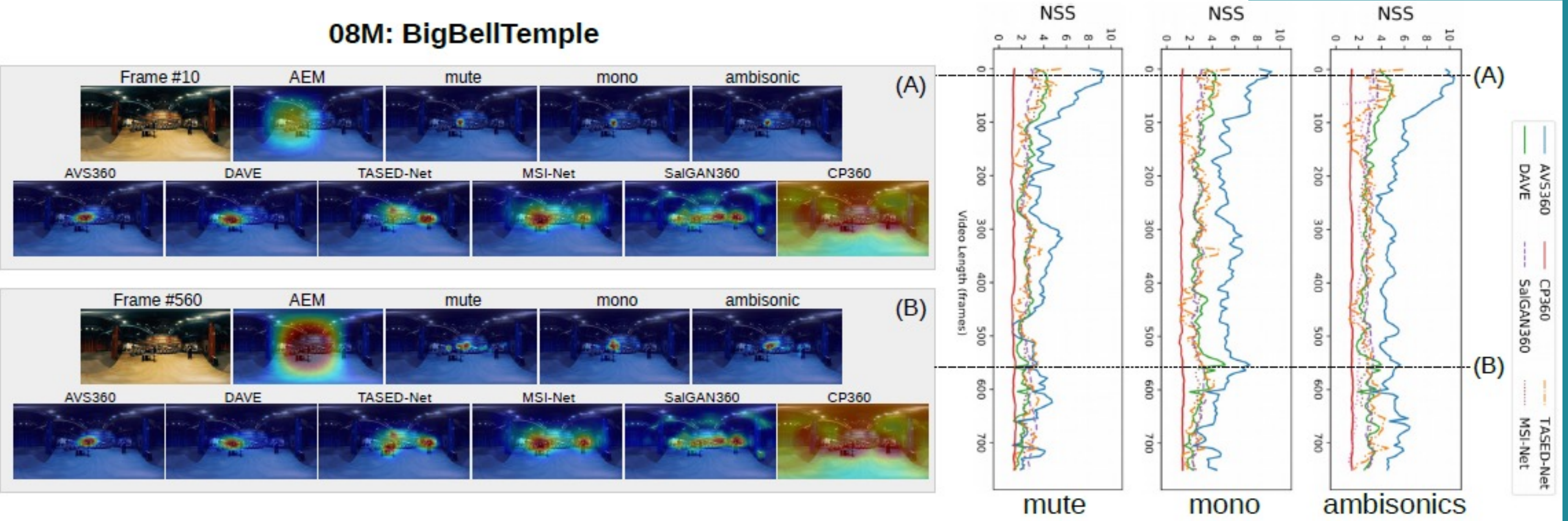


EXAMPLES

Category Music:

08M: BigBellTemple

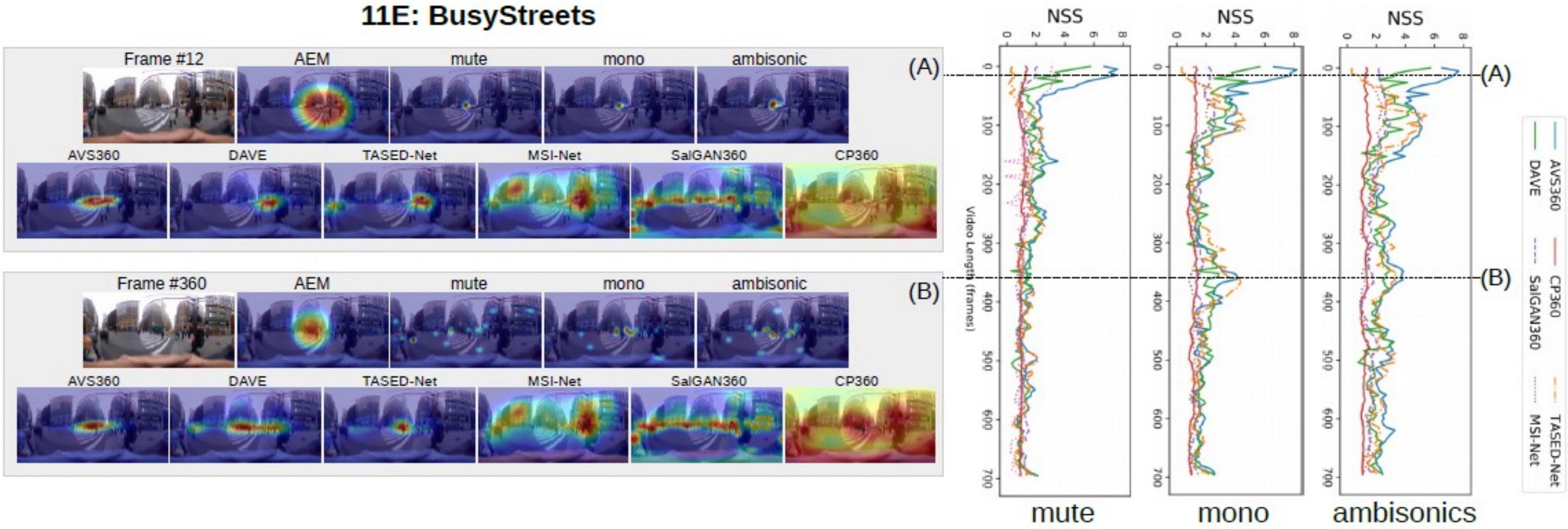
08M: BigBellTemple



EXAMPLES

Category Environment

11E: BusyStreets



Trinity College Dublin
The University of Dublin

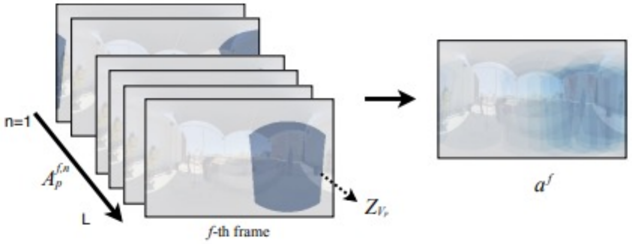
V-SENSE

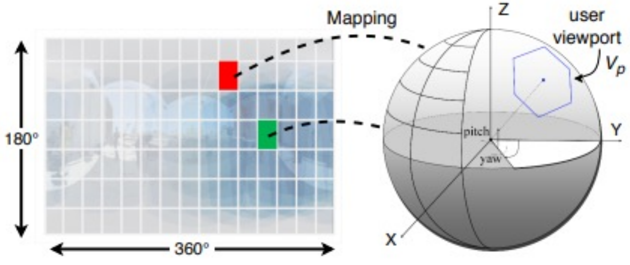
Transformer-based Long-Term Viewport Prediction in 360˚ Video: Scanpath is All You Need

Fang-Yi Chao, Cagri Ozcinar, Aljosa Smolic

Best Paper Award at IEEE MMSP, October 2021

Introduction



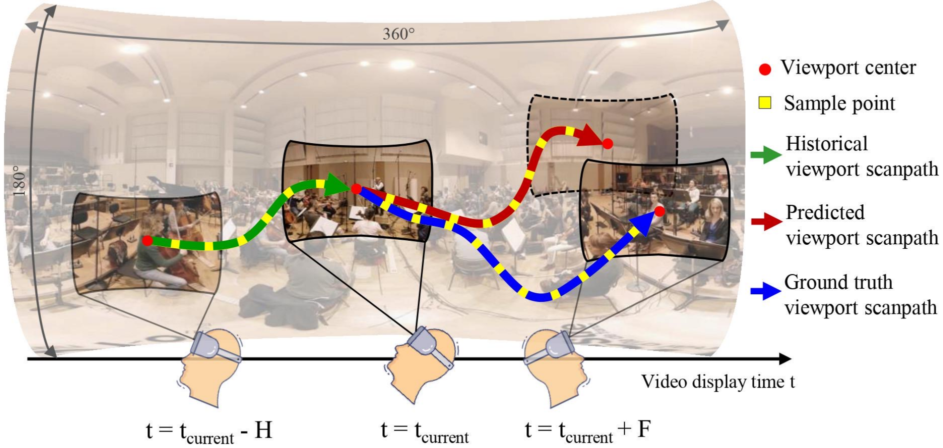


- Viewport-based adaptive streaming [8], which streams only the user's viewport of interest with high quality and streams the rest part with lower quality, has emerged as the primary technique to save bandwidth over the best-effort Internet.

- Thus, users' viewport prediction in the forthcoming seconds becomes an essential task for informing the streaming decisions in the VR system.

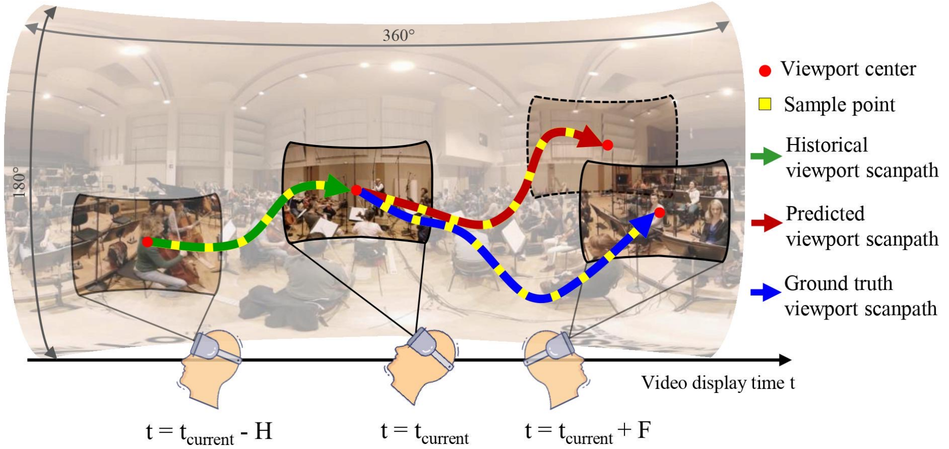
- Our goal is to predict a viewer's viewport center trajectory (*i.e.*, scanpath) in the following F seconds given the user's historical viewport center trajectory in the previous H seconds.

Problem formulation



- We define $\{P_t\}_{t=0}^{T}$ as a viewport scanpath of a viewer consuming a 360˚ video in duration T.
- It can be represented in
  - Polar coordinates $\{P_t = [\theta_t, \phi_t]\}_{t=0}^{T}$ where $[-\pi < \theta \leq \pi, -\pi/2 < \phi \leq \pi/2]$
  - Cartesian coordinates $\{P_t = [x_t, y_t, z_t]\}_{t=0}^{T}$ where $[-1 < x \leq 1,\ -1 < y \leq 1, -1 < z \leq 1]$.
- Let F denote output prediction window length and H denote input historical window length.
- In every time stamp t, the model predicts the future viewport scanpath, $\hat{P}_{t+s}$, for all prediction steps $s \in [1, F]$ with the given historical information P_{t-h} for all past steps $h \in [0, H]$.

Problem formulation



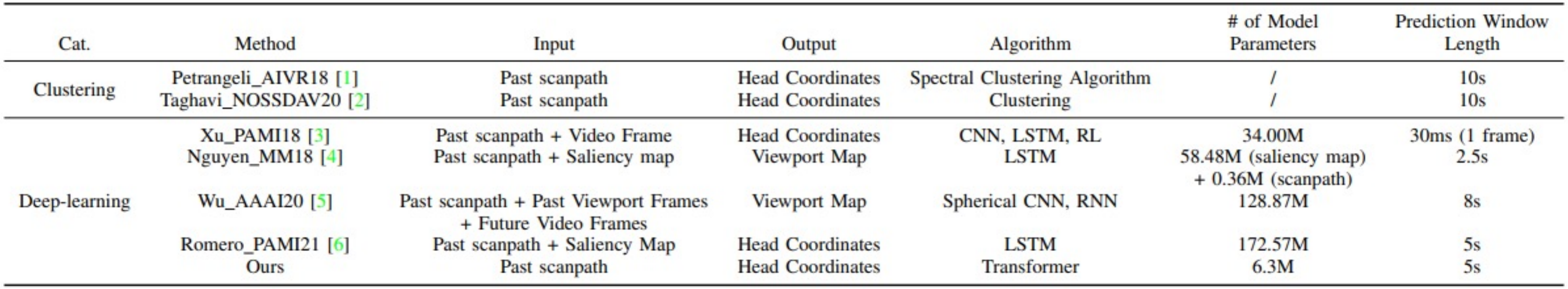
- We can formalize the problem as finding the best model f_F^* :

$$f_F^* = \arg\min E_t[D(f_F\{P_t\}_{t=t-H}^{t}, \{P_t\}_{t=t+1}^{t+F})] \qquad (1)$$

- where $D(\cdot)$ measures the geometric distance between the predicted viewport center positions and corresponding ground truth in each time step s, and E_t computes the average distance of every prediction step in interval $t \in [t+1, t+F]$.

Related work

TABLE I: Taxonomy of existing viewport prediction methods

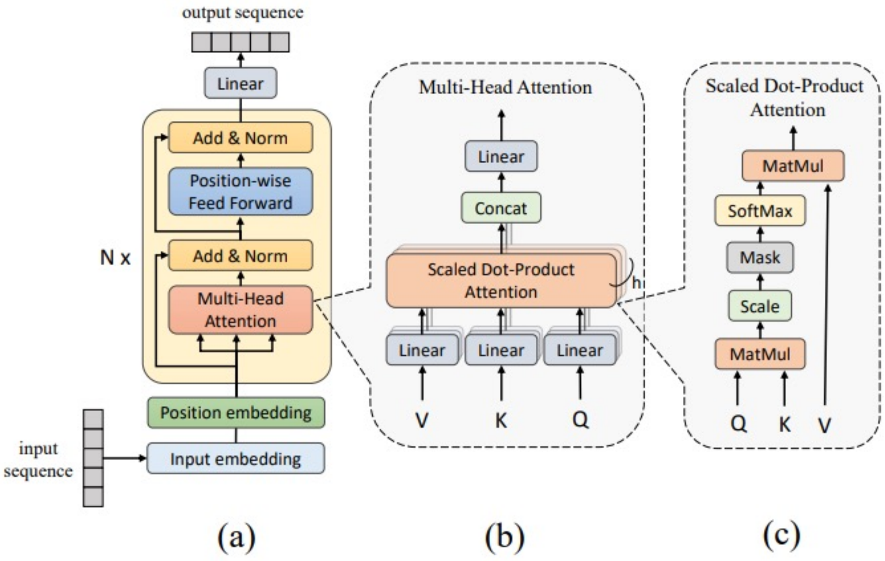


Cat.	Method	Input	Output	Algorithm	# of Model Parameters	Prediction Window Length
Clustering	Petrangeli_AIVR18 [1]	Past scanpath	Head Coordinates	Spectral Clustering Algorithm	/	10s
	Taghavi_NOSSDAV20 [2]	Past scanpath	Head Coordinates	Clustering	/	10s
Deep-learning	Xu_PAMI18 [3]	Past scanpath + Video Frame	Head Coordinates	CNN, LSTM, RL	34.00M	30ms (1 frame)
	Nguyen_MM18 [4]	Past scanpath + Saliency map	Viewport Map	LSTM	58.48M (saliency map) + 0.36M (scanpath)	2.5s
	Wu_AAAI20 [5]	Past scanpath + Past Viewport Frames + Future Video Frames	Viewport Map	Spherical CNN, RNN	128.87M	8s
	Romero_PAMI21 [6]	Past scanpath + Saliency Map	Head Coordinates	LSTM	172.57M	5s
	Ours	Past scanpath	Head Coordinates	Transformer	6.3M	5s

- **Clustering-based methods:**
  - Pros: Have relatively less computation.
  - Cons: The clusters of every video are content-dependant. It requires collecting viewing trajectories from multiple users for any 360˚ video.

- **Deep-learning-based methods:**
  - Pros: Can directly be applied to any 360˚ video using the trained model.
  - Cons: Their complex architectures, which have many learnable parameters in the models, require heavy computation and lead to high latency in the streaming system.

Method: Transformer-based VPT360



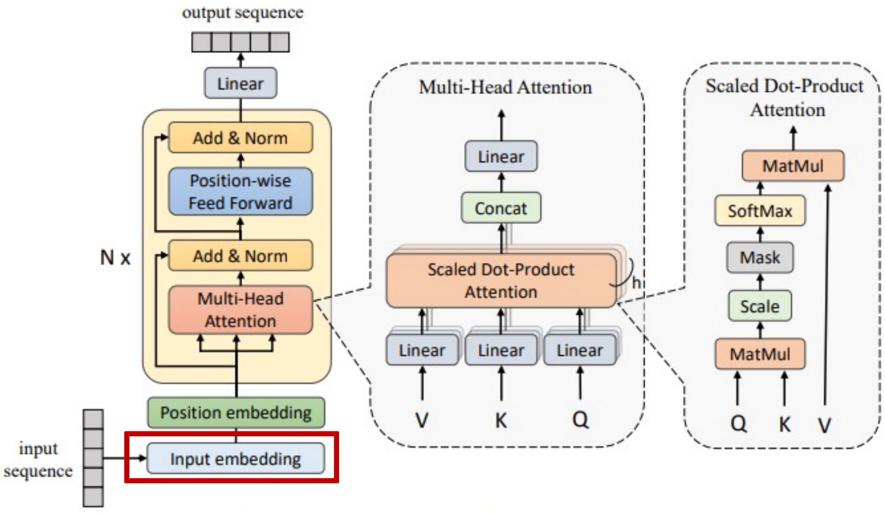
Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

- Our transformer-based model uses only the viewport scanpath without requiring any other content information (e.g., video frames, saliency maps, etc.) to reduce the computational cost and attain superior results compared to existing methods.
- Unlike RNN, which processes sequential data in order, transformers simultaneously take account of multiple elements in the input sequence and attribute different weights to model the impacts between each element.
- This architecture achieves better long-term dependency modeling and larger-batch parallel training compared to RNNs.

Proposed method



Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

Input embedding:

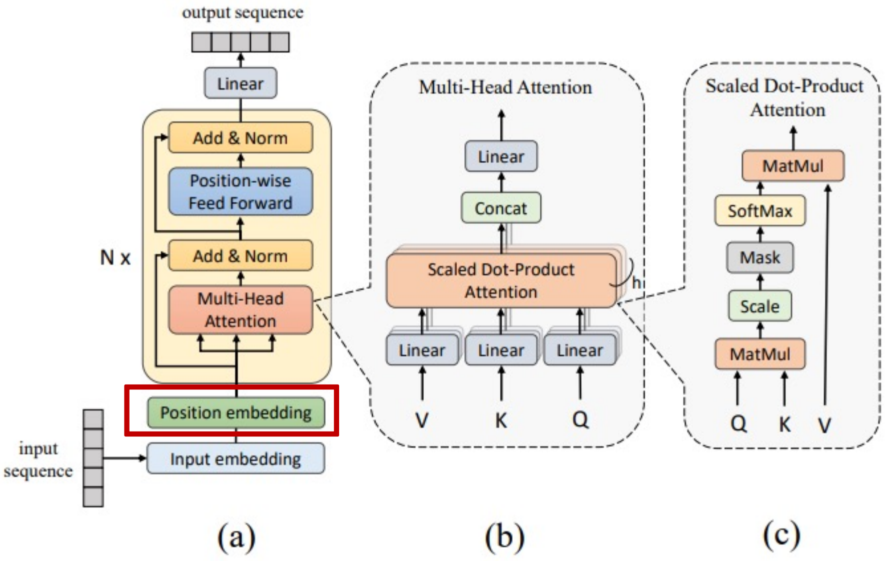
- The transformer block shown in Fig. (a) processes a set of scanpath embeddings $\{e_t\}_{t=t-H}^{t}$ as input, and output a set of updated embeddings $\{e_t'\}_{t=t-H}^{t}$ with temporal dependencies.

- We can create the query, key, and value matrices $Q \in \mathbf{R}^{d_k \times d_{model}}$, $K \in \mathbf{R}^{d_k \times d_{model}}$, $V \in \mathbf{R}^{d_v \times d_{model}}$ respectively, from the given input sequence with the functions:

$$Q = f_Q\left(\{e_j\}_{j=1}^{t}\right), K = f_K\left(\{e_j\}_{j=1}^{t}\right), V = f_V(\{e_j\}_{j=1}^{t}) \quad (2)$$

where f_Q, f_K and f_V are the corresponding query, key and value functions which linearly project the input sequence.

Proposed method



Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

Position embedding:

- Since there is no recurrent unit in transformer layer to capture temporal features, we exploit positional embedding to infuse the relative or absolute position information of the elements in the input sequence.

- We use the summation of input sequence with sine and cosine functions [9] of different frequencies:

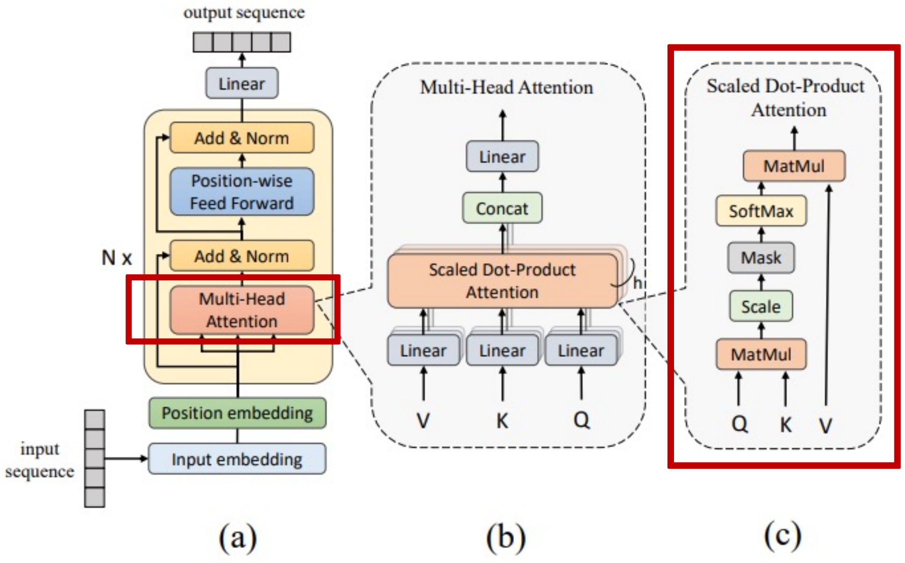
$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{3}$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{4}$$

where pos denotes the position and i denotes the dimension. This sinusoidal function allows the model to attend by relative positions easily.

- We also tried the learnable position embedding as used in [10], but found that this led to worse performance in our case. We analyze the effect of the position embedding in experiments.

Proposed method



Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

Scaled Dot-Product Attention

- The attention weights between each element can be calculated with the scaled dot-product attention defined as:

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{Q,K^T}{\sqrt{d_k}}\right)V \qquad (5)$$

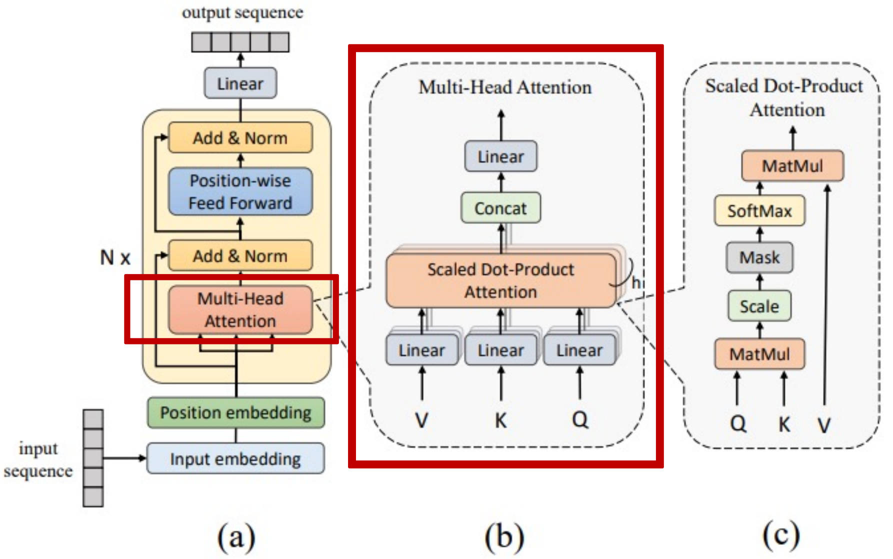
- It can be regarded as the matrix V multiply with the weights calculated by the matrix Q and K.

- The weights are defined by how each element of the sequence Q is influenced by all the other elements in the sequence K.

- the softmax function normalizes the weights to yield a distribution between 0 and 1. Those weights are then applied to all the elements in the sequence in V.

- The scale factor $\sqrt{d_k}$ is to avoid overly large values of the inner product, especially when the dimensionality is high.

Proposed method



Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

Multi-Head Self-Attention Module:

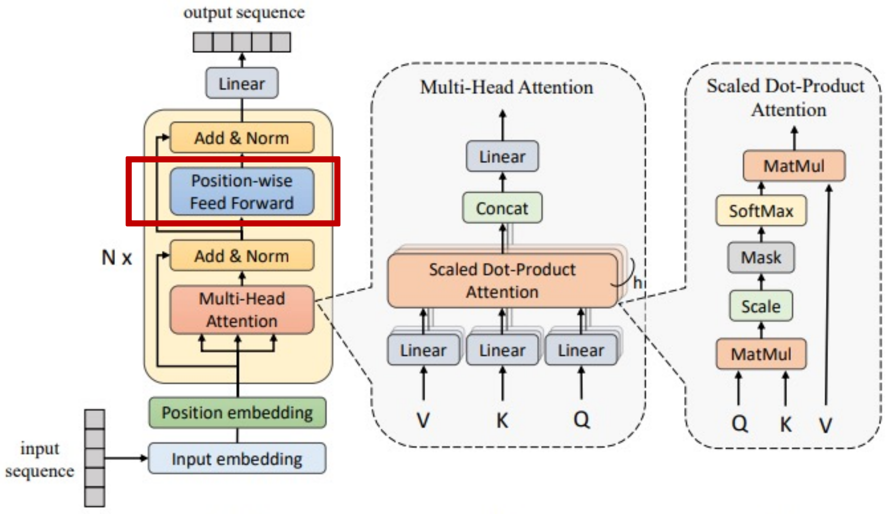
- The attention mechanism, can be repeated multiple times with linear projections of Q, K, and V.
- This multi-head attention benefits the model to learn from different representations of Q, K, and V by jointly attending to information from different representation subspaces at other positions.
- These linear representations are done by multiplying Q, K, and V by weight matrices W^Q, W^K, W^V that are learned during the training process.

$$\text{MultiHead}(Q,K,V) = \text{Concat}\left(\left[head_{j=1}^{h}\right]\right)W^O \qquad (6)$$

$$\text{where } head_j = \text{Att}(QW_j^Q, KW_j^K, VW_j^V) \qquad (7)$$

- The projections are parameter matrices $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_k}$, and $W_i^O \in R^{hd_v \times d_{model}}$.

Proposed method



(a)

Architecture of our transformer-based VPT360 model

(b)

Multi-Head Attention Module

(c)

Scaled Dot-Product Attention

Position-wise Feed-Forward Network

- After self-attention sub-layers aggregate all input embeddings with adaptive weights, each layer contains a fully connected feed-forward network to consider interactions between different dimensions.

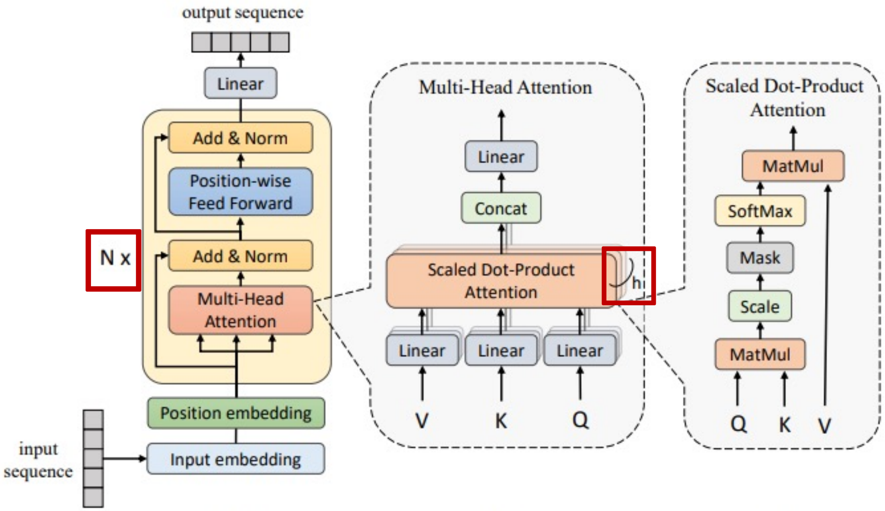
- It consists of two linear transformations with a $ReLU$ activation in between and is applied to each position separately and identically.

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \tag{8}$$

- where $W_1 \in R^{d_{model} \times 4d_{model}}$, $W_2 \in R^{d_{model} \times 4d_{model}}$, $b_1 \in R^{4d_{model}}$ and $b_2 \in R^{4d_{model}}$ are learnable weights and shared across all positions.

- Note that while the linear transformations are the same across different positions, they use different weights in different layers.

Proposed method



(a) Architecture of our transformer-based VPT360 model

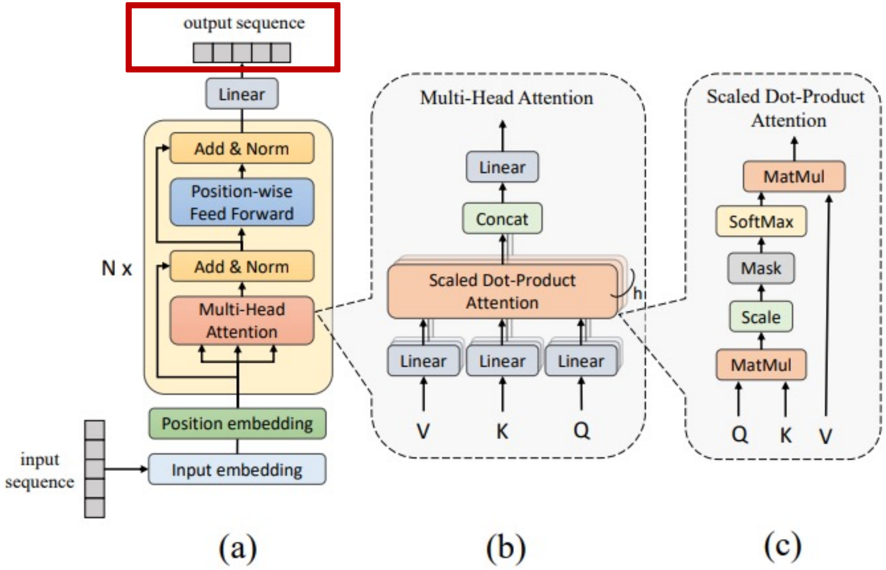
(b) Multi-Head Attention Module

(c) Scaled Dot-Product Attention

VPT360

- Our transformer employs $N = 1$ layer encoder with $h = 8$ multi-head attention
- Model length $d_{model} = 512$, where $d_k = d_v = \frac{d_{model}}{h} = 64$ following the vanilla transformer [9].

Proposed method



Architecture of our transformer-based VPT360 model

Multi-Head Attention Module

Scaled Dot-Product Attention

Combination Loss Function

- Our loss function is then defined as the combination of position MSE and motion velocity MSE as:

$$L = \alpha MSE(P, \hat{P}) + \beta MSE(V, \hat{V}) \tag{9}$$

where $\hat{V}$ denotes the motion velocity in predicted position $\hat{P}$, and V denotes the motion velocity in ground truth position P.

- The motion velocity V is the root mean square value of the position in current moment and the position in the last moment, where

$$\begin{aligned} V &= \sqrt{(P_t - P_{t-1})^2} \\ &= \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2 + (z_t - z_{t-1})^2}. \end{aligned} \tag{10}$$

- The hyper-parameters α and β are used to balance the scale of two loss components. We set $(\alpha, \beta) = (0.75, 0.25)$ by experiments.

Experiments

- **Settings:**
  - We use Cartesian coordinates (i.e., *x, y, z*) rather than Polar coordinates (i.e., θ, ϕ) to represent viewport position since the former retains continuous between ±1 in x, y, z dimensions on the sphere, while the latter has a periodic issue which −π = π in θ coordinate.
  - We set the input historical window length $H = 1$ second and output prediction window length $F = 5$ seconds since the professional streaming systems (e.g., Facebook) download video segments at least 5 seconds before the playout time [14].
  - The sample rate is 25 elements per second, implying that the model inputs a 25-element sequence and outputs a 125- element sequence.

- **Dataset:**

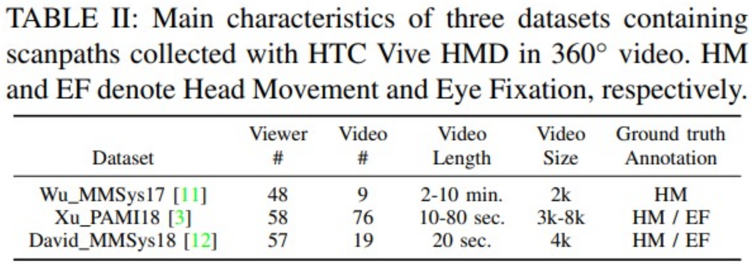


TABLE II: Main characteristics of three datasets containing scanpaths collected with HTC Vive HMD in 360° video. HM and EF denote Head Movement and Eye Fixation, respectively.

Dataset	Viewer #	Video #	Video Length	Video Size	Ground truth Annotation
Wu_MMSys17 [11]	48	9	2-10 min.	2k	HM
Xu_PAMI18 [3]	58	76	10-80 sec.	3k-8k	HM / EF
David_MMSys18 [12]	57	19	20 sec.	4k	HM / EF

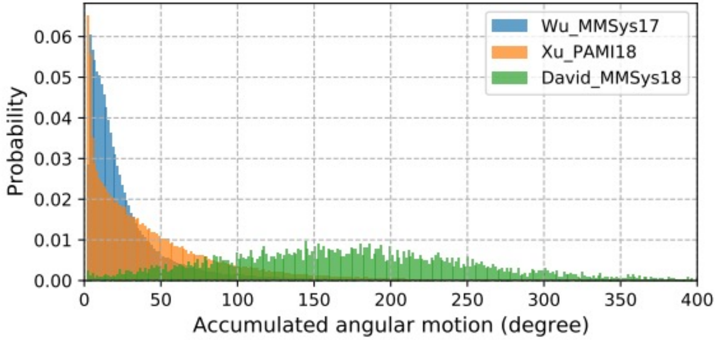


Fig. 3: Histogram of accumulative angular motion in every 5-second scanpath in three datasets. Accumulated angular motion is the summation of absolute head rotation angle in a given scanpath.

Experiments

Three Evaluation Metrics:

- Average great-circle distance: the distance between the predicted point $\hat{P}_t = (\hat{\theta}_t, \hat{\phi}_t)$ and groudtruth point $P_t = (\theta_t, \phi_t)$ on a sphere. The lower value indicates the higher prediction accuracy.
- Mean overlap: the average overlap ratio of intersection over the union between predicted and ground truth viewport area in a given prediction window. The higher value indicates the higher prediction accuracy.
- The average ratio of overlapping tiles measures accuracy in terms of the percentage of overlapping tiles between predicted and ground truth viewports. We follow Nguyen_MM18 [4] to set (9, 16) tiles in a frame. The higher value indicates the higher prediction accuracy.
- The FoV size is set to 100˚ .

Baseline:

- No-prediction method: the repetition of the last element in the input scanpath as an output scanpath.

Experiments

Ablation study

- 1^{st} row: The impact of sinusoidal/learnable positional embedding (denoted SPE/LPE) and combination loss function (denoted L_{pos} and $L_{pos+vel}$). From the results, we can see that sinusoidal outperforms learnable positional embedding in all 5-second sequences.

- 2^{nd} row: The results of different input historical window lengths. It shows that the shorter window length leads to better short-term prediction but worse long-term prediction. We set the window length to 1 second as it achieves the best results most of the time.

- 3^{rd} row: The effects of different transformer encoder (denoted Enc) layers. We discover that, in our work, only using one-layer encoder performs satisfactorily. More layers of encoder do not improve the results.

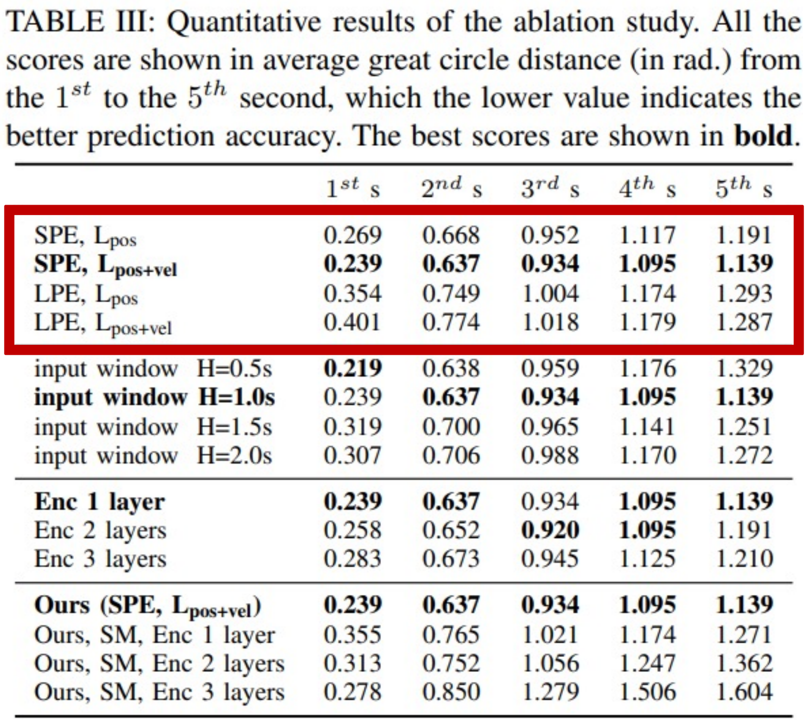


TABLE III: Quantitative results of the ablation study. All the scores are shown in average great circle distance (in rad.) from the 1^{st} to the 5^{th} second, which the lower value indicates the better prediction accuracy. The best scores are shown in **bold**.

	1^{st} s	2^{nd} s	3^{rd} s	4^{th} s	5^{th} s
SPE, L_{pos}	0.269	0.668	0.952	1.117	1.191
SPE, $L_{pos+vel}$	**0.239**	**0.637**	**0.934**	**1.095**	**1.139**
LPE, L_{pos}	0.354	0.749	1.004	1.174	1.293
LPE, $L_{pos+vel}$	0.401	0.774	1.018	1.179	1.287
input window H=0.5s	**0.219**	0.638	0.959	1.176	1.329
input window H=1.0s	0.239	**0.637**	**0.934**	**1.095**	**1.139**
input window H=1.5s	0.319	0.700	0.965	1.141	1.251
input window H=2.0s	0.307	0.706	0.988	1.170	1.272
Enc 1 layer	**0.239**	**0.637**	0.934	**1.095**	**1.139**
Enc 2 layers	0.258	0.652	**0.920**	**1.095**	1.191
Enc 3 layers	0.283	0.673	0.945	1.125	1.210
Ours (SPE, $L_{pos+vel}$)	**0.239**	**0.637**	**0.934**	**1.095**	**1.139**
Ours, SM, Enc 1 layer	0.355	0.765	1.021	1.174	1.271
Ours, SM, Enc 2 layers	0.313	0.752	1.056	1.247	1.362
Ours, SM, Enc 3 layers	0.278	0.850	1.279	1.506	1.604

Experiments

Ablation study

- Referring to other existing methods using saliency maps to improve the prediction, we integrate ground truth saliency maps of future frames to see if it contributes to better prediction.
- We use the method proposed in Romero_PAMI21 [6] to flatten the saliency maps of the next frame into one dimension and concatenate it with position embedded input sequence. We then use the transformer encoder to encode the concatenation of position embedded sequence and flattened saliency map.
- 4th row: The results of integrating ground truth saliency maps (denoted SM) with encoders in a different number of layers. We can see that the performance is not improved by simply combining the saliency maps. A better integration method is required.

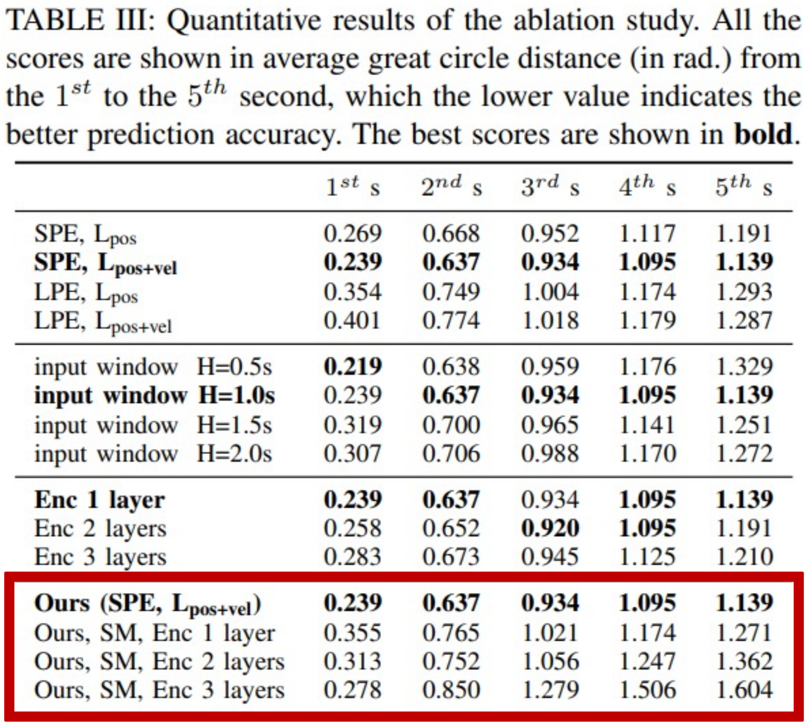
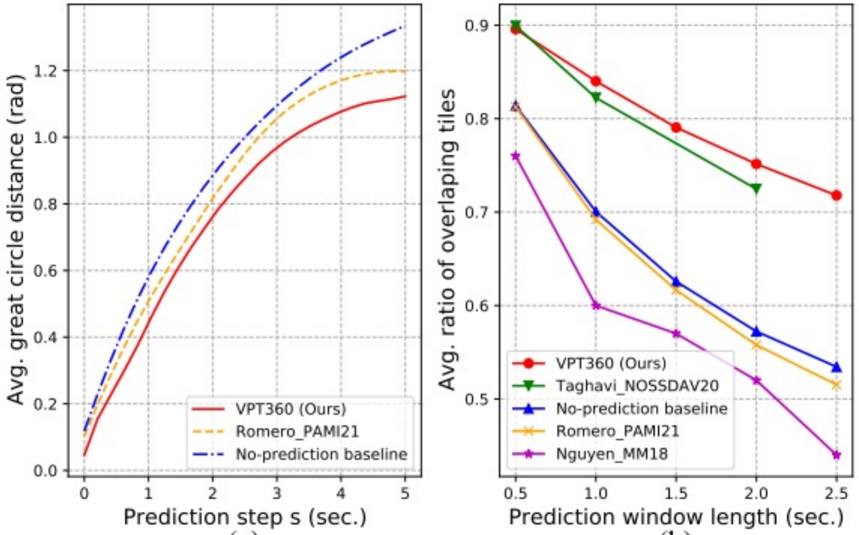


TABLE III: Quantitative results of the ablation study. All the scores are shown in average great circle distance (in rad.) from the 1^{st} to the 5^{th} second, which the lower value indicates the better prediction accuracy. The best scores are shown in **bold**.

	1^{st} s	2^{nd} s	3^{rd} s	4^{th} s	5^{th} s
SPE, L_{pos}	0.269	0.668	0.952	1.117	1.191
SPE, $L_{pos+vel}$	**0.239**	**0.637**	**0.934**	**1.095**	**1.139**
LPE, L_{pos}	0.354	0.749	1.004	1.174	1.293
LPE, $L_{pos+vel}$	0.401	0.774	1.018	1.179	1.287
input window H=0.5s	**0.219**	0.638	0.959	1.176	1.329
input window H=1.0s	0.239	**0.637**	**0.934**	**1.095**	**1.139**
input window H=1.5s	0.319	0.700	0.965	1.141	1.251
input window H=2.0s	0.307	0.706	0.988	1.170	1.272
Enc 1 layer	**0.239**	**0.637**	0.934	**1.095**	**1.139**
Enc 2 layers	0.258	0.652	**0.920**	**1.095**	1.191
Enc 3 layers	0.283	0.673	0.945	1.125	1.210
Ours (SPE, $L_{pos+vel}$)	**0.239**	**0.637**	**0.934**	**1.095**	**1.139**
Ours, SM, Enc 1 layer	0.355	0.765	1.021	1.174	1.271
Ours, SM, Enc 2 layers	0.313	0.752	1.056	1.247	1.362
Ours, SM, Enc 3 layers	0.278	0.850	1.279	1.506	1.604

Results

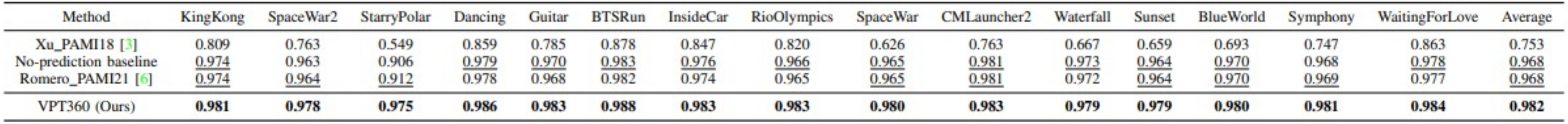
Comparison with the state of the arts



(a) (b)

Fig. 4: Comparison results on (a) David_MMSys18 and (b) Wu_MMSys17 dataset, respectively.

TABLE IV: Comparison with Xu_PAMI18: Mean Overlap scores of FoV prediction, prediction window length F ≈ 30ms (1 frame). The best score is shown in **bold** and the second-best score is shown in underline.



Method	KingKong	SpaceWar2	StarryPolar	Dancing	Guitar	BTSRun	InsideCar	RioOlympics	SpaceWar	CMLauncher2	Waterfall	Sunset	BlueWorld	Symphony	WaitingForLove	Average
Xu_PAMI18 [3]	0.809	0.763	0.549	0.859	0.785	0.878	0.847	0.820	0.626	0.763	0.667	0.659	0.693	0.747	0.863	0.753
No-prediction baseline	0.974	0.963	0.906	0.979	0.970	0.983	0.976	0.966	0.965	0.981	0.973	0.964	0.970	0.968	0.978	0.968
Romero_PAMI21 [6]	0.974	0.964	0.912	0.978	0.968	0.982	0.974	0.965	0.965	0.981	0.972	0.964	0.970	0.969	0.977	0.968
VPT360 (Ours)	**0.981**	**0.978**	**0.975**	**0.986**	**0.983**	**0.988**	**0.983**	**0.983**	**0.980**	**0.983**	**0.979**	**0.979**	**0.980**	**0.981**	**0.984**	**0.982**

Results

Comparison with the state of the arts

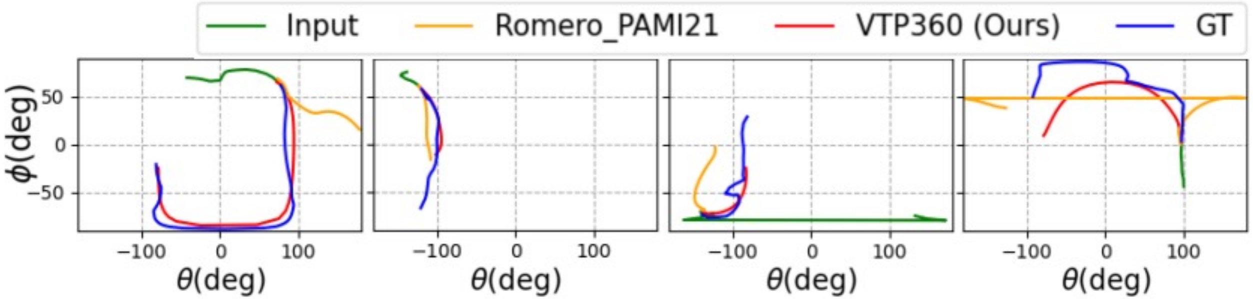


Fig. 5: Four examples of viewport scanpath predicted by our VPT360 and Romero_PAMI21 on the David_MMSys18 dataset.

Conclusion

- We introduced a novel transformer-based long-term viewport prediction method for 360° video, namely VPT360.
- We process the user's viewport scanpath as a time-dependent sequence and model the time dependencies to predict future viewport scanpath.
- By exploiting the self-attention mechanism in the transformer to compute the impact between every two elements in a sequence, we efficiently model long-term time dependencies in the viewport scanpath without any other video content information.
- Our ablation study validated the usage of sinusoidal position embedding, combination loss function, and 1-layer transformer encoder processing 1-second viewport scanpath contributes to the highest accuracy in 5- second prediction.
- Our VPT360 requires the least learnable parameters and achieves the highest accuracy on short-term and long-term prediction over three widely-used datasets compared with other state-of-the-art methods.
- In future work, we intend to develop an effective yet simple method to integrate the saliency maps of 360° video to increase the prediction accuracy and benefit the streaming system.

Trinity College Dublin

The University of Dublin

V-SENSE

Volumetric Video – 6DoF

Professor Aljosa Smolic

SFI Research Professor of Creative Technologies

Trinity College Dublin
The University of Dublin

V-SENSE

User Behaviour Analysis of Volumetric Video in Augmented Reality

Emin Zerman, Radhika Kulkarni, Aljosa Smolic

V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

Trinity College Dublin
The University of Dublin

V-SENSE

2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

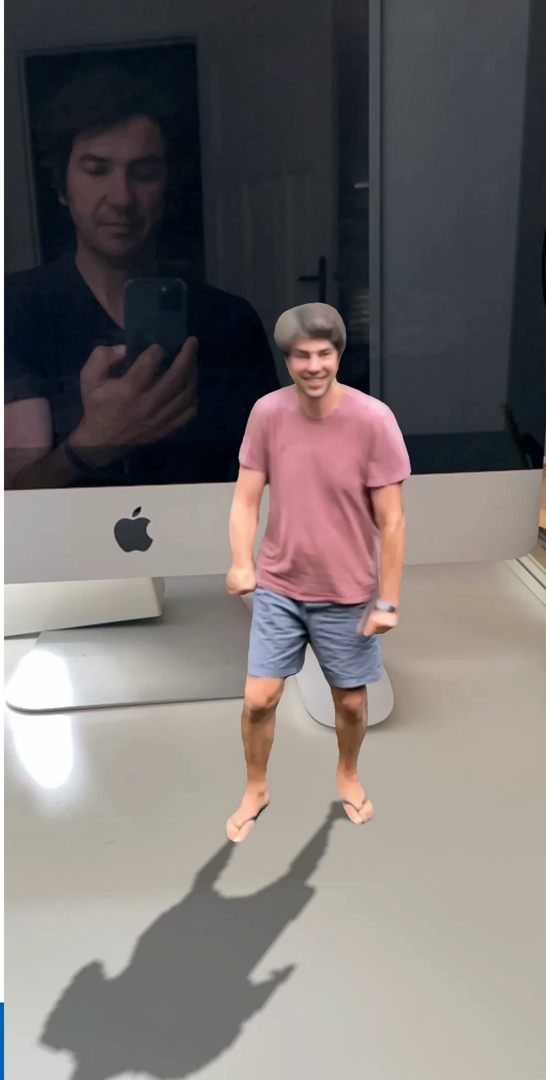
VOLOGRAMS

Volu: Volumetric video content creation with a single mobile phone, available to everyone!

https://www.volograms.com

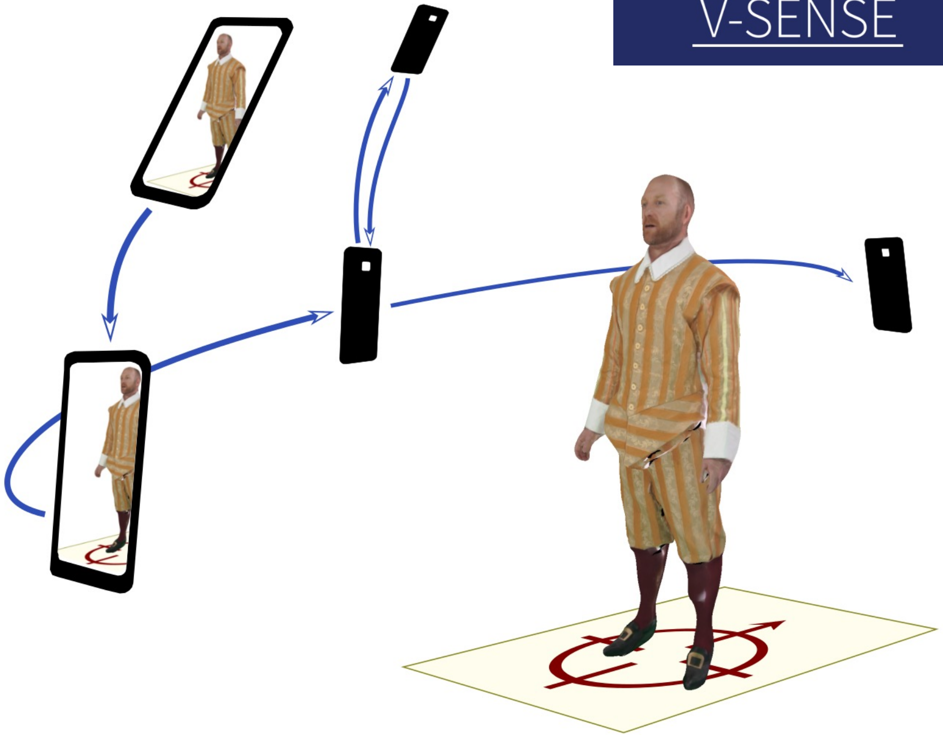
https://youtu.be/mlADjAIlSYs

https://www.volograms.com/volu



Remote User Behaviour Study

- A mobile AR application for Android smartphones was developed.
- User behaviour data was collected remotely.
- We provide a detailed analysis of user navigation in a marker-based AR scenario

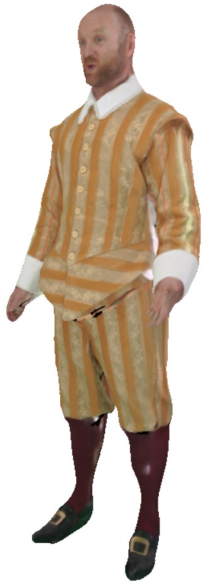


User Study – Volumetric Data

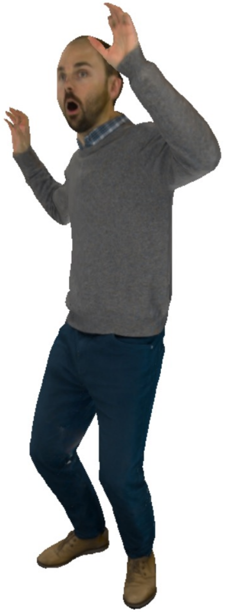
Two volumetric videos, represented as dynamic 3D meshes with texture information:

- "Sir Frederick"
  - 60-seconds long,
  - showing a man telling a story for the visitors of a castle
  - ∼25K polygons and 1024 × 1024 pixel texture maps
- "Nico"
  - 7-seconds long,
  - a sample VV showing a surprised man
  - ∼16K polygons and 4096 × 4096 pixel texture maps

- Both videos were 30 fps.



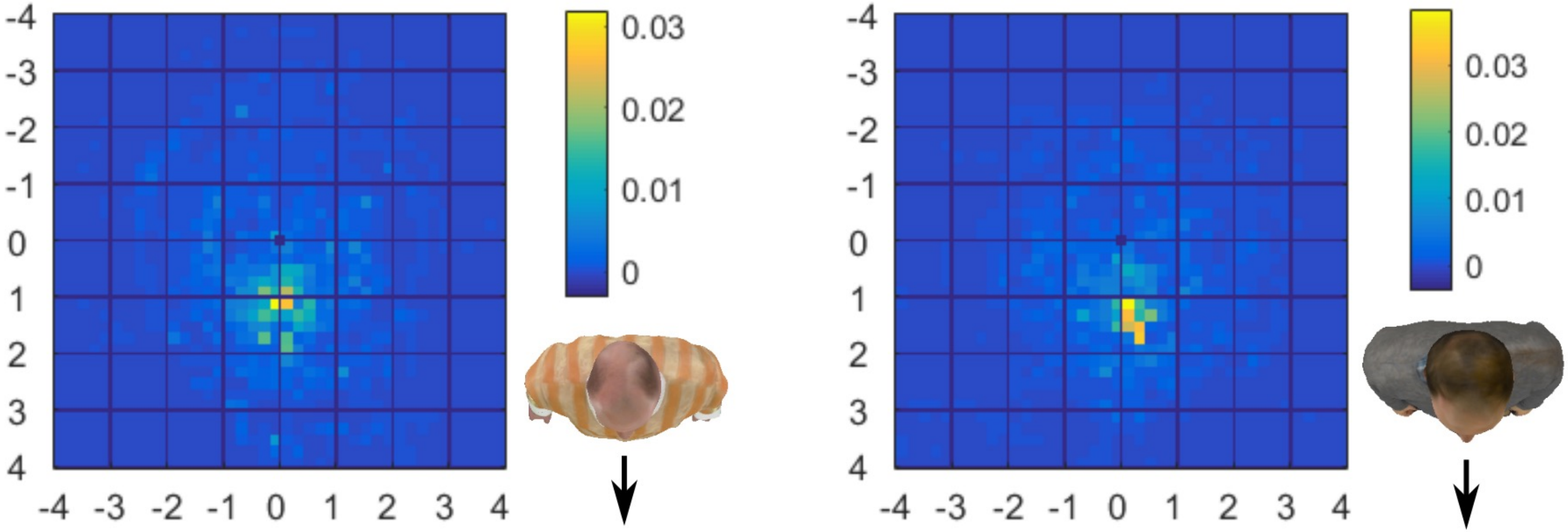
"Sir Frederick"



"Nico"

Quantitative Analysis Results

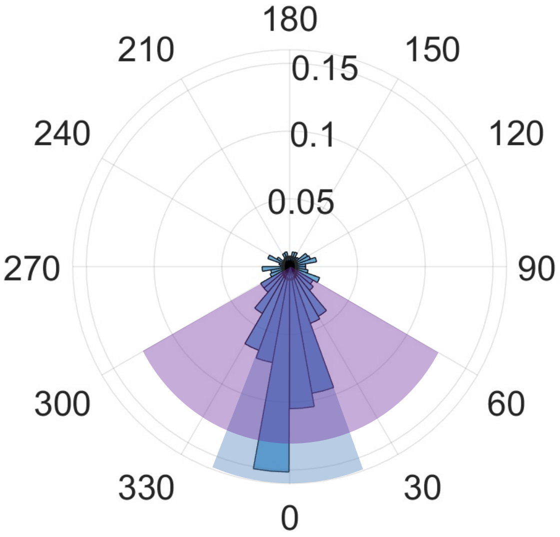
2D histogram of relative positions, with respect to the VV



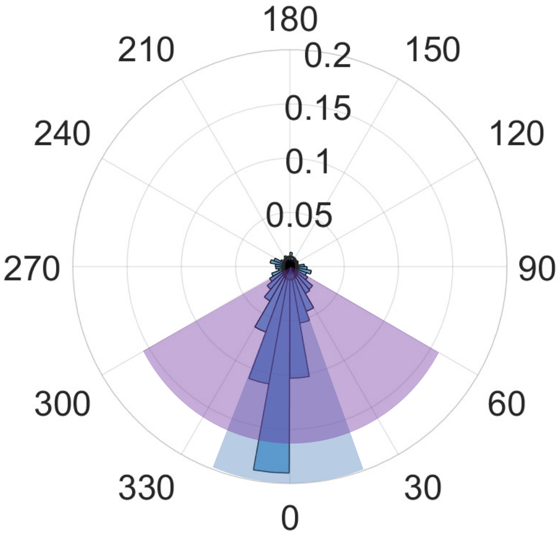
Quantitative Analysis Results

Distribution of users' relative viewpoints

- 44% of the time looking at the front
  - (±20° difference from centre)
- 72% of the time looking at
  - a larger frontal arc of ±60°



"Sir Frederick"



"Nico"

Quantitative Analysis Results

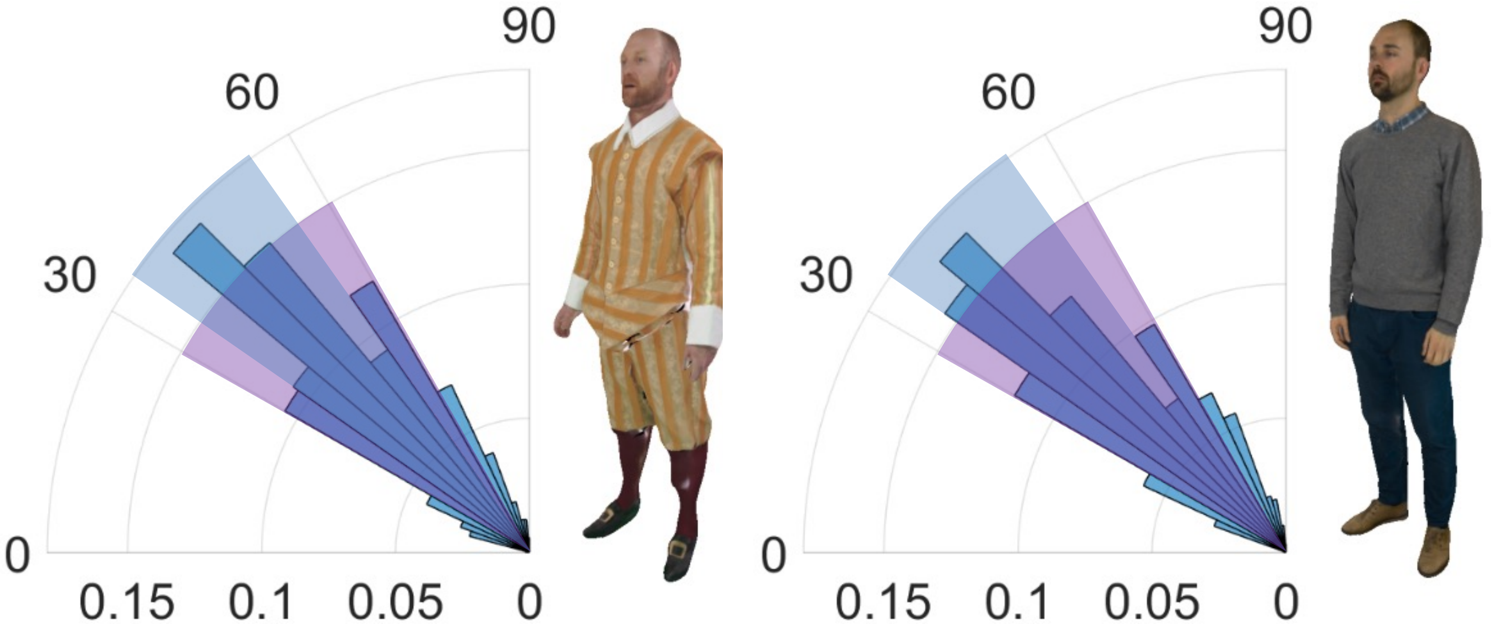
Distribution of users’ relative vertical viewpoints

44% of time

- [35°, 55°]

74% of time

- [30°, 60°]

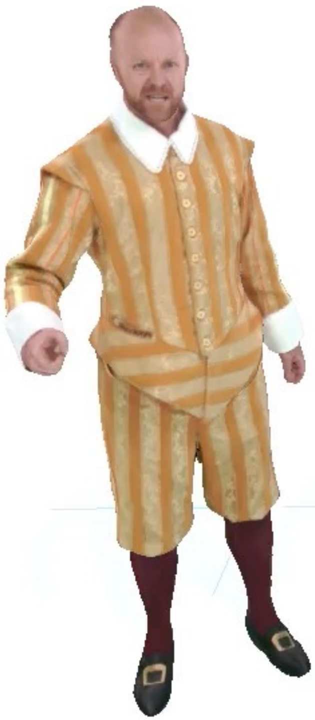


Trinity College Dublin
The University of Dublin

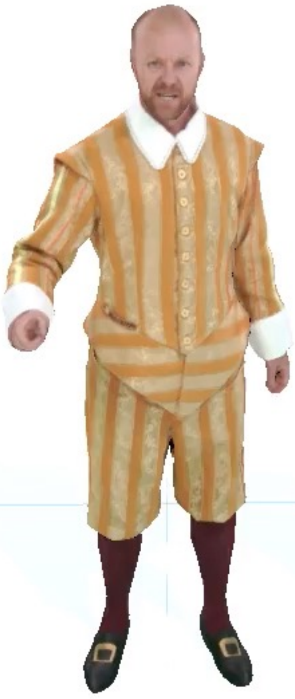
V-SENSE

Volumetric Video Quality

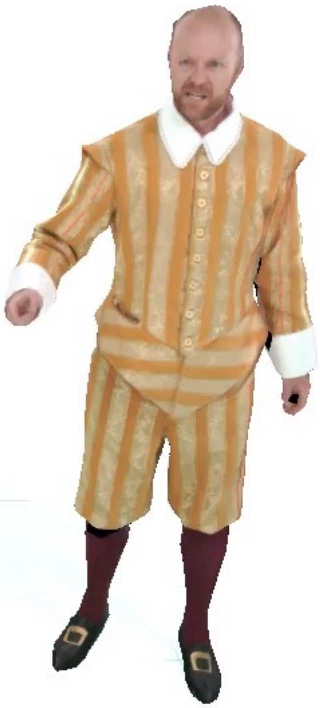
3D Mesh Resolution



50k polys/frame
~1GB / minute



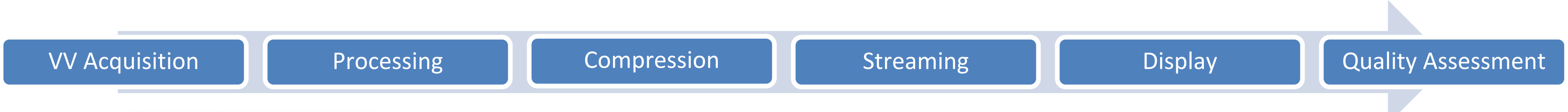
15k polys/frame
~400MB / minute



4k polys/frame
~150MB / minute

Content Delivery Pipeline for Volumetric Video

- From content acquisition to display and quality assessment, several steps of the content delivery pipeline for free-viewpoint videos are designed and analysed.



VV Acquisition

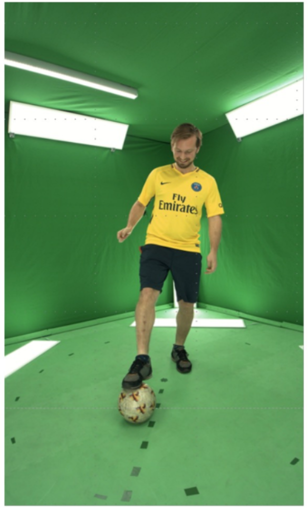
Processing

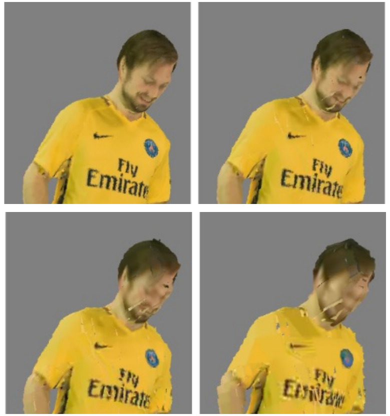
Compression

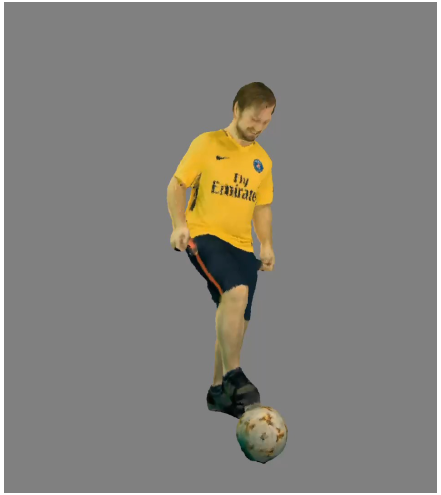
Streaming

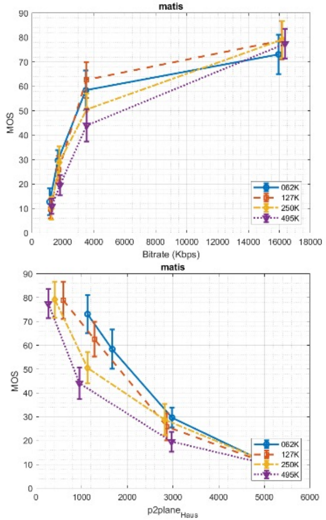
Display

Quality Assessment









- Zerman, Emin; Gao, Pan; Ozcinar, Cagri; Smolic, Aljosa; “*Subjective and objective quality assessment for volumetric video compression*”, **IS&T Electronic Imaging, Image Quality and System Performance XVI, 2019**.

Volumetric Video

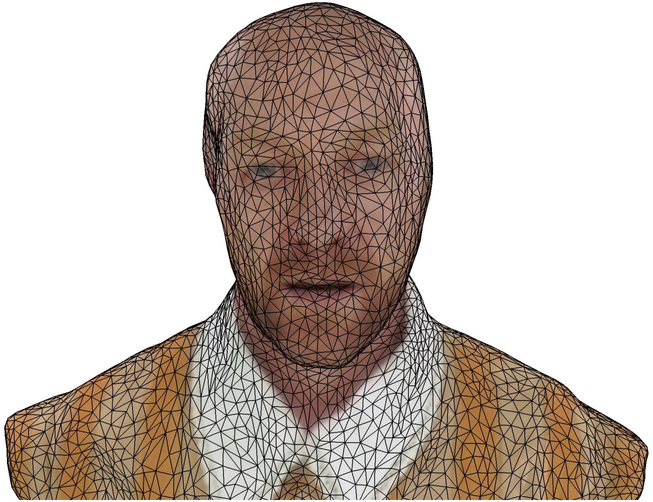
How is it stored?

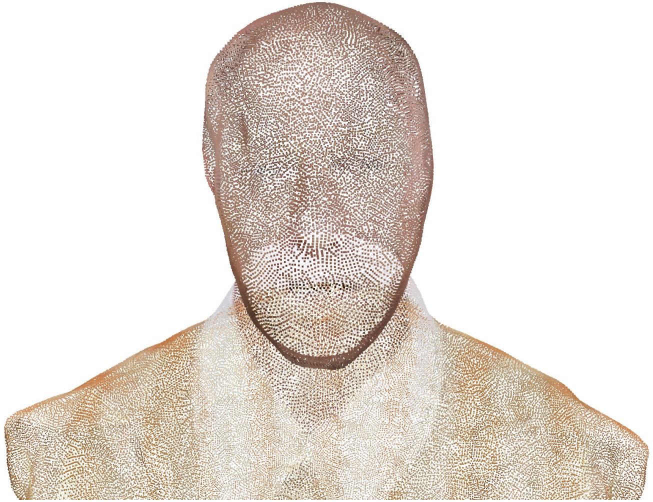
Textured polygonal meshes

- Vertices and Faces
- Texture atlas

Coloured point clouds

- Points
- Attributes (e.g., colour, normal, etc.)





Related Work

Quality assessment

Polygonal meshes:

- Doumanoglou et al. (2019) an Christaki et al. (201 different open sour compression algorit
- Google's Draco was found the best performing mesh compression method, among:
  - Corto
  - Draco
  - O3DGC
  - OpenCTM

Coloured clouds

- d Alexiou et al. (2019) nance of MPEG point methods on static clouds.
- Zerman et al. (20) and Gonçalvez et al. 2 (i.e., V-PCC).
- TMC2) was found to erforming method.

There are **no** publicly available large QA databases for VV which can be used for understanding of QA for VV!

Textured meshes and point clouds were **not** compared in the literature!

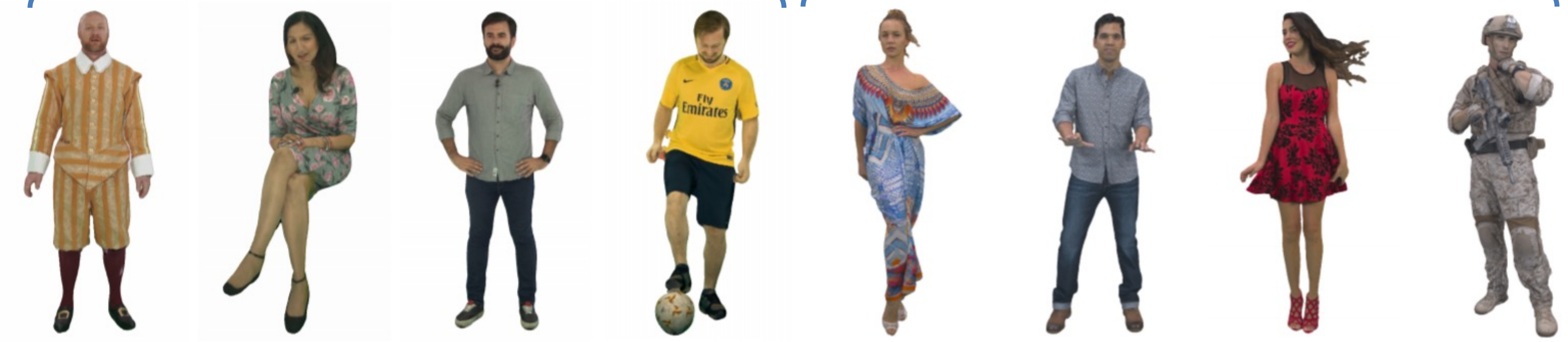
Creating vsenseVVDB2 Database

Both textured meshes and coloured PCs

V-SENSE Data

Only coloured point clouds

8i Point Clouds



(a) AxeGuy [v:25K / p:405K]

(b) LubnaFriends [v:25K / p:402K]

(c) Rafa2 [v:25K / p:406K]

(d) Matis [v:25K / p:406K]

(e) Longdress [p:765K]

(f) Loot [p:784K]

(g) Redandblack [p:729K]

(h) Soldier [p:1.06M]

Zerman, Emin; Ozcinar, Cagri; Gao, Pan; Smolic, Aljosa

Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression

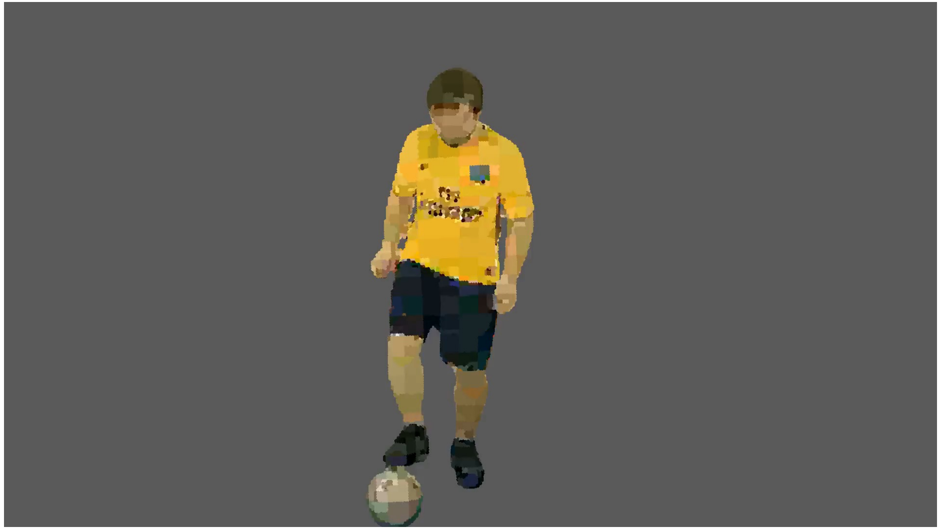
Twelfth International Conference on Quality of Multimedia Experience (QoMEX), IEEE Athlone, Ireland.

Subjective Data Collection

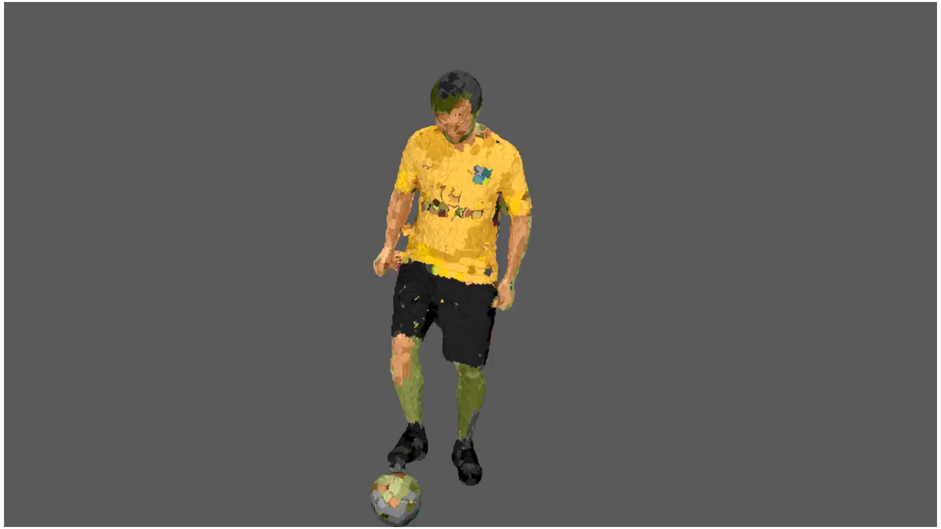
Sample stimuli

Here are two heavily compressed sequences

TMC1-RAHT (Point cloud)



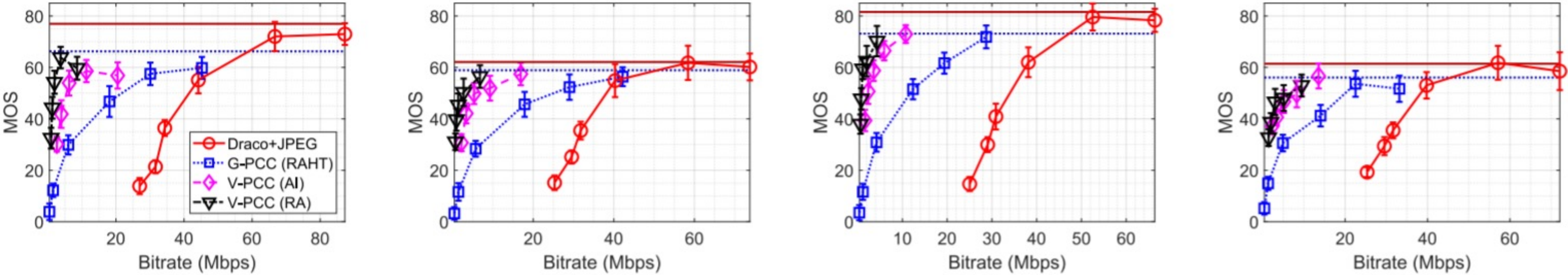
Draco + JPEG (Mesh)



Results

Mesh vs. Point Cloud

Textured mesh seems to be better than point clouds in high-bitrate cases, whereas point cloud compression is better in limited bitrate cases.



(a) AxeGuy (b) LubnaFriends (c) Rafa2 (d) Matis

Zerman, Emin; Ozcinar, Cagri; Gao, Pan; Smolic, Aljosa

Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression

Twelfth International Conference on Quality of Multimedia Experience (QoMEX), IEEE Athlone, Ireland.

Trinity College Dublin
The University of Dublin

V-SENSE

Swift @ Library

Jonathan Swift: AR application for the Long Room

Enhancing museum visitor experience

- The Library of Trinity College Dublin curators were interested in using immersive imaging
- Augmented Reality improves visitor experience while preserving the cultural heritage site

User studies for validation

- Conducted in The Long Room in the Old Library in Trinity College Dublin
- Outside regular opening hours
- Apple iPad & Microsoft HoloLens

O'Dwyer, N., Ondej, J., Pagés R., Amplianitis, K., and Smolic, A. (2018). **Jonathan Swift: Augmented Reality Application for Trinity Library's Long Room**. In: Rouse R., Koenitz H., Haahr M. (eds*) Interactive Storytelling. ICIDS 2018. Lecture Notes in Computer Science*, vol 11318. Springer, Cham.

Young, G. W. (2019). **Demonstration of the Jonathan Swift Experience: An augmented and virtual reality application for TCD's Long Room library [Interactive AR Experience]**. *The 13th Annual Irish Human-Computer Interaction (HCI) Conference*, NUI, Galway.

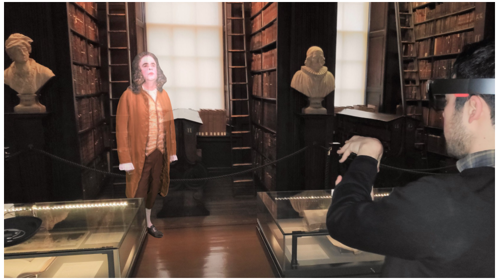
Experiment Methodology

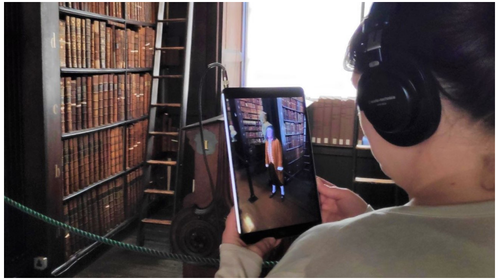
Participants

- TCD Library and Faculty of Arts & Humanities
- 17 volunteers between ages 25-65

Procedure

- Briefing
- Presentation (at least 3 mins each)
  - Both tablet and HMD
  - Randomized order
- Questionnaire
  - 5-point Likert scale & Preference
  - "Why did you give this score?"

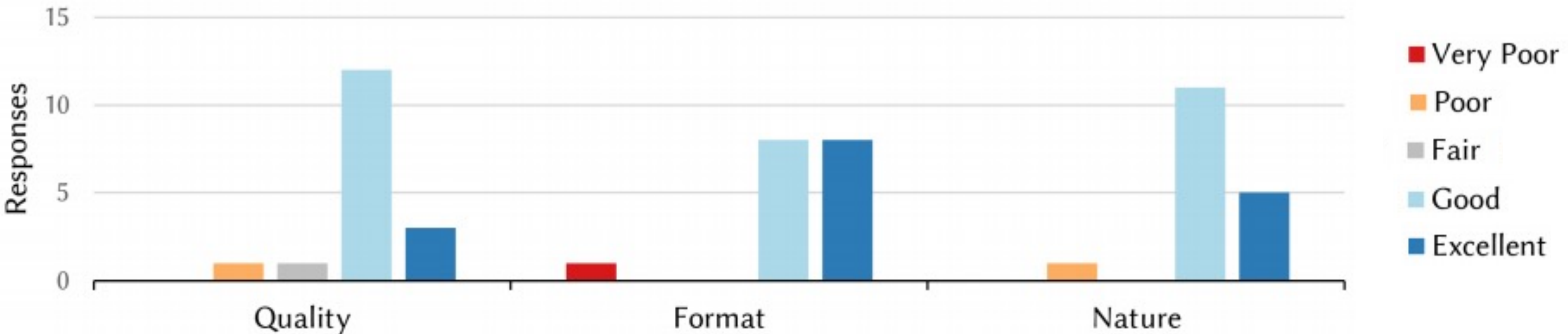




Zerman, E., O'Dwyer, N., Young, G. W, and Smolic, A. (2020). **A Case Study on the Use of Volumetric Video in Augmented Reality for Cultural Heritage**. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction (NordiCHI '20)*, Association for Computing Machinery (ACM), Tallinn, Estonia.

O'Dwyer, N., Zerman, E., Young, G. W, Smolic, A., Dunne, S., and Shenton, H. (2021). **Volumetric Video in Augmented Reality Applications for Museological Narratives: A user study for the Long Room in the Library of Trinity College Dublin**. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 14(2), 1-20.

Quantitative Analysis – Visitor Experience

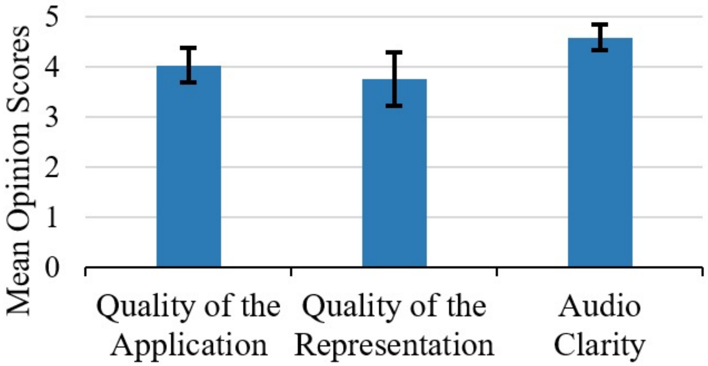


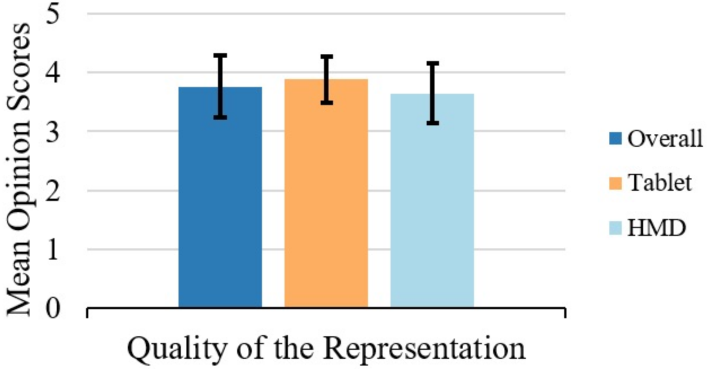
Quality: The overall quality of the application, including the performances of an actor disclosing some historical information.

Format: The type of technology used for information disclosure and how it compares to existing museum technologies like audio guides.

Nature: The whimsical and humorous nature of the story, and its playful mode of delivery, versus formal, pertinent information that is expected in traditional museum settings

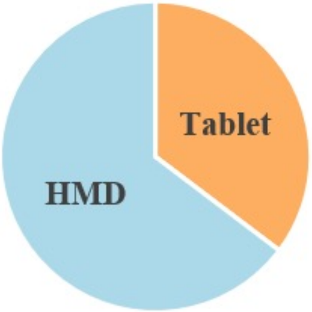
Quantitative Analysis – Technical Aspects





The development quality and the technology were inspected in terms of:

- Overall quality
- Visual quality
- Audio clarity



No differences found between HMD and Tablet

People preferred HMD in terms of immersion even though Tablet had better visual quality.

Qualitative Analysis

Technological Limitations:

- Participant comments highlighted shortcomings of the current state-of-the-art technology

“[it] was difficult to see the full body”

“…the AR only showed parts of him at a time (narrow field of vision), but it’s a prototype”

- Volumetric Video Reconstruction

“…the face was also a bit flat”

“contours of face, e.g. nose, not so successful”

“feet were floating on [the] image of Swift”.

“there was some slight distortion & flickering of the image”

Qualitative Analysis

Immersion:

- The AR application was successful in presenting the participants with an overall quality experience

- Immersion & the nature of delivery

“personal conversation” with a “very lifelike figure of Swift”,

“[Swift] was given a very ‘characteristic’ tone, that is, one feels they’re in front of a person from the past”

- Device-related comments

“Even though the quality of representation on HoloLens was not as good as iPad, the HoloLens was still a more immersive compelling experience”

“HoloLens was a bit heavy”

Trinity College Dublin
The University of Dublin

V-SENSE

Light Fields

Professor Aljosa Smolic
SFI Research Professor of Creative Technologies

Trinity College Dublin
The University of Dublin

V-SENSE

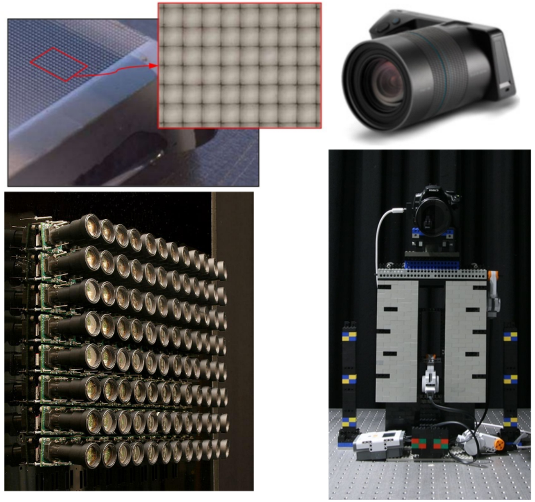
Focus Guided Light Field Saliency Estimation

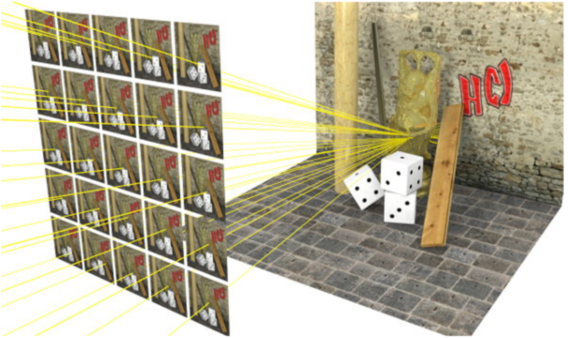
2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

Authors: Ailbhe Gill, Emin Zerman, Martin Alain, Mikael Le Pendu, Aljosa Smolic

Light Fields

- Light rays travelling within a 3D-space captured using a light field camera.
- Can be stored in 2D as an array of **sub-aperture images or** a stack of **refocused images**





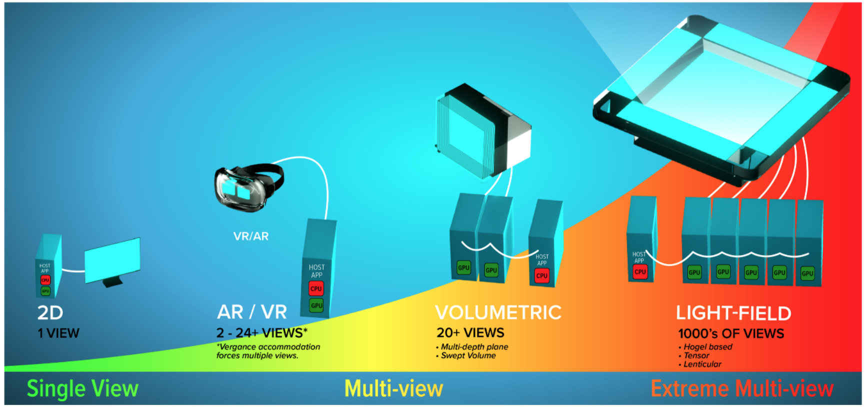
http://lightfield.stanford.edu/lfs.html
https://raytrix.de/

http://www.lightfield-info.com/

Where could visual attention prediction in light fields be of benefit?

The **higher dimensionality** of light fields brings with it **challenges** where visual attention prediction could be of benefit:

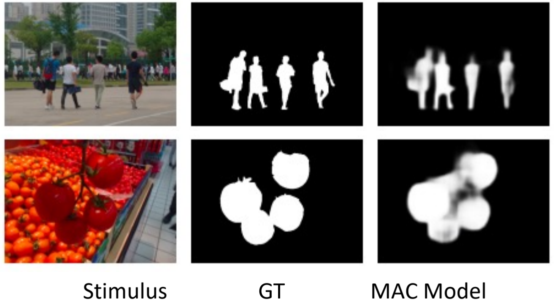
- Rendering for displays
- Compression
- Quality assessment

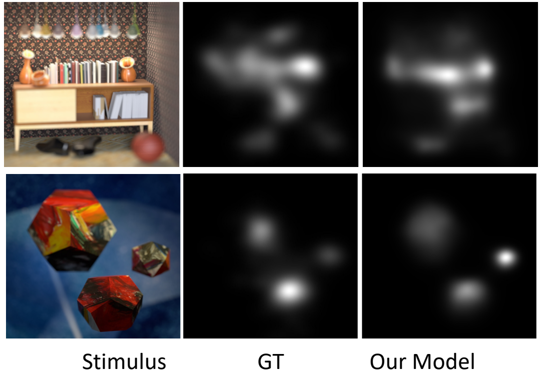


http://www.fovi3d.com/technology#light-field-display-multiview-computation

Light Field Saliency Estimation

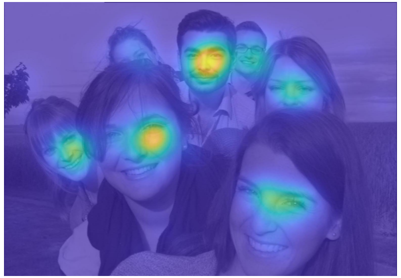
	Previous Work	Our Work
Saliency Estimation	Salient Object Detection	Visual Attention Prediction
Ground Truth	**Binary map** obtained by human segmentation of all-in-focus images	**Probability map** obtained by tracking eye fixations of various renderings
Aim	Detect and **segment semantic objects** that stand out in the scene	**Predict all regions of visual interest** and the extent by which they attract visual attention
Output	**2D saliency map** with salient object segmented	**Four-dimensional saliency field and focus map** used to generate saliency map of any LF rendering





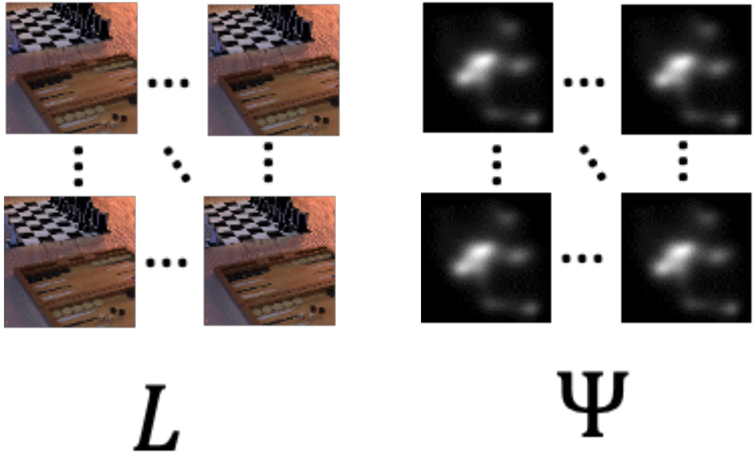
Zhang, Jun, et al. "**Light field saliency detection with deep convolutional networks.**" *IEEE Transactions on Image Processing* 29 (2020): 4421-4434.
Piao, Yongri, et al. "**Exploit and Replace: An Asymmetrical Two-Stream Architecture for Versatile Light Field Saliency Detection.**" AAAI. 2020.

Saliency Field: Ψ



- A saliency map assigns a probability of visual importance to every pixel of an image.

- **Light field saliency should assign a probability of visual importance to every ray of a light field**



Gill, Ailbhe; Zerman, Emin; Alain, Martin; Le Pendu, Mikael; Smolic, Aljosa
Focus Guided Light Field Saliency Estimation Inproceedings
In: QoMEX, IEEE 2021.

Rendered Light Field Dataset

How to choose a Data Set?

Light Fields chosen:

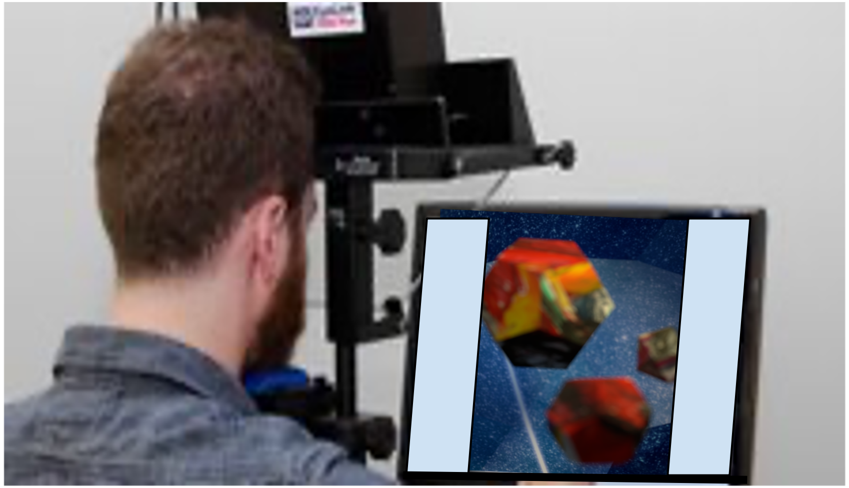
- contained multiple regions or objects with high colour contrast
- contained regions with great edge density and local contrast at varied depths and spatial locations.
- were acquired using different methods/ camera types

https://v-sense.scss.tcd.ie/research/light-fields/visual-attention-for-light-fields/

Eye-tracking Data Collection

Experiment

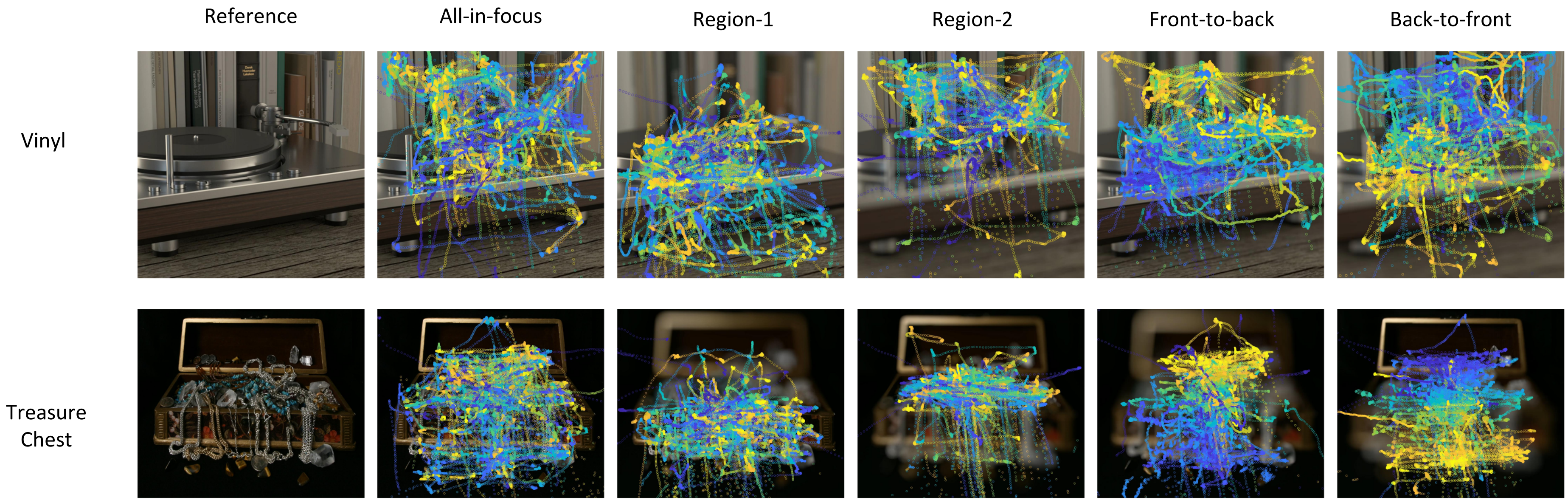
- Used the SR Research Eyelink 1000 plus
- All-in-focus videos played to avoid first time viewing bias
- Calibration step
- Full set of videos displayed in randomised order to each participant
- **Eye fixation data of 21 participants was recorded for five renderings of 20 light fields**



SR Research Eyelink, https://www.sr-research.com/eyelink-1000-plus/

Results

Scanpaths



Reference

All-in-focus

Region-1

Region-2

Front-to-back

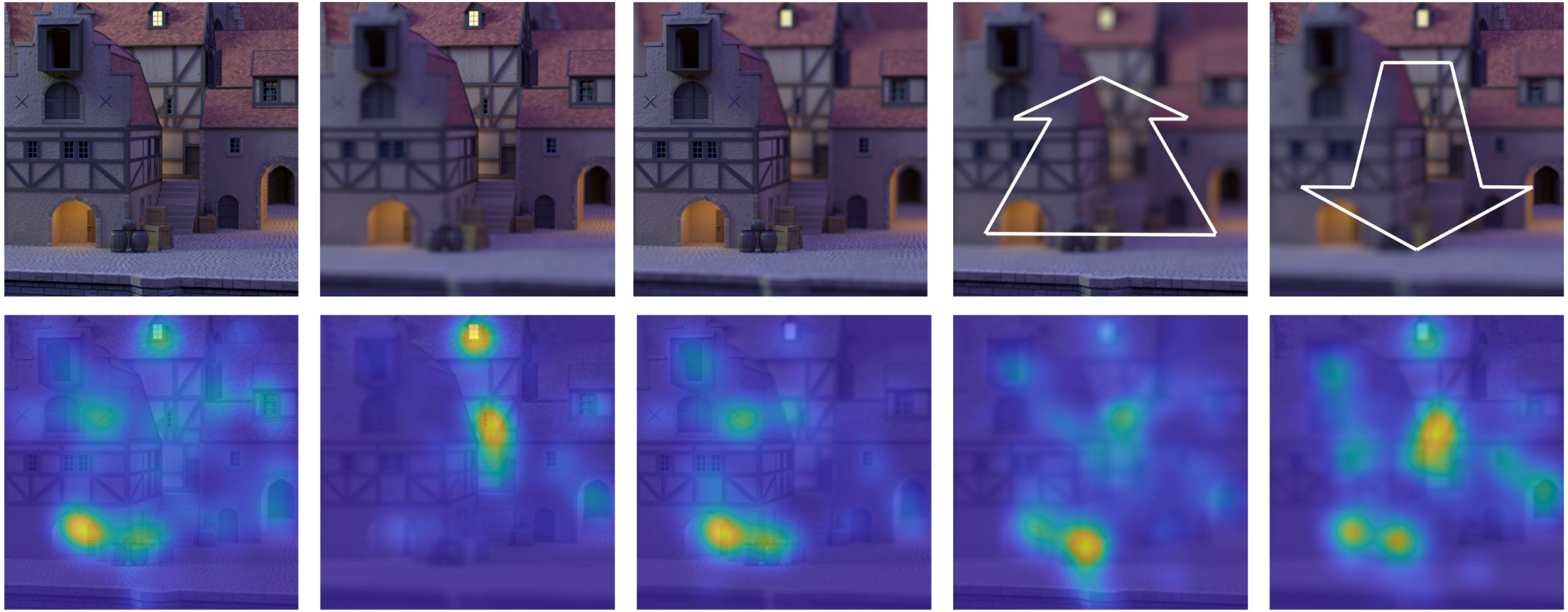
Back-to-front

Vinyl

Treasure Chest

Results

Average Saliency Maps



Medieval

All-in-focus

Region-1

Region-2

Front-to-back

Back-to-front

Reference

Saliency Map

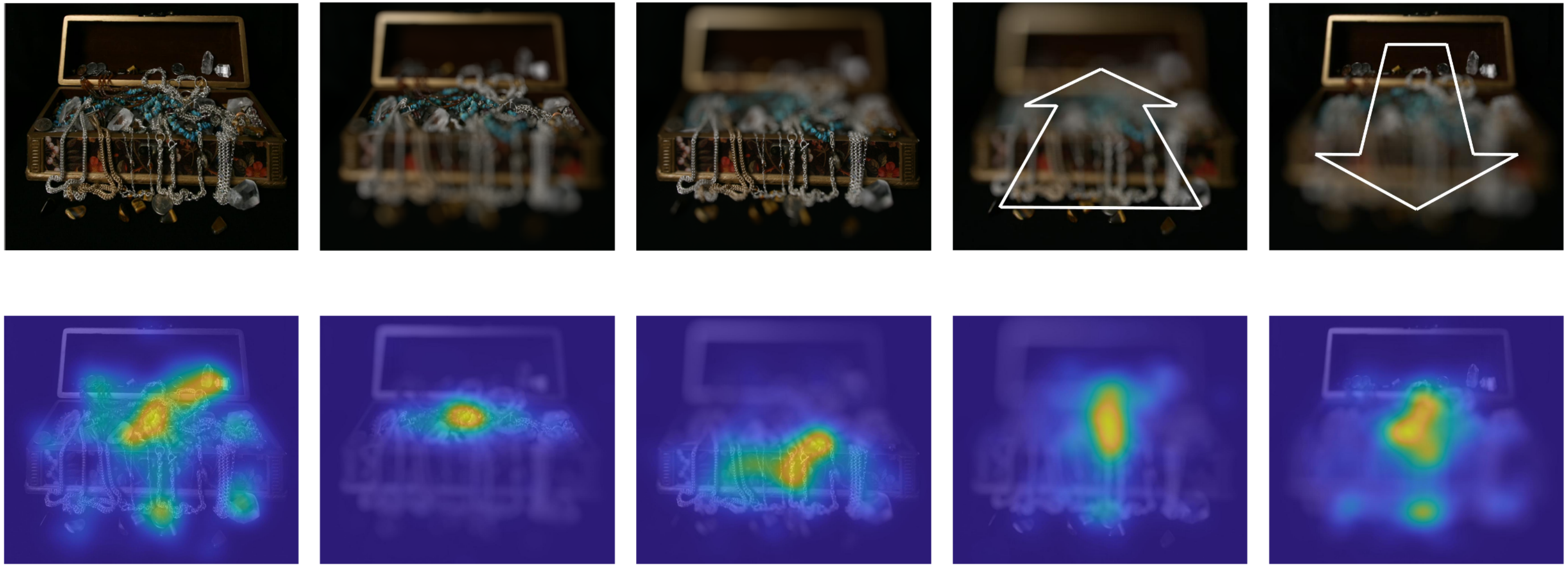
Results

Average Saliency Maps

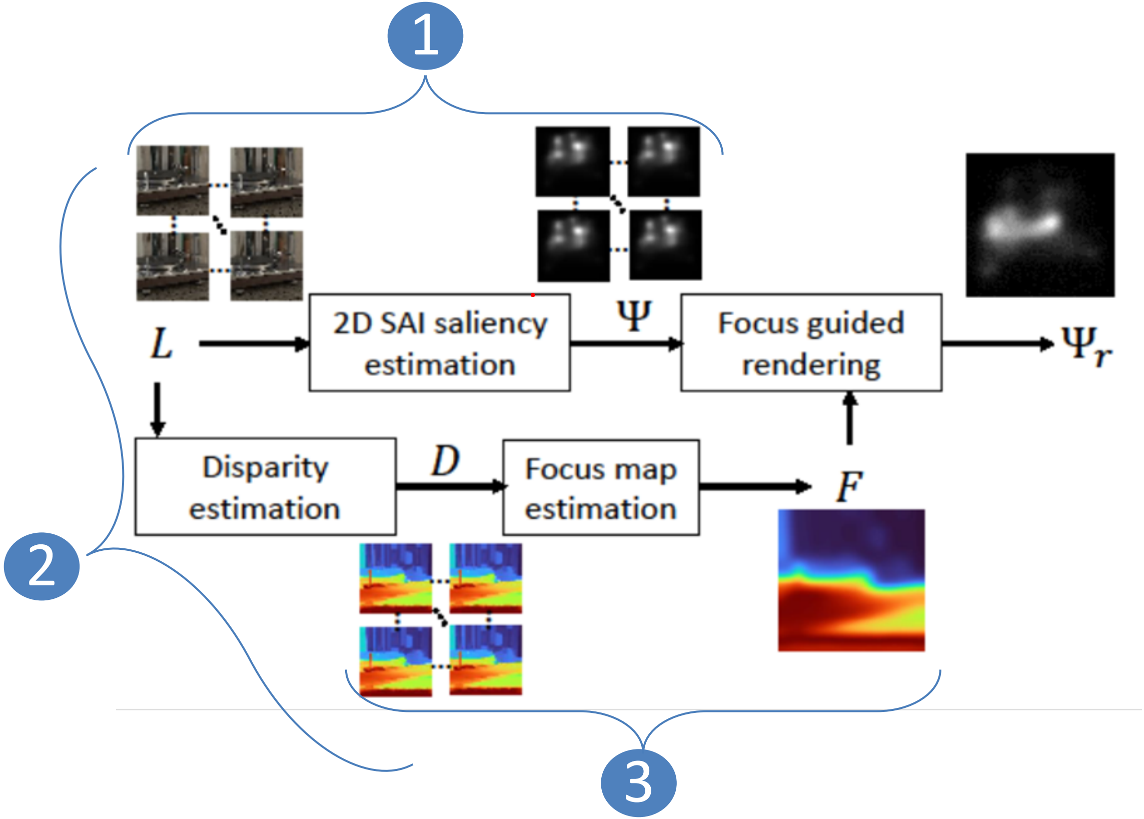
Treasure Chest | All-in-focus | Region-1 | Region-2 | Front-to-back | Back-to-front

Reference

Saliency Map



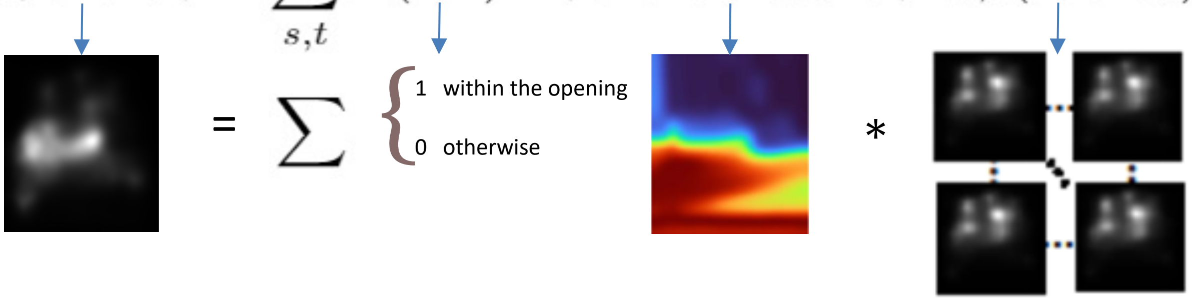
Focus Guided Saliency Estimation Pipeline



Focus Guided Rendering

- The shift and sum algorithm is used to weight the saliency sub-aperture images (SAIs) with the focus map.

$$\Psi_r(u, v, \delta_F) = \sum_{s,t} A(s,\, t)\, F(s, t, u_F,\, v_F,\, \delta_F)\Psi_{s,t}(u_F,\, v_F)$$



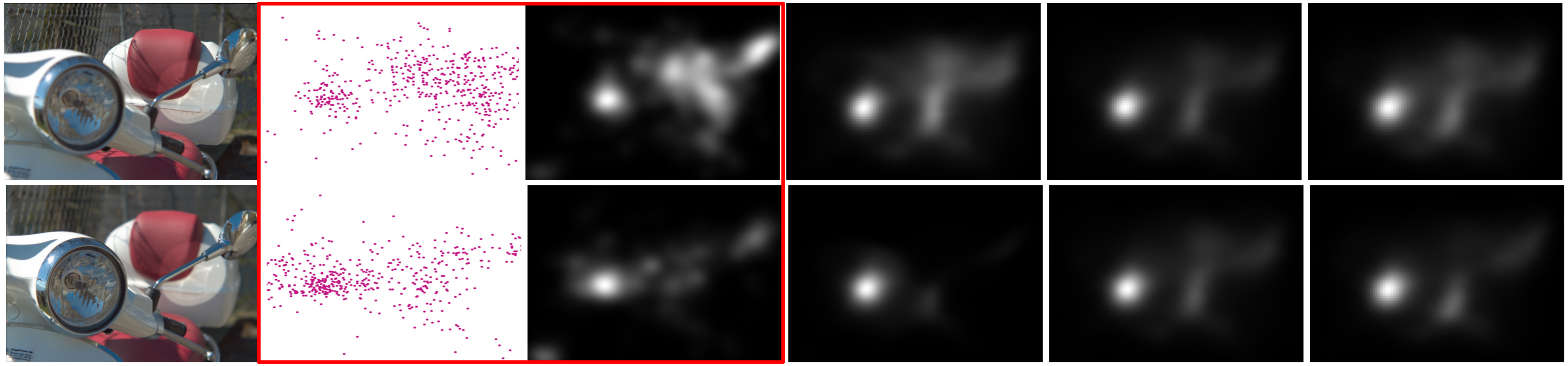
$$= \sum \begin{cases} 1 & \text{within the opening} \\ 0 & \text{otherwise} \end{cases} \quad *$$

- The algorithm can be simplified to:

$$\Psi_r(u, v, \delta_F) = F_r(u, v, \delta_F) \sum_{s,t} A(s,\, t)\, \Psi_{s,t}(u_F,\, v_F)$$

Qualitative Results – Vespa

Ground truth data



Region-1

Region-2

Stimulus

GT Fixations

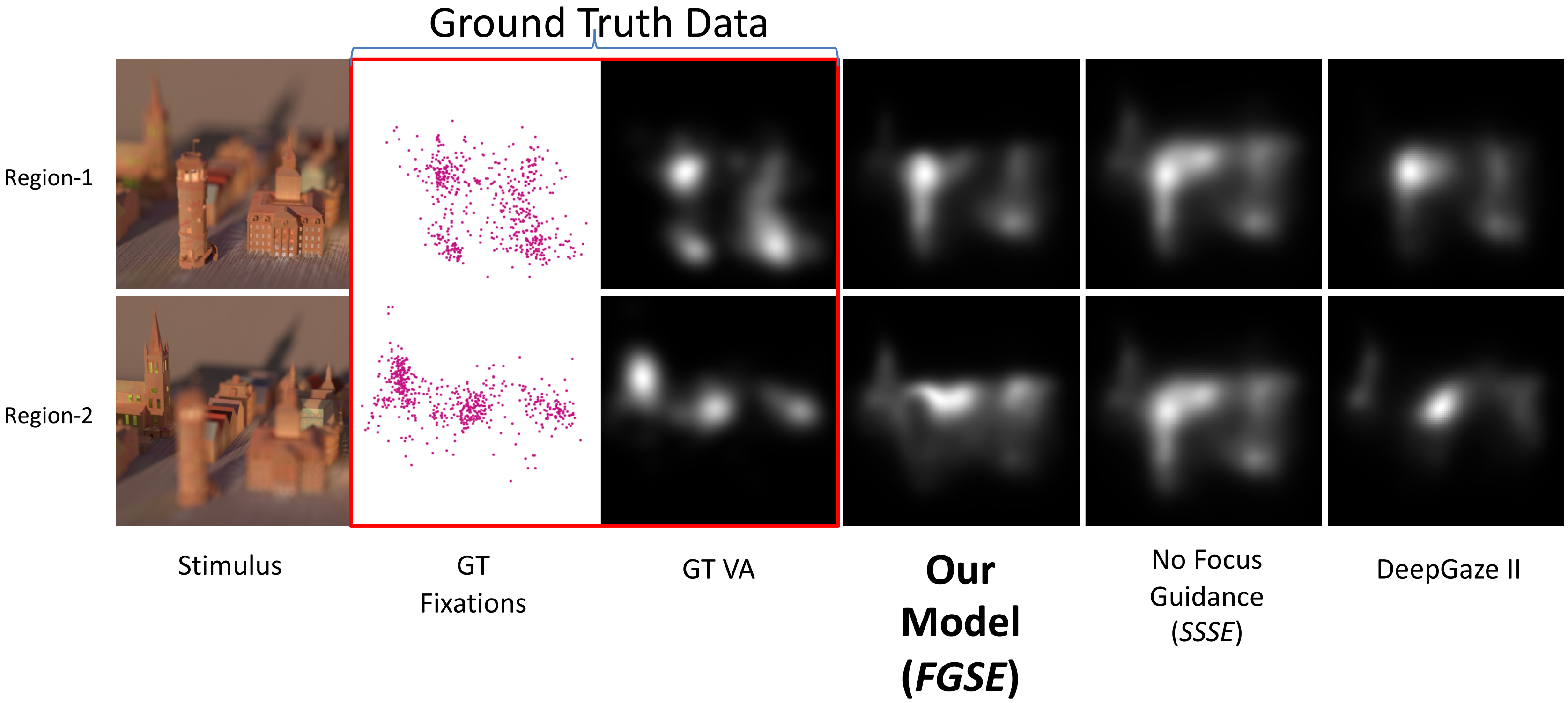
GT VA

Our Model (*FGSE*)

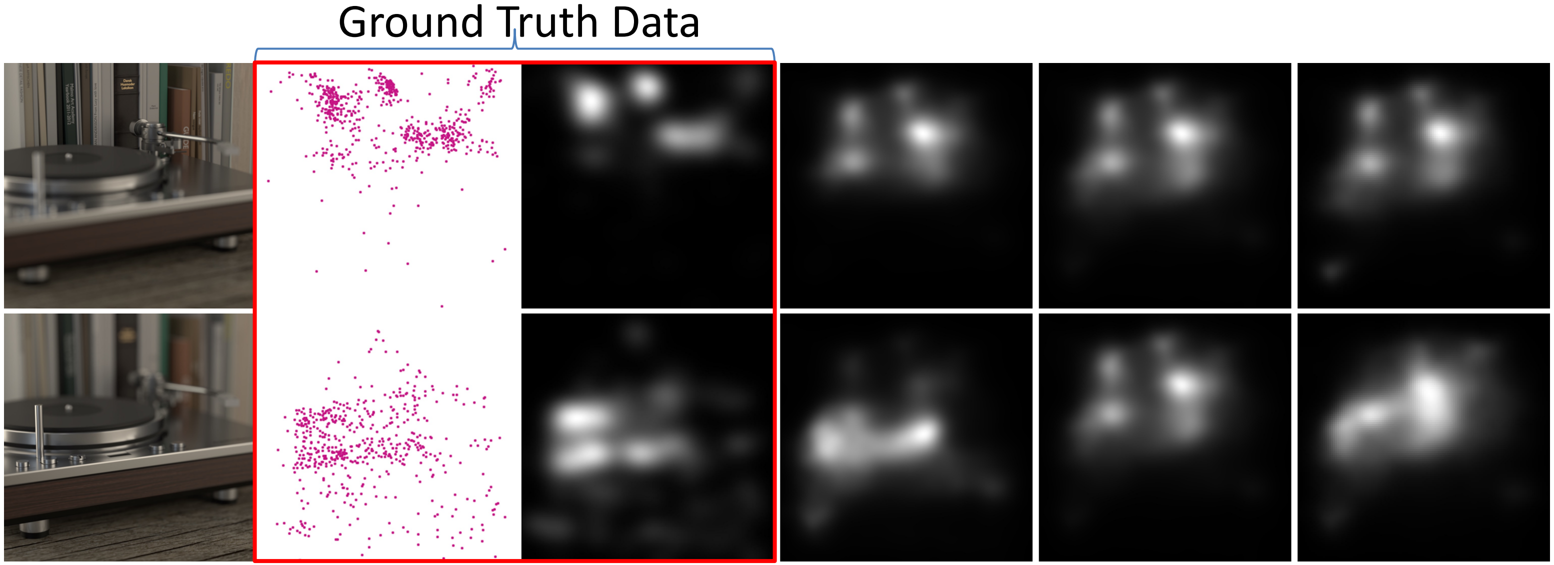
No Focus Guidance (*SSSE*)

DeepGaze II

Qualitative Results – Tower



Qualitative Results – Vinyl



Ground Truth Data

Region-1

Region-2

Stimulus

GT Fixations

GT VA

Our Model (*FGSE*)

No Focus Guidance (*SSSE*)

DeepGaze II

Quantitative Results

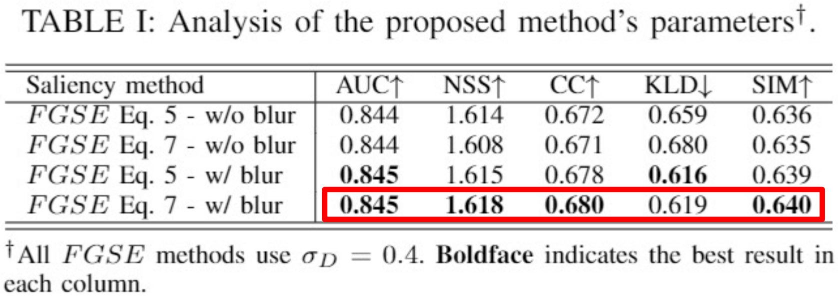


TABLE I: Analysis of the proposed method's parameters[†].

Saliency method	AUC↑	NSS↑	CC↑	KLD↓	SIM↑
$FGSE$ Eq. 5 - w/o blur	0.844	1.614	0.672	0.659	0.636
$FGSE$ Eq. 7 - w/o blur	0.844	1.608	0.671	0.680	0.635
$FGSE$ Eq. 5 - w/ blur	**0.845**	1.615	0.678	**0.616**	0.639
$FGSE$ Eq. 7 - w/ blur	**0.845**	**1.618**	**0.680**	0.619	**0.640**

[†] All $FGSE$ methods use $\sigma_D = 0.4$. **Boldface** indicates the best result in each column.

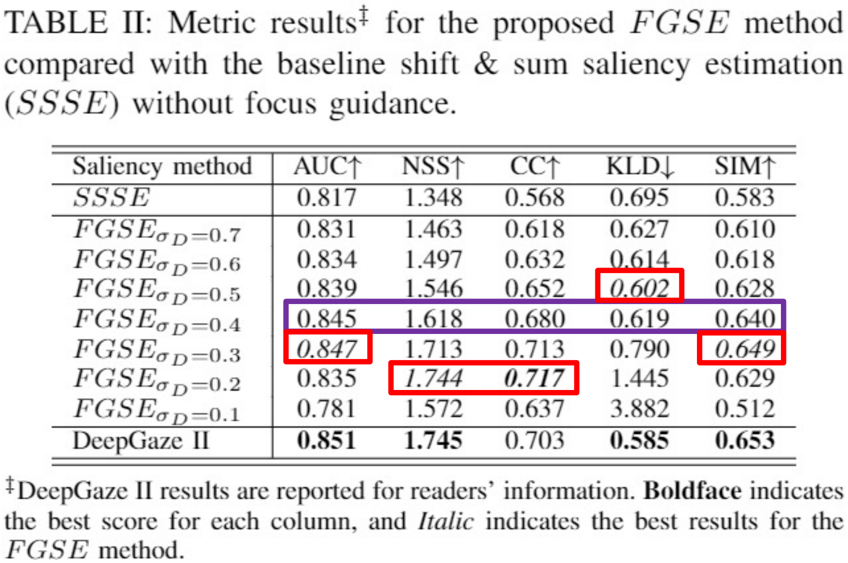


TABLE II: Metric results[‡] for the proposed $FGSE$ method compared with the baseline shift & sum saliency estimation ($SSSE$) without focus guidance.

Saliency method	AUC↑	NSS↑	CC↑	KLD↓	SIM↑
$SSSE$	0.817	1.348	0.568	0.695	0.583
$FGSE_{\sigma_D=0.7}$	0.831	1.463	0.618	0.627	0.610
$FGSE_{\sigma_D=0.6}$	0.834	1.497	0.632	0.614	0.618
$FGSE_{\sigma_D=0.5}$	0.839	1.546	0.652	*0.602*	0.628
$FGSE_{\sigma_D=0.4}$	0.845	1.618	0.680	0.619	0.640
$FGSE_{\sigma_D=0.3}$	*0.847*	1.713	0.713	0.790	*0.649*
$FGSE_{\sigma_D=0.2}$	0.835	*1.744*	***0.717***	1.445	0.629
$FGSE_{\sigma_D=0.1}$	0.781	1.572	0.637	3.882	0.512
DeepGaze II	**0.851**	**1.745**	0.703	**0.585**	**0.653**

[‡] DeepGaze II results are reported for readers' information. **Boldface** indicates the best score for each column, and *Italic* indicates the best results for the $FGSE$ method.

Trinity College Dublin
The University of Dublin

V-SENSE

Many Thanks
smolica@tcd.ie