Video Quality: A Nexus of Video Engineering and Visual Neuroscience

Al Bovik

Mile High Video

Denver

May 2023

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

How many distortions can you find?


© 1967 P. Patterson

Video Quality Issues are Pervasive

- **Every day:**
  - 80% of Internet traffic is pictures and videos

USER GENERATED CONTENT

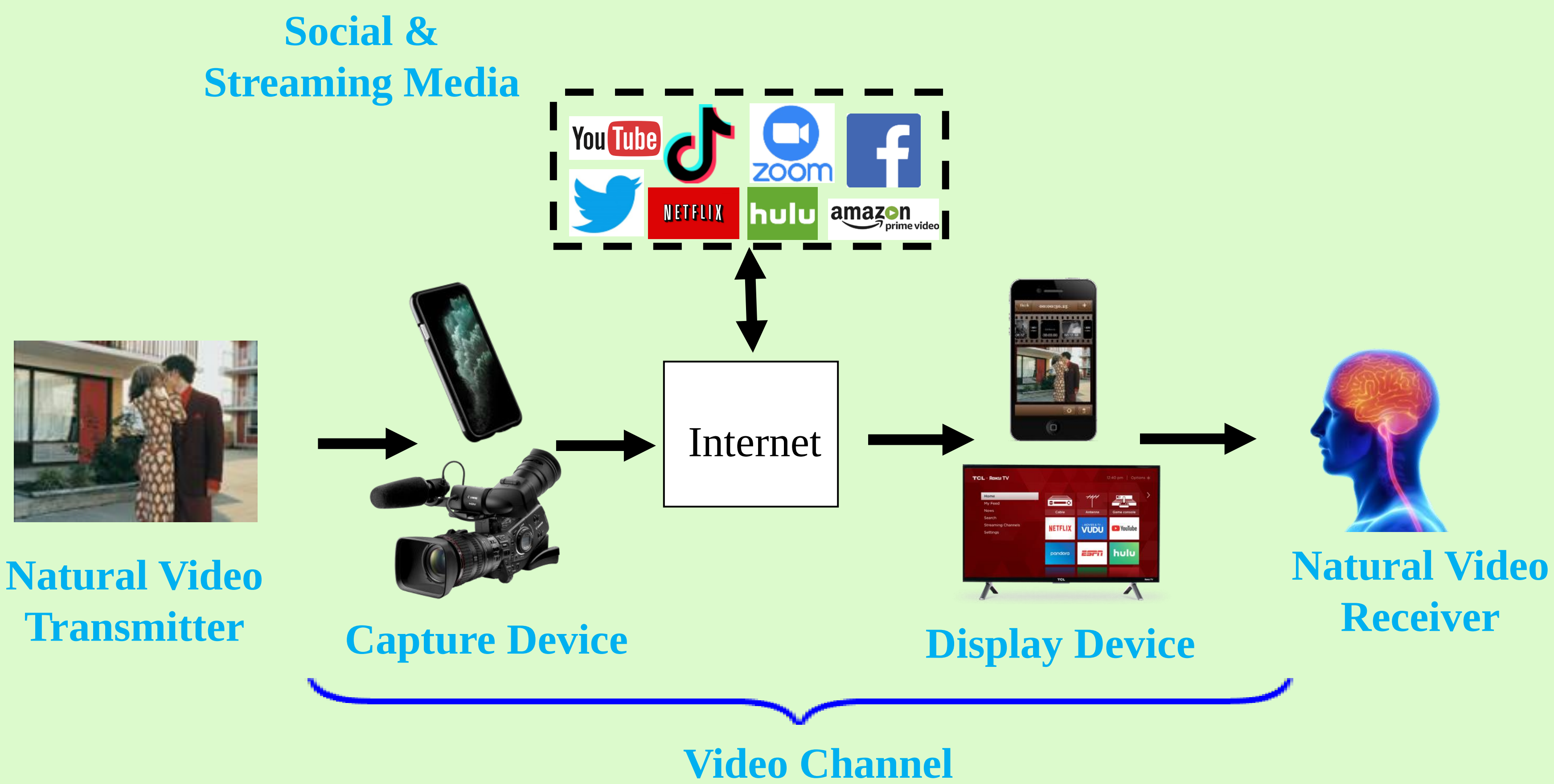
STUDIO GENERATED CONTENT

- Pictures and videos suffer from an **extreme diversity** of **distortion types** and **severities**

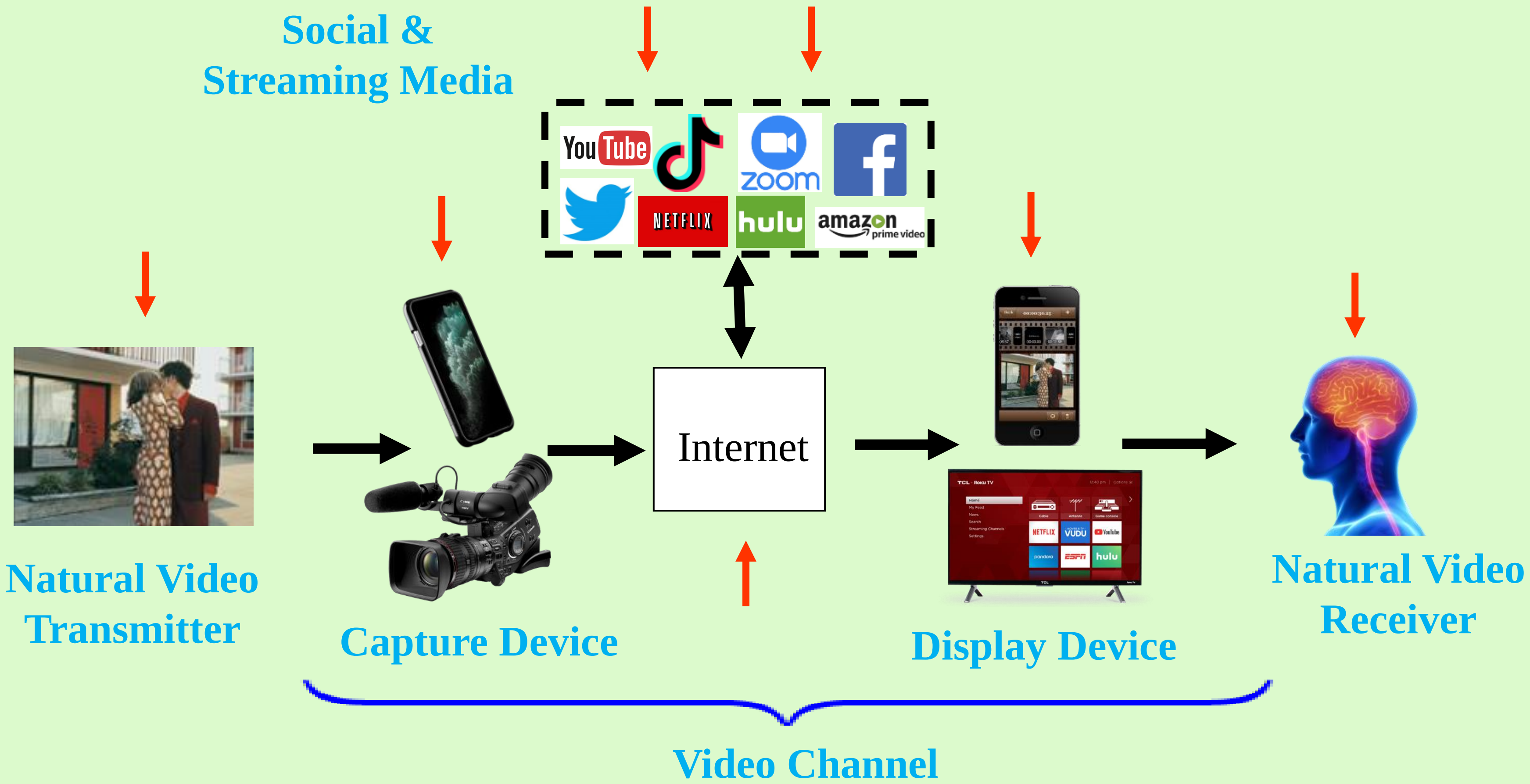
- These often occur in **complex combinations** of **degradations** creating **new visual distortions**.

- Distortions affect **user experience** and **bandwidth usage.**

Today's Video Communication System



Sources of Video Distortion



Video Quality

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.** Too complex to model, too many **distortion variations,** too many **distortion combinations,** too **hard to map to perception.**

Video Quality Prediction is Hard! Can we?

Yes, because

Videos are Special

and because distortion changes their specialness

Since our

Brains expect Specialness

we can model and predict the perception of distortion

Special Property 1: Reciprocal Law

- The **power spectra** of **videos** are pretty **reliably modeled as obeying reciprocal power laws:**

$$E\left[\left|F(U)\right|\right] \propto \frac{1}{U^{\alpha}}$$

where U = is spatial or temporal frequency.

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

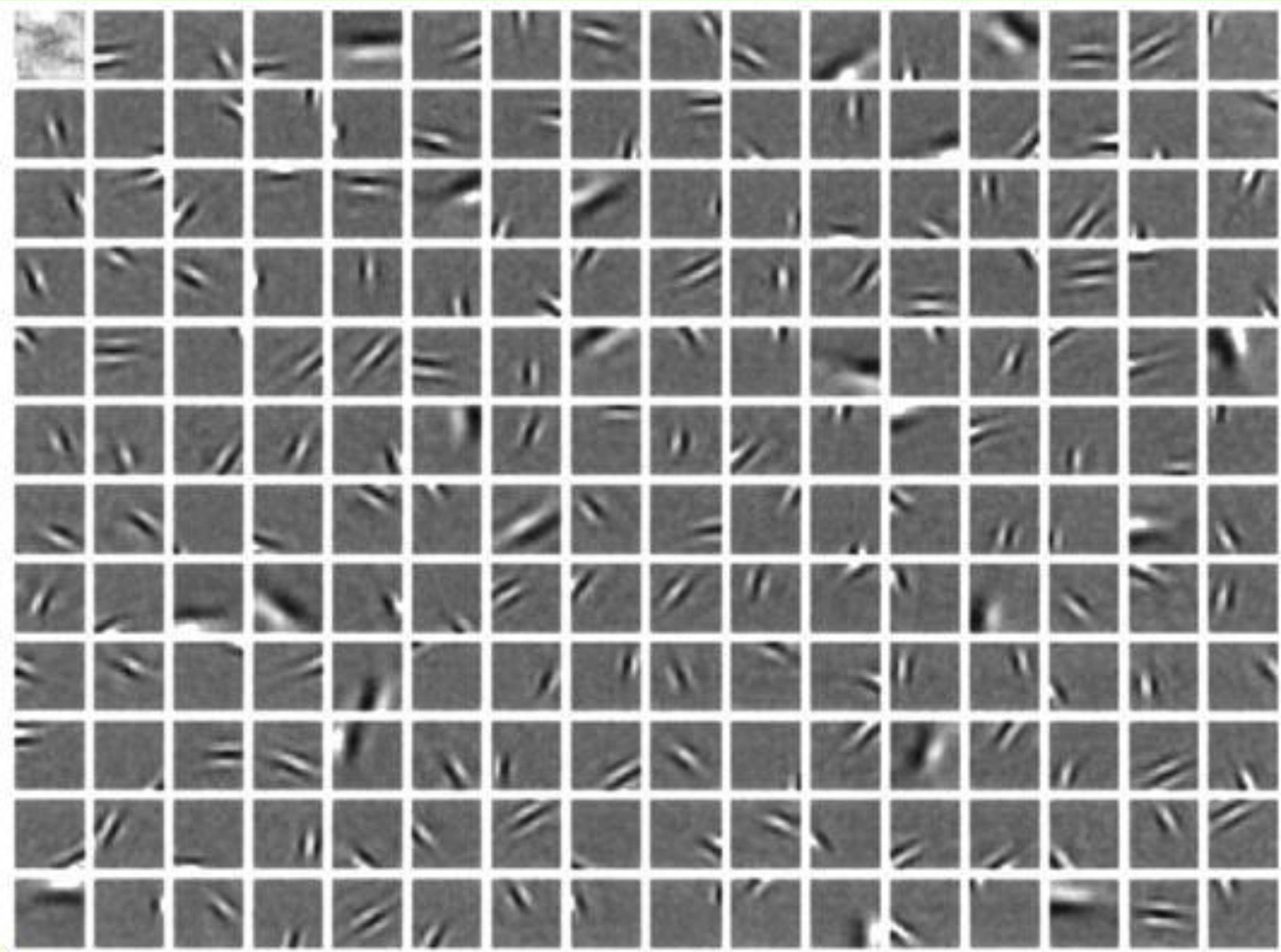
- **Functions** satisfying these are **uniquely self-similar:**

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

Example: Alpine Sled

Videos are **self-similar** and **multiscale.** So is **perception** of **them**.

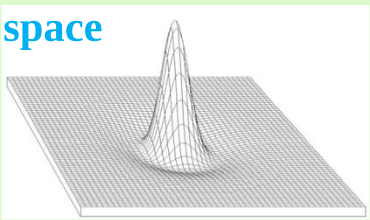
Tolhurst, et al "Amplitude spectra of natural images," *Ophthal. & Physiol Optics*, 1992.

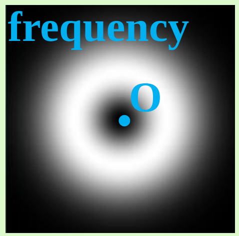


C
ir

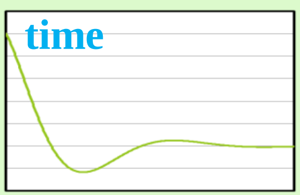
Bandpass Retino-Cortical Filters

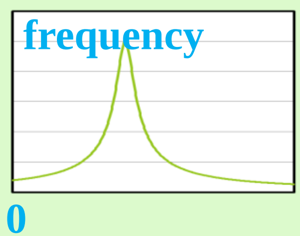
- Sparse coding of pictures and videos resemble **bandpass receptive field profiles** of **neurons** along **retino-cortical pathway.**



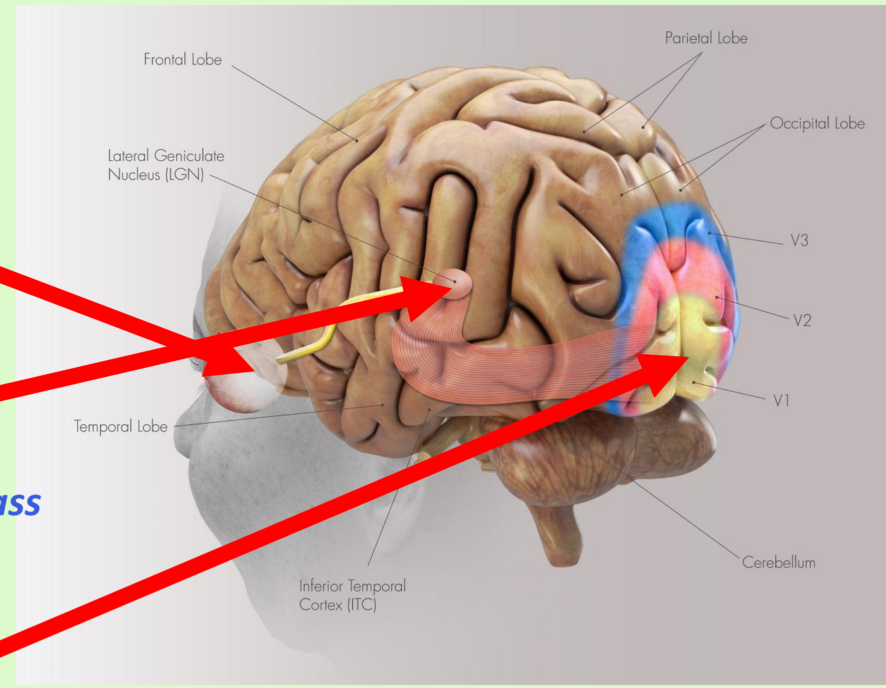


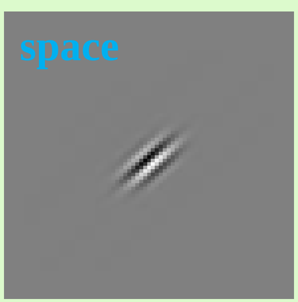
Spatial bandpass predictive coding by retinal ganglion cells …

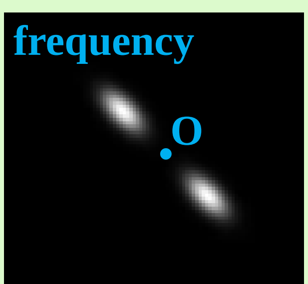


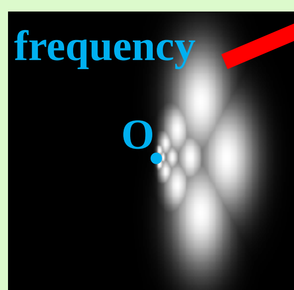


… temporal bandpass coding in LGN …









Bandpass decompositions in visual cortex …

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

Special Property 3: Gaussian Law

- **Bandpass videos are reliably modeled** as obeying **gaussian scale mixture (GSM) models.** If (f = video)

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

  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

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

- Bandpass processing also **decorrelates** (it's differencing!)

- **Dividing** by local space/time/scale energies (**estimates of z**) further **decorrelates** & **gaussianizes.**

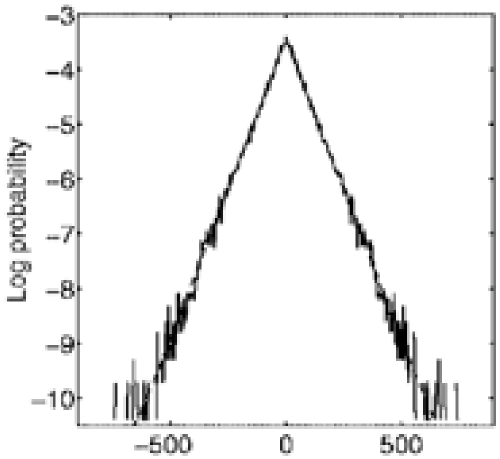
G re

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

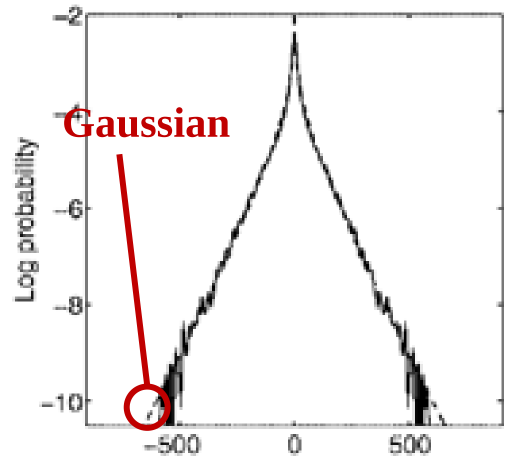
  p

- **ML est** oeff. γ:

mandrill

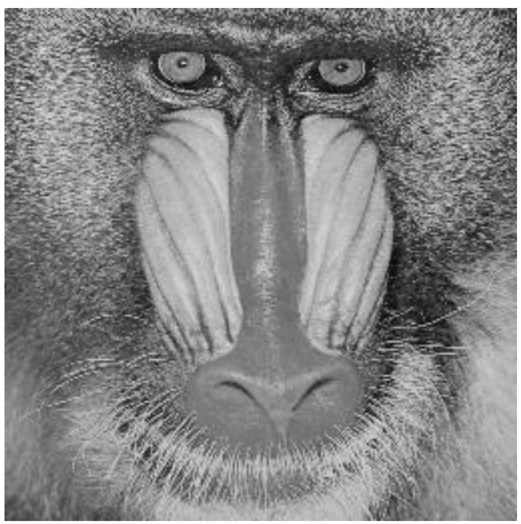


boats



Bandpass, divisively normalized pictures

- **Dividi on**) yields approxi



Original images

The rema ty
visual sig s
this regul

Ruderman, The sta 1994.

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

Dual Nature of Sensory Neurons

Dual (evolutionary) **bandpass & normalized processing** in sensory neurons explains **perceptual contrast masking** (incl. of distortions).

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

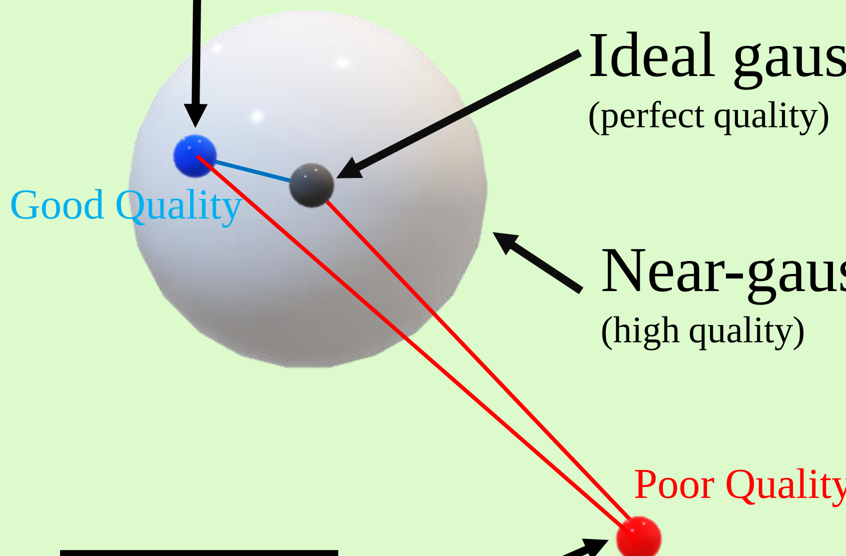
(Very) General Quality Measurement Concept

Perceptual Processing Model

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

Distort

Perceptual Processing Model

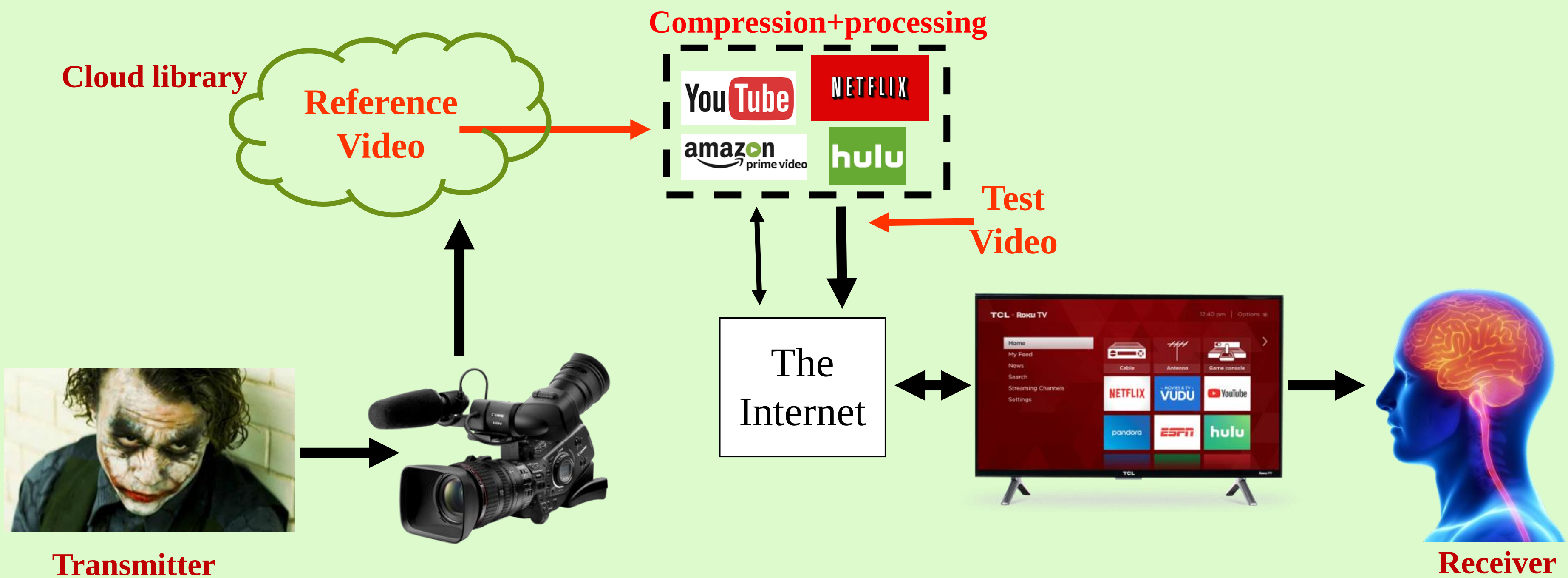


Ideal gaussian
(perfect quality)

Near-gaussian
(high quality)

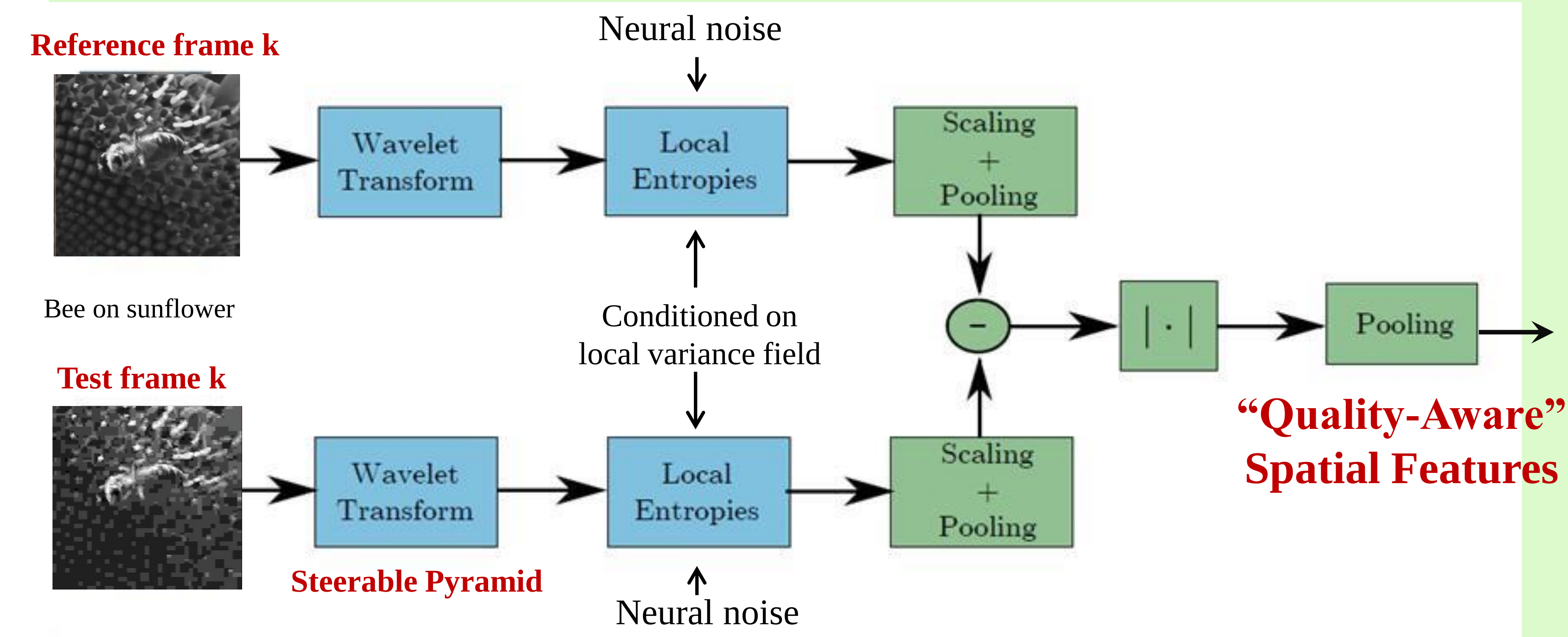
How to define perceptual quality distances?

"Reference" Video Quality Prediction



- Need accurate **models of transmitter.**
- Need accurate **models of the receiver.**
- These models **are dual/combined.**

Measure Spatial Information Loss



"Quality-Aware" Spatial Features

Bandpass, multiscale cortical model

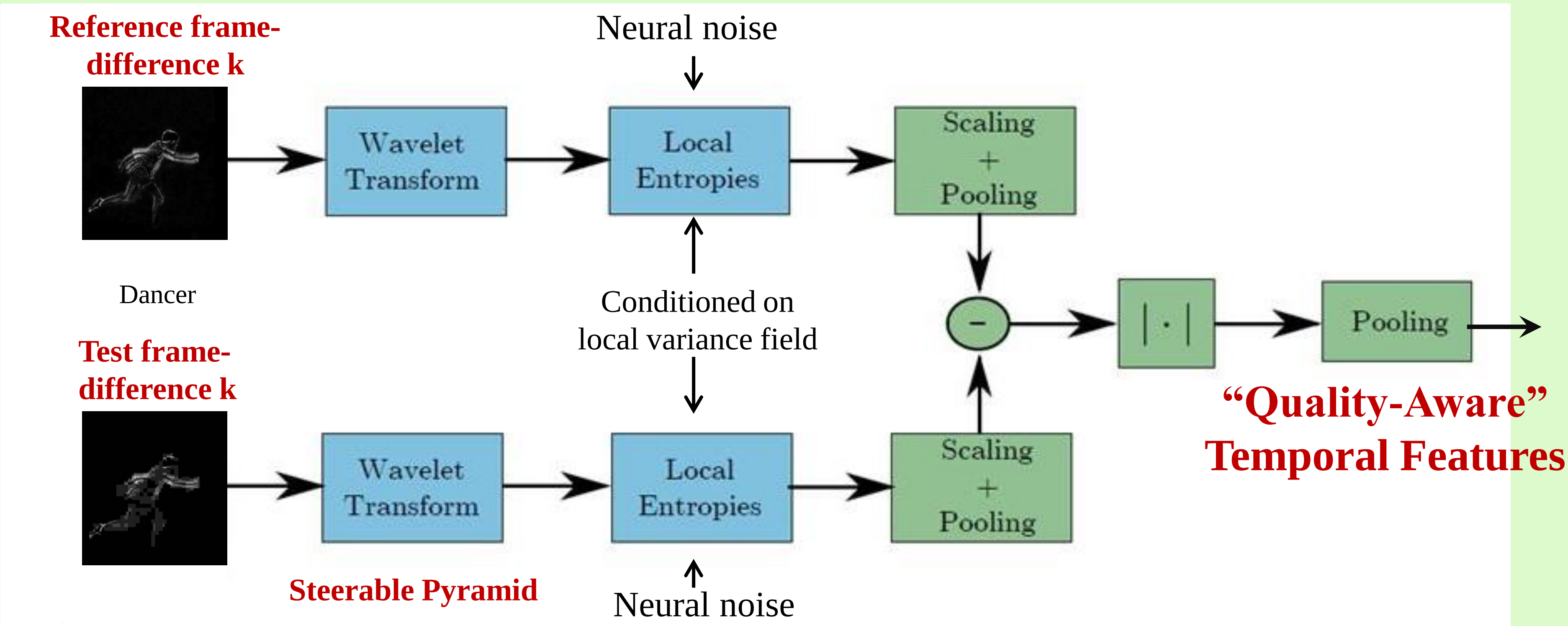
GSM model

Scale by local variance field
Pool & Difference

H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, 2006.

R. Soundararajan and A. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, *IEEE Transactions on Circuits and Systems*, 2013.

Measuring Temporal Information Loss



Bandpass, multiscale LGN-cortical model

GSM model

Scale by local variance field
Pool & Difference

Frame differences are bandpass and also strongly obey the GSM law, violated by local **temporal distortion**.

GSM Model of Frames and Frame Differences

- Bandpass **video frames** and **frame differences** are **modeled** as **GSM,** which is a **regular, reliable model** of both.

- All are modeled as **noisy GSM vectors:**

$$g = z\gamma + w$$

$$z = \text{variance field} \qquad \gamma \sim \eta(0, \mathbf{K}) \qquad w \sim \eta(0, \sigma_w^2 \mathbf{I})$$

where **w** models **visual uncertainty** - neural noise and other perceptual imperfections.

- This **same model** is **independently applied** on each element of

{frames, frame differences} x {original, distorted}

Conditional Entropies

- Find the **ML estimate(s)**

$$\hat{z} = \arg\max_{z} \left\{ \log\left[p\left(g|z\right)\right]\right\} = \sqrt{\frac{g^T \tilde{\mathbf{K}}^{-1} g^T}{N}}$$

Compute on:
- **Original frames**
- **Distorted frames**

- **Frame Differences**
- **Distorted Frame Differences**

yielding **conditional entropies**

$$H\left(g|z=\hat{z}\right) = \frac{1}{2}\log\left[\left(2\pi e\right)^N \left|\hat{z}^2\tilde{\mathbf{K}} + \sigma_W^2 \mathbf{I}\right|\right]$$

which are **perceptually scaled** by log variance:

$$\alpha = \log\left(1+\hat{z}_{\text{space}}^2\right)\cdot\log\left(1+\hat{z}_{\text{time}}^2\right)\cdot H\left(g|z=\hat{z}\right)$$

to numerically **stabilize** and highlight **higher local energy** of either **content** or **distortion.**

Popular Algorithms Derived from These Models

- Visual Information Fidelity (VIF, 2006): no variance scaling, additive pooling.

- Space-Time Reduced Reference Entropic Differencing (ST-RRED, 2012): temporal aspect, variance scaling, additive pooling.

- NETFLIX's Visual Multi-Method Fusion (VMAF, 2016): VIF/ST-RRED features + “detail” feature pooled using SVM.

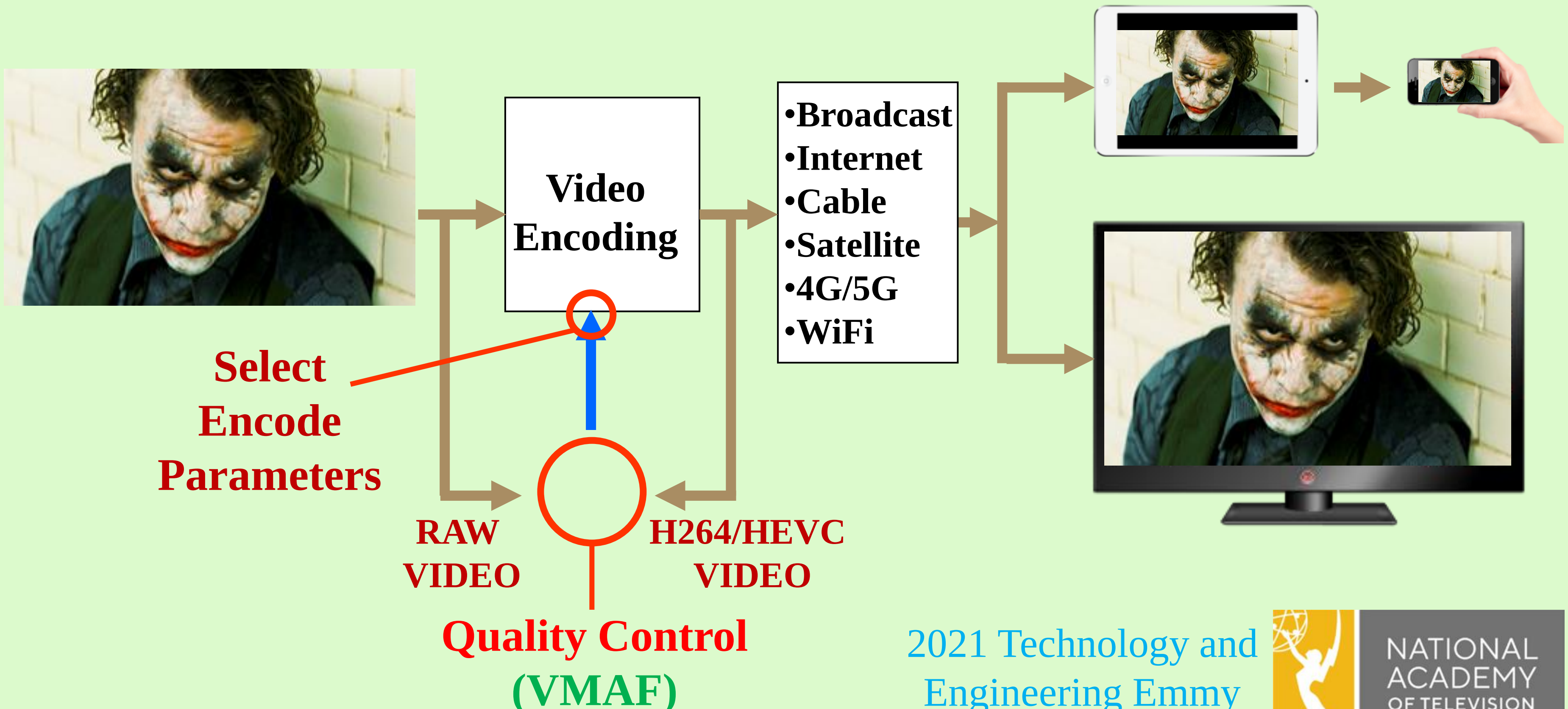
- Algorithms from these models are used at the global scale by most broadcast, streaming, and social media providers, affecting billions of viewers, significantly reducing bandwidth consumption and the carbon footprint of the Internet.

ST-RRED Map 1

ST-RRED Map 2

Use Case: Encoder Control of Broadcast, Streaming, and Sharing

- **Compressed videos** streams are 80% of **US Internet bits**
- Most **encodes** are **quality controlled** by **VIF/VMAF**

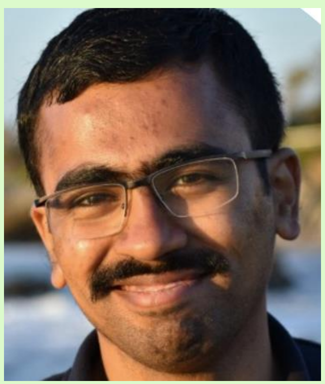


2021 Technology and Engineering Emmy Award

NATIONAL ACADEMY OF TELEVISION ARTS & SCIENCES

New Use Case 1:

High/Variable Frame Rates (VFR)



Pavan Madhusudana

P.C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A.C. Bovik, “Making video quality models sensitive to frame rate distortions,” *IEEE Signal Processing Letters*, vol. 29, pp. 897-901, 2022.

Video Frame Rate

- **Older frame rates** of **30 frames/sec** (fps) and **24 fps (cinema)** are now largely **superseded by 60 fps.**

- However, **even 60 fps is inadequate** in the presence of **high object** and/or **camera motions.**

- This is becoming of **pressing importance** since live **sports content** is being delivered by YouTube, Amazon Prime Video, and others.

Quality vs Frame Rate

- How does **frame rate** affect **perceived quality?**

- Given a bandwidth target, can we **optimize** the **compression / framerate vs. perceptual quality tradeoff?**

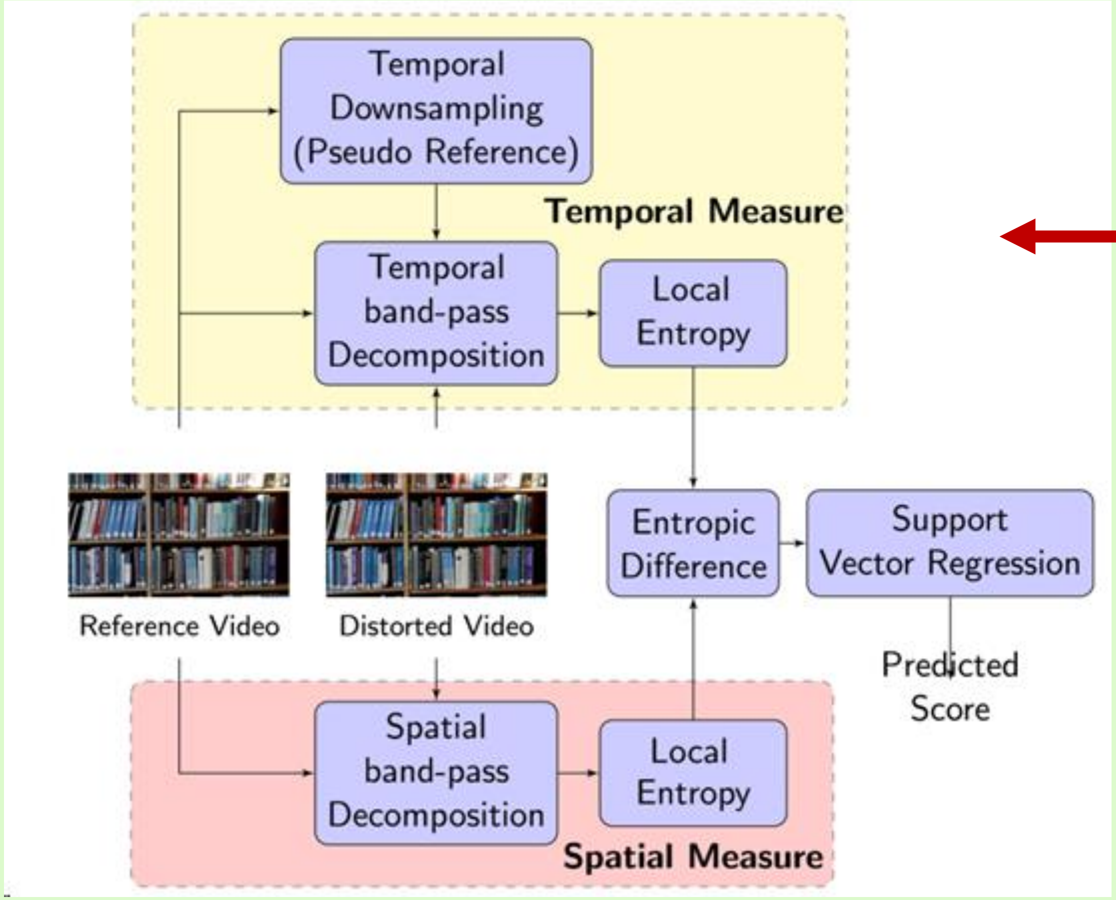
High Motion Example

24 fps

60 fps

The Same Statistical Models Apply in Time

- **Temporal bandpass** videos obey the same **statistical laws** with **high regularity.**



Temporal BP Filters

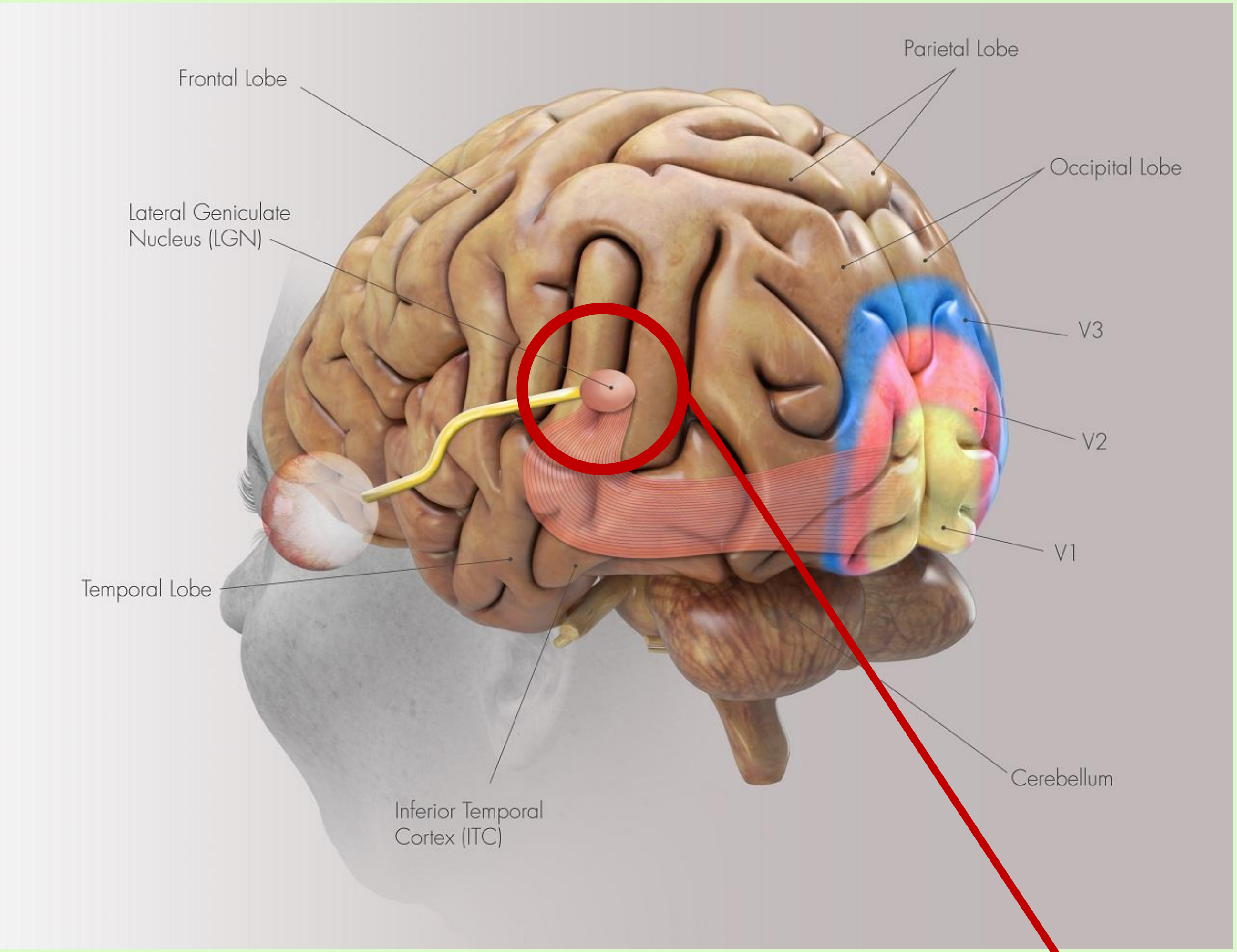
- Given a video $V(\mathbf{x}, t)$ compute K **temporal bandpass responses**

$$B_k(\mathbf{x}, t) = V(\mathbf{x}, t) * b_k(t)$$

$K = 1, \ldots, k.$

- We use **Daubechies biorthogonal-2.2** wavelet filters.

LGN

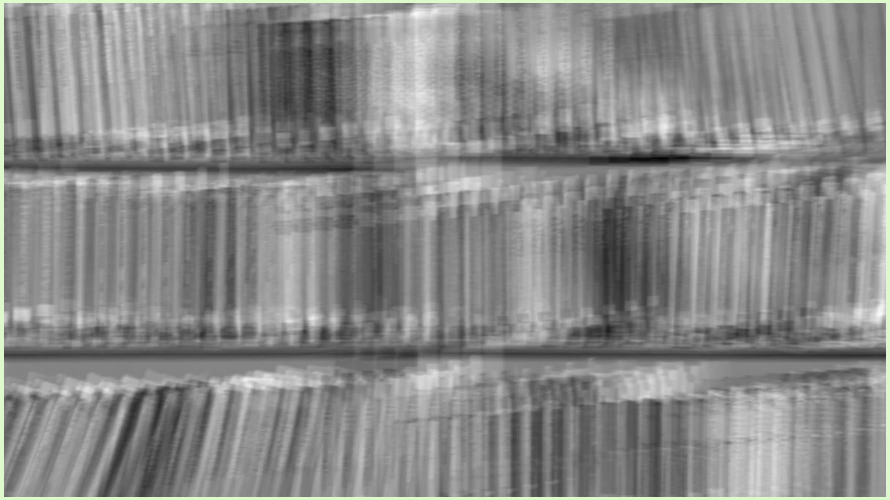


Simple linear model of temporal visual processing in thalamus area LGN.

Effects of Frame Rate on Temporal BP Responses

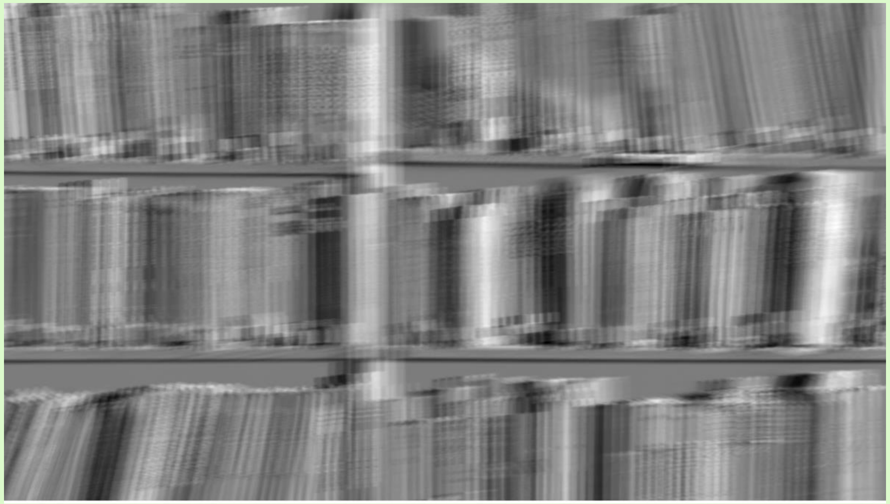
$$B_k(\mathbf{x}, t) = V(\mathbf{x}, t) * b_k(t)$$

No compression applied



24 fps

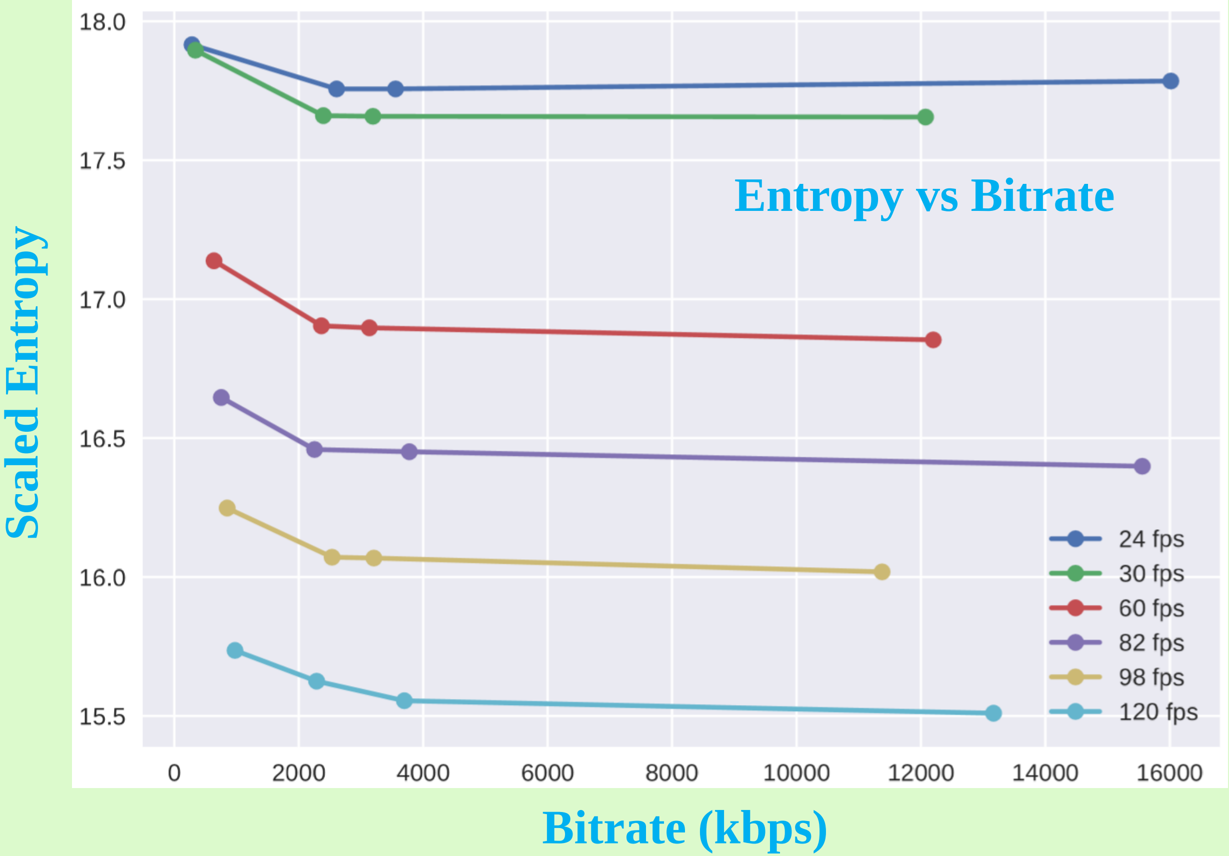
The **BP statistics** are also **greatly affected . . .**



60 fps

Frame Rate Biases Bandpass Entropy

- Frame rate **biases temporal bandpass entropy.**
- The bias is **part of measuring temporal distortion** must be **removed** when measuring **spatial distortion.**



Bias Removal

- **Three videos** are needed to form **temporal quality features**:
  - The 120 fps **reference video**
  - the **distorted video** (compressed & changed frame rate)
  - a **pseudo-reference** (**PR**) **video** for **entropy bias removal.**

- The **PR** video is the reference **down-sampled** to the **distorted video frame rate. NO** spatial (compression) **distortion.**

Generalized Space-Time (GST) Video Quality Features

- Defined in terms of the **scaled entropies** of reference (**R**), distorted (**D**), and **PR** videos:

$$\varepsilon \; = \; \log\left[1+\hat{z}_k^2\left(B_k\right)\right]\cdot h\left(B_k\right)$$

- For each **temporal BP filter** (indexed $k = 1, \ldots, K$), the $\mathbf{GST}_{kt}$ at frame t is

$$GST_{kt} \; = \; \left(1+\left|\varepsilon_{kt}^{D}-\varepsilon_{kt}^{PR}\right|\right)\cdot\left(\frac{\varepsilon_{kt}^{R}+1}{\varepsilon_{kt}^{PR}+1}-1\right)$$

- **Absolute difference:** Measures compression distortion as if R and D have the ***same*** frame rate.
- **Ratio term:** Measures frame rate distortion as if there were no compression.

How to Use GST Features

- If **R and D** have the **same frame rate,** then

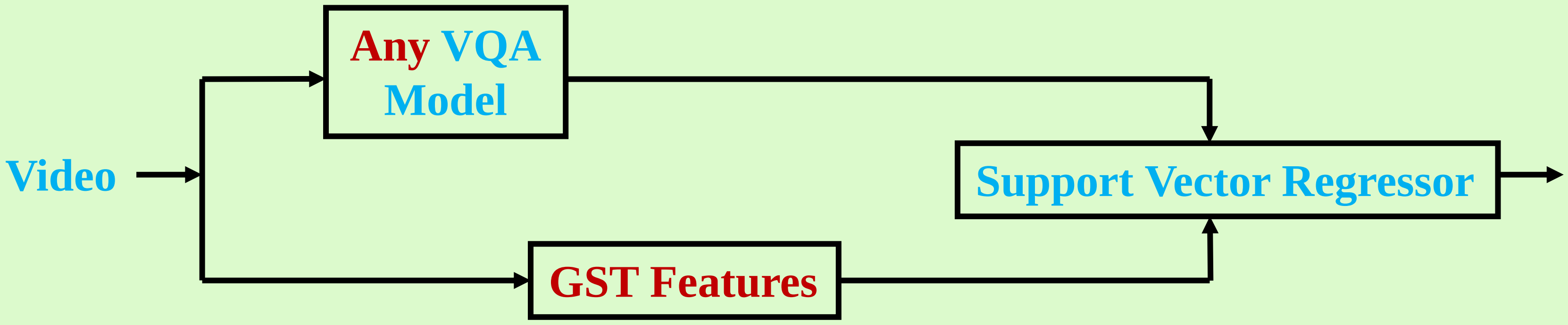
$$\mathrm{GST}_{kt} = \left| \varepsilon^{D}_{kt} - \varepsilon^{PR}_{kt} \right|$$

- If **D** is **not compressed/distorted,** then

$$\mathrm{GST}_{kt} = \left| \frac{\varepsilon^{R}_{kt} + 1}{\varepsilon^{PR}_{kt} + 1} - 1 \right|$$

- **GST = 0** only when **D = PR = R.**

How to Use GST Features: Dual Channel Solution



- **VQA features** can be drawn from **ANY** leading VQA model: SSIM, VMAF, NIQE, even PSNR
- **Neurostatistical GST** features are highly predictive of temporal distortions.

Performance Enhancements

- We **enhanced top models** with **GTS features.**
- Tested on largest "Variable Frame Rate / Compression" **subjective quality** database.

Median Correlations Against Human Quality Judgments Over 200 Train-Test Splits of Leading Models on the UT-LIVE/YouTube HFR Database

	SROCC ↑	PLCC ↑
SSIM	0.5566	0.5418
GST SSIM	0.7576	0.7700
MS-SSIM	0.5742	0.5512
GST MS-SSIM	0.8128	0.8179
ST-RRED	0.6394	0.6073
GST ST-RRED	0.8029	0.8144
SpEED	0.6051	0.5206
GST SpEED	0.8276	0.8346
VMAF	0.7782	0.7419
GST VMAF	**0.8658**	**0.8723**

- **Enormous bandwidth savings on high-motion (sports, action) content**
- **By perceptually optimizing (pushing) compression/framerate.**

SROCC = Spearman's Rank-Order Correlation Coefficient
PLCC = Pearson's Linear Correlation Coefficient

By Bitrate

- **Performances** at **bitrates** 24, 30, 60, 92, 98, 120 fps?
- **GST** especially effective at **low bitrates**
- Also effective at **high bitrates**

Median Correlation Over 200 Train-Test Splits of Leading Models, By Bitrate, on UT-LIVE/YouTube HFR Database

	24 fps		30 fps		60 fps		82 fps		98 fps		120 fps	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
SSIM	0.266	0.222	0.283	0.189	0.382	0.302	0.371	0.362	0.537	0.497	0.867	0.833
GST SSIM	0.386	0.635	0.461	0.715	0.516	0.722	0.698	0.835	0.743	0.833	0.797	0.817
MS-SSIM	0.305	0.260	0.296	0.238	0.416	0.338	0.439	0.393	0.578	0.561	0.706	0.696
GST MS-SSIM	0.505	0.682	0.495	0.769	0.579	0.796	0.704	0.844	0.775	0.841	0.832	0.838
ST-RRED	0.305	0.275	0.296	0.206	0.612	0.613	0.584	0.513	0.650	0.604	0.755	0.696
GST ST-RRED	0.518	0.698	0.421	0.664	0.580	0.814	0.684	0.833	0.752	0.816	0.888	0.885
SpEED	0.432	0.273	0.410	0.233	0.439	0.292	0.546	0.390	0.578	0.471	0.758	0.739
GST SpEED	0.645	0.744	0.616	0.759	0.577	0.745	0.723	0.789	0.787	0.827	0.860	0.867
VMAF	0.250	0.368	0.362	0.471	0.630	0.680	0.734	0.793	**0.860**	0.868	0.818	0.816
GST VMAF	**0.748**	**0.805**	**0.743**	**0.833**	**0.773**	**0.836**	**0.786**	**0.880**	**0.860**	**0.899**	**0.881**	**0.903**

GST allows streamers to further adaptively **adjust frame rate** vs **compression** to improve bandwidth budgets and environmental impact.

New Use Case 2

High Dynamic Range (HDR)

Josh Ebenezer

Zaixi Shang

J.P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A.C. Bovik, “Making video quality assessment models robust to bit depth,” *IEEE Signal Processing Letters*, to appear.

High Dynamic Range
(or bit depth)

- Older **Standard Dynamic Range (SDR)** videos represent luminances and colors with 8 bits each (24 bits total).

- Fine on **old dim televisions:** SDR is limited to **100 nits*** while modern **HDR** is mastered at **1000-4000 nits.**

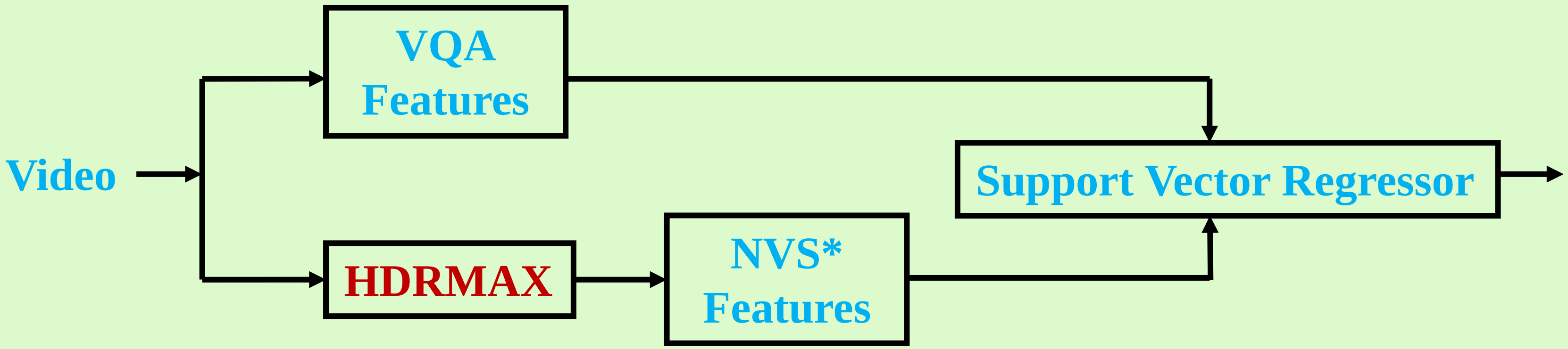
- **Modern standards** like HDR10, Dolbyvision, HDR10+ now pervasive, enabling content with
  - Darker blacks and brighter whites
  - Wider ranges of colors

- **HDR uses 25% more bits!** Increased data volume, **more compression** needed! Which impacts **perceived quality.**

- **For a given bitrate, more distortions.**

*candela / meter^2 (cd/m^2), candela measures luminous intensity

Dual Channel Solution



- **VQA features** can be drawn from **ANY** leading VQA model: SSIM, VMAF, NIQE, even PSNR
- **Neurostatistical distortion models** of **HDRMAX responses** are also regular and sensitive to distortion.
- **What is HDRMAX?**

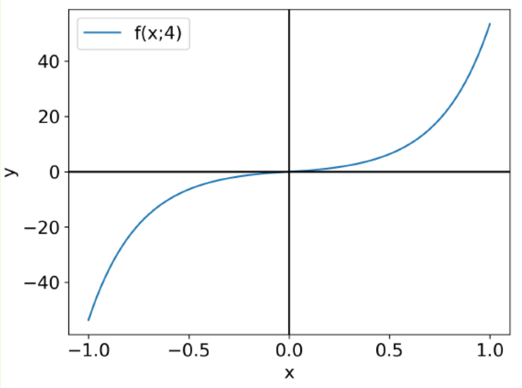
*NVS = Natural video statistics

What is HDRMAX?

- **Simple:** Linearly map the video values (luminances and/or chrominances) to [-1, 1].

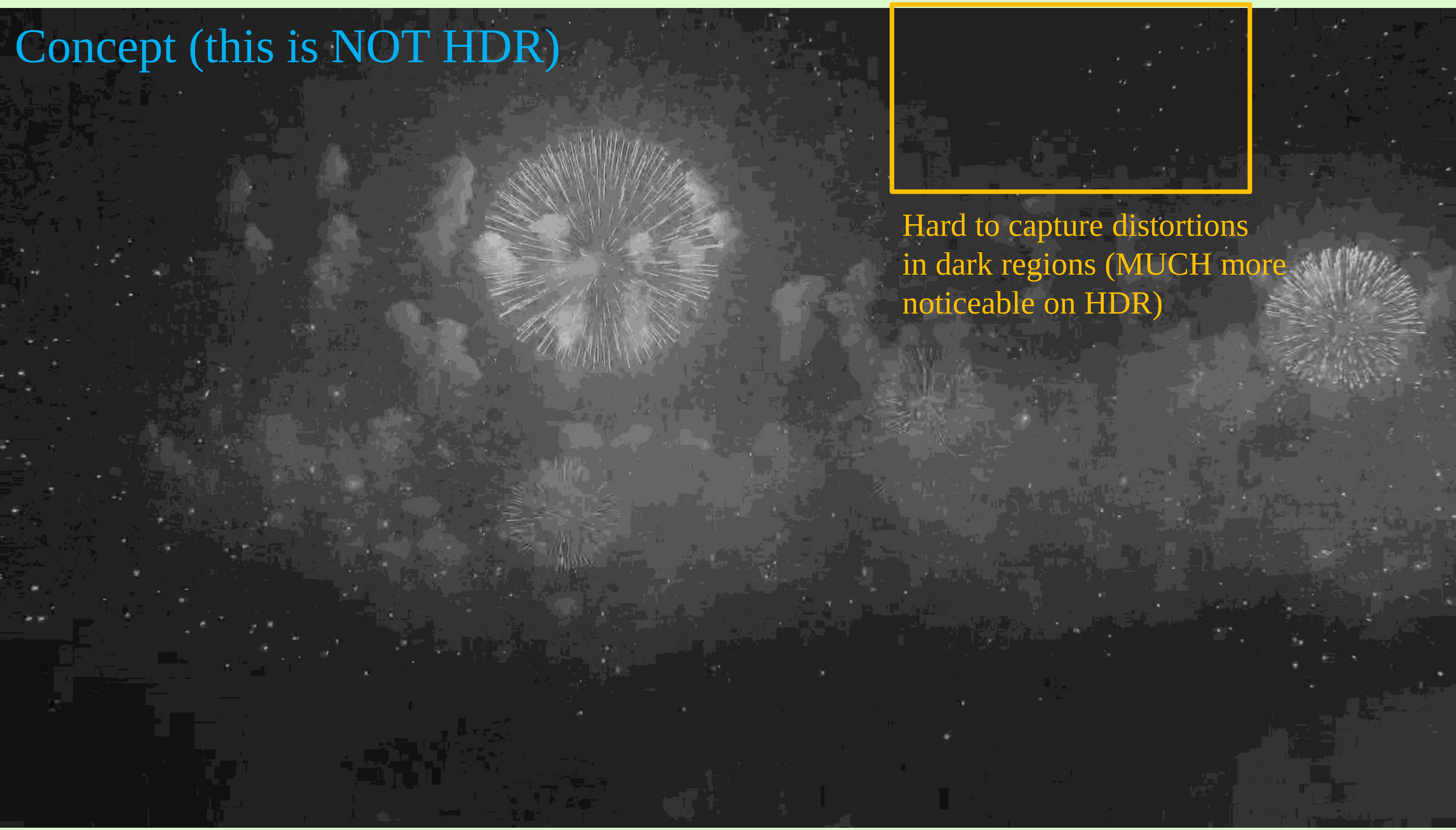
- Then apply the heuristic **expansive nonlinearity**

$$f(x;\delta) = \begin{cases} \exp(\delta x) - 1 & x > 0 \\ 1 - \exp((-\delta x)) & x < 0 \end{cases}$$

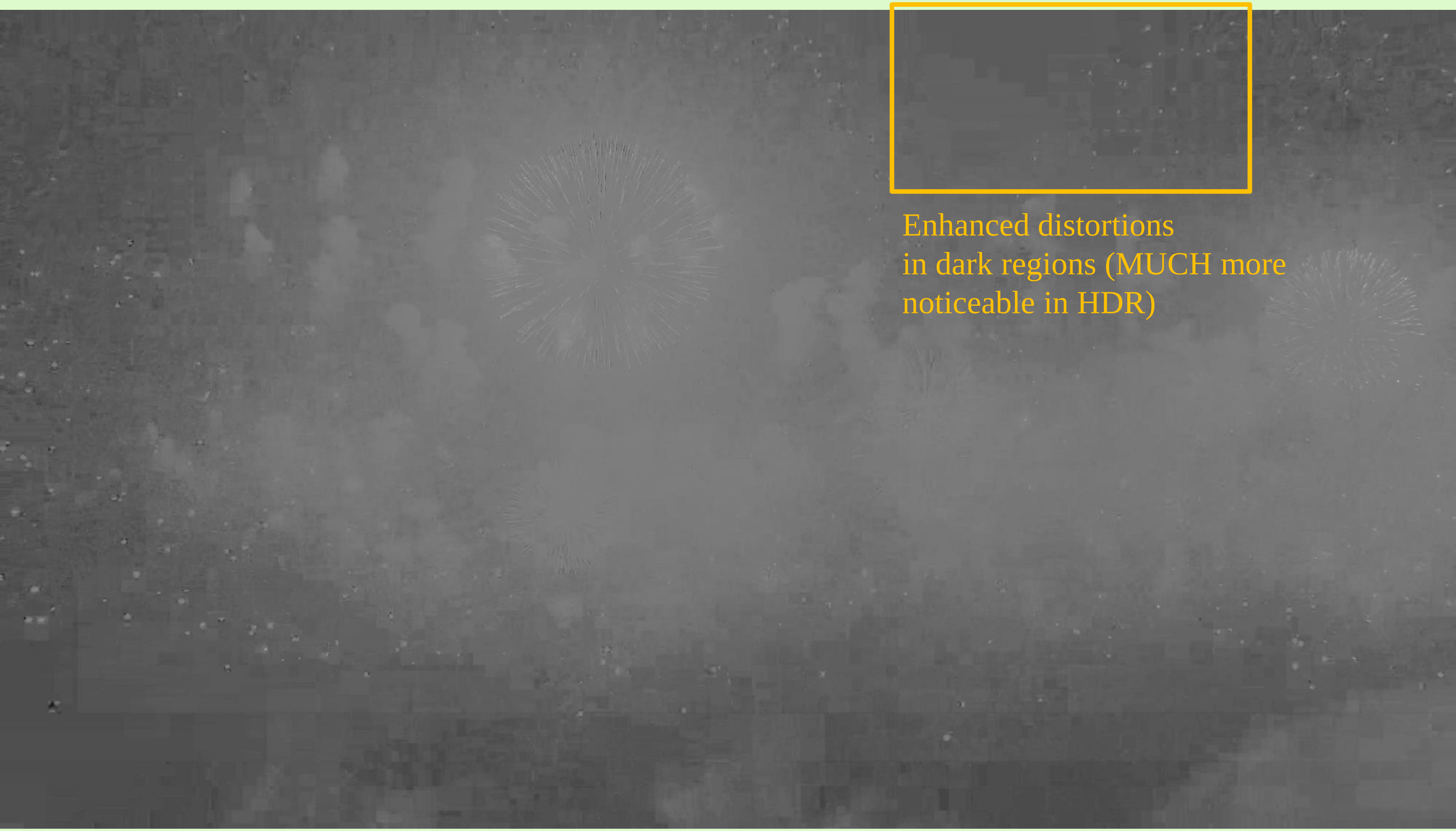


- **Midrange luminances** (or chrominances) are **crushed** to **near zero.**
- **"HDR" regions** (and VQA responses to them) **now dominate.**
- Still obey **natural video statistic models.**
- Even **better performance** is obtained when HDRMAX is applied on a **patch-wise basis** (WxW patches)

Problem


Concept (this is NOT HDR)
Hard to capture distortions
in dark regions (MUCH more
noticeable on HDR)

HDRMAX


Enhanced distortions
in dark regions (MUCH more
noticeable in HDR)

Improvements are Dramatic

MEDIAN SROCC, LCC, AND RMSE ON 10 BIT VQA DATABASES OBTAINED USING FR MODELS. STANDARD DEVIATIONS ARE SHOWN IN PARENTHESES. THE BEST PERFORMING ALGORITHM IS BOLD-FACED.

Dataset	LIVE HDR		LIVE ETRI (SDR 10 bit)		Livestream (SDR 8 bit)	
Algorithm	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
SSIM	0.5208(0.1611)	0.4898(0.1595)	0.3568(0.2625)	0.3358(0.2395)	0.6539(0.0927)	0.6584(0.0832)
SSIM+HDRMAX	0.7771(0.0866)	0.7529(0.0964)	0.8485(0.0733)	0.8301(0.0741)	0.7521(0.0714)	0.7689(0.0619)
MS-SSIM	0.6007(0.1228)	0.5810(0.1260)	0.5234(0.2336)	0.5319(0.2279)	0.7306(0.1097)	0.7377(0.1083)
MS-SSIM+HDRMAX	0.7645(0.0838)	0.7258(0.0868)	0.7519(0.1399)	0.7297(0.1328)	0.7397(0.0712)	0.7724(0.0681)
ST-RRED	0.6863(0.0700)	0.6569(0.0744)	0.7500(0.0853)	0.7587(0.0933)	0.6122(0.0738)	0.6273(0.0637)
ST-RRED+HDRMAX	0.7896(0.0607)	0.7595(0.0603)	0.8628(0.0889)	**0.8535(0.0840)**	**0.7685(0.0690)**	**0.7902(0.0630)**
SpEED-QA	0.611(0.1243)	0.6196(0.1066)	0.7031(0.1485)	0.7179(0.1565)	0.5561(0.0481)	0.5891(0.0454)
SpEED-QA+HDRMAX	0.7581(0.0921)	0.7107(0.0993)	0.8597(0.0971)	0.8355(0.0907)	0.6519(0.0416)	0.6642(0.0374)
VMAF	0.6753(0.0493)	0.6086(0.0583)	0.5617(0.0919)	0.5069(0.0844)	0.6424(0.0574)	0.7050(0.0498)
VMAF+HDRMAX	**0.8528(0.0543)**	**0.8342(0.0632)**	**0.8654(0.1076)**	0.8417(0.0996)	0.7050(0.0853)	0.7120(0.0944)

MS-SSIM and VMAF are global standards that control the quality of >70% of Internet bits

HDRMAX allows streamers to adaptively **adjust bit depth** vs **compression** to conserve video quality, bandwidth, and the environment.

Summary

- Accurate video quality prediction, **unsolved since Edison's Kinetograph,** has become possible.

- By **modeling the statistical responses of visual neurons** to distortion – not by measuring distortion directly.

- What about **Deep Learning / AI?**
  - **Less gain** on the **Referenc**e problem, which is "more tractable"
  - It is the **key to** the **No-Reference** problem

LIVE's Current Sponsors

NSF

NETFLIX

YouTube

amazon prime video

Meta

Video Clarity
Tools for Video Analysis

REALITY LABS

ShareChat

U.S. DEPARTMENT OF COMMERCE
NTIA
NATIONAL TELECOMMUNICATIONS & INFORMATION ADMINISTRATION

Viasat

ERICSSON

Questions?