Realistic Image Synthesis

- Perception: Image Quality Metrics -

Philipp Slusallek
Karol Myszkowski
Gurprit Singh
Corentin Salauin

Outline

- **Questions of Appearance Preservation**
- **Basic characteristics of Human Visual System in image perception**
- **Daly's Visible Differences Predictor (VDP)**
- **Metric for rendering artifacts**
  - Full-reference CNN-based metric

Image Quality Metrics

- **Application examples which require metrics of the image quality as perceived by the human observer**
  - Lossy image compression and broadcasting
  - Design of image input/output devices
    - scanners, cameras, monitors, printers, and so on
  - Watermarking
  - Computer graphics, medical visualization

Questions of Appearance Preservation

- **The concern is not whether images are the same**
- **Rather the concern is whether images appear the same.**

How much computation is enough?

How much reduction is too much?

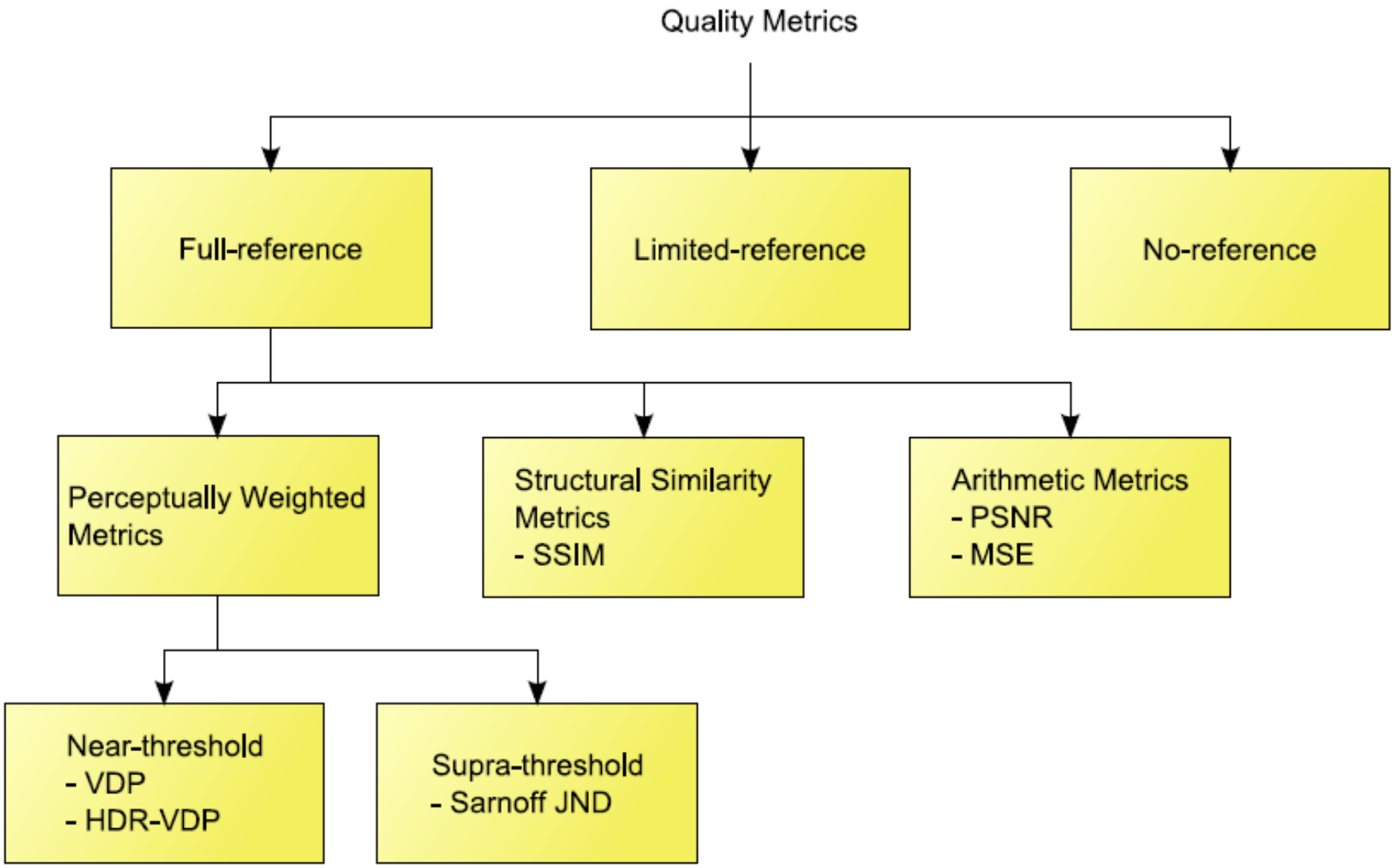
Subjective Methods

- **The best results can be obtained when human observers are involved**
  - Carefully controlled observation conditions
  - Representative number of participants
    - Averaging individual visual characteristics
    - Limiting the influence of emotional reactions
- **Very costly**
- **Limited use in practical routine applications**

Objective Methods

- **Usually rely on the comparison of images against the reference image**
  - – Measure perceivable differences between images, but an absolute measure of the image quality is difficult to obtain
  - – Not always in good agreement with the subjective measures
  - \+ Good repeatability of results
  - \+ Easy to use
  - \+ Low costs

Classification of Objective Quality Metrics



Classification of Objective Quality Metrics

- **Full-reference (FR)** where the reference image is available as it is typical in image compression, restoration, enhancement and reproduction applications.
- **Limited-reference (RR)** where a certain number of features characteristic for the image is extracted and made available as reference through a back-channel with reduced distortion. To avoid the back-channel transmission, known in advance and low magnitude signals, such that their visibility is prevented (as in watermarking), are directly encoded into an image and then the distortion of these signals is measured after the image transmission on the client side.
- **No-reference (NR)** which are focused mostly on detecting distortions which are application specific and predefined in advance such as blockiness (typical for DCT encoding in JPEG and MPEG), and ringing and blurring (typical for wavelet encoding in JPEG2000).

Full-reference Quality Metrics (1)

- **Pixel-based Metrics** with the mean square error (MSE) and the peak signal-to-noise ratio (PSNR) difference metrics as the prominent examples. In such a simple framework the HVS considerations are usually limited to the choice of a perceptually uniform color space such as CIELAB and CIELUV, which is used to represent the reference and distorted image pixels.
- **Structure-based Metrics** with the *Structural SIMilarity (SSIM) index* one of the most popular and influential quality metric in recent years. Since the HVS is strongly specialized in learning about the scenes through extracting structural information, it can be expected that the perceived image quality can be well approximated by measuring structural similarity between images.

Full-reference Quality Metrics (2)

- **Perception-based Fidelity Metrics** the *visible difference predictor* (VDP) and the *Sarnoff visual discrimination model* (VDM) as the prominent examples. These contrast-based metrics are based on advanced models of early vision in the HVS and are capable of capturing just visible (near threshold) differences or even measuring the magnitude of such (supra-threshold) differences and scale them in JND (just noticeable difference) units.

Pixel–based Metrics: Mean Square Error

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \frac{1}{n}\sum_{i,j}(P_{ij} - Q_{ij})^2$$

$$\mathrm{PSNR} = 20\log_{10}\frac{\mathrm{Pixel}_{Max}}{\mathrm{MSE}}$$

Reference image (P)

Compared images (Q)

Pixel–based Metrics: Mean Square Error

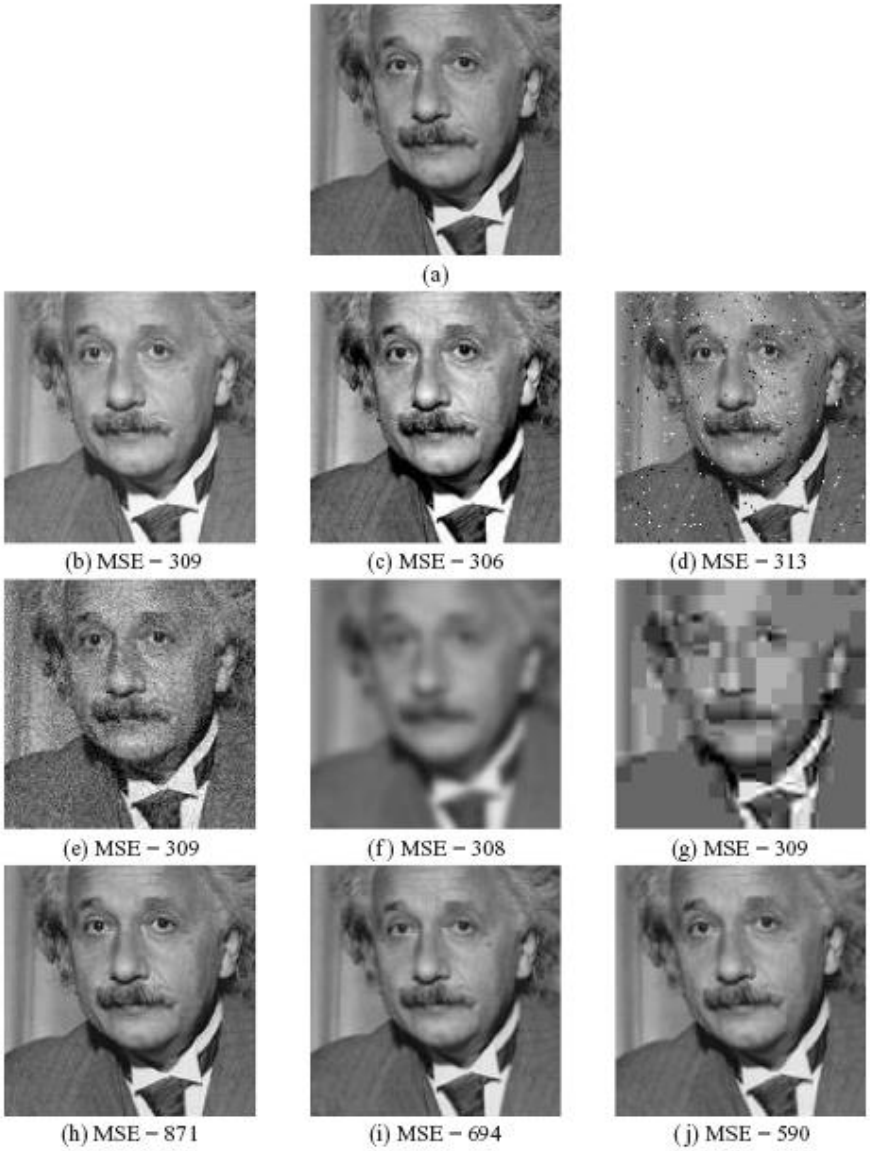
$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \frac{1}{n}\sum_{i,j}(P_{ij} - Q_{ij})^2$$

$$\mathrm{PSNR} = 20\log_{10}\frac{\mathrm{Pixel}_{Max}}{\mathrm{MSE}}$$

Reference image (P)

Compared images (Q)

Pixel–based Metrics: Mean Square Error



Einstein image altered with different types of distortions:
(a) “original image”;
(b) mean luminance shift;
(c) a contrast stretch;
(d) impulsive noise contamination;
(e) white Gaussian noise contamination;
(f) blurring;
(g) JPEG compression;
(h) a spatial shift (to the left);
(i) spatial scaling (zooming out);
(j) a rotation.

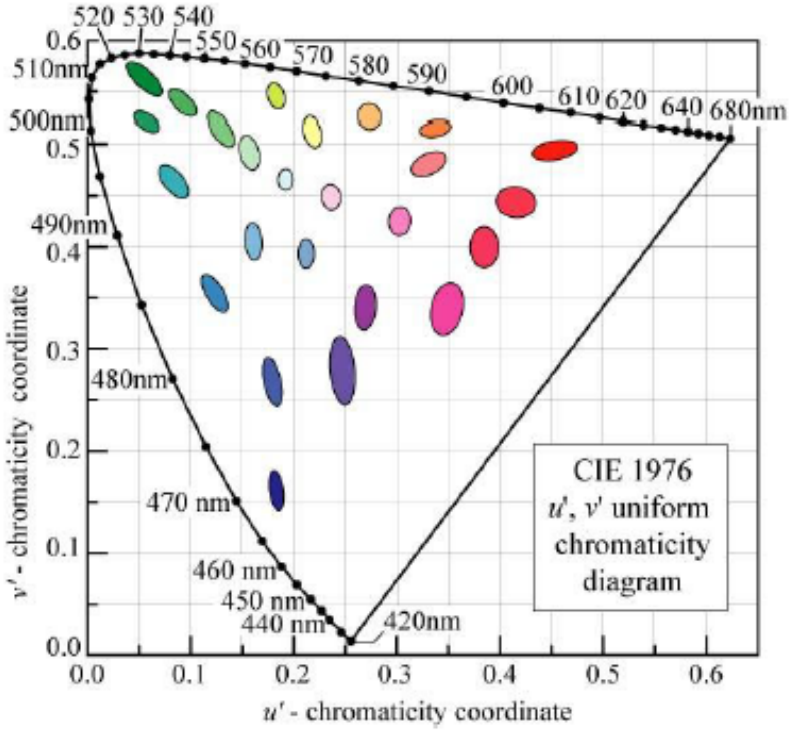
Note that images (b)–(g) have almost the same MSE values but drastically different visual quality. Also, note that the MSE is highly sensitive to spatial translation, scaling, and rotation [Images (h)–(j)].

Color Appearance Spaces

- CIE 1976 L*u*v* and L*a*b*
  - Color (X, Y, Z) reflected by a surface under known illuminant (X_n, Y_n, Z_n) ("white point")
  - $f(r) = \begin{cases} r^{1/3} & \text{if } r > 0.008856 \quad \text{(log-like)} \\ 7.787r + 16/116 & \text{otherwise} \end{cases}$
  - $L^* = 116\, f(Y/Y_n) - 16$
  - $u' = 4X / (X+15Y+3Z)$
    $v' = 9Y / (X+15Y+3Z)$
  - $u^* = 13\, L^* (u'-u_n')$
    $v^* = 13\, L^* (v'-v_n')$
  - $a^* = 500\, [f(X/X_n) - f(Y/Y_n)]$
    $b^* = 200\, [f(Y/Y_n) - f(Z/Z_n)]$
  - Euclidean distances ΔE^*_{uv} and ΔE^*_{ab}

Color Appearance Spaces

- u'v' chromaticity diagram
  - Deformed ellipses
- CIELUV and CIELAB
  - Close to uniform
  - Useful for practical color differences
  - Not perfect



Full-reference Quality Metrics

- **Structure-based Metrics** with the *Structural SIMilarity (SSIM) index* one of the most popular and influential quality metric in recent years.
- Since the HVS is strongly specialized in learning about the scenes through extracting structural information, it can be expected that the perceived image quality can be well approximated by measuring structural similarity between images.

Structural SIMilarity (SSIM) index

- The SSIM index decomposes similarity estimation into three independent comparison functions: **luminance**, **contrast**, and **structure**.
- The **luminance** comparison function l(x, y) for an image pair x and y is specified as:

$$l(x,y) = l(\mu_x, \mu_y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad \text{where} \quad \mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The **contrast** comparison function c(x, y) is specified as:

$$c(x,y) = c(\sigma_x, \sigma_y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad \text{where} \quad \sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2}$$

- The **structure** comparison function s(x, y) is specified as:

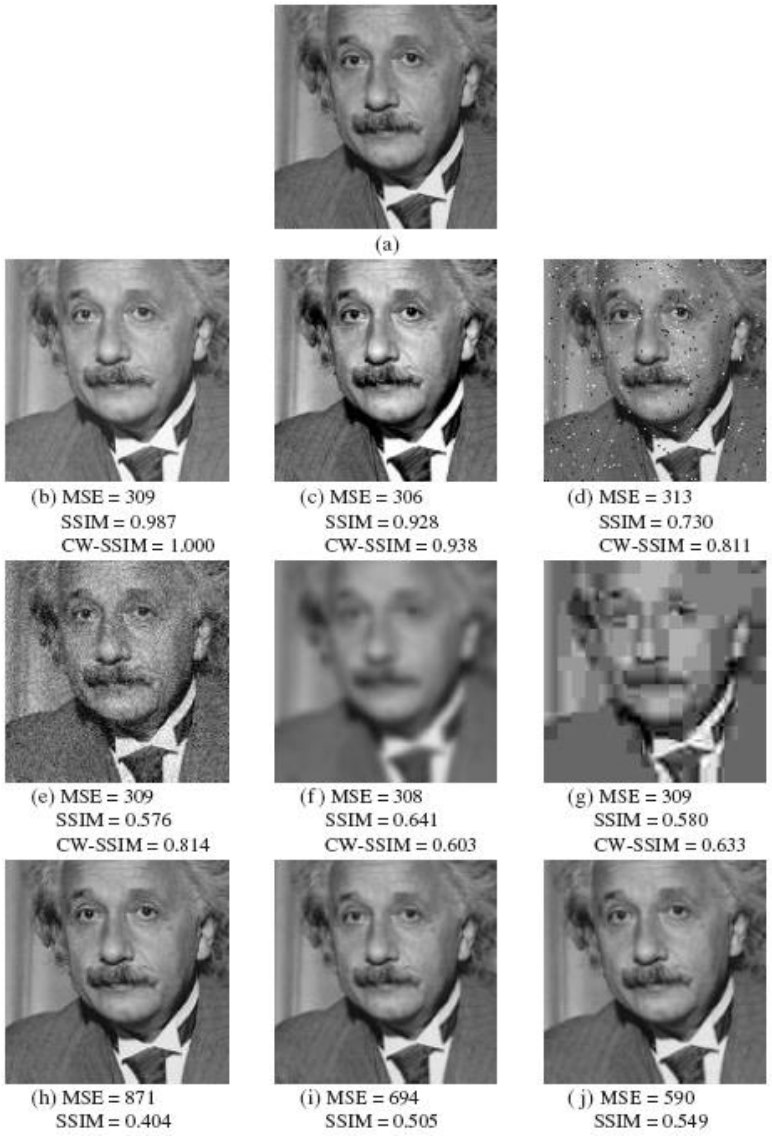
$$s(x,y) = s\left(\frac{x-\mu_x}{\sigma_x}, \frac{y-\mu_y}{\sigma_y}\right) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad \text{where} \quad \sigma_{xy} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}$$

- The three comparison functions are combined in the SSIM index:

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

- To obtain a local measure of structure similarity all statistics μ, σ are computed within a local 8 × 8 window which slides over the whole image.

Structural SIMilarity (SSIM) index



Einstein image altered with different types of distortions:
(a) “original image”;
(b) mean luminance shift;
(c) a contrast stretch;
(d) impulsive noise contamination;
(e) white Gaussian noise contamination;
(f) blurring;
(g) JPEG compression;
(h) a spatial shift (to the left);
(i) spatial scaling (zooming out);
(j) a rotation.

Images (b)–(g) drastically different visual quality and SSIM captures well such quality degradation. Also, note that the SSIM is highly sensitive to spatial translation, scaling, and rotation [Images (h)–(j)].

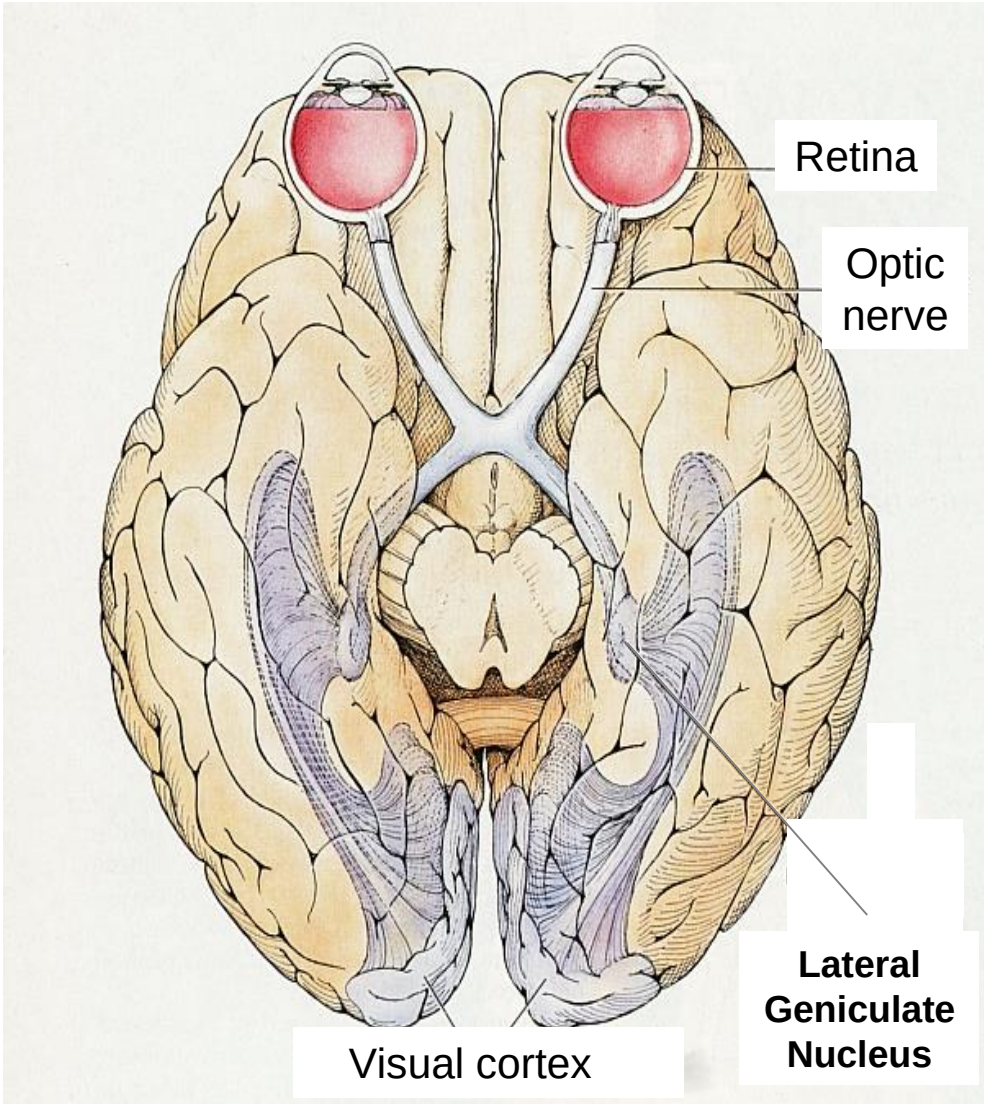
Human Visual System (HVS)

vs. Image Quality Metrics

- **Anatomy and physiology of visual pathway determine its sensitivity on various image elements.**
- **Basic HVS characteristics must be taken into account to estimate perceivable differences between images.**
- **Complete model of image perception has not been elaborated so far.**

Visual Pathway

- Functionality of visual pathway from retina to the visual cortex are relatively well understood.
- Modeling on the physiological level too complex.
- Behavioral models acquired through psychophysical experiments are easy to use.



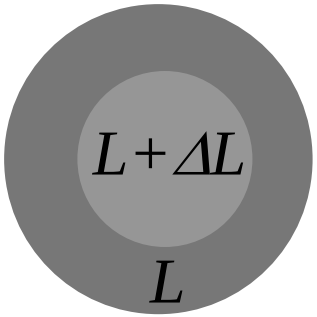
Important Characteristics of the HVS

- **Visual adaptation**
- **Temporal and spatial mechanisms (channels) which are used to represent the visual information at various scales and orientations as it is believed that primary visual cortex does.**
- **Contrast Sensitivity Function which specifies the detection threshold for a stimulus as a function of its spatial and temporal frequencies.**
- **Visual masking affecting the detection threshold of a stimulus as a function of the interfering background stimulus which is closely coupled in space and time.**

Visual Adaptation

Ernst Heinrich Weber
[From wikipedia]

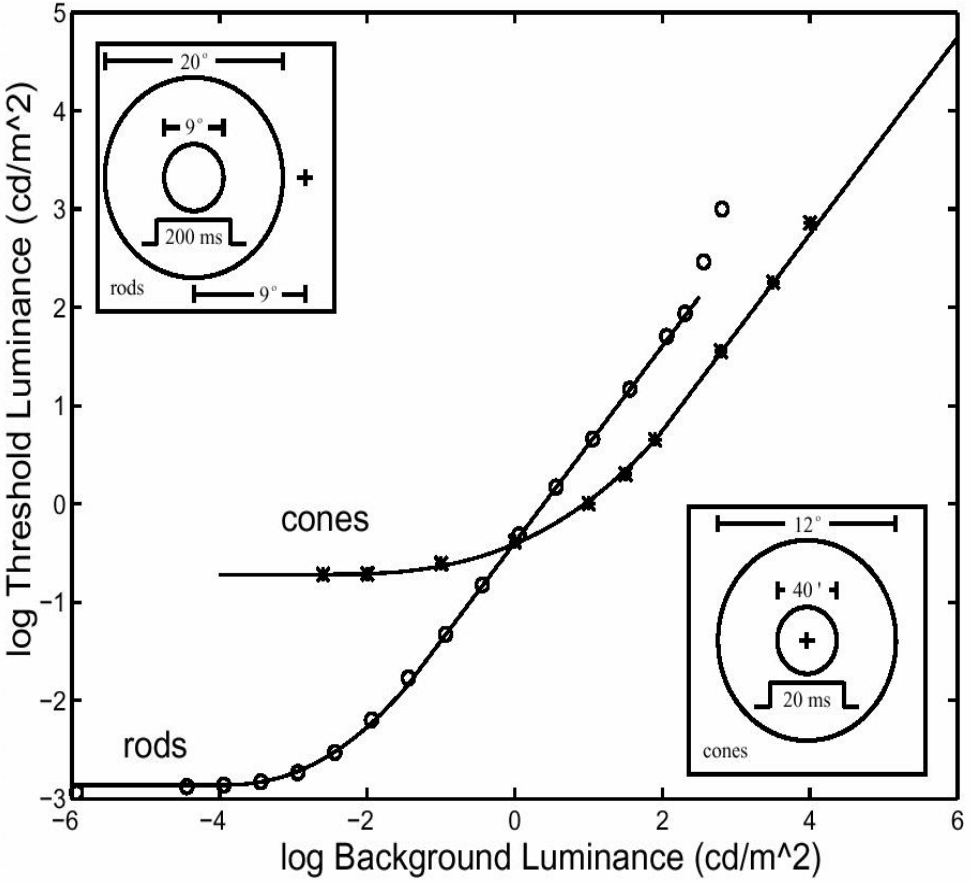
- Adaptation of visual system to various levels of background luminance



- Weber's law:

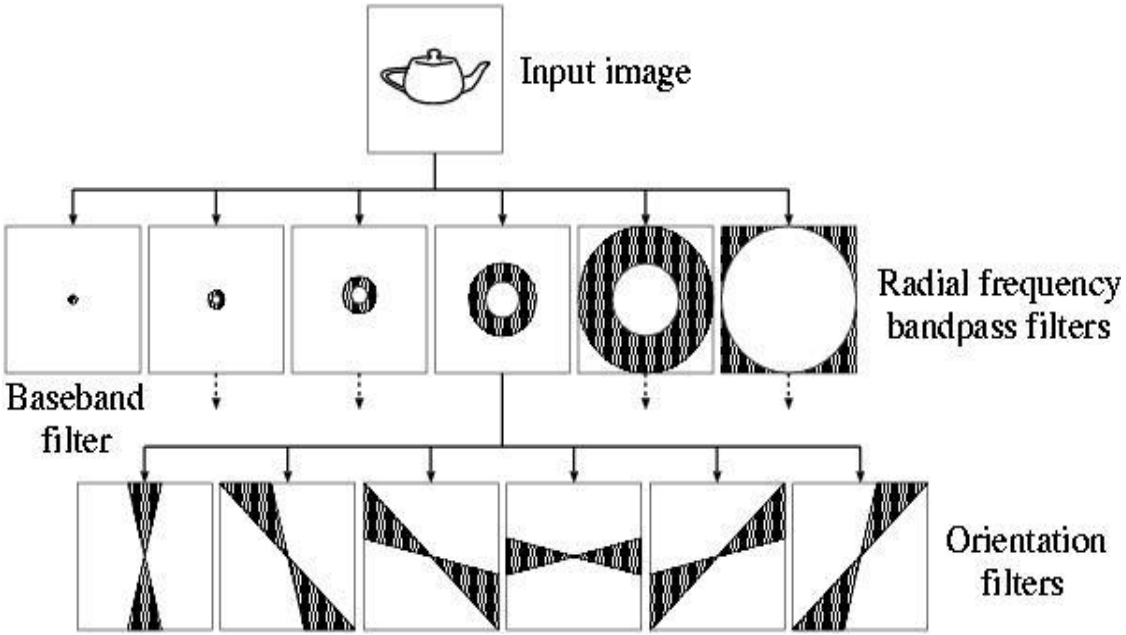
$$\frac{\Delta L}{L} = const$$

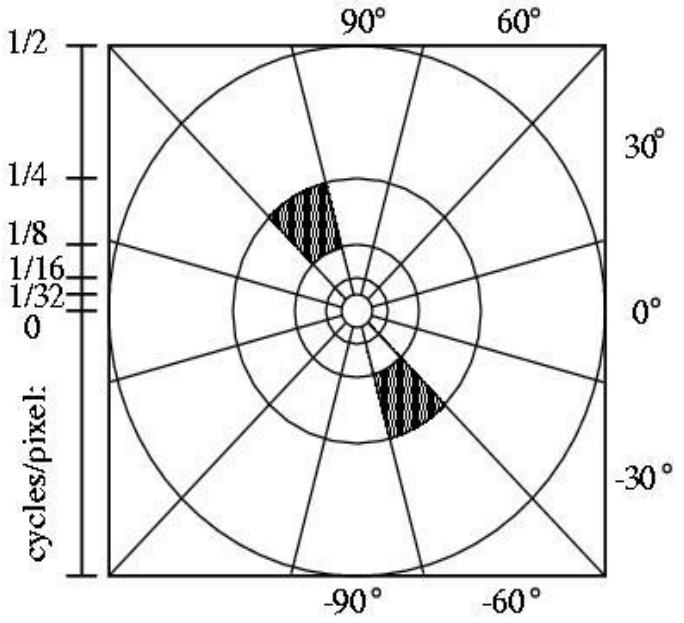
TVI – Threshold *versus* Intensity function



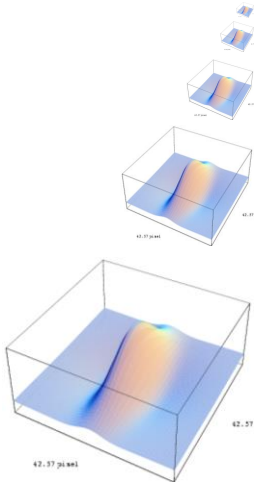
Ferwerda et al.

Cortex Transform: Filter Bank

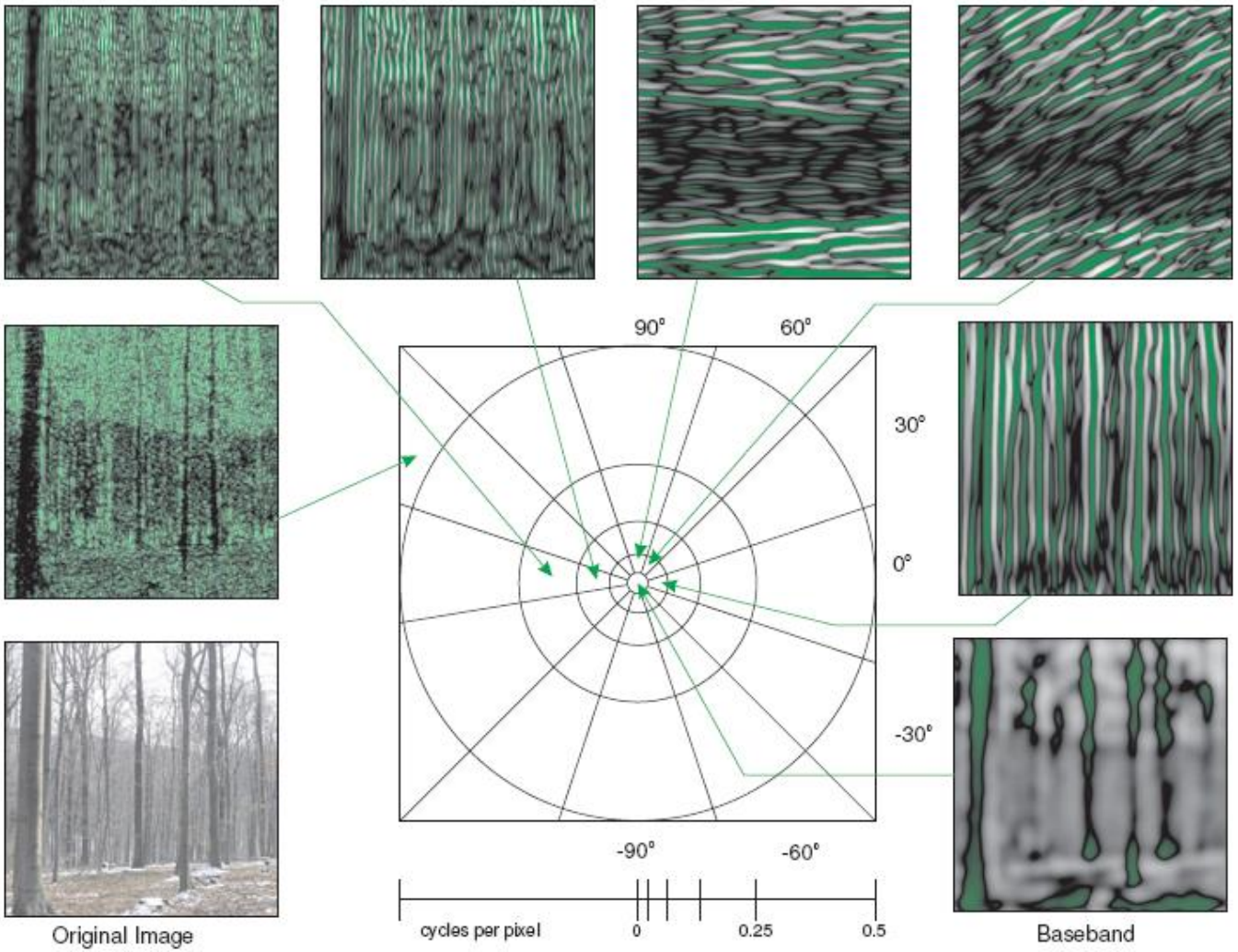




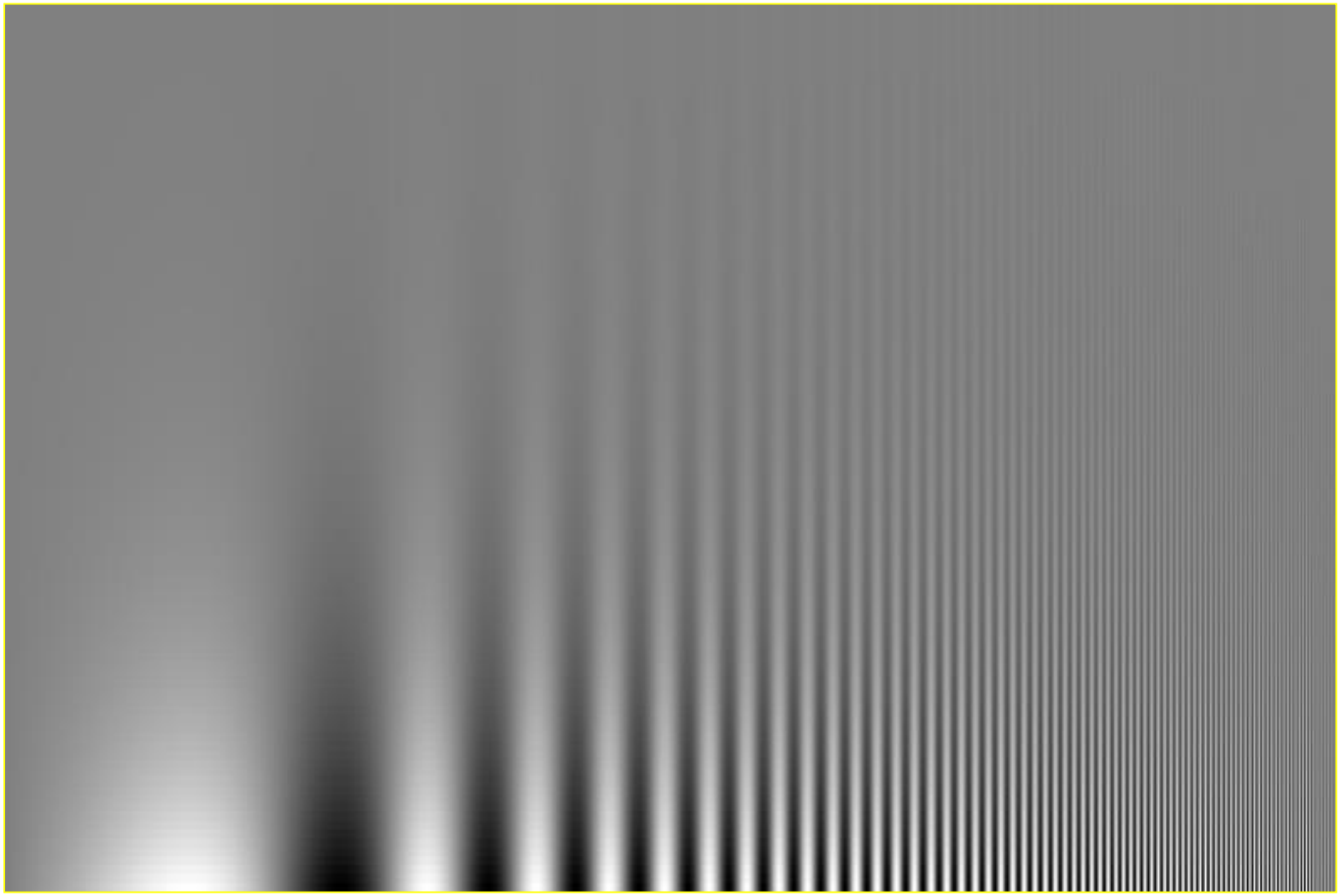
Filter bank examples: Gabor functions (Marcelja80), steerable pyramid transform (Simoncelli92), Discrete Cosine Transform (DCT), difference of Gaussians (Laplacian) pyramids (Burt83, Wilson91), Cortex transform (Watson87, Daly93).



Cortex Transform: Frequency and Orientation Bands

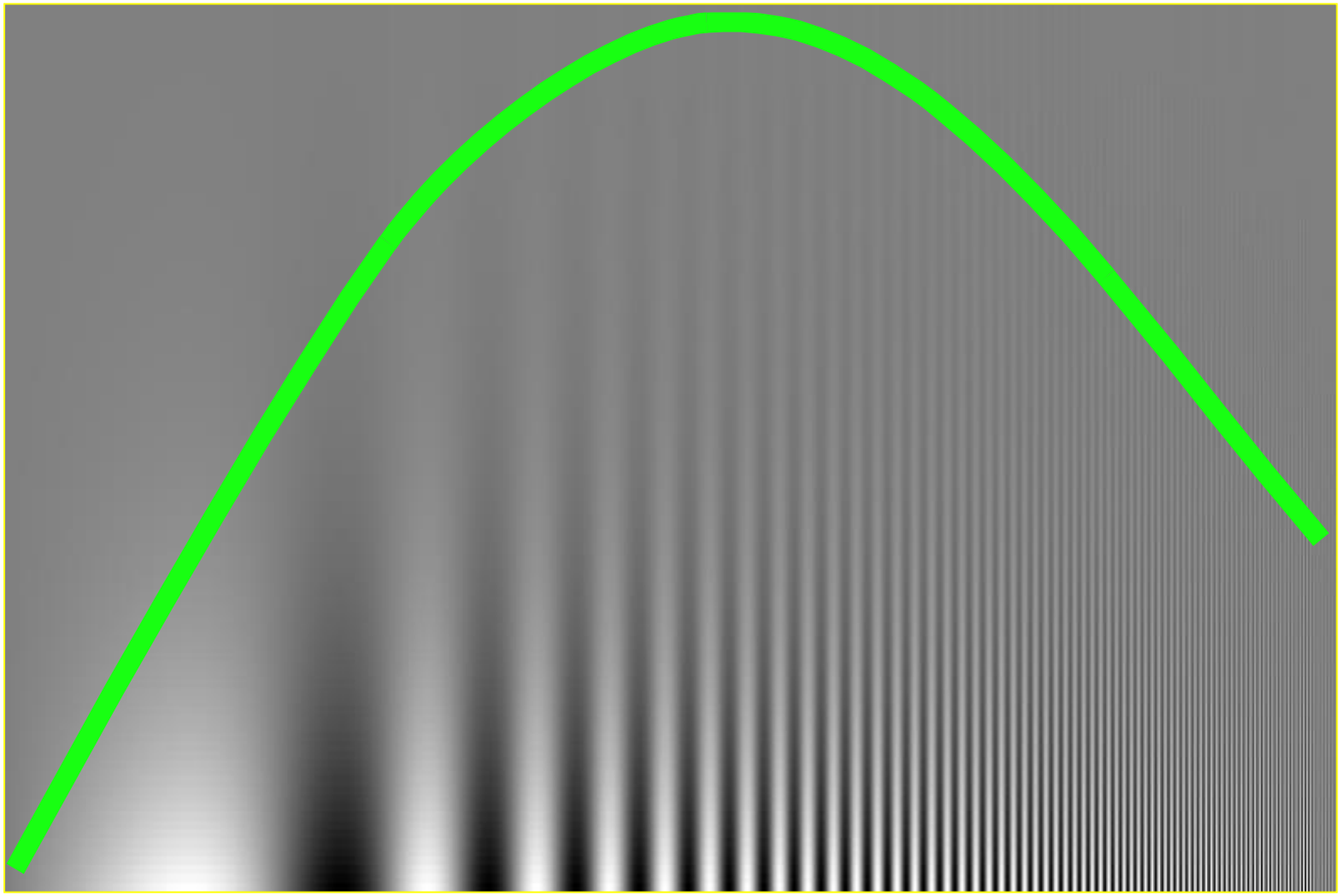


Contrast Sensitivity Function



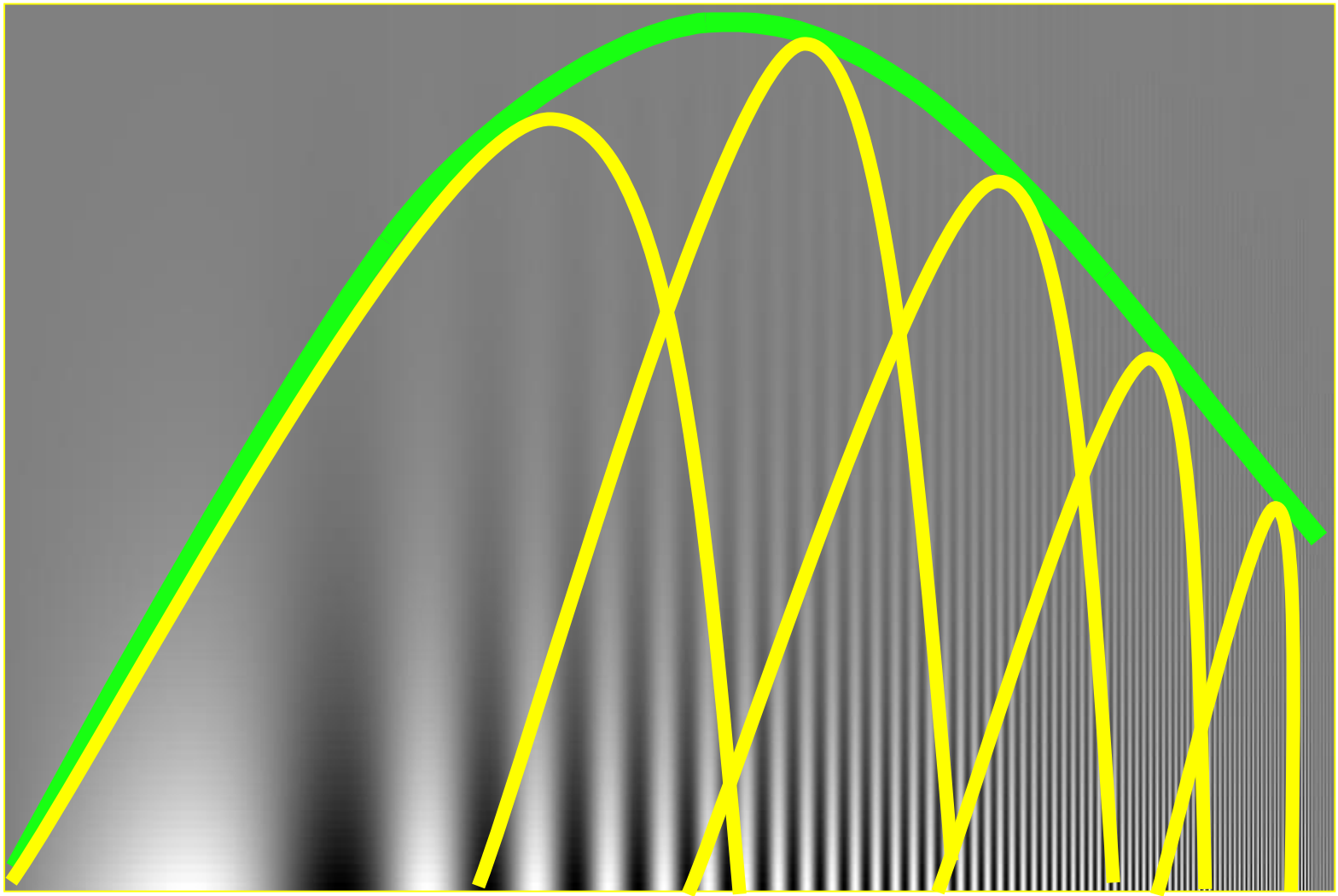
J.G. Robson CSF chart

Contrast Sensitivity Function



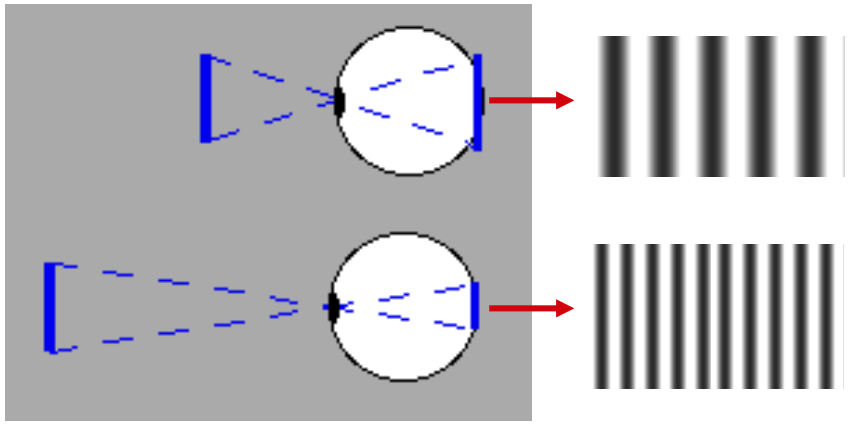
J.G. Robson CSF chart

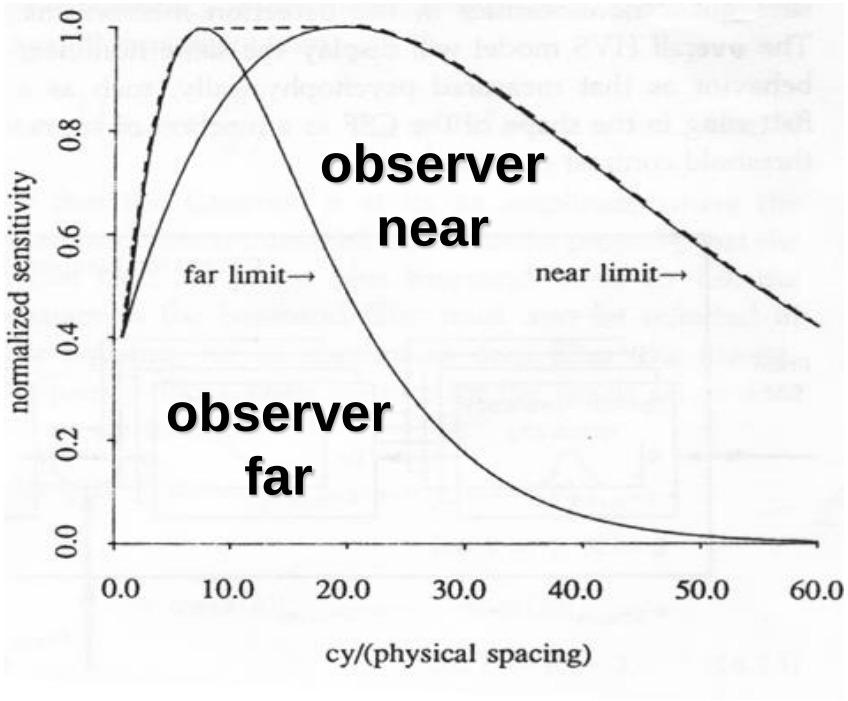
Contrast Sensitivity Function (CSF)



J.G. Robson CSF chart

CSF *versus* Observation Distance

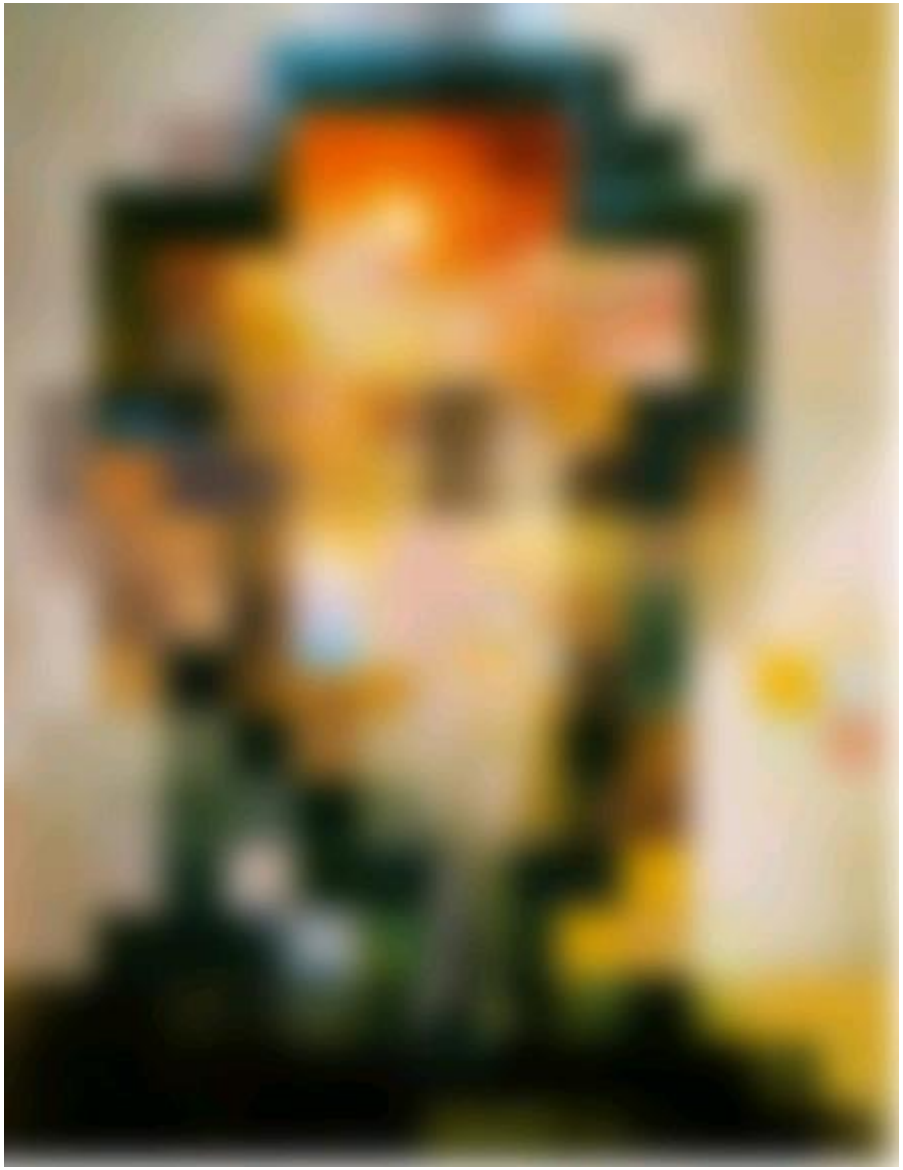




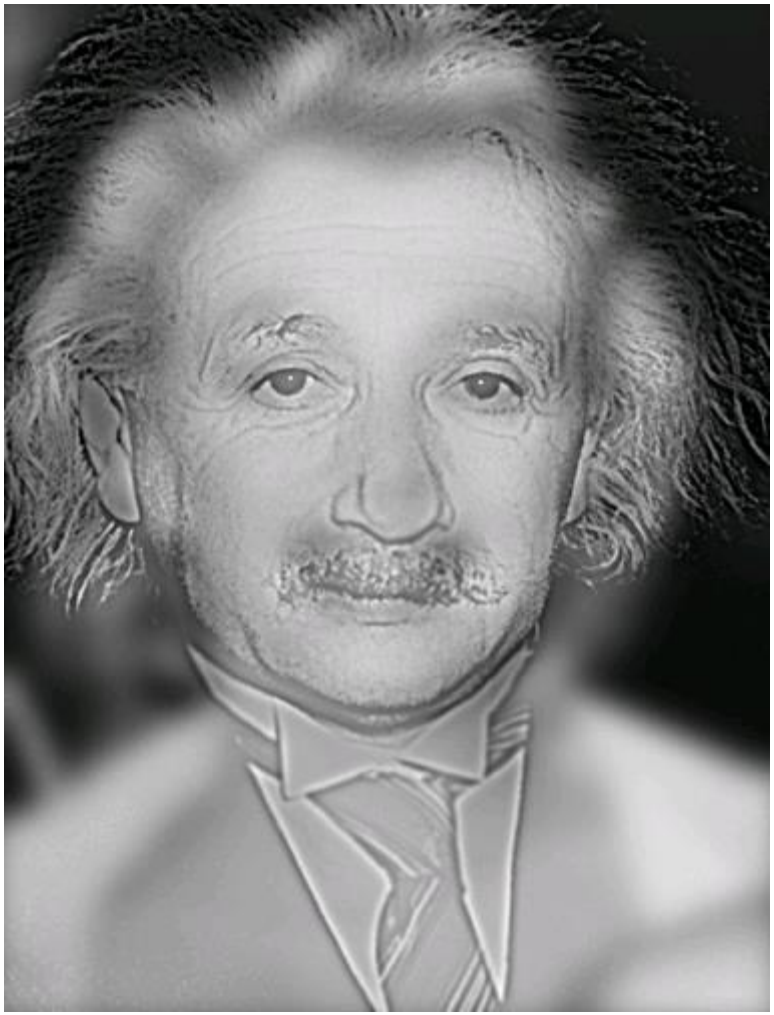
- **Spatial frequencies projected on the retina increase proportionally to the observation distance.**
- **Image elements represented by low (high) spatial frequencies might become visible (invisible) with the increase of the observation distance.**

To estimate conservatively the image quality for variable observer positions the envelope of CSFs for the extreme observer locations can be used.

Lincoln illusion



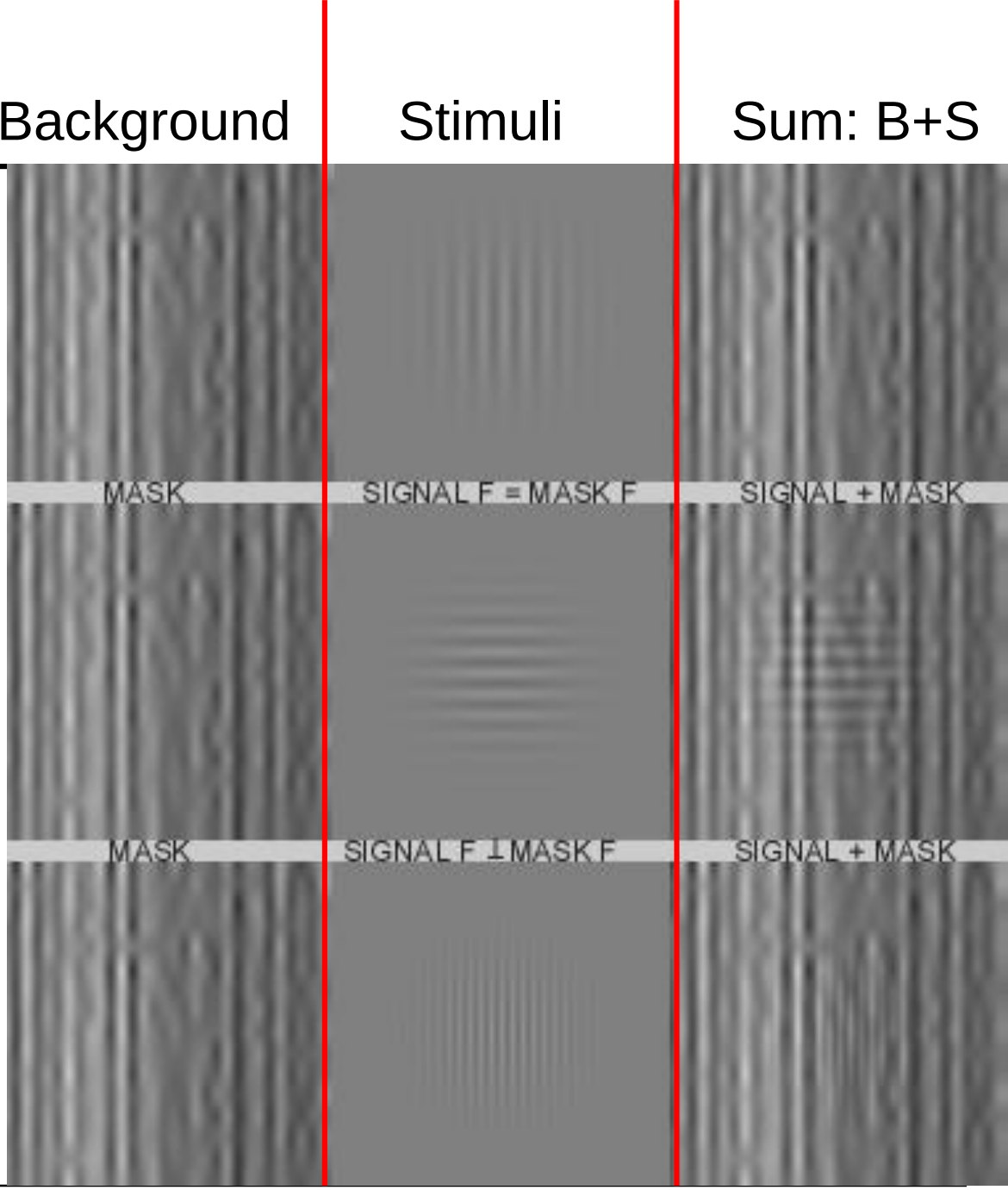
Hybrid Images



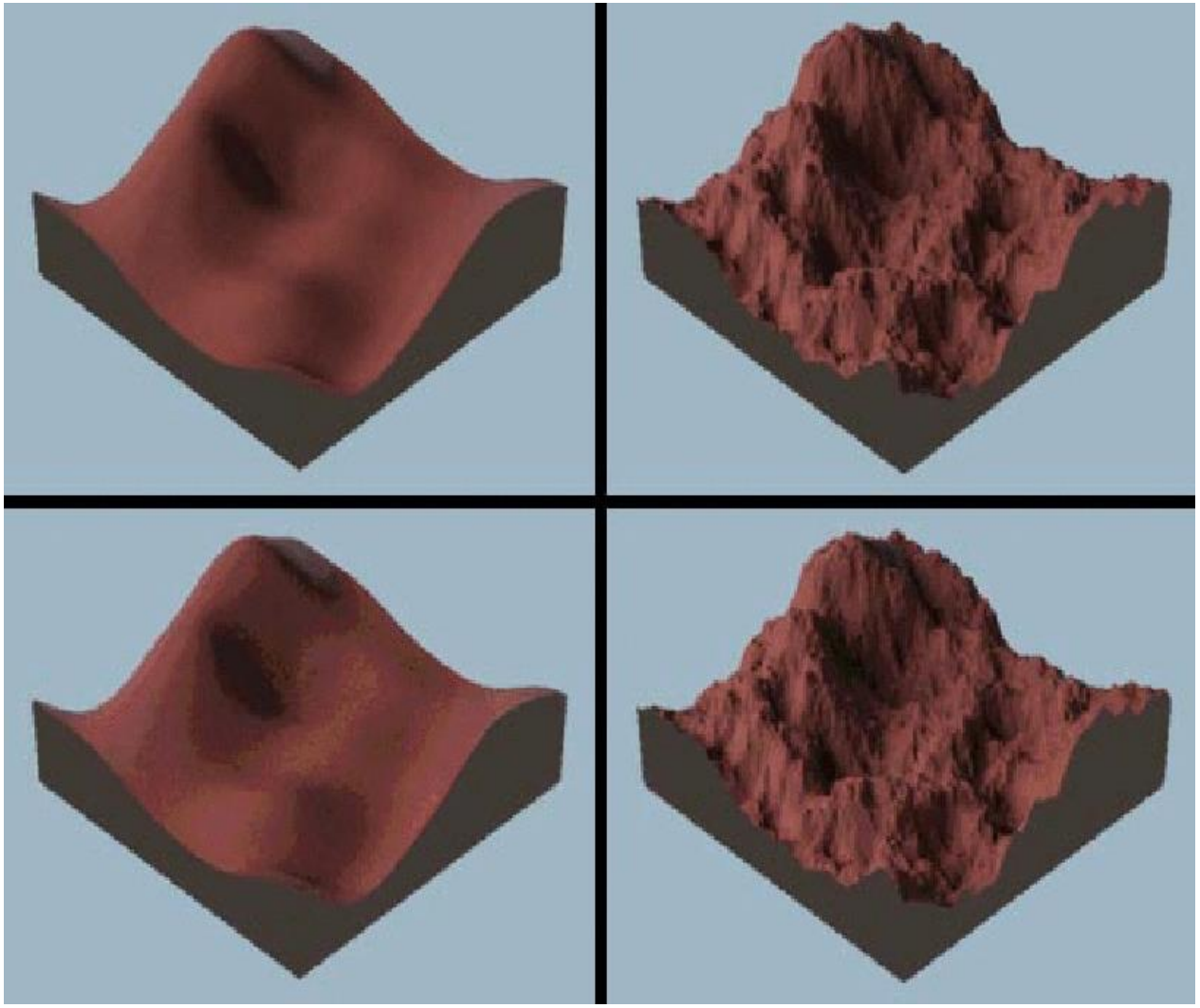
Hybrid Images

Visual Masking

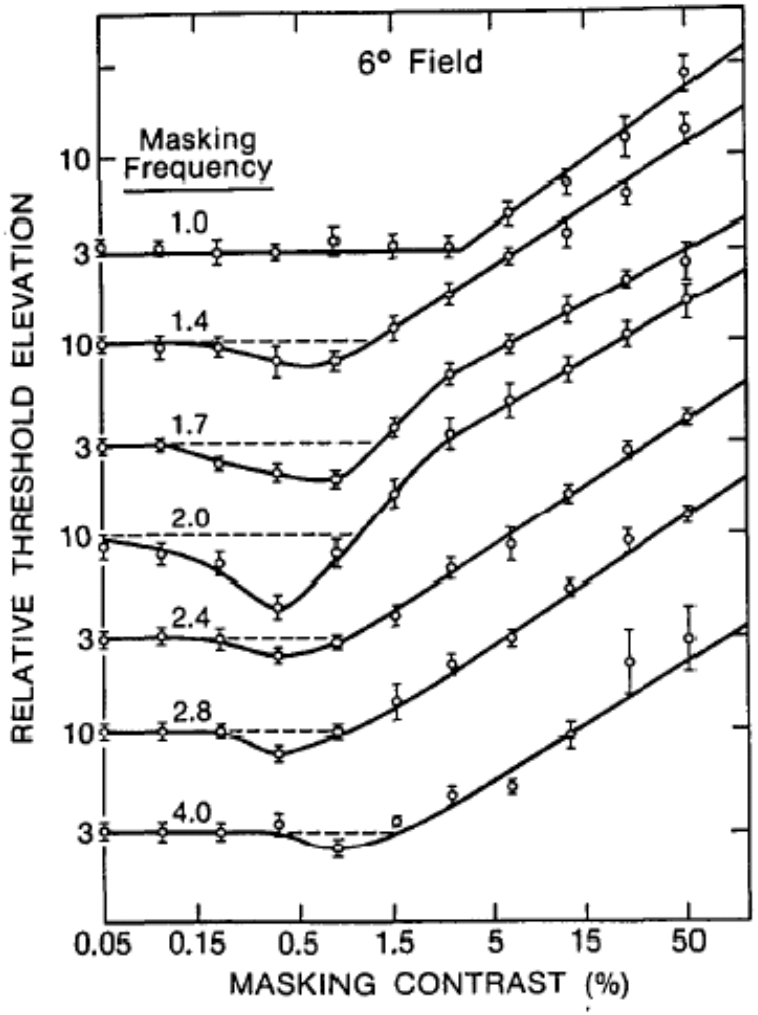
- **Strong masking: similar spatial frequencies**
- **Weak masking: different orientations**
- **Weak masking: different spatial frequencies**



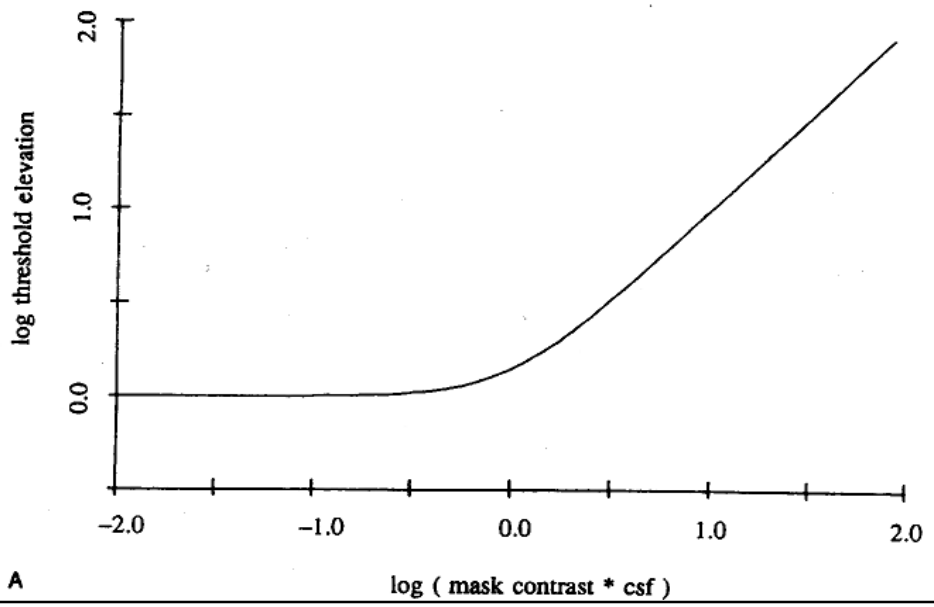
Visual Masking Example



Visual Masking Model

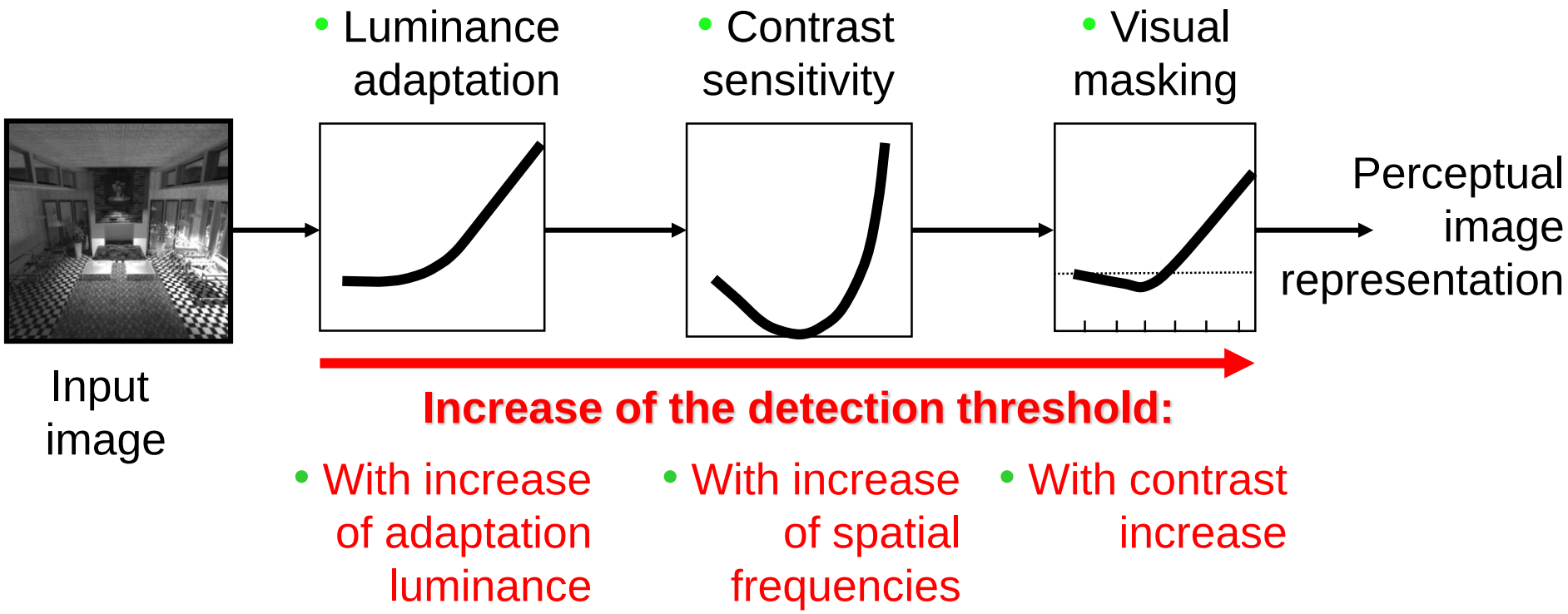


- **Masking is strongest between stimuli located in the same perceptual channel, and many vision models are limited to this intra-channel masking.**
- **The following threshold elevation model is commonly**

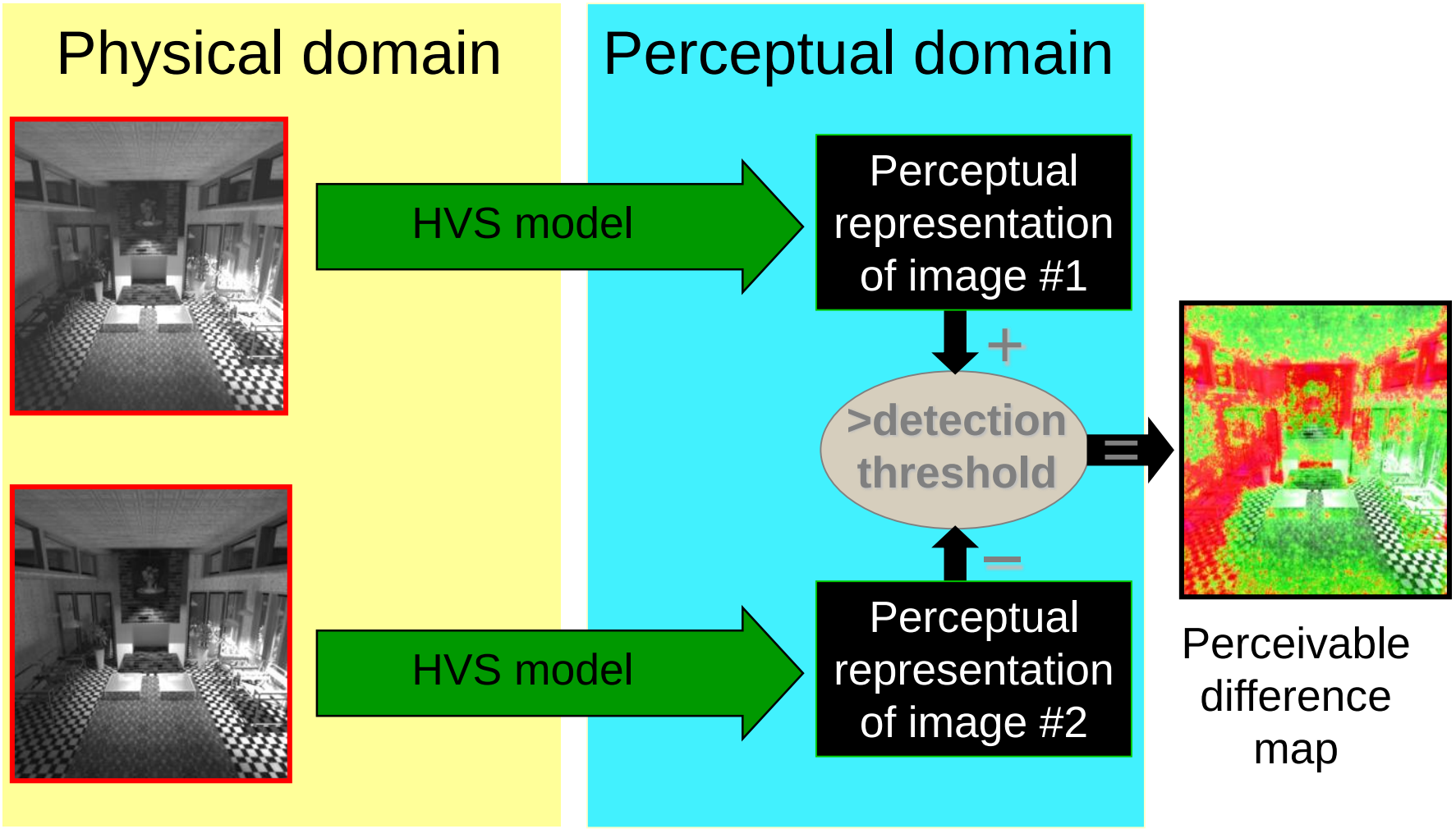


Typical HVS Model

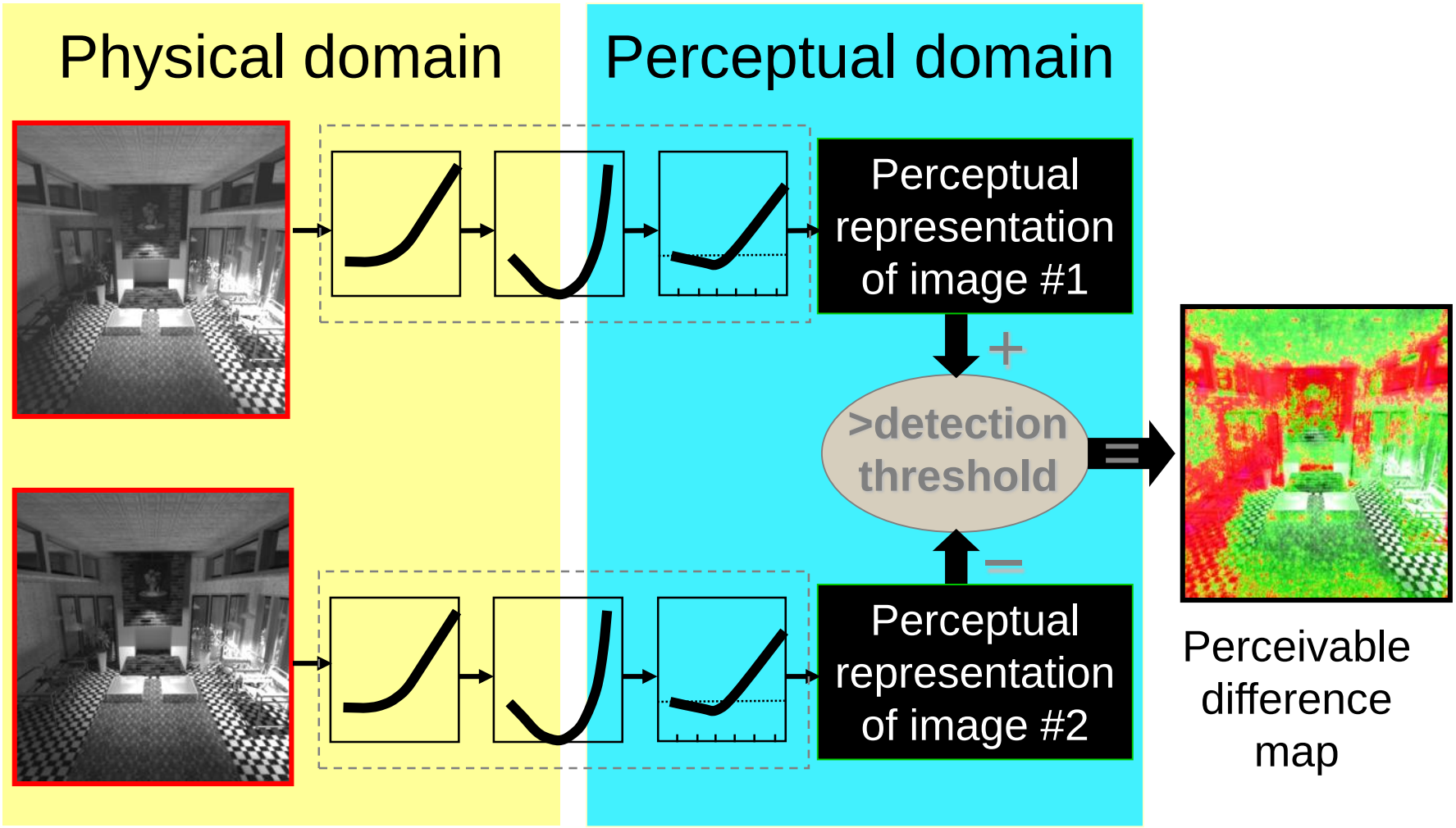
Detection of perceivable differences between images strongly depends on the following characteristics of the human visual system:



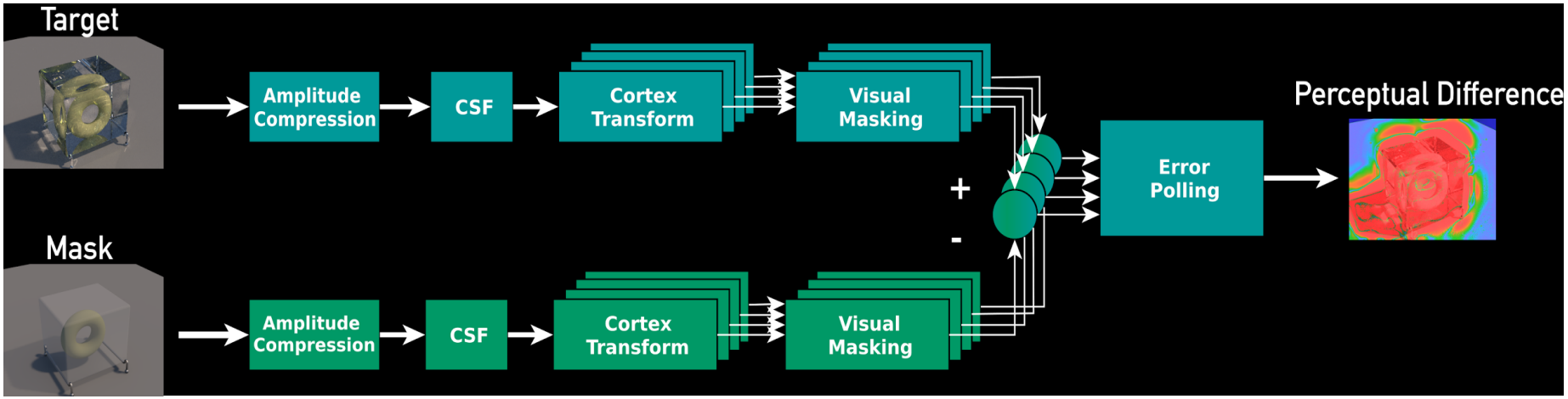
Perceivable Differences Predictor



Perceivable Differences Predictor

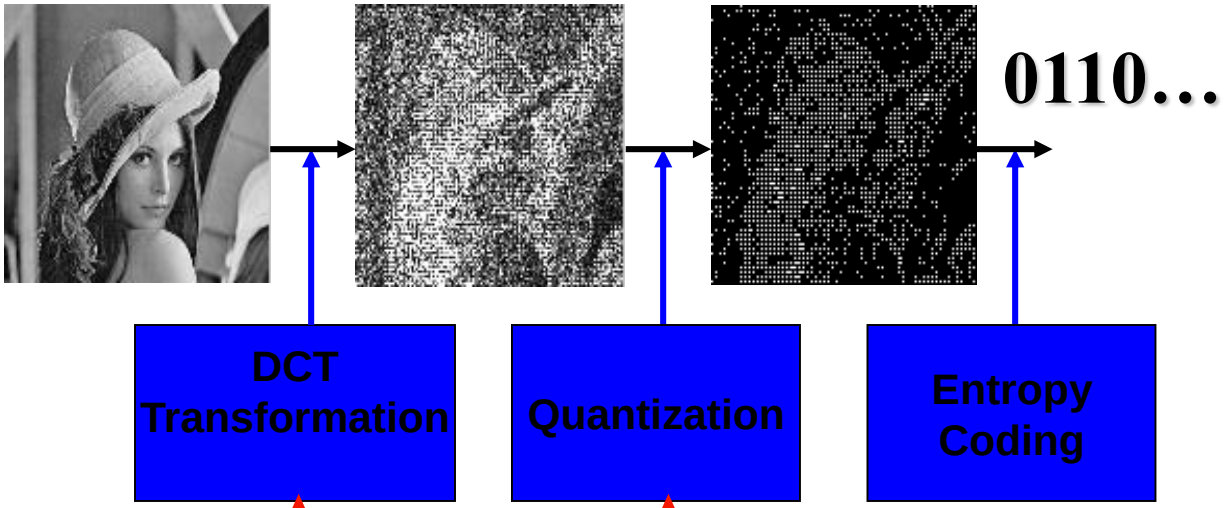


Daly's Visible Differences Predictor



Application Example –

Lossy Image Compression



16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Quantization matrix in JPEG [Annex K]

1 Image representation obtained as the result of DCT transformation should approximate the image representation in the Visual Cortex.

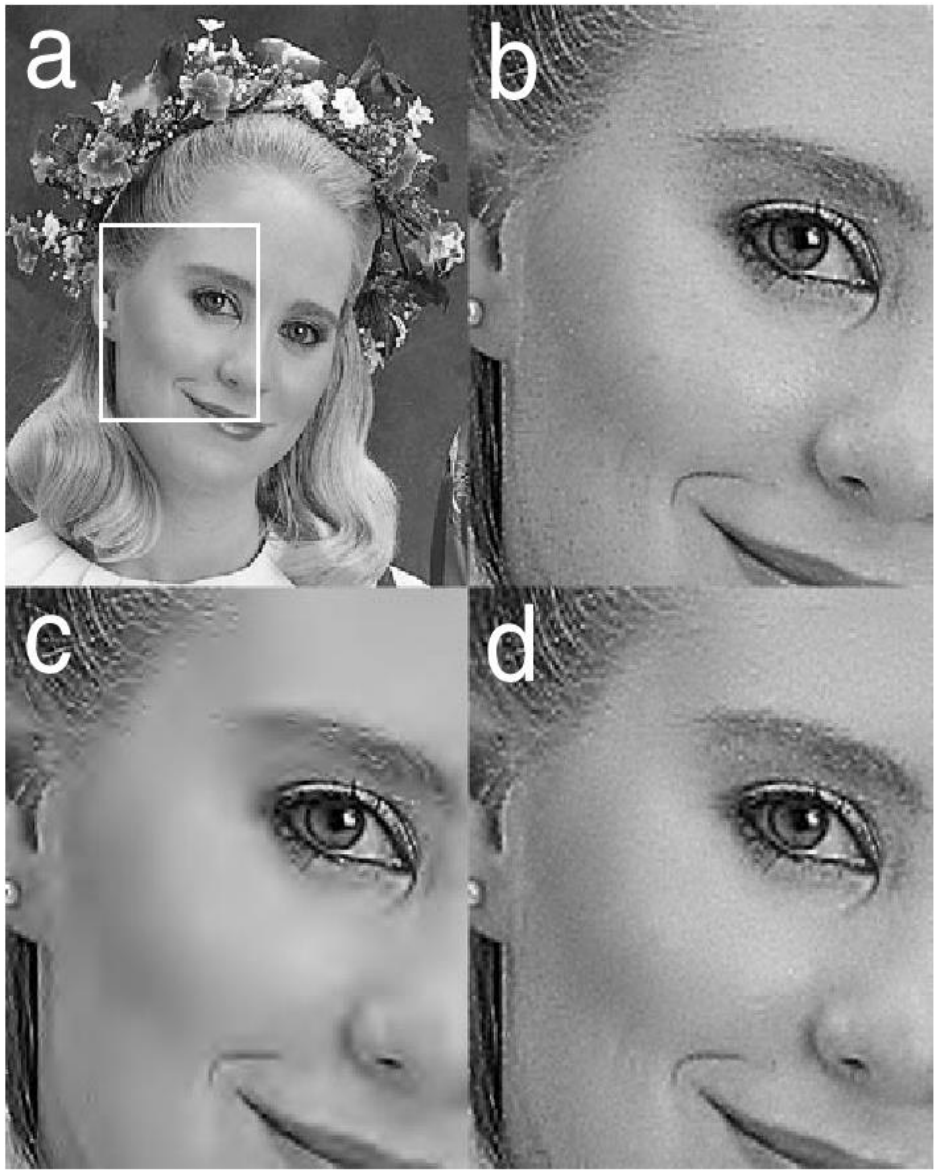
2 Perceivability of image distortions resulting from the quantization should be measured and controlled by a perceptual error metric.

JPEG 2000

a,b – original image,

c – standard JPEG 2000 algorithm controlled by a metric minimizing the MSE. The missing skin texture appears blurred and unnatural to the human observer. Exact reproduction of spatial detail, e.g., hair of the woman is less important due to visual masking by strong textures.

d – JPEG 2000 controlled by a perceptual image quality metric.



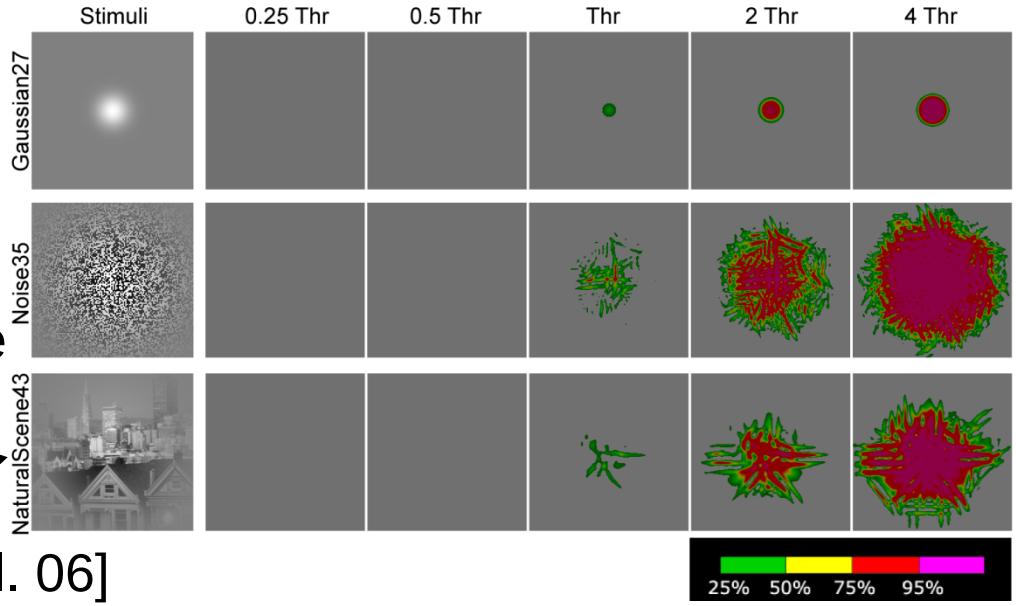
Evaluation of Image Quality Metrics

- Mostly only photos/real videos
- Focus on compression/transmission related artifacts
- Subjective responses: only overall quality (MOS)

Mean Opinion Score (MOS)		
MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Evaluation of Image Quality Metrics

- **Input images + Subjective responses = dataset**
- **Datasets**
  - Simpler evaluations
  - Reproducible evaluations
  - Should comprise typical artifacts
  - Should be publicly available
- **IMAGES**
  - Modelfest [Watson 99]
  - LIVE image db [Sheikh et al. 06]
  - TID (Tampere Image Database) [Ponomarenko et al. 09]
- **VIDEOS**
  - VQEG FRTV Phase 1 [VQEG ‘00]
  - LIVE video db [Seshadrinathan et al. 09]



Evaluation of Image Quality Metrics

Dataset	# of Reference Images	# of Distortion Types	Total # of Distorted Images
KADID [Lin et al. 2019]	81	25	10125
TID2013 [Ponomarenko et al. 2015]	25	24	3000
PIPAL [Jinjin et al. 2020]	200	40	23200
CSIQ [Larson and Chandler, 2010]	30	6	866

- **The earlier datasets (KADID, CSIQ, TID2013) have traditional artifacts such as:**
  - Gaussian noise
  - Gaussian blur
  - JPEG Artifacts
  - Contrast change
  - Color shifting
  - Brightness change

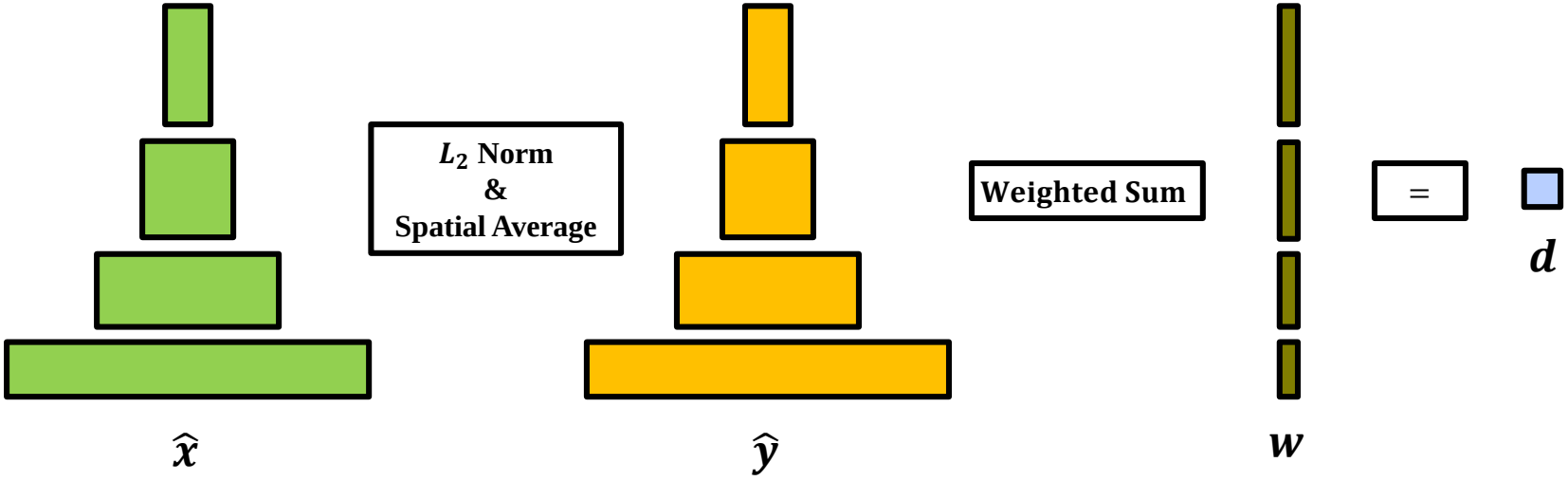
- **The newest dataset (PIPAL) additionally contains:**
  - Traditional Denoising
  - CNN-based Denoising
  - CNN-based Superresolution
  - GAN-based Superresolution

Feature-Based Metrics - LPIPS

- Computes the Mean Square Error (MSE) between the extracted VGG features $\hat{x}$ and $\hat{y}$:

$$d = \sum_{i=0}^{M} \sum_{j=0}^{N_i} w_{ij} \cdot MSE_{ij}(\hat{x}, \hat{y})$$

- where M denotes the number of VGG layers, N_i denote the number of channels in the layer i, and w_{ij} are learnable weights indicating the perceptual importance of each channel.

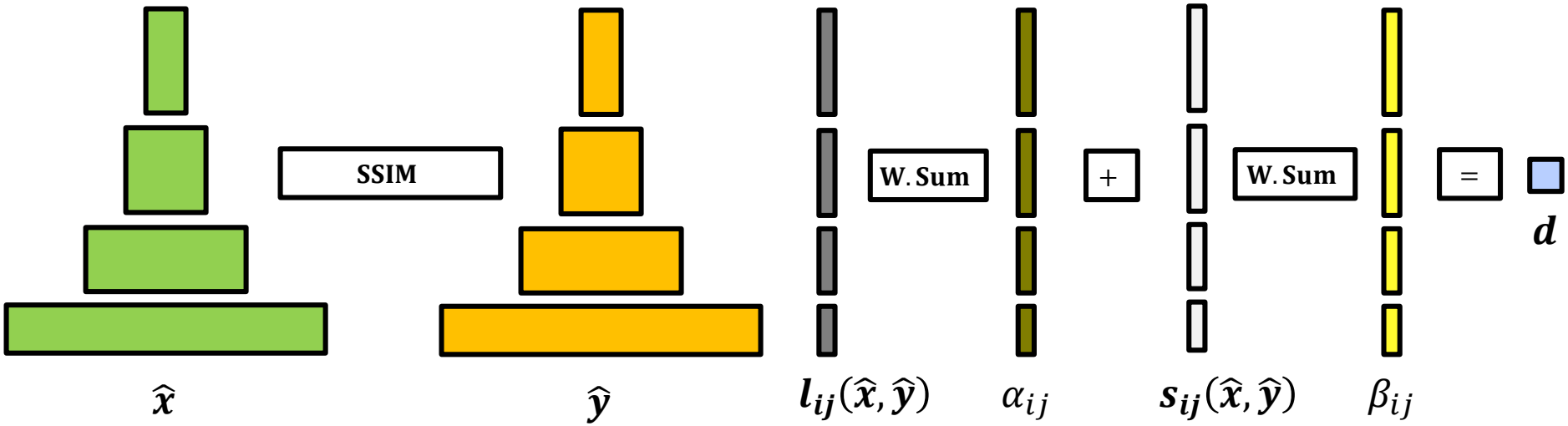


Feature-Based Metrics - DISTS

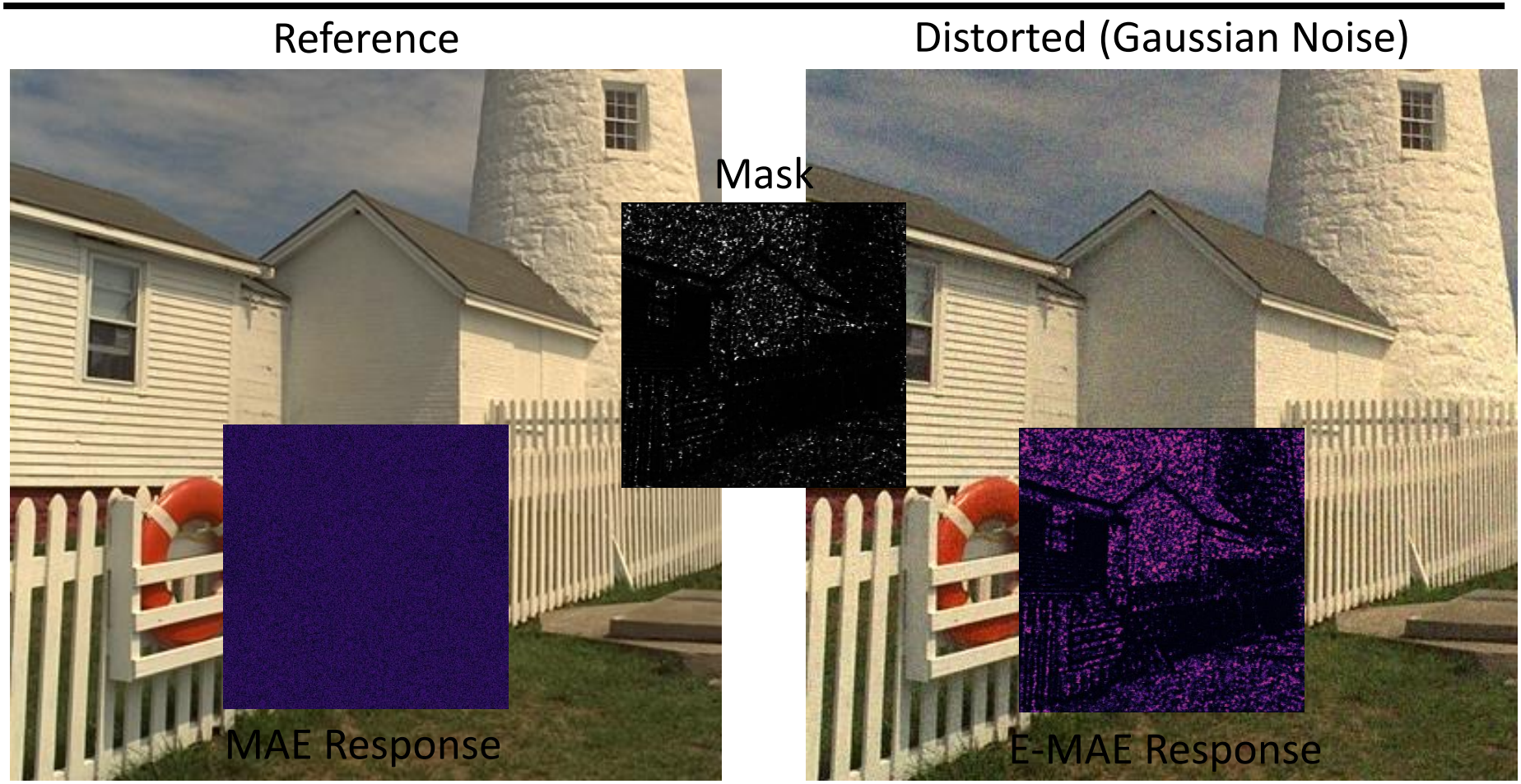
- Computes the global SSIM between the extracted VGG responses $\hat{x}$ and $\hat{y}$. The SSIM components $l(\hat{x},\hat{y})$ and $s(\hat{x},\hat{y})$ are computed for each VGG channel: $l(\hat{x},\hat{y}) = \frac{2\mu_{\hat{x}}\mu_{\hat{y}}+c_1}{\mu_{\hat{x}}^2+\mu_{\hat{y}}^2+c_1}$ and $s(\hat{x},\hat{y}) = \frac{2\sigma_{\hat{x}\hat{y}}+c_2}{\sigma_{\hat{x}}^2+\sigma_{\hat{y}}^2+c_2}$

$$d = 1 - \sum_{i=0}^{M}\sum_{j=0}^{N_i}[\alpha_{ij}\cdot l_{ij}(\hat{x},\hat{y}) + \beta_{ij}\cdot s_{ij}(\hat{x},\hat{y})]$$

- where *M* denotes the number of VGG layers, N_i denote the number of channels in the layer *i*, and α_{ij} and β_{ij} are positive learnable weights indicating the perceptual importance of each SSIM component.



Learning Visual Masking Model



Visual Masking Modeling

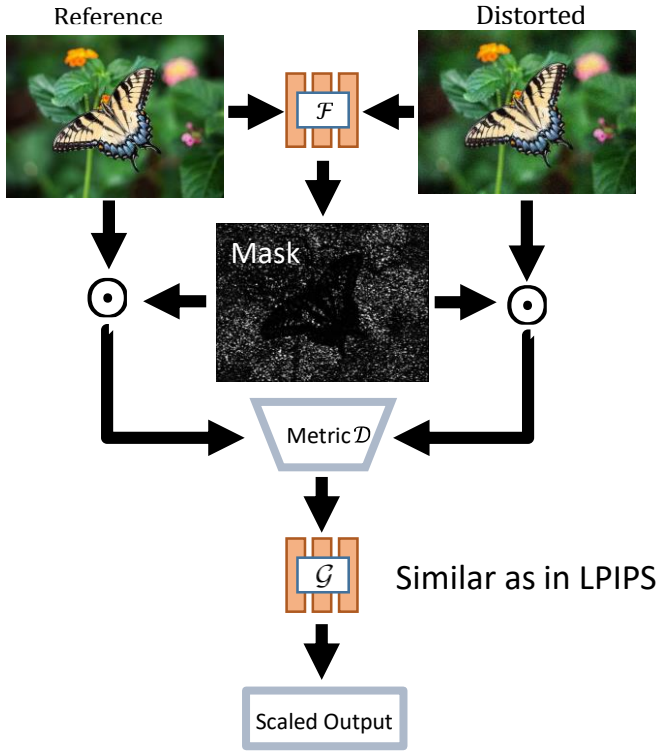
1. **Improving the perceptual prediction accuracy for existing full-reference image quality metrics**
2. **Learning visual masking model**
   - specific for each metric
   - adapting to all distortion types
3. **Using existing Mean Opionion Score (MOS) datasets for training**
   - MOS datasets provide just one value per-image
   - Desirable learning content-dependent (per-pixel) masking model

Enhancing Classic Metrics

- **Classic Metrics (Image Space)**
  1. MAE
     - L_1 Difference
  2. PSNR
     - Mean Squared Error
  3. SSIM
     - Luminance, Contrast, Structure
  4. MS-SSIM
     - Multiscale SSIM
  5. FLIP
     - Recent statistical metric

Enhancing Classic Metrics

- **Given images** $Reference$ **and** $Distorted$**:**



Similar as in LPIPS

Kadid Dataset

$$Loss = \|\mathcal{G}(\mathcal{D}(Mask \odot Ref, Mask \odot Dist)) - q\|_2^2$$

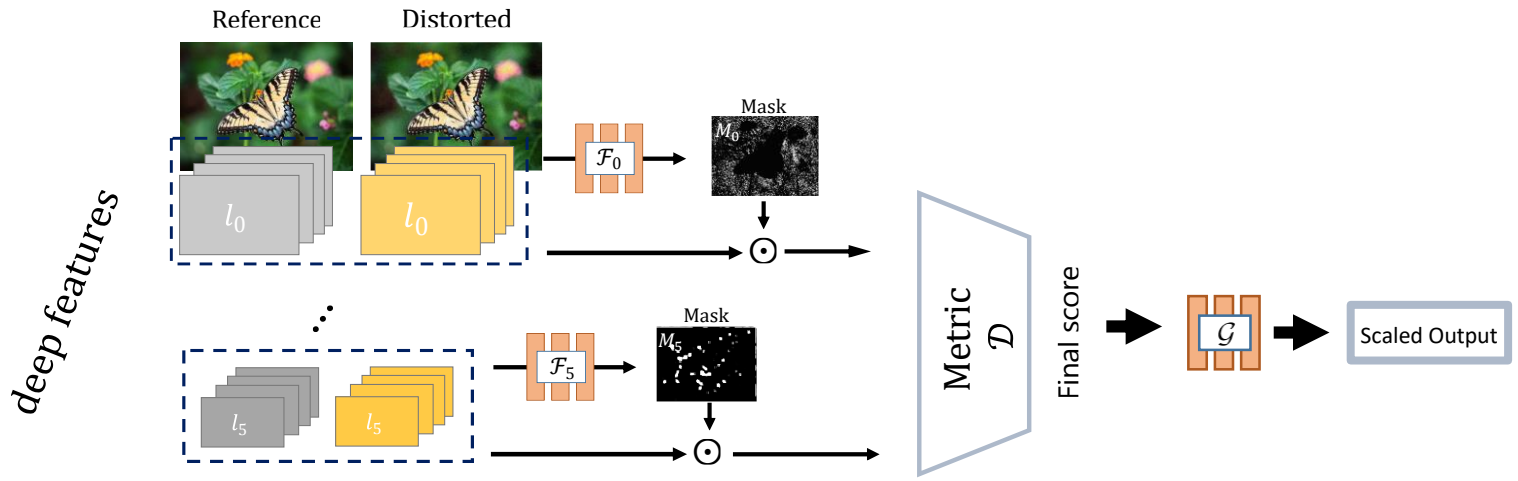
Enhancing Learning-based Metrics

- **Classic Metrics (Image Space)**
  1. MAE
     - L_1 Difference
  2. PSNR
     - Mean Squared Error
  3. SSIM
     - Luminance, Contrast, Structure
  4. MS-SSIM
     - Multiscale SSIM
  5. FLIP
     - Recent statistical metric

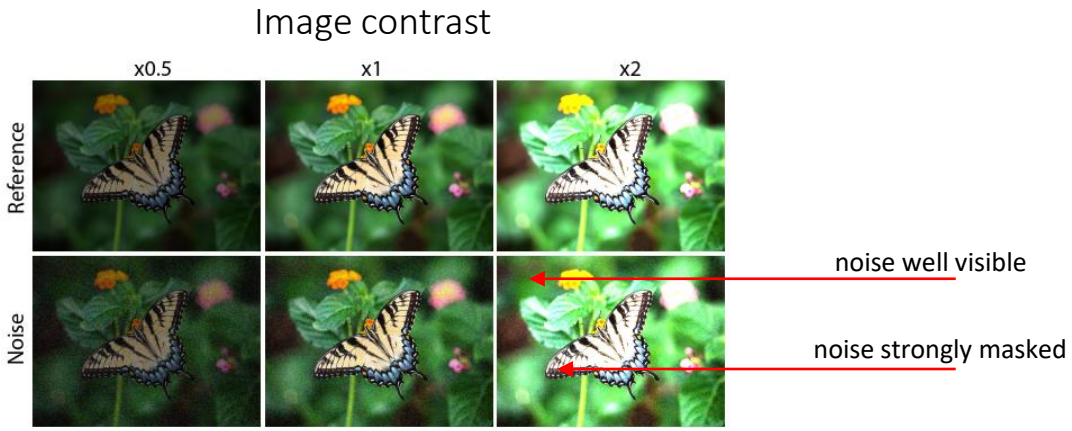
- **Learning-based Metrics (Feature Space)**
  1. VGG
     - L_1 Difference
  2. LPIPS
     - L_2 Difference
  3. DISTS
     - Luminance, Contrast, Structure
  4. DeepWSD
     - Wasserstein Distance

Enhancing Learning-based Metrics

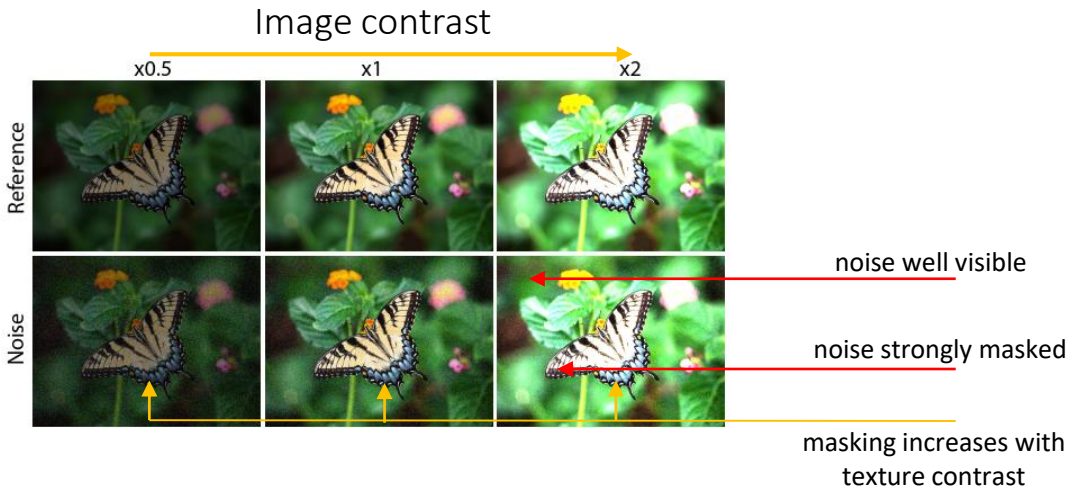
- Given images $Reference$ and $Distorted$ and their corresponding deep features:



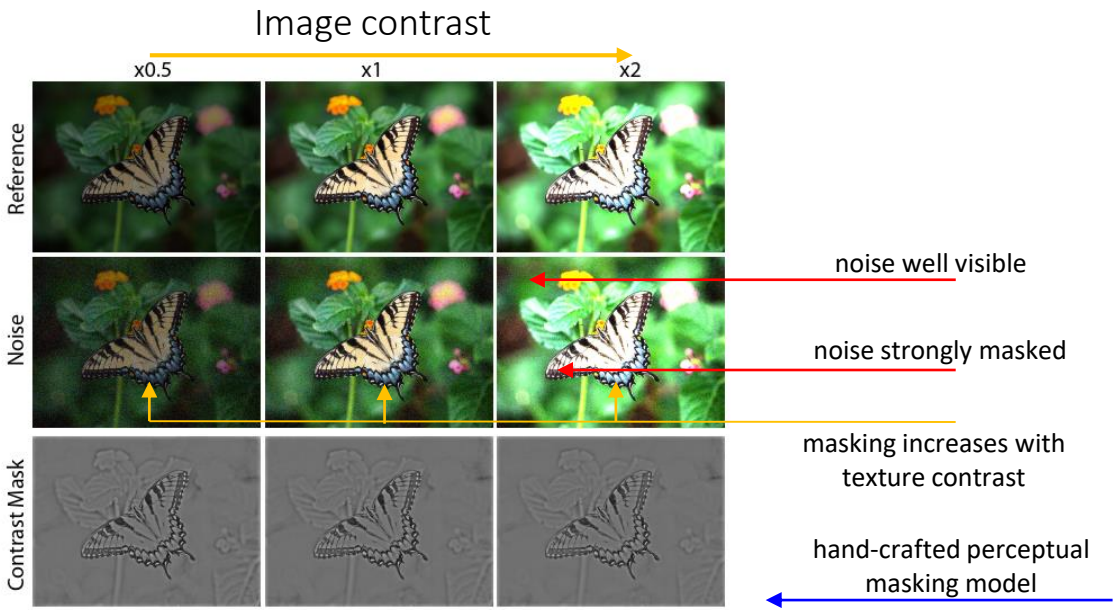
Classic vs. Learned Visual Masking



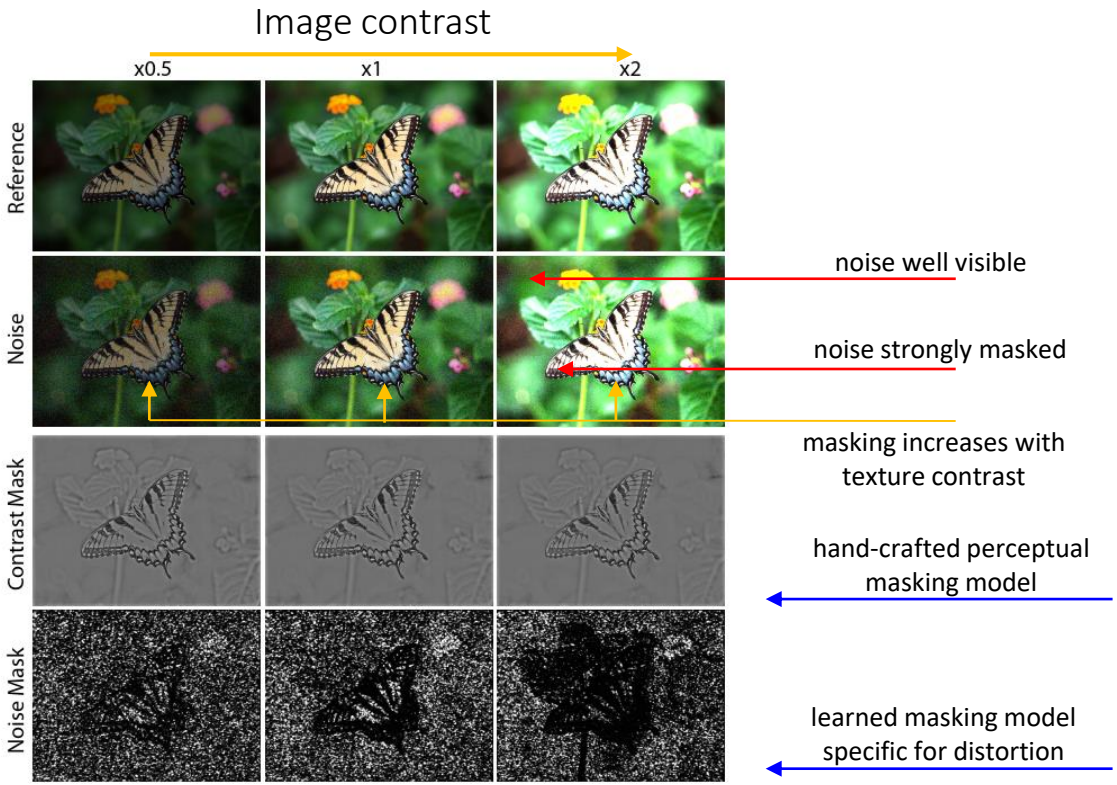
Perception: Visual Masking



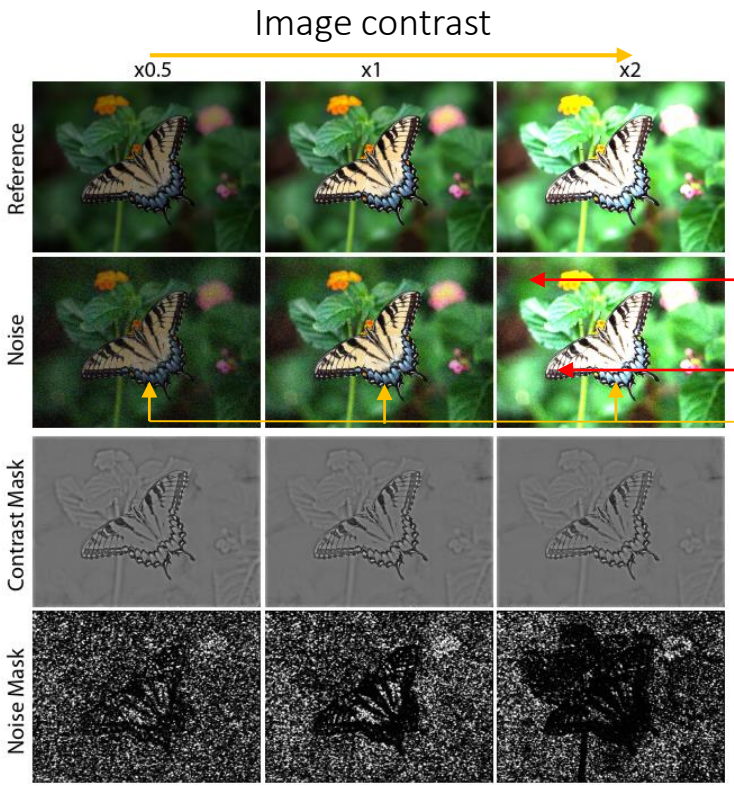
Perception: Visual Masking



Perception: Visual Masking



Perception: Visual Masking



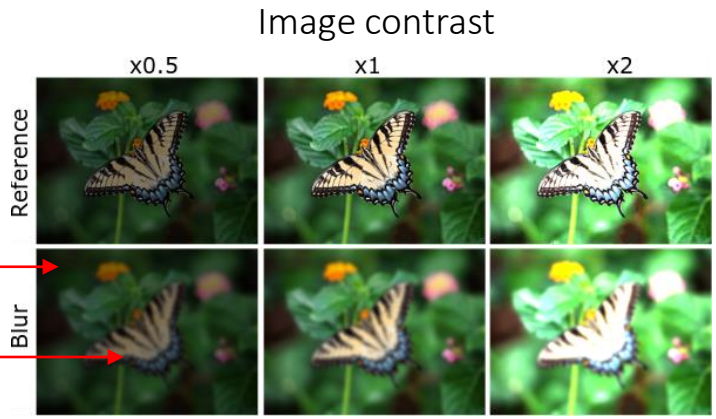
noise well visible
blur strongly masked

noise strongly masked
blur well visible

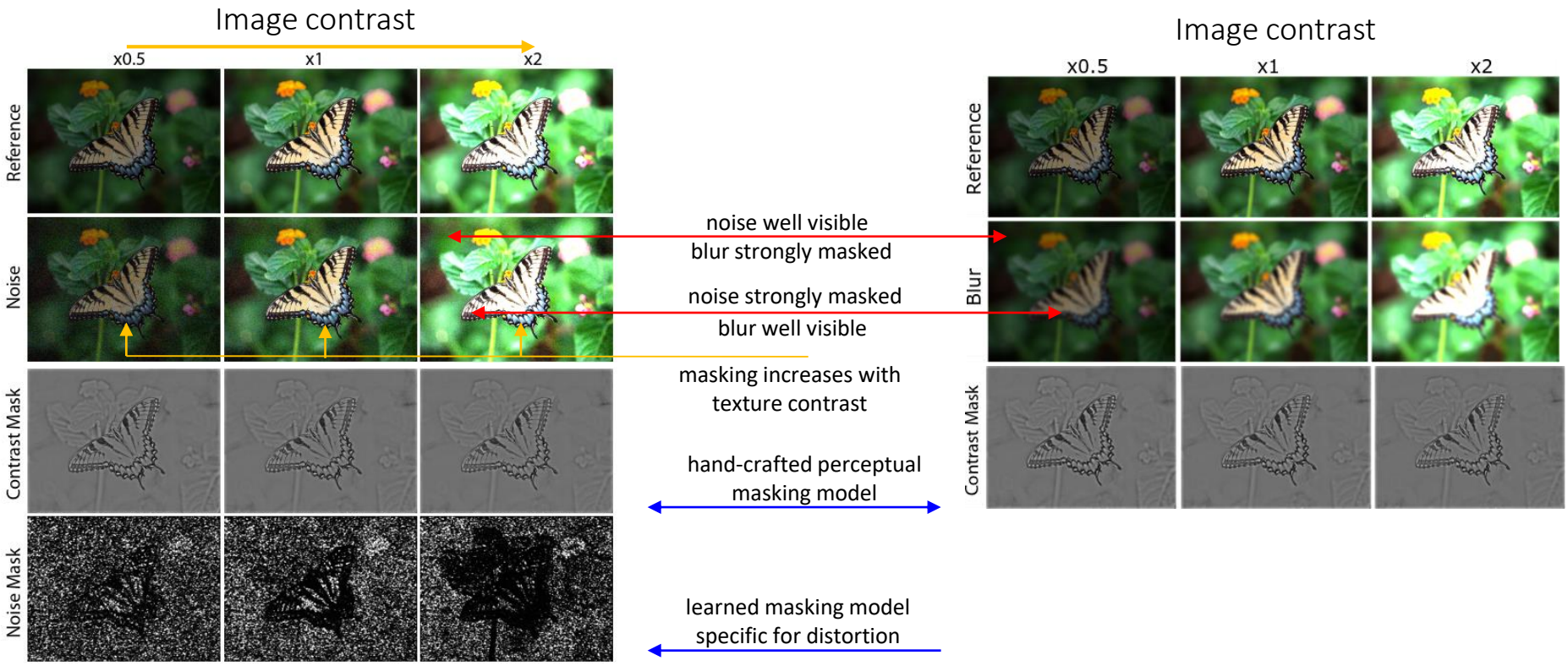
masking increases with
texture contrast

hand-crafted perceptual
masking model

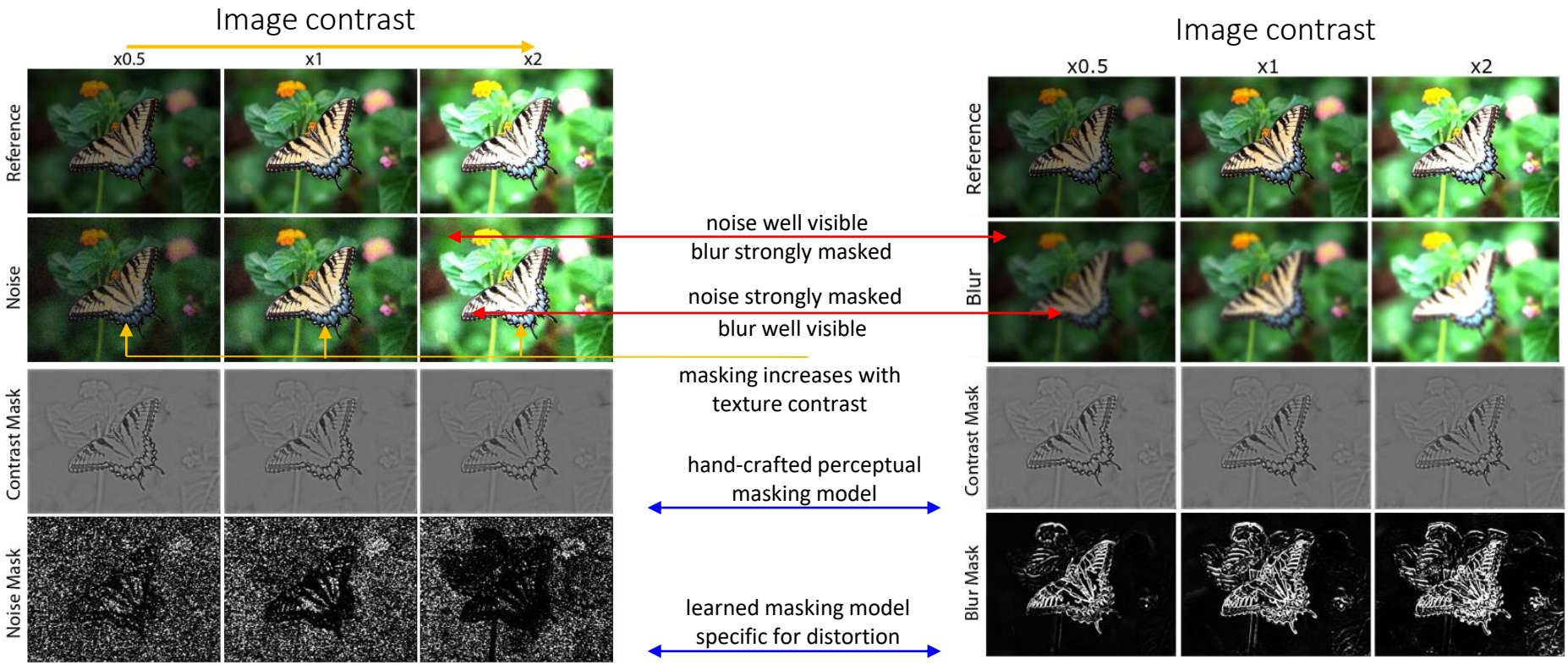
learned masking model
specific for distortion



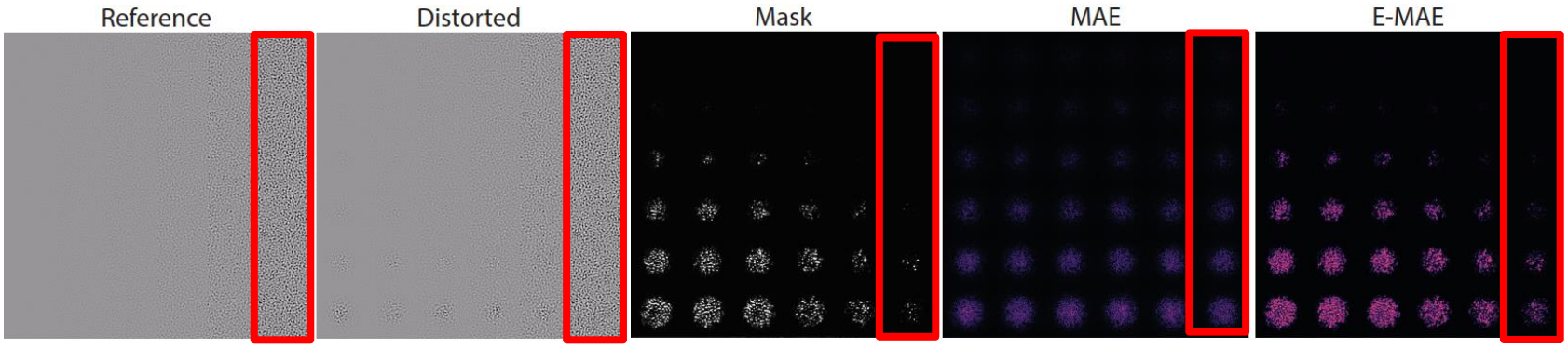
Perception: Visual Masking



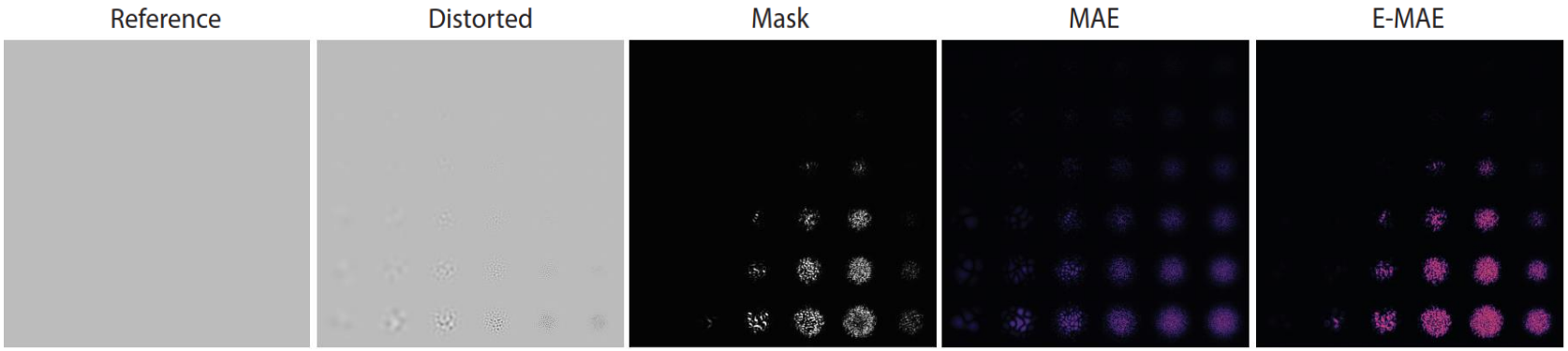
Perception: Visual Masking



Masking by Background of Increasing Contrast

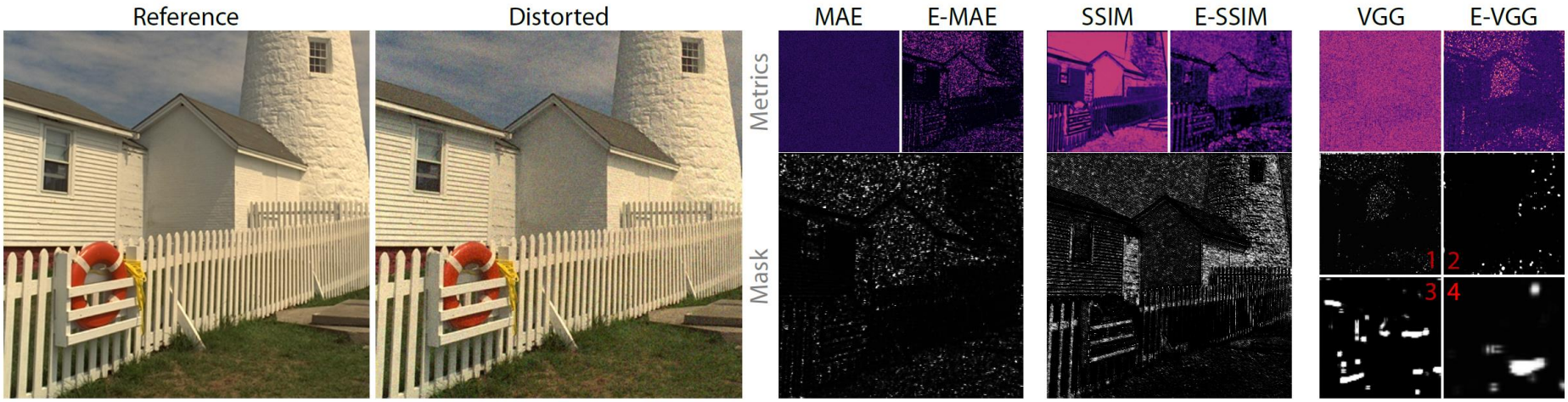


Contrast Sensitivity Function Reproduction



Visual Masking Comparison for Different Metrics

- **Enhanced standard image quality metrics with learned masking model better predict noise visibility by the human observer**



Quantitative Results

PLCC: Pearson Linear Correlation Coefficient
- Linear correlation

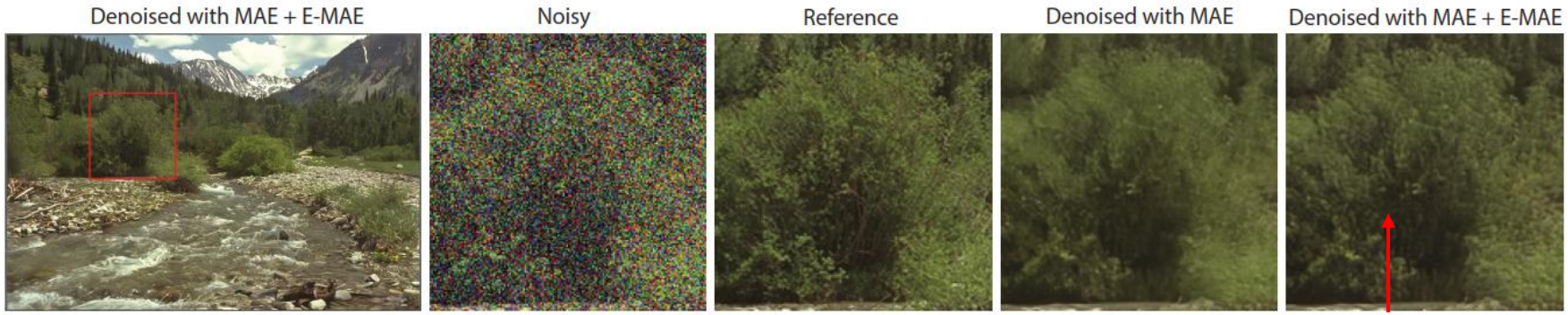
SRCC: Spearman's Rank Correlation Coefficient
- Monotonicity

KRCC: Kendall Rank Correlation Coefficient
- Ordinal association

		TID			PIPAL			CSIQ	
	Metric	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
Classic Metrics	MAE	0.639	0.627	0.409	0.458	0.443	0.304	0.819	0.801	0.599
	E-MAE	0.857	0.863	0.673	0.597	0.606	0.429	0.871	0.917	0.738
	PSNR	0.726	0.714	0.540	0.468	0.456	0.314	0.851	0.837	0.645
	E-PSNR	0.855	0.844	0.656	0.637	0.629	0.446	0.901	0.910	0.728
	SSIM	0.697	0.663	0.479	0.550	0.534	0.373	0.848	0.863	0.665
	E-SSIM	0.842	0.868	0.677	0.671	0.656	0.469	0.869	0.910	0.732
	MS-SSIM	0.820	0.813	0.616	0.584	0.538	0.379	0.826	0.841	0.642
	E-MS-SSIM	0.806	0.825	0.621	0.642	0.634	0.453	0.862	0.895	0.709
Learning-Based Metrics	FLIP	0.591	0.537	0.413	0.498	0.442	0.306	0.731	0.724	0.527
	E-FLIP	0.859	0.858	0.666	0.621	0.612	0.434	0.871	0.902	0.715
	VGG	0.853	0.820	0.639	0.643	0.610	0.432	0.938	0.952	0.804
	E-VGG	0.895	0.889	0.710	0.695	0.675	0.485	0.914	0.938	0.776
	LPIPS	0.803	0.756	0.568	0.640	0.598	0.424	0.944	0.929	0.769
	E-LPIPS	0.884	0.876	0.689	0.705	0.678	0.490	0.922	0.933	0.771
	DISTS	0.839	0.811	0.619	0.645	0.626	0.445	0.947	0.947	0.796
	E-DISTS	**0.903**	**0.915**	**0.725**	**0.725**	**0.697**	**0.507**	0.932	0.925	0.753
	DeepWSD	0.879	0.861	0.674	0.593	0.584	0.409	**0.949**	**0.961**	**0.821**
	E-DeepWSD	0.905	0.892	0.710	0.704	0.672	0.485	0.937	0.937	0.775

Denoising: Loss Driven by Visual Masking

- Restormer [Zamir et al. 2021] denoising guided by E-MAE produces better visual quality
  - less blurred details in high-contrast textures



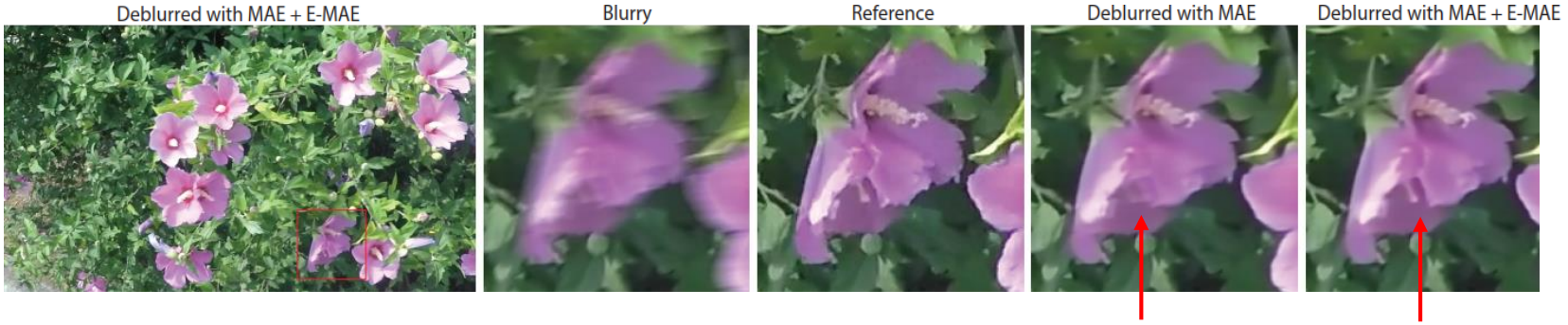
- Such visual quality improvement is predicted by the LPIPS metric, but NOT by PSNR scores.

Noise magnitude →

	σ = 15				σ = 25				σ = 50				σ = 60			
Loss	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓
MAE	34.36	**0.94**	0.058	0.0343	**31.94**	0.90	0.092	0.0849	**28.82**	**0.84**	0.163	3.187	**28.02**	**0.81**	0.182	4.258
MAE + E-MAE	**34.37**	**0.94**	**0.055**	**0.0145**	31.92	**0.91**	**0.087**	**0.0263**	28.71	**0.84**	**0.152**	**0.790**	27.88	**0.81**	**0.167**	**1.035**

Deblurring: Loss Driven by Visual Masking

- Restormer [Zamir et al. 2021] deblurring guided by E-MAE produces better visual quality
  - better deblurring in high-contrast textures



- Such visual quality improvement is agreed by PSNR, SSIM, and LPIPS metrics

Metric	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓
MAE	31.70	0.92	0.1030	0.0192
MAE + E-MAE	**31.78**	**0.93**	**0.1018**	**0.0184**