VQEG

Predicting local distortions introduced by AV1 using Deep Features

Andréas Pastor[1], Lukáš Krasula[2], Xiaoqing Zhu[2], Zhi Li[2], Patrick Le Callet[1,3]

[1] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F–44000 Nantes, France

[2] Netflix Inc., Los Gatos, CA, USA

NETFLIX Nantes Université LS2N IUF institut universitaire de France

The problem we are trying to answer

- Video encoding is driven by measures (SSE, SAD) to assess the visibility of distortion locally, but these **pixel-based measures** are not well tuned to how humans perceive distortions, but efficient to compute.

The problem we are trying to answer

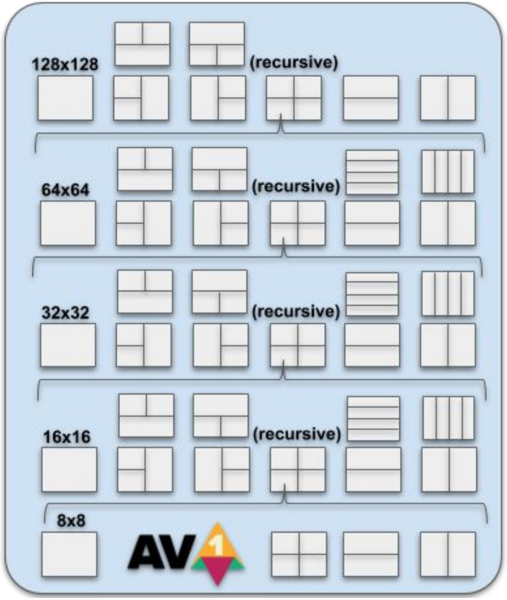
- Video encoding is driven by measures (SSE, SAD) to assess the visibility of distortion locally, but these **pixel-based measures** are not well tuned to how humans perceive distortions, but efficient to compute.
- Our goal is to correct these measurements at a local horizon in a video to improve the overall quality and reduce bitrate usage: **What is this local horizon?**

The problem we are trying to answer

- Video encoding is driven by measures (SSE, SAD) to assess the visibility of distortion locally, but these **pixel-based measures** are not well tuned to how humans perceive distortions, but efficient to compute.
- Our goal is to correct these measurements at a local horizon in a video to improve the overall quality and reduce bitrate usage: **What is this local horizon?**
- Requirement: a ground truth dataset to drive the research development and metric creation. **What is this ground truth data? How can we leverage Deep Features extracted from Neural Network to correct SSE?**

Connecting video encoding and localized Human Visual System perception with "Perceptual Unit"

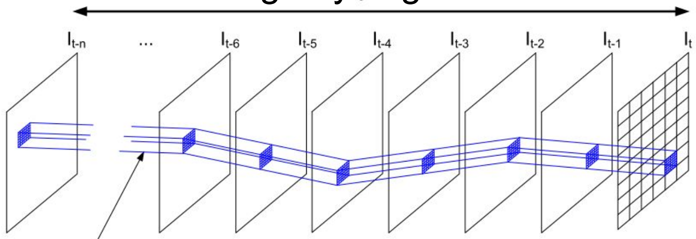
- Video encoders make decisions on Coding Units (CUs): mode selection, partionating, transform, filters …



Connecting video encoding and localized Human Visual System perception with “Perceptual Unit”

- Video encoders make decisions on Coding Units (CUs): mode selection, …
- A gaze performed by an human eye is:
  - **spatially located**, around foveated view: 1° of visual angle, 60ppd under standard viewing condition
  - **temporally located**: gaze fixation movement ~200ms
  - aligned along the direction of an object: **pursuit**

Average eyes gaze duration

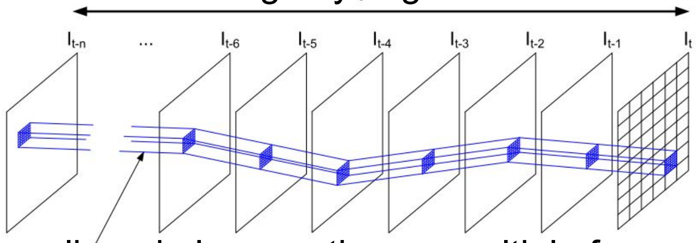


A spatio-temporal tube aligned along motion on multiple frames

Connecting video encoding and localized Human Visual System perception with “Perceptual Unit”

- Video encoders make decisions on Coding Units (CUs): mode selection, …
- A gaze performed by an human eye is:
  - **spatially located**, around foveated view: 1° of visual angle, 60ppd under standard viewing condition
  - **temporally located**: gaze fixation movement ~200ms
  - aligned along the direction of an object: **pursuit**
- **Perceptual Unit (PU):** same spatio-temporal horizon as a gaze on which we want to model how humans perceive distortion to drive CUs encoding

Average eyes gaze duration

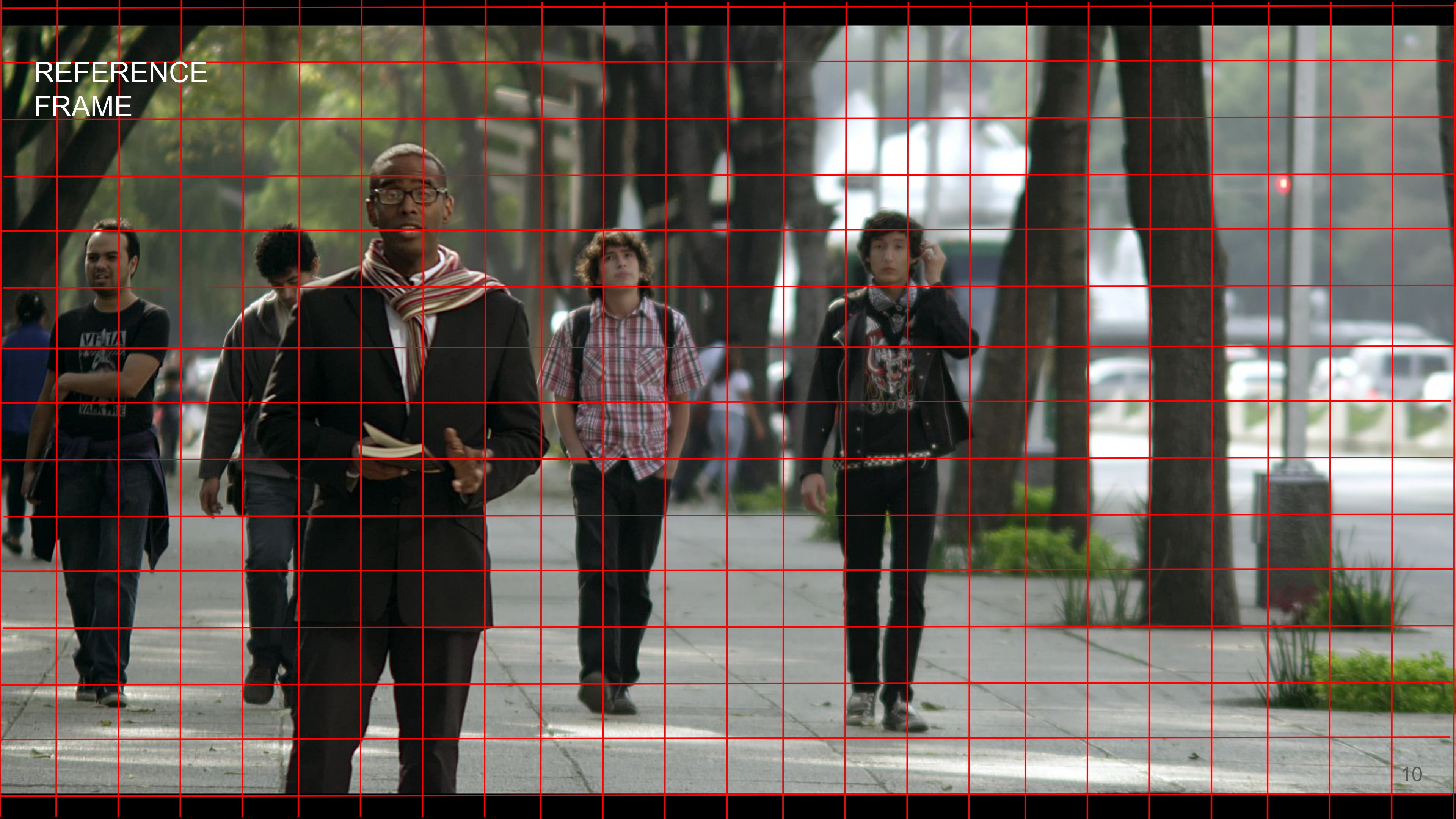


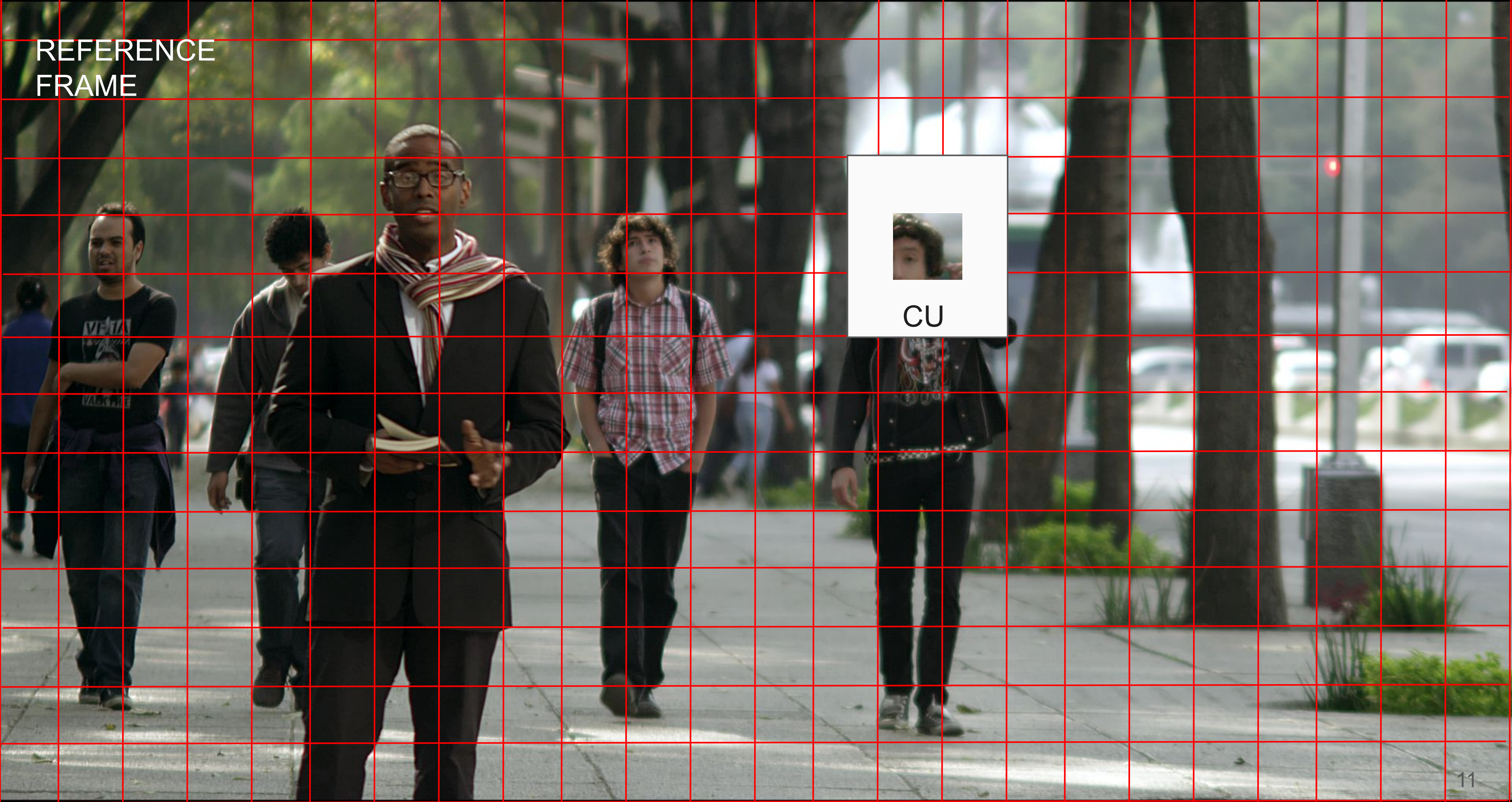
A spatio-temporal tube aligned along motion on multiple frames

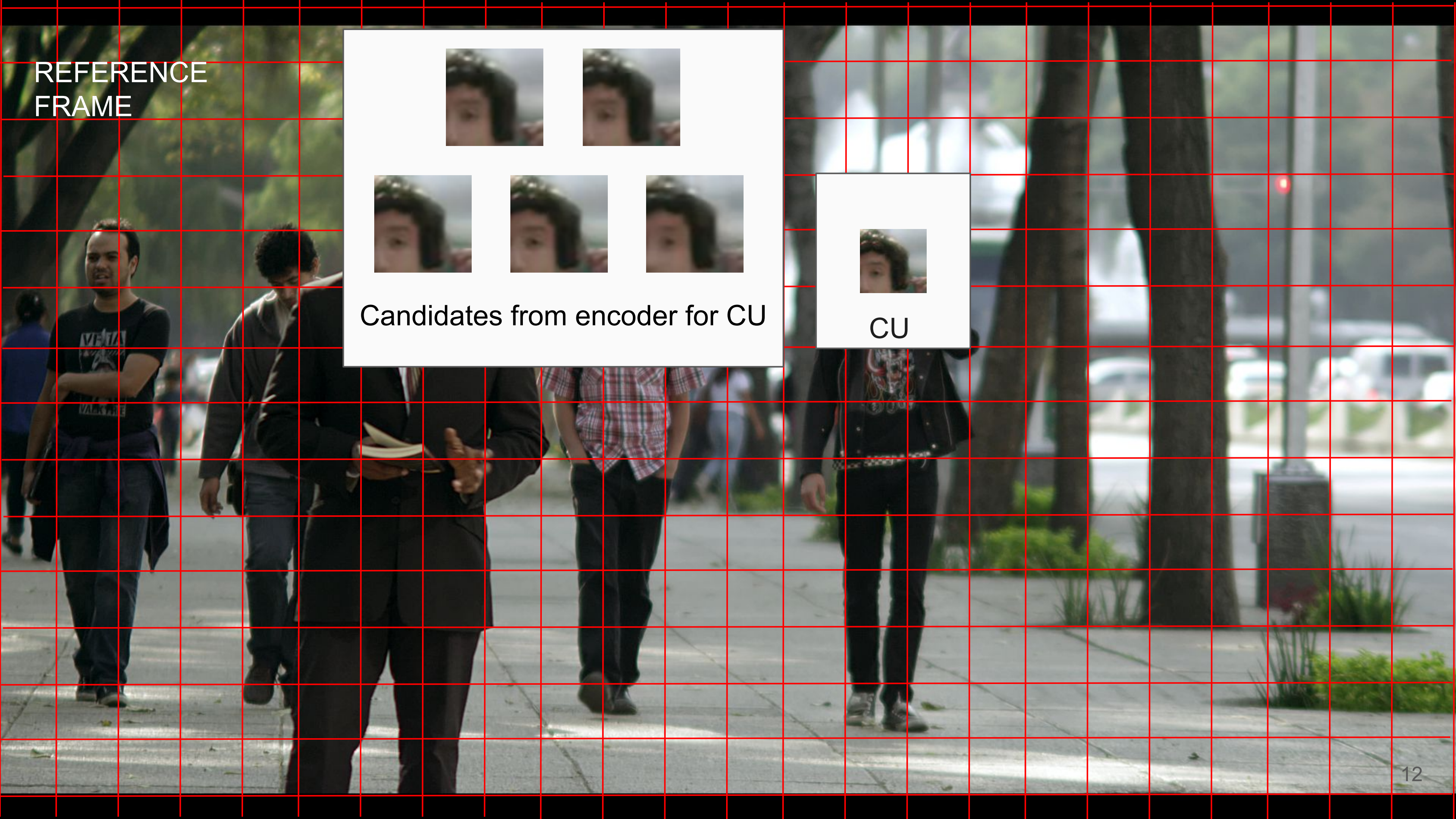
Visual example

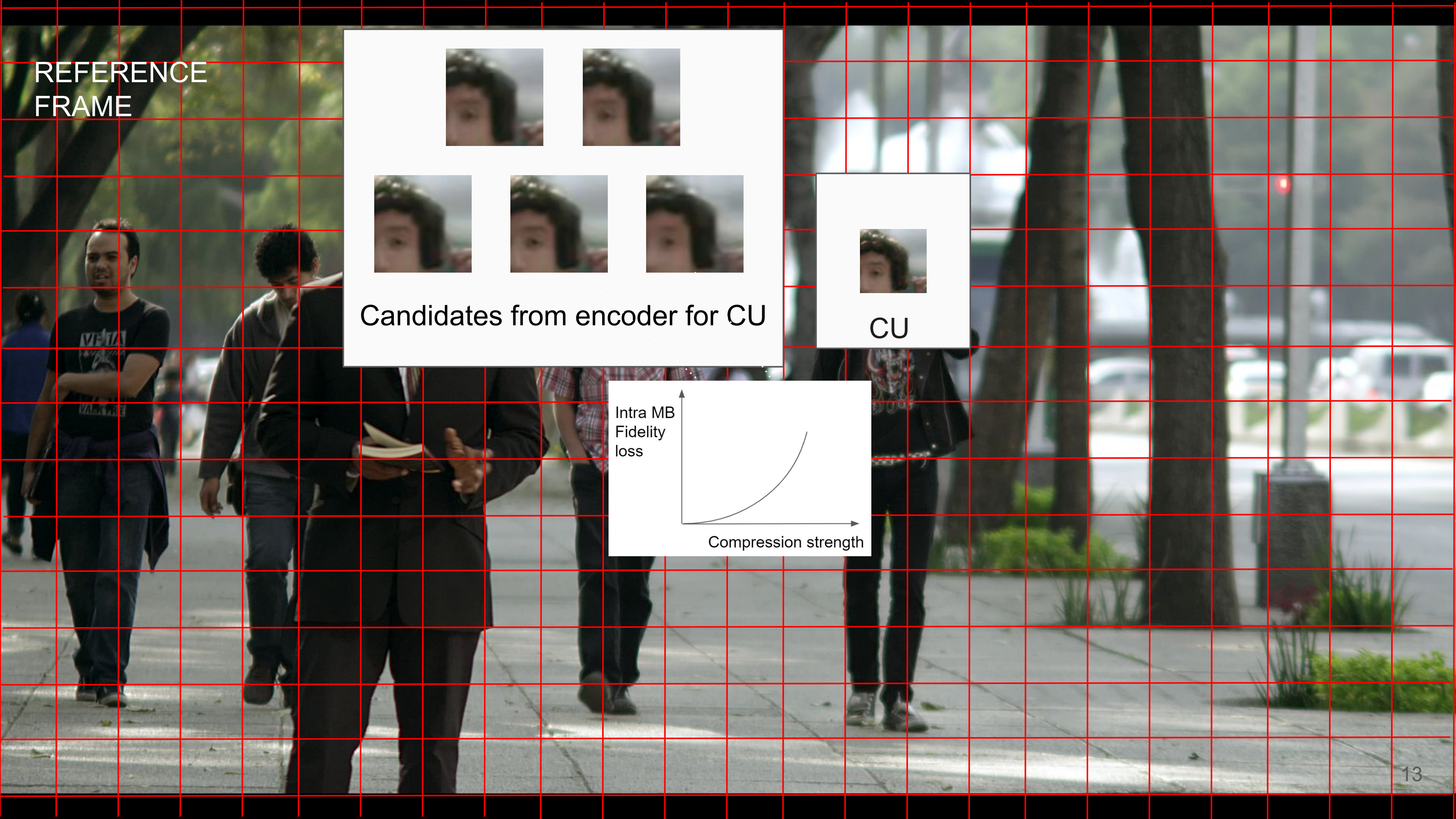
Perceptual Units and Perceptual Difference curves in encoding process

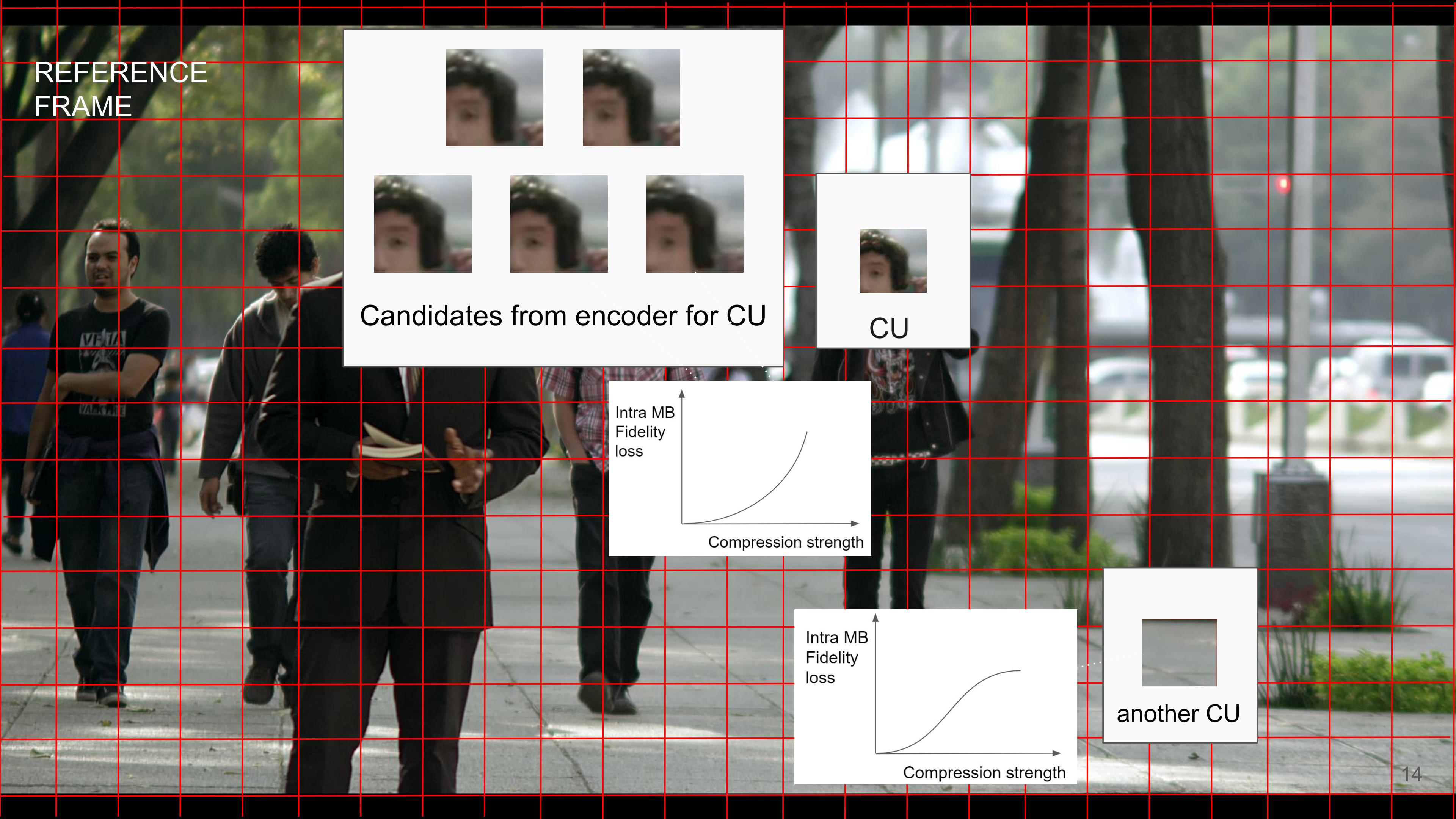
REFERENCE
FRAME

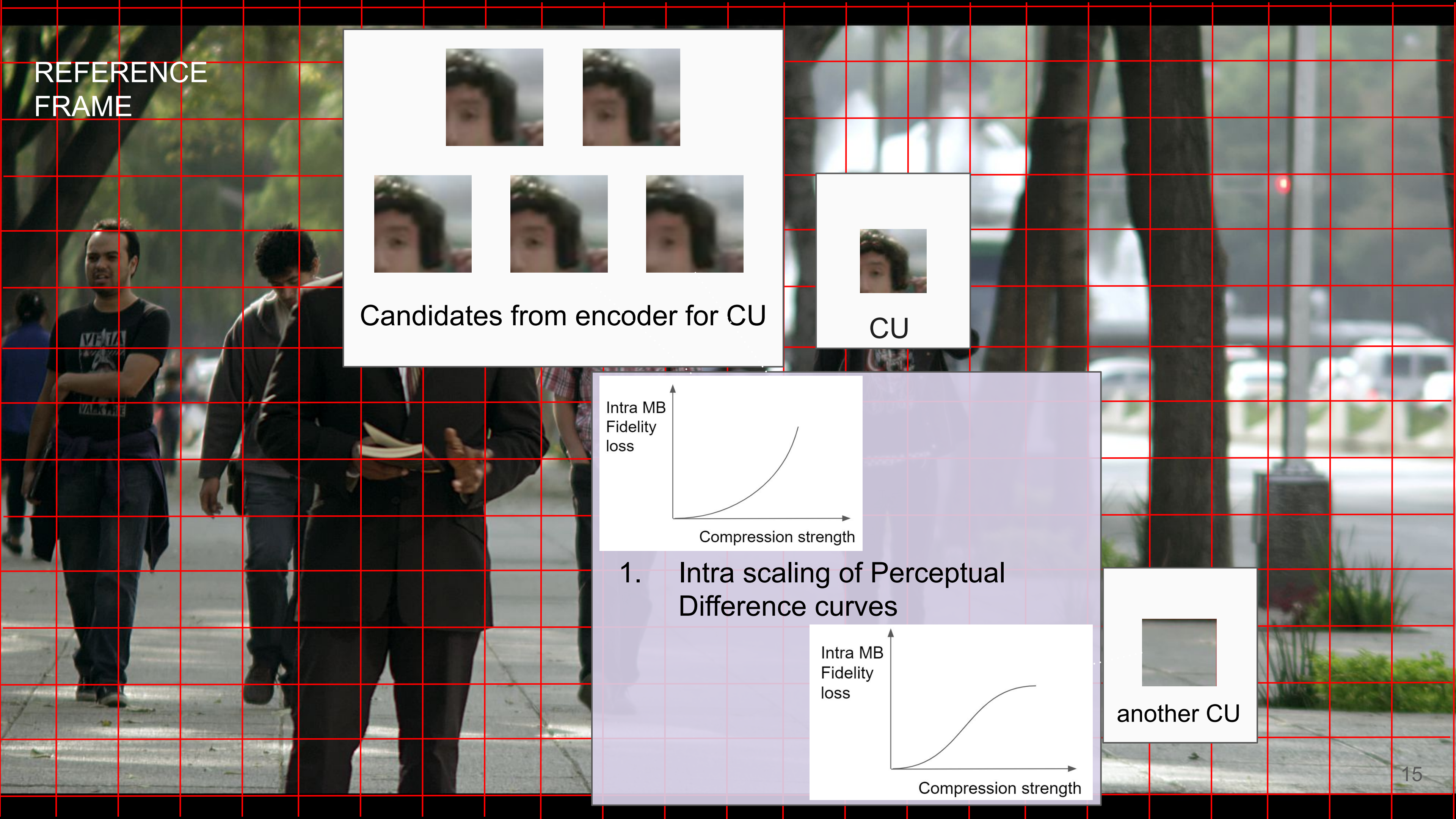

REFERENCE
FRAME

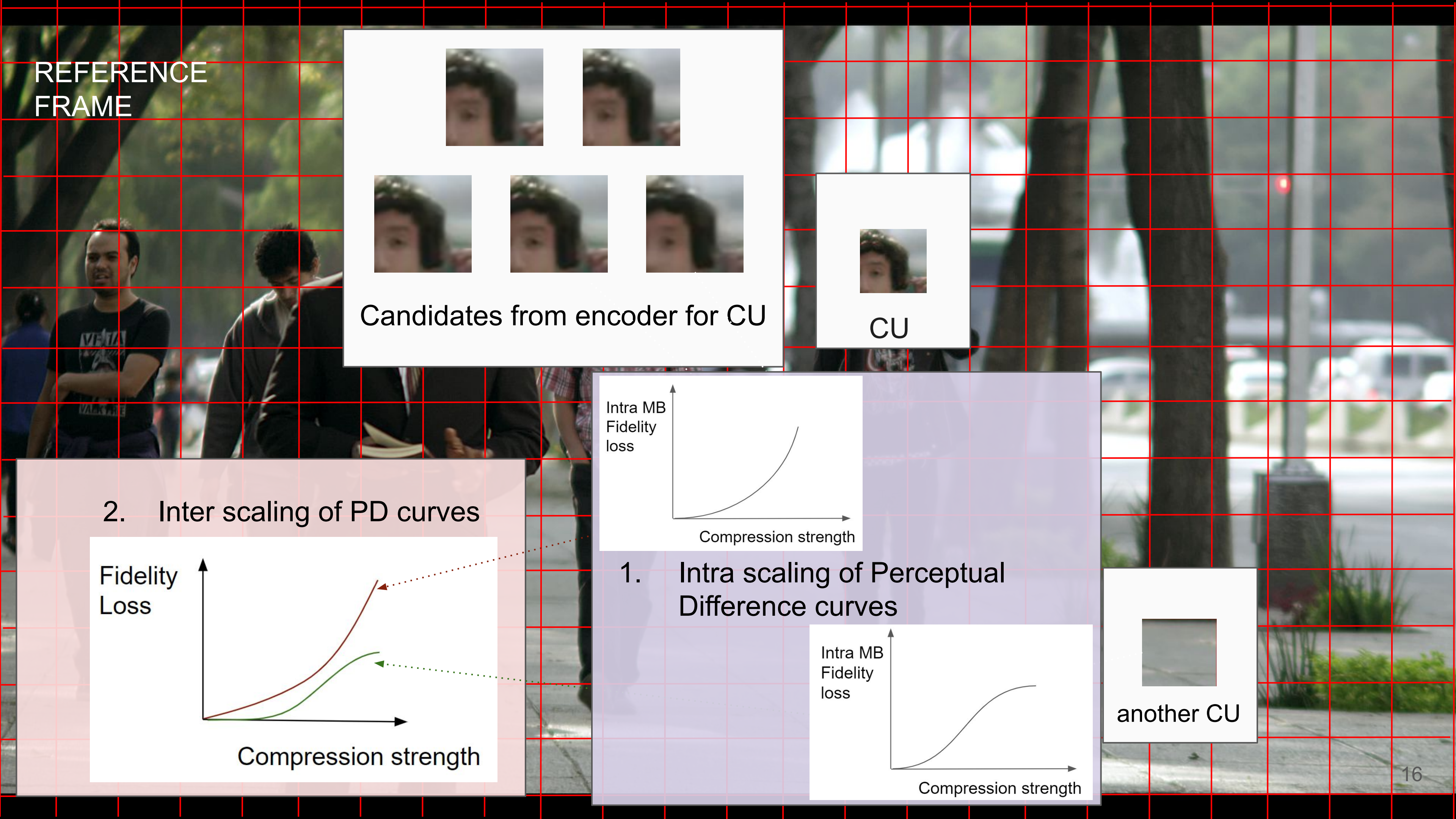

REFERENCE
FRAME
CU


REFERENCE
FRAME
Candidates from encoder for CU
CU


REFERENCE
FRAME
Candidates from encoder for CU
CU
Intra MB
Fidelity
loss
Compression strength


REFERENCE FRAME
Candidates from encoder for CU
CU
Intra MB Fidelity loss
Compression strength
Intra MB Fidelity loss
Compression strength
another CU


REFERENCE
FRAME
Candidates from encoder for CU
CU
Intra MB
Fidelity
loss
Compression strength
1. Intra scaling of Perceptual
Difference curves
Intra MB
Fidelity
loss
Compression strength
another CU


REFERENCE FRAME
Candidates from encoder for CU
CU
Intra MB Fidelity loss
Compression strength
1. Intra scaling of Perceptual Difference curves
Intra MB Fidelity loss
Compression strength
another CU
2. Inter scaling of PD curves
Fidelity Loss
Compression strength

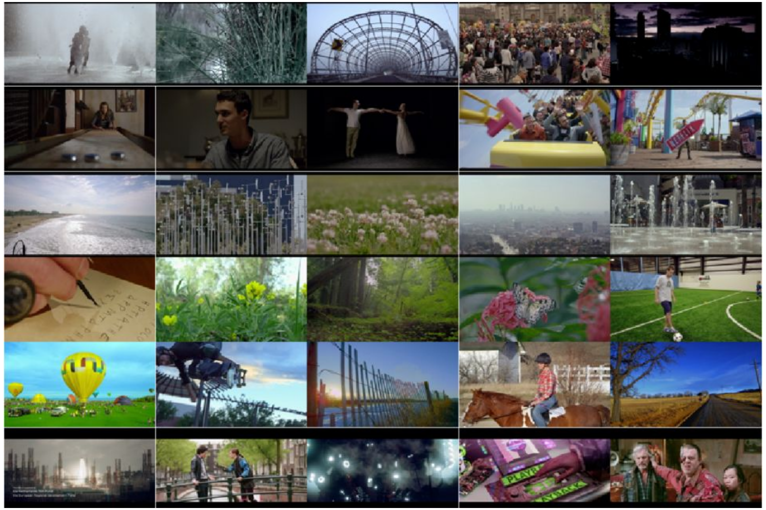
Dataset creation of tube-contents

Content selection and data collection

Content creation: encoding

To select tube-contents, we followed these steps:

- **Step 1**: Encoding of sources (SRCs).
    - 115 SRCs from VideoSet dataset[1] @1080p 30fps
    - Encoding with libaom AV1 in Random Access mode at fixed QP
    - 31 Processed Video Sequences (PVS): encoded with --cq-level ranging from 3 to 63, step of 2

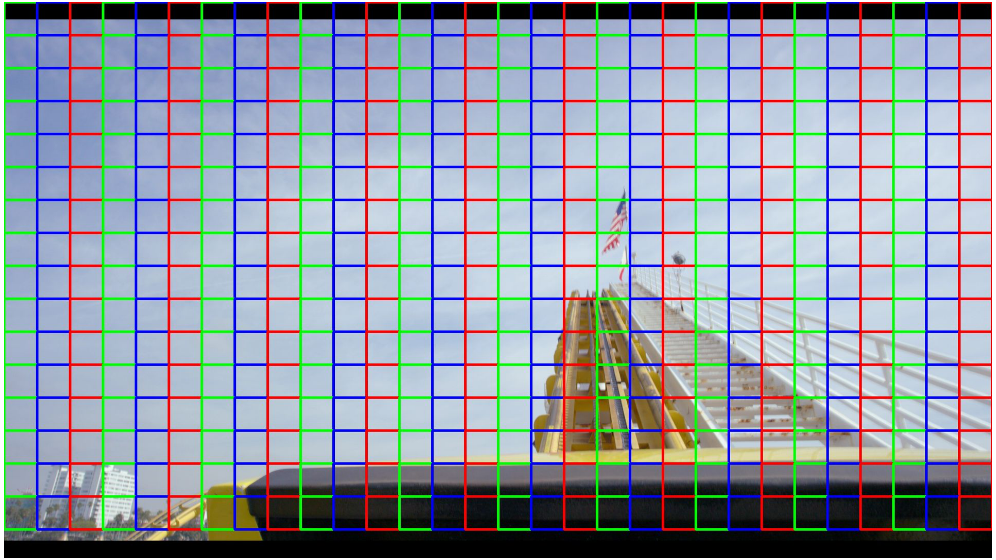


[1] Haiqiang Wang, Ioannis Katsavounidis, Xin Zhou, Jiwu Huang, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, Yun Zhang, Jeonghoon Park, Jiantong Zhou, Shawmin Lei, Sam Kwong, C.-C. Jay Kuo, December 29, 2016, "VideoSet", IEEE Dataport, doi: https://dx.doi.org/10.21227/H2H01C

Content creation: tube-content extraction

To select tube-contents, we followed these steps:

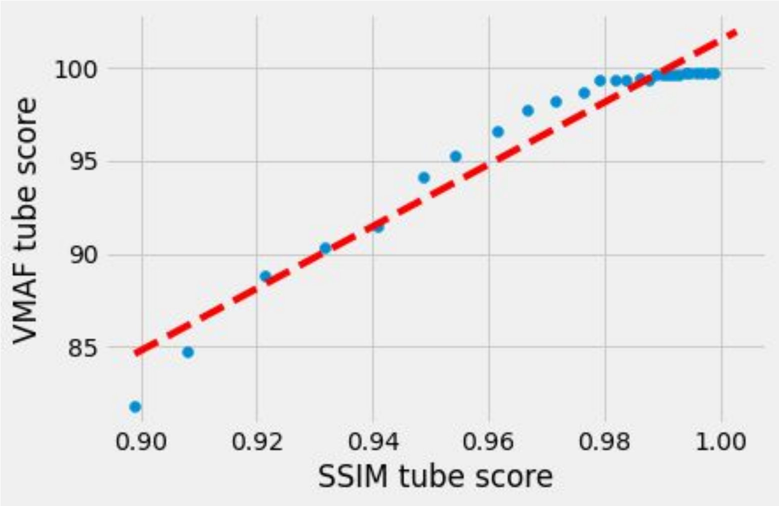
- **Step 2**: Extraction of tube-contents aligned on the motion: tube size = a PU (64x64px, 400ms)
    - A tube-content: a reference tube and 31 distorted version of it from PVS
    - 100K tube-contents extracted from the 115 SRCs



Clustering of tube-contents

To select tube-contents, we followed these steps:

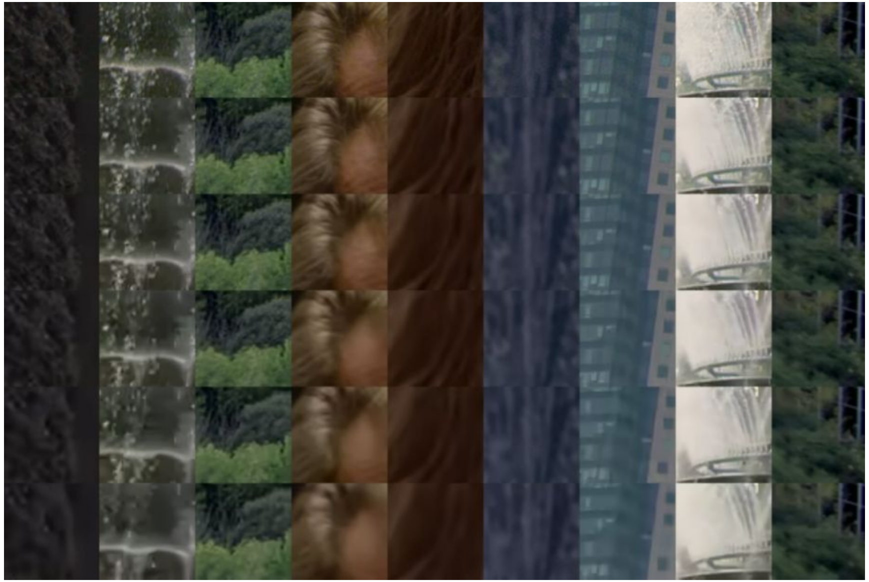
- **Step 3**: Clustering of the 100K tube-contents from the response of quality metrics.
    - Quality metrics used: VMAF, SSIM, PSNR, LPIPS
    - Feature extraction from the relation (red line) in all pairs of quality metrics (slope, intercept, error)
    - 96 clusters are learned with K-Means



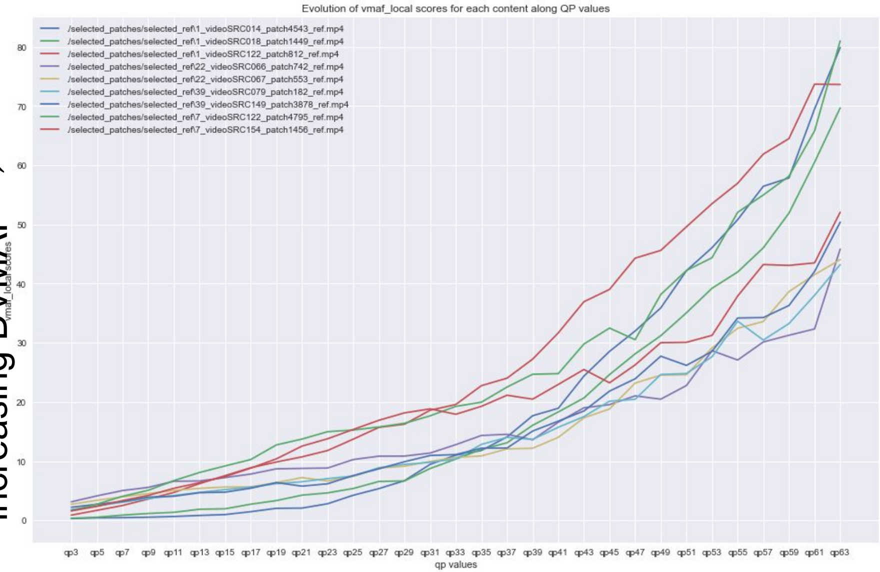
Tube-contents selection for subjective evaluation

To select tube-contents, we followed these steps:

- **Final step**: 268 tube-contents (2+ per cluster) sampled.
  - Per tube-content: 6 distortion levels out of the 31 available are selected using VMAF
    - VMAF as a fidelity proxy for distortion level spacing selection (DVMAF = 100 - VMAF)



increasing DVMAF →

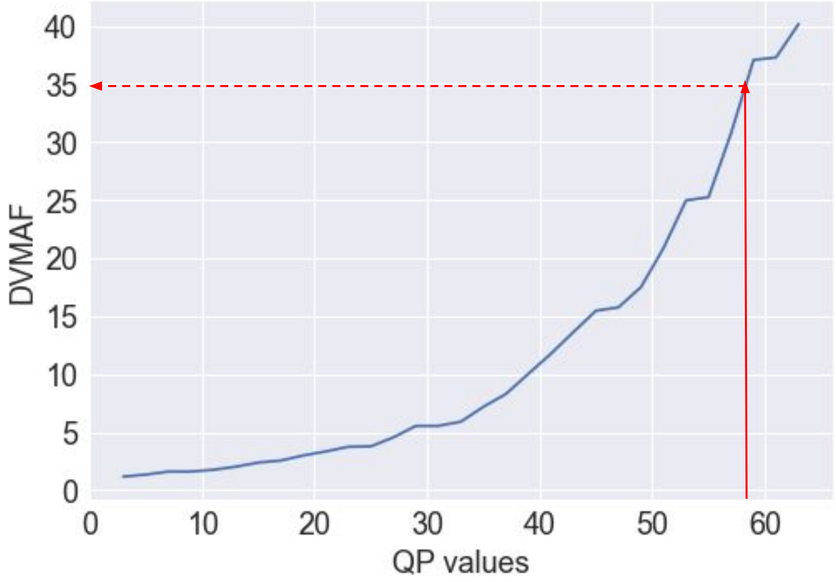


increasing QPs values →

Tube-contents selection for subjective evaluation

To select tube-contents, we followed these steps:

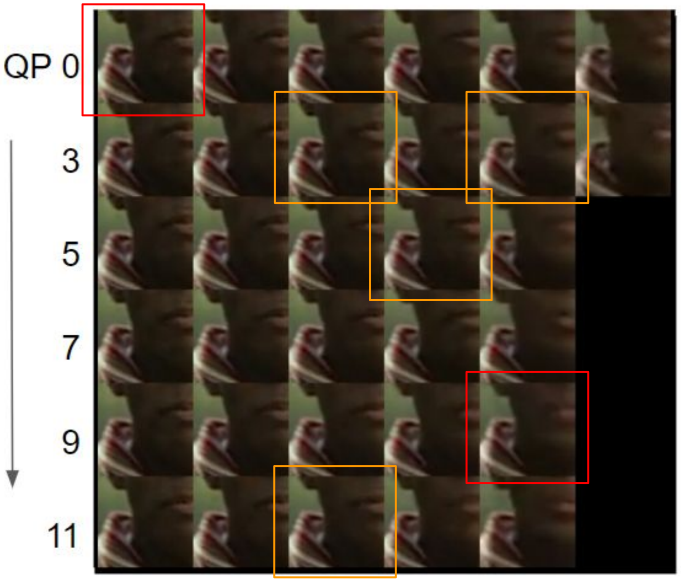
- **Final step**: 268 tube-contents (2+ per cluster) sampled.
    - Per tube-content: 6 distortion levels out of the 31 available are selected using VMAF
        - VMAF as a fidelity proxy for distortion level spacing selection (DVMAF = 100 - VMAF)

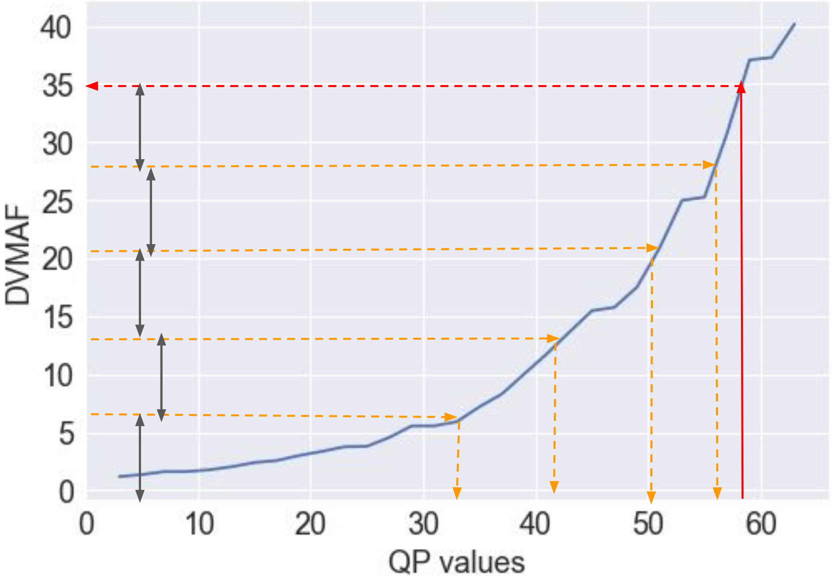


Tube-contents selection for subjective evaluation

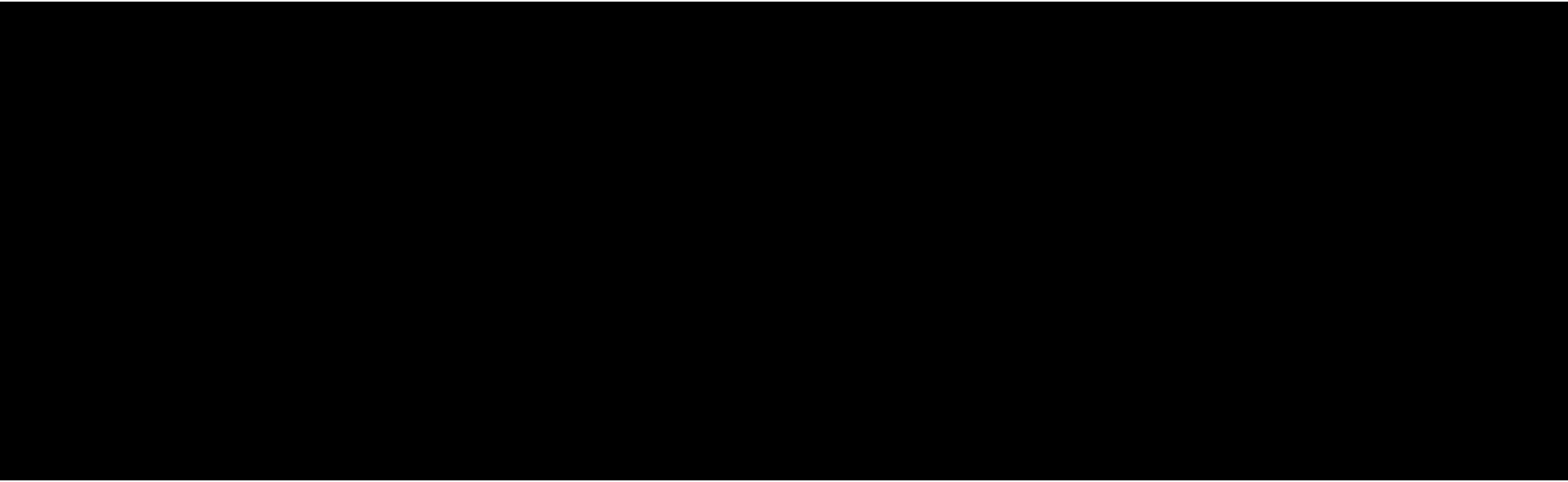
To select tube-contents, we followed these steps:

- **Final step**: 268 tube-contents (2+ per cluster) sampled.
    - Per tube-content: 6 distortion levels out of the 31 available are selected using VMAF
        - VMAF as a fidelity proxy for distortion level spacing selection (DVMAF = 100 - VMAF)



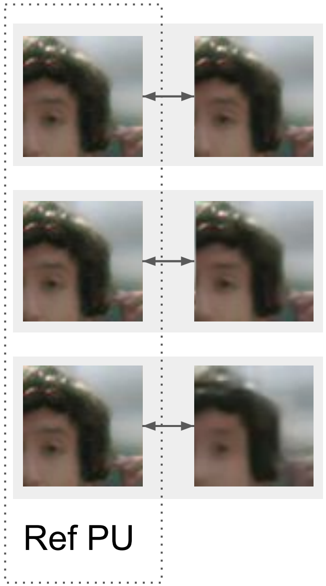


Example of tube-contents and distortion levels?



What kind of subjective data are we trying to collect on a PU?

A fidelity loss evaluation: How much distortions the human eyes can perceive between a reference PU and an encoded/compressed/distorted version of it?



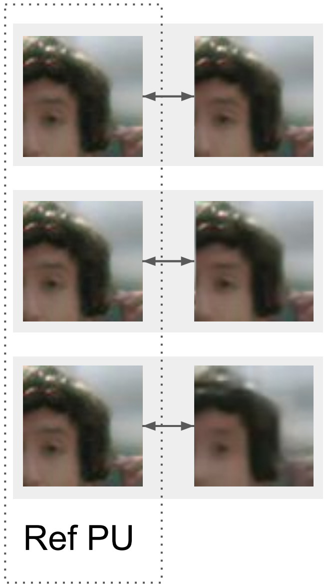
Not noticeable distortion (d = 0)

Noticeable distortion (d > 0)

Very noticeable distortion (d >> 0)

What kind of subjective data are we trying to collect on a PU?

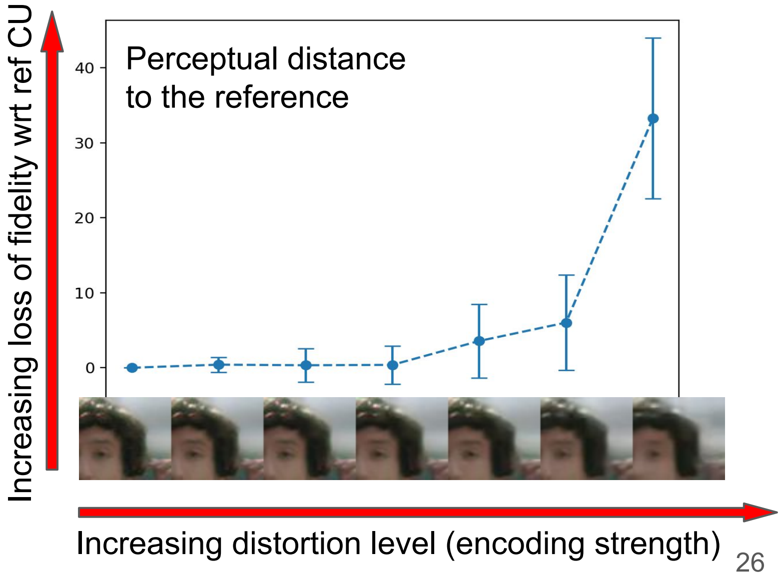
A fidelity loss evaluation: How much distortions the human eyes can perceive between a reference PU and an encoded/compressed/distorted version of it?



Not noticeable distortion (d = 0)

Noticeable distortion (d > 0)

Very noticeable distortion (d >> 0)



Collecting Ground Truth Efficiently

- Available subjective methodologies:
  - Pairwise comparison, (with boosting strategies as ARD, Hybrid-MST[1], ASAP[2] …)
  - **Quadruplets, triplets, 2-AFC, … with boosting strategies AFAD[3]**

- From subjective judgments to perceptual continuum:
  - Bradley-Terry, Thurstonian models, …
  - **Maximum Likelihood Difference Scaling MLDS[4] solvers**

[1] Li, J., Mantiuk, R., Wang, J., Ling, S., & Le Callet, P. (2018). Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation. *Advances in neural information processing systems*, *31*.
[2] Mikhailiuk, A., Wilmot, C., Perez-Ortiz, M., Yue, D., & Mantiuk, R. K. (2021, January). Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 2559-2566). IEEE.
[3] A. Pastor, L. Krasula, X. Zhu, Z. Li and P. Le Callet, "Improving Maximum Likelihood Difference Scaling Method To Measure Inter Content Scale, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2045-2049, doi: 10.1109/ICASSP43922.2022.9746681.
[4] Knoblauch, K., & Maloney, L. T. (2008). MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*, 1-26.

Quadruplet “intra” and “inter-content” comparison

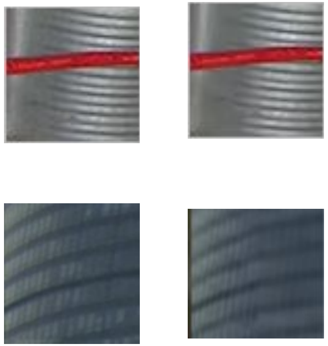
- Participants perform subjective annotations on “intra” and “inter-content” quadruplets
- 50 000 judgments collected, 25 000 “intra” and 25 000 “inter” from naïves observers
- Experiment in crowdsourcing and observers annotated 40 quadruplets per session (~7min)

“INTRA”

Where do you perceive a greater difference between the lower two and the upper two patches?

“INTER”

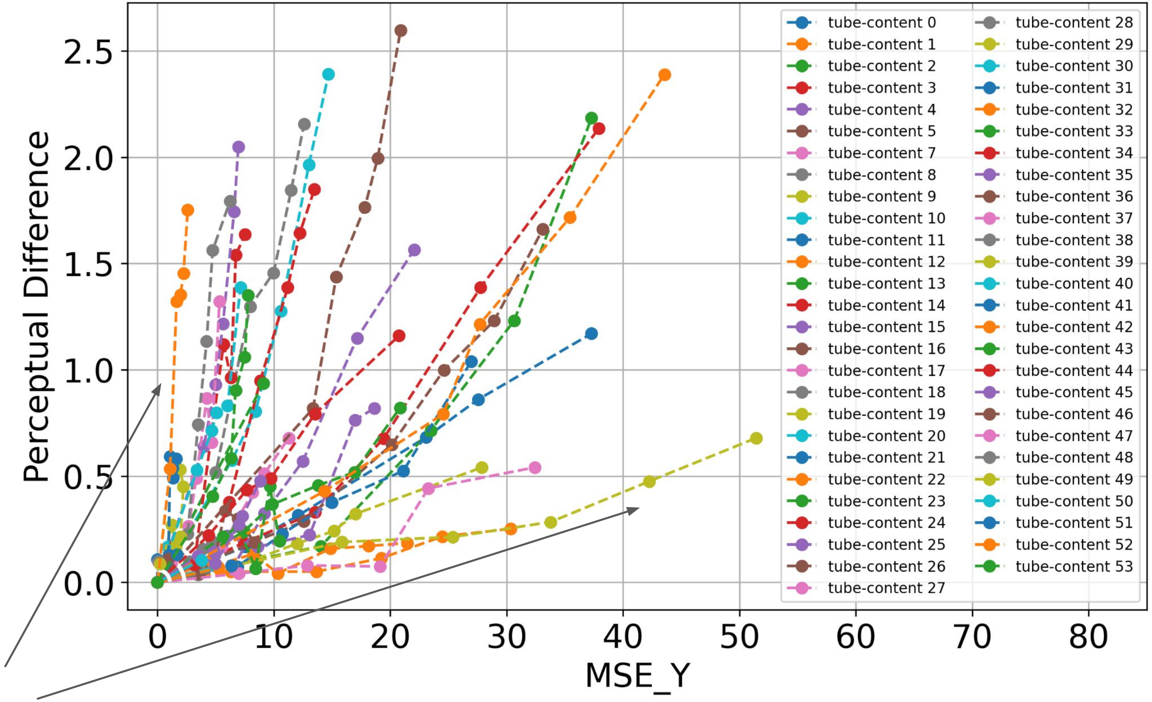
Where do you perceive a greater difference between the lower two and the upper two patches?



Example of PD–MSE curves obtained

Here, the 54 PD–MSE curves in the test set of dataset (20%):

- MSE distortion on X-axis
- subjective perceptual difference from observers on Y-axis

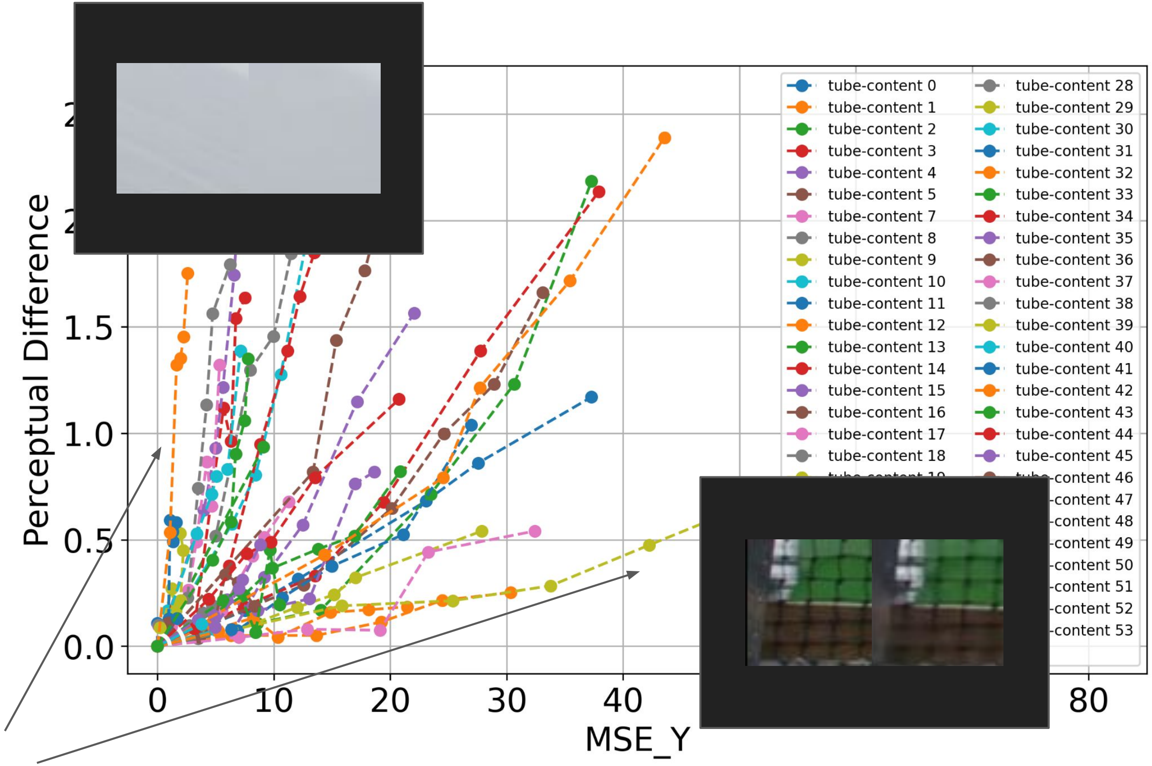


Example of under and over estimated distortions if we use MSE_Y as a PD predictor

Example of PD–MSE curves obtained

Here, the 54 PD–MSE curves in the test set of dataset (20%):

- MSE distortion on X-axis
- subjective perceptual difference from observers on Y-axis



Example of under and over estimated distortions if we use MSE_Y as a PD predictor

Per tube weighting of MSE from Deep Semantic Features

PD-curve modelisation, proposed model, training and performances

proposed model for PD–MSE curve prediction

- step 0: model PD–MSE curves
- step 1: extract deep learning features from references tubes
- step 2: perform dimensionality reduction with PCA
- step 3: use SVM from topK PCA features pooling and predict PD-curves slopes

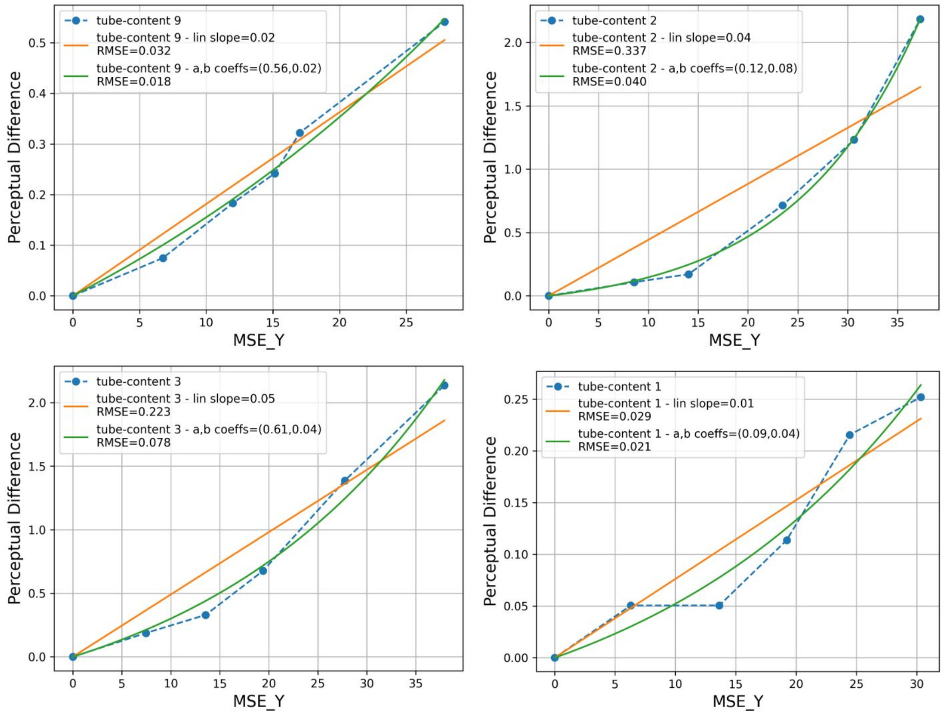
Step 0: modeling of PD–MSE curves

Use prior knowledge to simplify and model PD–MSE curves with linear function (orange) or exp function (green)

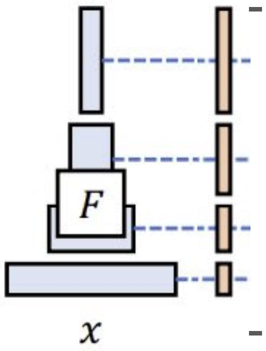
$$PD'_{score} = A \times MSE_Y$$

$$PD'_{score} = A \times (e^{B \times MSE_Y} - 1)$$

Train models to predict linA, and expA + expB



Step 1: extract Deep Learning features from reference tubes



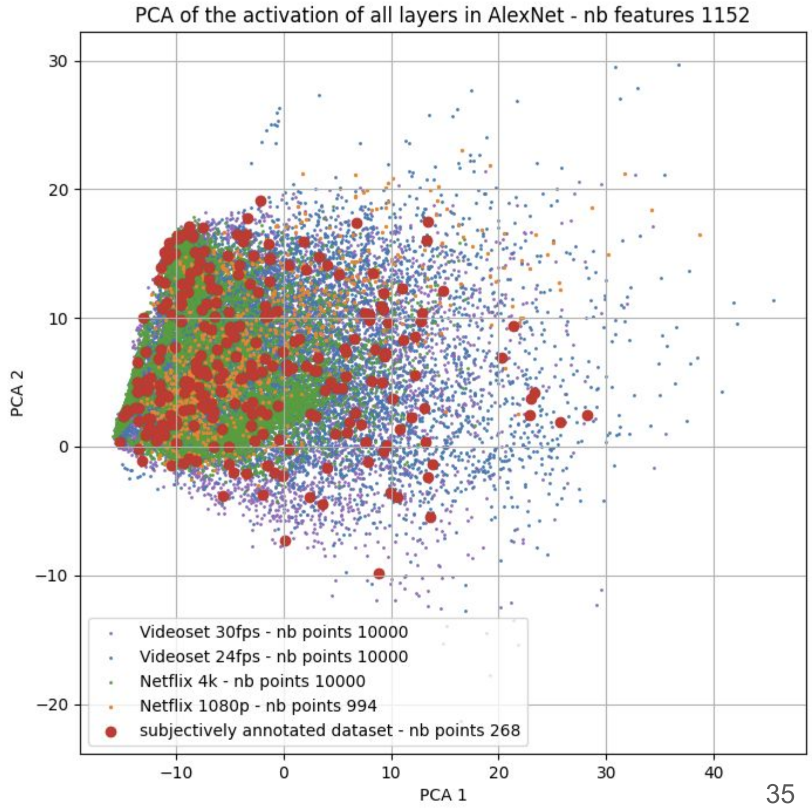
- Why extract DL features from reference tubes only?
    - as we aim to correct MSE, a cheap statistic available during encoding
- Process for each reference tube:
    - pass each frame patch in Neural Network backbone (AlexNet, VGG, …)
    - get each layer filter activation
    - average them along spatial dimension
    - then compute temporal average and temporal std
    - obtain finally 2 vectors of 1152 features (AlexNet) per reference tube
- Perform the operation over the 100K tubes of the database

Step 2: perform dimensionality reduction with PCA

Goal: reduce 1152 features vectors to K features to ease model training on limited data

Use PCA to learn a projection from extracted features from 100K unlabeled tube-contents

Use the learned projection to extract top K Principal Components of train set features



Training options

$$mse_{i,j} = MSE(Tube_{i,0}, Tube_{i,j}) = MSE(Tube_{i,ref}, Tube_{i,j})$$

Learn SVM pooling to predict a subjective score for content i, distortion j:

$$PD_{i,j} = SVM(pca_i^1, pca_i^2, \ldots, mse_{i,j})$$

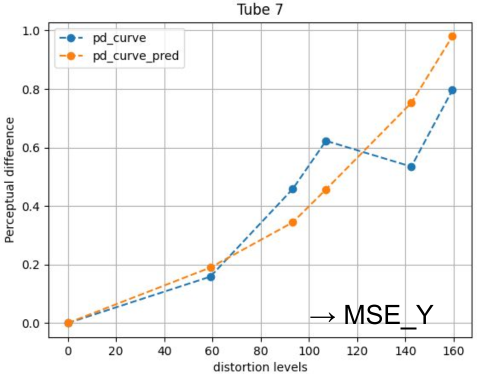
Learn SVM pooling to predict slope of linear fitting

$$Slope_i = SVM(pca_i^1, pca_i^2, \ldots)$$
$$PD_{i,j} = Slope_i \times mse_{i,j}$$

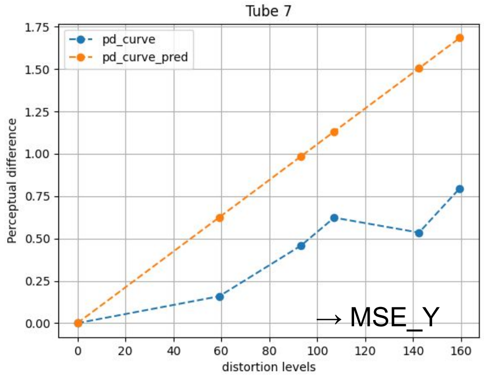
Learn 2 SVMs to predict a,b coeff of exp fitting:

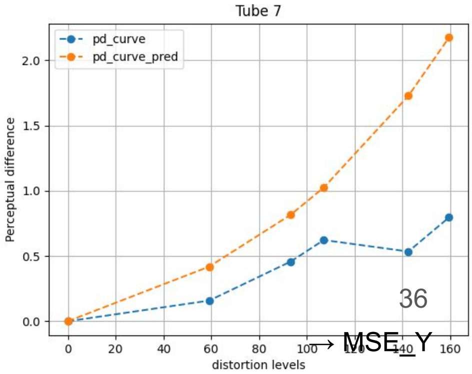
$$a_i = SVM_a(pca_i^1, pca_i^2, \ldots)$$
$$b_i = SVM_b(pca_i^1, pca_i^2, \ldots)$$
$$PD_{i,j} = a_i \times (e^{b \times mse_{i,j}} - 1)$$

No constraint



Add our prior knowledge





Performance of all metrics on test set

Comparison with Full Reference metric (classic and Deep Learning based) and “Reference-only/MSE corrector” metrics

Prior modeling of the PD–MSE curves increases performances

TABLE II
FULL-REFERENCE AND REFERENCE-ONLY METRICS SCORES ON DATASET TEST SET. * INDICATE PERFORMANCES OF RETRAINED METRICS.

Type	Metrics	PLCC	SRCC	KRCC	RMSE
Full-Reference IQA/VQA no semantic	$PSNR_{CB}$	0.472	0.594	0.428	0.535
	$PSNR_{CR}$	0.447	0.539	0.376	0.539
	$PSNR_{Y}$	0.517	0.685	0.507	0.526
	SSIM [4]	0.629	0.763	0.586	0.481
	VIF [22]	0.693	0.780	0.603	0.431
	DLM [23]	0.846	0.869	0.696	0.321
	VMAF [8]	0.833	0.867	0.694	0.335
	VMAF*	*0.875*	*0.900*	*0.747*	*0.291*
DL Full-Reference IQA semantic	LPIPS-vgg [1]	0.711	0.795	0.631	0.420
	LPIPS-squeeze	0.674	0.785	0.622	0.445
	LPIPS-alex	0.628	0.754	0.588	0.470
	DISTS [3]	0.787	0.851	0.671	0.369
Reference-Only no semantic	WPSNR [5]	0.618	0.819	0.642	0.483
	XPSNR [6]	0.665	0.828	0.652	0.461
	libaom tune=ssim	0.653	0.795	0.614	0.476
DL Reference-Only VQA	our model (raw)	**0.844**	0.878	0.714	0.336
	our model (lin)	**0.843**	**0.888**	**0.721**	0.328
	our model (exp)	**0.852**	**0.888**	**0.728**	**0.316**

Conclusion

- Human perception is important to drive encoding algorithms (AV1, …)
- Creation of a dataset of 268 tube-contents with inter-content scaling
- Benchmark of existing quality metrics
- Creation of a metric to correct MSE
- Ongoing next steps:
    - Perceptually tuned Rate Distortion Optimization in libaom
    - going from local to global video scale distortion prediction