Viewport-driven Multi-metric Fusion Approach for 360° Video Quality Assessment

EPFL: Roberto Azevedo, Pascal Frossard

YouTube Media Algorithm: Neil Birkbeck, Ivan Janatra, Balu Adsumilli

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

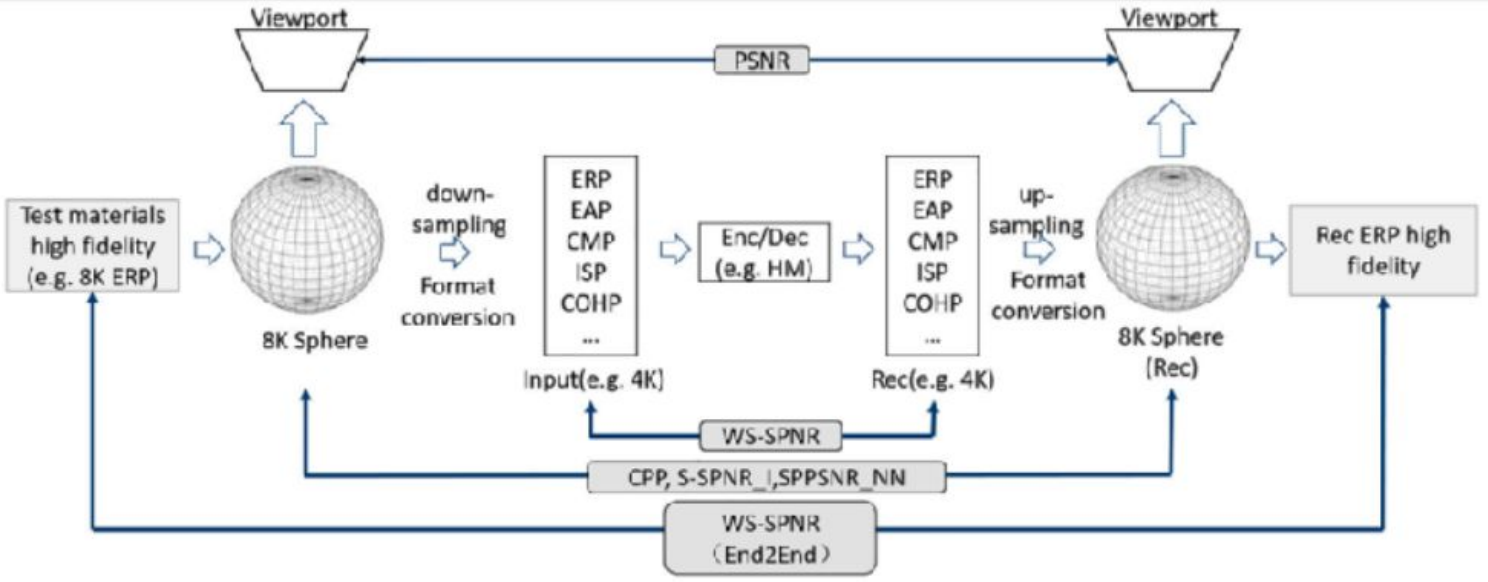
Background

- Subjective and objective quality for 360° videos still an open problem
  - VR headset -> Increased level of immersion -> Changes the QoE perspective
- Follow up from our previous study with more limited dataset (Azevedo et al., 2020)
  - Individual metrics computed on **viewports** correlates better with subjective scores than metrics computed on the projection domain...
  - ...but no single metric performs best across all distortion types
- Objective: Build a multi-metric model (e.g. VMAF for 2D videos) for 360-degree VQA

Roberto Azevedo, Neil Birkbeck, Ivan Janatra, Balu Adsumilli, and Pascal Frossard, "Subjective and viewport-based objective quality assessment of equiangular cubemap 360° videos," Electronic Imaging 2020.

Related work

Error-based metrics



Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," Signal Processing, May 2018.

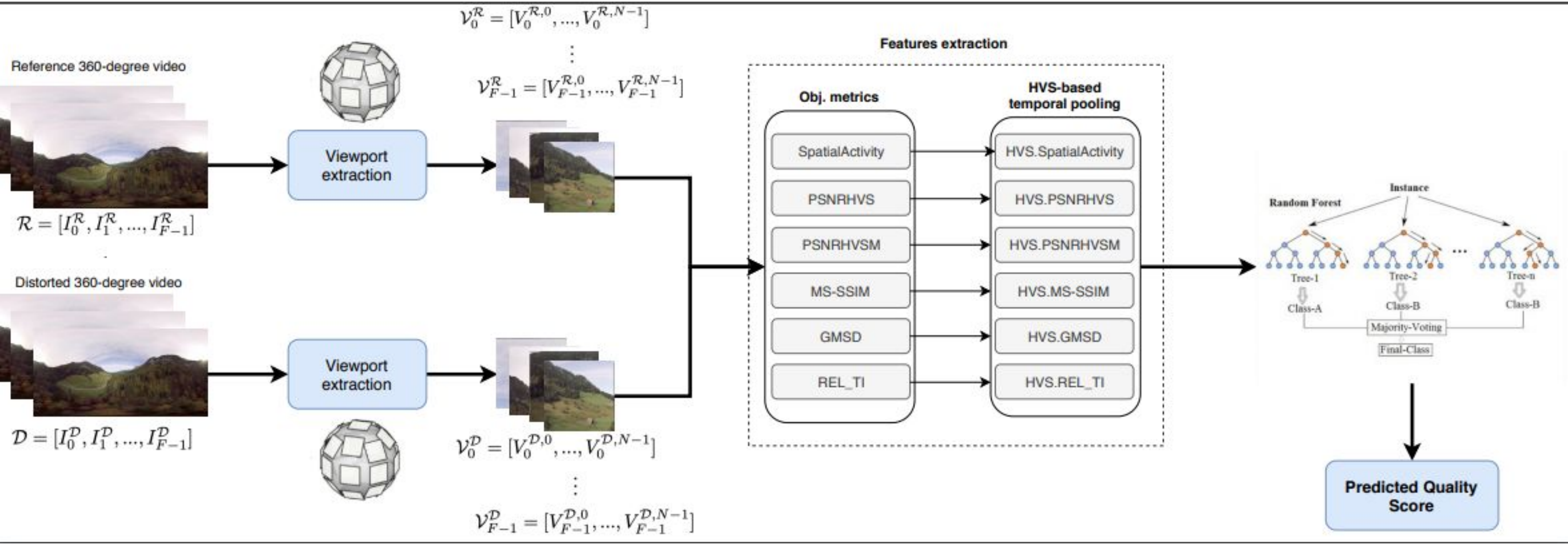
Related work

Deep learning

- MC360I3D (image-only)
- DeepVR-IQA (image-only)
- V-CNN (video, viewport-based CNN)

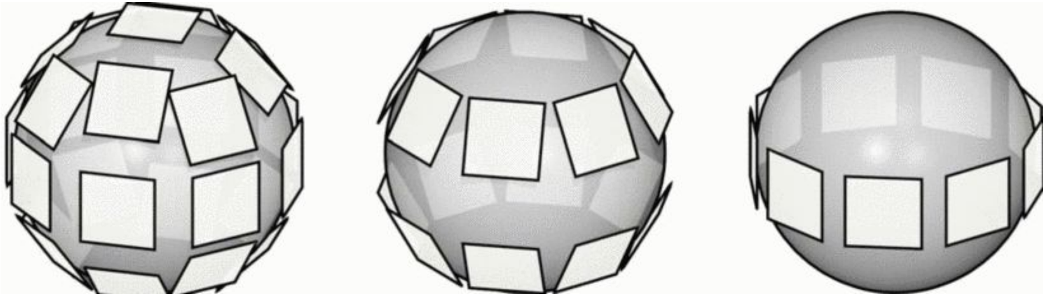
Y. Sun et al., "Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video," IEEE Signal Process. Lett., 2017
H. G. Kim et al., "Deep Virtual Reality Image Quality Assessment with Human Perception Guider for Omnidirectional Image," IEEE Trans. on Circuits and Syst. for Video Tech., 2019.
C. Li et al., "Viewport Proposal CNN for 360° video quality assessment," June 2019.

General Approach



Viewports Sampling

We tried 3 viewport sampling modes x 3 FOV (30°, 40°, 50°)



Uniform Tropical Equatorial

Rendered collage

... as with our previous study, Uniform 40° seems to perform best.

N. Birkbeck, C. Brown, and R. Suderman. “Quantitative evaluation of omnidirectional video quality,” in Proc. 9th QoMEX, pages 1–3, 2017.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

YouTube

Viewports Sampling

Example - Uniform 40°

Objective Metrics

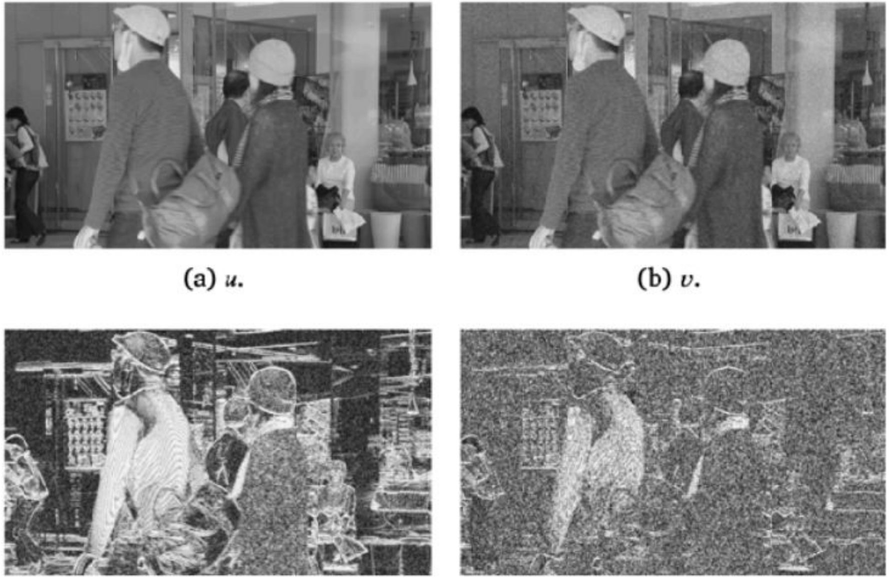
Spatial Activity

$$S(z) = \sqrt{(G_1 * z)^2 + (G_1^\top * z)^2},$$

$$G_1 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix},$$

$$s = S(u) - S(v).$$

$$\mathrm{SA}(v, u) = \sqrt{\frac{1}{MN} \sum_{i,j} |s_{ij}|^2},$$



(a) u. (b) v. (c) $S(u)$. (d) $S(v)$.

P.G. Freitas et al., "Using multiple spatio-temporal features to estimate video quality," Signal Processing Image Commun., May 2018.

Objective Metrics

PSNR-HVS and PSNR-HVS-M

- PSNR-HVS
  - Divides image in 8x8 non-overlapping blocks, and
  - Applies weight on the difference based on contrast sensitivity function (CSF)
- PSNR-HVS-M
  - Like PSNR-HVS, with additional contrast masking multiplier applied to the DCT coefficients difference

N. Ponomarenko et al., “On between-coefficient contrast masking of DCT basis functions,” in 3rd Intern. Workshop on Video Processing and Quality Metrics, 2007.

Objective Metrics

SSIM and MS-SSIM

- SSIM
  - Luminance $l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$
  - Contrast $c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$
  - Structure $s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$

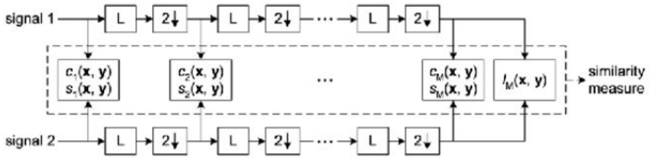
$$SSIM(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^\alpha \cdot [c(\mathbf{x},\mathbf{y})]^\beta \cdot s(\mathbf{x},\mathbf{y})]^\gamma$$

$\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C2)}{(\mu^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)}$$

- MS-SSIM

$$MSSSIM(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^\alpha \cdot \prod_{m=1}^{M} [c_m(\mathbf{x},\mathbf{y})]^\beta \cdot [s_m(\mathbf{x},\mathbf{y})]^\gamma$$



L: low-pass filtering; 2 ↓: downsampling by 2.

Z. Wang et al., “Multiscale structural similarity for image quality assessment,” in The 37th Asilomar Conf. on Signals, Systems, Computers 2003

Objective Metrics

Gradient-magnitude Similarity Deviation (GMSD)

$$\mathrm{GMS}(u,v) = \frac{2 \cdot m(u) \cdot m(v) + c}{m(u)^2 + m(v)^2 + c},$$

$$\mathrm{GMSD}(u,v) = \sqrt{\frac{1}{NM} \sum_{i,j} \left(\mathrm{GMS}(u,v) - \overline{\mathrm{GMS}(u,v)} \right)^2},$$

$$m(z) = \sqrt{(z * G_2)^2 + (z * G_2^\top)^2}. \qquad G_2 = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix},$$

$$\overline{\mathrm{GMS}(u,v)} = \frac{1}{NM} \sum_{i,j} \mathrm{GMS}(u,v).$$

(a) Original.

(b) Distorted.

(c) GMS map.

W. Xue et al., “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” IEEE Trans. On Image Process., Feb 2014

Objective Metrics

Relative change in Temporal Information

- Current 360-VQA approaches don't seem to incorporate temporal effects

$$TI[F_n] = std(\Delta F_n), \text{ where } \Delta F_n = F_n - F_{n-1}$$

$$TI_{rel}[F_n] = \frac{|TI_{ref}[F_n] - TI_{dist}[F_n]|}{TI_{ref}[F_n]}$$

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Temporal Pooling

- Metrics computed per frame, then pooled. Why?
  - Smooth effect
  - Asymmetric effect
  - Recency effect

$$Q^n_{LP}(f) = \begin{cases} Q^n_{LP}(f-1) + \boxed{\alpha} \cdot \Delta Q(f), \text{if } \Delta Q^n \leq 0 \\ Q^n_{LP}(f-1) + \boxed{\beta} \cdot \Delta Q(f), \text{if } \Delta Q^n > 0 \end{cases}$$

$$Q^n_{pool} = \frac{1}{F} \sum_{f=1}^{F} \left(Q^n_{LP}(f) \cdot \boxed{ln(\gamma} \cdot f + 1) \right)$$

Use α = 0.03, β = 0.2, γ = 1000.

Y. Lu, M. Yu, G. Jiang, "Low-complexity Video Quality Assessment Based on Spatio-Temporal Structure," Information and Software Tech. 2019.

Regression

- Generate feature vector containing each combination of pooled metric and viewport
- Use these to learn non-linear mapping w/ subjective scores
- Tested both SVR and RFR, ended up using RFR
- Run the following:
  - Our method (projection, VP collage, and VP domains)
  - PSNR (projection and VP collage domains)
  - S-PSNR
  - WS-PSNR
  - MS-SSIM (projection and VP collage domains)
  - VMAF (projection and VP collage domains)

Experiments

- We ran two experiments:
  - Fixed train-test set: use single fixed 80% train/validation set and 20% test set, prescribed by Dataset.
  - Cross-validation: in each of the 1000 runs, split Dataset to 80% train/validation set and 20% test set, and run as Fixed.

Dataset

VQA-ODV

- Contains 60 ref + 180 impaired equirect sequences.
    - Ref videos have varying resolutions (4k-8k), varying length (10-23s), varying fps (24-30fps)
    - Impaired videos use H.265 encoding with 3 QP levels (27, 32, 42)
- Rating from 221 subjects, divided into 10 groups
    - Use single-stimulus with hidden reference
    - Has MOS and DMOS
- Using HTC Vive as HMD; take HMD resolution into account when sampling viewport

C. Li, M. Xu, X. Du, and Z. Wang, “Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video: A Large-Scale Dataset and a Deep Learning Model,” in ACM MM, Seoul, Republic of Korea, 2018.

Dataset

Group 1
Group 2
Group 9
Group 3
Group 4
Group 5
Group 6
Group 10
Group 7
Group 8

Fixed train-test sets

- For our method: Run group shuffle cross-validation on training set to find best RF hyper-parameters, train the model on training set and test on the test set
- For comparison metrics: Fit a 4-parameter logistic function on the training set, and compute its function with the test set

C. Li et al., "Viewport Proposal CNN for 360° video quality assessment," June 2019.

Fixed train-test sets

Results

Metric	PLCC	SROCC	RMSE
PSNR (Proj.)	0.72495	0.73797	8.176
PSNR (VP-Collage)	0.76222	0.76345	7.5824
S-PSNR	0.75138	0.7704	7.7557
WS-PSNR	0.74328	0.56056	7.9501
MS-SSIM (Proj.)	0.76005	0.78867	7.8741
MS-SSIM (VP-Collage)	0.81719	0.84144	7.0024
VMAF (Proj.)	0.79657	0.79382	7.2481
VMAF (VP-Collage)	0.84483	0.85637	6.271
Ours (Proj.)	0.85629	0.86873	6.3588
Ours (VP-Collage)	0.89867	0.87439	5.7256
Ours (VP)	0.92575	0.91712	4.9954

VP-Collage domain generally outperforms projection domain

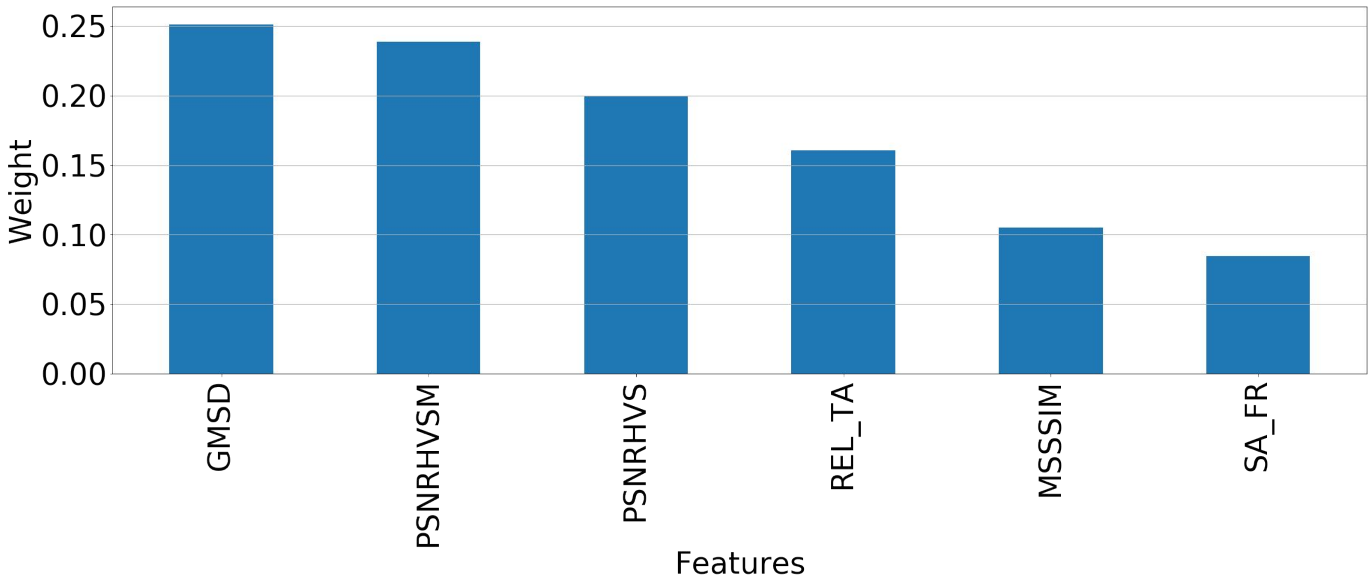
Our method outperforms VMAF due to selection of individual metrics and improved temporal pooling

best

Fixed train-test sets

Results

Average viewport features importance in our viewport method (for VQA-ODV)



Cross-Validation

Results

Metric	PLCC	SROCC	RMSE
PSNR (Proj.)	0.57156	0.61873	9.8249
PSNR (VP-Collage)	0.64746	0.68579	9.1224
S-PSNR	0.6246	0.66731	9.3461
WS-PSNR	0.59803	0.64501	9.5983
MS-SSIM (Proj.)	0.75004	0.77535	7.9351
MS-SSIM (VP-Collage)	0.76405	0.79113	7.758
VMAF (Proj.)	0.74692	0.76673	7.9631
VMAF (VP-Collage)	0.78085	0.79802	7.5147
Ours (Proj.)	**0.81728**	**0.82901**	**6.8716**
Ours (VP-Collage)	**0.82676**	**0.82647**	**6.7376**
Ours (VP)	**0.86778**	**0.86769**	**5.9367**

VP-Collage domain generally outperforms projection domain

Our method outperforms VMAF due to selection of individual metrics and improved temporal pooling

best

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

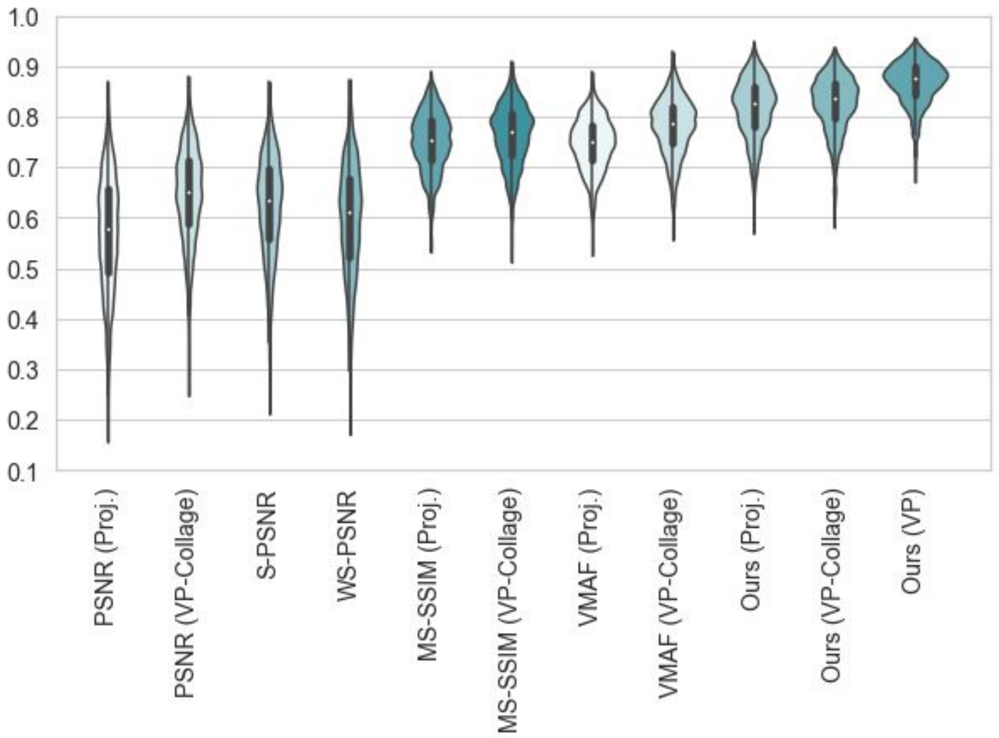
YouTube

Cross-Validation

Results

Our method (VP) has:

- smallest range of value
- best average
- higher density



Conclusion

- Viewport-based MMF achieves very good results compared to other objective metrics
  - Even just MMF (without viewport) outperforms single metrics
  - Metrics of separate viewports outperforms metrics of collaged viewports
  - Not as training-data-hungry as deep learning techniques
- Using viewport means it should also work for other projections
- Using multimetric means other individual metric can be added if the type of distortion in the dataset is known

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Future work

- Verify our method on multiple datasets
- Verify our method on different projections
- Consider visual attention data (available on VQA-ODV dataset)

Questions / Discussion

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE