Trinity College Dublin

The University of Dublin

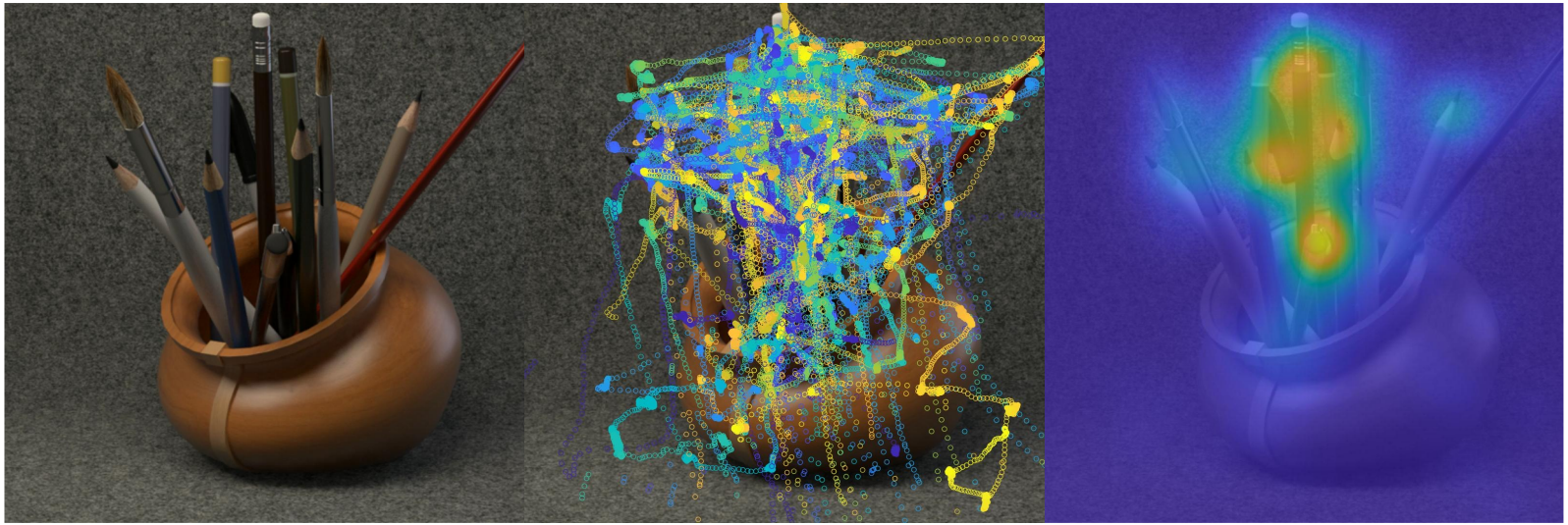
V-SENSE

Perception and Quality of Immersive Media

Professor Aljosa Smolic
SFI Research Professor of Creative Technologies

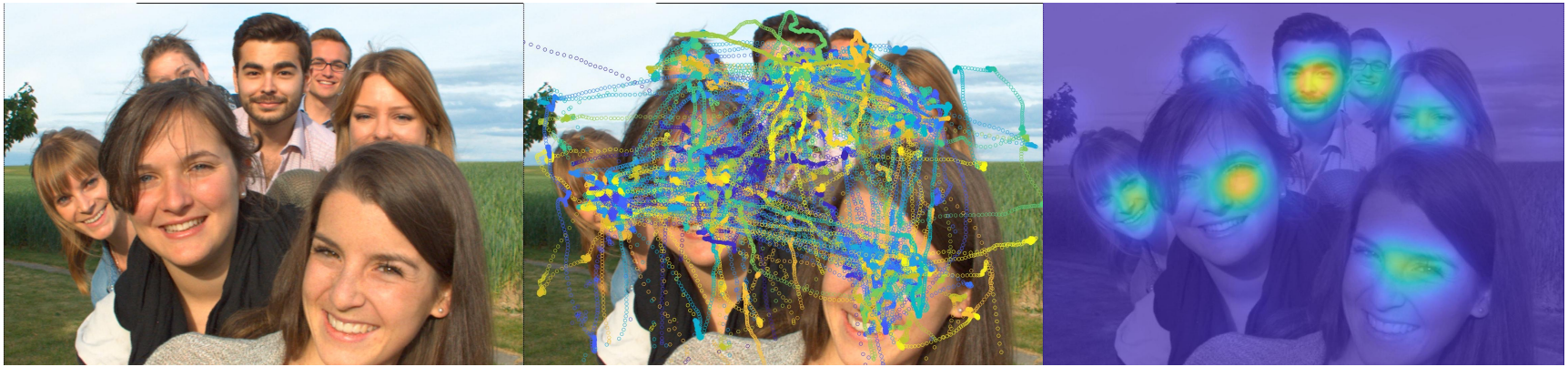
Visual Attention

Where people look when viewing a visual scene.



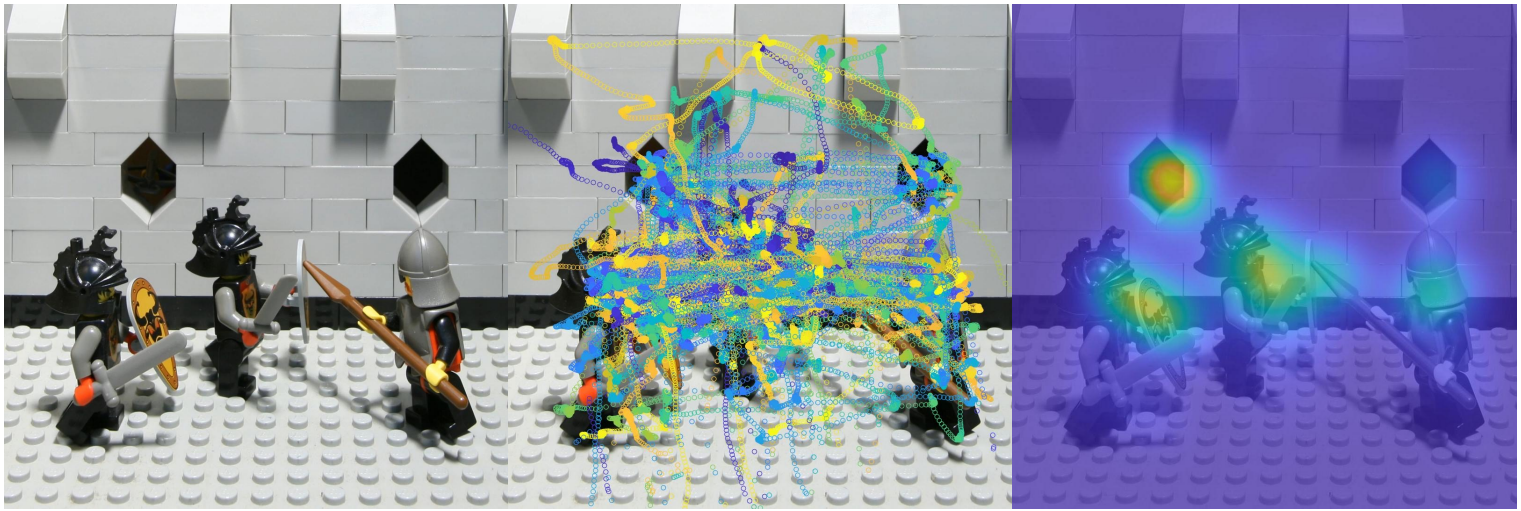
Visual Attention

Where people look when viewing a visual scene.



Visual Attention

Where people look when viewing a visual scene.



Purpose Visual Attention and Saliency

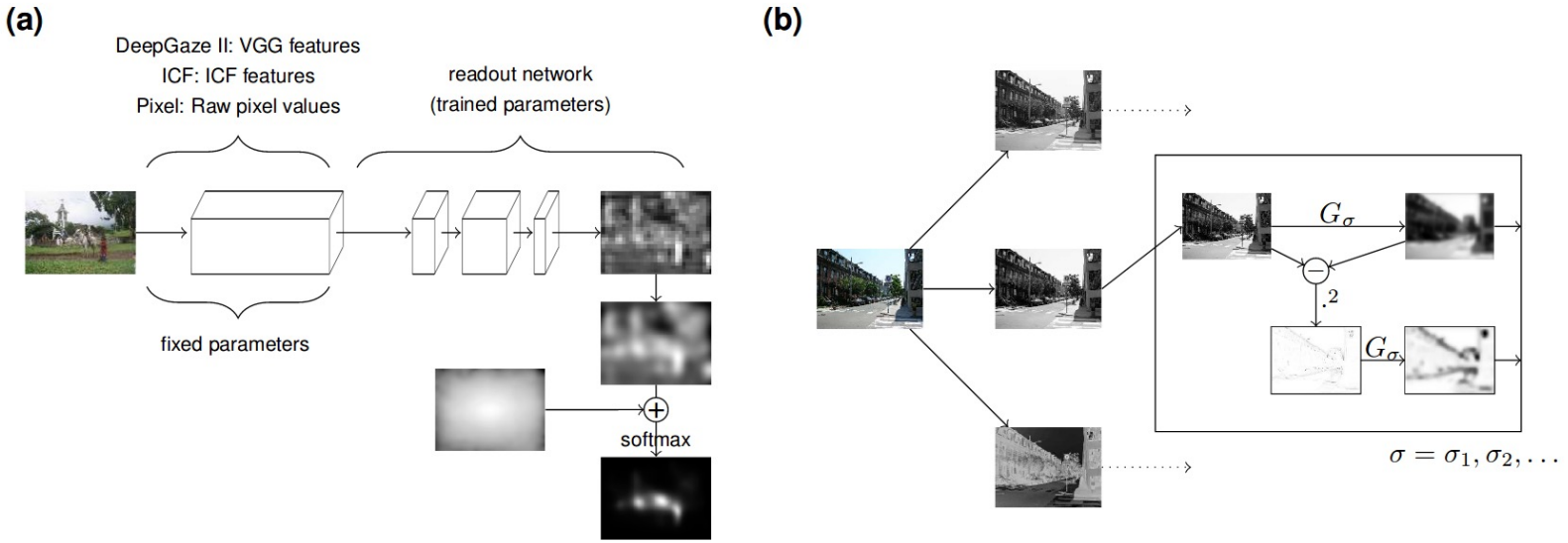
- **Understanding human perception**
- **Analysing/evaluating properties of the content**
- **Assigning resources to important parts of the content**
  - Coding/compression, streaming
  - Rendering
- **Quality assessment**
- **Optimizing algorithms driven by perceptual priority**

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

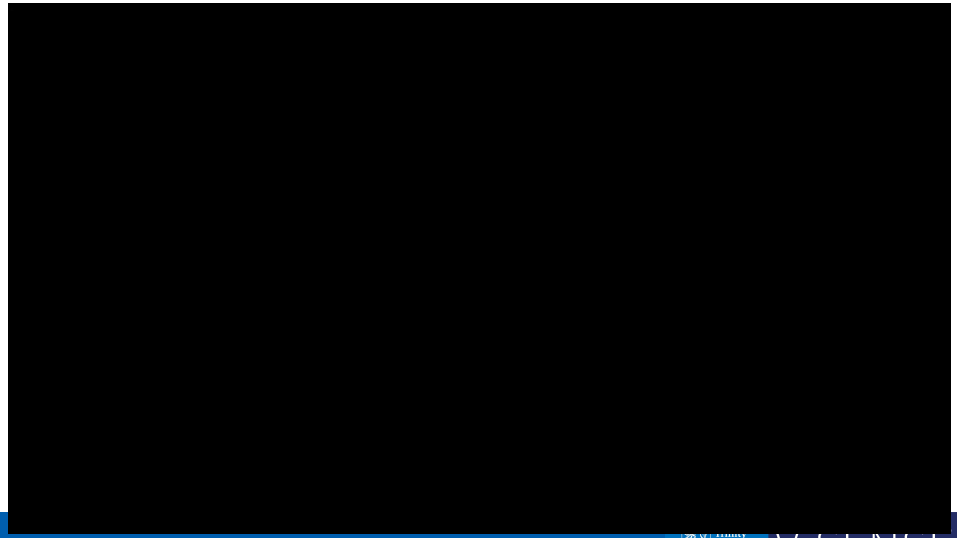
Viewing characteristics: free look around in 6DoF

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

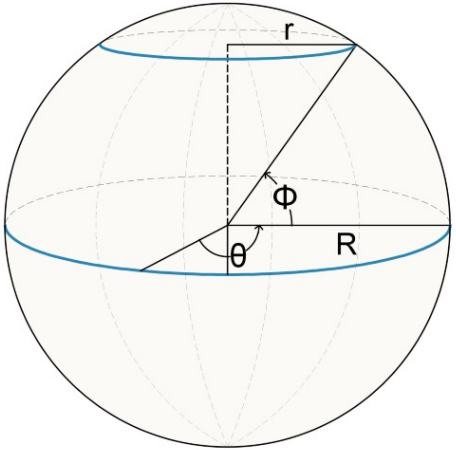
The University of Dublin

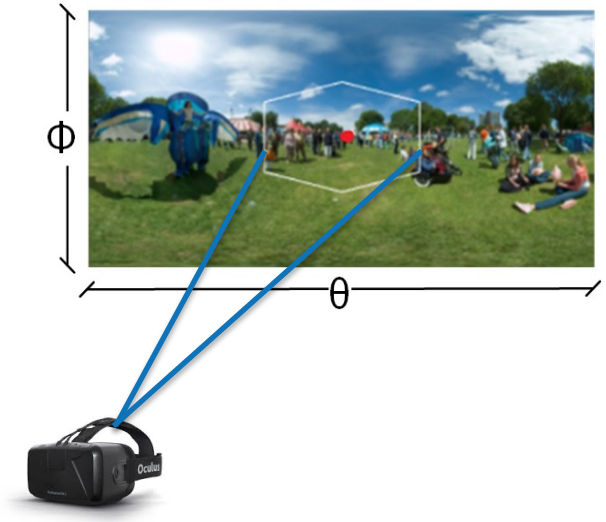
V-SENSE

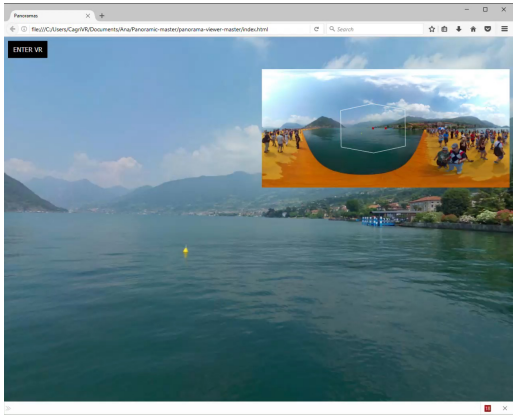
Omnidirectional Video – 3DoF

Professor Aljosa Smolic
SFI Research Professor of Creative Technologies

Omnidirectional Images (ODIs) in VR







- 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

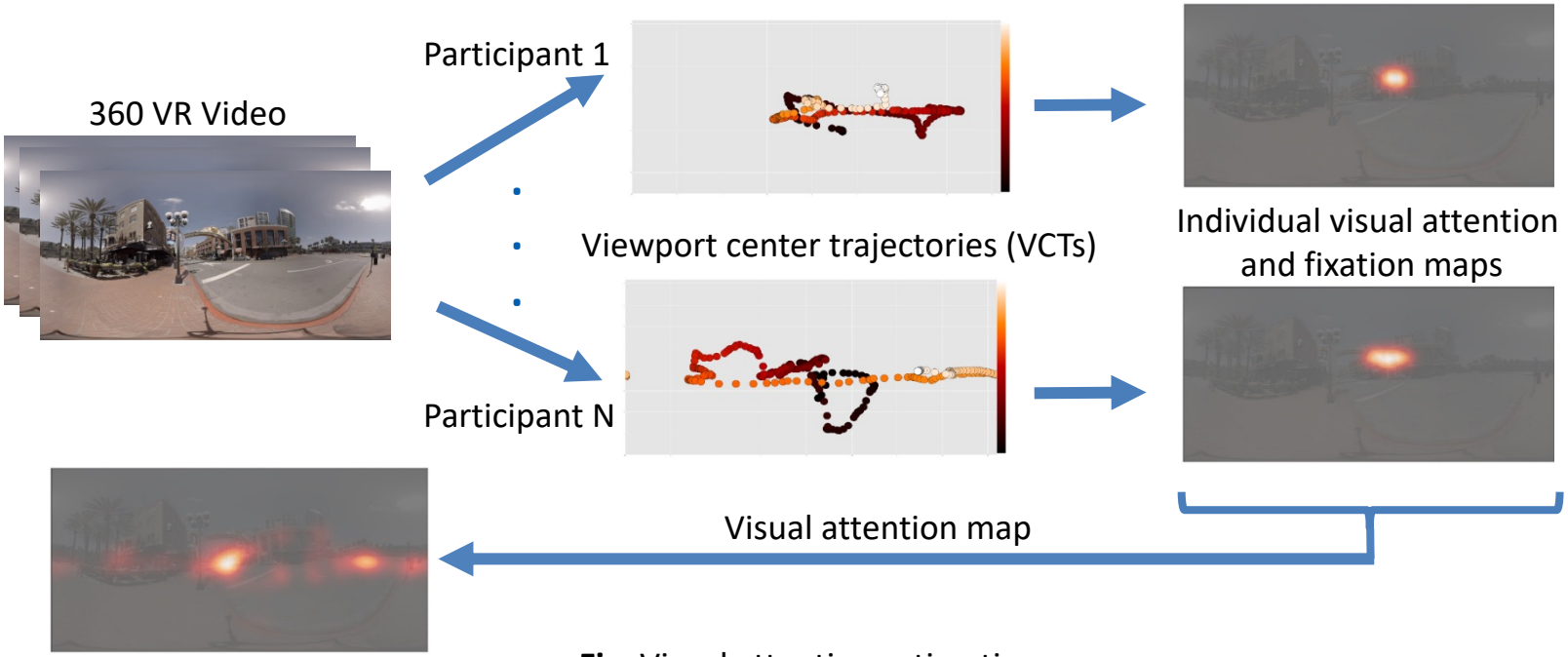


Fig. Visual attention estimation.

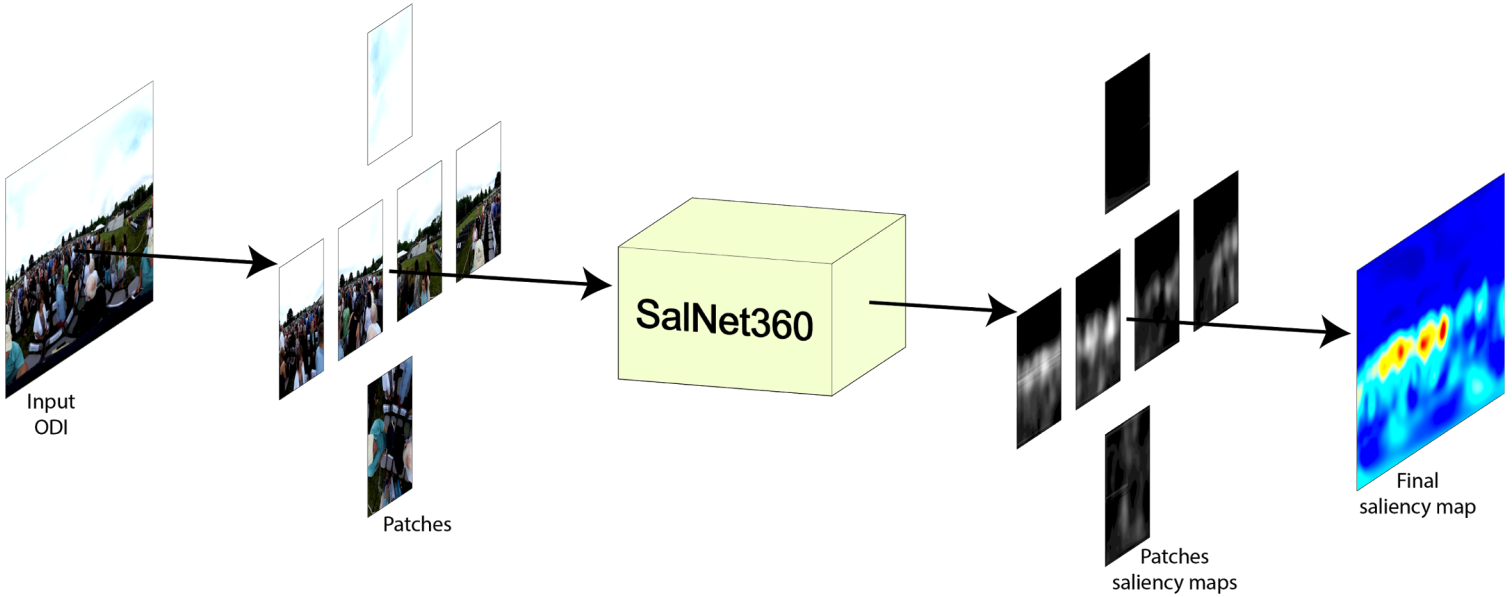
- Abreu, Ana De; Ozcinar, Cagri; Smolic, Aljosa; "Look around you: saliency maps for omnidirectional images in VR applications," **9th IEEE International Conference on Quality of Multimedia Experience (QoMEX), 2017.**
- Ozcinar, Cagri; Smolic, Aljosa; "Visual Attention in Omnidirectional Video for Virtual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

SalNet360


Input
ODI
Patches
SalNet360
Patches
saliency maps
Final
saliency map

Modelling of Visual Attention

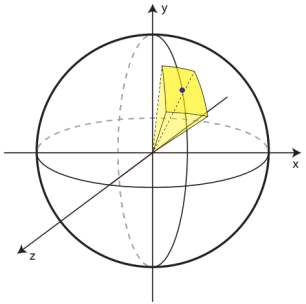


Fig. Sliding frustum used to create multiple patches.

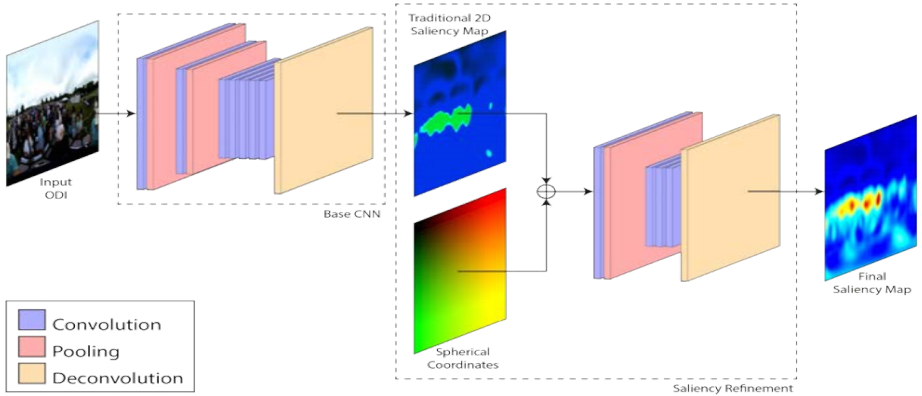
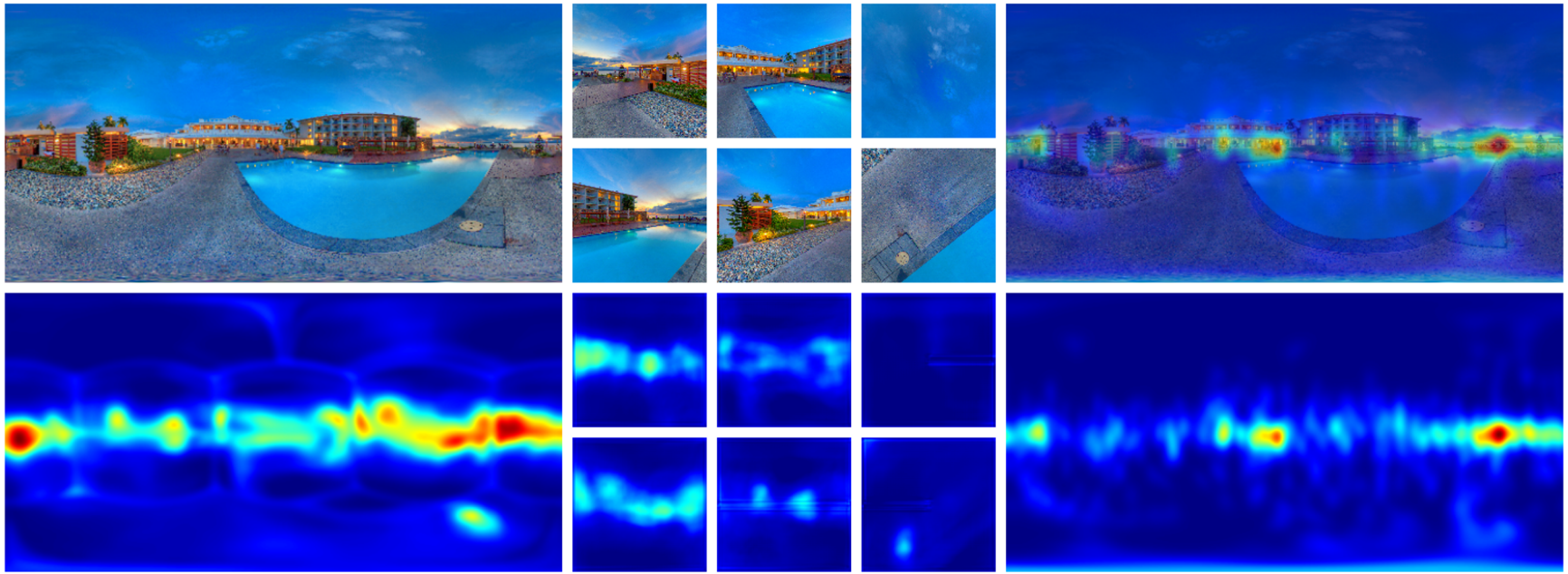


Fig. Network architecture of the SalNet360.

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

Results



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

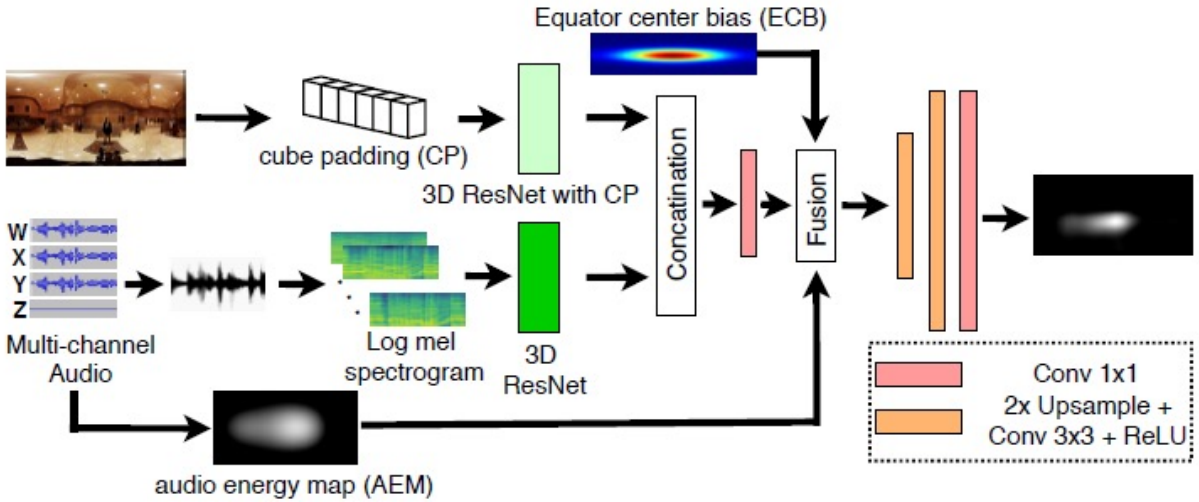
[1]Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
[2]V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland

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

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

AVS360

- **Network architecture**



Chao, Fang-Yi; Ozcinar, Cagri; Wang, Chen; Zerman, Emin; Zhang, Lu; Hamidouche, Wassim; Deforges, Olivier; Smolic, Aljosa
Audio-Visual Perception of Omnidirectional Video for Virtual Reality Applications
2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2020, ISBN: 978-1-7281-1486-6.

RESULTS

- Comparison to the state of the arts

Cat.	Models	*mute* NSS	*mute* CC	*mono* NSS	*mono* CC	*ambisonics* NSS	*ambisonics* 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	1.40	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	1.33	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).

Trinity College Dublin

The University of Dublin

V-SENSE

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

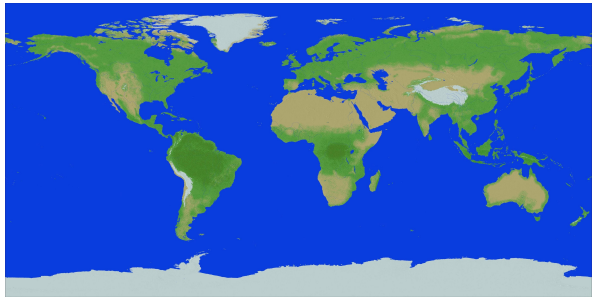
Simone Croci, Emin Zerman, and Aljosa Smolic

Unique Aspects of ODV

1. Spherical nature but stored in planar representations

360°

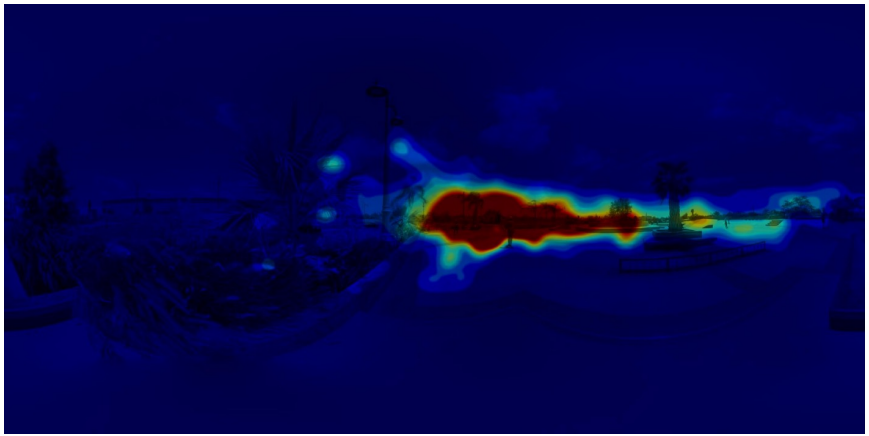
Projection



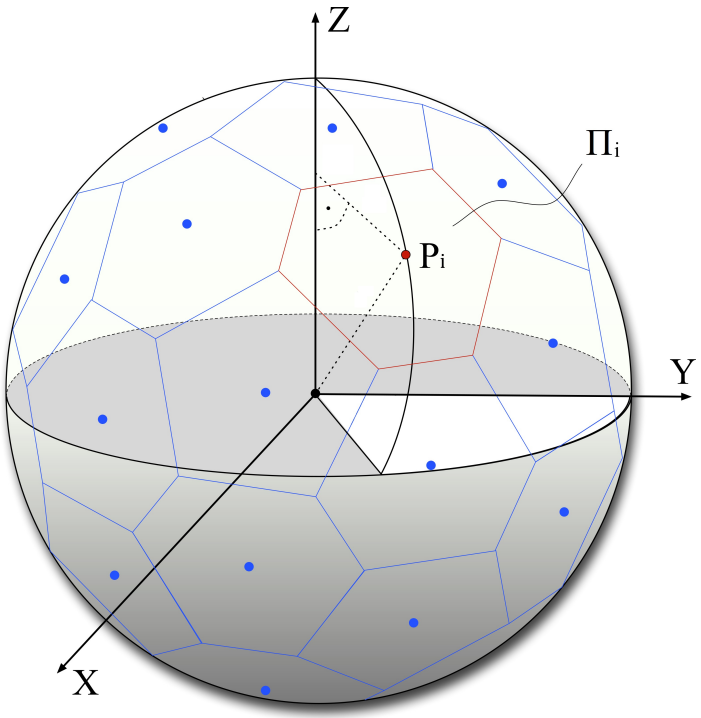
Unique Aspects of ODV

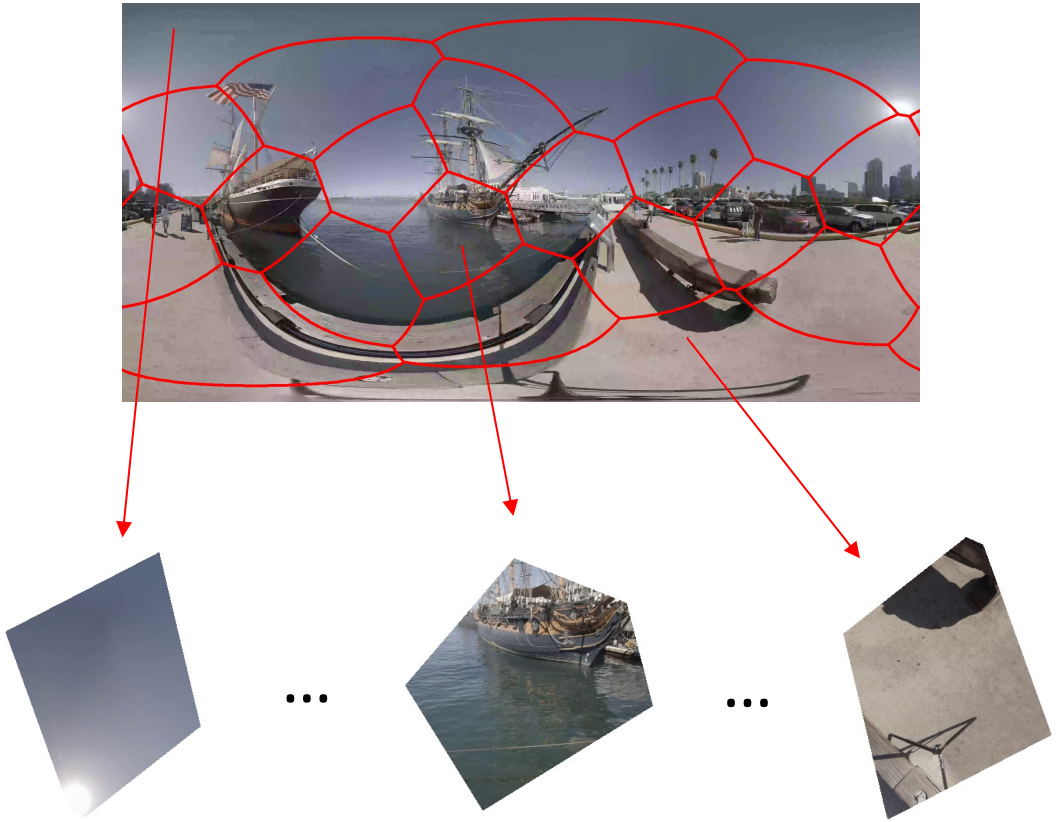
2. Viewing characteristics: free look around, only viewport

Visual Attention

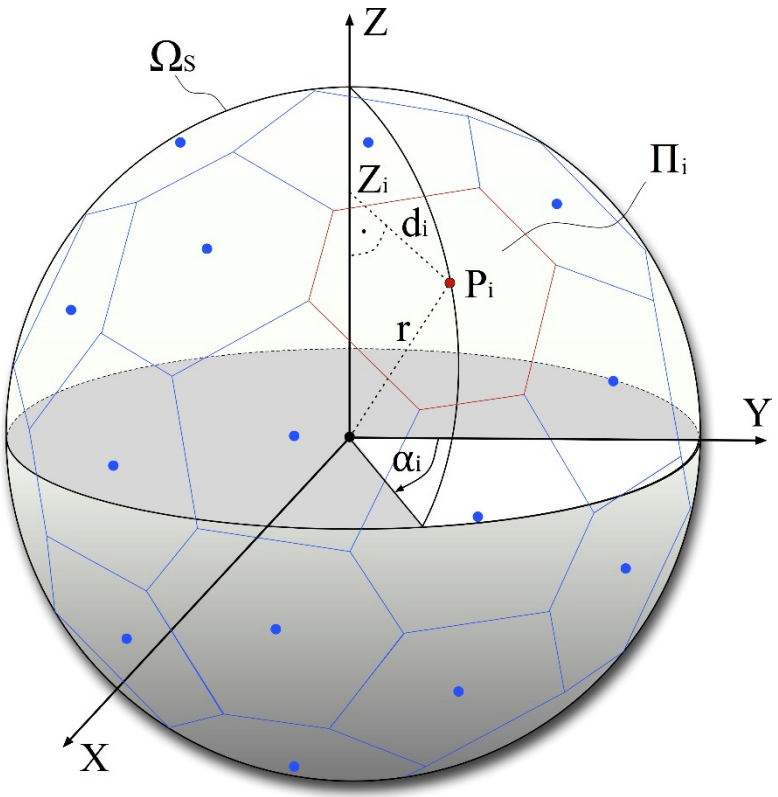


Voronoi Patch Extraction





Voronoi Patch Extraction



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

$$\alpha_i = i\pi \cdot \left(3 - \sqrt{5}\right)$$

$$Z_i = \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2i}{N-1}\right)$$

$$d_i = \sqrt{1 - Z_i^2}$$

$$X_i = d_i \cdot cos(\alpha_i)$$

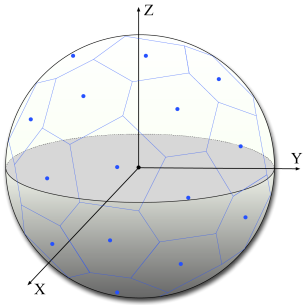
$$Y_i = d_i \cdot sin(\alpha_i)$$

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

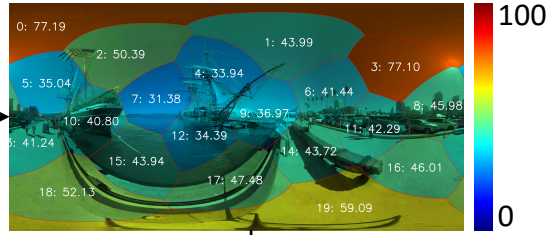
Spherical Voronoi Diagram

2D Video Metrics

PSNR, SSIM, MS-SSIM, VMAF, ...

VI-PSNR, VI-SSIM, VI-MS-SSIM, VI-VMAF, ...

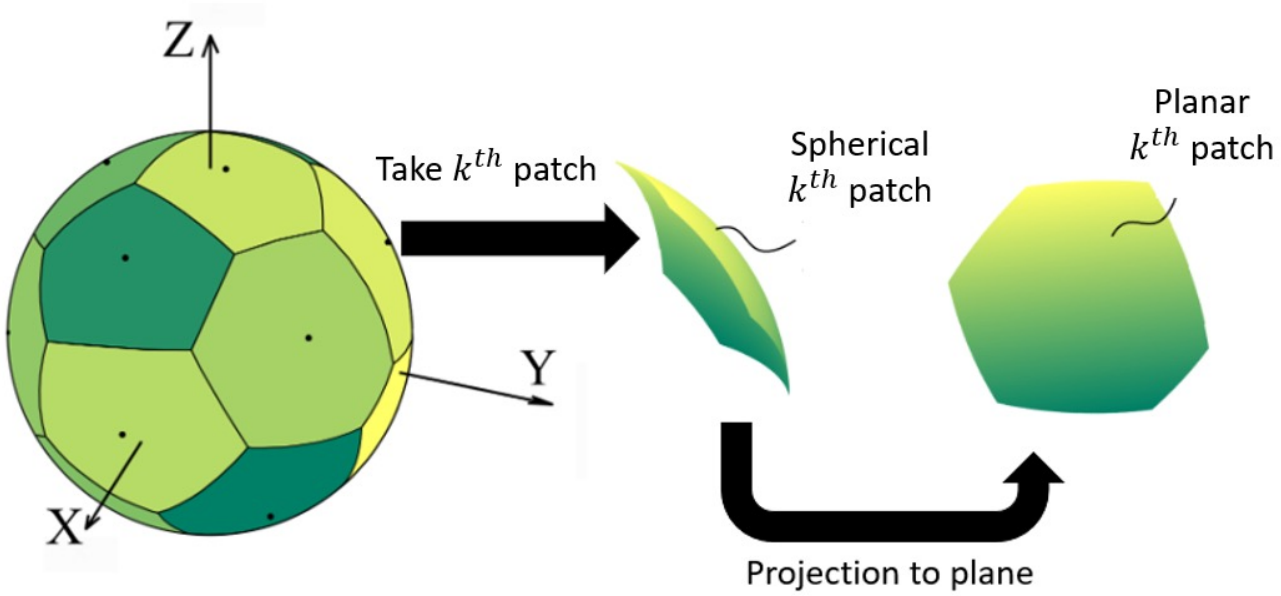
Patch Scores S_i

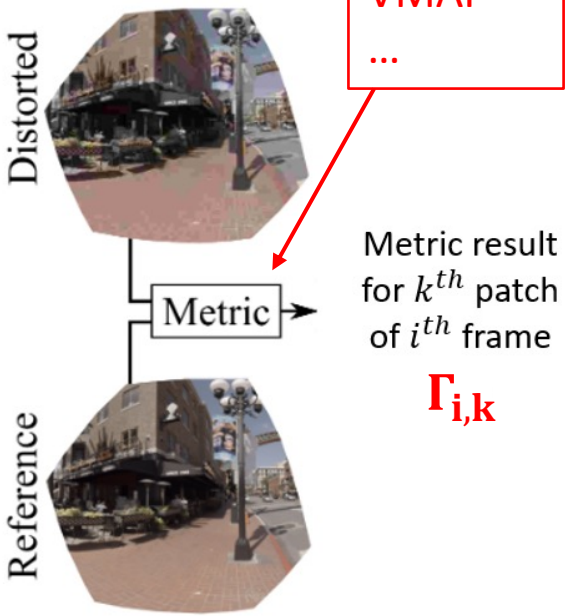


Arithmetic Mean $\frac{1}{N}\sum_{i=1}^{N} S_i$

Final Score

VI-VA-METRIC Framework





VI-VA-METRIC Framework

Score of frame *i*:

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

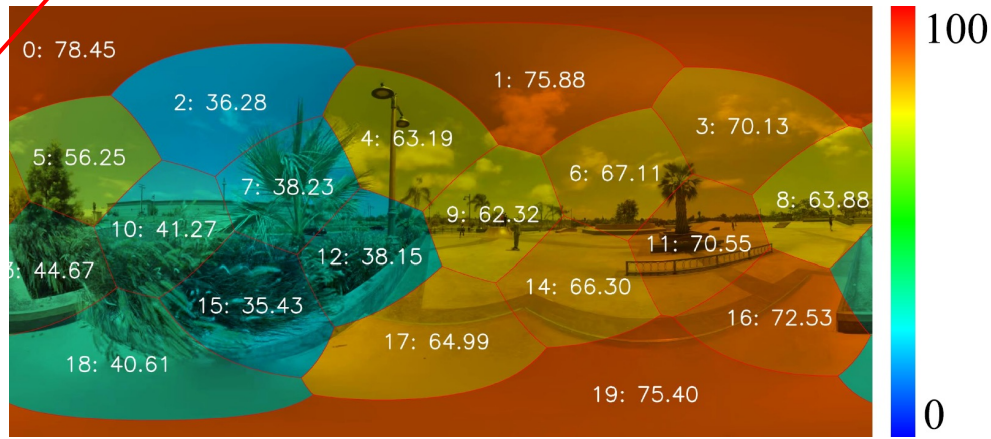
$$T_i' = \frac{\sum_{k=1}^{M} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} \nu_{i,k}}$$

$\Gamma_{i,k}$ 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*:

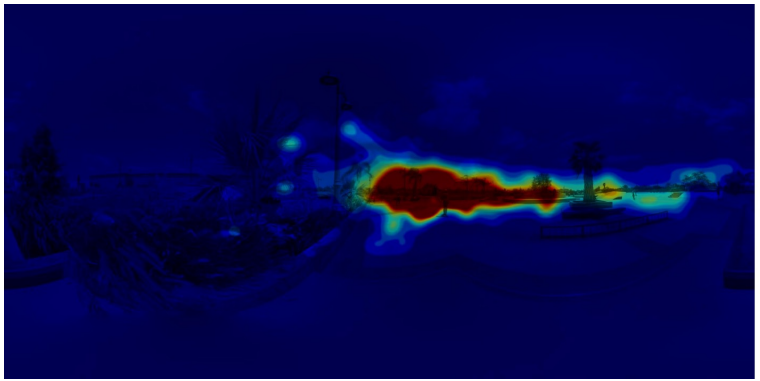
$$T_i = \frac{\sum_{k=1}^{M} \Gamma_{i,k}}{M}$$

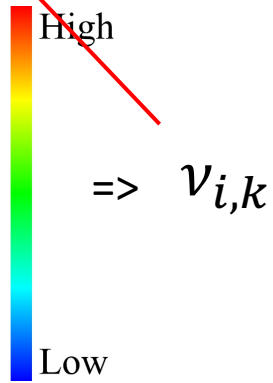
$$T_i' = \frac{\sum_{k=1}^{M} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} \nu_{i,k}}$$

$\Gamma_{i,k}$ Patch score

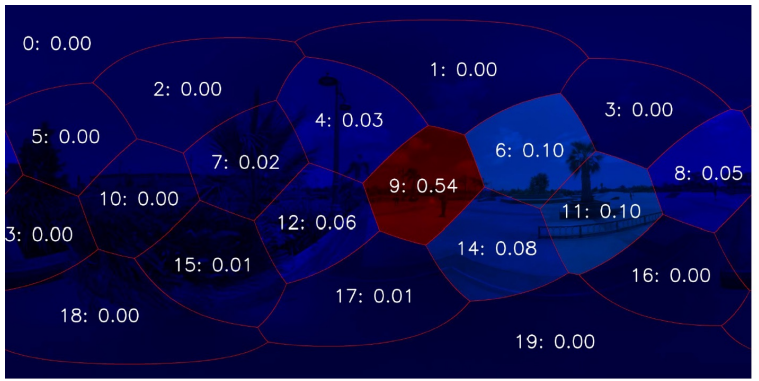
$\nu_{i,k}$ Visual attention weight

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





=> $\nu_{i,k}$



VI-VA-METRIC Framework

Final score from temporal pooling of frame scores

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

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

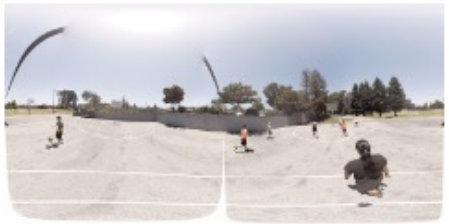
P: arithmetic mean, harmonic mean, min, median, p-th percentile, …

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)

Trinity College Dublin

The University of Dublin

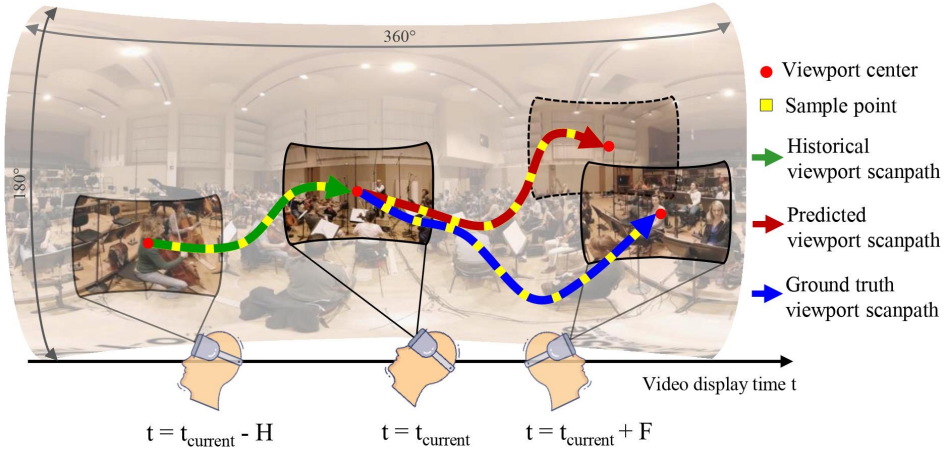
V-SENSE

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

Fang-Yi Chao, Cagri Ozcinar, Aljosa Smolic

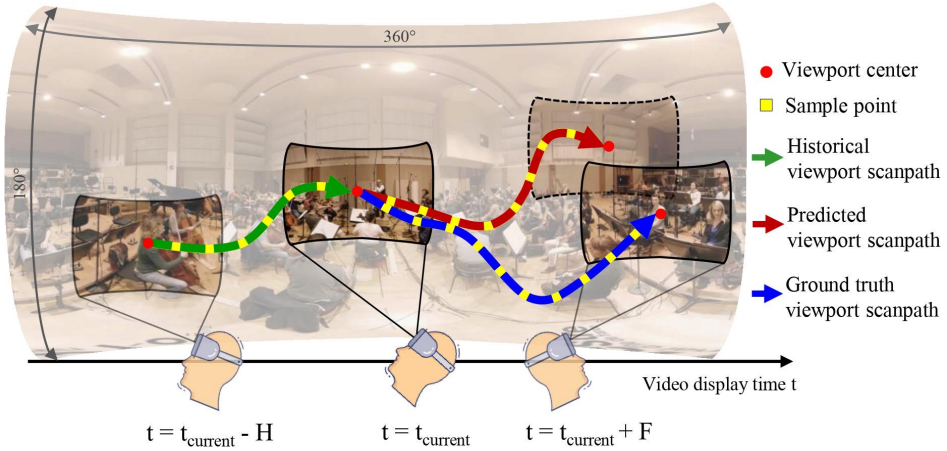
MMSP, October 2021

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

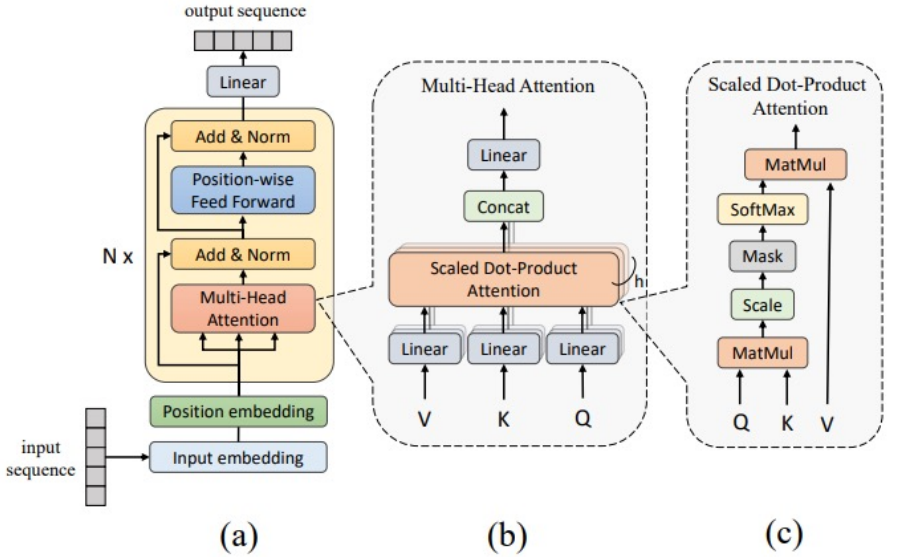


- We can formalize the problem as finding the best model f_F^* :

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

- 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 $t \in [t+1, t+F]$.

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.

Results

Comparison with the state of the arts

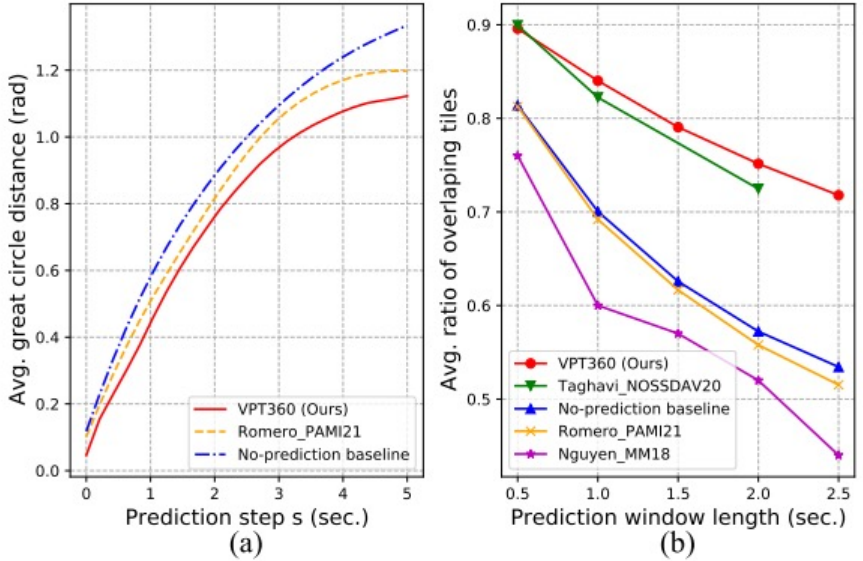


Fig. 4: Comparison results on (a) David_MMSys18 and (b) Wu_MMSys17 dataset, respectively.

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**

Trinity College Dublin

The University of Dublin

V-SENSE

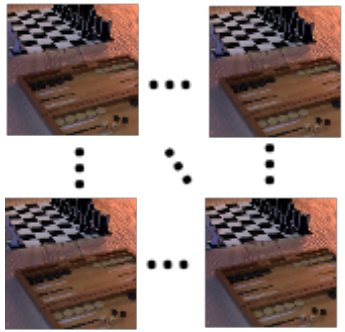
Focus Guided Light Field Saliency Estimation

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

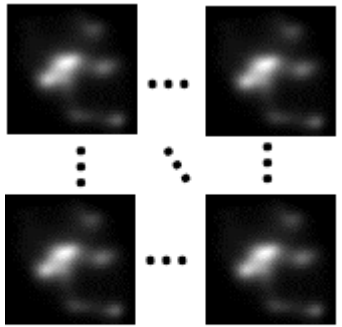
Authors: Ailbhe Gill, Emin Zerman, Martin Alain, Mikael Le Pendu, Aljosa Smolic

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**



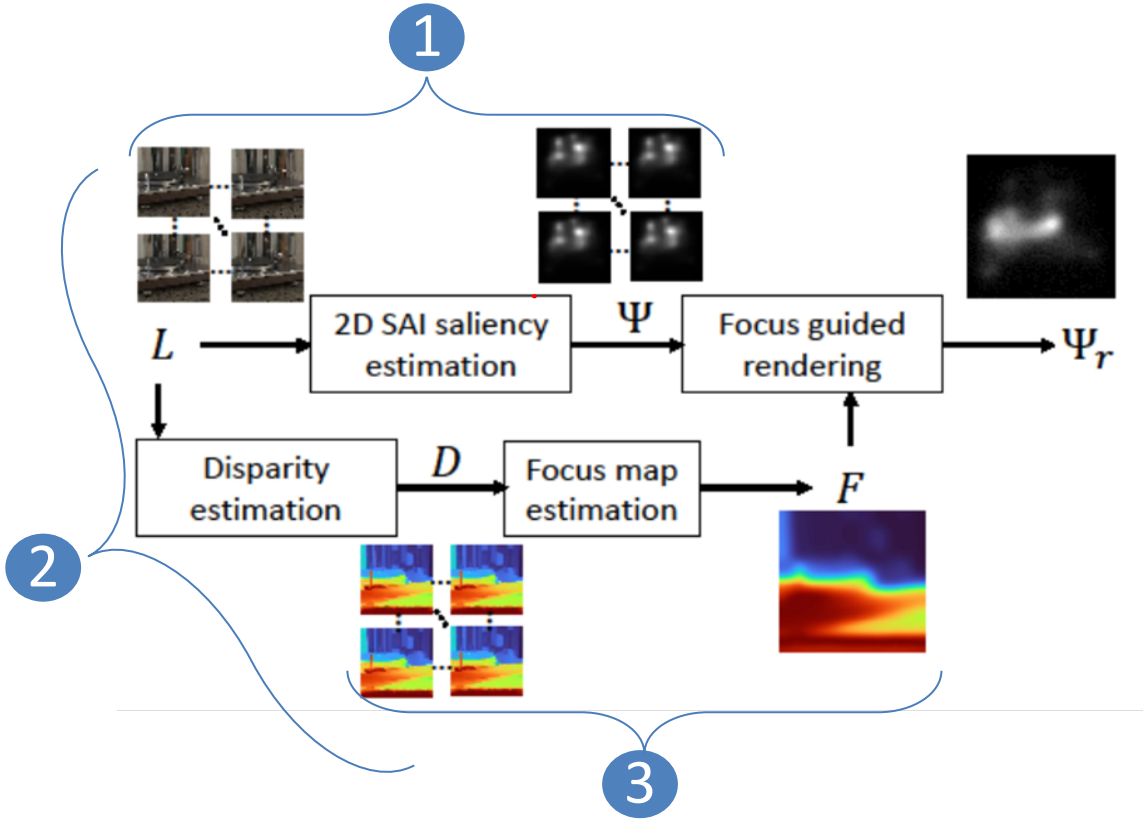
L



Ψ

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

Focus Guided Saliency Estimation Pipeline



Trinity College Dublin

The University of Dublin

V-SENSE

Many Thanks

smolica@tcd.ie