We need a better perceptual similarity metric

Lubomir Bourdev

WaveOne, Inc.

CVPR Workshop and Challenge on Learned Compression

June 18th 2018

Challenges in benchmarking compression

- Measurement of perceptual similarity
- Consideration of computational efficiency
- Choice of color space
- Aggregating results from multiple images
- Ranking of R-D curves
- Dataset bias
- Many more!

Challenges in benchmarking compression

- **Measurement of perceptual similarity**
- Consideration of computational efficiency
- Choice of color space
- Aggregating results from multiple images
- Ranking of R-D curves
- Dataset bias
- Many more!

Why perceptual similarity is critical now?

- **Perceptual similarity is not a new problem**

■ Manos and Sakrison, 1974 ■ Girod, 1993 ■ Teo & Heeger, 1994 ■ Eskicioglu and Fisher, 1995 ■ Eckert and Bradley, 1998 ■ Janssen, 2001 ■ Wang, 2001 ■ Wang and Bovik, 2002 ■ Wang et al., 2002 ■ Pappas & Safranek, 2000 ■ Wang et al., 2003 ■ Sheikh et al., 2005 ■ Wang and Bovik, 2009 ■ Wang et al., 2009 ■ Many more…

Why perceptual similarity is critical now?

- **Perceptual similarity is not a new problem**

  ■ Manos and Sakrison, 1974 ■ Girod, 1993 ■ Teo & Heeger, 1994 ■ Eskicioglu and Fisher, 1995 ■ Eckert and Bradley, 1998 ■ Janssen, 2001 ■ Wang, 2001 ■ Wang and Bovik, 2002 ■ Wang et al., 2002 ■ Pappas & Safranek, 2000 ■ Wang et al., 2003 ■ Sheikh et al., 2005 ■ Wang and Bovik, 2009 ■ Wang et al., 2009 ■ Many more...

- **Today we have new much more powerful tools**
  - Deep nets can exploit any weaknesses in the metrics

Why perceptual similarity is critical now?

- **Perceptual similarity is not a new problem:**

  ■ Manos and Sakrison, 1974 ■ Girod, 1993 ■ Teo & Heeger, 1994 ■ Eskicioglu and Fisher, 1995 ■ Eckert and Bradley, 1998 ■ Janssen, 2001 ■ Wang, 2001 ■ Wang and Bovik, 2002 ■ Wang et al., 2002 ■ Pappas & Safranek, 2000 ■ Wang et al., 2003 ■ Sheikh et al., 2005 ■ Wang and Bovik, 2009 ■ Wang et al., 2009 ■ Many more…

- **Today we have new much more powerful tools**
  - Deep nets can exploit any weaknesses in the metrics
  - Nets get penalized if they do better than the metric

How do we measure quality assessment?

How do we measure quality assessment?

- **Idea 1: Stick to traditional metrics**
  - MSE, PSNR
  - SSIM, MS-SSIM [Wang et. al. 2003]

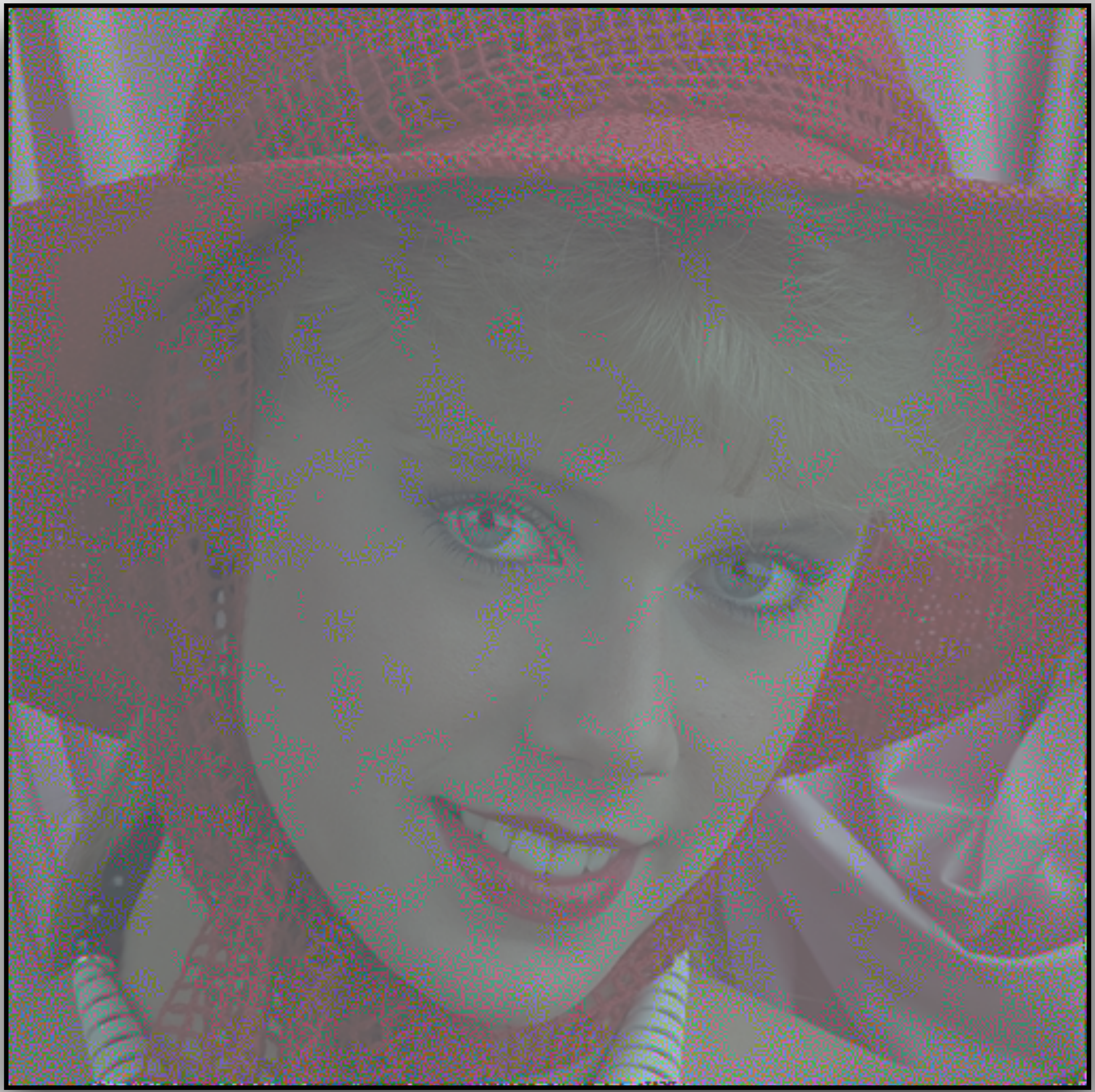
- **Simple, intuitive way to benchmark performance**

How do we measure quality assessment?

- **Idea 1: Stick to traditional metrics**
  - MSE, PSNR
  - SSIM, MS-SSIM [Wang et. al. 2003]
- **Simple, intuitive way to benchmark performance**
- **However, they are far from ideal**

Min PSNR on MS-SSIM isocontour

Target



MS-SSIM: 0.99
PSNR: 11.6dB

Min PSNR on MS-SSIM isocontour

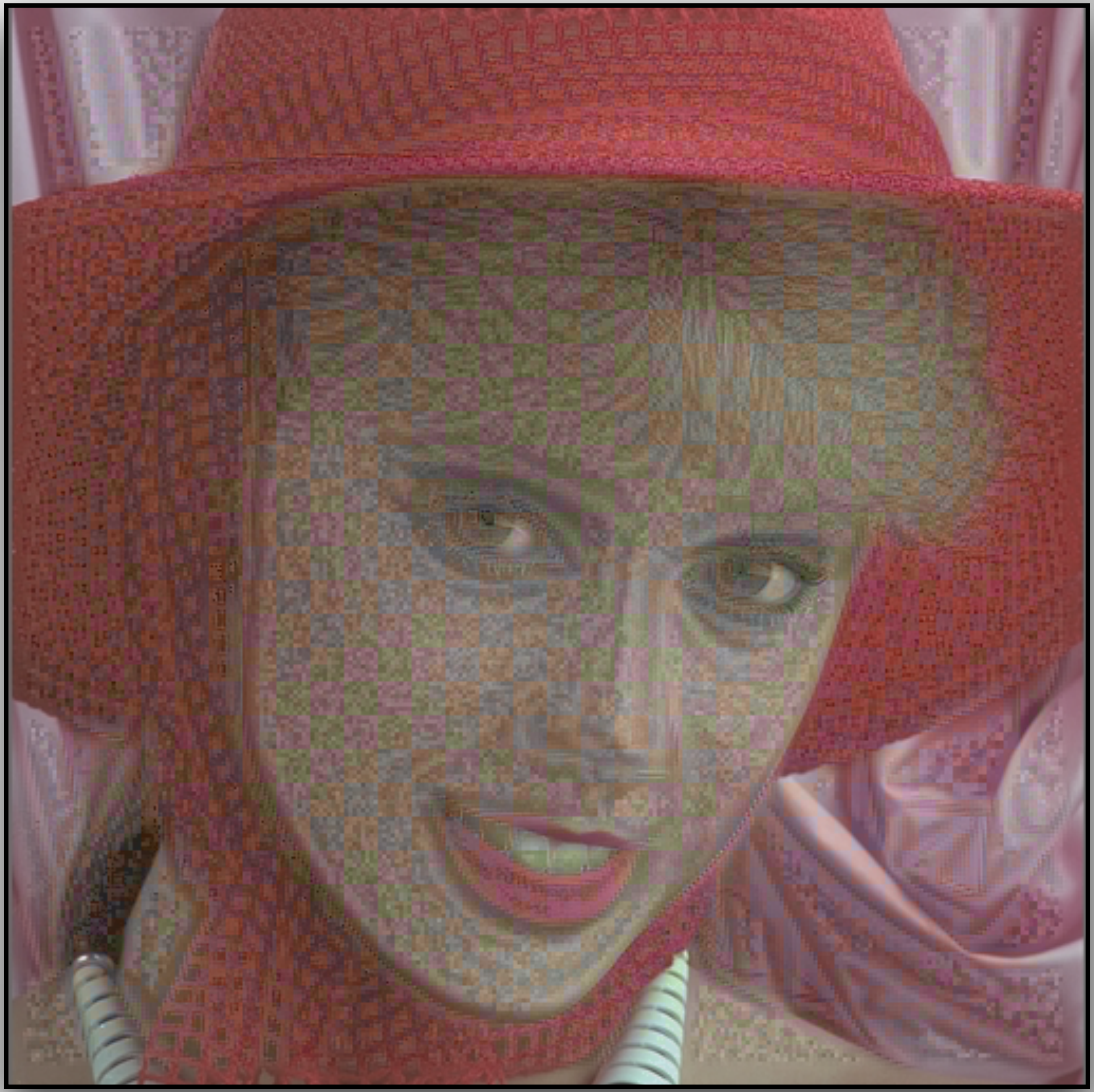
Target

MS-SSIM: 0.997

PSNR: 14.4dB

Min MS-SSIM on PSNR isocontour

Target



PSNR: 30dB

MS-SSIM: 0.15

Min MS-SSIM on PSNR isocontour

Target

PSNR: 40dB

MS-SSIM: 0.90

Min MS-SSIM on PSNR isocontour

Target

PSNR: 40dB

MS-SSIM: 0.90

Idea 2: Maybe we should maximize both?

Is maximizing PSNR + MS-SSIM the right solution?

Is maximizing PSNR + MS-SSIM the right solution?

~200
bytes

Is maximizing PSNR + MS-SSIM the right solution?

~200 bytes

Generic WaveOne (no GAN)

Domain-aware Adversarial model

Is maximizing PSNR + MS-SSIM the right solution?

~200 bytes

Generic WaveOne (no GAN)

MS-SSIM: 0.93

PSNR: 25.9

Domain-aware Adversarial model

MS-SSIM: 0.89

PSNR: 23.0

Is maximizing PSNR + MS-SSIM the right solution?

~200 bytes

Generic WaveOne (no GAN)

MS-SSIM: 0.93

PSNR: 25.9

Domain-aware Adversarial model

MS-SSIM: 0.89

PSNR: 23.0

Idea 3: Maybe we should use GANs?

GANs are very promising

GANs are very promising

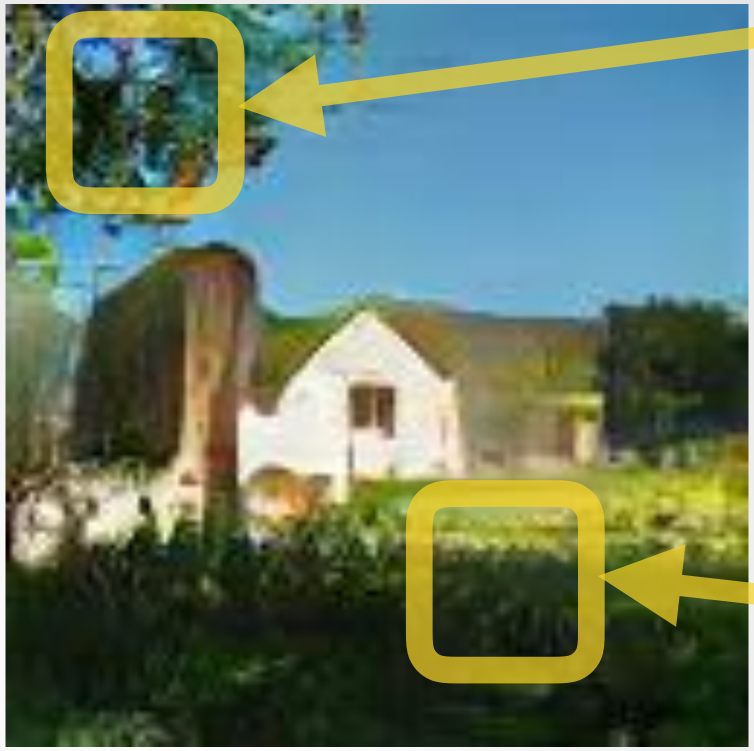
- **Reconstructions visually appealing (sometimes!)**
- **Generic and intuitive objective:**
  - Similarity function of the difficulty of distinguishing the images by an expert

GANs are very promising

- **Reconstructions visually appealing (sometimes!)**
- **Generic and intuitive objective:**
  - Similarity function of the difficulty of distinguishing the images by an expert
- **Unfortunately the loss is different for every network and evolves over time**

What makes people prefer the right image?

What makes people prefer the right image?



Looks like leaves

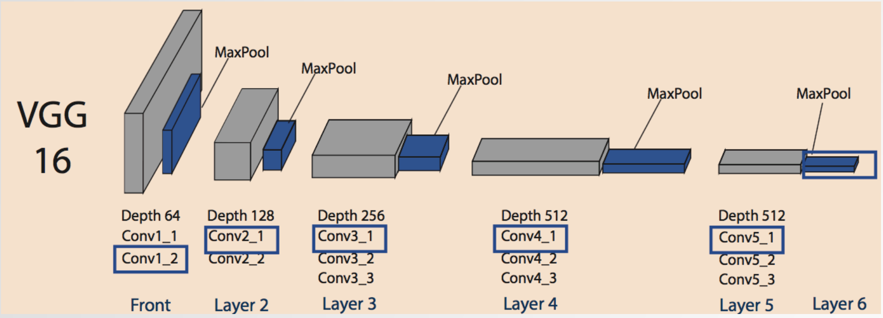
Looks like grass

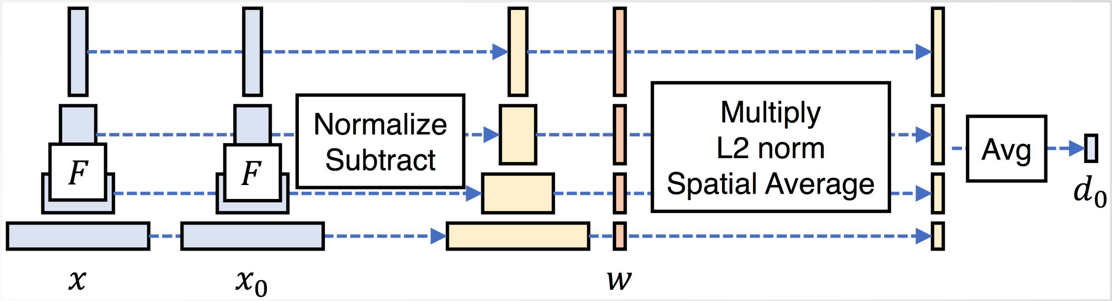
What makes people prefer the right image?

Idea 4: Maybe we should use semantics?

Losses based on semantics

- Intermediate layers of pre-trained classifiers capture semantics [Zeiler & Fergus 2013]



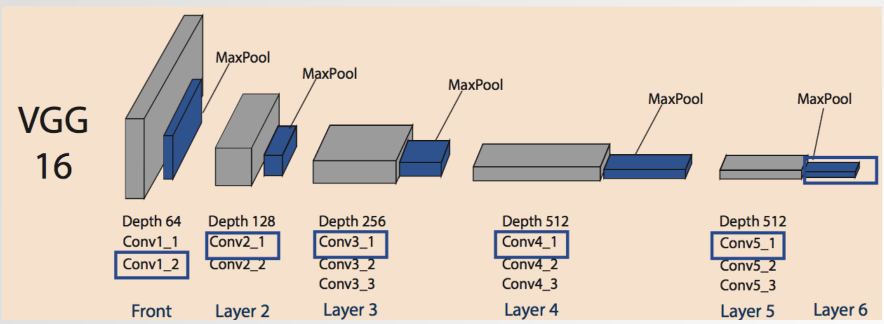


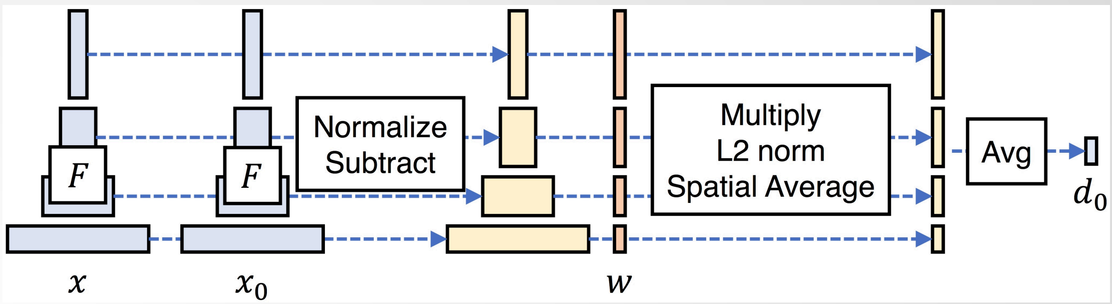
[Zhang et al, CVPR18]

- Significantly better correlation to MoS vs traditional metrics

Losses based on semantics

- **Intermediate layers of pre-trained classifiers capture semantics [Zeiler & Fergus 2013]**





[Zhang et al, CVPR18]

- **Significantly better correlation to MoS vs traditional metrics**

- **However, arbitrary and over-complete**
  - Millions of parameters
  - Trained on unrelated task
  - Which nets? Which layers? How to combine them?

Idea 5: Attention-driven metrics

Where the bandwidth goes

Where people look

Idea 5: Attention-driven metrics

Where the bandwidth goes

Where people look

- **All existing metrics treat every pixel equally**
  - Clearly suboptimal

Idea 5: Attention-driven metrics

Where the bandwidth goes

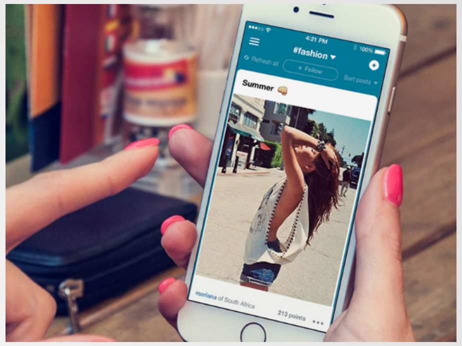
Where people look

- **All existing metrics treat every pixel equally**
  - Clearly suboptimal
- **But defining importance is another open problem**

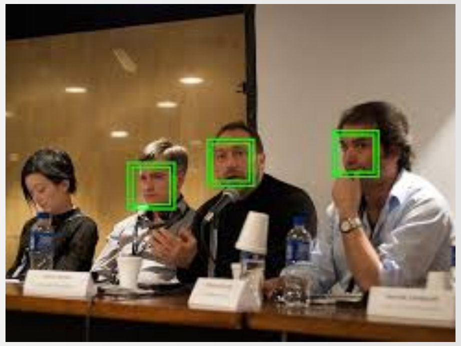
Idea 6: Task-driven metrics

- **A/B testing compression variants based on feature**

  - **Goal**: Social sharing
  - **Measure**: user engagement

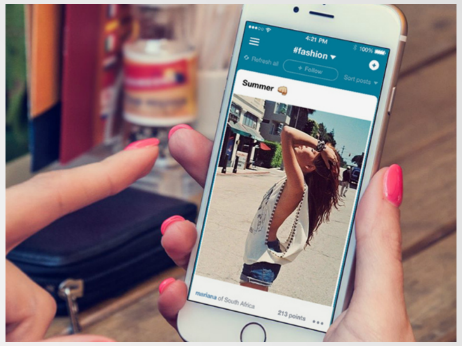


  - **Goal**: ML on the cloud
  - **Measure**: performance on the ML task

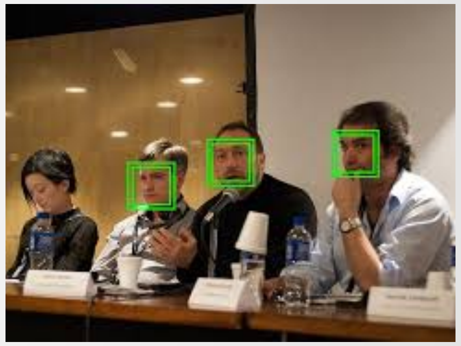


Idea 6: Task-driven metrics

- **A/B testing compression variants based on feature**
  - **Goal**: Social sharing
  - **Measure**: user engagement



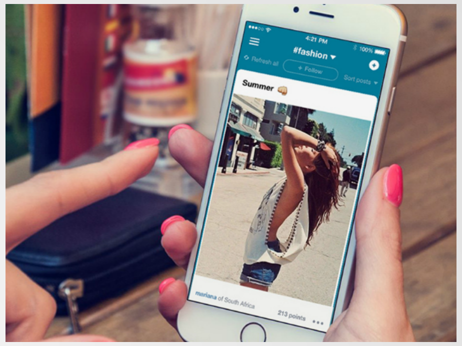
  - **Goal**: ML on the cloud
  - **Measure**: performance on the ML task



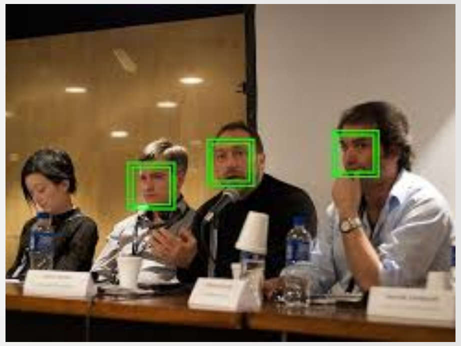
- **Solves the "right" problem**

Idea 6: Task-driven metrics

- **A/B testing compression variants based on feature**
  - **Goal**: Social sharing
  - **Measure**: user engagement



  - **Goal**: ML on the cloud
  - **Measure**: performance on the ML task



- **Solves the "right" problem**
- **However, not accessible, not repeatable, not back-propagatable**

Idea 7: when all fails, ask the experts

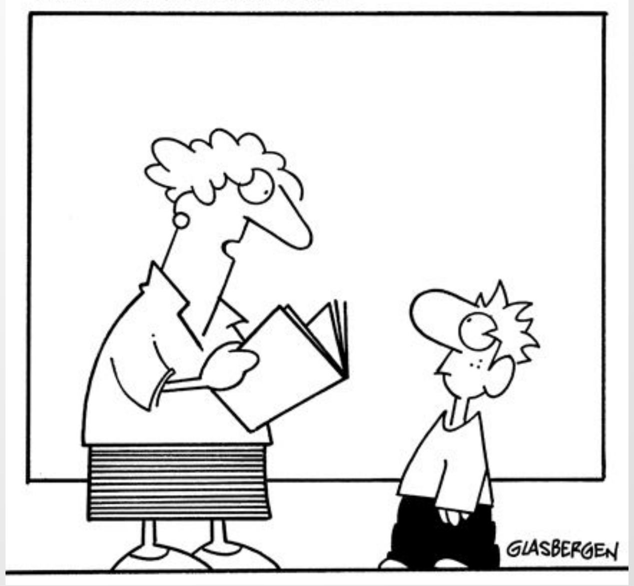
Idea 7: when all fails, ask the experts

- **Humans are the gold standard for perceptual fidelity**

Idea 7: when all fails, ask the experts

- **Humans are the gold standard for perceptual fidelity**

- **Challenges**
  - Hard to construct objective tests
  - Can’t back-propagate through humans
  - Expensive to evaluate (both time & money)
  - Non-repeatable



“On a scale from 0 to 1, how different are these two pixels? Only another 999,999 comparisons to go!”

Conclusion

- **The impossible wishlist for ideal quality metric:**
  - Simple and intuitive
  - Repeatable
  - Back-propagatable
  - Content-aware
  - Efficient
  - Importance-driven
  - Task-aware

Conclusion

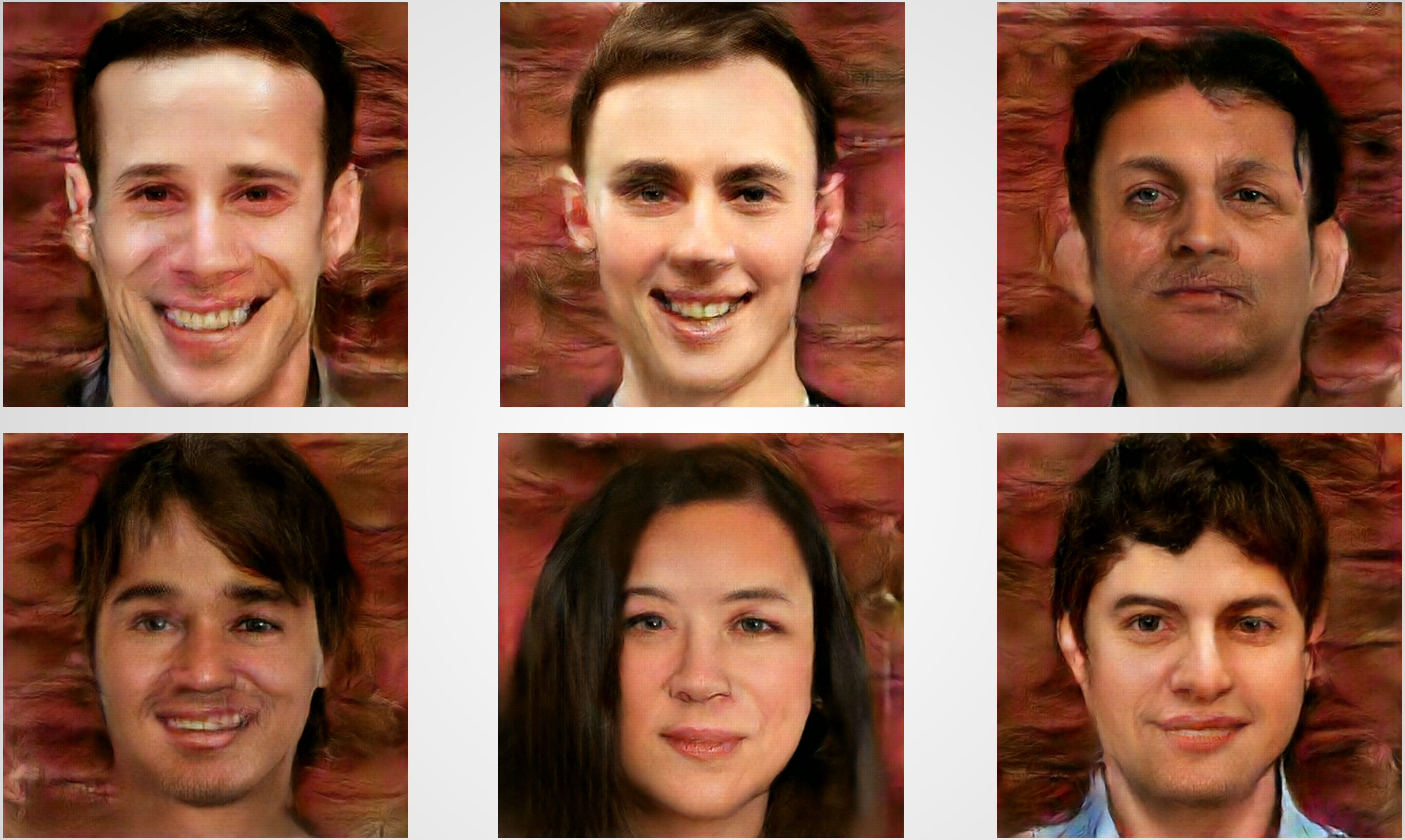
- **The impossible wishlist for ideal quality metric:**
  - Simple and intuitive
  - Repeatable
  - Back-propagatable
  - Content-aware
  - Efficient
  - Importance-driven
  - Task-aware
- **Improving quality metrics is critical in the neural net age**

Conclusion

- **The impossible wishlist for ideal quality metric:**
  - Simple and intuitive
  - Repeatable
  - Back-propagatable
  - Content-aware
  - Efficient
  - Importance-driven
  - Task-aware
- **Improving quality metrics is critical in the neural net age**

The wrong metrics lead to good solutions to the wrong problem!

Thanks to my team!



The WaveOne team, compressed to ***0.01 BPP***,
using GAN specializing on frontal faces

http://wave.one

WaveOne