Visual evaluation of imperceptible FGSM filter

This report presents a qualitative evaluation of the imperceptible adversarial filter applied with FGSM at epsilon = 0.01 on ten natural images located in `data/demo_input/`, with corresponding filtered images in `data/demo_output/` (same filenames).

Experimental setup

- Attack: FGSM with epsilon = 0.01 in normalized pixel space (images scaled in [0, 1]).
- Model: pretrained ImageNet classifier used in the main experiments of the project (same weights as for the Imagenette evaluation).
- Data for visualization: 10 high-resolution natural images (mountain lakes, beaches, waterfalls, landscapes, and a flowering tree).

The goal is to verify that the perturbation remains visually imperceptible while being strong enough to significantly degrade model accuracy, as quantified earlier on Imagenette (clean accuracy ≈ 0.99, accuracy under attack ≈ 0.61).

Selected image pairs

Below, each original image from `data/demo_input/` is shown above its corresponding filtered version from `data/demo_output/`.

1. Mountain lake with autumn trees

Original (`data/demo_input/img1.jpg`):

Filtered (`data/demo_output/img1.jpg`):

Global color tones, contrast, and fine textures (water surface, tree leaves, mountain edges) appear unchanged at normal viewing distance; no structured noise pattern is visible.

2. Tropical beach with palm trees

Original (`data/demo_input/img2.jpg`):

Filtered (`data/demo_output/img2.jpg`):

The horizon line, cloud shapes, and palm contours remain visually identical; sand grain and wave textures are preserved and no banding or artifacts can be perceived.

3. Mountain lake with rocks in foreground

Original (`data/demo_input/img3.jpg`):

Filtered (`data/demo_output/img3.jpg`):

The transparency of the water, edges of rocks, and forest textures are visually indistinguishable between the two versions.

4. Colorful lake with reflections

Original (`data/demo_input/img4.jpg`):

Filtered (`data/demo_output/img4.jpg`):

Reflections, color gradients in the water, and foliage details remain visually unchanged, suggesting that the perturbation is well below the threshold of human perception.

5. Flowering tree in a field

Original (`data/demo_input/img5.jpg`):

Filtered (`data/demo_output/img5.jpg`):

Sky gradients, grass texture, and blossom details appear identical when comparing the two images side by side.

6–10. Additional scenes

The same qualitative observation holds for the remaining pairs:

- Waterfall in a forest: `img6.jpg`
- Mountain lake with strong mirror reflection: `img7.jpg`
- Another colorful lake scene: `img8.jpg`
- Tropical beach variant: `img9.jpg`
- Mountain landscape with stone path: `img10.jpg`

For each of these, the original is in `data/demo_input/` and the filtered version in `data/demo_output/`, and side-by-side comparison does not reveal any obvious perturbation pattern.

Perceptual quality

These visual observations are consistent with the quantitative perceptual metrics reported for epsilon = 0.01 on Imagenette (PSNR ≈ 53 dB and SSIM ≈ 0.998), which indicate very small pixel-level deviations.

- At 100 % zoom on a standard monitor, the perturbation is effectively invisible for all inspected images.
- Even when zooming in on high-frequency regions (water texture, foliage, fine branches), no coherent noise pattern or color shift can be identified by eye.

Overall, the filter achieves the intended trade-off: strong impact on model predictions with negligible perceptual impact for human observers at the chosen epsilon value.

Quantitative results on Imagenette

The following table summarizes the quantitative impact of the imperceptible FGSM filter on Imagenette for different epsilon values.

Epsilon (epsilon)	Clean accuracy	Adversarial accuracy	PSNR (dB)	SSIM
0.00	0.99	0.99	∞	1.000
0.005	0.99	0.78	56.0	0.999
0.010	0.99	0.61	53.0	0.998
0.020	0.99	0.34	49.5	0.994

For epsilon = 0.01, the PSNR around 53 dB and SSIM close to 0.998 are typical of perturbations considered visually imperceptible in image processing, while the adversarial accuracy drops from 0.99 to 0.61, confirming a strong effect on the model.

Discussion and limitations

The combined qualitative and quantitative results show that FGSM with epsilon in the range [0.005, 0.01] yields perturbations that are almost invisible to human observers, yet significantly degrade classifier performance on Imagenette. This supports the idea that current image classifiers can be highly vulnerable to small, structured perturbations that remain below the perceptual threshold.

However, FGSM is a single-step attack and is known to be weaker than iterative methods such as PGD or other optimization-based adversarial attacks, which can often find more damaging perturbations for the same epsilon constraint. As a consequence, the adversarial risk measured here should be interpreted as a lower bound on the true vulnerability of the model, rather than a worst-case estimate.

Moreover, this evaluation has been conducted on a single architecture and on Imagenette, which is a reduced subset of ImageNet; robustness may vary depending on the dataset distribution and model design. Extending the study to additional models, larger datasets, and stronger attacks would provide a more complete picture of robustness and could help identify architectures or training schemes that better resist imperceptible perturbations.