Trinity College Dublin
The University of Dublin

V-SENSE

VIVA-Q: Omnidirectional Video Quality Assessment based on Voronoi Patches and Visual Attention

Simone Croci, Emin Zerman, and Aljosa Smolic

ODV Pipeline

Storytelling/
Pre-production

Acquisition

Coding

VP9 AV1
H.264 MPEG-4/AVC
HEVC

Delivery/
Streaming

MPEG DASH
HLS

Display

Full-reference Quality Assessment

Unique Aspects of ODV

1. Spherical nature but stored in planar representations

Projection

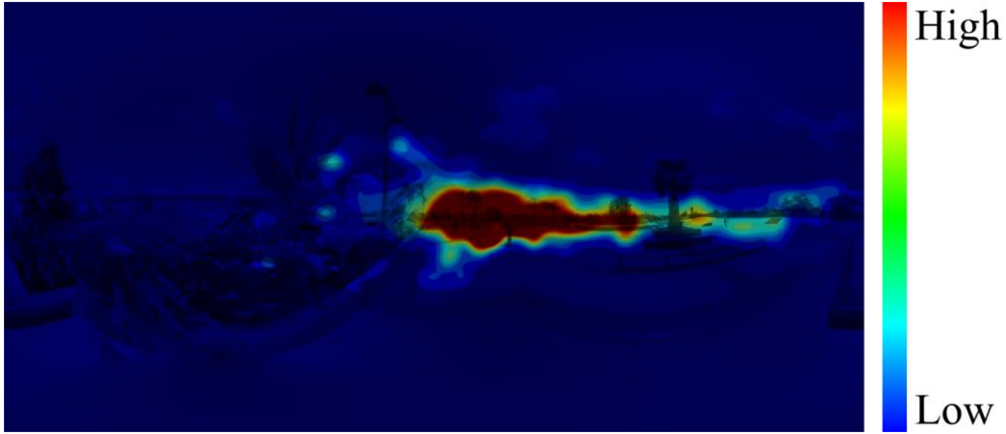
Unique Aspects of ODV

2. Viewing characteristics: free look around, only viewport

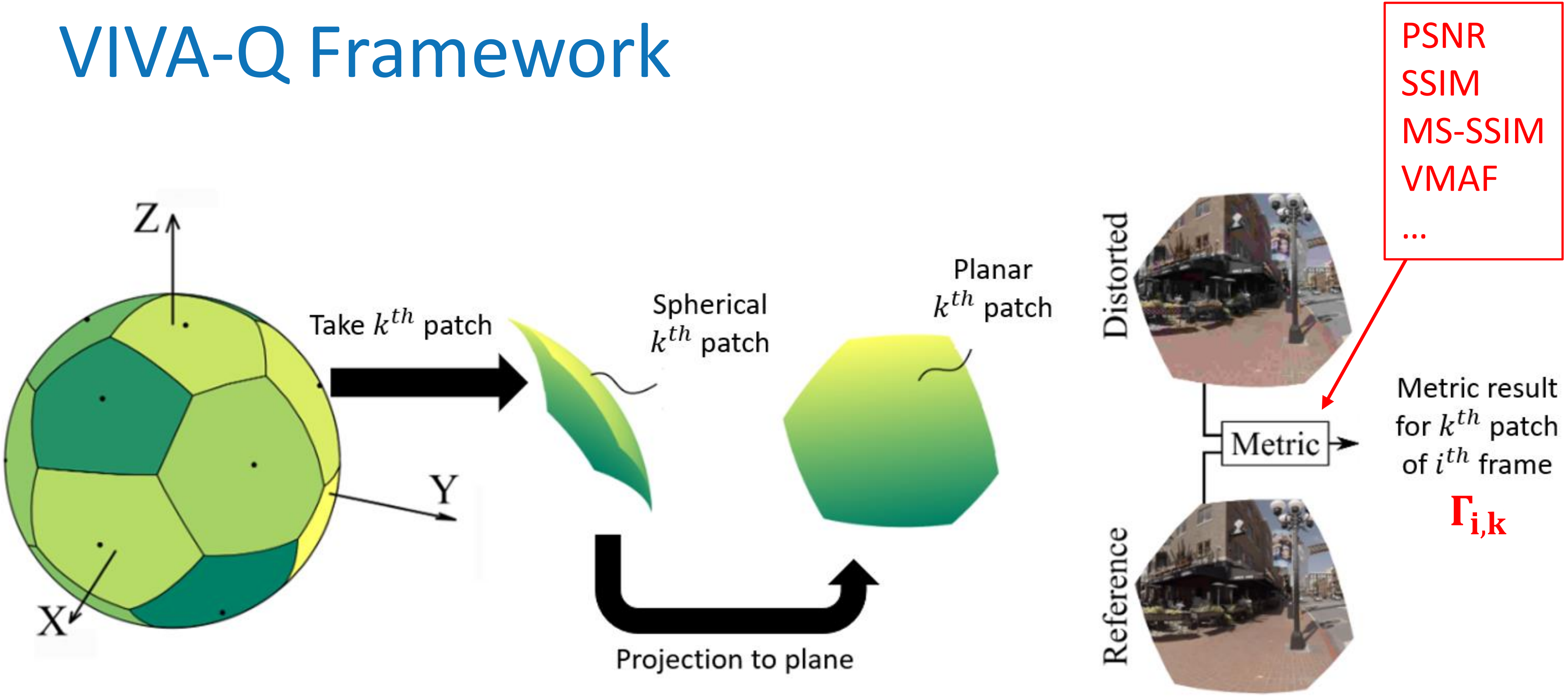
Unique Aspects of ODV

2. Viewing characteristics: free look around, only viewport

Visual Attention



VIVA-Q Framework



VIVA-Q Framework

Score of frame *i*:

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

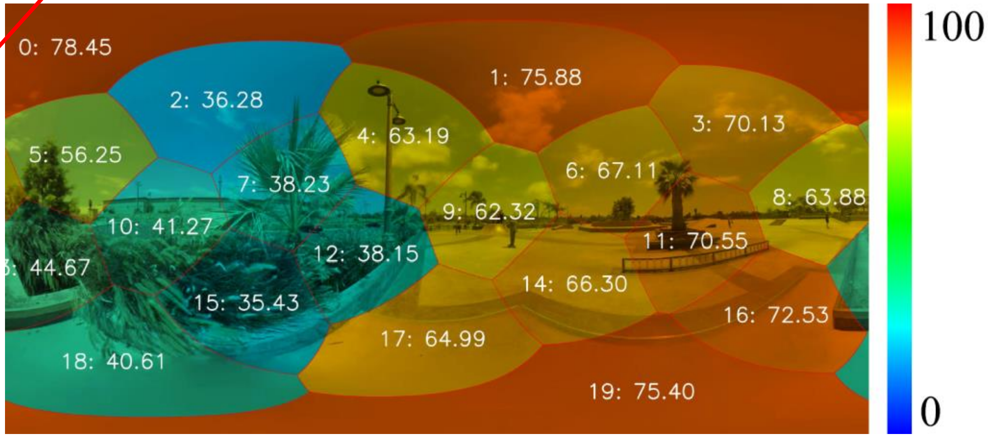
$$T_i' = \frac{\sum_{k=1}^{M} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} \nu_{i,k}}$$

$\Gamma_{i,k}$ Patch score

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

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

$\Gamma_{i,k}$



VIVA-Q Framework

Score of frame *i*:

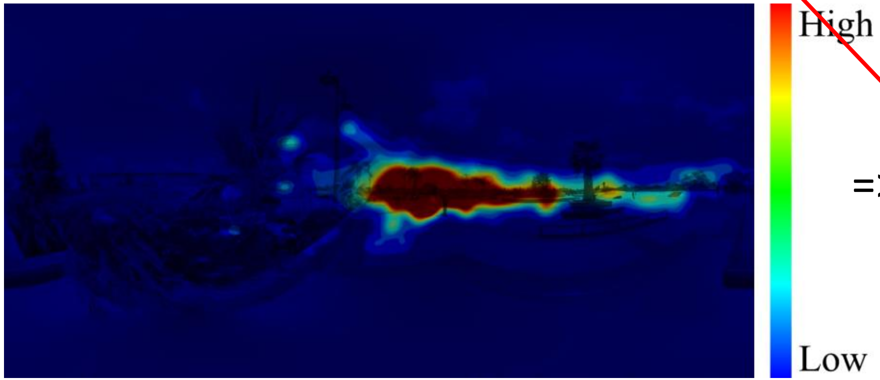
$$T_i = \frac{\sum_{k=1}^{M} \Gamma_{i,k}}{M}$$

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

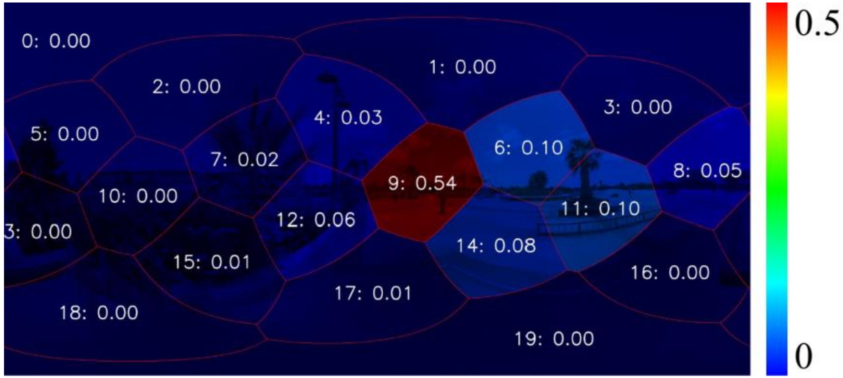
$\Gamma_{i,k}$ Patch score

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

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



=> $\nu_{i,k}$



VIVA-Q Framework

Final score from temporal pooling P of frame scores

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

$$\text{VIVA-Q} = P(T'_1, T'_2, \ldots, T'_N)$$

P: <u>arithmetic mean</u>, 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*

ODV Dataset

Adaptive Streaming System Distortions

1. Scaling: 8128 x 4064, 3600 x 1800, 2032 x 1016
2. Compression:
   - HEVC/H.265 (libx265 codec): two-pass encoding with the video buffering verifier method
   - Five target bitrates selected by experts

=> 120 distorted ODVs

Subjective Experiments

- **M-ACR-HR**[1]

Stimulus (10 sec)	Mid-Gray (3 sec)	Stimulus (10 sec)	Voting

  - [0,100] continuous grading scale

- **Apparatus:** HTC Vive + Virtual Desktop

[1] Singla et al., "Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution", Proceedings of the on Thematic Workshops of ACM Multimedia, 2017

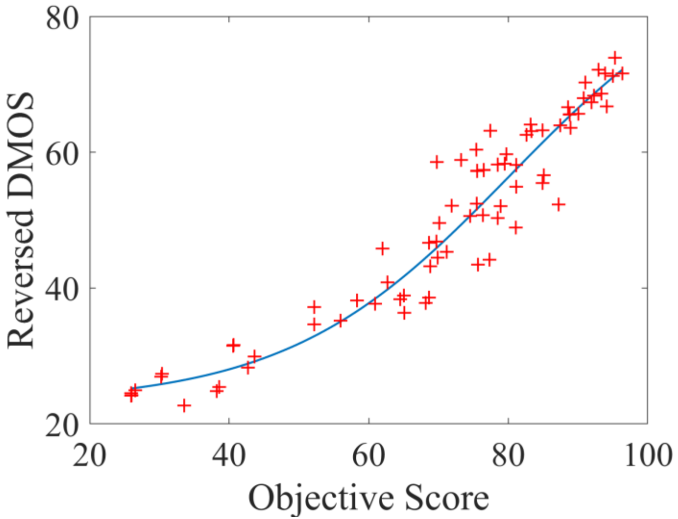
Comparative Analysis

- **Metrics:**
  - VI-Q: VI-PSNR, VI-SSIM, VI-MS-SSIM, VI-VMAF

    VIVA-Q: VIVA-PSNR, VIVA-SSIM, VIVA-MS-SSIM, VIVA-VMAF
    - 20 patches with 10 pix/deg resolution
  - Traditional video: PSNR, SSIM, MS-SSIM, VMAF[1]
    - Formats: equirectangular proj. (ERP), cubemap proj. (CMP)
  - ODV: S-PSNR-I[2], S-PSNR-NN[2], WS-PSNR[3], CPP-PSNR[4]

[1]Li et al., "Toward a practical perceptual video quality metric", Netflix Tech Blog, 2019
[2]Yu et al., "A framework to evaluate omnidirectional video coding schemes", ISMAR, 2015
[3]Sun et al., "Weighted-to-spherically-uniform quality evaluation for omnidirectional video", Signal Process. Lett., 2017
[4]Zakharchenko et al., "Quality metric for spherical panoramic video", Proc. SPIE, 2016

Comparative Analysis

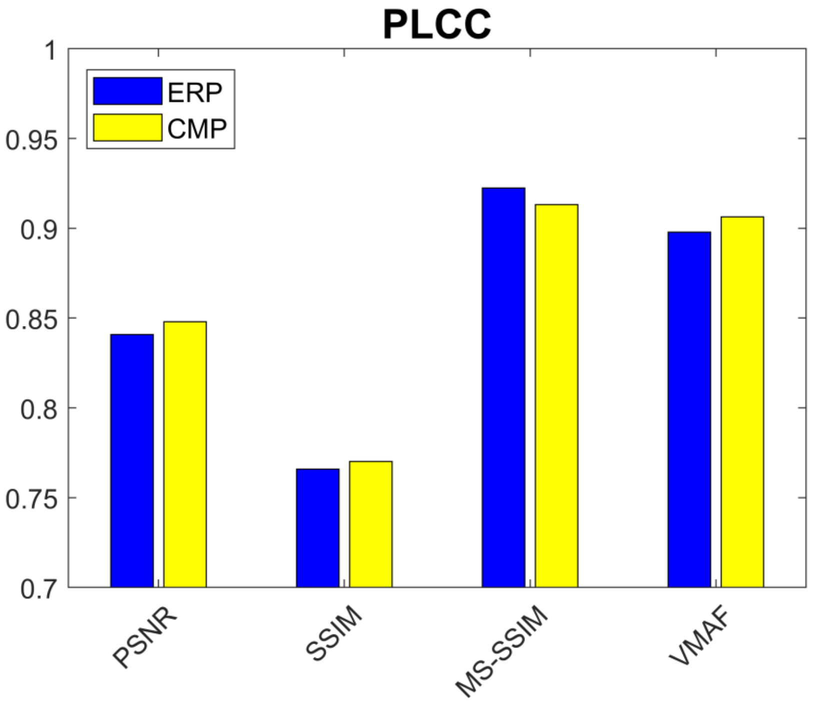


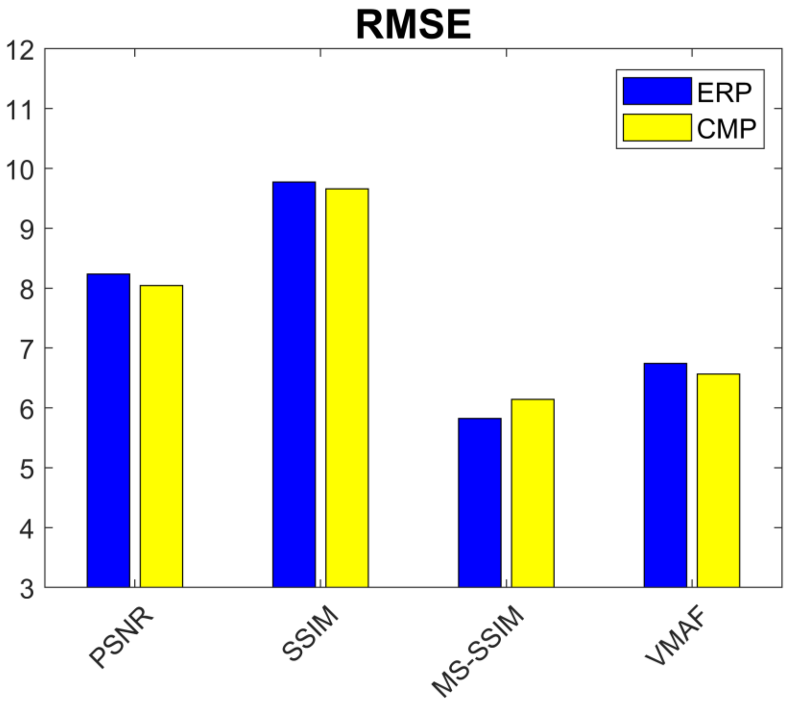
Correlation Analysis:

- **Logistic function:** $S' = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{S - \beta_3}{\|\beta_4\|}}} + \beta_2$
- **Performance metrics**
  - Pearson's linear correlation coefficient (PLCC)
  - Spearman's rank ordered correlation coefficient (SROCC)
  - Root mean squared prediction error (RMSE)
  - Mean absolute prediction error (MAE)

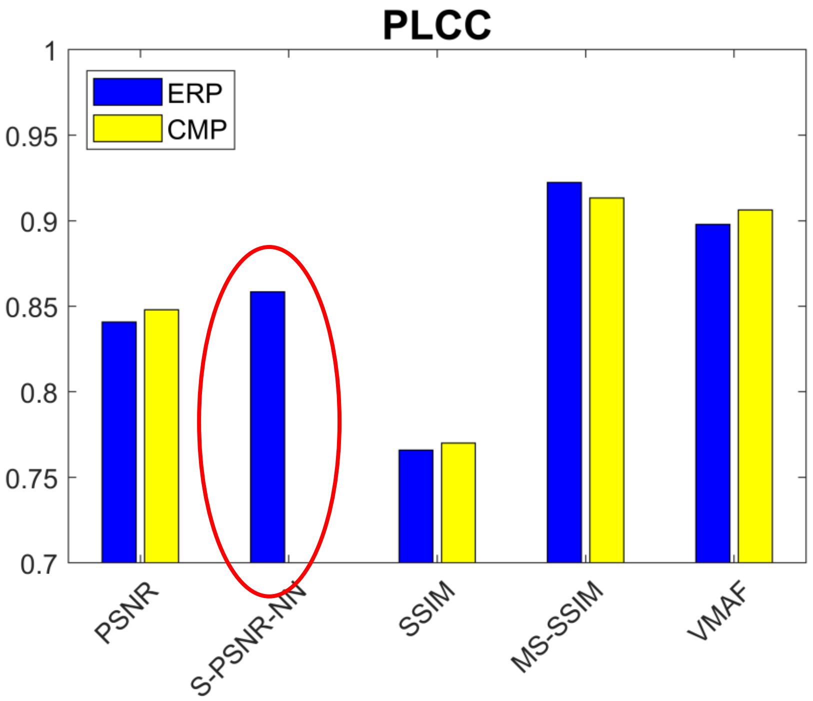
Metrics	PLCC	SROCC	RMSE	MAE
PSNR_{ERP}	0.8408	0.8237	8.2326	6.3169
PSNR_{CMP}	0.8480	0.8323	8.0419	6.2085
S-PSNR-I	0.8580	0.8438	7.8207	5.9715
S-PSNR-NN	0.8584	0.8433	7.8066	5.9648
WS-PSNR	0.8582	0.8430	7.8107	5.9772
CPP-PSNR	0.8579	0.8439	7.8200	5.9779
SSIM_{ERP}	0.7659	0.7551	9.7734	7.7396
SSIM_{CMP}	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
VIVA-PSNR	0.8876	0.8712	7.1818	5.5072
VIVA-SSIM	0.9106	0.9007	6.4345	4.8097
VIVA-MS-SSIM	0.9676	0.9635	3.8982	3.1526
VIVA-VMAF	**0.9773**	**0.9717**	**3.3753**	**2.5948**

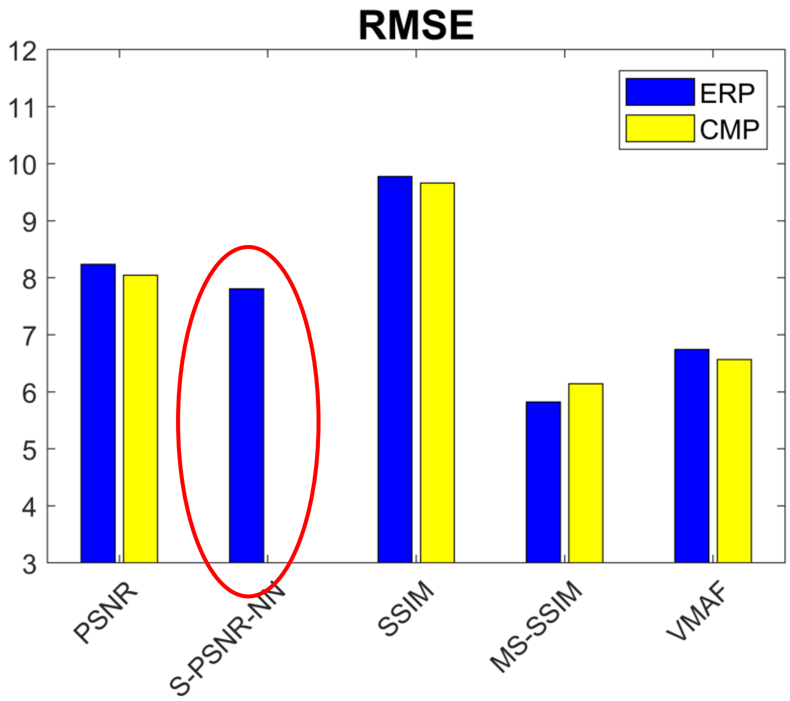
Standard Video Metrics



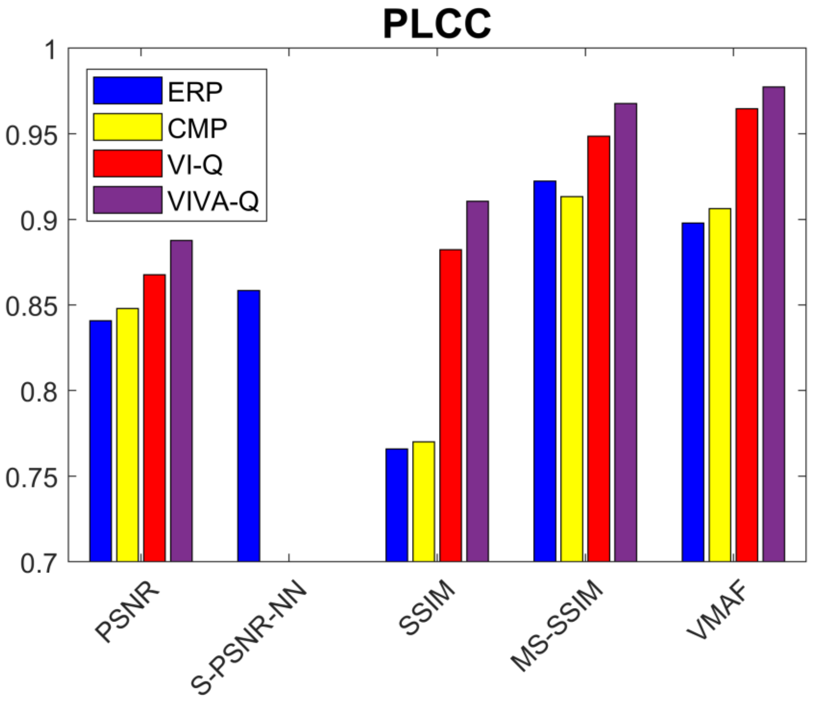


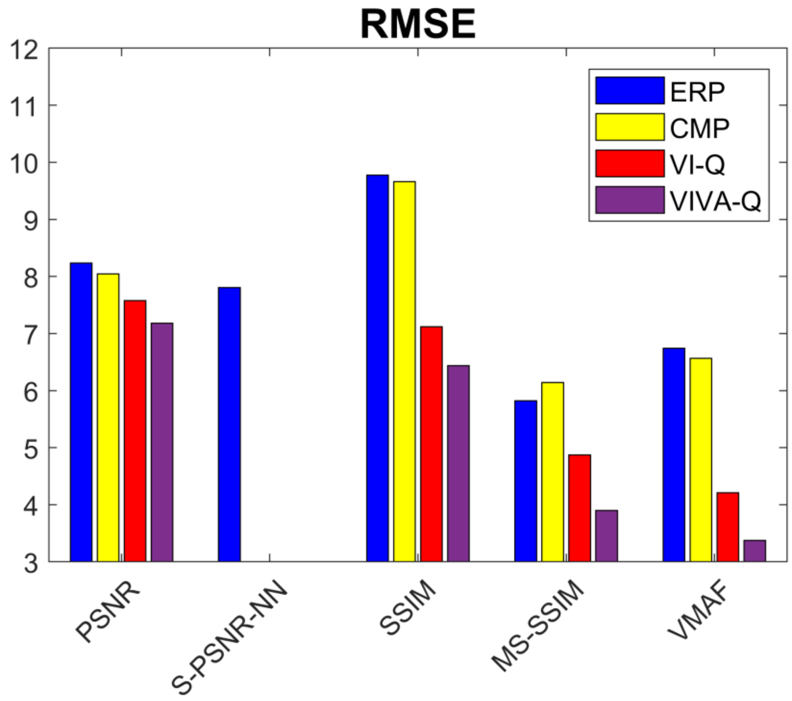
S-PSNR-NN


PLCC
1
0.95
0.9
0.85
0.8
0.75
0.7
ERP
CMP
PSNR
S-PSNR-NN
SSIM
MS-SSIM
VMAF


RMSE
12
11
10
9
8
7
6
5
4
3
ERP
CMP
PSNR
S-PSNR-NN
SSIM
MS-SSIM
VMAF

Voronoi patches and Visual Attention





	2K		*4K*		*8K*	
Metrics	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR_{ERP}	0.7388	0.6139	0.8360	0.8343	0.9202	0.9183
PSNR_{CMP}	0.7517	0.6203	0.8431	0.8450	0.9221	0.9163
S-PSNR-I	0.7634	0.6469	0.8568	0.8615	0.9304	0.9228
S-PSNR-NN	0.7649	0.6433	0.8570	0.8574	0.9300	0.9227
WS-PSNR	0.7650	0.6366	0.8570	0.8574	0.9299	0.9230
CPP-PSNR	0.7638	0.6432	0.8567	0.8615	0.9302	0.9230
SSIM_{ERP}	0.6996	0.5570	0.7703	0.7951	0.8600	0.8482
SSIM_{CMP}	0.7011	0.5591	0.7714	0.7878	0.8565	0.8484
$\mathrm{MS\text{-}SSIM}_{ERP}$	0.8841	0.7992	0.9150	0.9351	0.9652	0.9478
$\mathrm{MS\text{-}SSIM}_{CMP}$	0.8673	0.7824	0.9071	0.9276	0.9583	0.9446
VMAF_{ERP}	0.9202	0.8735	0.9203	0.9071	0.9515	0.9240
VMAF_{CMP}	0.9226	0.8790	0.9309	0.9156	0.9567	0.9285
VI-PSNR	0.7640	0.6321	0.8660	0.8769	0.9358	0.9247
VI-SSIM	0.8346	0.7109	0.8794	0.9060	0.9367	0.9249
VI-MS-SSIM	0.8642	0.8807	0.8140	0.9437	0.9767	0.9557
VI-VMAF	0.9627	0.9287	0.9577	0.9458	0.9789	0.9500
VIVA-PSNR	0.7960	0.6644	0.9050	0.9006	0.9451	0.9321
VIVA-SSIM	0.8434	0.7326	0.9200	0.9321	0.9593	0.9392
VIVA-MS-SSIM	0.9529	0.9105	0.8332	**0.9674**	0.9829	**0.9634**
VIVA-VMAF	**0.9762**	**0.9493**	**0.9737**	0.9625	**0.9862**	0.9593

Findings

- **VI-Q and VIVA-Q better than ERP and CMP**
  - Low projection distortion of Voronoi patches
- **VIVA-Q better than VI-Q**
  - Visual attention is important
- **Best: VIVA-VMAF**

Conclusions

- **VIVA-Q framework**
  - Metrics based on Voronoi patches and visual attention
- **ODV Dataset with 8 reference and 120 distorted ODVs**
  - Subjective scores and visual attention data
- **Comparative analysis**
  - VIVA-VMAF achieves state-of-the-art performance

Suggestions

- **VIVA-Q as standard recommendation**
- **Extension of ODV Dataset**
  - More contents
  - Different types of distortions
  - Subjective quality scores and visual attention data

Trinity College Dublin
The University of Dublin

V-SENSE

Many Thanks!

- Contact: crocis@tcd.ie
- Paper: Croci et al., “Visual Attention-Aware Quality Estimation Framework for Omnidirectional Video using Spherical Voronoi Diagram”, QUX 2020
- Code & Dataset: https://v-sense.scss.tcd.ie/research/voronoi-based-objective-metrics/