The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

Richard Zhang[12] Phillip Isola[13] Alexei A. Efros[1] Eli Shechtman[2] Oliver Wang[2]
[1]Berkeley Artificial Intelligence Research [2]Adobe Research [3]OpenAI

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Problem Statement

Evaluate how well metrics correspond with human perceptual judgments.

1) Collect a large-scale perceptual similarity dataset
2) Deep features across training objectives outperform widely-used perceptual metrics (e.g., SSIM)
3) Train new metric (LPIPS) on perceptual judgments
➔ Try it: richzhang.github.io/PerceptualSimilarity/

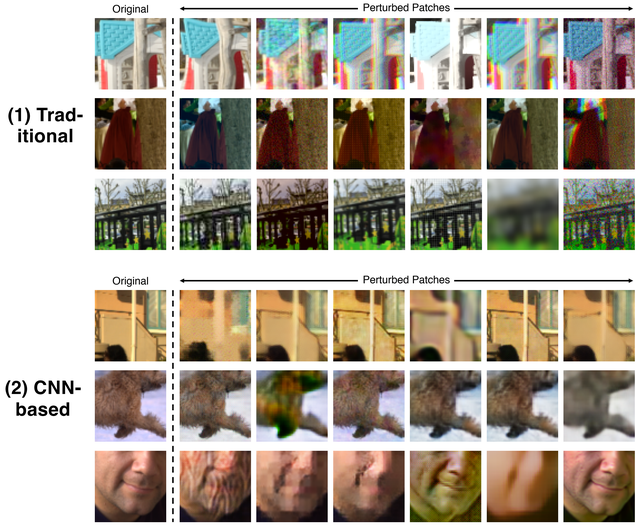
Two Alternative Forced Choice (2AFC)

Goal: Collect large-scale set of human perceptual judgments on distortions

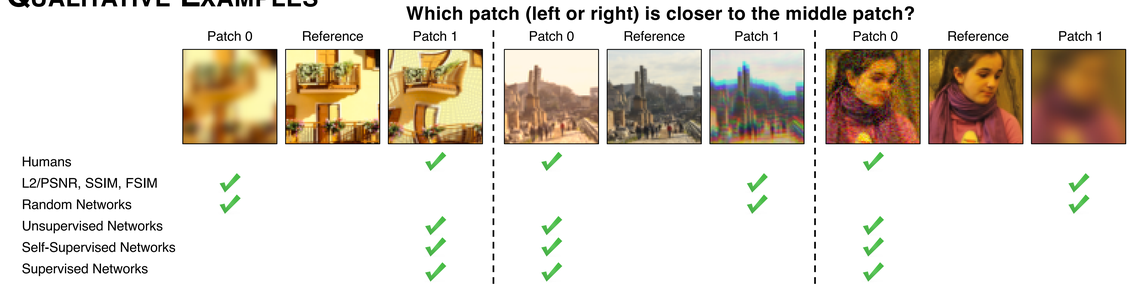
Procedure: Sample a patch. Distort it twice. Ask human which is smaller.

Distortions for Train&Val:
(1) Traditional distortions noise, photometric, blur, warps, compression
(2) CNN-Based distortions Randomly generated denoising autoencoders by varying hyperparameters

Distortions for Val only:
(3) Real algorithms Outputs from superresolution, frame interpolation, video deblurring, colorization algorithms



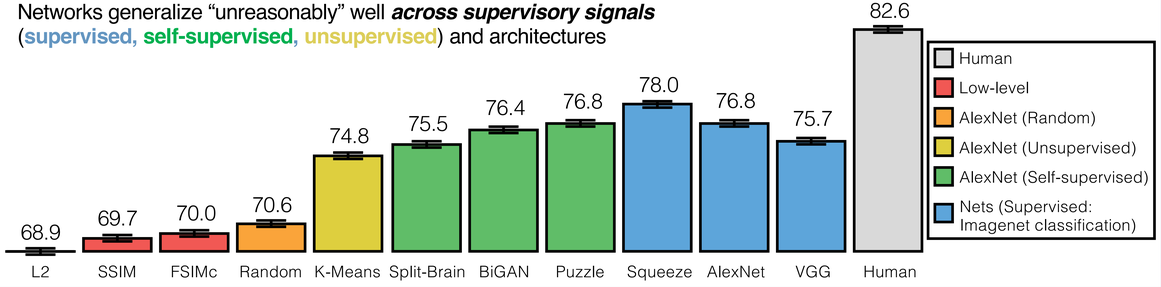
Qualitative Examples



How Perceptual are Off-the-shelf Networks?

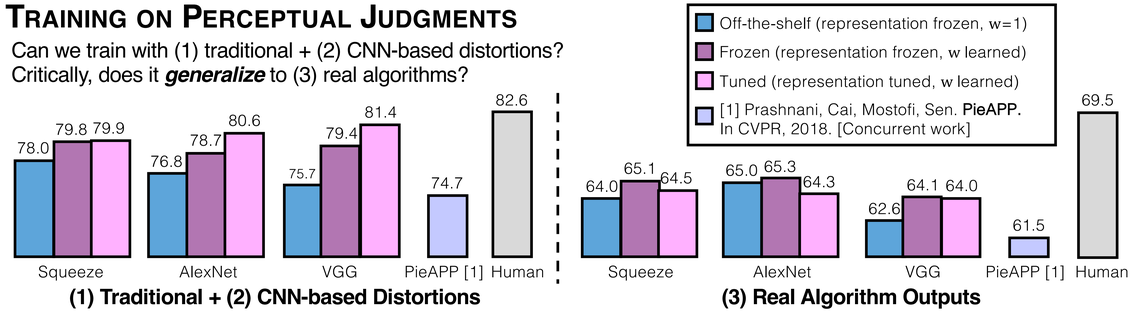
Test on **(1) Traditional + (2) CNN-based Distortions** with off-the-shelf networks

Networks generalize "unreasonably" well ***across supervisory signals*** (supervised, self-supervised, unsupervised) and architectures



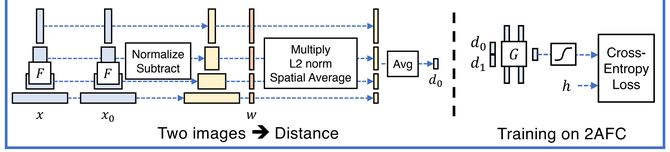
Training on Perceptual Judgments

Can we train with (1) traditional + (2) CNN-based distortions?
Critically, does it ***generalize*** to (3) real algorithms?



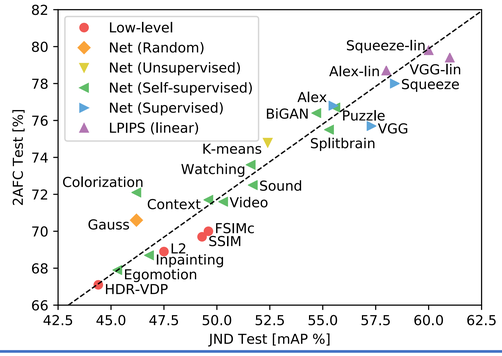
Off-the-shelf networks already perform well. Training a **linear layer** (LPIPS) on top yields small performance boost. But fine-tuning through representation ***leads to overfitting on specified distortion distribution***.

Deep Networks as a Perceptual Metric

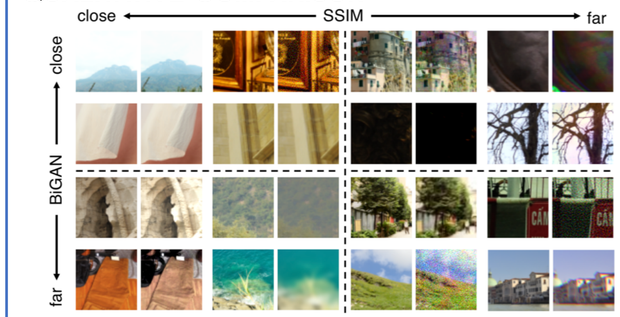


Just Noticeable Differences (JND)

Goal: Validate 2AFC with less "cognitively penetrable" test
Procedure: Ask human if 2 patches are identical or not
Evaluation: High correlation (ρ=.928) between 2AFC and JND



Qualitative Comparison



Though SSIM was ***not designed to handle geometric distortions***, it is commonly used, even when such distortions are a large factor.