Supplemental Material for E-LPIPS: Robust Perceptual Image Similarity via Random Transformation Ensembles

Markus Kettunen
Aalto University

Erik Härkönen
Aalto University

Jaakko Lehtinen
Aalto University
NVIDIA

1 Adversarial Images

The paper shows adversarial attacks (A1) and (A2) against both LPIPS and E-LPIPS, and shows that E-LPIPS resists the attacks significantly better. Figure 1 shows additional examples for attack (A1), and Figure 2 shows additional examples for attack (A2).

2 Effect of Ensemble Size

Each sample of $d_{\text{E-LPIPS}}$ requires applying an input transformation to both images. In Figure 3 we show that using only a small number of fixed input transformations will seriously deteriorate the robustness of E-LPIPS, and an effectively infinite number of different input transformations will maximize robustness.

We also confirm this result numerically for the 30 image-pair dataset used in the paper: The attacks pull the images, on average, to L_2 distance 14.5 ± 0.7 for ensembles of size 1, 11.7 ± 0.7 for size 16, 6.9 ± 0.5 for size 256, 4.7 ± 0.4 for size 4096 and 2.9 ± 0.2 for the full infinite ensemble.

3 Eigenvalues of Hessians

In the paper we study the local behavior of the distance functions to prove the severe ill-conditioning of LPIPS in the close neighborhood of an image x_0. We evaluate the largest and smallest eigenvalues of the Hessian H of the distance function $f_{x_0}(v) = d(x_0, x_0 + v)$. The Hessian is a very large matrix even for small images, $(3 \times H \times W) \times (3 \times H \times W)$, and cannot in practice be computed. Computing the Hessian-vector-product Hv is much easier, and implementations are readily available in TensorFlow and many other frameworks. We use this to find the largest eigenvalue of H by power iteration [2, 6], repeatedly applying $u_{k+1} = Hu_k/\|Hu_k\|$ from an initial guess. Since E-LPIPS is stochastic, we increase the number of samples geometrically by a factor of 1.35 (chosen empirically) for each iteration to guarantee convergence.

The smallest eigenvalues are much harder to find. We do it by inverse iteration, i.e. applying power iteration to H^{-1}: the smallest eigenvalue of H is the reciprocal of the largest eigenvalue of H^{-1}. While of course H^{-1} is intractable, we evaluate $H^{-1}v$ by solving the linear system $Hx = v$ with the conjugate gradient method. We increase the number of conjugate gradient iterations geometrically with factor 1.35 (chosen empirically) to ensure convergence. This is a very heavy process even without the stochasticity of H, and as described the paper, we evaluate all results involving the smallest eigenvalues in lower resolution (64×64) and use a weakened version of E-LPIPS with a fixed ensemble of input transformations. This weakening worsens the conditioning of E-LPIPS, meaning that the corresponding numbers in the paper – which show orders of magnitude improvement for E-LPIPS – are a lower bound for the improvement from the full method.

Preprint. Under review.

3.1 Statistics of Eigenvalues

In the paper we show the mean, variance, skewness and kurtosis of the eigenvalues of the Hessian matrix. Our Hessians H are too large to allow for extracting the full spectrum of eigenvalues, but these statistical descriptors can be easily computed from the raw moments $\sum_i \lambda_i^k / N$ which we can obtain by Monte Carlo sampling. Since we can evaluate the product Hv for any vector v, the raw moments are given by the Monte Carlo average of $v^T H^k v$ over the unit sphere:

$$\frac{1}{M}\sum_{j=1}^{M} \hat{v}_j^T H^k \hat{v}_j \approx \fint_{|v|=1} v^T H^k v \,\mathrm{d}v = \frac{1}{N}\sum_{i=1}^{N} \lambda_i^k. \tag{1}$$

We were unable to locate this result in existing scientific literature so we provide a proof:

Theorem 1. *Let H be a real symmetric positive semi-definite $N \times N$ matrix and (λ_i) its eigenvalues. The mean of its eigenvalues equals the average integral of $v^T H v$ over the N-dimensional unit sphere,*

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i = \fint_{|v|=1} v^T H v \,\mathrm{d}v. \tag{2}$$

Proof. Let's start with the integral on the right-hand-side. Since H is a real symmetric matrix, it allows for an eigendecomposition with orthogonal matrices: $H = SDS^T$. This allows us to write $v^T H v = v^T S D S^T v = (S^T v)^T D (S^T v)$. With the change of variables $x = S^T v$, which has a unit Jacobian determinant and retains the norm, we have

$$\fint_{|v|=1} v^T H v \,\mathrm{d}v = \fint_{|x|=1} x^T D x \,\mathrm{d}x. \tag{3}$$

Let us write x as a sum of the standard basis vectors, $x = \sum_i x_i \hat{e}_i$. Matrix D is diagonal with $D_{ii} = \lambda_i$, so by orthogonality of the standard basis vectors we have

$$\fint_{|x|=1} x^T D x \,\mathrm{d}x = \sum_{j=1}^{N}\sum_{k=1}^{N} \fint_{|x|=1} (x_j \hat{e}_j^T)(\lambda_k x_k \hat{e}_k) \,\mathrm{d}x = \sum_{k=1}^{N} \lambda_k \fint_{|x|=1} x_k^2 \,\mathrm{d}x. \tag{4}$$

We note that the coordinate x_k is not an independent variable, but a function of x: $x_k = \hat{e}_k^T x$. The integral is also rotation invariant and we can use any coordinate of x in the integrand. Therefore, if we take the average over all coordinates, we get

$$\sum_{k=1}^{N} \lambda_k \fint_{|x|=1} x_k^2 \,\mathrm{d}x = \sum_{k=1}^{N} \frac{\lambda_k}{N} \sum_{j=1}^{N} \fint_{|x|=1} x_j^2 \,\mathrm{d}x = \sum_{k=1}^{N} \frac{\lambda_k}{N} \fint_{|x|=1} \sum_{j=1}^{N} x_j^2 \,\mathrm{d}x = \frac{1}{N}\sum_{i=1}^{N} \lambda_i. \tag{5}$$

This concludes the proof. □

The immediate follow-up is that means of λ_i^k may be evaluated by integrating $v^T H^k v$ over the unit sphere, where $H^k v$ may be evaluated by repeated multiplication by H.

4 Visualizing Largest Eigenvectors

We include a video visualizing randomly sampled perturbations in the space spanned by the 16 largest eigenvectors of LPIPS and E-LPIPS, computed for 6 images from the OpenImages [5] dataset at a resolution of 128×128 pixels. The perturbations have been normalized to emphasize visual characteristics instead of magnitude. The results of E-LPIPS seem to adapt more to the image content and to be a better match to human judgment, whereas the corresponding perturbations of LPIPS seem more random. Figure 4 shows two examples.

5 Attacks' Sensitivity to Modification

Our adversarial attacks against LPIPS appear to survive minor changes, typically losing around half of their potency at JPEG compression 92 without chroma sub-sampling, dropping color precision from 8 bits to 5 bits per channel, tiny Gaussian blur ($\sigma = 0.5$ px), small amounts of Gaussian noise ($\sigma = 0.01$ for images in range $[0, 1]$), small rotations (0.1 – 0.2 degrees for images of 400×400), and moderate changes to brightness (10 – 40%) and contrast (0.3 – 2.6 in PIL).

6 Pairwise Averages

We study the geometric properties of the distance metrics by seeking averages (barycenters) of input images A and B (Figures 5, 6 and 7). The L_2 barycenter is obtained as the pixelwise mean; other results are obtained through numerical optimization by solving

$$\arg\min_x \; d(A, x)^2 + d(B, x)^2. \tag{6}$$

We find that the E-LPIPS geometry yields images that on a cursory inspection look quite close to the pixelwise means, but a closer look reveals significantly less ghosting and better melding of local image features; a more faithful image hybrid overall. Both the SqueezeNet and VGG versions of LPIPS yield midpoints that clearly (though informally) fall further off the natural image manifold. In addition to high sensitivity to small perturbations in pixel space, the less robust metrics have a larger "blind spot" that gives only weak pressure for the mean images to look natural, offering a different view into the behavior of adversarial attacks in the paper.

7 Barycenters of Noisy and Shifted images

In the paper we show that unlike the L_2 barycenter, the E-LPIPS barycenter for images corrupted with Gaussian noise actually resembles all of the noisy inputs. We show more examples of this in Figures 8, 9 and 10. We also note that the E-LPIPS barycenter is not just copying parts from the different input images but its barycenter has a noise pattern not present in any of the inputs.

We repeat this test for random shifts in figures 11, 12 and 13, and note that while the L_2 barycenter results in a blurry image, the E-LPIPS is a much sharper image (i.e., more similar to the inputs).

8 Geodesics

We solve for discrete geodesics of 8 intermediate frames by minimizing

$$\arg\min_{x_1,\ldots,x_8} \sum_{i=0}^{8} d(x_i, x_{i+1})^2, \quad \text{with } x_0 = a,\ x_9 = b. \tag{7}$$

The free frames $x_1, \ldots, x_8$ are initialized to random noise. The optimization is performed with Adam with a learning rate of 0.002 over 40 000 steps.

As information propagates only between pairs of individual frames, and the only sources of fixed supervisory signal are the first and last frames, the optimization problem is difficult. We hierarchically precondition the problem in two ways. First, the spatial dimensions of each image are represented as four-level Laplacian pyramid expansions [1]. Second, the residual low-pass and intermediate band- and high-pass representations in the pyramid are treated as 3D tensors over image x, y and the frame index i (time). The time evolution of the pyramid coefficients is encoded in an unnormalized[1] Haar wavelet basis of dimension 8.

This hierarchical representation over both space and time couples the frames together and shapes the optimization landscape in a way that encourages explaining large-scale and slow variations compactly. We use the same preconditioner for all metrics.

The supplemental material contains videos that compare the E-LPIPS, LPIPS and L_2 results.

[1] Each basis function contains only values -1, 0, and 1.

Distortion	E-LPIPS (ours)	LPIPS [10] (VGG)	LPIPS [10] (SqueezeNet)
Traditional	74.75	76.90	77.78
CNN	82.94	82.79	83.51
Colorization	62.83	62.08	64.81
Superres.	71.43	68.96	70.97
Deblur	60.19	58.90	60.62
Frame int.	62.83	62.27	62.73
Mean (all)	69.16	68.65	70.07

Table 1: Accuracy (percent) of predicting human answers to two-alternative forced choice (2AFC) visual similarity questions, evaluated over various forms of image distortions. Our metric performs on par with that of Zhang et al. [10]. The mean human score is 73.9, while simple metrics such as pixel-wise L_2 distance and SSIM achieve a score of 63%.

9 Denoising and Super-Resolution Results

We study whether the increased robustness of E-LPIPS translates into improved robustness or performance when used as a training loss for neural networks. To test this, we train a blind denoiser and a $4\times$ super-resolution network. The network is described in Figure 14.

Both networks use weight normalization [9] with He initialization [3], and are optimized with Adam [4] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. Learning rate is constant 0.0005 with geometric ramp-up in the beginning.

The result images for the denoising test are displayed in Figure 15, with numerical data in Table 2. The super-resolution result images are displayed in Figure 16, and the numerical data in Table 3.

For training we use the 60% of the ImageNet 2012 validation dataset [8]. We report our results for the 24-image Kodak Lossless True Color Image Suite dataset.

To summarize our neural network training results, we find that E-LPIPS seems to slightly improve results for both denoising and super-resolution networks, although the improvement is quite small. We also find that E-LPIPS increases robustness over LPIPS: As we state in the paper, neural networks with enough expressive power may learn to cheat LPIPS, resulting in nonsensical images which still report low LPIPS distance to the ground truth (Figure 17). We have never observed this with E-LPIPS.

10 2AFC Results

The paper includes a study on the E-LPIPS and LPIPS metrics' accuracy in predicting human opinion in a two-alternative forced choice (2AFC) test. Table 1 shows a breakdown of the results into different distortion categories. See Zhang et al. [10] for category descriptions.

References

[1] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[2] C. De Sa, B. He, I. Mitliagkas, C. Ré, and P. Xu. Accelerated stochastic power iteration. *arXiv preprint arXiv:1707.02670*, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society.

[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[5] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[6] R. Mises and H. Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(2):152–164, 1929.

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 9351:234–241, 2015.

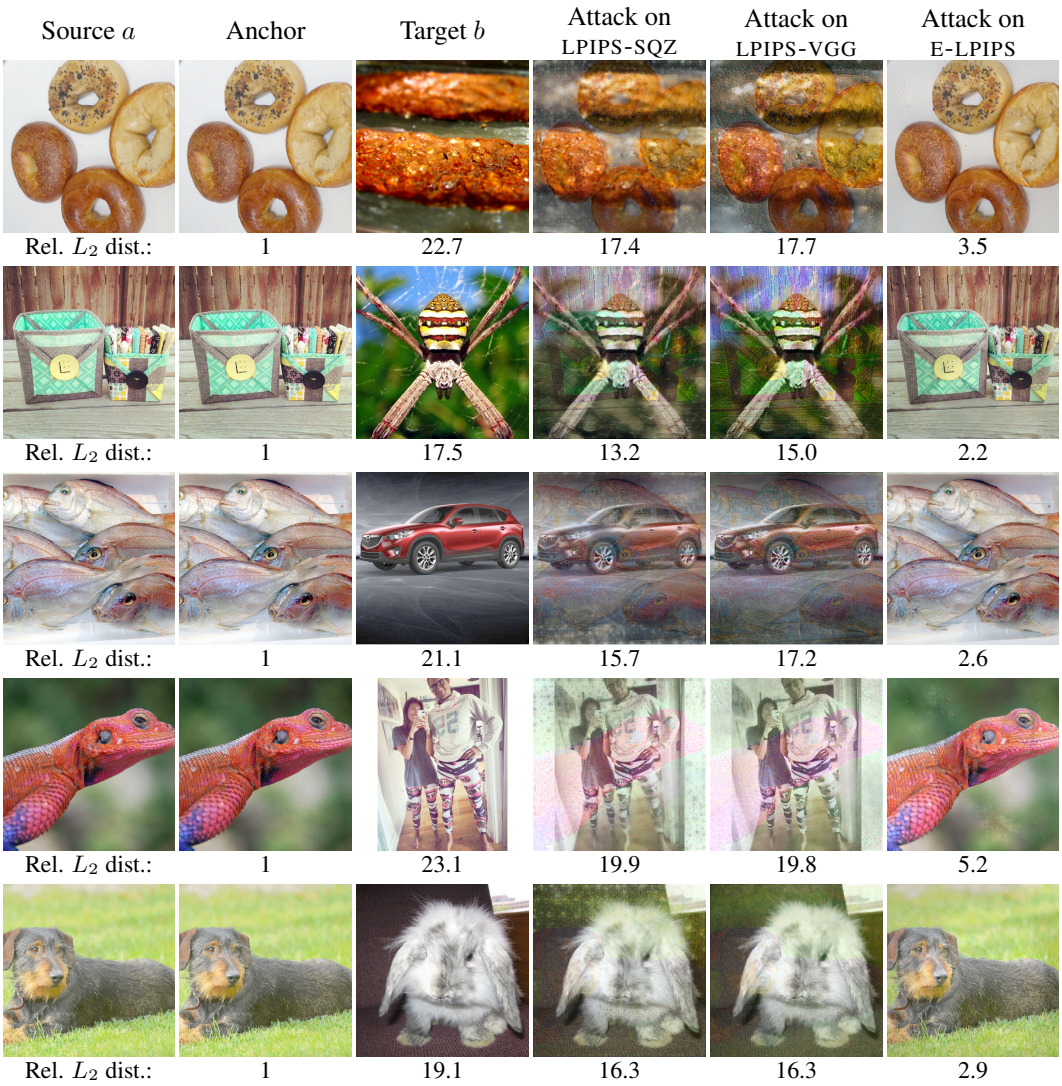


Figure 1: More examples for Attack (A1). The attack keeps the image at the anchor's distance from the input according to the LPIPS or E-LPIPS metric, and pulls the images as close to the target as possible, in L_2 distance. Both LPIPS metrics yield badly to the attack, typically ending up resembling the target more than input. The E-LPIPS result stays much closer to the source both visually and by relative L_2 distance, which is desirable.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[9] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *CoRR*, abs/1602.07868, 2016.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

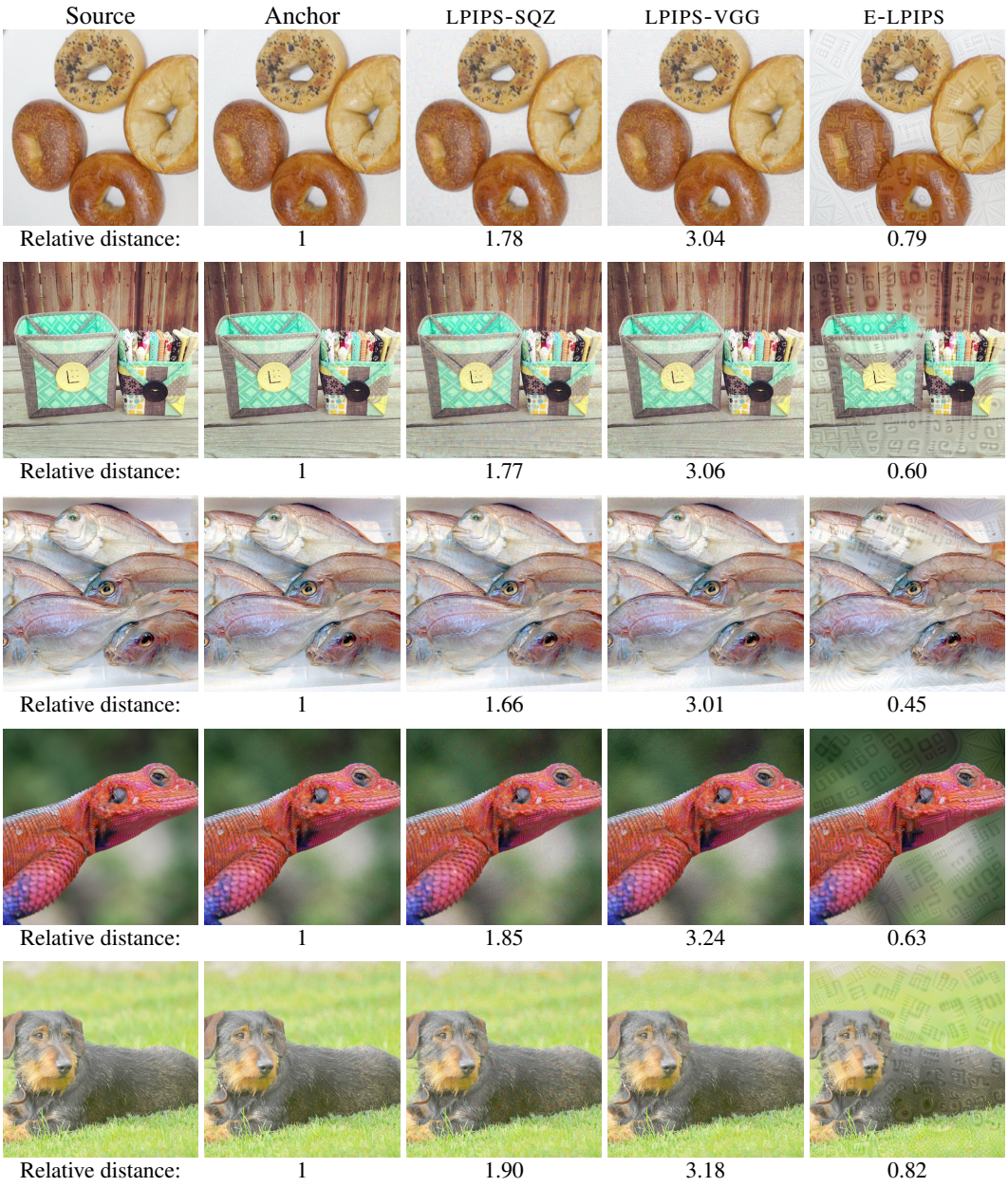
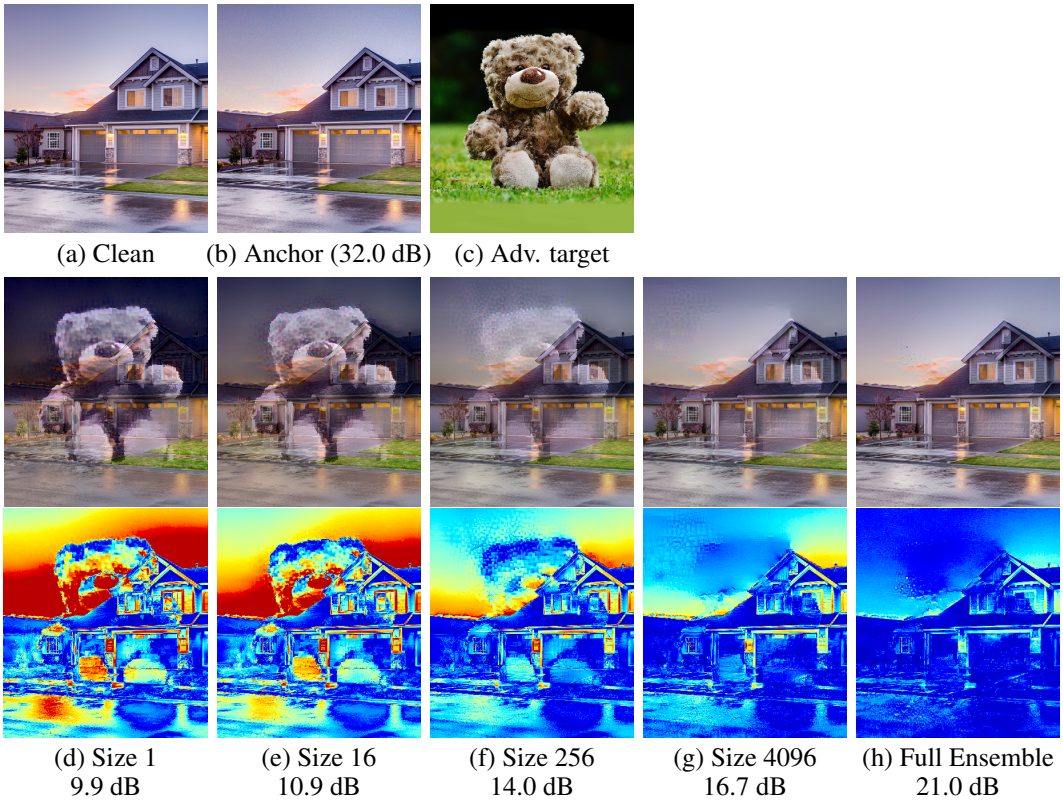


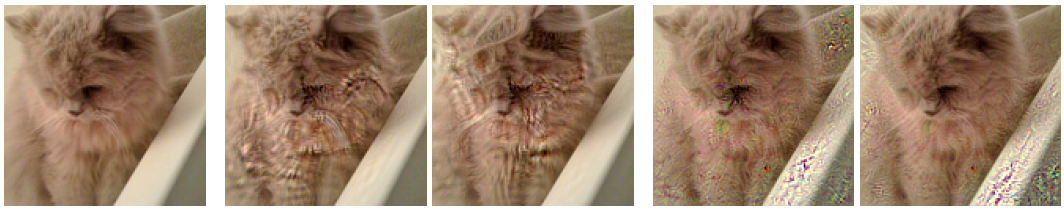
Figure 2: More examples for Attack A2. Both LPIPS metrics allow the image to be pushed far away in distance by modifications that are small both visually and in L_2 sense. In contrast, the attack is unable to increase the E-LPIPS distance nearly as much; furthermore the visual change is much more clearly visible at the same L_2 distance, which is desirable. The relative distance reported below the images is normalized such that 1 is the mean distance between the different images in the dataset.



(a) Clean (b) Anchor (32.0 dB) (c) Adv. target

(d) Size 1 9.9 dB (e) Size 16 10.9 dB (f) Size 256 14.0 dB (g) Size 4096 16.7 dB (h) Full Ensemble 21.0 dB

Figure 3: An ablation study on E-LPIPS with different sized ensembles. Images (d-h) (middle row) have been optimized as close to (c) as possible in terms of d_{L_2} while constraining their E-LPIPS distance from (a) to the distance between (a) and its noisy copy (b). Column (h) shows the attack result against the full ensemble of all input transformations, while (d-g) show the result for different numbers of fixed input transformations. All cases include dropout. The top row of (d-h) shows the full images, and the bottom row shows the error distributions and PSNR of the attack results. As the size of the ensemble grows the PSNR numbers increase and the errors in the images diminish. This indicates that using the full spectrum of random transformations in the ensemble results in increased robustness even compared to a large fixed set of transformations.



Input E-LPIPS dir. 1 E-LPIPS dir. 2 LPIPS dir. 1 LPIPS dir. 2

Figure 4: Two random directions from the subspace of largest eigenvectors added on top of the input image for E-LPIPS and LPIPS. The E-LPIPS perturbations would seem to bring the image farther from the input in subjective human opinion, which is desirable. We recommend the reader to watch the video in the supplemental material.

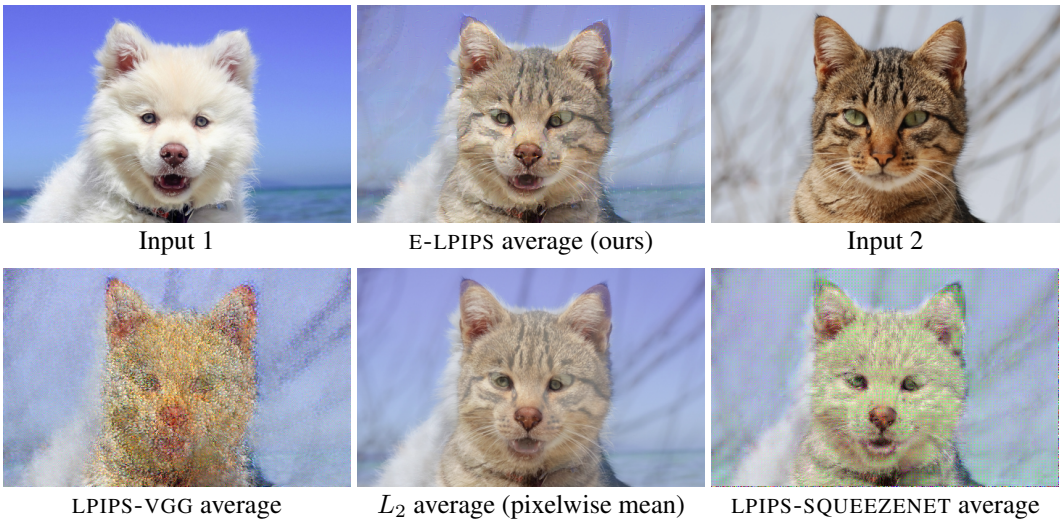


Figure 5: Averages of a dog and a cat. Compare how protruding the snout is, the mouth's shape, the details in the fur texture, and overall sharpness.

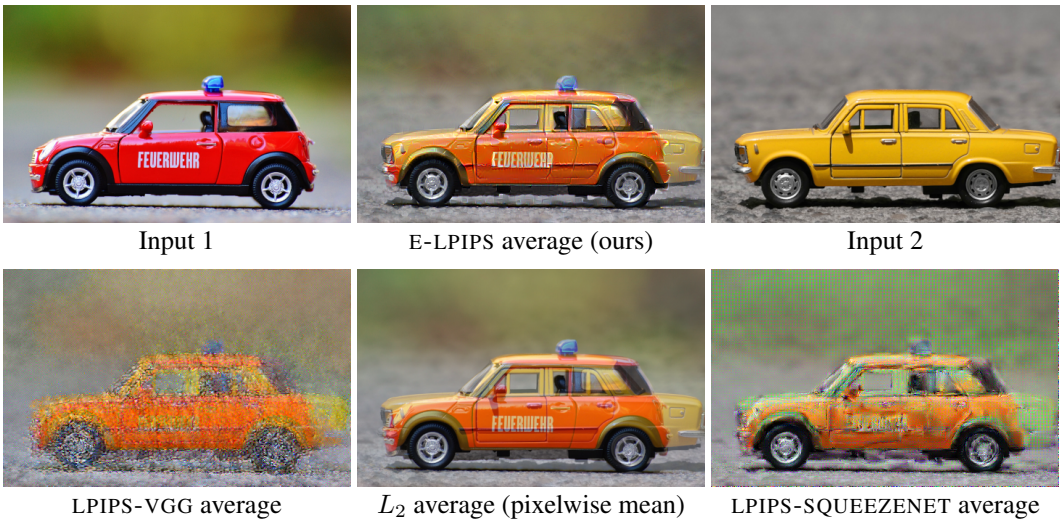


Figure 6: Averages of a red car and a yellow car. Compare the melding of the hood, windshield, the underside of the car and the general sharpness of details.

	Traditional Metrics			Neural Metrics		
Error metric → Network loss ↓	L1 ($\times 10^2$)	L2 ($\times 10^3$)	Mean color L1 ($\times 10^3$)	LPIPS-SQZ ($\times 10^2$)	LPIPS-VGG ($\times 10^2$)	Ours ($\times 10^2$)
L2	1.74 ± 0.18	0.62 ± 0.12	0.65 ± 0.11	1.20 ± 0.11	1.80 ± 0.15	0.96 ± 0.14
LPIPS-SQZ	2.15 ± 0.20	0.91 ± 0.15	$\mathbf{1.90 \pm 0.34}$	0.84 ± 0.07	1.69 ± 0.10	0.96 ± 0.12
LPIPS-VGG	2.05 ± 0.19	0.81 ± 0.14	3.62 ± 0.55	$\mathbf{0.88 \pm 0.08}$	1.44 ± 0.09	$\mathbf{0.86 \pm 0.11}$
Ours	$\mathbf{2.01 \pm 0.19}$	$\mathbf{0.78 \pm 0.13}$	2.42 ± 0.32	0.89 ± 0.08	$\mathbf{1.56 \pm 0.11}$	0.81 ± 0.11

Table 2: Means of errors and two sigma confidence intervals for our denoising networks, averaged over the Kodak test set. Numbers whose training and result metrics correspond are grayed out and should not be considered for a fair comparison. The best numbers for each metric are colored dark green, and those that fit in its confidence interval are colored light green. Others are colored red. The numbers are averages of four snapshots of the network from a single training run.

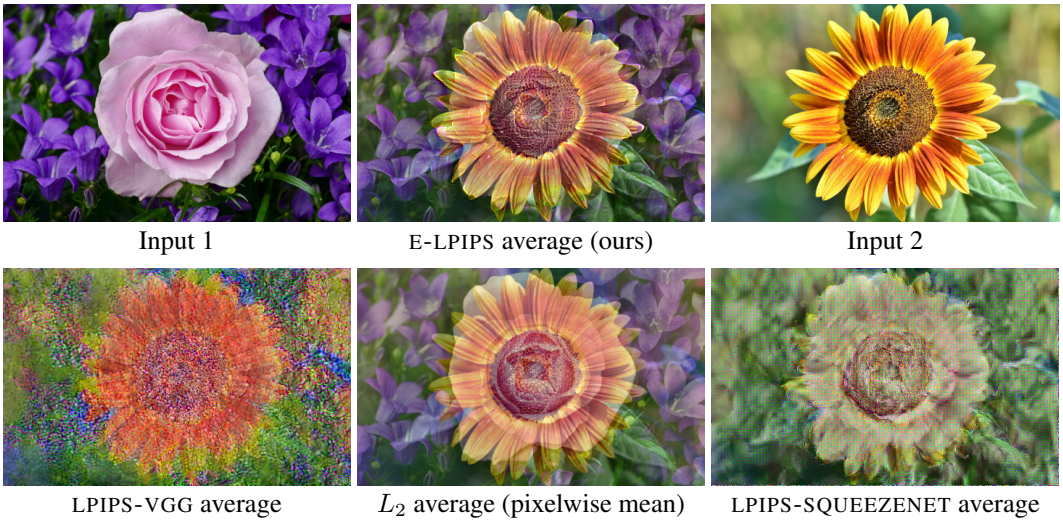


Figure 7: Averages of a rose and a sunflower. See the clarity of the hybrid's shape. Also compare the contrast in the background.

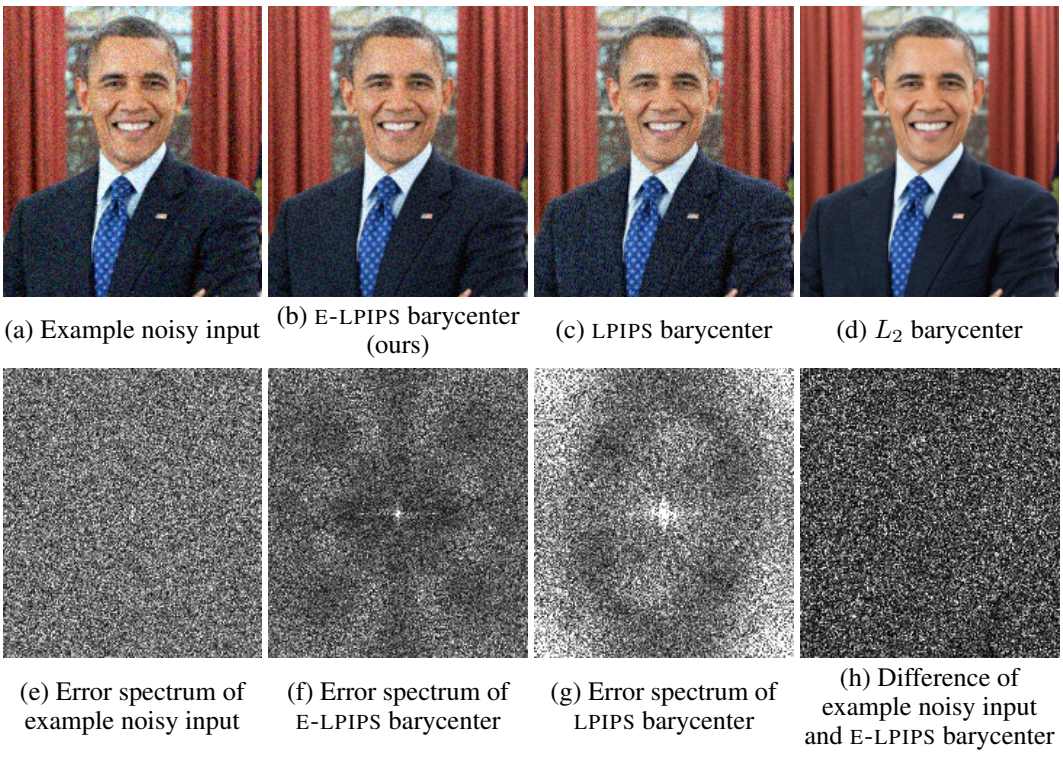
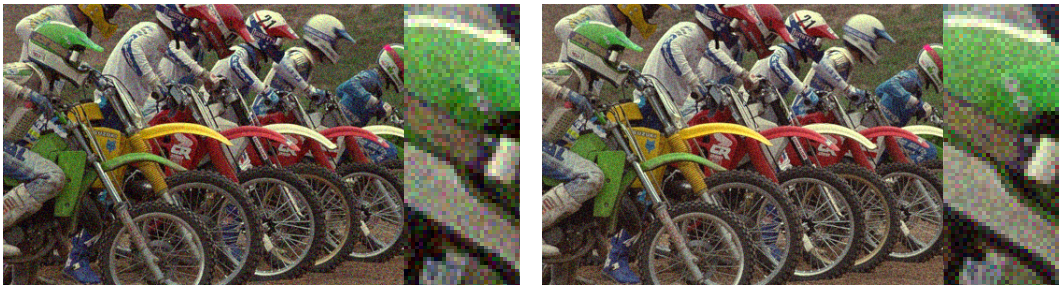


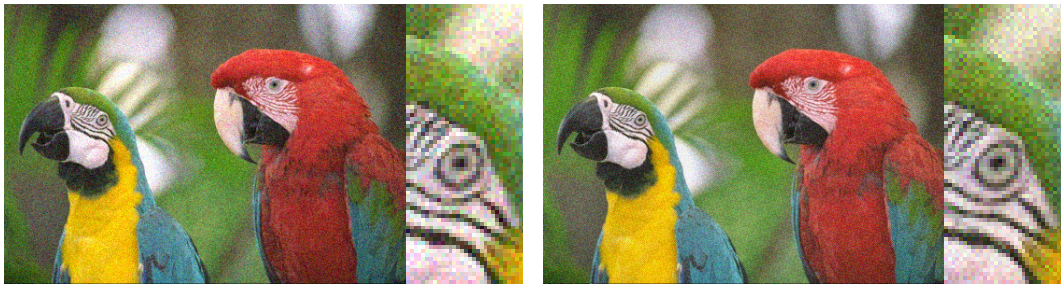
Figure 8: *Top row:* Barycenters of 10 random instances of an image corrupted with Gaussian noise. (a) One of the noisy inputs. (b) The E-LPIPS barycenter simultaneously resembles all noisy inputs with only a small decrease in noise magnitude. (c) The LPIPS barycenter features clear high frequency structure. (d) The L_2 barycenter is the mean of the noisy images and is much closer to the noise-free image. *Bottom row:* (e) The power spectrum of the noise in the inputs is white (uniform). (f) The power spectrum of the noise in the E-LPIPS barycenter differs slightly from uniform, so the noise in the pixels is not completely uncorrelated. (g) The LPIPS barycenter's high frequency structure is reflected in the error spectrum as strong high frequencies. (h) No clear patterns can be seen when comparing the noisy inputs and the E-LPIPS barycenter, indicating that the E-LPIPS barycenter is not directly copied from the inputs.



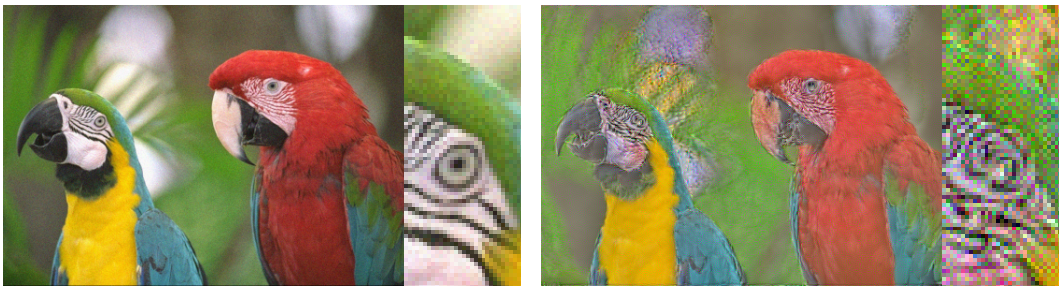
Example noisy input | E-LPIPS barycenter (ours)

L_2 barycenter | LPIPS barycenter

Figure 9: Barycenters of 10 random instances of noisy images. The E-LPIPS barycenter retains a relatively similar look to the noisy inputs. The L_2 barycenter differs by averaging out the noise. The LPIPS distance barycenter does not guarantee a meaningful image.



Example noisy input | E-LPIPS barycenter (ours)

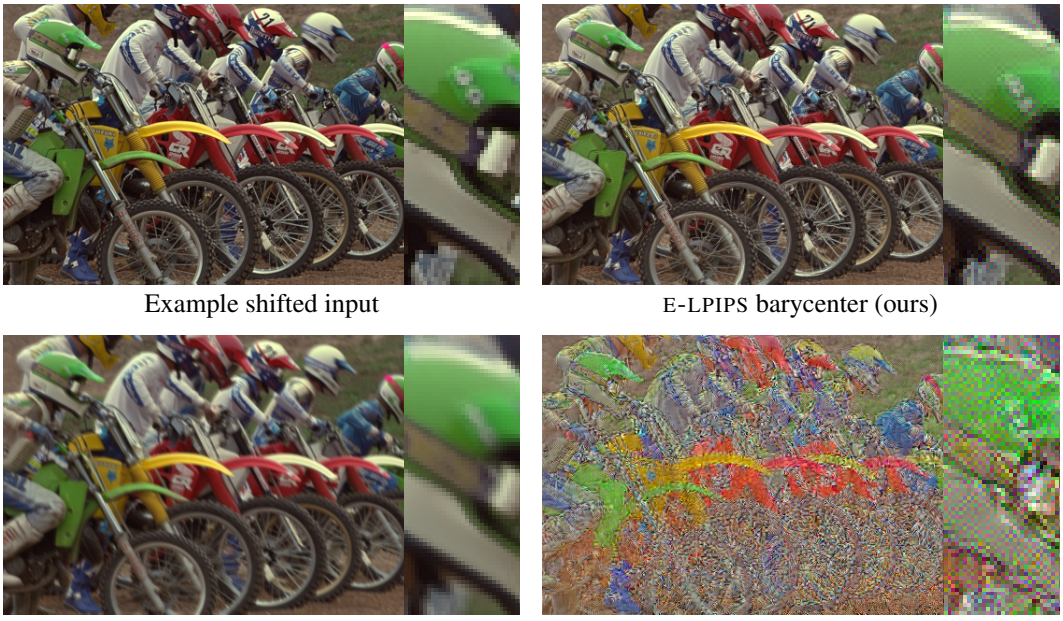


L_2 barycenter | LPIPS barycenter

Figure 10: Barycenters of 10 random instances of noisy images. The E-LPIPS barycenter retains a relatively similar look to the noisy inputs. The L_2 barycenter differs by averaging out the noise. The LPIPS distance barycenter does not guarantee a meaningful image.

(a) Example of shift (b) E-LPIPS barycenter (ours) (c) L_2 barycenter (d) LPIPS barycenter

Figure 11: Barycenter of 10 slightly shifted images. (a) One of the shifted inputs. (b) The E-LPIPS barycenter captures the essence of the inputs with only a small amount of loss of detail. (c) The L_2 barycenter is the pixelwise mean of the inputs, resulting in a blurry image. (d) The LPIPS distance barycenter does not guarantee a meaningful image.



Example shifted input E-LPIPS barycenter (ours)

L_2 barycenter LPIPS barycenter

Figure 12: Barycenters of 10 random instances of slightly shifted images. The E-LPIPS barycenter retains a relatively similar look to the shifted inputs. The L_2 barycenter results in a blurry image. The LPIPS distance barycenter does not guarantee a meaningful image.

	Traditional Metrics			Neural Metrics		
Error metric → Network loss ↓	L1 ($\times 10^2$)	L2 ($\times 10^3$)	Mean color L1 ($\times 10^3$)	LPIPS-SQZ ($\times 10^2$)	LPIPS-VGG ($\times 10^2$)	Ours ($\times 10^2$)
L2	2.72 ± 0.51	2.46 ± 0.77	0.72 ± 0.09	2.51 ± 0.30	2.97 ± 0.24	2.23 ± 0.36
LPIPS-SQZ	**3.79 ± 0.65**	**4.21 ± 1.24**	2.16 ± 0.33	1.13 ± 0.12	2.46 ± 0.17	1.66 ± 0.29
LPIPS-VGG	**3.48 ± 0.60**	**3.64 ± 1.08**	3.77 ± 0.86	**1.35 ± 0.15**	1.77 ± 0.13	**1.48 ± 0.23**
Ours	**3.40 ± 0.60**	**3.50 ± 1.06**	**1.98 ± 0.28**	**1.34 ± 0.13**	**2.02 ± 0.14**	1.32 ± 0.22

Table 3: Means of errors and two sigma confidence intervals for our super-resolution networks, averaged over the Kodak test set. Numbers whose training and result metrics correspond are grayed out and should not be considered for a fair comparison. The best numbers for each metric are colored dark green, and those that fit in its confidence interval are colored light green. Others are colored red. The numbers are averages of four snapshots of the network from a single training run.

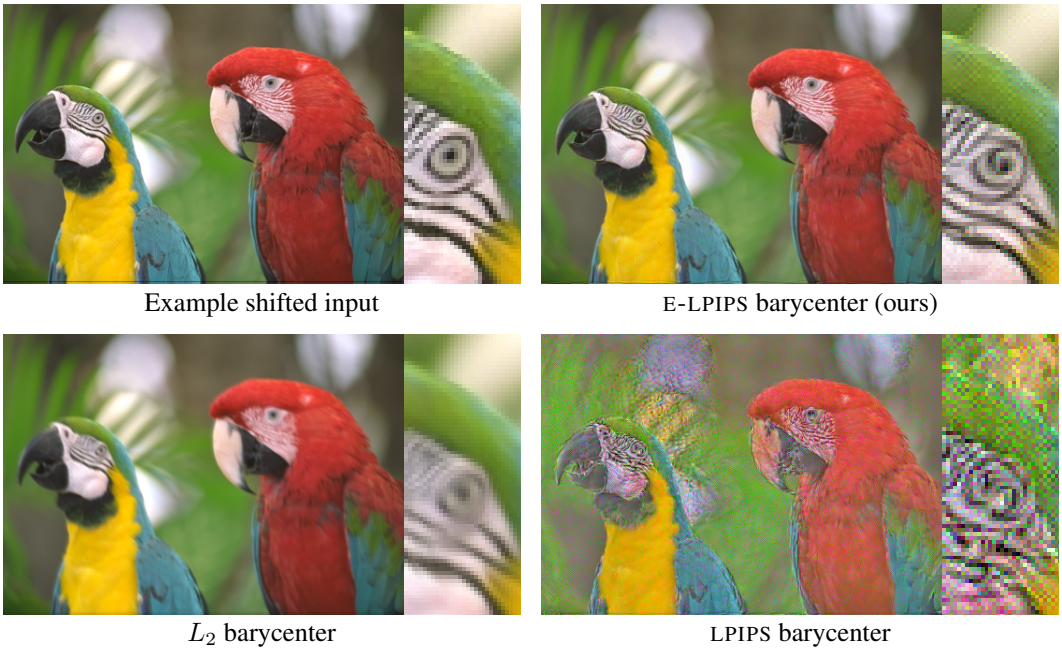


Figure 13: Barycenters of 10 random instances of slightly shifted images. The E-LPIPS barycenter retains a relatively similar look to the shifted inputs. The L_2 barycenter results in a blurry image. The LPIPS distance barycenter does not guarantee a meaningful image.

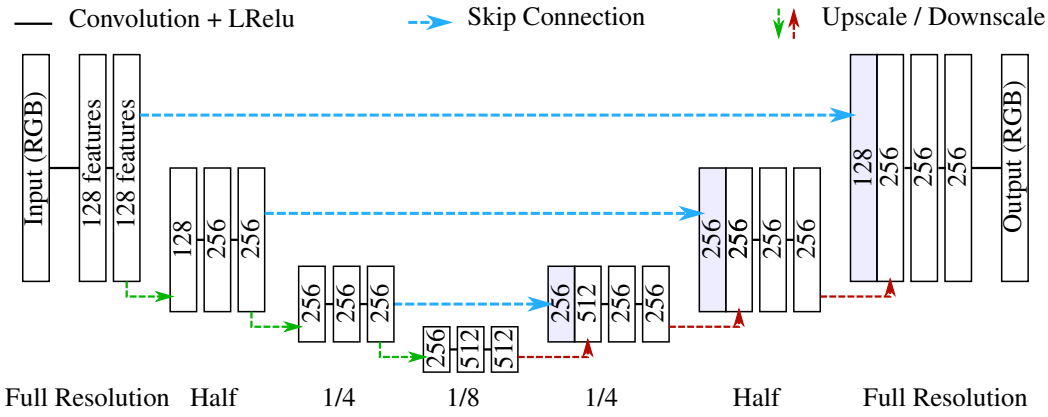


Figure 14: Architecture of our "U-net" [7] networks used in the denoising and super-resolution tests. The networks receive RGB images as input and repeatedly convolve them with 3x3 convolutions, upping the feature depth to 128 in the beginning and to 256 and 512 features in the subsequent resolutions. Each convolution is followed by a leaky ReLU with factor 0.01. At the halfway of each resolution the image is downscaled to half resolution with average pooling, and the same process repeats until 1/8-resolution. The result is then upscaled bilinearly, and the previous result of the higher resolution – before downsampling – is concatenated to the feature dimension. The result again goes through convolutional layers as described above, and the process continues until the end of the network. A final 1x1 convolution with linear activation transforms the result into RGB. For super-resolution the input given to the network is first upscaled to 4x size with the Lanczos filter. The total parameter count is 10 million.

Figure 15: Results from training a U-net for blind denoising using different loss functions. The reader is encouraged to look at the full-sized images included in the supplemental material.

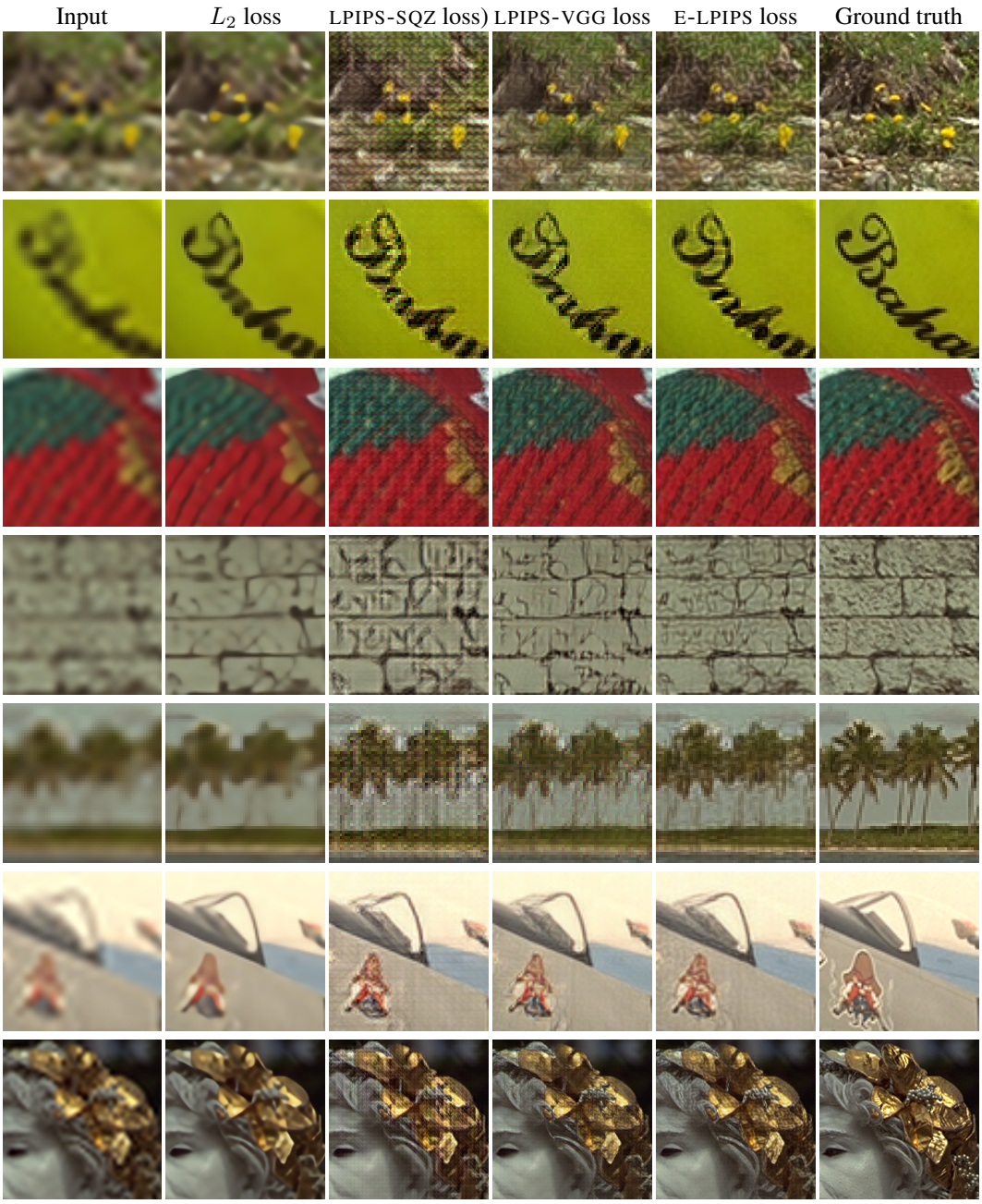


Figure 16: Results from training a U-net for 4-fold single-image super-resolution using different loss functions. The reader is encouraged to look at the full-sized images included in the supplemental material.

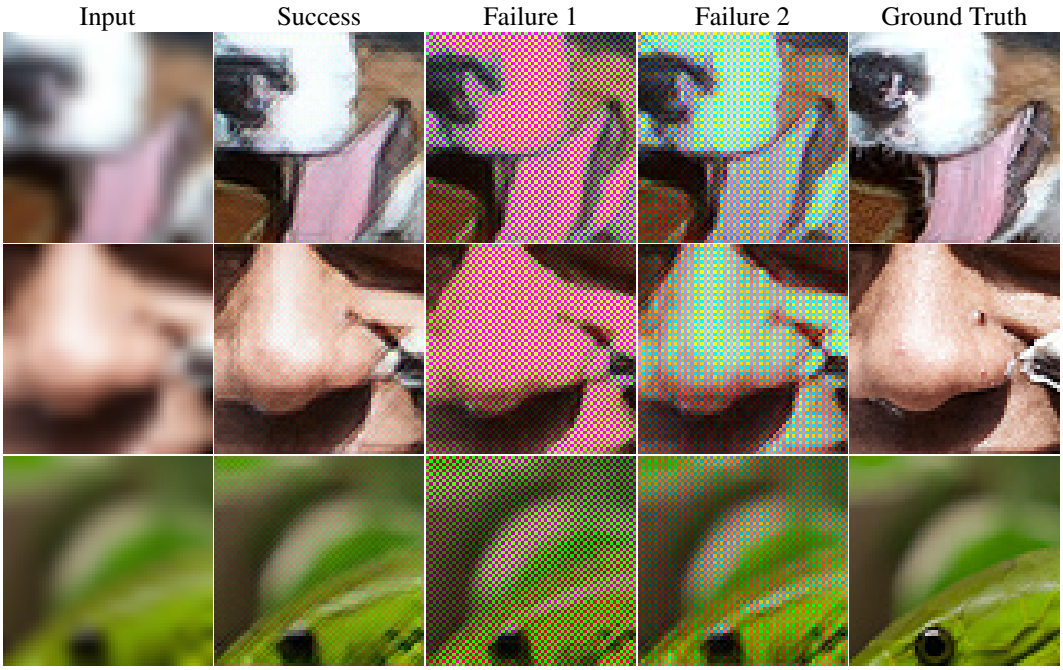


Figure 17: When training a $4\times$ super-resolution network with convolution transpose unpooling and LPIPS-VGG training loss, we find that over 10% of training runs converge to spurious local minima such as the ones shown above, forcing a restart. Our understanding is that the network learns an adversarial attack on LPIPS by tuning the pixel colors such that the activations are correct starting from the first layer that LPIPS actually studies. This does not happen with E-LPIPS. The LPIPS distances of the failures to the ground truth are equal to those of a successful training run.