ACM multimedia
Chengdu, China OCT 20-24 2021

Part III: Evaluation of IQA Models

An Analysis-by-Synthesis Approach

Kede Ma

Department of Computer Science
香港城市大學
City University of Hong Kong

Outline

- Standard Approach and Its Caveats
- MAximum Differentiation (MAD) Competition [Wang and Simoncelli, 2008]
  - Group MAximum Differentiation (gMAD) Competition [Ma et al., 2016, 2020]
  - MAximum Discrepancy (MAD) Competition for Visual Recognition
- Comparison of IQA Models for Optimization of Image Processing Systems
- Eigen-Distortion Analysis of Perceptual Representations
- Discussion

Standard Approach

for Evaluating IQA Models

Standard Approach

Main Steps

1. Select a set of images from *the image domain of interest*

2. Collect the MOS for each image via psychophysical experiments (i.e., subjective user studies)

3. Compare the goodness of fit among the competing IQA models (i.e., sort by **average** performance)

    - Spearman rank correlation coefficient - prediction monotonicity

    - Pearson linear correlation coefficient - prediction linearity

    - Mean squared error - prediction accuracy

$$\mathrm{SRCC} = 1 - \frac{6\sum_i d_i^2}{M(M^2-1)}$$

$$\mathrm{PLCC}(x, y) = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}}$$

$$\mathrm{MSE}(x, y) = \frac{1}{M}\sum_i (x_i - y_i)^2$$

Caveats

- *Sampling bias* due to the extremely sparse distribution of the selected samples in the image space
  - I.e., the curse of dimensionality
- *Algorithmic bias* due to potentially overfitting the selected samples
  - The dataset creation precedes the algorithm development
- *Subjective bias* due to potentially cherry-picking test results

A Detour

Debiased Subjective Assessment of Real-World Image Enhancement [Cao et al., 2021]

MAximum Differentiation (MAD) Competition

for Evaluating IQA Models

MAD Competition

[Wang and Simoncelli, 2008]

- A methodology for comparing computational models of perceptual quantities
- Inspired by "analysis by synthesis," a core idea in the Pattern Theory by Ulf Grenander
- Main idea: Efficiently and automatically selecting stimuli (e.g., images) that are likely to falsify the computational model in question
- Originally demonstrated using two perceptual quantities: contrast and image quality

Another Detour

Pattern Theory [Grenander, 1970, Mumford, 1994]

- Definition: The analysis of the patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of **reconstructing** the processes, objects and events that produced them
- Plain English: If one wants to test whether a computational method relies on *intended* features for a specific task, the set of features should be tested in a *generative* (not a *discriminative*) way
- Well demonstrated in the context of texture analysis [Julesz, 1962]
    - Texture discrimination vs texture synthesis

MAD Competition

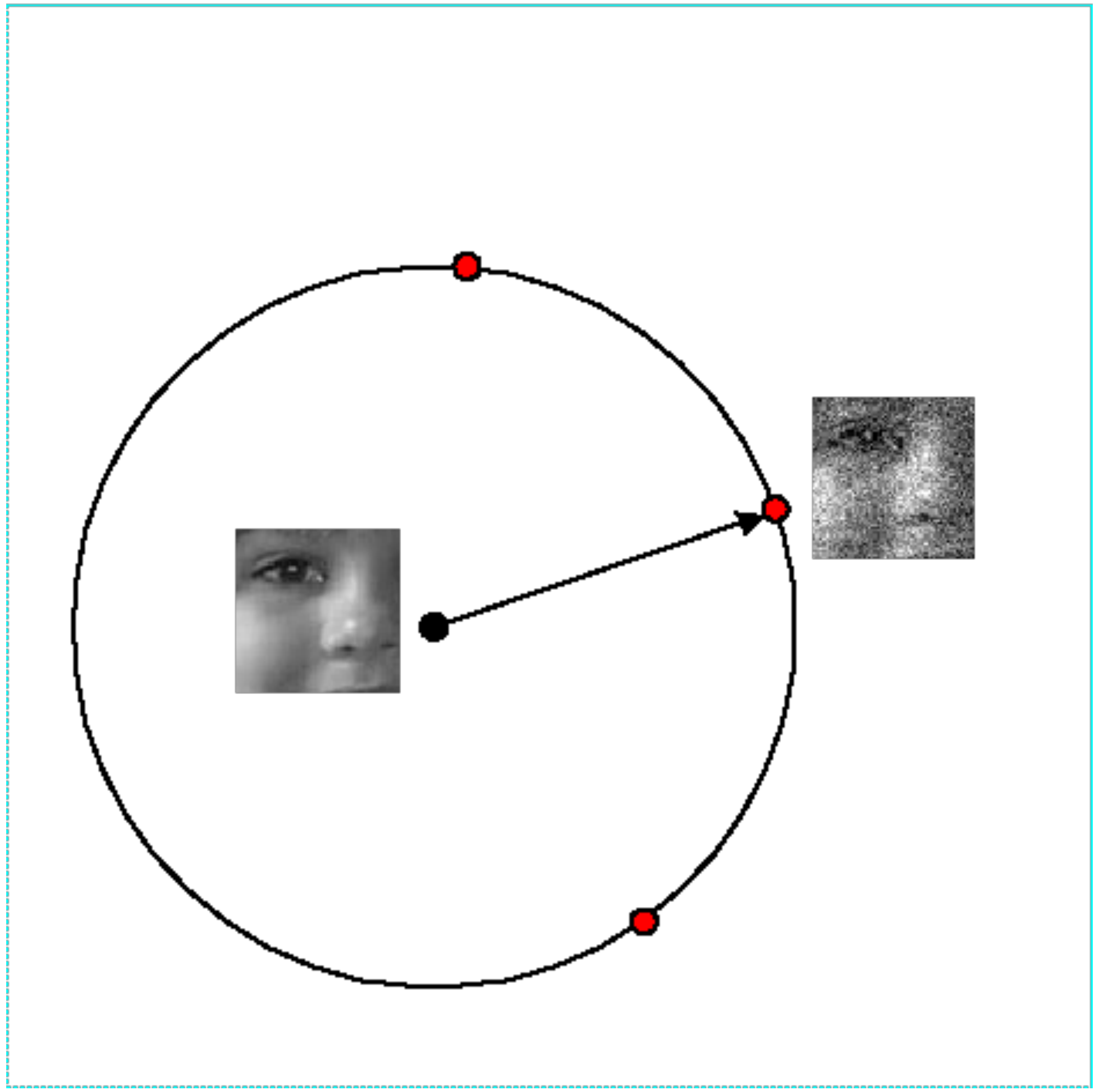


Image Credit: Wang

MAD Competition

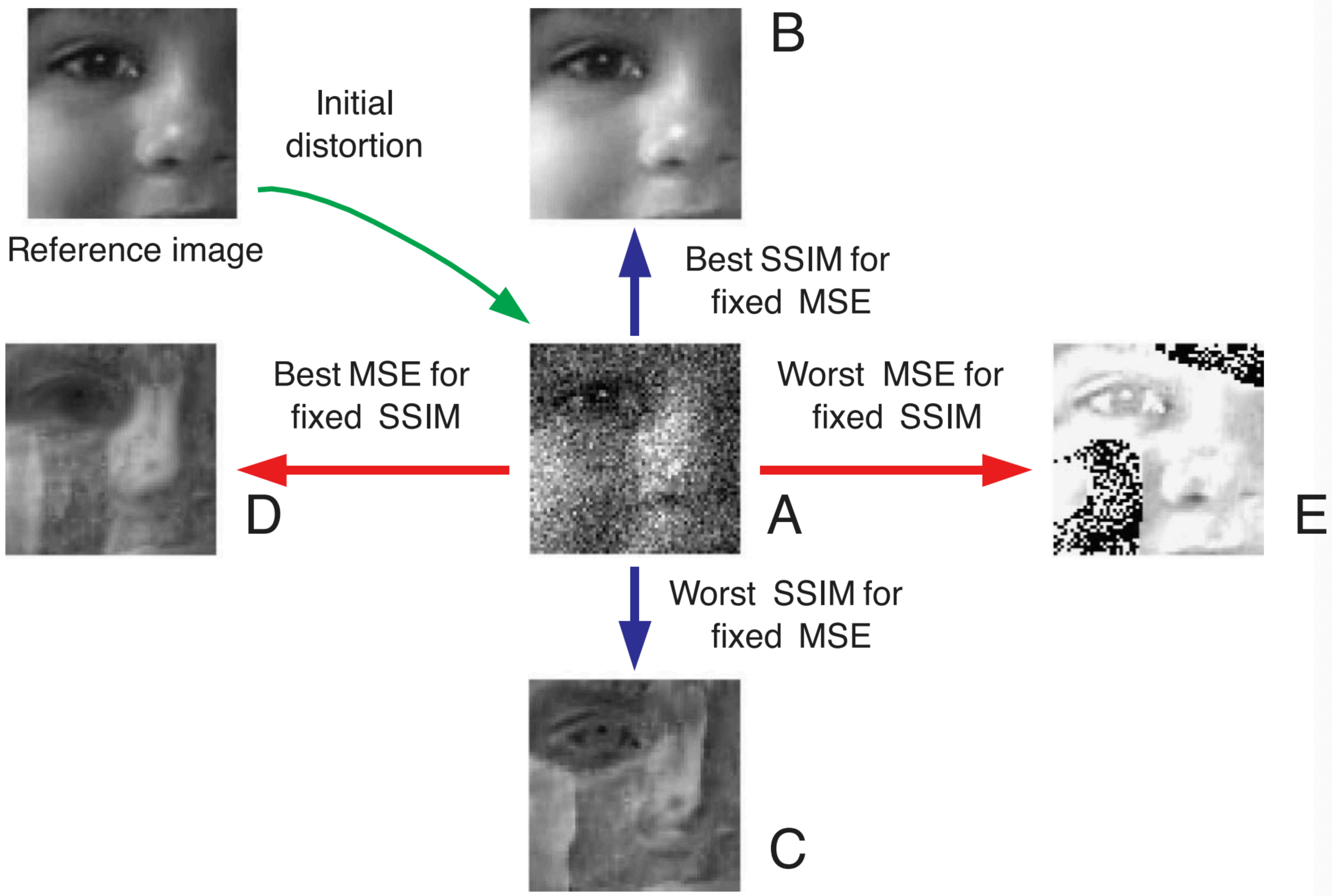


Image Credit: Wang

MAD Competition

Math Formulation

$$(x^\star, y^\star) = \mathrm{argmax}_{x,y}\, f_1(x) - f_1(y)$$

$$\text{subject to} \quad f_2(x) = f_2(y) = \alpha$$

- f_i for $i \in \{1,2\}$ represents an IQA model (with a larger value indicating higher predicted quality)
  - f_1 and f_2 can be treated as “attacker” and “defender,” respectively
  - The roles of f_1 and f_2 should be switched

Connection to Adversarial Perturbations in Classification

MAD Competition:

$$(x^\star, y^\star) = \text{argmax}_{x,y} f_1(x) - f_1(y)$$

$$\text{subject to} \quad f_2(x) = f_2(y) = \alpha$$

Adversarial Perturbations:

$$x^\star = \text{argmax}_x \, \text{logit}_t(x) - \text{logit}_p(x)$$

$$\text{subject to} \quad \ell_\infty(x, x_{\text{init}}) \leq \alpha$$

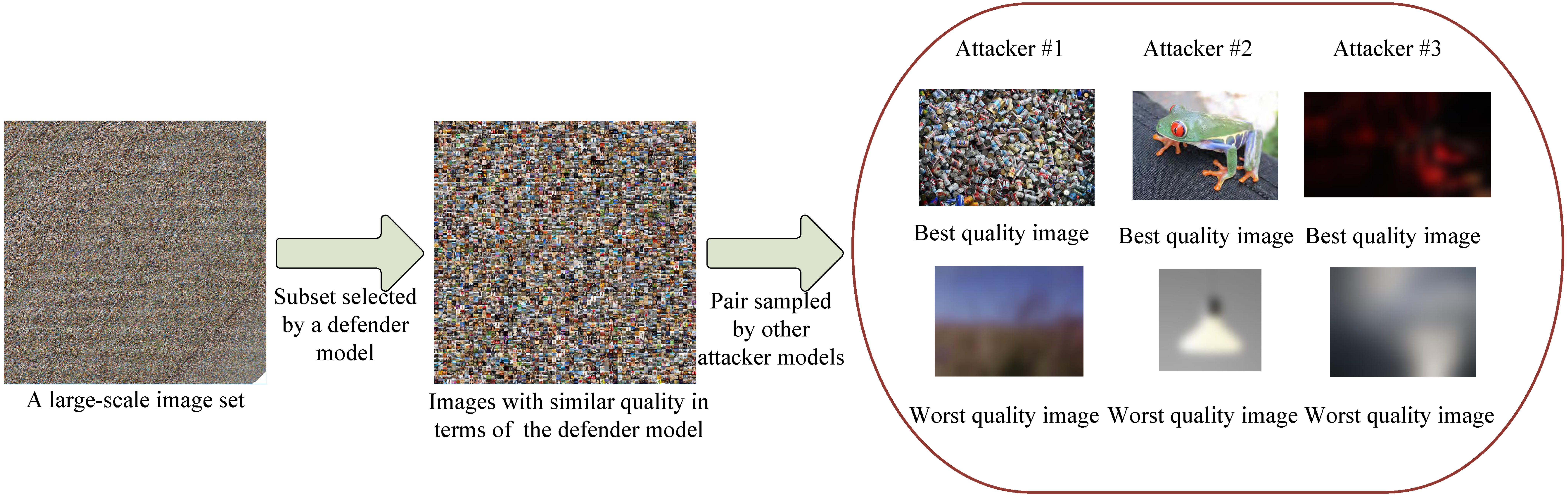
- Here we consider *targeted* adversarial attack
- MAD competition is constrained at the α-level set of f_2
- Adversarial attack is constrained within the ℓ_∞-ball centered at the initial point

Limitations of MAD Competition

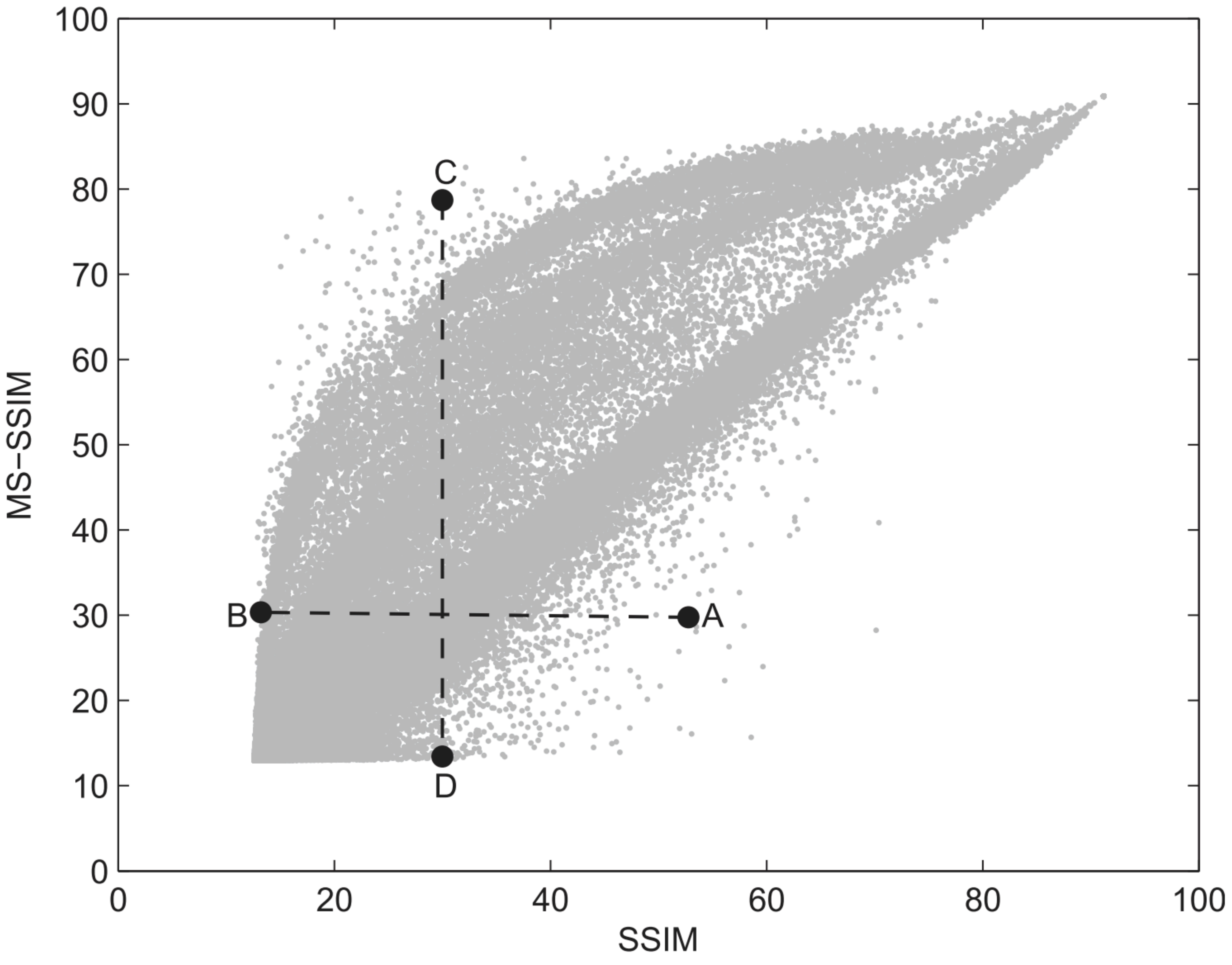
- Require solving constrained optimization problems by projected gradient ascent/decent algorithms
  - Computationally costly
  - Stuck in bad local maxima/minima
- MAD-generated stimuli may be highly unnatural
  - Of less practical relevance

Group MAD (gMAD) Competition
[Ma et al., 2016, 2020]

- A discrete instantiation of MAD competition for comparing multiple models



gMAD Competition
Scatter Plot



gMAD Competition

Pairwise Comparison to Global Ranking

$$\text{argmax}_{\mu} \sum_{ij} a_{ij} \log \left(\Phi(\mu_i - \mu_j) \right)$$

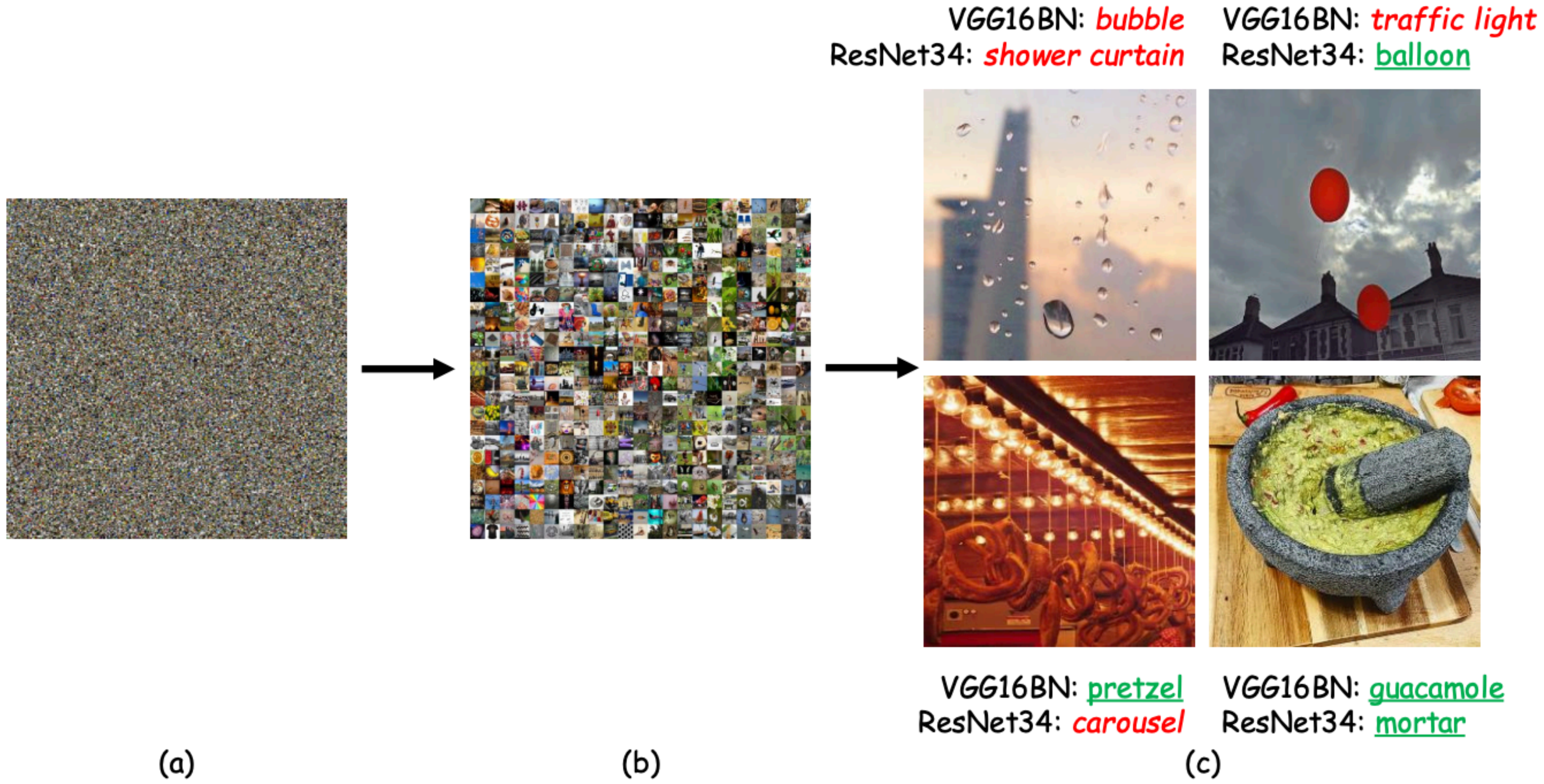
$$\text{s.t.} \sum_{i} \mu_i = 0$$

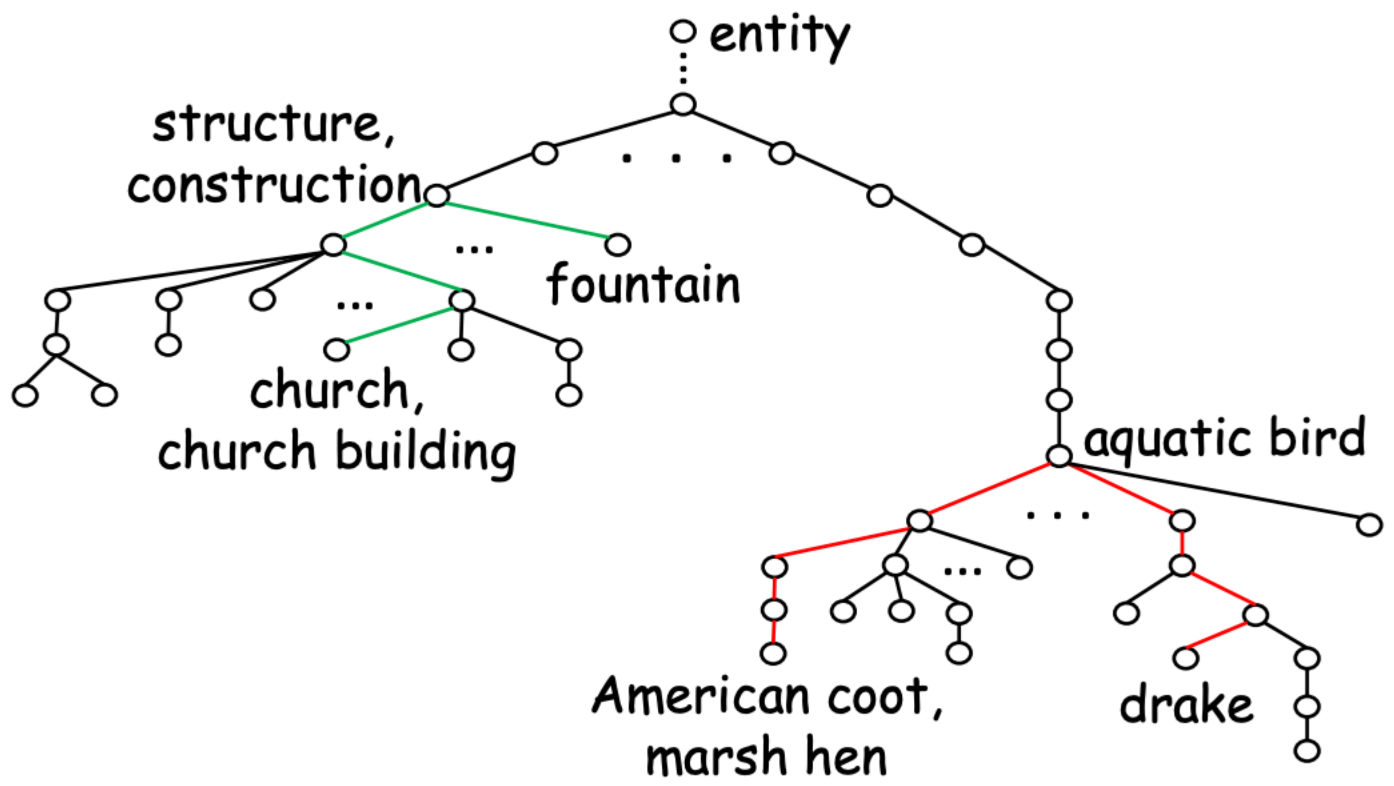
gMAD Competition
Visual Result

(a) (b) (c) (d)

Another Detour

MAximum Discrepancy (MAD) Competition for Image Classification [Wang et al. 2021]





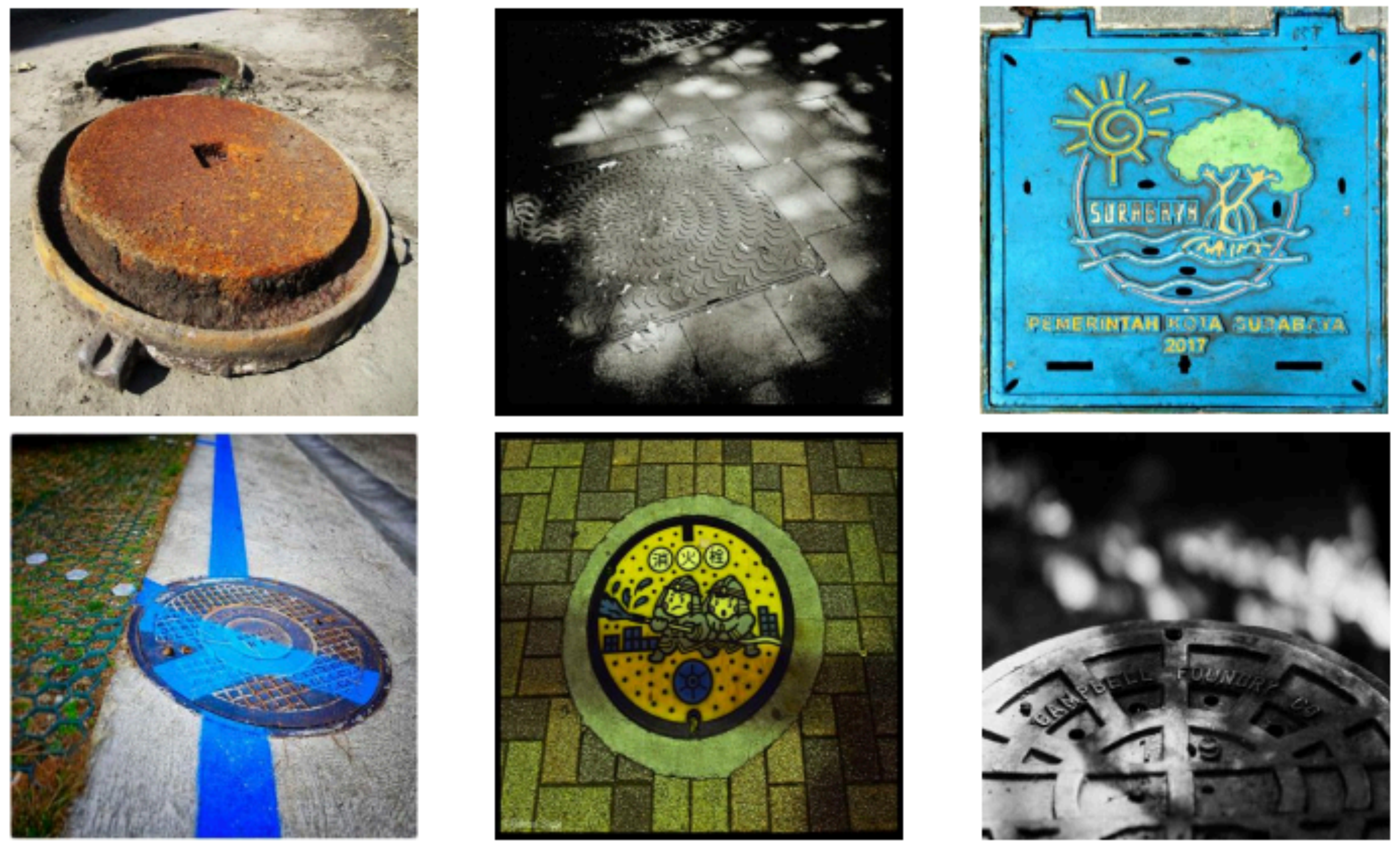
MAD Competition for Image Classification

Visual Comparison

ResNet34: Dutch oven
EfficientNet-B7: manhole cover

ResNet34: spider web
EfficientNet-B7: manhole cover

ResNet34: mailbox, letter box
EfficientNet-B7: manhole cover



ResNet101: sundial
NASNet-A-Large: manhole cover

ResNet101: doormat
NASNet-A-Large: manhole cover

ResNet101: sundial
NASNet-A-Large: barbell

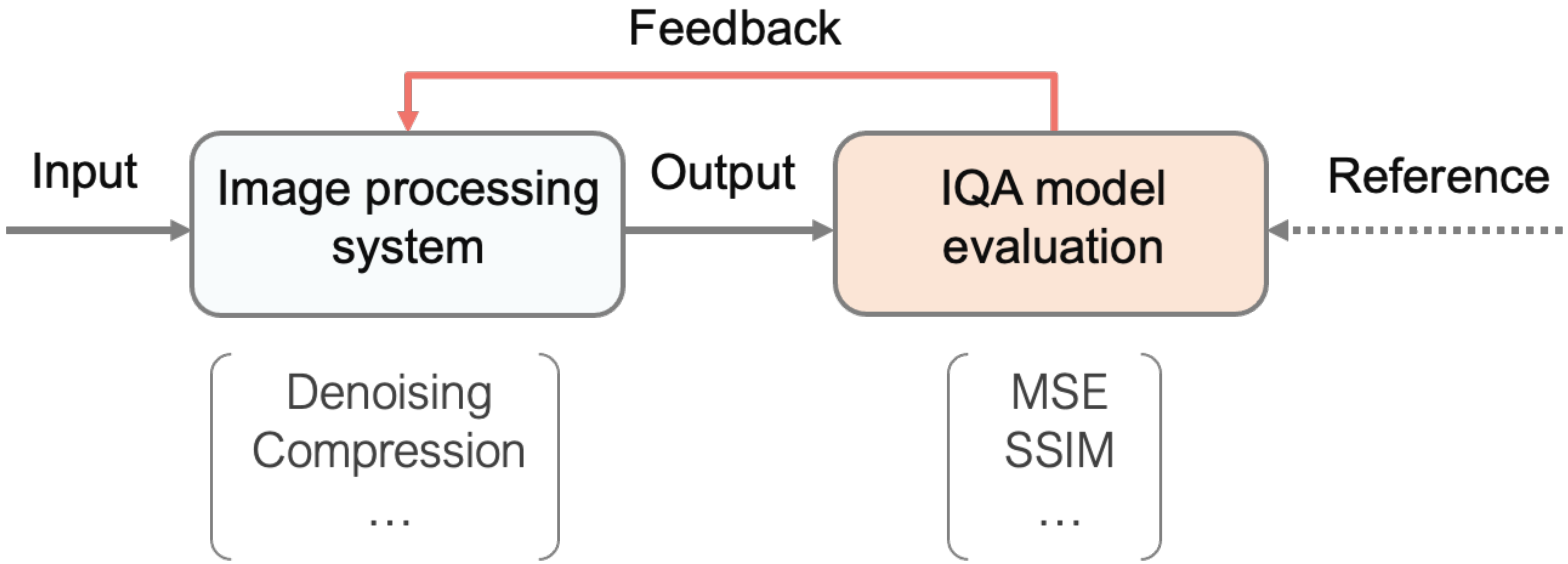
(a)

(b)

Comparison of IQA Models for Optimization of Image Processing Systems

Diagram of IQA-based Optimization

- A highly promising application of IQA models is to use them as objectives for the design and optimization of new image processing algorithms

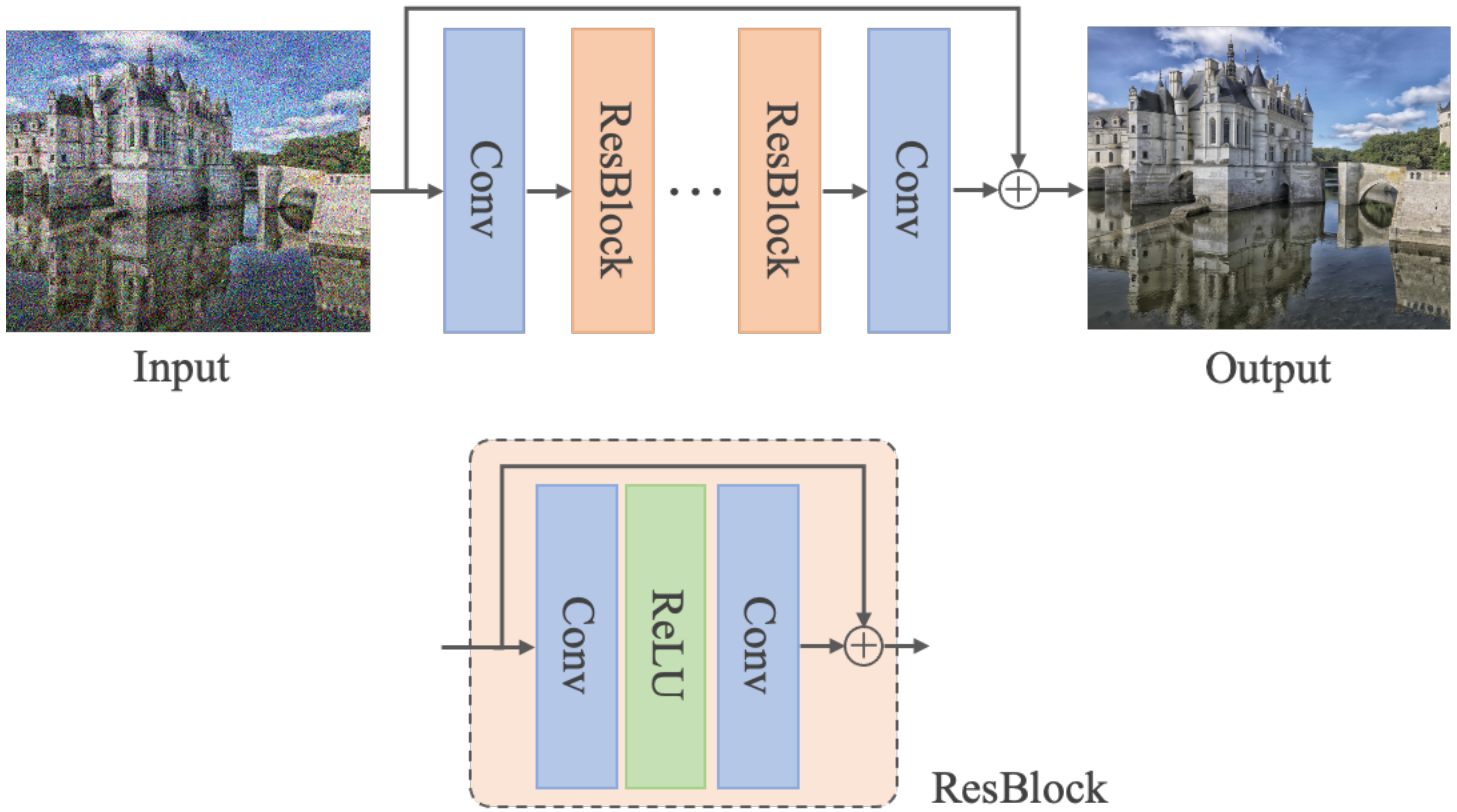


A Comprehensive Benchmark
[Ding et al., 2020]

- Evelen IQA models
  - MAE, MS-SSIM, VIF, CW-SSIM, MAD, FSIM, GMSD, VSI, NLPD, LPIPS, DISTS
- Four low-level vision tasks
  - Image denoising
  - Blind image deblurring
  - Single image super-resolution
  - Lossy image compression

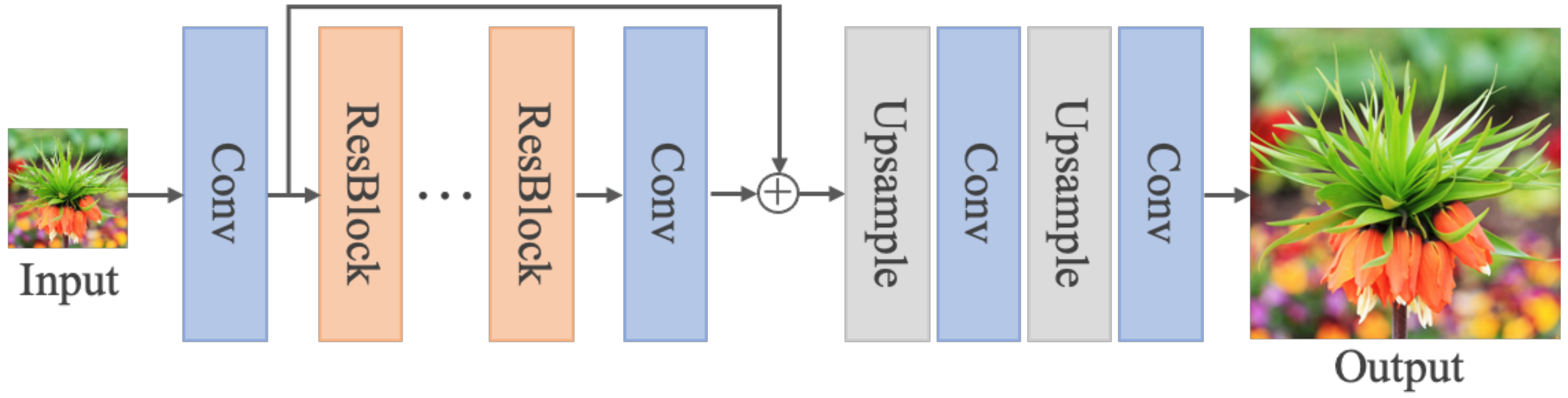
A Comprehensive Benchmark

- Network architecture for denoising and deblurring

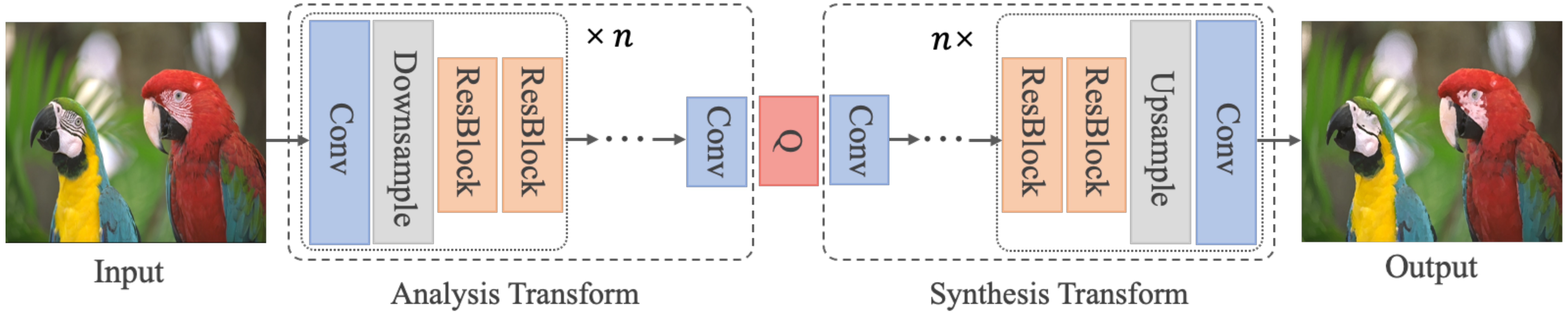


A Comprehensive Benchmark

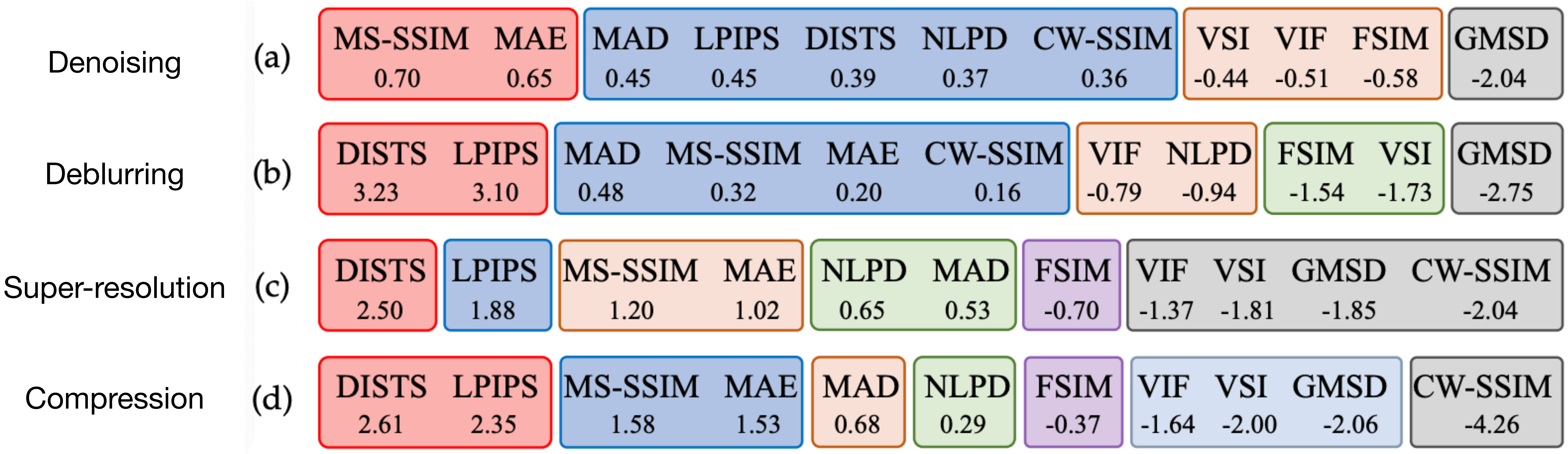
- Network architecture for super-resolution:



- Network architecture for compression:

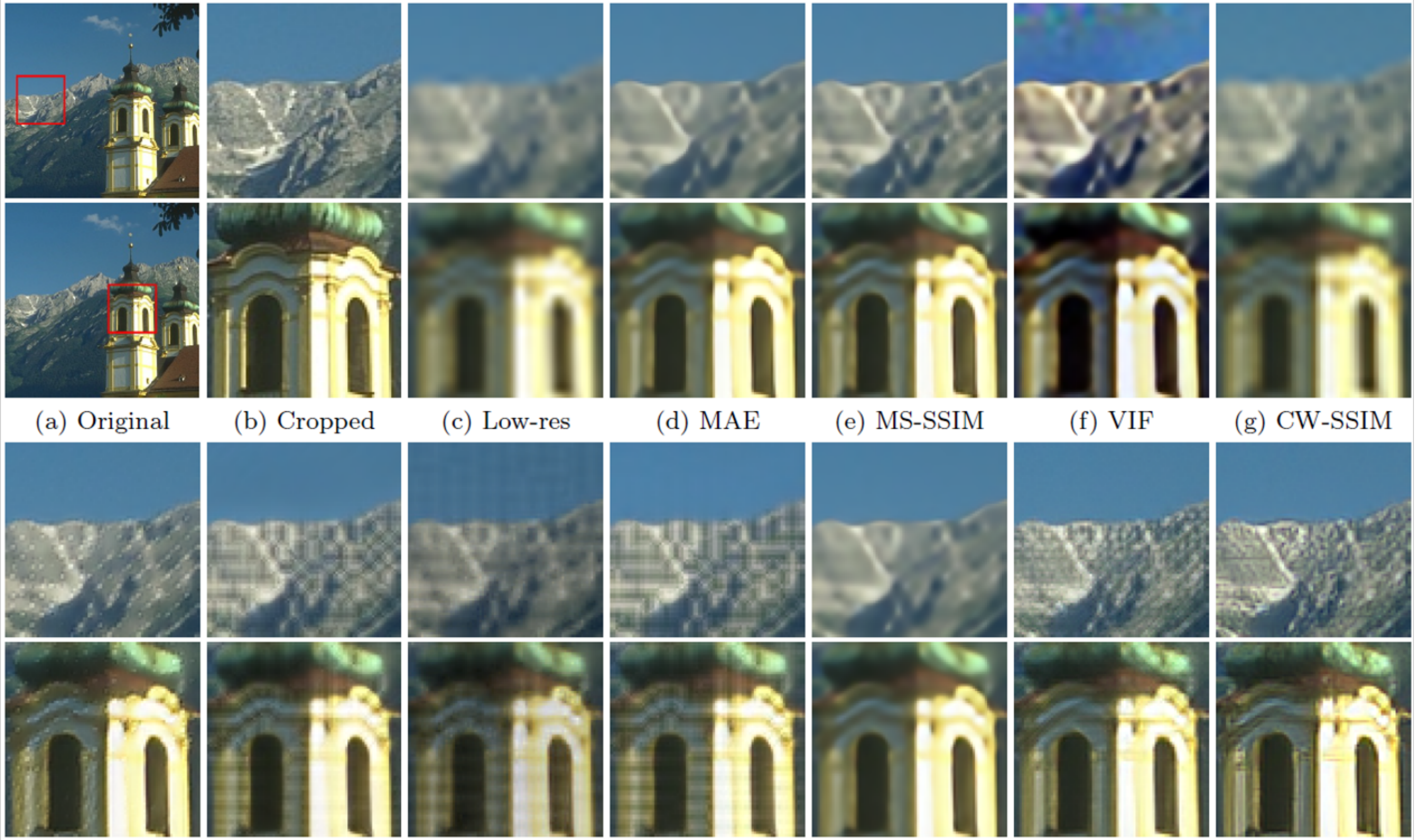


A Comprehensive Benchmark
Subjective Result



A Comprehensive Benchmark

Visual Result of Super-resolution



(a) Original (b) Cropped (c) Low-res (d) MAE (e) MS-SSIM (f) VIF (g) CW-SSIM

(h) MAD (i) FSIM (j) GMSD (k) VSI (l) NLPD (m) LPIPS (n) DISTS

Eigen-Distortion Analysis of Perceptual Representations

Eigen-Distortion Analysis of Image Representations

[Berardino et al., 2018]

- A computational method for comparing image representations when explaining perceptual sensitivity in humans

- Use Fisher information to predict model sensitivity to local image perturbations

$$J(\mathbf{x}) = \frac{\partial(\mathbf{f}(\mathbf{x}))^T}{\partial \mathbf{x}} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$$

- Compute the eigenvectors of the Fisher information matrix with largest and smallest eigenvalues
  - Correspond to the model-predicted most- and least-noticeable distortion directions

Eigen-Distortion Analysis of Image Representations

- Ratio of thresholds for model-generated extremal distortions will be larger for models that are more similar to the human subjects

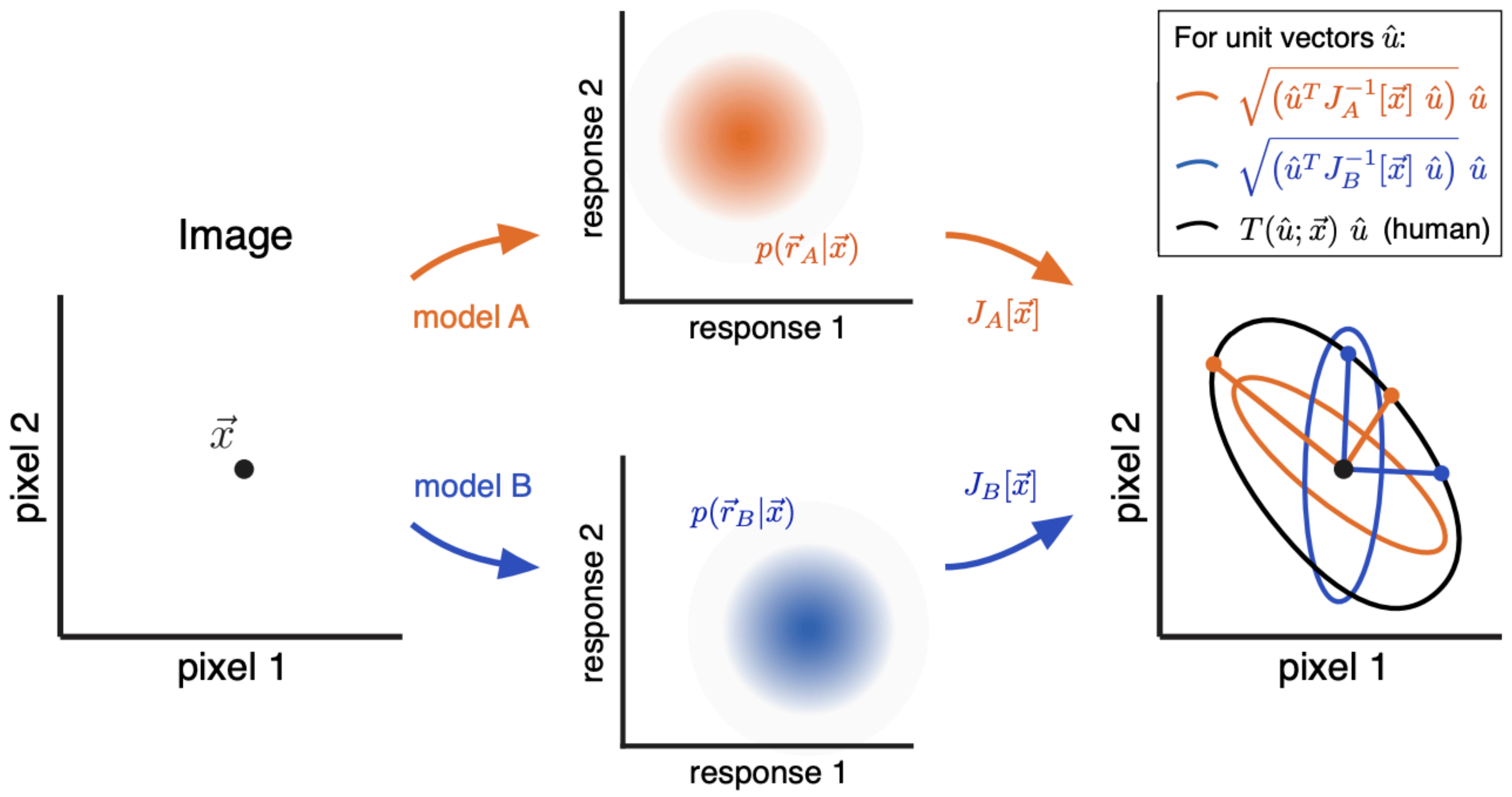
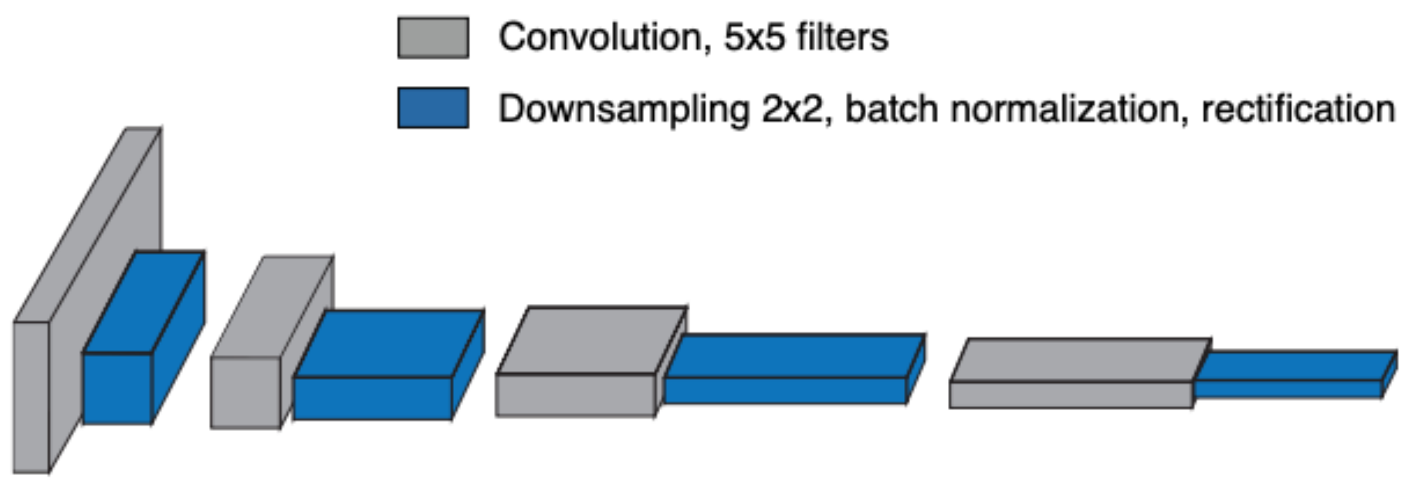
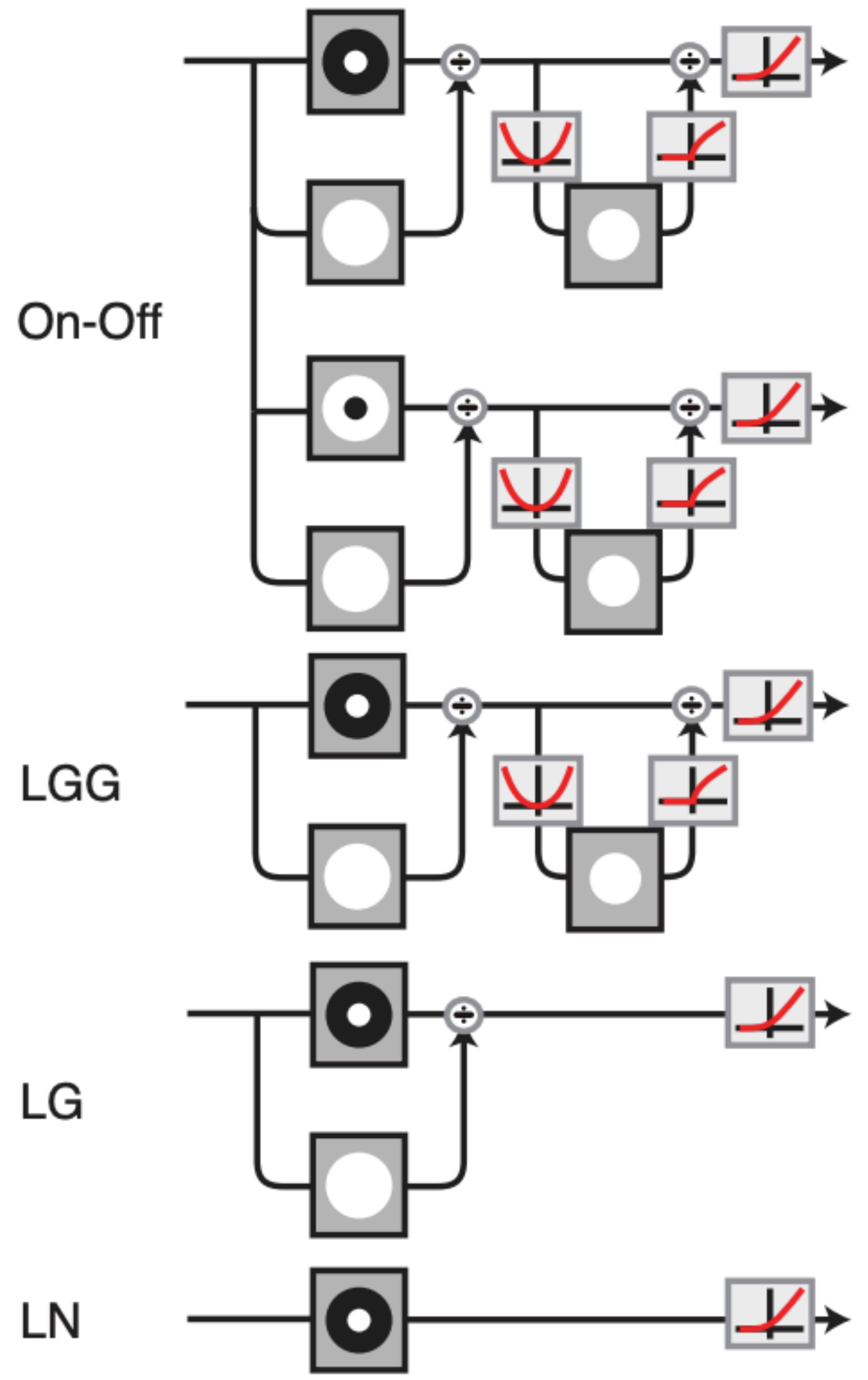


Image Credit: Berardino

Eigen-Distortion Analysis of Image Representations

- Simple bio-inspired models provide substantially better predictions of human sensitivity than either the CNN, or any combination of layers of VGG16





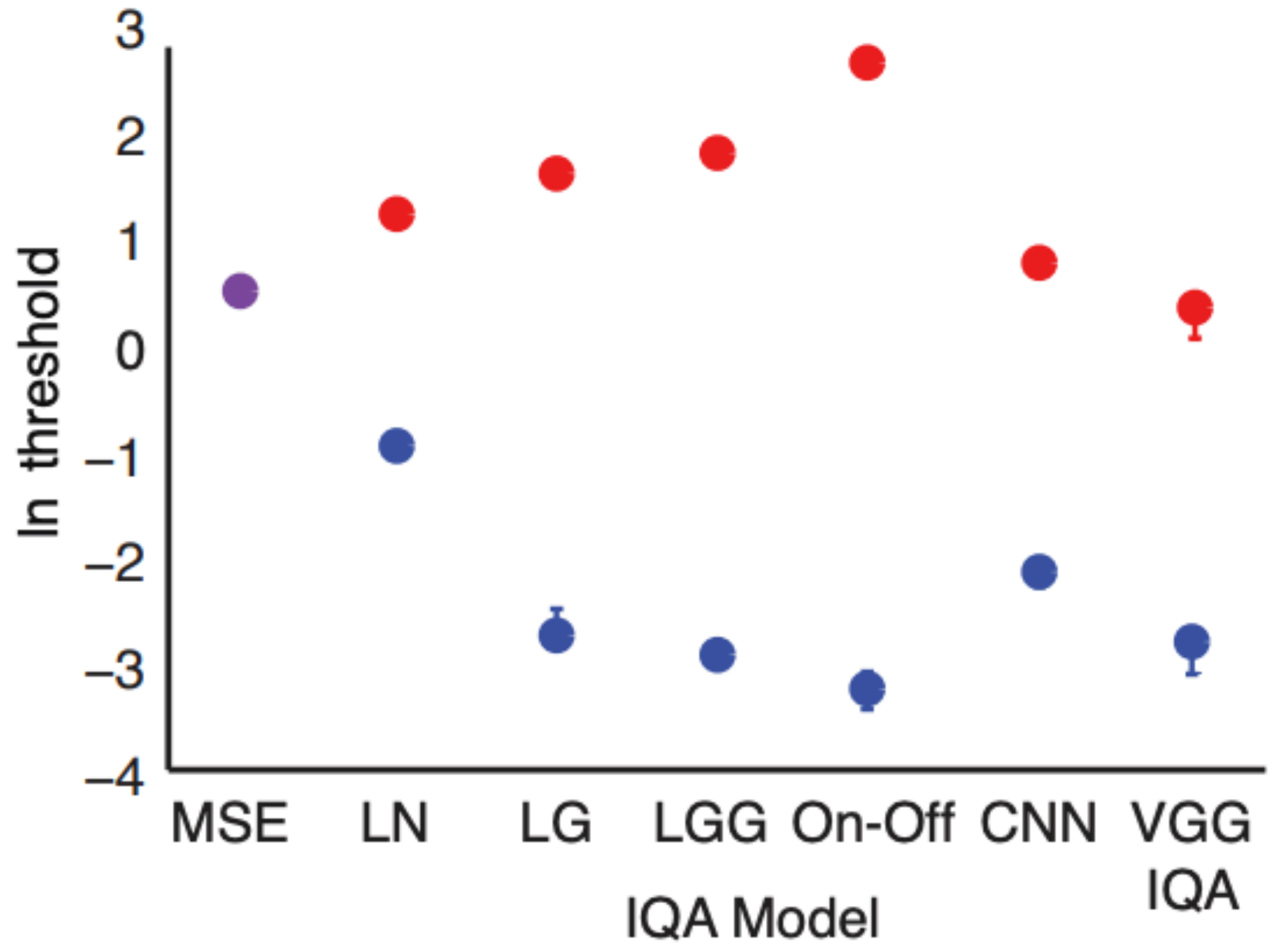


Image Credit: Berardino

Discussion

Discussion

- Fixed-set accuracy vs adaptive-set generalization
- Scale of human ratings
- Image quality $p(y \mid x)$ vs image prior $p(x)$
  - Question: Is it reasonable to test no-reference IQA models in the framework of maximum a posteriori based image restoration?