Interactive Analysis of CNN Robustness
— Supplemental Document —

Stefan Sietzen*, Mathias Lechner, Judy Borowski,
Ramin Hasani, and Manuela Waldner

A Extended and Additional Exploration Scenarios

In addition to the exploration scenario shown in the main paper, we show here results of the following exploratory analyses:

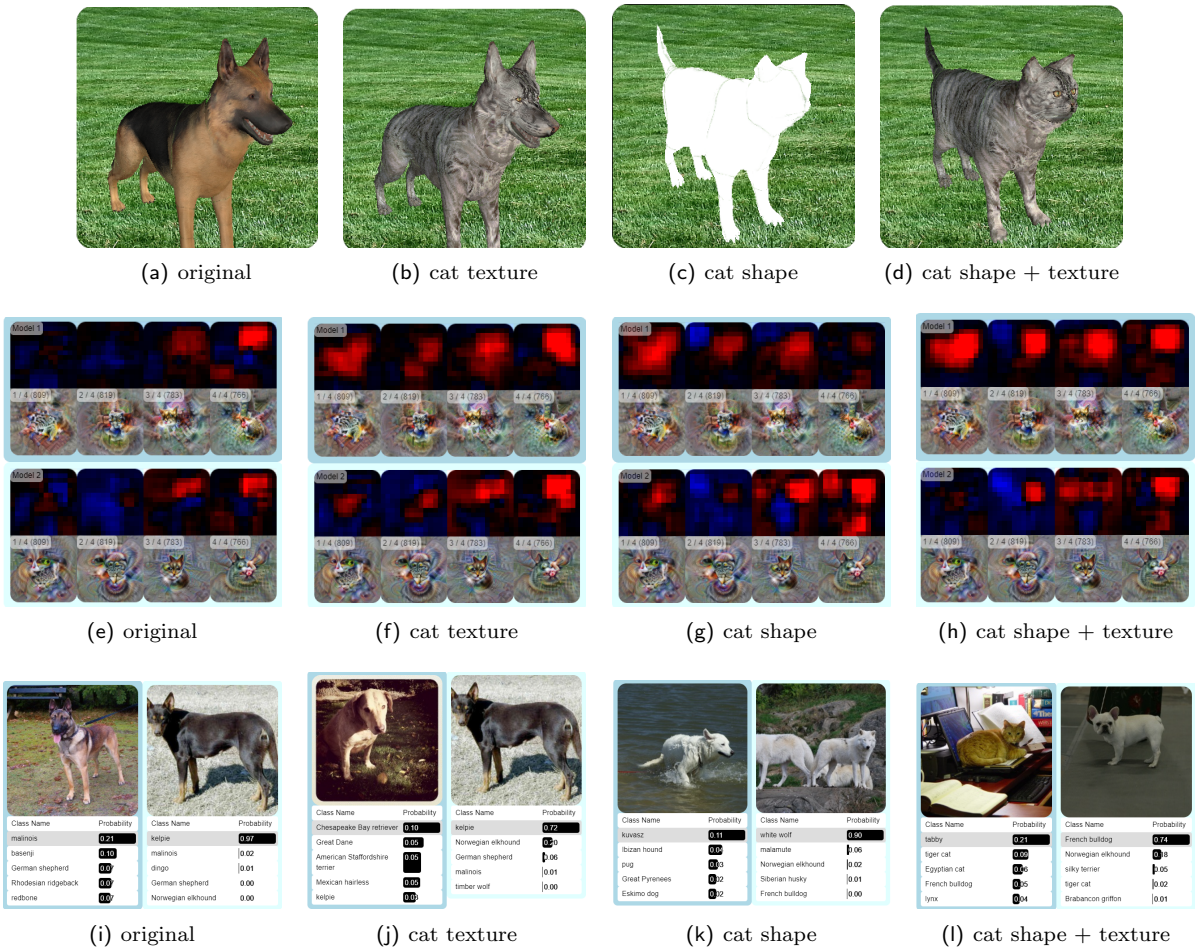
- texture influence (Section A.1 and Section A.2),
- shape sensitivity (Section A.3),
- low frequency information (Section A.4),
- high frequency information (Section A.5),
- adversarial attacks (Section A.6),
- fading to black (Section A.7),
- geometric transformations (Section A.8),
- as well as geometric transformations in combination with background modifications (Section A.9).
- development of activations & feature visualizations during fine-tuning (Section A.10).

For the respective scenarios, we compare the standard model (Inception-V1 trained on ImageNet) with the Stylized-ImageNet trained model (Inception-V1 fine-tuned with Stylized-ImageNet [1]) in Sections A.1, A.2, A.3, A.4, A.5 and with the adversarially trained model (Inception-V1 adversarially fine-tuned [2, 5]) in Section A.6, A.7, A.8, A.9. Finally, we show the development of feature visualizations during fine-tuning for both models in Section A.10.

*stefan.sietzen@gmx.at

A.1 Texture vs. Shape

Figure A.1 shows a comparison between the standard trained model and the model trained by Stylized-ImageNet. The middle row shows how cat-related neurons get activated by morphing the texture and shape, respectively. Please note how the standard model (top row in Figure A.1 e-h) gets strongly activated by the cat texture, while the respective neurons of the Stylized-ImageNet trained model (bottom row in Figure A.1 e-h) seem to get more activated by shape changes. Also note that the Stylized-ImageNet trained model never predicts a cat.



(a) original (b) cat texture (c) cat shape (d) cat shape + texture

(e) original (f) cat texture (g) cat shape (h) cat shape + texture

(i) original (j) cat texture (k) cat shape (l) cat shape + texture

Figure A.1: Object morphing to assess the texture vs. shape conflict: input scene (a-d), activations of four cat-related neurons in mixed4e (e-h) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (i-l) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.2 Patch Shuffling

Figure A.2 shows the effect of randomly shuffling image patches on the standard and the Stylized-ImageNet trained model. The dog is still predicted correctly by both models when shuffling 2×2 image patches (Figure A.2 g). The Stylized-ImageNet trained model is more sensitive to patch shuffling than the standard model (Figure A.2 h-i). Note how the activations of neuron 429 in layer mixed4b (third column in Figure A.2 d-f) follow certain regions in the dog face.

(a) $k = 2$ (b) $k = 5$ (c) $k = 8$

(d) $k = 2$ (e) $k = 5$ (f) $k = 8$

(g) $k = 2$ (h) $k = 5$ (i) $k = 8$

Figure A.2: Patch shuffling: increasing the number of randomly shuffled image patches k (a-c), activations of dog-relevant neurons in mixed4b (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.3 Silhouette

Figure A.3 investigates the models' shape sensitivity by analyzing the dog's silhouette against a red background in different poses. The activations of dog-related neurons in mixed4d (Figure A.3 e-h) show that the standard model seems to be more sensitive to pose changes. Indeed, the predictions (Figure A.3 i-l) of the standard model fluctuate with the pose modifications, while the Stylized-ImageNet trained model consistently predicts "white wolf" with a very high probability.

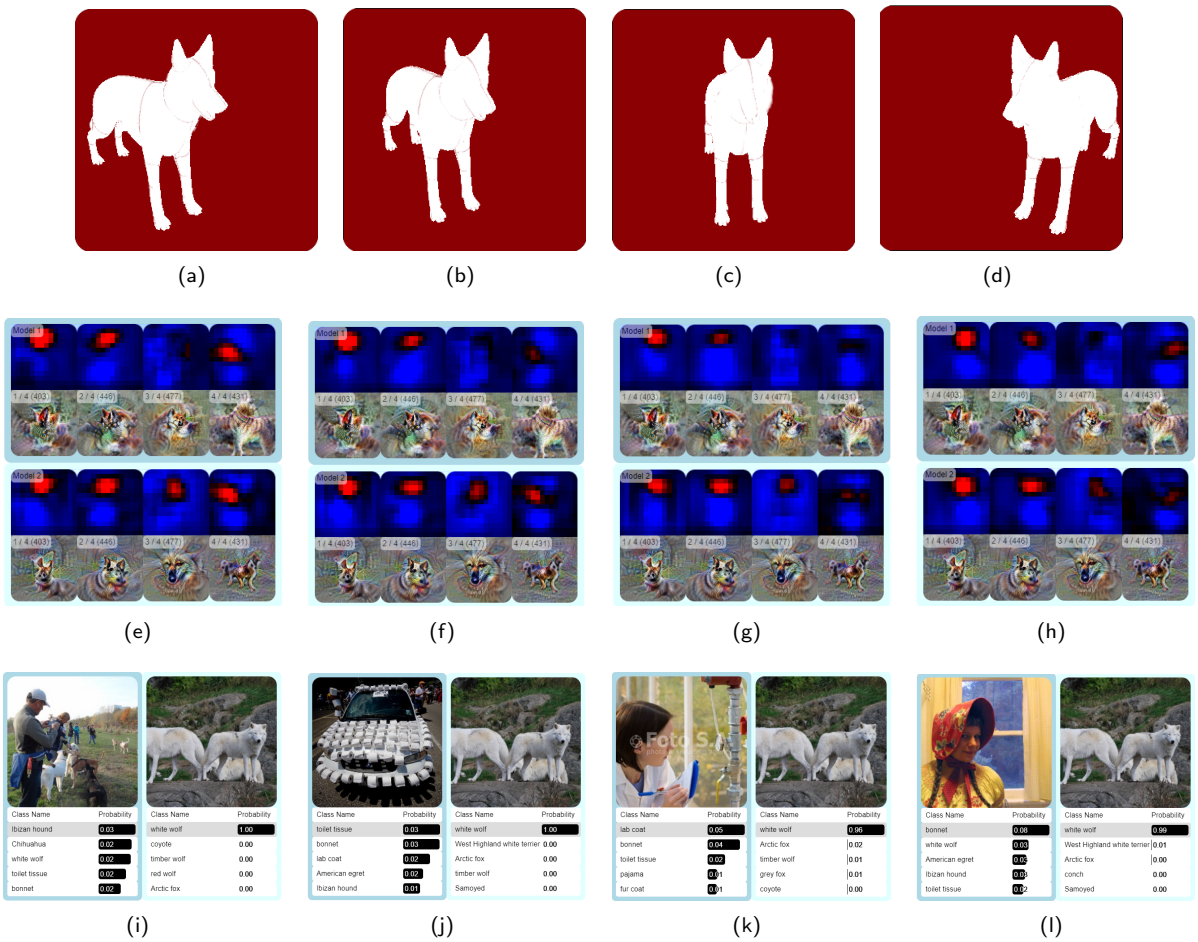


Figure A.3: Analyzing shape influence: different silhouette poses as input (a-d), activations of four dog-related neurons in mixed4d (e-h) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (i-l) by the standard model (left) and the Stylized-ImageNet trained model (right).

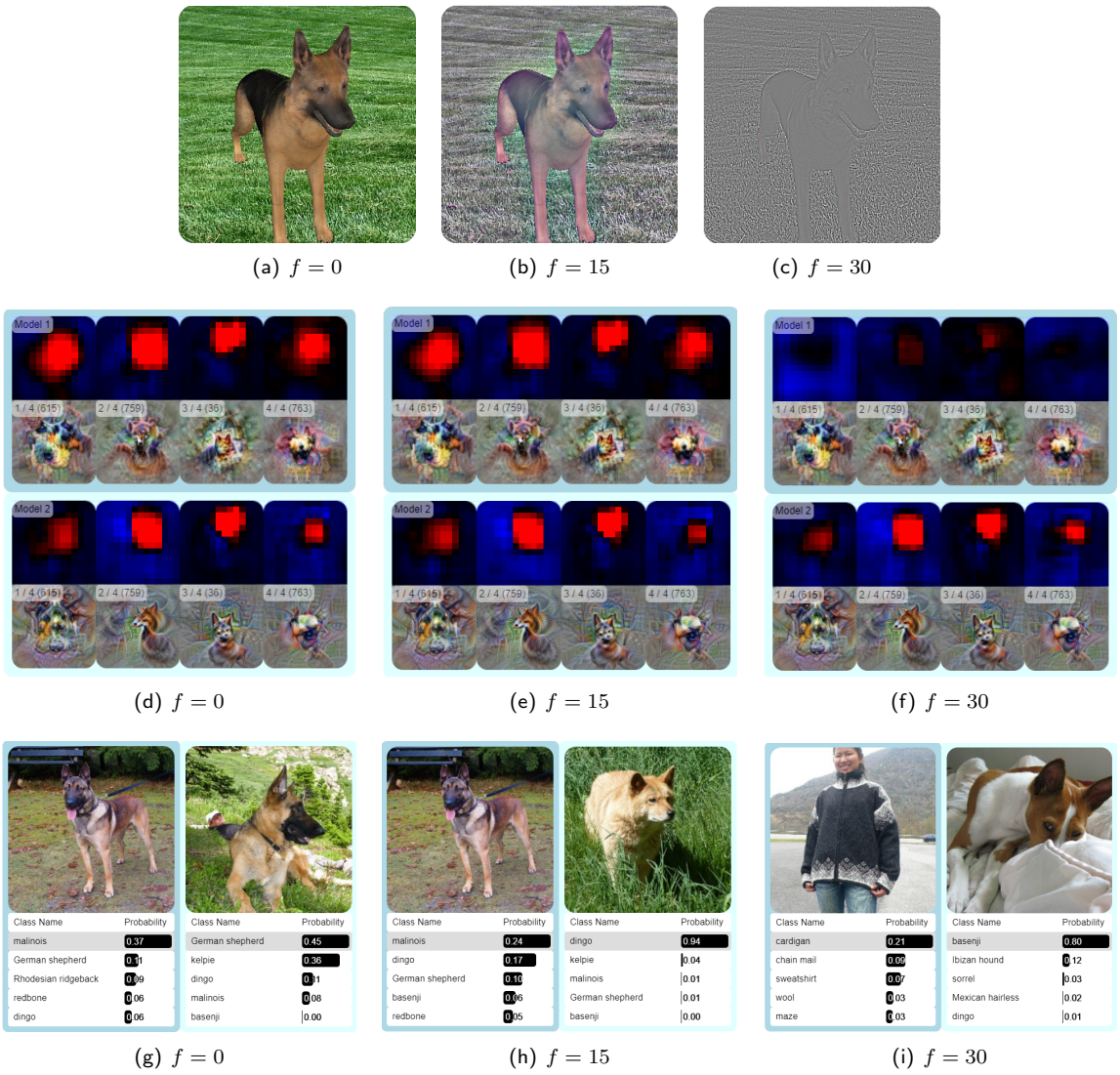
A.4 Blur

In the case study, a suspicion of one user was that the model trained on Stylized-ImageNet would be more sensitive to high-frequency information. Blurring the image would therefore disturb this model more heavily than the standard model. Figure A.4 illustrates that this is not necessarily the case: Activations of dog-related neurons gradually degrade by applying blur for both models (Figure A.4 d-f). At a high blur level, both models have very uncertain predictions (Figure A.4 i).

Figure A.4: Blurring the image: input image with gradual blur (a-c), activations of oriented dog heads in mixed4a (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.5 High-Pass Filtering

In contrast to low-pass filtering shown in the previous section, here we show the effects of high-pass filtering on the standard trained model and the Stylized-ImageNet trained model (Figure A.5). As the frequency threshold is increased to a high level, the activations of dog-related neurons considerably decrease for the standard model (Figure A.5 f, top row), while the same neurons are still highyly activated for the Stylized-ImageNet trained model (Figure A.5 f, bottom row), and it also still predicts a canine (Figure A.5 i, right column).



(a) $f = 0$ (b) $f = 15$ (c) $f = 30$

(d) $f = 0$ (e) $f = 15$ (f) $f = 30$

(g) $f = 0$ (h) $f = 15$ (i) $f = 30$

Figure A.5: Applying a high-pass filter on the input image: input image with increasing frequency threshold (a-c), activations of oriented dog-related neurons in mixed4e (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

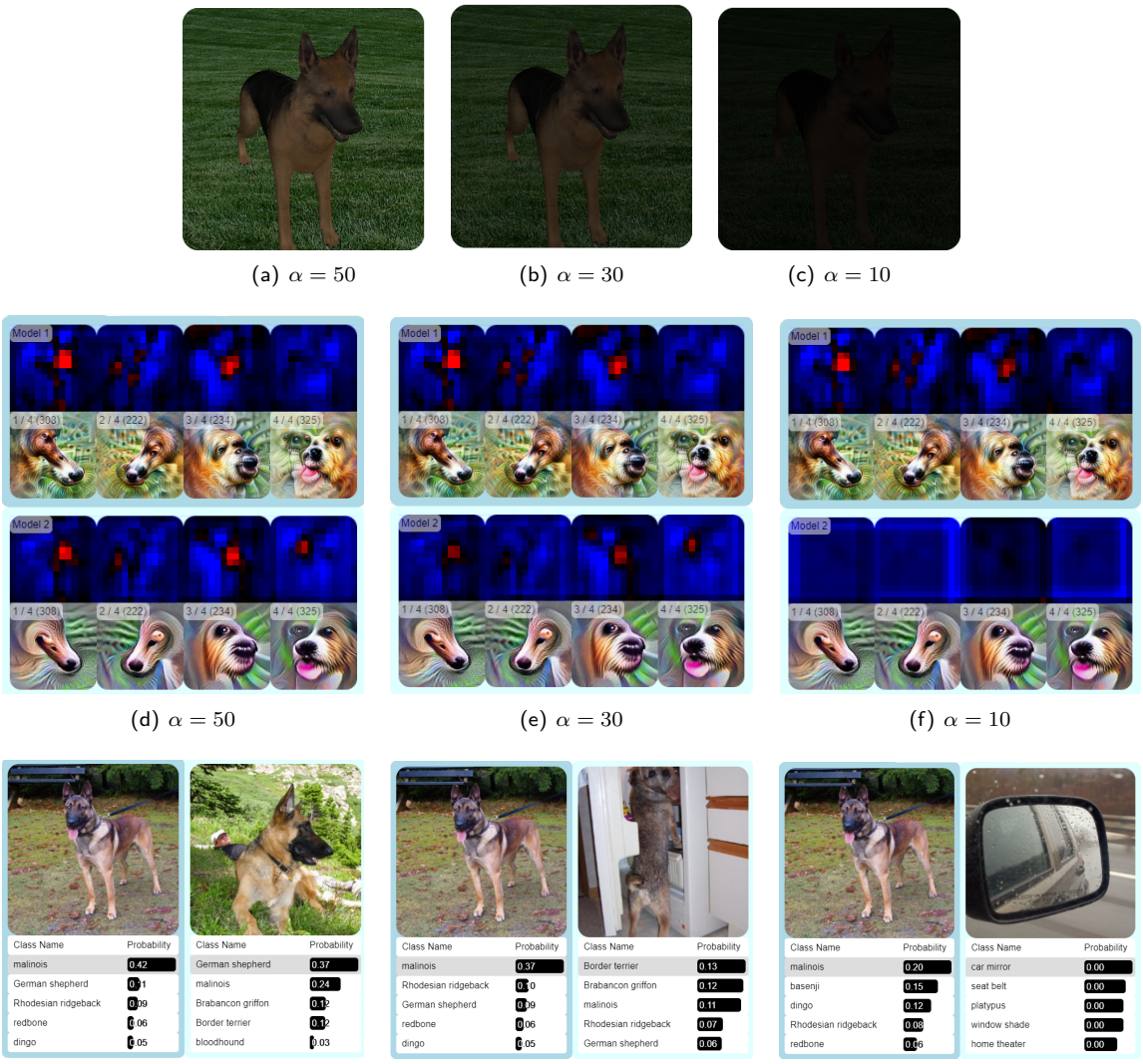
A.6 Adversarial Attack

Figure A.6 shows an adversarial attack (target class "Egyptian cat") for a standard and an adversarially trained model. Figure A.6 c-d, bottom row shows how the activations of the adversarially trained model remain unaffected, while the cat-related neurons get strongly activated by the attack for the standard model (Figure A.6 c-d, top row). Consequently, the standard model's prediction switches to the attack's target class (Figure A.6 f, left column), while the adversarially trained model still predicts a German shepherd with very high confidence.

Figure A.6: Adversarial attack with target class "Egyptian cat": input image before (a) and after (b) a successful attack, activations of cat-related neurons (c,d) by the standard model (top) and the adversarially trained model (bottom), and the predictions (e,f) by the standard model (left) and the adversarially trained model (right).

A.7 Alpha

Figure A.7 illustrates the *saturation* effect [3] by gradually blending the input image to black. Note how the activations of the standard model (Figure A.7 d-f, top row) remain high and the predictions (Figure A.7 g-i, left column) remain correct even though the α is already very low on the last image so that humans can no longer perceive any content. The adversarially trained model, however, is more sensitive to the reduced image contrast.



(a) $\alpha = 50$ (b) $\alpha = 30$ (c) $\alpha = 10$

(d) $\alpha = 50$ (e) $\alpha = 30$ (f) $\alpha = 10$

(g) $\alpha = 50$ (h) $\alpha = 30$ (i) $\alpha = 10$

Figure A.7: Reducing the overall alpha and blending into black: input image with different alpha values (a-c), activations of oriented dog heads in mixed4a (d-f) by the standard model (top) and the adversarially trained model (bottom), and the predictions (g-i) by the standard model (left) and the adversarially trained model (right).

A.8 Rotation

Figure A.8 compares two models' sensitivities to viewpoint changes. From a side view (Figure A.8 a-b), both models fairly reliably predict a race car (Figure A.8 i-j). However, when looking at the car from a front-top view (Figure A.8 c), predictions are getting unstable for both models (Figure A.8 k) – in particular for the adversarially trained model, which does not predict any car-like object as top-5 target (Figure A.8 k, right). Looking at the activations of neurons that are important for the prediction of race cars in mixed4b (Figure A.8 e-h), it seems that wheels play a very important role. As the car is rotated and wheels disappear, the activations of these neurons decrease considerably (Figure A.8 g-h).

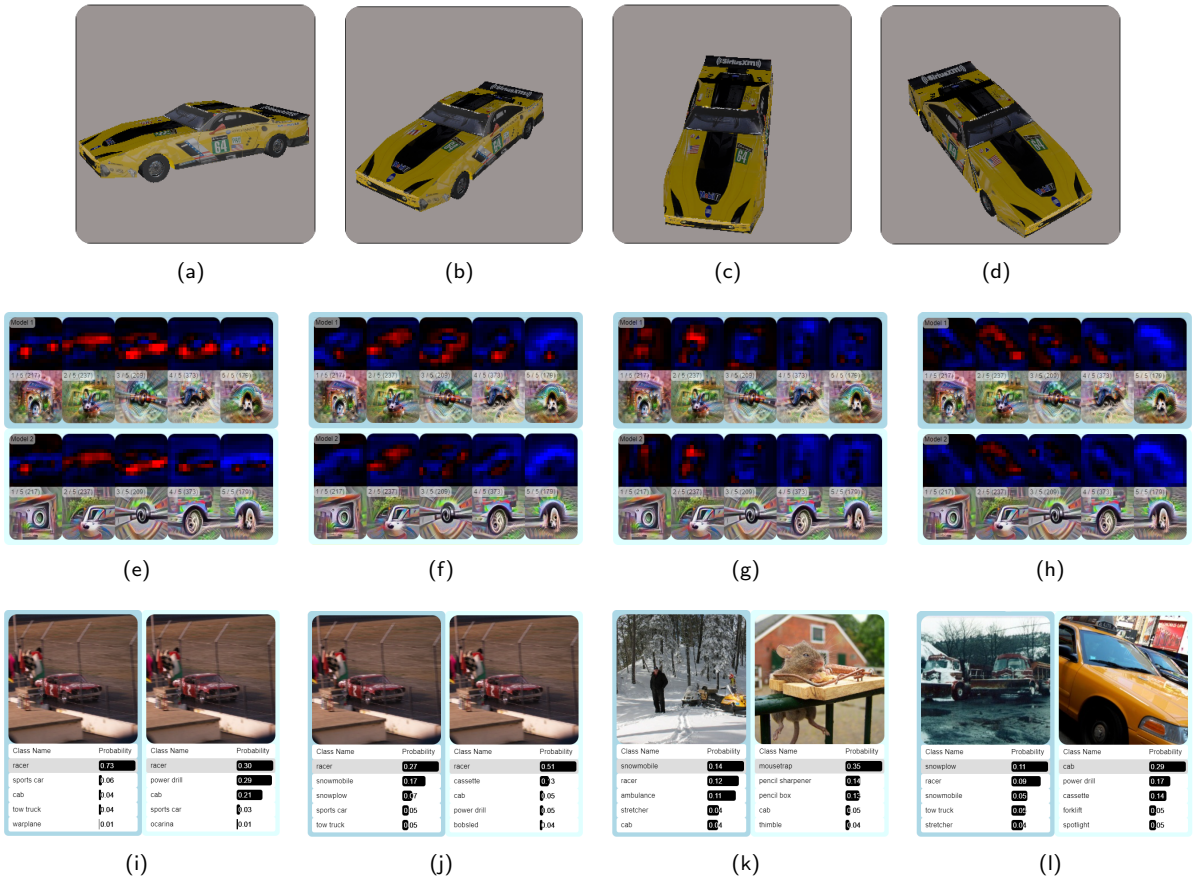
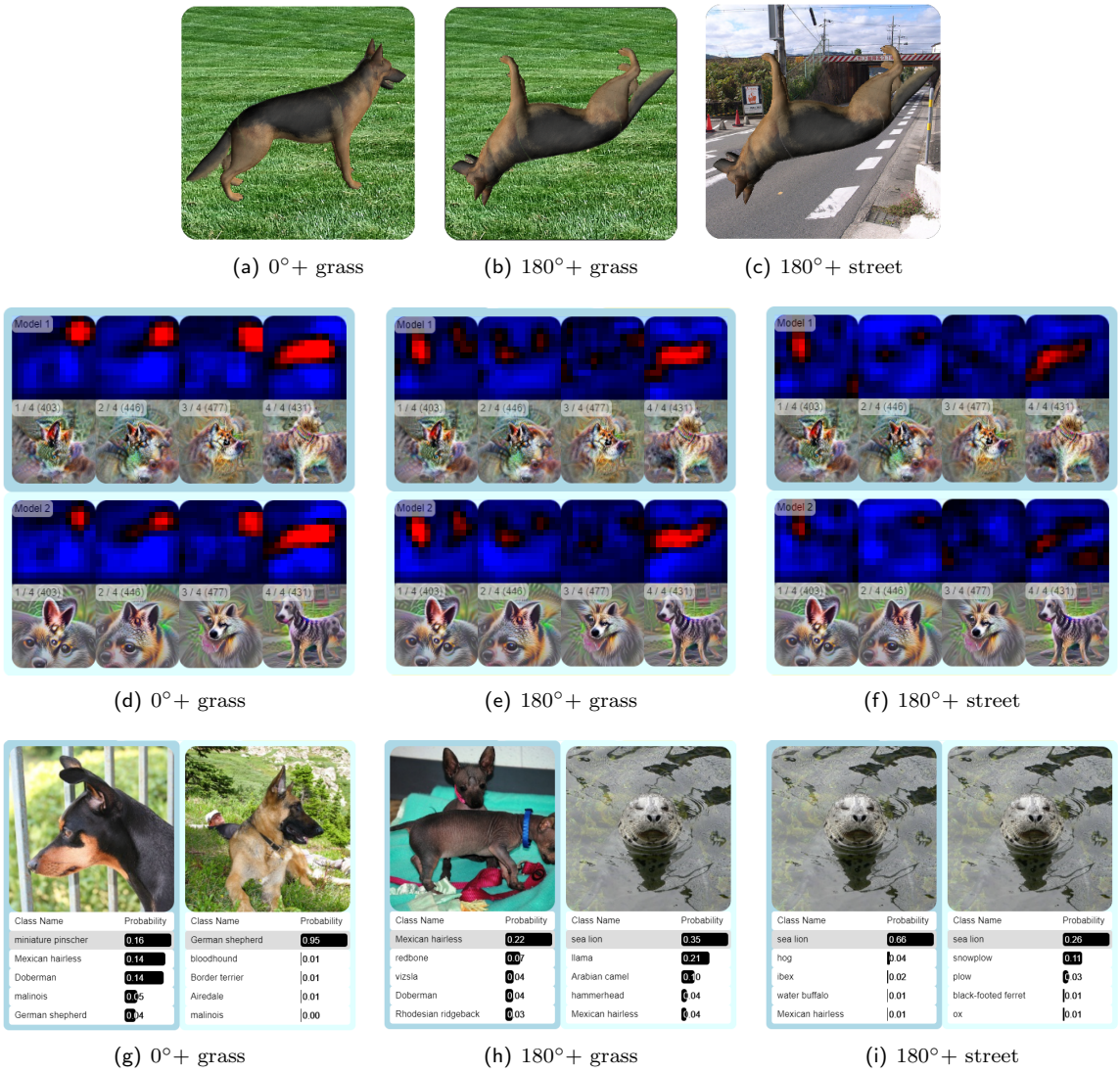


Figure A.8: Orbiting around the race car: input scene from different camera angles (a-d), activations of four race car-related neurons in mixed4b (e-h) by the standard model (top) and the adversarially trained model (bottom), and the predictions (i-l) by the standard model (left) and the adversarially trained model (right).

A.9 Roll + Background

To assess the sensitivity to the context, we analyze the influence of the main object's rotation *and* the background image in Figure A.9. While the standard model still predicts dog breeds after a 180°rotation (Figure A.9 h, left column), the adversarially trained model has a tendency to predict sea animals (Figure A.9 h, right column). After changing the background, none of the models predicts a dog (Figure A.9 i). The adversarially trained model also has artifacts in the top-5 in case of a street background (Figure A.9 i, right column). Also, note how the dog-related activations decrease once the object is rotated (Figure A.9 e) and the background is swapped (Figure A.9 f) for both models.



(a) 0°+ grass (b) 180°+ grass (c) 180°+ street

(d) 0°+ grass (e) 180°+ grass (f) 180°+ street

(g) 0°+ grass (h) 180°+ grass (i) 180°+ street

Figure A.9: Rotating the main object and changing the background: input image with different rotations and / or background images (a-c), activations of dog-related neurons in mixed4d (d-f) by the standard model (top) and the adversarially trained model (bottom), and the predictions (g-i) by the standard model (left) and the adversarially trained model (right).

A.10 Model Fine-Tuning

To better understand what happens during fine-tuning, users can compare models at various intermediate checkpoints of the fine-tuning process. This is similar to the transfer learning visualizations by Szabo et al. [4]. However, they investigated transfer learning with different datasets, while we fine-tuned the models with variants of the same dataset.

Perturber provides 17 selected checkpoints during adversarial fine-tuning and seven selected checkpoints during Stylized-ImageNet fine-tuning. We chose to include more checkpoints for the adversarial fine-tuning because it appears more dynamic compared to the Stylized-ImageNet fine-tuning (as can be seen in Figure A.10), and generating the required data is computationally expensive.

Figure A.10 shows activations and feature visualizations of neuron 222 in layer mixed4a at various fine-tuning checkpoints. During Stylized-ImageNet fine-tuning, the activations and feature visualizations stay relatively consistent. During intermediate steps of adversarial fine-tuning, however, the positive response vanishes before reappearing after iteration 10K. The corresponding feature visualizations also reflect this phenomenon by becoming less similar to a dog head intermediately before assuming the appearance of a smoother version of a dog head than before fine-tuning.

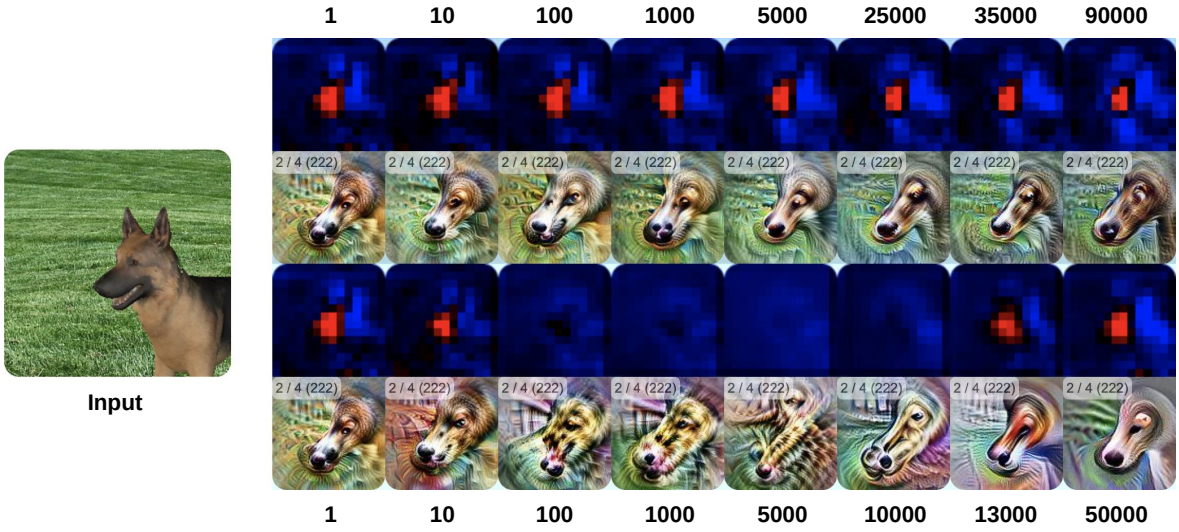


Figure A.10: Activations and feature visualizations of neuron 222 in layer mixed4a at selected fine-tuning checkpoints. Numbers above and below show the fine-tuning iteration.

B Case Study Observations & Feedback

Here, we list all observations and feedback recorded during the case studies. Table B.1 shows the research focus of the individual study participants.

Table B.2 lists all reported observations. Some of these observations are visual confirmations of known facts, some observations are highly speculative, some are just descriptions of what the users saw.

In Table B.3, finally, we list all suggestions for future improvements mentioned by the users.

Table B.1: Research focus of the case study participants.

User	Research Focus
P1	Understanding vision in humans and machines, with a special focus on Deep Learning interpretability and feature visualizations.
P2	On the interface between psychophysics and deep learning, in particular understanding how object recognition differs between humans and machines.
P3	Detection and interpretation of failure cases of computer vision models.
P4	Learning more robust, safe, and verifiable machine learning models.
P5	Designing interpretable deep learning models.

Table B.2: Observations reported by the participants in our case studies, including references to corresponding exemplary scenarios.

User	Observation	Reference
Geometric Perturbations		
P3	The adversarially trained model is robust to translation when the dog is viewed from the side, while the standard model fluctuates.	
P3	Zooming into the race car makes the Stylized-ImageNet trained model predict a school bus, which is incorrect.	
P4	Rotation affects the class output of the adversarially trained model more than that of the standard model.	Section A.8
P5	The adversarially trained model tends to misclassify the scene more often upon viewpoint changes than the standard model.	Section A.8
P5	The adversarially trained model seems to be less sensitive to object distance (zoom).	
Scene Perturbations		
P3	Background blur makes the adversarially trained model less consistent. The adversarially trained model seems to use the background more than the standard model.	Section 6.2 (main paper)
P5	Background significantly alters the decisions made by the adversarially trained model. This is less apparent for the standard model.	Section 6.2 (main paper)
Object Morphing		
P1	The cat is predicted surprisingly "late".	Section A.1
P2	Predictions first change to another dog class before they switch to a cat.	Section A.1
P3	Activations for dog-related neurons do not necessarily peak at "pure" dog images.	
Frequency Decomposition		
P1	The strong influence of frequency decomposition operations on the class predictions is surprising.	
P2	The Stylized-ImageNet trained model is more robust under blur than expected.	Section A.4
Patch Shuffling		
P2	The target class is soon difficult to predict for a human, but it is still correctly predicted by the model	Section A.2
Adversarial Attacks		
P2	Adversarial attacks only affect the standard model.	Section A.6
P4	Adversarial attacks change the activations of the early layers very little. The activations seem to change more on the later layers.	
P4	An adversarial attack on a car scene towards "badger" leads to fur neurons getting activated.	
Complex Perturbations		
P2	A small rotation in combination with an unusual background is sufficient to disturb the adversarially trained model.	Section A.9
P3	The untextured dog head can be quite certainly predicted by the adversarially trained model, but leads to a hammerhead prediction upon close-up for the standard model. Texture makes predictions more certain, but blurred texture (coloring) also helps.	
P3	Rotation has a strong influence in combination with low texture influence and zooming.	
Feature Visualizations & Activation Maps		
P1, P3	Feature visualizations of the adversarially trained model look more "intuitive" / "cartoonish".	
P2	Eye detectors react to surprisingly many regions in the street background image.	
P4	The lack of differences of activation maps between models is suprising. The only major differences were observable during adversarial attacks.	

Table B.3: Suggestions for improvement provided by the participants of the study.

User	Suggestion for Improvement
Scene Perturbations	
P1	Allow users to upload custom background images.
P2	Support background rotation.
P3	Object texture could be more detailed.
Adversarial Attacks	
P4	Change the scene behind the adversarial attack instead of adversarial attack being tied to a fixed image.
Feature Visualizations	
P1	Show dataset examples (i.e., strongly activating examples from the training data) instead of / in addition to feature visualizations.
P3	Support different feature visualizations.
General Suggestions	
P1	Provide more guidance through the interface, otherwise it can be overwhelming.
P1, P3	Provide more 3D models.
P3	Show dataset examples with similar activations as the current input scene.
P5	Perform a grid search to systematically generate input images and store the results for quantitative evaluation.

References

[1] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *International Conference on Learning Representations (ICLR)* (2019).

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *arXiv:1706.06083 [cs, stat]* (Sept. 2019). arXiv: 1706.06083.

[3] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. "Visualizing the impact of feature attribution baselines". In: *Distill* 5.1 (2020), e22.

[4] Róbert Szabó, Dániel Katona, Márton Csillag, Adrián Csiszárik, and Dániel Varga. "Visualizing Transfer Learning". In: *arXiv:2007.07628 [cs]* (July 2020). arXiv: 2007.07628.

[5] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy". en. In: *arXiv:1805.12152 [cs, stat]* (Sept. 2019). arXiv: 1805.12152.