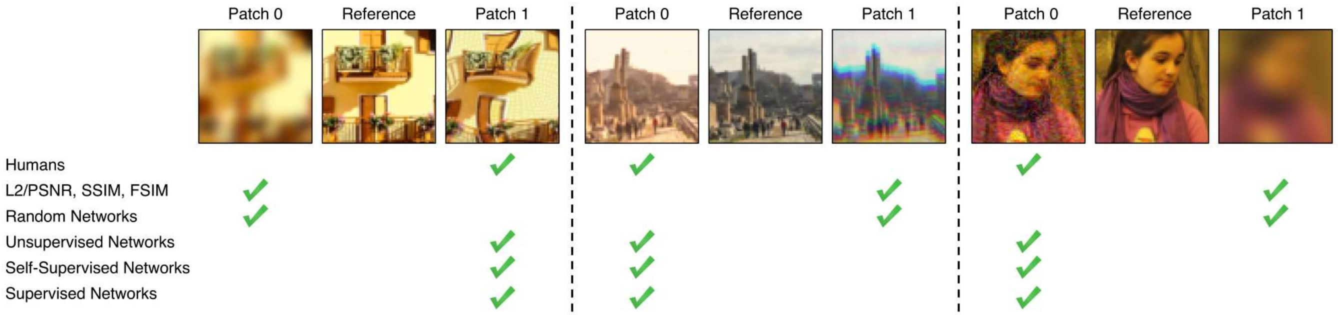
🍌 Spencer's Wiki

Perceptual Similarity

> The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.

> 🌏 **Source**
>
> Downloadable at: Open Access - CVPR 2018. Source code is available at: GitHub - richzhang/PerceptualSimilarity.

Brief Introduction



The paper argues that widely used image quality metrics like SSIM and PSNR mentioned in [[image-quality-assessment]] are *simple and shallow* functions that may fail to account for many nuances of human perception. The paper introduces a new dataset of human perceptual similarity judgments to systematically evaluate deep features across different architectures and tasks and compare them with classic metrics.

Findings of this paper suggests that *perceptual similarity is an emergent property shared across deep visual representations.*

Main contributions

In this paper, the author provides a hypothesis that perceptual similarity is not a special function all of its own, but rather a *consequence* of visual representations tuned to be predictive about important structure in the world.

- To testify this theory, the paper introduces a large scale, highly varied perceptual similarity dataset containing 484k human judgments.
- The paper shows that deep features trained on supervised, self-supervised, and unsupervised objectives alike, model low-level perceptual similarity surprisingly well, outperforming previous, widely-used metrics.
- The paper also demonstrates that network architecture alone doesn't account for the performance: untrained networks achieve much lower performance.

The paper suggests that with this data, we can improve performance by *calibrating* feature responses from a pre-trained network.

Methodology

The perceptual similarity dataset

> `This content is less related to my interests. I'll cover them briefly.`

- Traditional distortions: photometric distortions, random noise, blurring, spatial shifts, corruptions.

- CNN-based distortions: input corruptions (white noise, color removal, downsampling), generator networks, discriminators, loss/learning.
- Distorted image patches.
- Superresolution.
- Frame interpolation.
- Video deblurring.
- Colorization.

Similarity measures

- 2AFC similarity judgments: [[2afc]] - Two-alternative forced choice.
- Just noticeable differences: [[jnd]]

Deep Feature Spaces

Network activations to distance

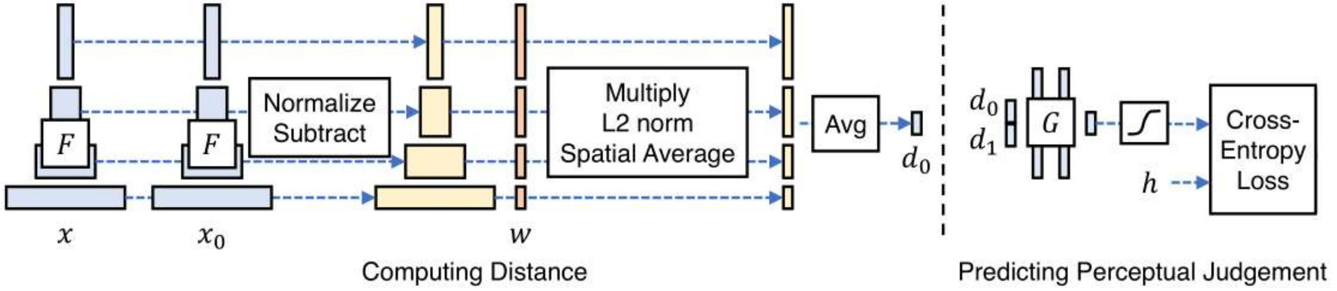


Figure 3: **Computing distance from a network** (Left) To compute a distance d_0 between two patches, x, x_0, given a network $\mathcal{F}$, we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector w, and take the ℓ_2 distance. We then average across spatial dimension and across all layers. (Right) A small network $\mathcal{G}$ is trained to predict perceptual judgment h from distance pair (d_0, d_1).

The distance between reference and distorted patches x and x_0 is calculated using this workflow and the equation below with a network $\mathcal{F}$. The paper extract feature stack from L layers and unit-normalize in the channel dimension. Then the paper scales the activations channel-wise and computes the ℓ_2 distance.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})\|_2^2$$

Training on this data

The paper considers the following variants:

- ***lin***: the paper keep pre-trained network weights fixed and learn linear weights w on top.
- ***tune***: the paper initializes from a pre-trained classification model and allow all the weights for network $\mathcal{F}$ to be fine-tuned.
- ***scratch***: the paper initializes the network from random Gaussian weights and train it entirely on the author's judgments.

Finally, the paper refer to these as variants of the proposed **Learned Perceptual Image Patch Similarity (LPIPS)**.

Experiments

Performance of low-level metrics and classification networks

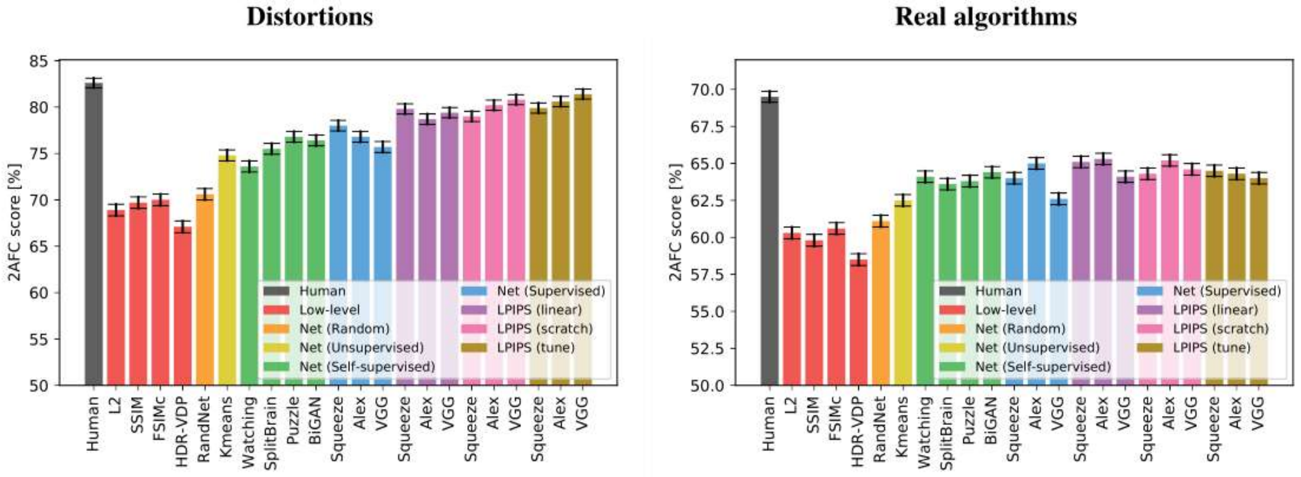


Figure 4: **Quantitative comparison.** We show a quantitative comparison across metrics on our test sets. (Left) Results averaged across our traditional and CNN-based distortions. (Right) Results averaged across our 4 real algorithm sets.

Figure 4 shows the performance of various low-level metrics (in red), deep networks, and human ceiling (in black).

Metrics correlate across different perceptual tasks

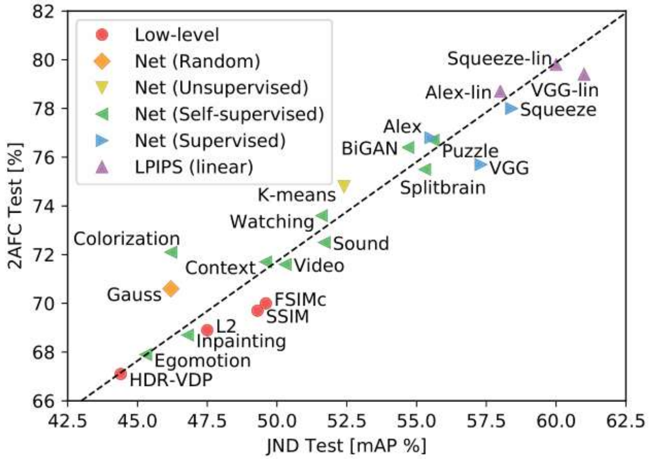


Figure 5: **Correlating Perceptual Tests.** We show performance across methods, including unsupervised [26], self-supervised [1, 43, 12, 56, 62, 41, 42, 40, 13, 63], supervised [27, 51, 20], and our perceptually-learned metrics (LPIPS). The scores are on our 2AFC and JND tests, averaged across traditional and CNN-based distortions.

The 2AFC distortion preference test has high correlation to JND: $\rho = .928$ when averaging the results across distortion types. This indicates that 2AFC generalizes to another perceptual test and is giving us signal regarding human judgments.

Where do deep metrics and low-level metrics disagree?

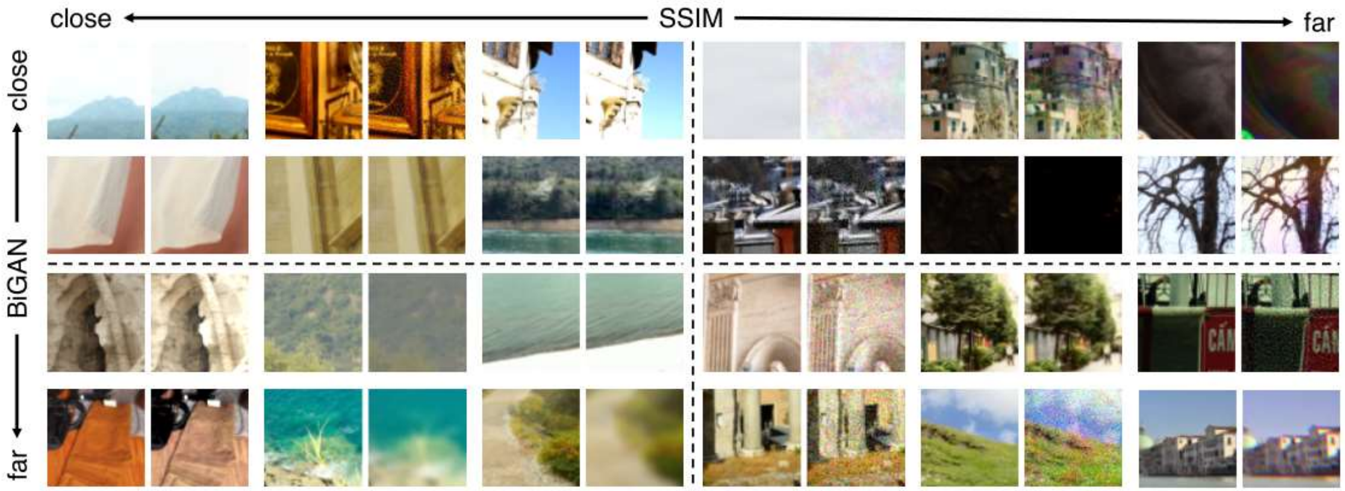


Figure 6: **Qualitative comparisons on distortions.** We show qualitative comparison on traditional distortions, using the SSIM [57] metric and BiGAN network [13]. We show examples where the metrics agree and siagree. A primary difference is that deep embeddings appear to be more sensitive to blur. Please see the appendix for additional examples.

Pairs which BiGAN perceives to be far but SSIM to be close generally contain some blur. BiGAN tends to perceive correlated noise patterns to be a smaller distortion than SSIM.

Conclusions

The stronger a feature set is at classification and detection, the stronger it is as a model of perceptual similarity judgments.

Features that are good at **semantic tasks**, are also good at **self-supervised and unsupervised tasks**, and also provide **good models of both human perceptual behavior and macaque neural activity.**

Referred in

- [[papers]]
  - | Paper Title | Publication | Source Code | | [[perceptual-similarity]] | CVPR 2018 | richzhang/PerceptualSimilarity | | [[pieapp]] | CVPR 2018 | prashnani/PerceptualImageError |