User-Generated Video Quality Prediction:

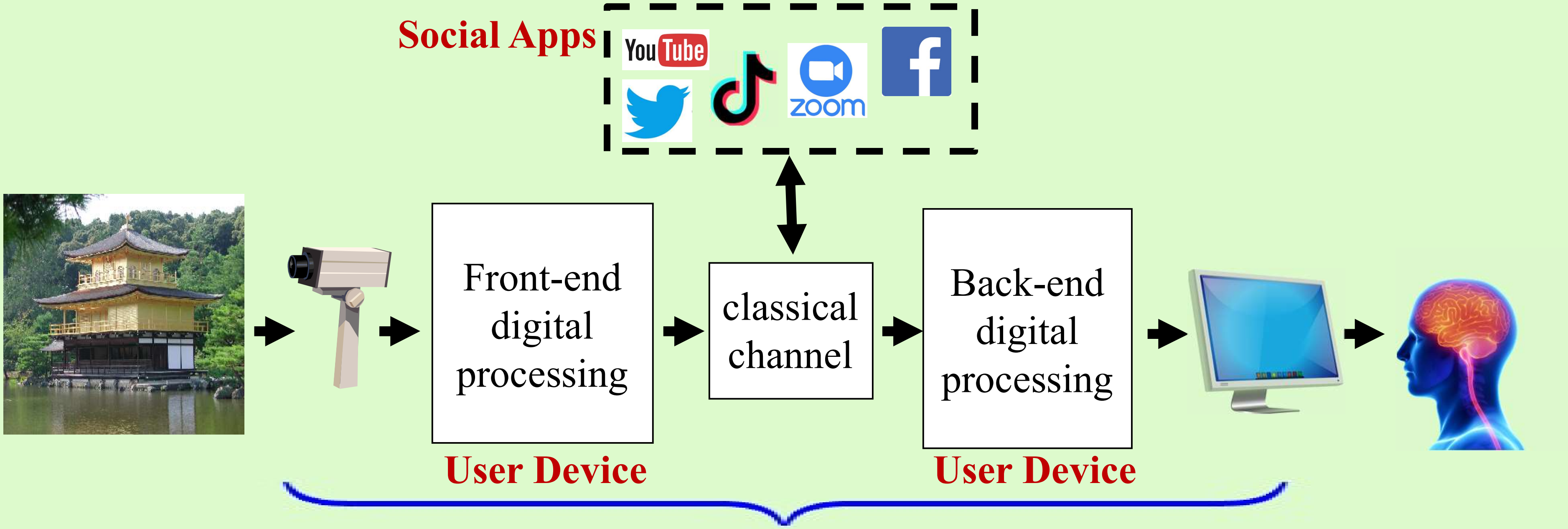
From Local to Global

Al Bovik

Data Compression Conference

March 24, 2021

Natural* Video Communication System



Social Apps

Front-end digital processing

classical channel

Back-end digital processing

User Device

User Device

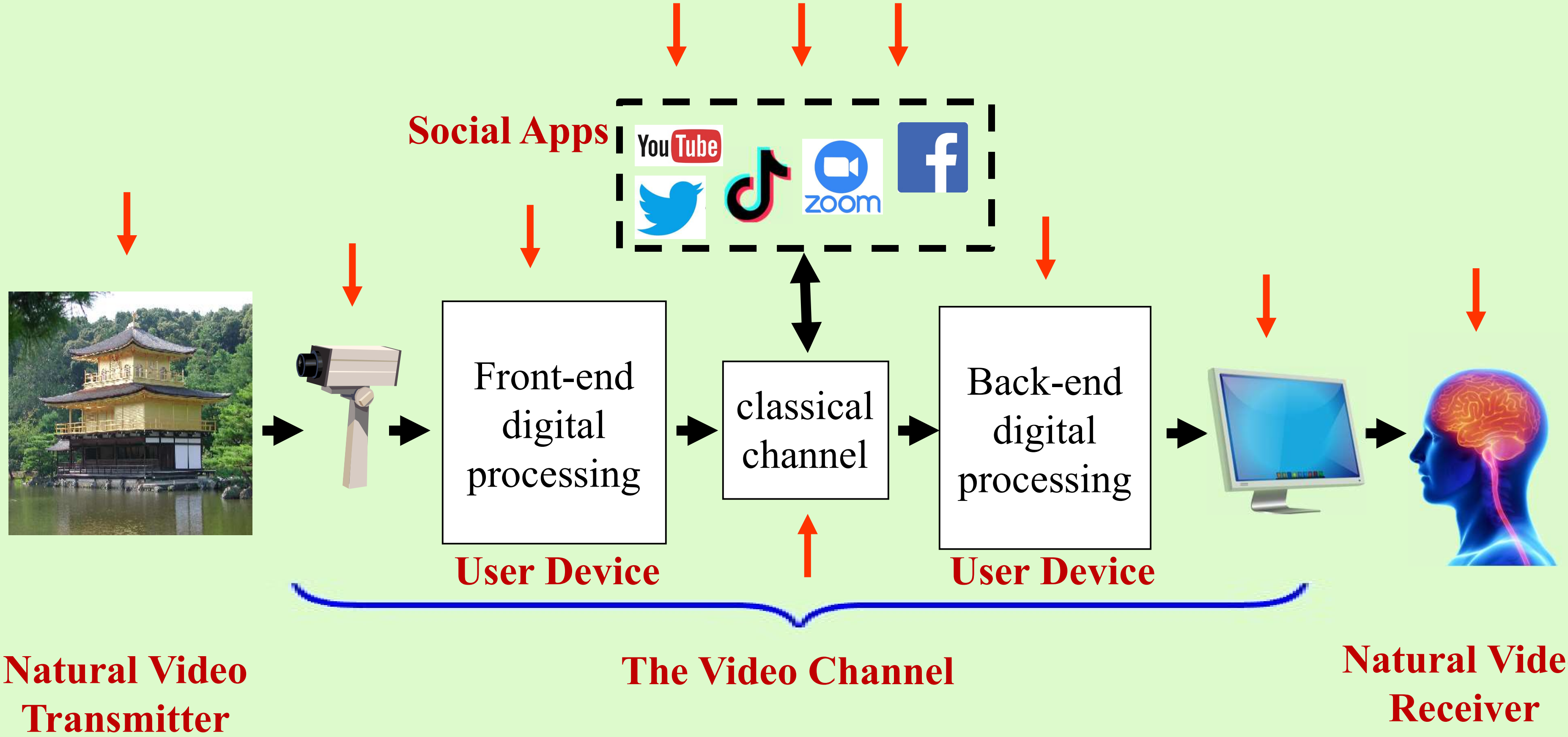
Natural Video Transmitter

The Video Channel

Natural Video Receiver

*Photographic

Sources of Video Distortion



Social Apps

Front-end digital processing

classical channel

Back-end digital processing

User Device

User Device

Natural Video Transmitter

The Video Channel

Natural Video Receiver

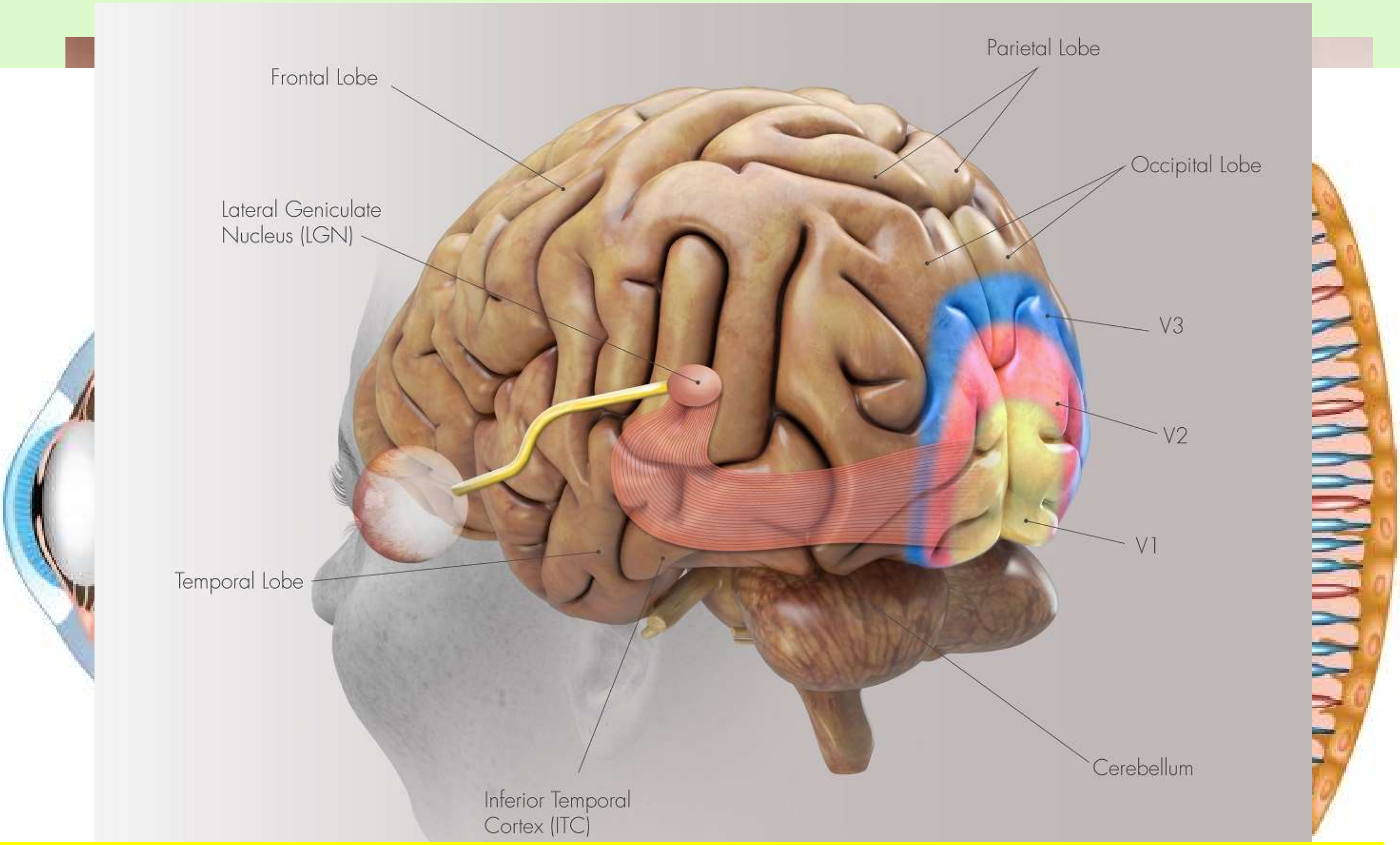
*Photographic

The Natural Video Transmitter

Frames or pictures from the natural video transmitter

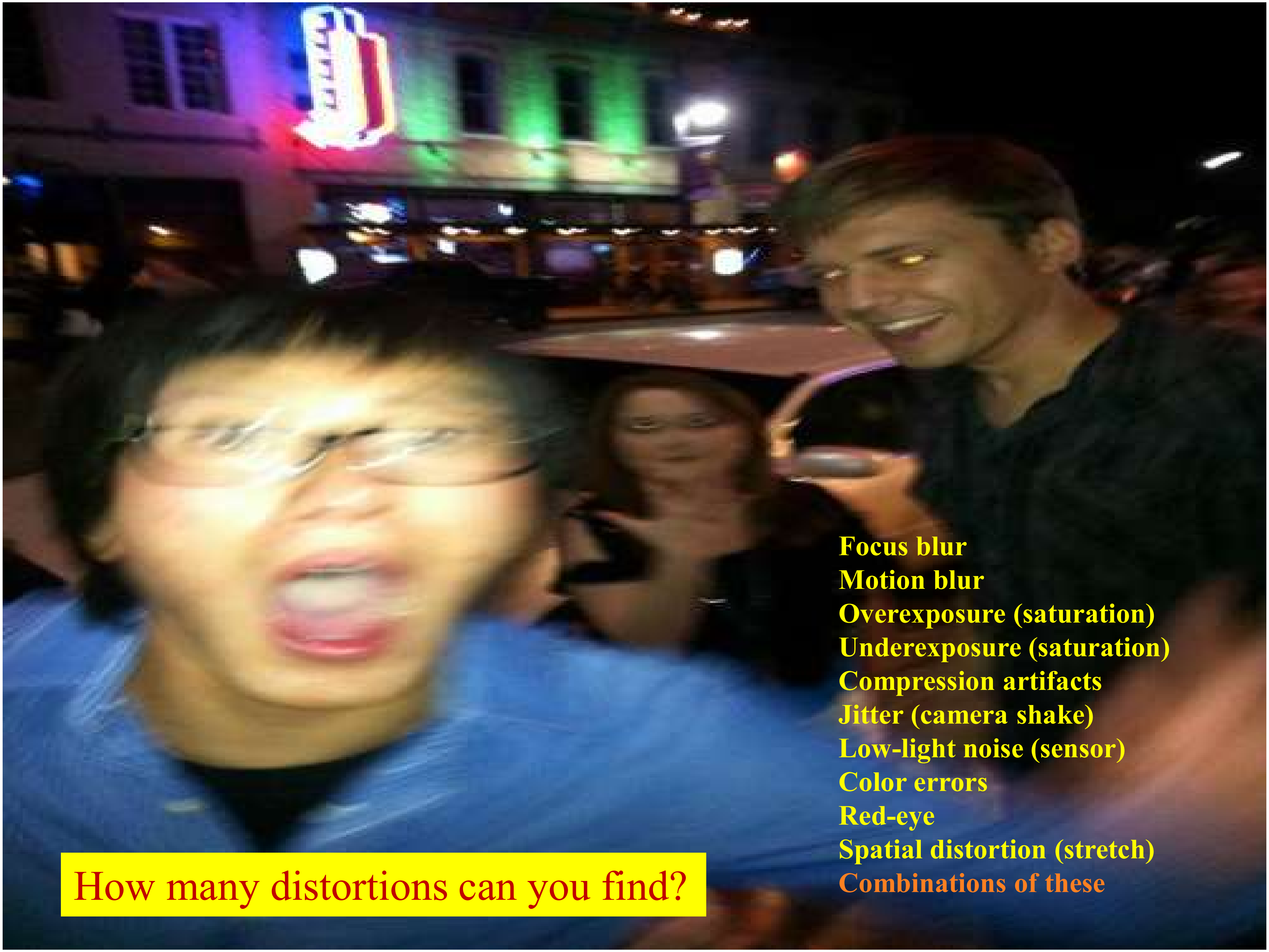
Video from the natural video transmitter

The Natural Image Receiver



The early visual pathway is largely devoted to “video compression”

Video Quality



How many distortions can you find?

Focus blur
Motion blur
Overexposure (saturation)
Underexposure (saturation)
Compression artifacts
Jitter (camera shake)
Low-light noise (sensor)
Color errors
Red-eye
Spatial distortion (stretch)
Combinations of these

Is this a good quality video?

Plethora of Distortions

“Mostly Spatial”

- Blocking artifacts
- Ringing
- Mosaicking
- False contouring
- Motion blur
- Optical blur
- Additive Noise
- Exposure
- Sensor noise
- Shake
- Color errors
- Many more

“Mostly Temporal”

- Ghosting
- Motion blocking
- Motion mismatches
- Mosquito noise
- Stutter
- Judder
- Texture Flutter)
- Jerkiness
- Temporal aliasing
- Smearing
- Many more

Decades of “distortion-specific” measurement didn’t work: couldn’t predict perceived quality well. Too complex to model, too many distortion variations, too many distortion combinations, too hard to map to perception.

UGC Video Quality Prediction is Really Hard! Can we?

Yes, because

Videos are Special

and because distortion changes their specialness

Special Property 1: Reciprocal Law

- The **power spectra** of **videos** f(**x**, t) ~ F(**U**) = F(U, V) and f(**x**, t) ~ F(W) are pretty reliably modeled:

$$E\left[\left|F(\mathbf{U})\right|^2\right] \propto \Omega^{-2\alpha} \qquad \Omega = \sqrt{U^2 + V^2} \qquad (1)$$

$$E\left[\left|F(W)\right|^2\right] \propto W^{-2\beta} \qquad (2)$$

  Ω, W = (radial) spatial, temporal frequency.

- Generally, $\alpha, \beta \in [0.8, 1.5]$ with $\alpha_{ave}, \beta_{ave} \approx 1.2$

- **Functions** (1) or (2) are **uniquely self-similar:**

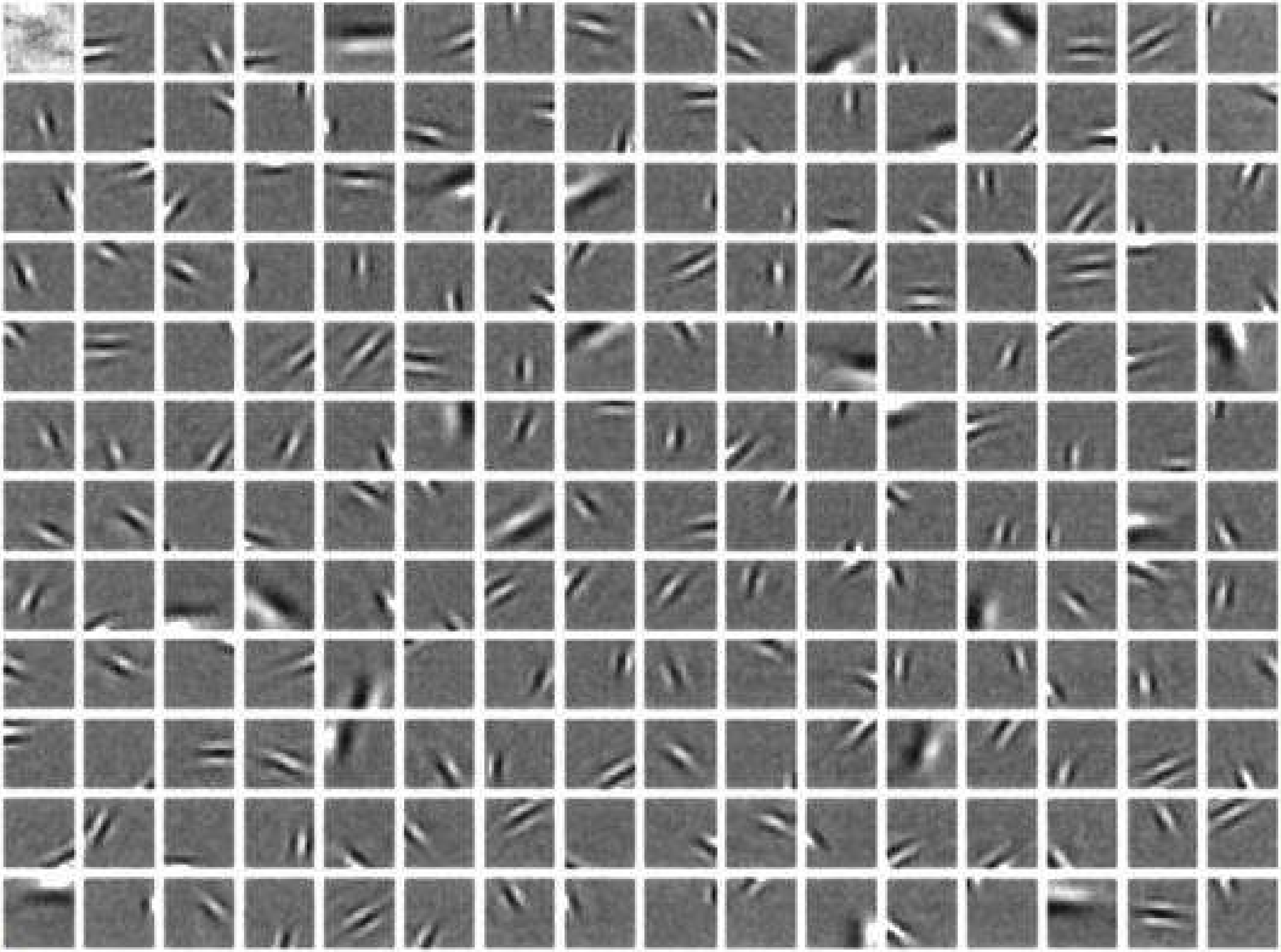
$$\left|F(s\mathbf{U})\right| \propto s^{-\beta} \left|F(\mathbf{U})\right|$$

Football

Alpine Sled

- Videos are **multiscale**, and so is **perception** of **them.**

Tolhurst, et al "Amplitude spectra of natural images," *Ophthal. & Physiol Optics,* 1992.



- sparse code for natural images, *Nature*, 1996.

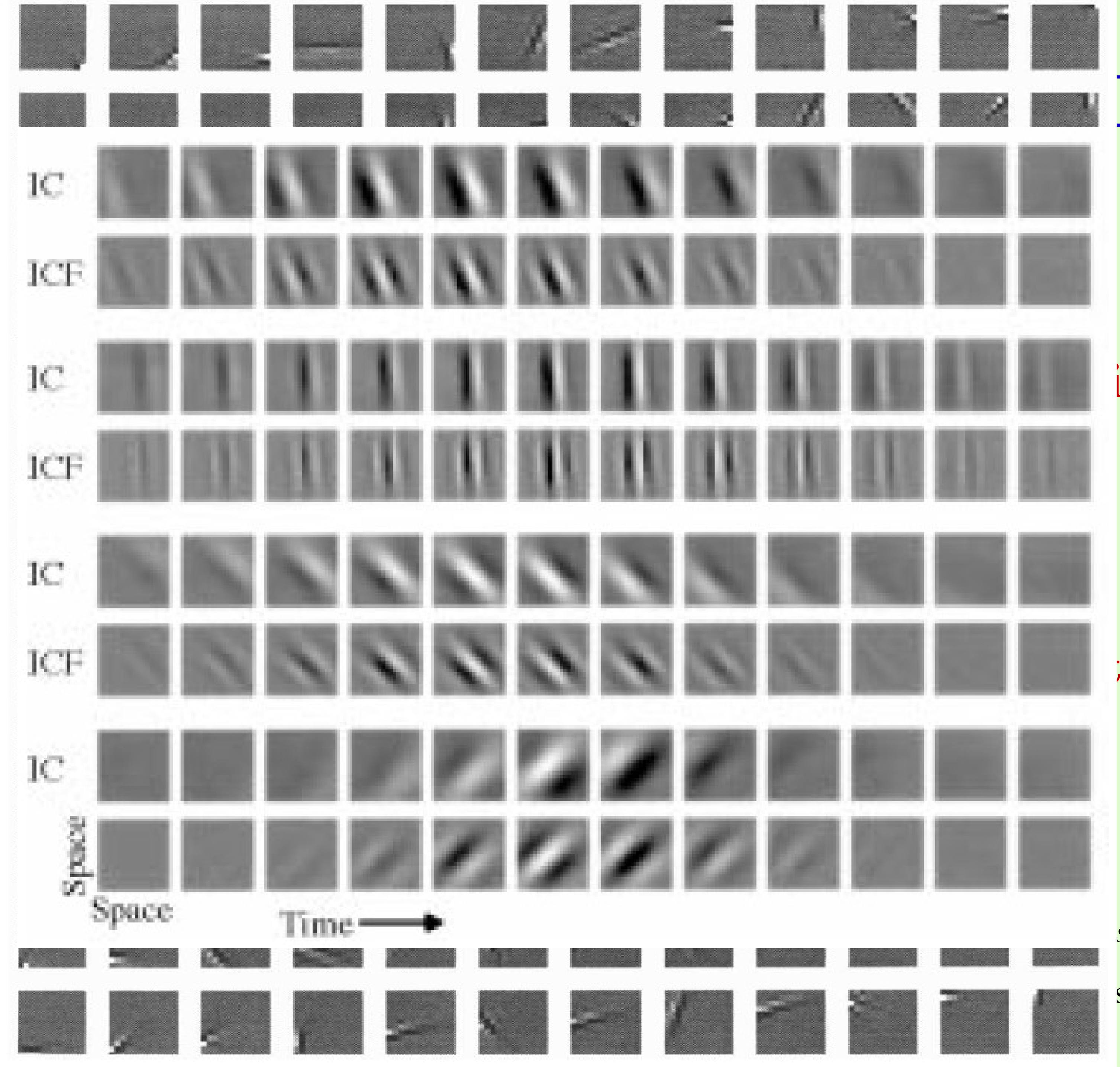
S C's

- I

- A ibution:

- N

- I tive f

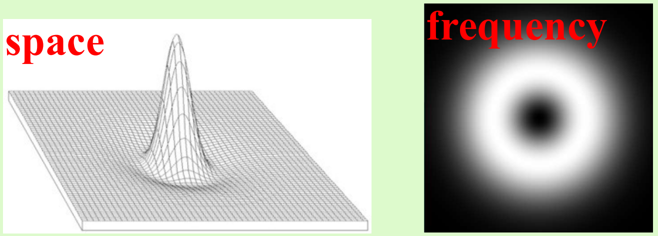


Bell & *search,* 1997.

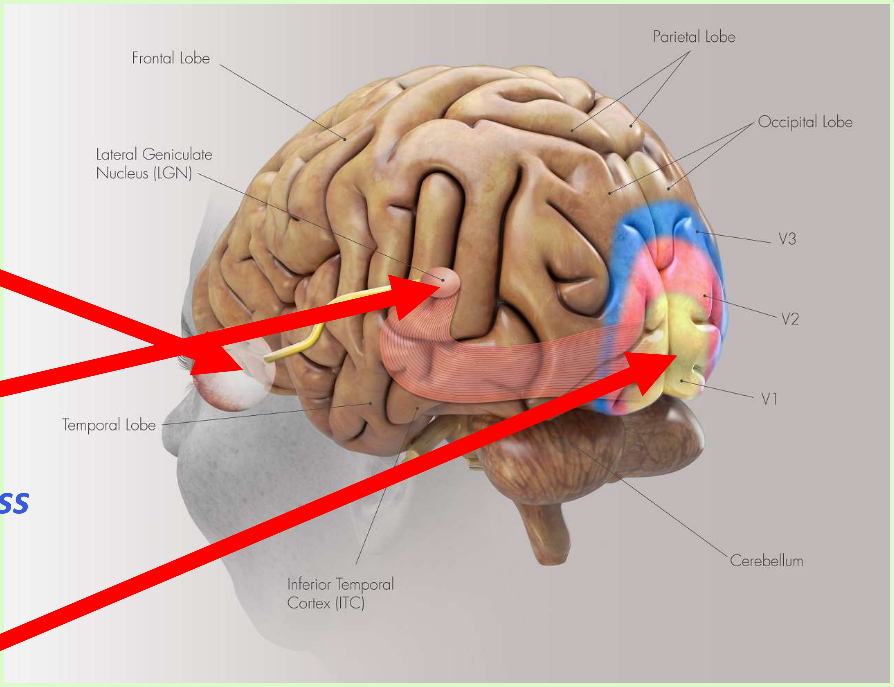
van Ha spatio-tempor

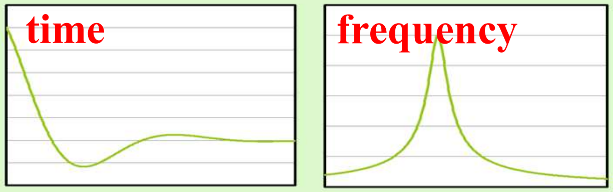
Bandpass Retino-Cortical Filters

- Sparse codes and IC’s of pictures and videos resemble **bandpass receptive field profiles** of **neurons** along **retino-cortical pathway.**

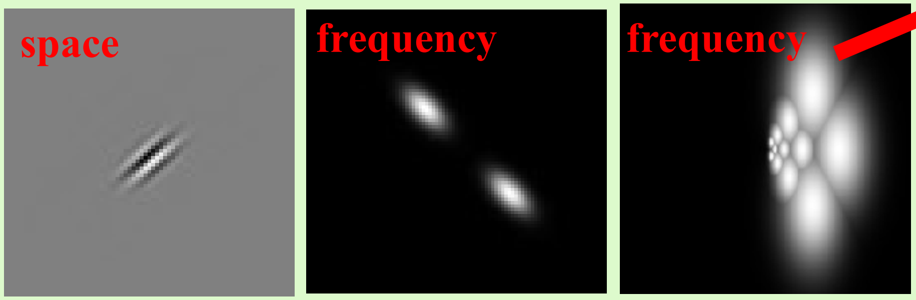


Spatial bandpass predictive coding by retinal ganglion cells …





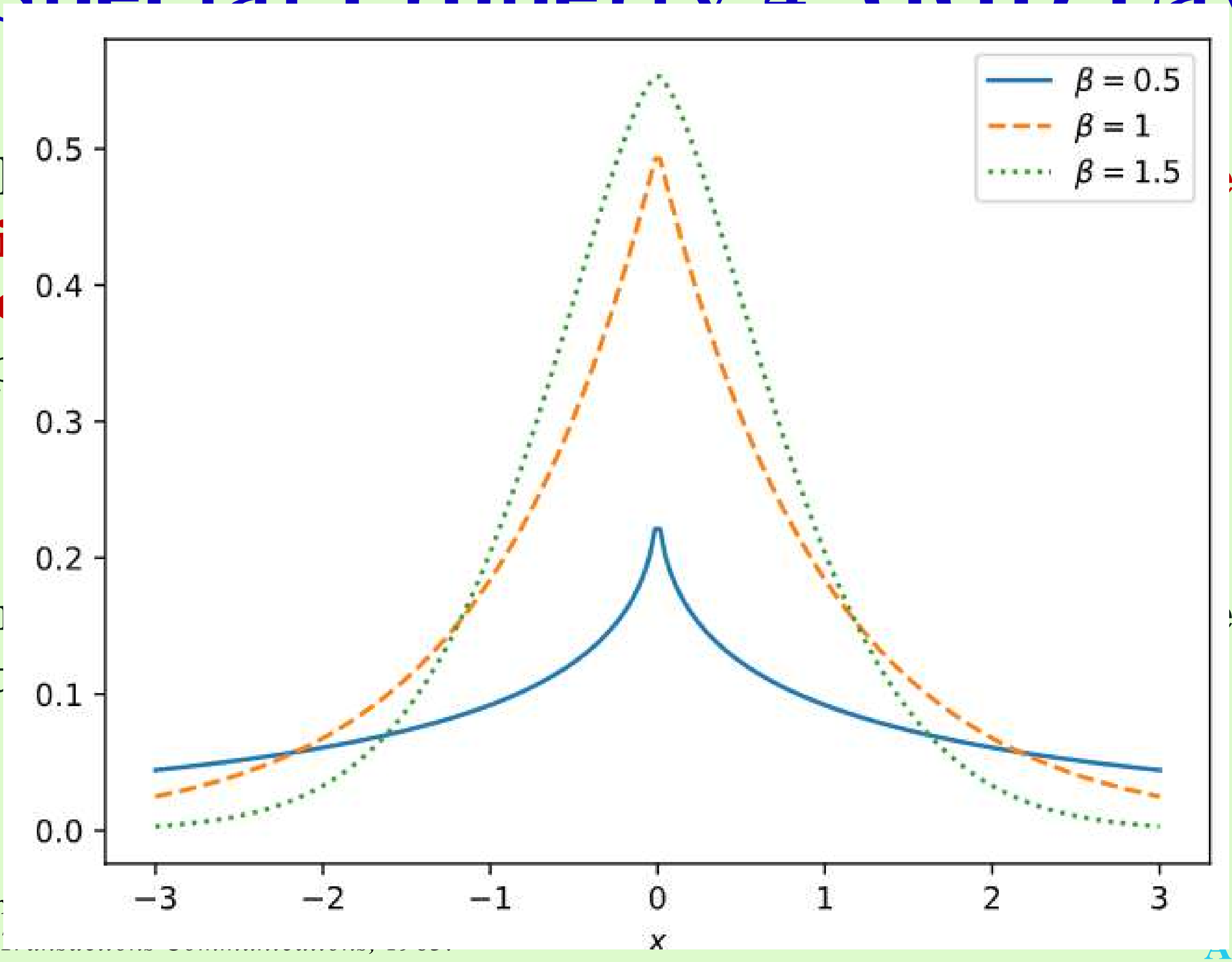
… temporal bandpass coding in LGN …



Bandpass decompositions in visual cortex …

- **Visual neurons “matched”** to **natural image structure** achieving **efficient representations.**
- Similar to **filters** in **early layers of deep nets!**

Special Property 4: GGD Law



- Th **ed vi ge** sp

a > 0

- Tr eff., et

Reinir
IEEE T

Mallat, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Transactions PAMI,* 1989.

Alpine Sled Basketball

Special Property 5: Gaussian Law

- An even more useful model of **bandpass videos f** is the **gaussian scale mixture (GSM).** If (h = BPF)

  $$g(\mathbf{m}) = f(\mathbf{m}) * h(\mathbf{m})$$

  then space/time/scale n'brhoods of g(**m**) are **well-modeled**

  $$\overline{g}(\mathbf{m}) \sim z(\mathbf{m}) \cdot \overline{\gamma}(\mathbf{m})$$

  where z(m) is a **scalar (variance) random field** and

  $$\overline{\boldsymbol{\gamma}}(\mathbf{m}) \sim \eta(0, C_{\overline{\gamma}}) \quad C_{\overline{\gamma}} = \text{near-diagonal covariance matrix of } \overline{\gamma}$$

- Implies **divisive normalization** by local space/time/scale energies further **decorrelates** & **gaussianizes.**

G re

- If $\overline{\mathbf{g}}(\mathbf{m},$ y

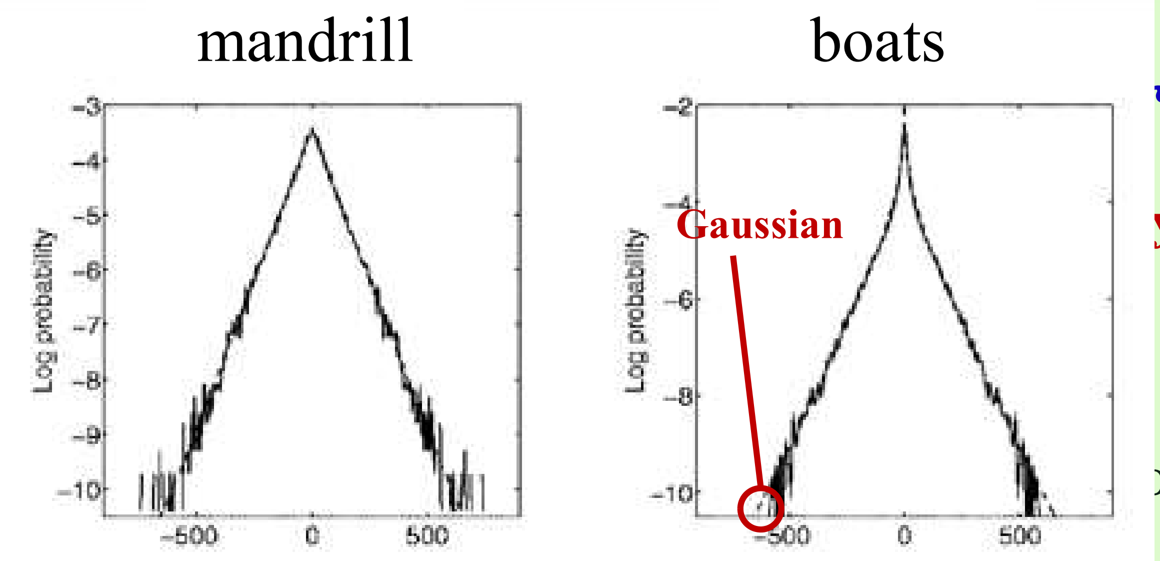
  p

- **ML est** oeff. ρ:

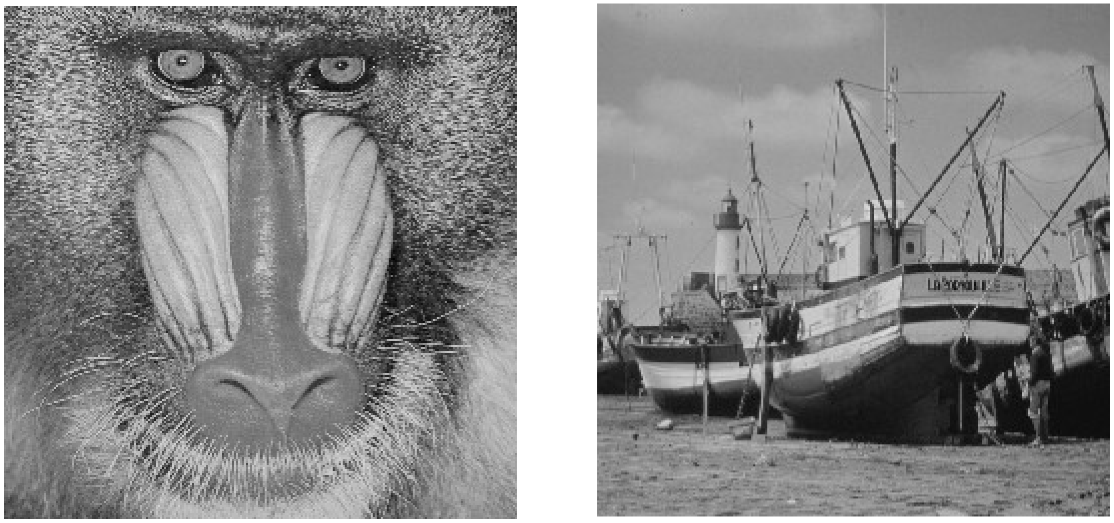
- **Dividi** **on**) yields approxi

- **The ur** **is a great r** **ng.**

- **The vi** **tortions alter it**



Bandpass, divisively normalized pictures



Original images

Ruderman, The st 1994.

M.J. Wainwright and ges,”
Advances in Neural I

Normalization of Sensory Neurons

- A lot like **layer normalization** in deep nets but localized.

Heeger, Normalization of cell responses in cat striate cortex, *Visual Neuroscience*, 1992.

Formulating
General Video Quality
Paradigms
by
Exploiting the Dual Nature Between Natural Video Statistics and Sensory Processing

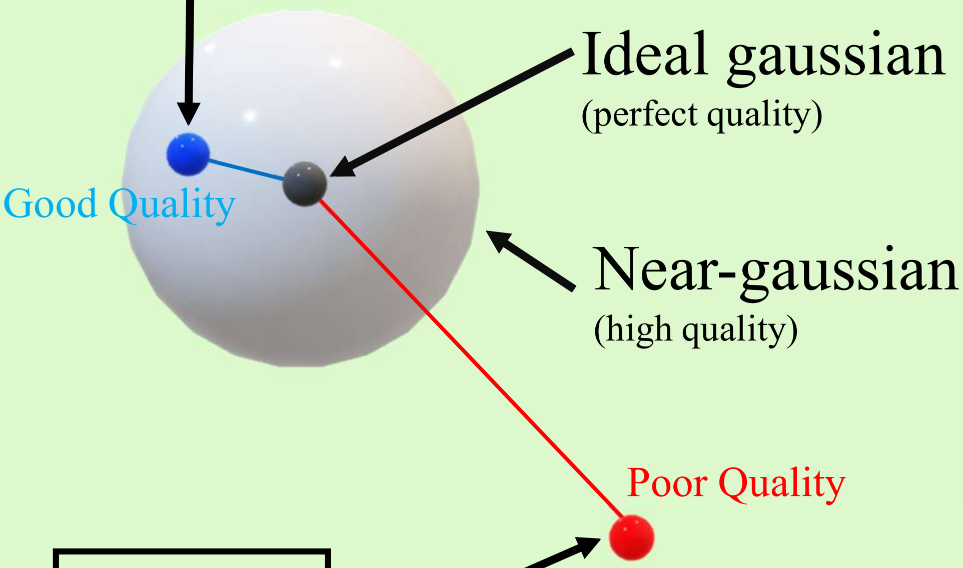
(Very) General Quality Measurement Concept

After **perceptual processing** (bandpass + normalize), **quality prediction** cast as statistical **distance measurement.**

Perceptual Processing Model

Distort

Perceptual Processing Model

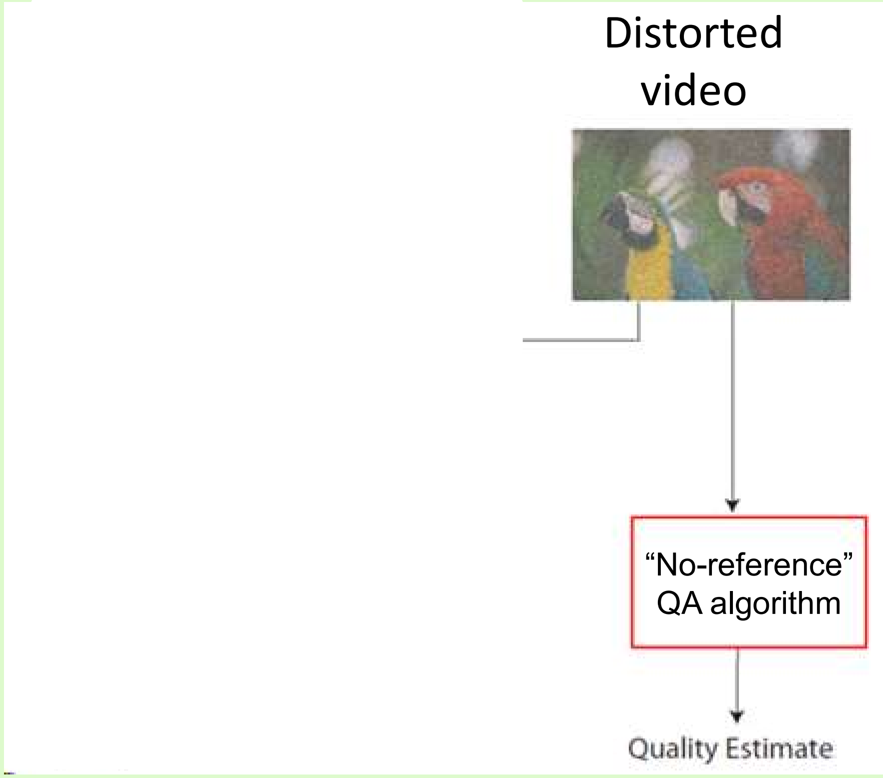


How to define perceptual quality distances?

Reference vs. No-Reference

"Reference" VQA:

- **Perceptually compare** videos against "pristine" **references**
- Really measures **"perceptual fidelity"**

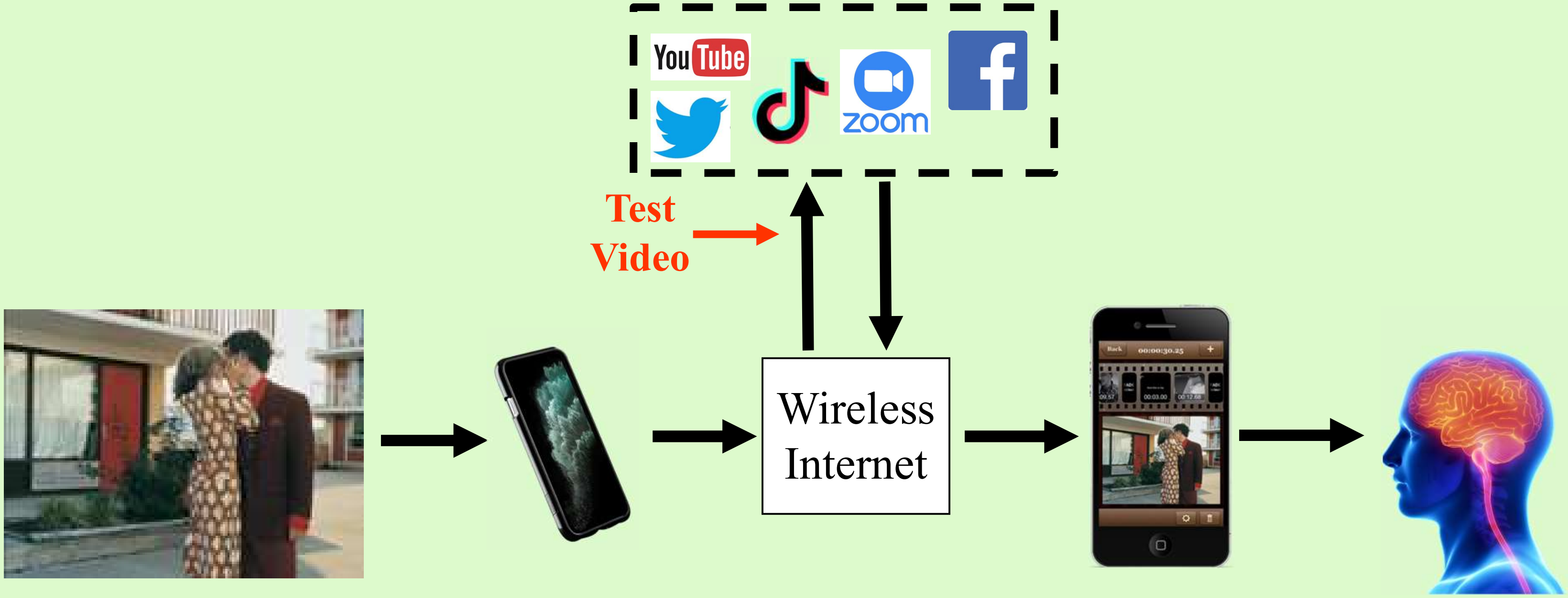


"No-Reference" VQA

- **No reference!**
- Also called **Blind VQA**
- Most common **UGC** scenario
- Pure **perceptual quality prediction**

No-reference (blind) VQA (especially of UGC) is a much harder, much sought-after problem.

No-Reference VQA



Test Video

Wireless Internet

This is what is required for UGC videos:
SSIM, VMAF, etc can’t be used.

BRISQUE

(Blind VQA)

Statistical Models

Gaussian Scale Mixture (GSM)

- **Bandpass** preprocess natural video

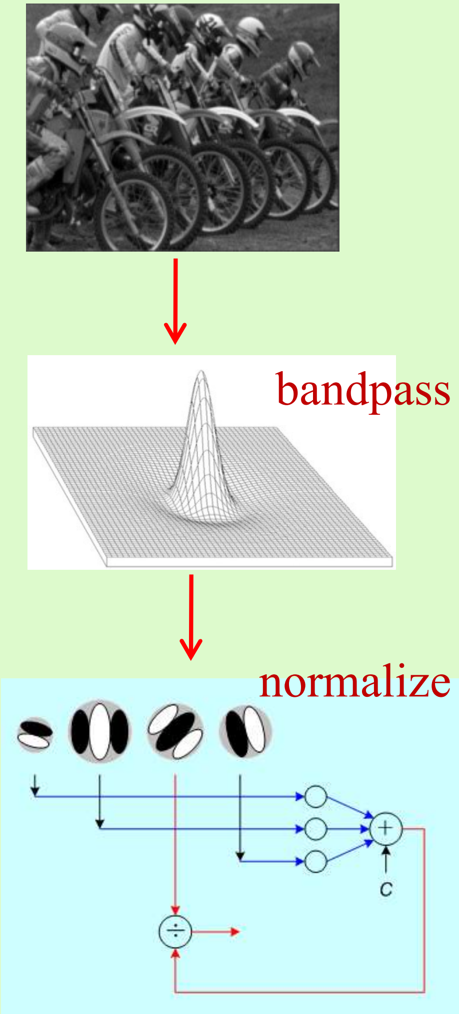
- Response **well-modeled** as

$$\bar{g}(\mathbf{m}) \sim z(\mathbf{m}) \cdot \bar{\gamma}(\mathbf{m})$$

$$\bar{\boldsymbol{\gamma}}(\mathbf{m}) \sim \eta(0, 1)$$

where z = **variance / correlation field**

- Estimate **local variance** z and normalize / decorrelate:



Images of the world have an essential UNDERLYING GAUSSIANITY

Natural Scene Statistic Model

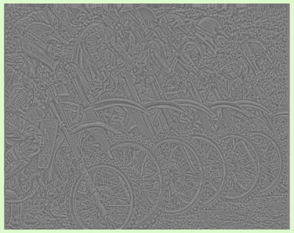
Gaussian Property:

If $$\text{MSCN}(\mathbf{x}) = \frac{f(\mathbf{x})-\mu(\mathbf{x})}{\sigma(\mathbf{x})+1}$$

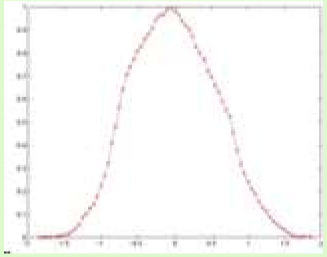
then $$\text{MSCN}(\mathbf{x}) \sim \frac{1}{\sqrt{2\pi}}\exp\left(-a^2/2\right)$$

$$\mu(\mathbf{x}) = \sum\sum w(\mathbf{y})f(\mathbf{x}\text{-}\mathbf{y}) \qquad \sigma(\mathbf{x}) = \sqrt{\sum\sum w(\mathbf{y})\left[f(\mathbf{x}\text{-}\mathbf{y})\text{-}\mu(\mathbf{x}\text{-}\mathbf{y})\right]^2}$$

video f



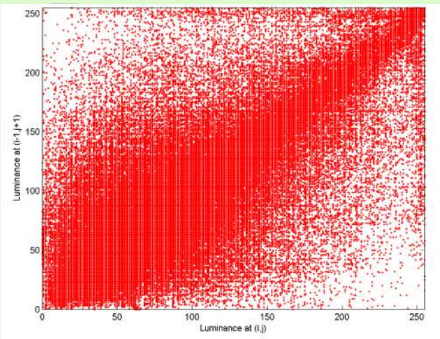
$\text{MSCN} = \frac{f-\mu}{\sigma+1}$



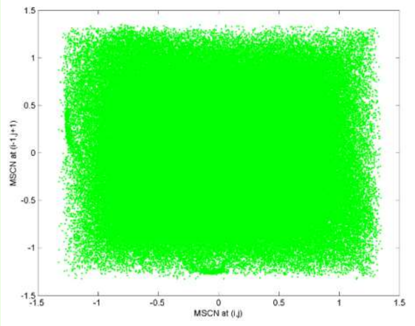
MSCN histogram

MSCN = “mean-subtracted, contrast normalized”:
a basic retinal model

Decorrelation



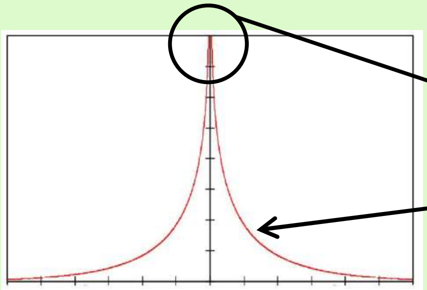
$f(\mathbf{x})$ vs $f(\mathbf{x}\pm1)$



$\text{MSCN}(\mathbf{x})$ vs $\text{MSCN}(\mathbf{x}\pm1)$

$$\text{MSCN}(\mathbf{x})\cdot\text{MSCN}(\mathbf{x}\pm1) \sim C_2\,K_0\left(\left|a\right|\right)$$

K_0 = modified Bessel function of the second kind



Small matter of infinity

Symmetric

Distortion Statistics

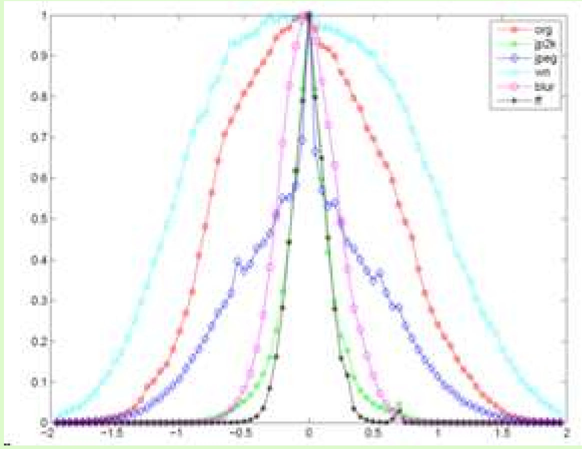
- Distortions **destroy gaussianity** of

$$\mathrm{MSCN}(\mathbf{x}) = \frac{f(\mathbf{x})-\mu(\mathbf{x})}{\sigma(\mathbf{x})+1}$$

- But most are well-modeled as **generalized gaussian (GGD)**

$$\mathrm{MSCN}_{\mathrm{distorted}}(\mathbf{x}) \sim C_2 \exp\left(-|a|/\sigma\right)^{\gamma}$$

Two distortion features

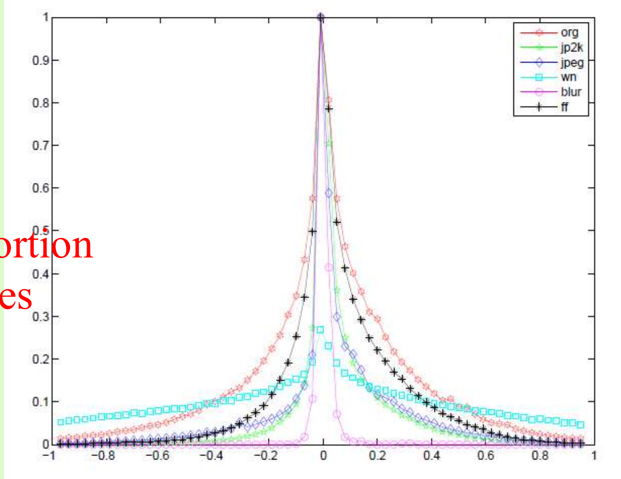


Point histogram of MSCN

- Distortions **introduce correlations**
- Hence **product distribution** becomes **asymmetric**
- Hence use an **asymmetric GG model** $(\mu \neq 0)$

$$\mathrm{MSCN}(\mathbf{x})\cdot\mathrm{MSCN}(\mathbf{x}\pm 1) \sim C_3 \begin{cases} \exp\left[-\left(a/\sigma_L\right)^{\gamma}\right]; \ a<0 \\ \exp\left[-\left(a/\sigma_R\right)^{\gamma}\right]; \ a\geq 0 \end{cases}$$

Four distortion features

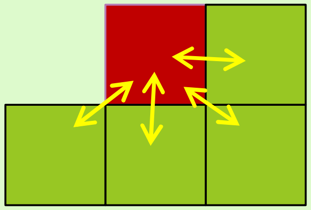


Pairwise product histogram of MSCN

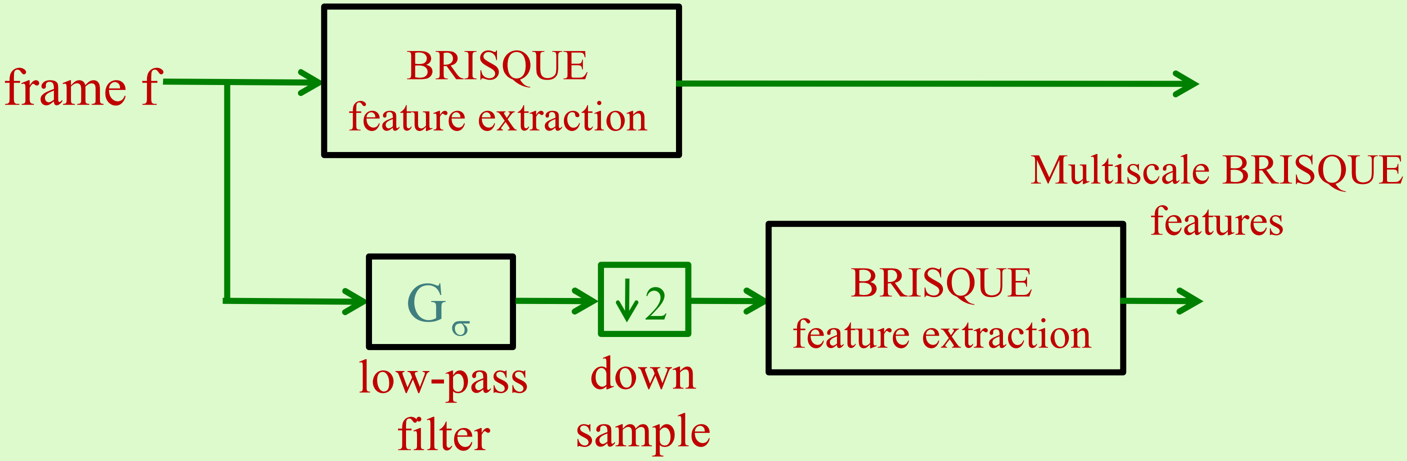
- When **no distortion,** expect $\sigma_L = \sigma_R$.

BRISQUE Features

- Univariate features: γ, σ **(2 features)**
- Product features η, γ, σ_L, σ_R along four orientations **(16 features)**



- Over **multiple scales** (just 2 in basic BRISQUE)



frame f

BRISQUE feature extraction

G_σ

low-pass filter

$\downarrow 2$

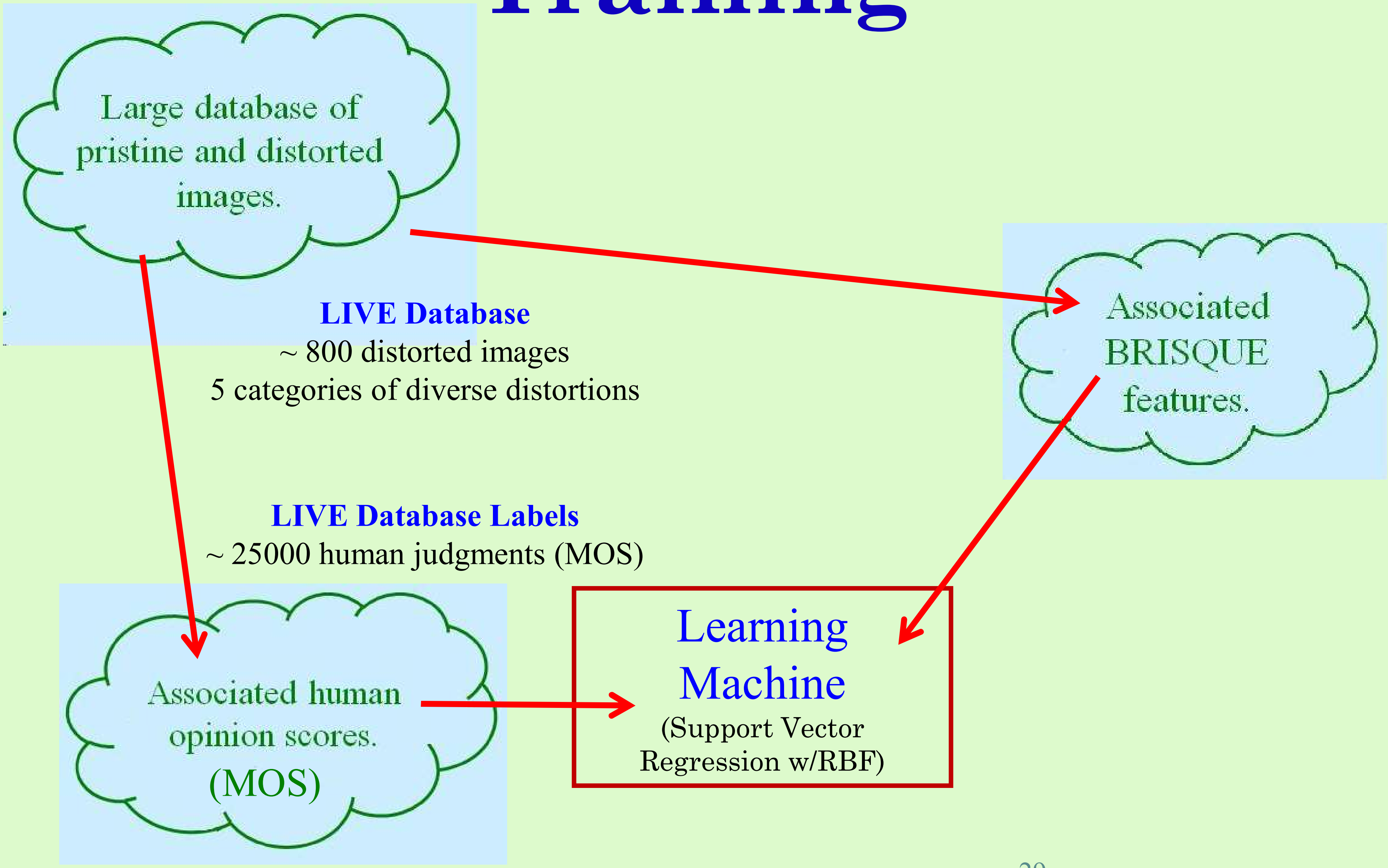
down sample

BRISQUE feature extraction

Multiscale BRISQUE features

36 features overall

Training



Large database of pristine and distorted images.

LIVE Database

~ 800 distorted images

5 categories of diverse distortions

Associated BRISQUE features.

LIVE Database Labels

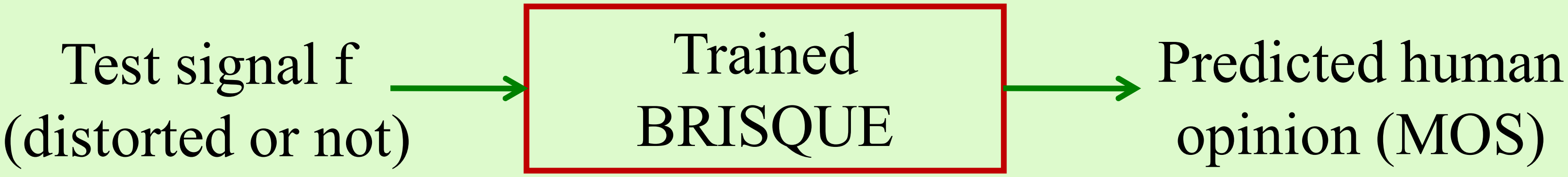
~ 25000 human judgments (MOS)

Associated human opinion scores. (MOS)

Learning Machine

(Support Vector Regression w/RBF)

Application



Test signal f (distorted or not)

Trained BRISQUE

Predicted human opinion (MOS)

But this is an old and easy database of single, synthetic distortions applied by the experimenters (us). BRISQUE does not do well on real UGC distortions.

Median linear correlation coefficient against real human opinions, 1000 train-test random divisions of the LIVE Image Quality Database

Comments

- BRISQUE and its derivative "NIQE" (unsupervised version) are marketed and **used worldwide.**

- **Example:** Quality-controlled **transcoding** of high-quality streaming video content in the cloud.

- **Performance is poor** on real-world **user-generated content** (UGC) – like much YouTube/Facebook content.

- We’ve created **"advanced BRISQUE"** models having dozens to 1000s of NSS features (time, color, scale, correlation distance, σ-field analysis, etc), with **some success.** One is called **VIDEVAL.**

Deep Blind Video Quality

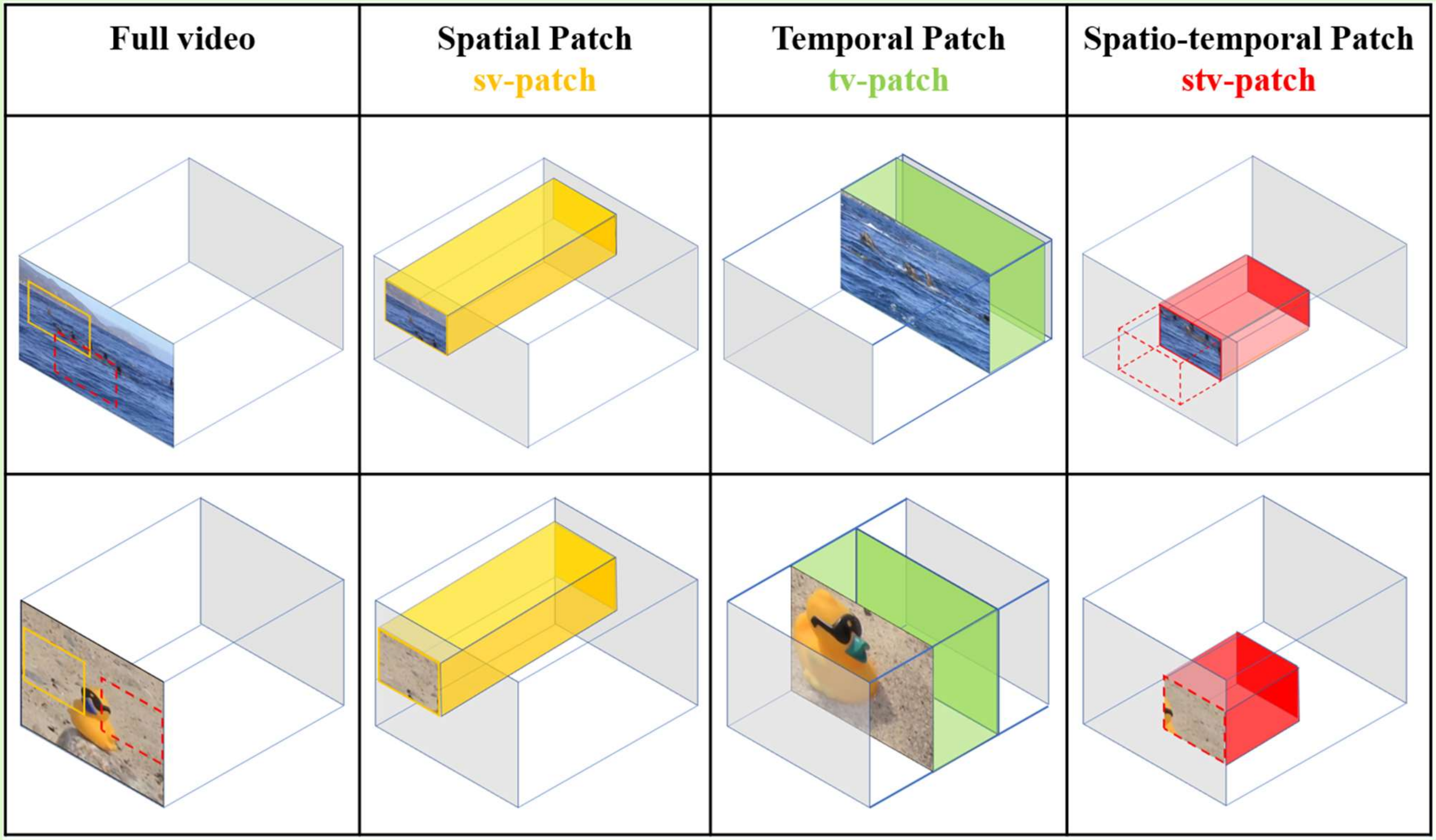
Zhenqiang Ying

Mani Mandal

- Col
  Res

- **Ur**
  - **V**
    l
    l
  - ((**v-**
    l **patio-**
    t
  - T **al**
    s nd
    l

LIVE-FB LSVQ Database Exemplar Patch Sampling



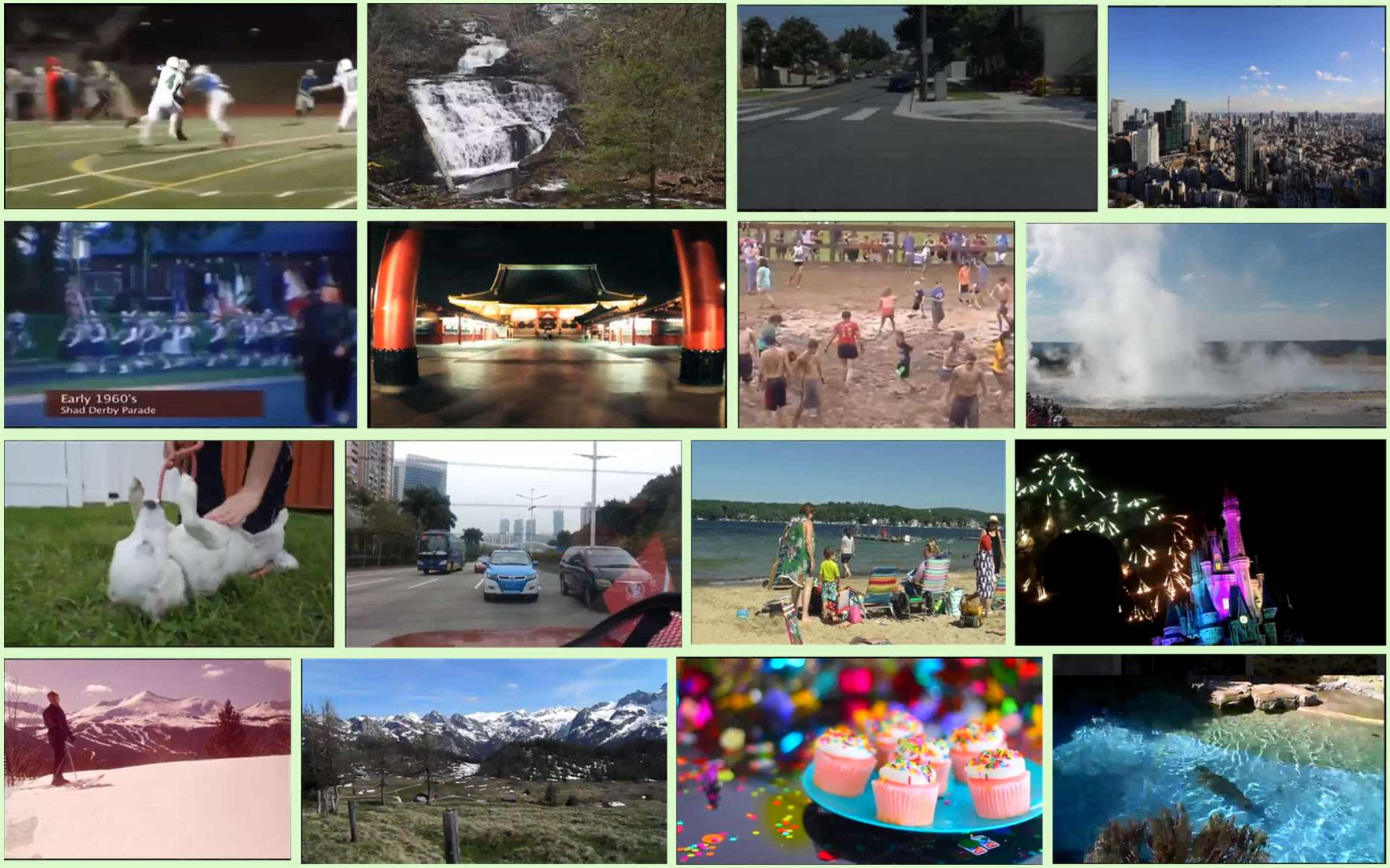
Full video

Spatial Patch
sv-patch

Temporal Patch
tv-patch

Spatio-temporal Patch
stv-patch

Exemplar Video Frames
LIVE-FB LSVQ Database



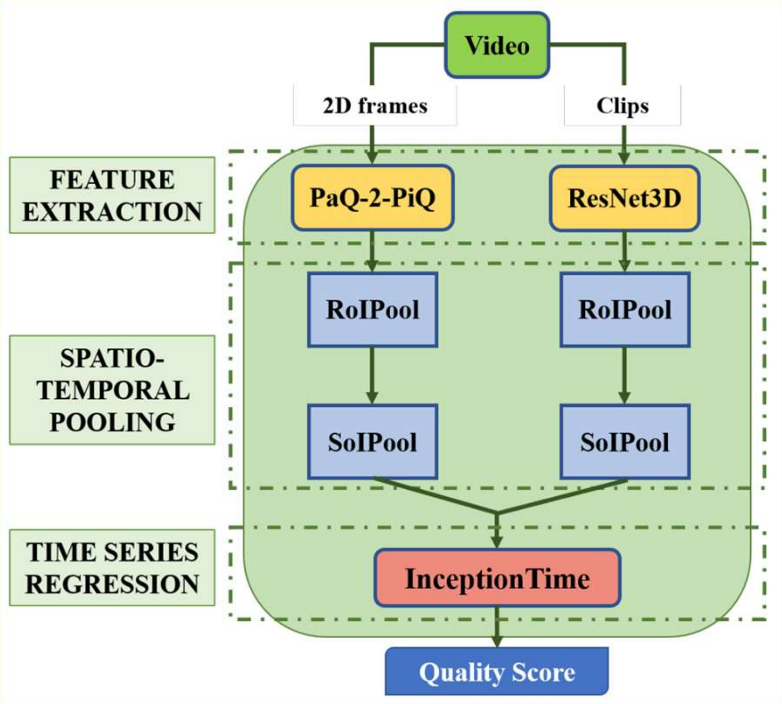
Patch-VQ or PVQ

(Patching Up Video Quality)

Ying, Mandal, Ghadiyaram, and Bovik, “Patch-VQ: 'Patching Up' the Video Quality Problem,” *Arxiv*, Nov. 2020; also *IEEE CVPR* 2021.

PatchVQ (PVQ)

PaQ-2-PiQ is a Resnet-18 image quality model fine-tuned on the LIVE-FB **Picture** Quality Database

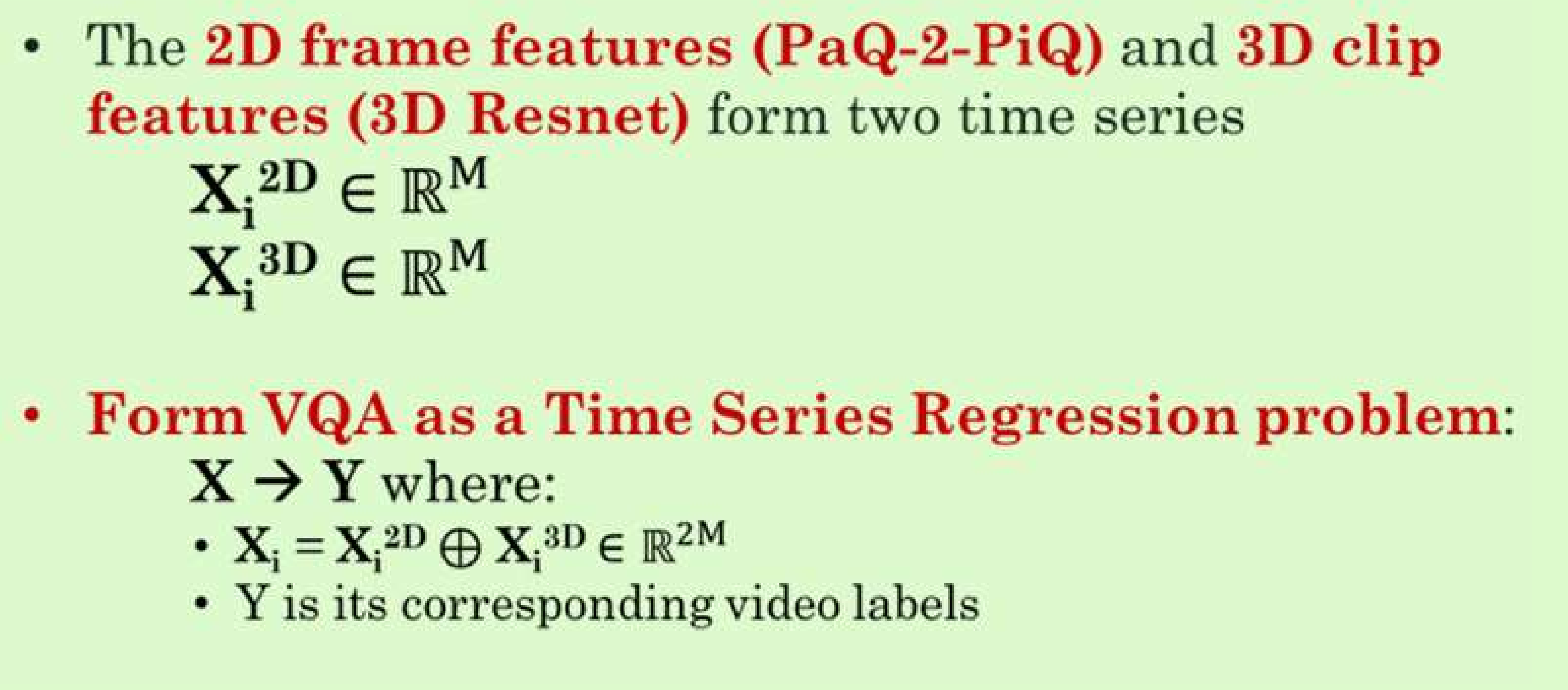


ResNet3D pretrained on **Kinetics-400** (action recognition DB)

- **Feature extractors:** "PaQ-2-PiQ" and ResNet 3D
- **4 "RoIs":** full video + 3 v-patches (16 coordinates)
- **4 "SoIs":** full video + 3 v-patches (8 coordinates)
- **InceptionTime** produces **video + patch scores**

Ying, Niu, Gupta, Mahajan, Ghadiyaram, Bovik, From patches to pictures (PaQ-2-PiQ), *IEEE CVPR* 2020.

Time Series of 2D + 3D Deep Features



- The **2D frame features (PaQ-2-PiQ)** and **3D clip features (3D Resnet)** form two time series

  $\mathbf{X_i}^{\mathbf{2D}} \in \mathbb{R}^{\mathrm{M}}$

  $\mathbf{X_i}^{\mathbf{3D}} \in \mathbb{R}^{\mathrm{M}}$

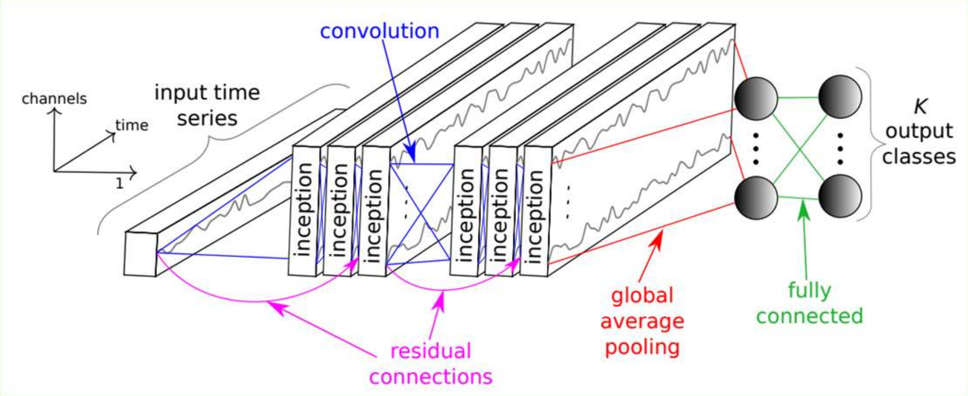
- **Form VQA as a Time Series Regression problem**:

  $\mathbf{X} \rightarrow \mathbf{Y}$ where:

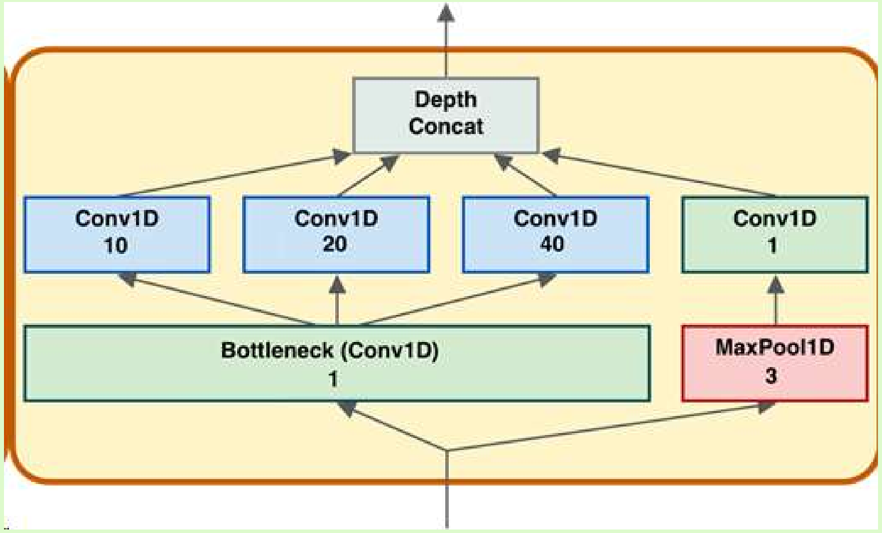
  - $\mathbf{X_i} = \mathbf{X_i}^{\mathbf{2D}} \oplus \mathbf{X_i}^{\mathbf{3D}} \in \mathbb{R}^{2\mathrm{M}}$
  - Y is its corresponding video labels

InceptionTime

- A SOTA DL model for **time series classification.**
- Major building block: **Inception module**



(K = 1: One output/video)



Inception modules used in InceptionTime. The number in each box is the **kernel size**.

1x1 convolutions reduce (channel) dim 128:32

Fawaz, *et al.*, "InceptionTime: Finding AlexNet for time series classification," *ArXiv*, Sep. 2019.

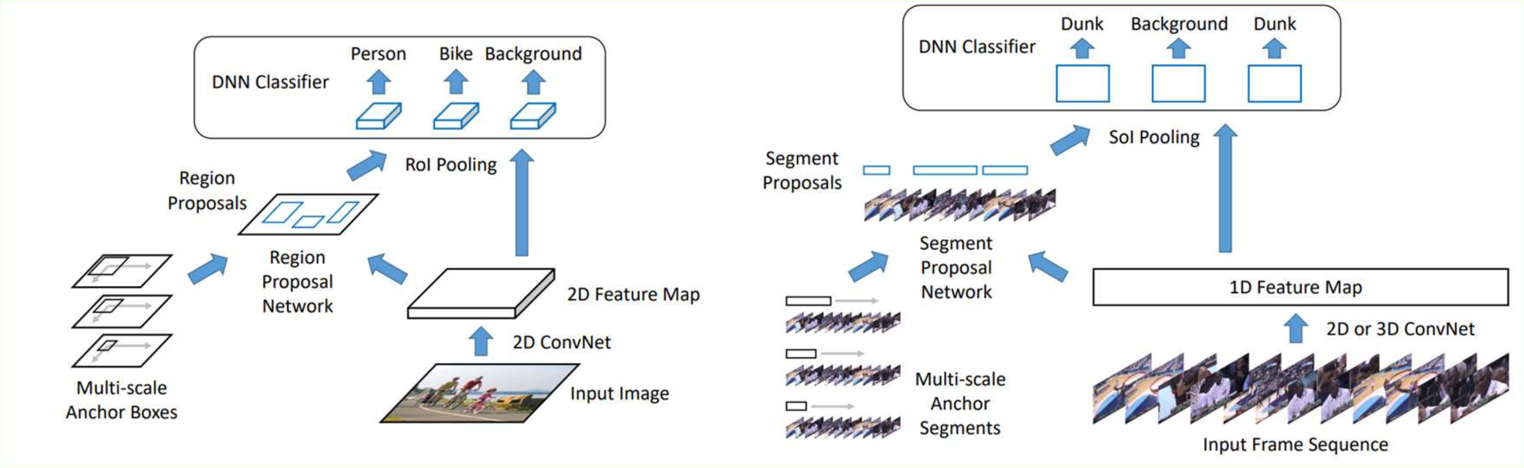
ROI-Pooling R-CNN

- **ROI pooling as** introduced in **R-CNN** (we use “Faster R-CNN”)
- **Simplified** since no need for region proposals (ROIs always specified).
- Learn on both **whole-video** and **v-patch** human labels.

Ren, He, Girshick, Sun, Faster R-CNN: “Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems,* 2015.

SoIPool

- **Inspired by TAL-Net***
  - Faster R-CNN (left) vs. TAL-Net (right)



- **Segment-of-interest pooling**
  - 1D version of **RoIPool** along **time axis**
  - Use **avg-pooling** instead of **max-pooling**

*C. Yu-Wei *et al.*, "Rethinking the faster R-CNN architecture for temporal action localization," *Computer Vision and Pattern Recognition,* 2018.

Training PVQ

- **V-patch locations/sizes** are **always known:**
  - **Training:** 4 locations: whole video, sv-patch, tv-patch, and stv-patch (from LIVE-FB LSVQ DB)
  - **PVQ Testing:** K = 4 pre-specified locations (**whole video** & any **3 v-patches**)

- **Quality prediction** of **whole videos** of **any size** and **any number** K of **v-patches.**

- **Training:** The 160K videos/v-patches were divided into
  - 72% for training
  - 19% for testing
  - 9% testing (≥1080p)

Testing PatchVQ

LIVE-FB LSVQ Database (2020)

Model	SROCC	LCC
BRISQUE	**.579**	**.576**
VIDEVAL	**.794**	**.783**
VSFA	**.801**	**.796**
PatchVQ	**.827**	**.828**

LIVE VQC Database (2018)

Model	SROCC	LCC
BRISQUE	**.524**	**.536**
VIDEVAL	**.630**	**.640**
VSFA	**.734**	**.772**
PatchVQ	**.770**	**.807**

- **BRISQUE:** Widely-used blind IQA model. NSS+SVM based.
- **VIDEVAL:** SOTA non-deep model based on fused features.
- **VSFA:** SOTA deep model. Resnet50+GRU (Gated Recurrent Units, like LSTM).

- **LIVE VQC** is a smaller (585 videos) real-world DB – **widely used** and accepted.
- **No** additional **fine-tuning.**
- Shows **generalization capability** since **trained on LIVE-FB**

PVQ Mapper: Perceptual Quality Map Predictor

Space-Time Quality Maps

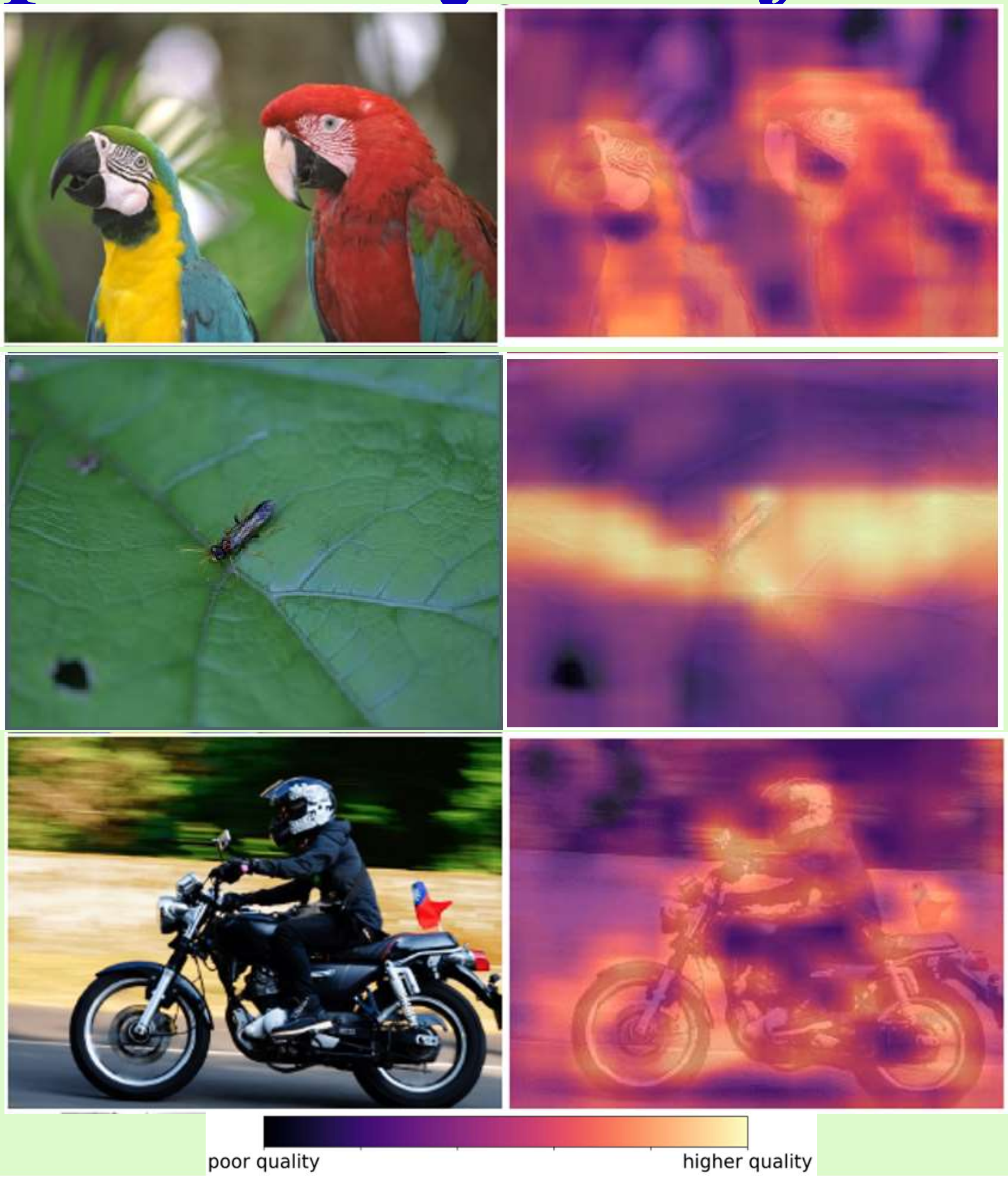
- **Application** of trained **PVQ Model to NxMxL video**

- **Spatial version: Partition frames** into 16x16 grid of 256 spatial patches, each **16 x N/16 x M/16**

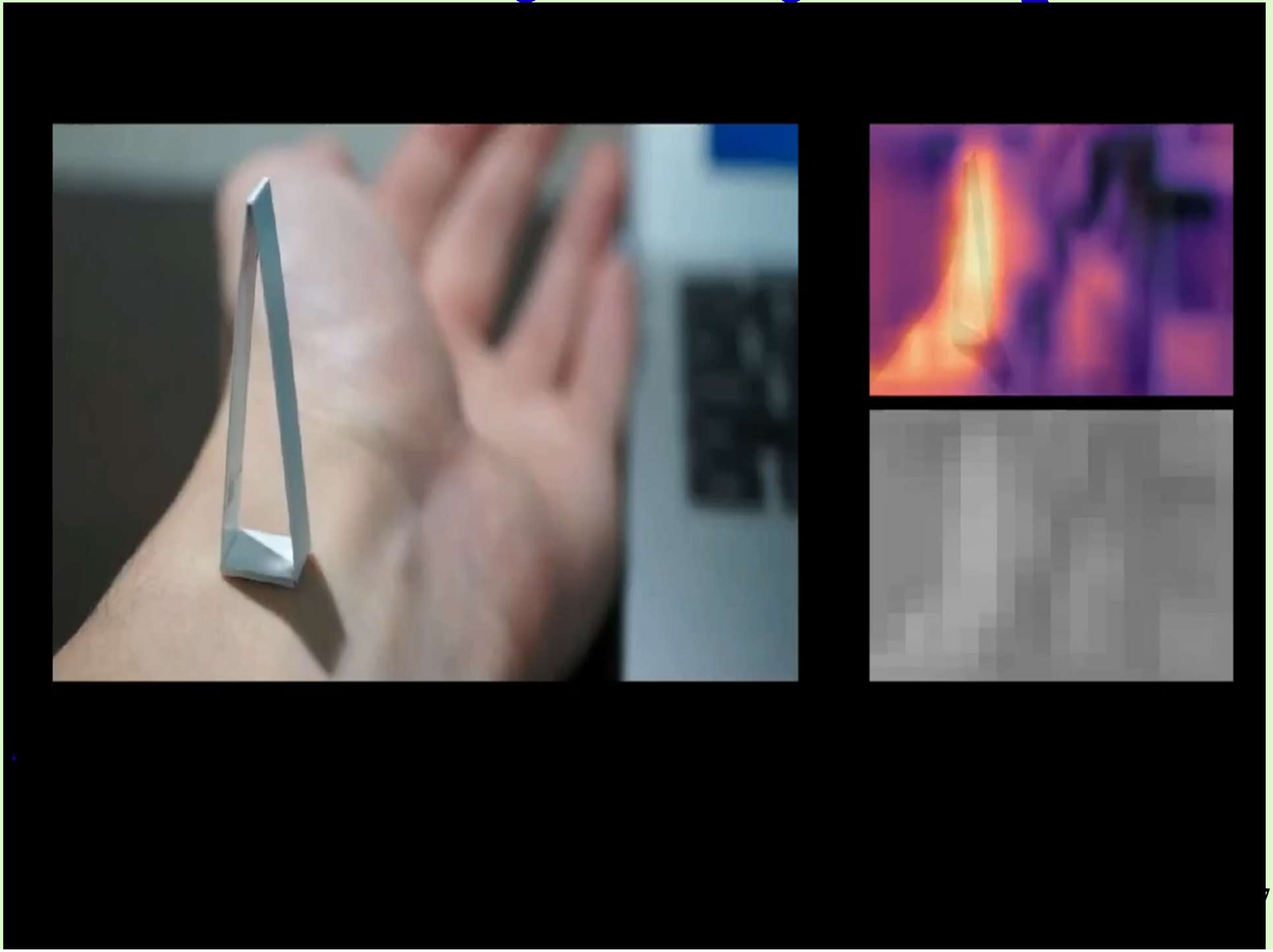
- **Space-time version:** Partition video
  - into 16-frames **clips,** calculate quality of each clip.
  - partition **frames** as above

- Produces a 16 x 16 **spatial quality map for each temporal clip**

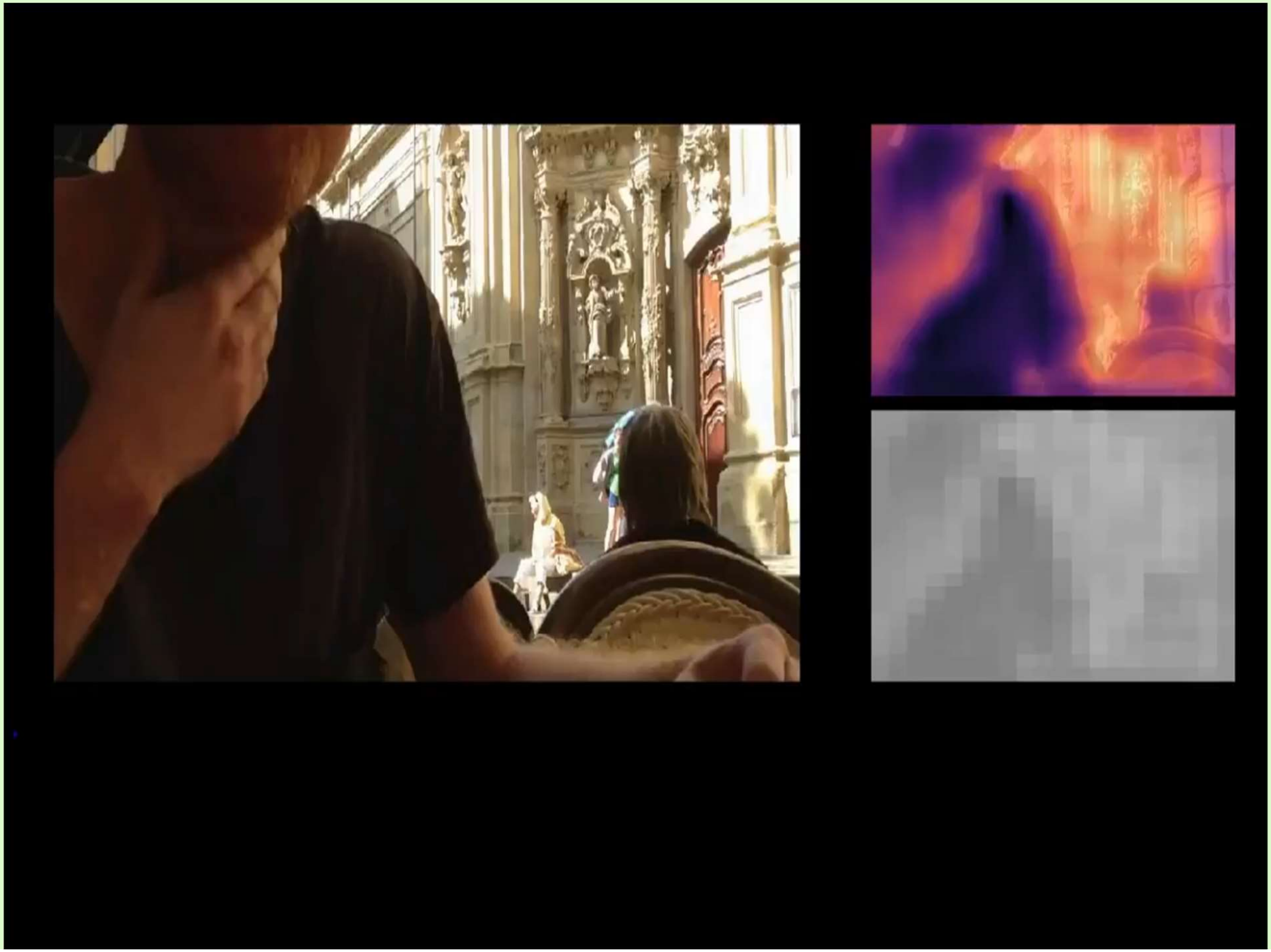
Spatial Quality Map



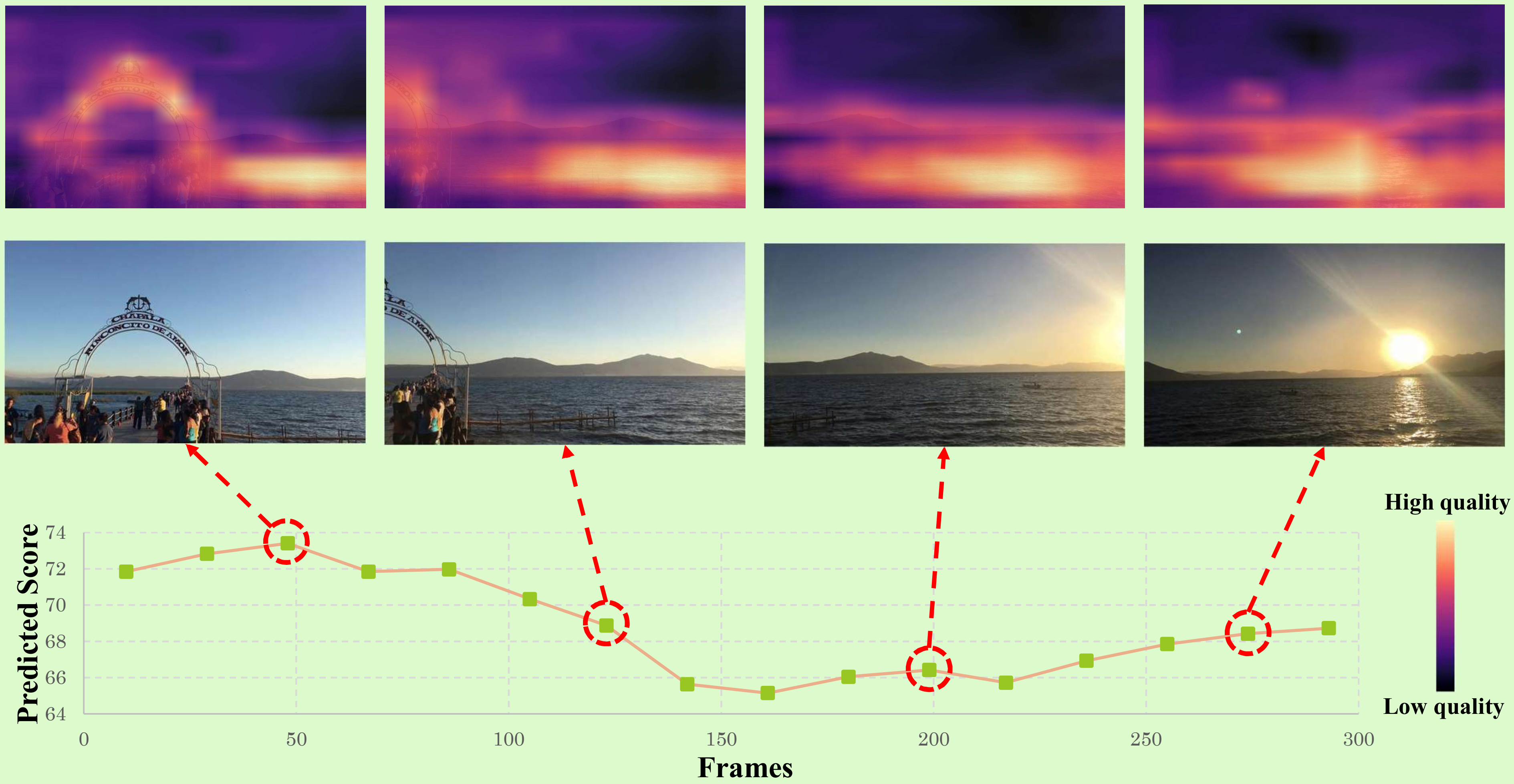
Frame Quality Map 1



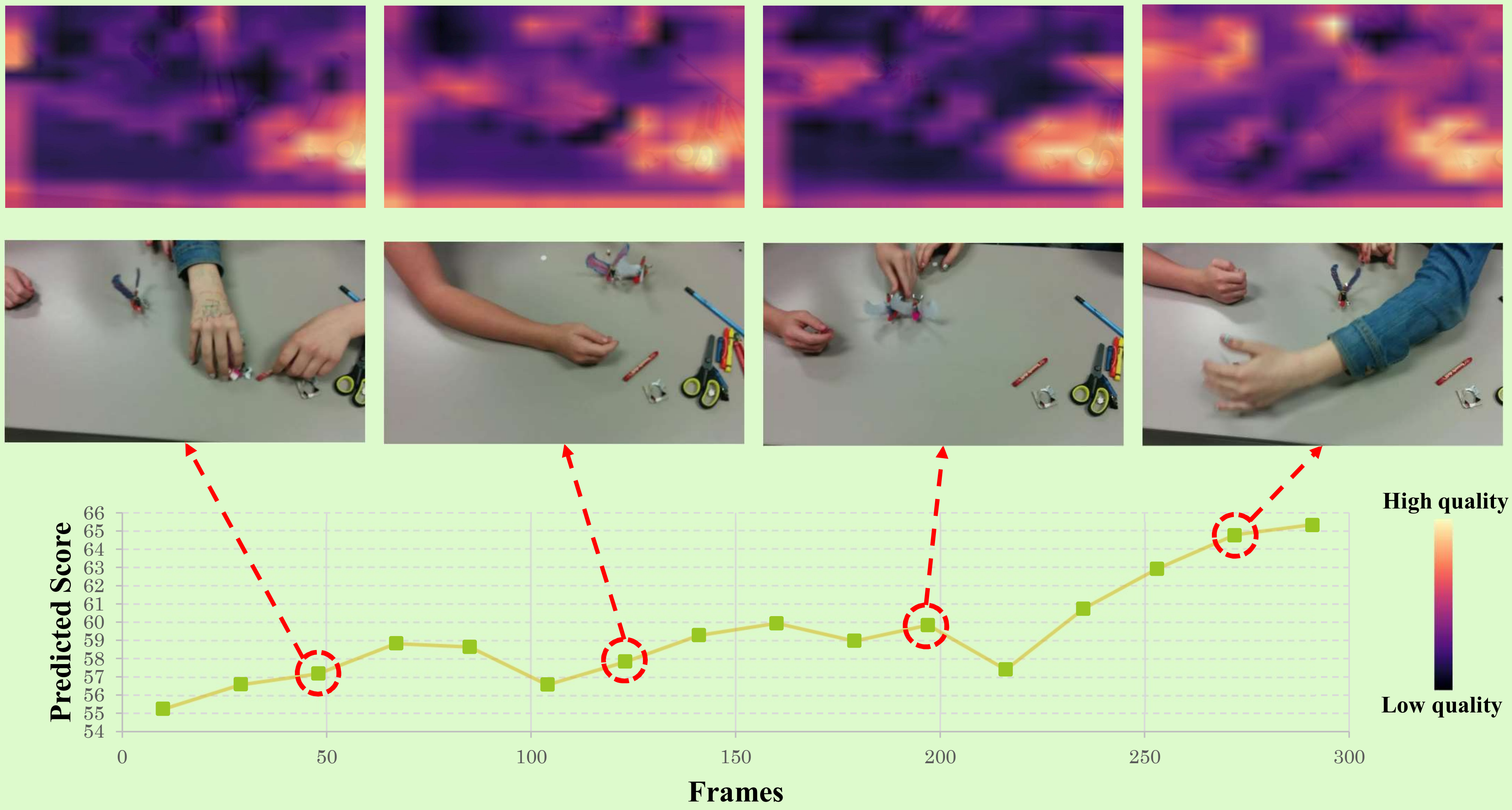
Example Quality Map 2



Space-Time Quality Map



Space-Time Quality Map



Can you identify the focus changes from the (dips in) temporal quality plot?

Test These Out Yourselves!

Online

DEMO

LIVE's Current Sponsors

NSF

NETFLIX

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

amazon prime video

f

Video Clarity
Tools for Video Analysis

oculus

ETRI

YouTube

Facebook AI Research

Questions?