Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks -Supplementary Material-

Jörg Wagner[1,2] Jan Mathias Köhler[1] Tobias Gindele[1,*] Leon Hetzel[1,*]
Jakob Thaddäus Wiedemer[1,*] Sven Behnke[2]
[1]Bosch Center for Artificial Intelligence (BCAI), Germany [2]University of Bonn, Germany
Joerg.Wagner3@de.bosch.com; behnke@cs.uni-bonn.de

The supplementary material provides details, additional results, and further comparisons.

A1. Defending against Adversarial Evidence

Our method produces explanations based on evidence in the image and suppresses hallucination of adversarial evidence. Without our adversarial defense the optimization can produce an explanation for any class (*i.e.* even for a class visually not present in the image).

To illustrate this differently to the experiment reported in Sec. 3.1 (Tab. 1 and Fig. 3), we show an alternative version of the evaluation, only using a black image as input. Fig. A1 shows an explanation for the adversarial class *iguana* with and without defense. For Tab. A1 we create explanations for each of the 998 ImageNet classes, using always the same black input image. We omit the predicted class of the black image and the class of the starting condition (image $\cdot$ zero mask). Without defense an explanation can always be generated due to hallucination of adversarial evidence. The results are comparable to the evaluation in the main paper.

A2. Implementation Details

Unless otherwise specified, the explanations are computed for the most-likely class using SGD with a learning rate of 0.1, running for 500 iterations. To improve optimization and avoid instabilities, we initialize the masks $\mathbf{m}$ with noise sampled for each pixel from a uniform distribution $\mathcal{U}(a, b)$. with $\mathcal{U}(0, 0.01)$ for the *generation* and *repression game* and $\mathcal{U}(0.99, 1)$ for the *preservation* and *deletion game*. We normalize the gradient using its maximum value to avoid large changes of individual mask pixels.

For the similarity metric $\varphi(\cdot, \cdot)$ we use the cross-entropy for the *generation* and *preservation game* and the negative probability for the *deletion* and *repression game*.

When computing an explanation for the most-likely class, we use a line-search for the parameter λ to determine its optimal value. Unless otherwise noted, we iteratively use 13 equally spaced λ values between 10^{-4} and 10^{-10} and stop when the resulting most-likely class of $\mathbf{e}_{ml}$ shifts (*deletion* and *repression game*) or achieves the highest probability among all classes (*preservation* and *generation game*). We use images of the ImageNet [26] validation set and pre-trained model weights.

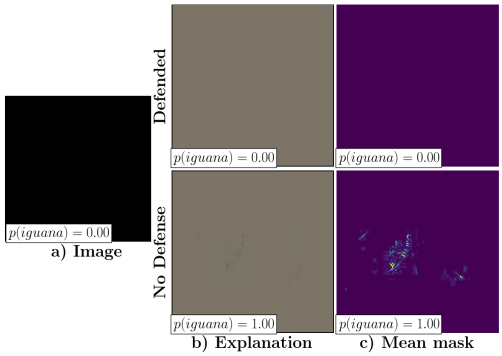


Figure A1: Explanation for the adversarial class *iguana* starting from a black image. An adversarial can only be computed without defense (*generation game*, *GoogleNet*). Mean masks are enhanced by a factor of 10.

Model	*GoogleNet*	*VGG16*	*AlexNet*	*ResNet50*
No Defense	100 %	100 %	100 %	100 %
Defended	0.0 %	0.1 %	0.1 %	0.0 %

Table A1: How often an adversarial class could be generated from a black image averaged over 998 ImageNet classes (*generation game*, $\lambda = 0$).

*contributed while working at BCAI. We additionally thank Volker Fischer, Michael Herman, Anna Khoreva for discussions and feedback.

A comparison of resulting masks for different learning rates and λ values for *GoogleNet* computed with the *deletion game* are shown in Fig. A2.

A higher λ value causes sparser masks due to a higher weighting of the sparsity invoking part $\|\mathbf{m}_{C_T}\|_1$ within the loss function (Eq. 2 and Eq. 3). Especially for higher λ values, the resulting masks are rather independent of the chosen learning rate of the SGD optimization.

A3. Qualitative Results

A3.1. Entropy of Reference Images

FGVis computes explanations $\mathbf{e}_{c_T}$ by optimizing for a perturbed version of the input image $\mathbf{x}$. The perturbation is modelled via a removal operator Φ [17, 14, 6, 11], which computes a weighted average between the image $\mathbf{x}$ and a reference image $\mathbf{r}$, using a mask $\mathbf{m}_{C_T}$:

$$\mathbf{e}_{c_T} = \Phi(\mathbf{x}, \mathbf{m}_{c_T}) = \mathbf{x} \cdot \mathbf{m}_{c_T} + (1 - \mathbf{m}_{c_T}) \cdot \mathbf{r}. \quad (7)$$

A good reference image $\mathbf{r}$ should carry little information and lead to a model prediction with a high entropy, meaning, ideally all classes are assigned the same softmax score (see '*Maximum (1000 classes)*' in Tab. A2 for the resulting maximum entropy). To compare references, we report their entropy for different models in Tab. A2.

For all models except *GoogleNet* the zero image reference has the highest entropy. Interestingly, for the zero image reference, the more recent architectures (*GoogleNet*, *ResNet50*) have a lower entropy. This indicates that these architectures do not assign a roughly equally distributed softmax score to all classes (as *AlexNet* or *VGG16*).

As expected, an increasing noise level σ_n for a Gaussian noise image as well as a decreasing standard deviation of the Gaussian blur filter σ_b reduces the entropy. Only *GoogleNet* does not fully follow this characteristic.

For comparison, we report the entropy for 1000 random ImageNet validation images for the different models.

Due to the high entropy as well as the low computational effort of a zero reference image, we choose this reference for FGVis.

A3.2. Class Discriminative / Fine-Grained

In Fig. A3 and Fig. A4 we show additional explanation masks for images containing two distinct objects. The objects are chosen from highly different categories to ensure little overlapping evidence. The explanations are computed using the *deletion game*, which generates the most pleasing class-discriminative explanations, and *GoogleNet*.

Note that FGVis discriminates well even if the two objects partially overlap. The figures additionally highlight the ability of FGVis to generate fine-grained explanations.

To determine λ we use for the most-likely class the strategy as described in Sec. A2. For the second class λ is optimized to significantly drop the softmax score of this class.

A3.3. Investigating Biases of Training Data

Learned objects. The coexistence of objects in images often results in a learned bias. In Fig. A5, we visualize such a bias for *GoogleNet* trained on ImageNet.

Sports equipment like hockey pucks or ping-pong balls frequently appear in combination with players. This bias is learned by the neural network and results in explanations that also contain pixels belonging to the players. Without deleting these pixels, the *deletion game* is not able to shift the class of the images.

Learned color. We quantitatively verify the color bias reported in Sec. 4.3 and show the 19 classes of ImageNet which are most and least affected by swapping the color in Tab. A3. We swap each of the three color channels *BGR* to either *RBG* or *GRB* and calculate the ratio of maintained true classifications on the validation data after the swap.

Fig. A6 shows explanations for the class school bus computed using the *preservation game* for *VGG*. The yellow color, also visible in the original images (Fig. A7), is dominant in most of the explanations.

Fig. A8 shows explanations for the class minivan computed using the *preservation game* for *VGG*. The original color of the car is not consistently preserved. Especially for

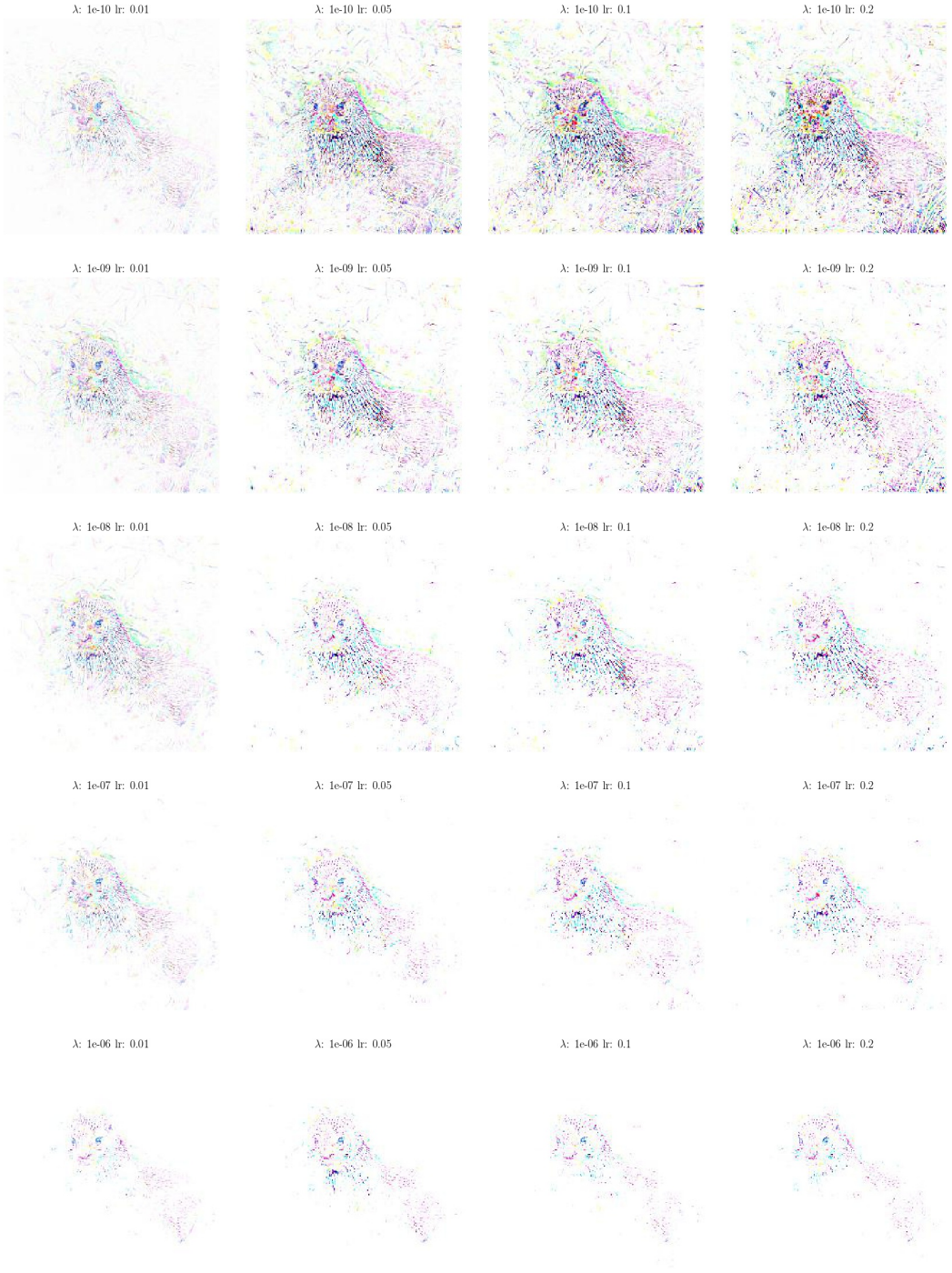
Reference image $\mathbf{r}$	*AlexNet*	*GoogleNet*	*VGG16*	*ResNet50*
Zero image	6.90	4.08	6.31	5.09
Gaussian noise image ($\sigma_n = 8$)	5.11 ± 0.16	4.62 ± 0.16	5.59 ± 0.09	4.56 ± 0.14
Gaussian noise image ($\sigma_n = 32$)	2.61 ± 0.29	4.67 ± 0.22	4.38 ± 0.23	4.07 ± 0.30
Blurred ImageNet image ($\sigma_b = 5$)	3.67 ± 1.12	3.15 ± 1.31	4.08 ± 1.43	2.38 ± 1.58
Blurred ImageNet image ($\sigma_b = 10$)	4.56 ± 0.88	4.09 ± 1.08	4.83 ± 0.86	3.22 ± 1.25
ImageNet image	1.73 ± 1.43	1.09 ± 1.14	1.06 ± 1.22	0.67 ± 0.91
Maximum (1000 classes)	6.91	6.91	6.91	6.91

Table A2: Entropy of reference images $\mathbf{r}$ for different models. The entropy is averaged over 1000 random instances of each reference image. Gaussian noise images are generated by independently sampling for each pixel from a Gaussian distribution with zero-mean and a standard deviation of σ_n. The blurred ImageNet images are computed using a Gaussian blur filter with a standard deviation of σ_b. For all random references we report the mean $\pm$ standard deviation of the entropy.

white or grey cars (original images in Fig. A9) the visible color in the explanation is reduced to a greenish-blue color.

Fig. A6 and A8 show all correctly classified images for school bus and minivan.

(a) Image



(b) Masks of class *otter*.

Figure A2: Comparison of resulting masks for different learning rates (lr) and λ values computed using the *deletion game* and *GoogleNet*.

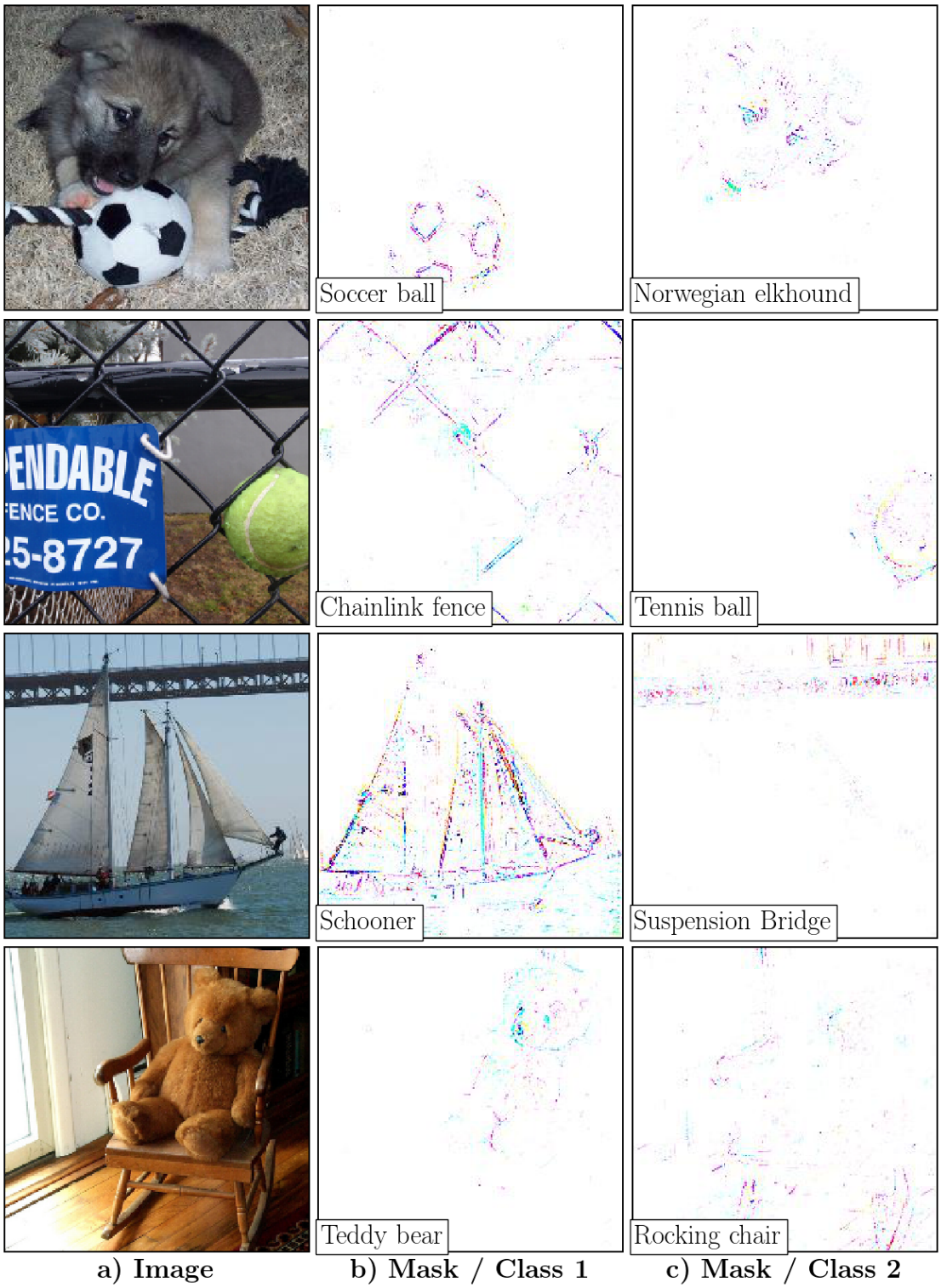


Figure A3: Explanation masks for images with multiple objects computed using the *deletion game* and *GoogleNet*. FGVis produces class discriminative explanations, even when objects partially overlap. Note that objects not belonging to either class, *e.g.* the rug in the top row, the blue sign on the chainlink fence, or the window in the bottom row vanish in the explanation. Additionally, FGVis is able to visualize fine-grained details down to the pixel level.

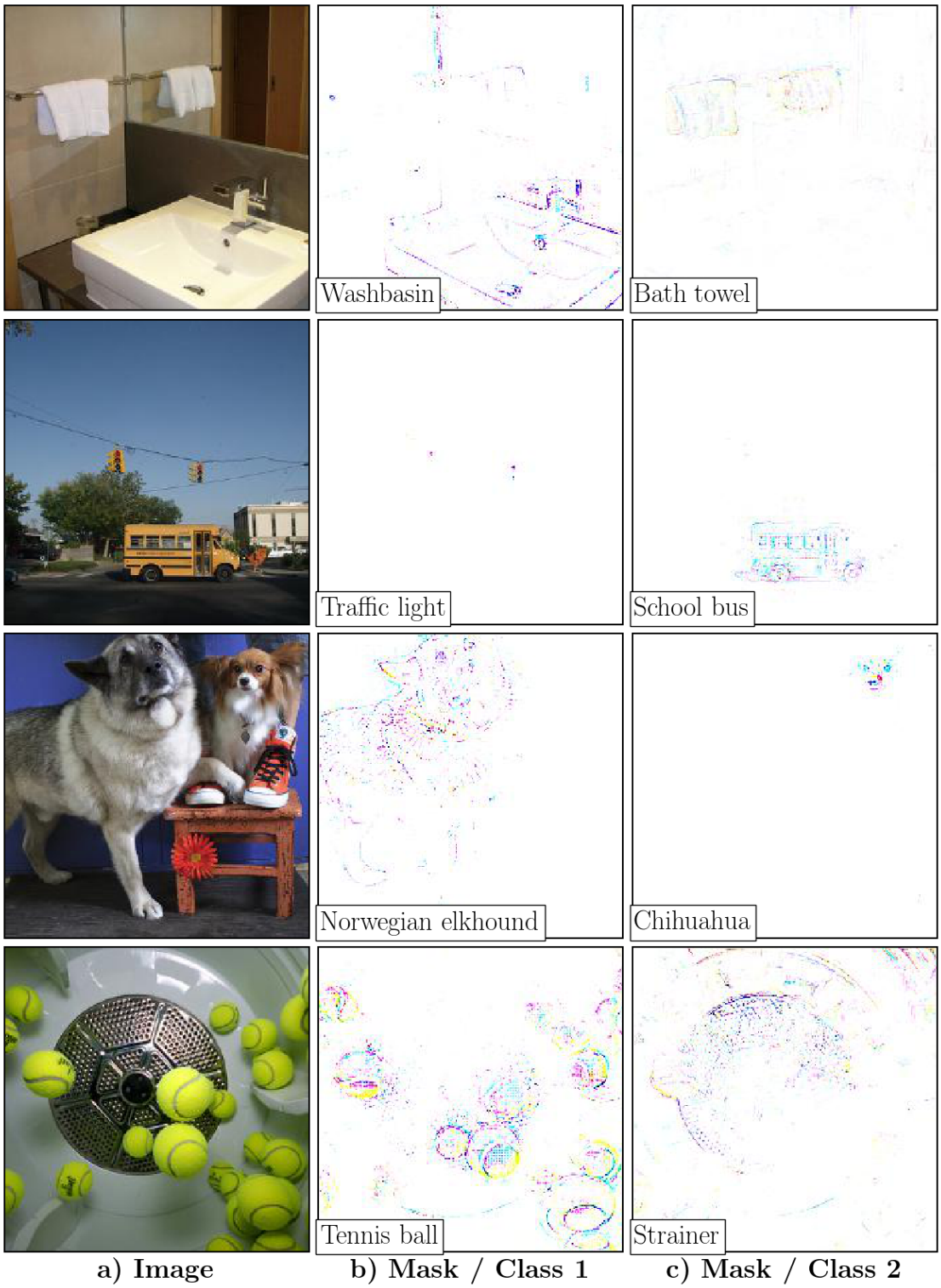


Figure A4: Explanation masks for images with multiple objects computed using the *deletion game* and *GoogleNet*. FGVis produces class discriminative explanations, even when objects partially overlap. This is especially visible in the last row where the tennis balls are almost all removed in the explanation mask for the class strainer.

Figure A5: Visual explanations computed using the *deletion game* for *GoogleNet*. For both classes (hockey puck and ping-pong ball) the explanation method has to additionally delete pixels of the players and the table tennis bat/ice-hockey stick to shift the prediction of the model. This clearly highlights a bias of the data towards images which contain a puck/ball, a player and sports equipment.

ID	Class name	#Images	Avg. RBG, GRB	RBG	GRB
168	redbone	31	0.00 %	0.00 %	0.00 %
964	potpie	28	0.00 %	0.00 %	0.00 %
159	Rhodesian ridgeback	35	0.00 %	0.00 %	0.00 %
930	French loaf	27	0.00 %	0.00 %	0.00 %
234	Rottweiler	42	1.19 %	0.00 %	2.38 %
214	Gordon setter	36	1.39 %	2.78 %	0.00 %
963	pizza, pizza pie	35	1.43 %	2.86 %	0.00 %
950	orange	35	1.43 %	2.86 %	0.00 %
184	Irish terrier	33	1.52 %	0.00 %	3.03 %
962	meat loaf, meatloaf	29	1.72 %	3.45 %	0.00 %
984	rapeseed	47	2.13 %	4.26 %	0.00 %
211	vizsla, Hungarian pointer	35	2.86 %	2.86 %	2.86 %
11	goldfinch, Carduelis carduelis	48	3.12 %	0.00 %	6.25 %
934	hotdog, hot dog, red hot	40	3.75 %	2.50 %	5.00 %
218	Welsh springer spaniel	39	3.85 %	2.56 %	5.13 %
191	Airedale, Airedale terrier	37	5.41 %	5.41 %	5.41 %
163	bloodhound, sleuthhound	18	5.56 %	5.56 %	5.56 %
961	dough	15	6.67 %	0.00 %	13.33 %
263	Pembroke, Pembroke Welsh corgi	41	7.32 %	7.32 %	7.32 %
...	...	...	...	...	...
779	school bus	42	8.33 %	9.52 %	7.14 %
...	...	...	...	...	...
656	minivan	21	83.33 %	71.43 %	95.24 %
...	...	...	...	...	...
528	dial telephone, dial phone	36	95.83 %	91.67 %	100.00 %
866	tractor	37	95.95 %	91.89 %	100.00 %
572	goblet	26	96.15 %	96.15 %	96.15 %
47	African chameleon, Chamaeleo chamaeleon	40	96.25 %	95.00 %	97.50 %
302	ground beetle, carabid beetle	27	96.30 %	96.30 %	96.30 %
463	bucket, pail	27	96.30 %	96.30 %	96.30 %
717	pickup, pickup truck	28	96.43 %	100.00 %	92.86 %
178	Weimaraner	44	96.59 %	93.18 %	100.00 %
669	mosquito net	44	96.59 %	97.73 %	95.45 %
661	Model T	46	96.74 %	97.83 %	95.65 %
769	rule, ruler	36	97.22 %	100.00 %	94.44 %
771	safe	40	97.50 %	97.50 %	97.50 %
829	streetcar, tram, tramcar, trolley, ...	41	97.56 %	97.56 %	97.56 %
713	photocopier	44	97.73 %	100.00 %	95.45 %
916	web site, website, internet site, site	47	97.87 %	100.00 %	95.74 %
423	barber chair	31	98.39 %	96.77 %	100.00 %
190	Sealyham terrier, Sealyham	39	98.72 %	97.44 %	100.00 %
340	zebra	47	100.00 %	100.00 %	100.00 %
545	electric fan, blower	37	100.00 %	100.00 %	100.00 %

Table A3: Ratio of maintained true classifications on the validation data of ImageNet after swapping color channels for the most and least affected 19 classes and minivan / school bus. Each of the three color channels *BGR* are swapped to either *RBG* or *GRB*. The class ID, class name, number of truly classified images before the color swap (#Images) and percentage of maintained classification after the swap for the average over *RBG* or *GRB* and each swap individually are reported. Most color-dependent classes are redbone or potpie. Most color-independent classes zebra or electric fan.

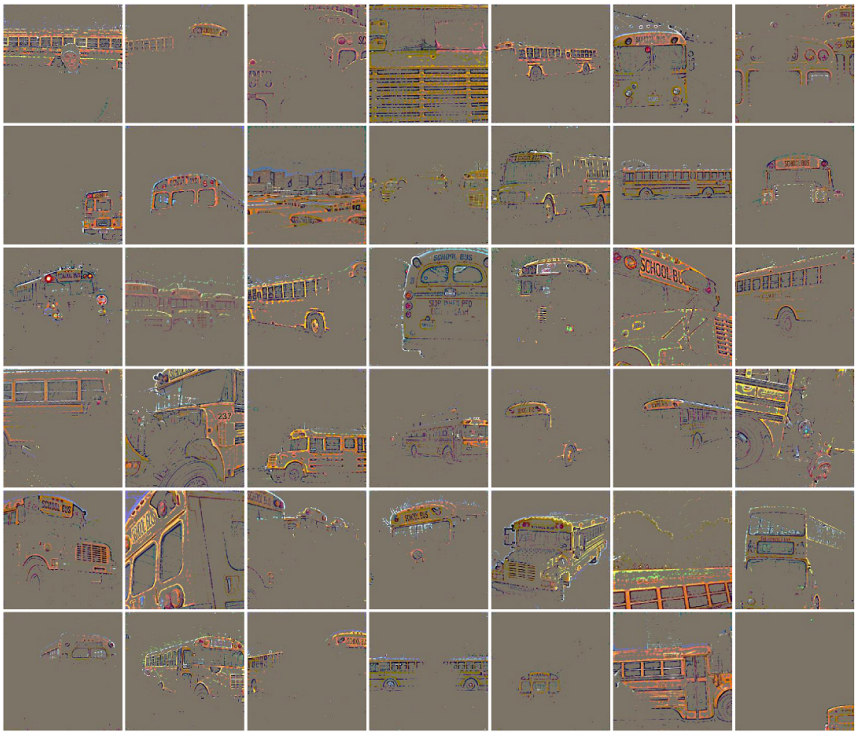


Figure A6: Explanations computed using the *preservation game* for *VGG* for the class school bus.

Figure A7: Input images for the explanations in Fig. A6

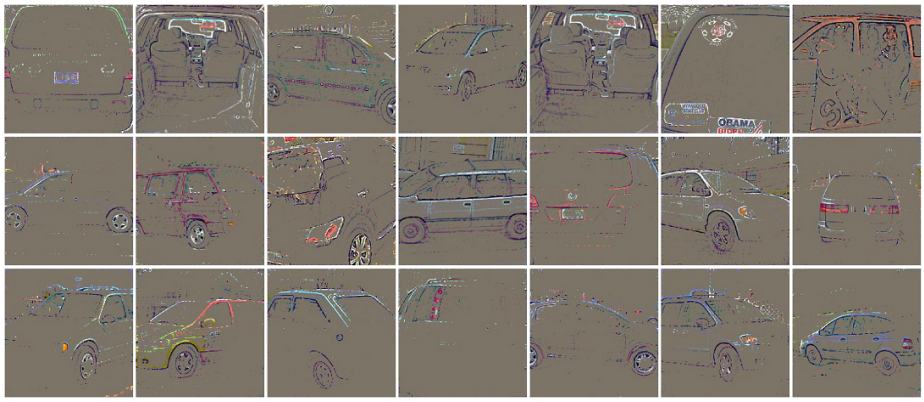


Figure A8: Explanations computed using the *preservation game* for *VGG* for the class minivan.

Figure A9: Input images for the explanations in Fig. A8

A3.4. Comparison of Networks

In Fig. A10 and Fig. A11 we compare the mask and explanation for four network architectures (*GoogleNet, VGG16, AlexNet, ResNet50*) using the *deletion game*. Respectively, in Fig. A12 and Fig. A13 we use the *preservation game* for the same comparison.

For all settings the explanations of *ResNet50* and *VGG16* are more dense, meaning more pixels have to be deleted/preserved to change/preserve the class prediction. This could be an indicator that these models are more robust, though, a detailed explanation would require further research. Besides, the grid-like pattern for the explanations from *ResNet50*, described in Sec. 4.1 are visible.

The importance of the color to classify the school bus (described in Sec. 4.3) can be seen in Fig. A13.

For *VGG16* we have observed that the pixels at the image edge are in many cases highlighted in the explanations. Furthermore, *VGG16* shows pronounced edges in the explanation compared to the other networks.

A3.5. Comparison of Games

In Fig. A14 and A15 the different game types (see Sec. 3.1) are visually compared for *GoogleNet*.

The resulting explanations for the *repression* and *deletion game* are qualitatively similar. The similarity among the two games is due to both using the same optimization with only a different starting condition $\mathbf{m} = 0$ for the *repression* vs. $\mathbf{m} = 1$ for the *deletion game*. The same observation holds for the *generation / preservation game.*

The explanations of the *repression* and *deletion game* are more sparse compared to the *generation / preservation game*. This is most likely due to the fact that only small parts of the image need to be suppressed to change the model output (*e.g.* shifting one breed of dog to another), though, to evoke a certain model output one needs to create sufficient amount of evidence for this class.

During the optimization only class pixels containing evidence towards the target class need to be changed for the *generation* and *deletion game*. After optimization most of the mask values stay zero for the *generation game* and one

for the *deletion game*. The optimized masks are thus similar to its starting conditions.

Vice versa, the opposite holds for the *preservation* and *repression game*.

A3.6. Further Examples

In Fig. A16, A17, A18, and A19 further explanations computed using FGVis are shown.

A4. Quantitative Results

A4.1. Faithfulness of Explanations

To evaluate the faithfulness of our approach, we use the deletion metric of Petsiuk *et al*. [32]. This metric measures how the removal of evidence affects the prediction of the used model. The metric assumes that an importance map is given, which ranks all image pixels with respect to their evidence for the predicted class c_{ml} (*i.e*. the most-likely class). We use the mean mask (see Sec. A3.5) as the pixel-wise importance map. The mean mask is computed for all images in the ImageNet validation dataset using the *deletion game* with a learning rate of 0.3 and a line-search to determine the λ value. We iteratively use 4 equally spaced λ values between 10^{-7} and 10^{-10} and stop when $y_e^{c_T} < 0.02 \cdot y_x^{c_T}$, where $y_e^{c_T}$ is the softmax score of class c_T given the explanation and $y_x^{c_T}$ the corresponding score given the image.

Using the importance map, the deletion curve is generated by successively removing pixels from the input image according to their importance and measuring the resulting probability of the class c_{ml} (see Fig. A20c). The removed pixels are set to zero, as proposed in Petsiuk *et al*. [32]. The fraction of removed pixels is increased in increments of 0.25% for the first 100 steps and in increments of 1% for the remaining 75 steps. In Fig. A20b, we visualize for an example image the binary masks used to successively set pixels to zero. For a clearer illustration, we reduced the number of deletion steps in this figure. The deletion metric is computed by measuring the *area under the curve* AUC of the deletion curve (see Fig. A20c) using the trapezoidal rule.

A4.2. Visual Explanation for Medical Images

Background of the disease: As people with diabetes have a high prevalence for RDR [47], a frequent retinal screening is recommended and deep learning algorithms have been successfully developed to classify fundus images ([8], [20], [3], [46]). The black box character of these algorithms can be reduced by visual explanation techniques as shown in [18].

Of the publicly available 88,702 images [15] from EyePACS [10], we us 80% for training and 20% for validation for a classifier with binary outcome (referable diabetic retinopathy (RDR) vs. non-RDR) which is later used for the weakly-supervised localization. We use a similar setup as in [18] to train the binary classifier (RDR vs. non-RDR).

Training was conducted with the same implementation settings as described in [18] using an adopted version of the CNN architecture proposed by [16] for classifying retinal images. We use leaky ReLUs as non-linearities and include batch normalization.

The DiaretDB1 dataset [25] used to evaluate the weakly-supervised localization is a dataset of 89 color fundus images collected at the Kuopio University Hospital, Finland. All images have a resolution of 1500x1152 pixels and are scaled to the input dimension of the model.

The dataset is ground truth marked by four experts. As proposed in [25] we consider pixels as lesions if at least three experts have agreed.

We use FGVis with a fixed $\lambda = 10^{-10}$ and a learning rate of 0.25 stopping if the softmax score for RDR falls below 10% with a maximum of 500 iterations.

In Fig. A21 retinal images overlaid with the ground truth (top row) are compared to our prediction (bottom row). To be consistent with [18] the masks m are binarized for better visualization and to be able to quantitatively report the sensitivity (see Tab. 3). Values greater or equal than 4% of the maximum are set to one, the remaining pixels to zero. The predicted pixels in the fine-grained masks m map to the ground truth. Note that FGVis detects these pixels as they are the important ones to be deleted to reduce the softmax score for RDR.

A medical expert would also look at mutations in the optic disk or blood vessels which additionally are an indicator for the disease [41]. These mutations are also highlighted by our method. They are not labelled in the ground truth markings leading to visual false positives (FPs).

The strength of FGVis to visualize fine-grained structures can be seen in the detection of red small dots (microaneurysm) which are the earliest sign of diabetic retinopathy [2]. As these often merely cover some pixels in the image, it is hard to detect them (zooming in Fig. A21 is necessary to spot these).

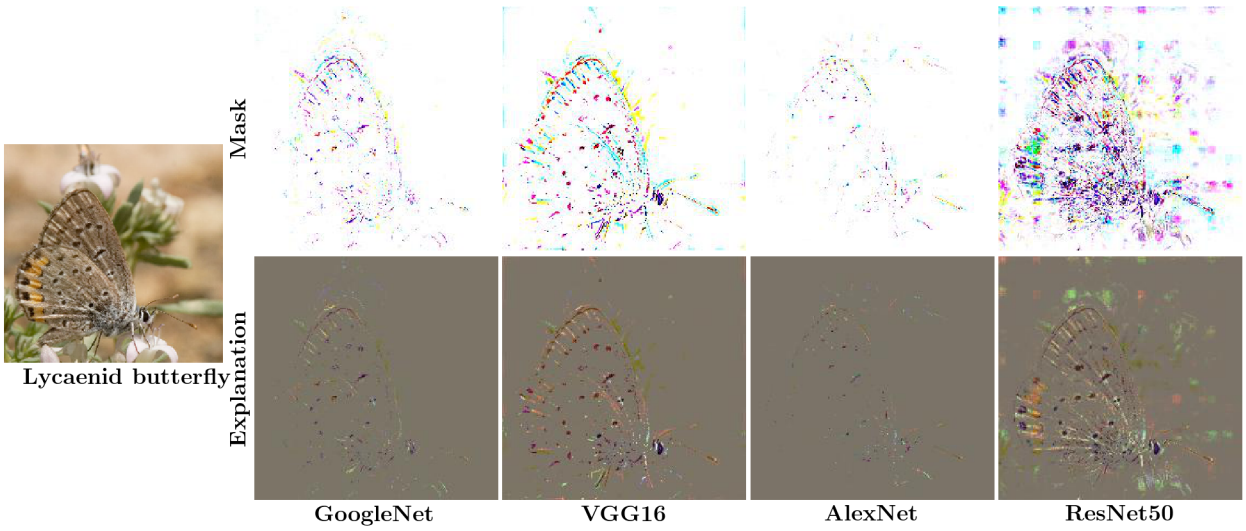


Figure A10: Masks and explanations computed using the *deletion game* for different networks.

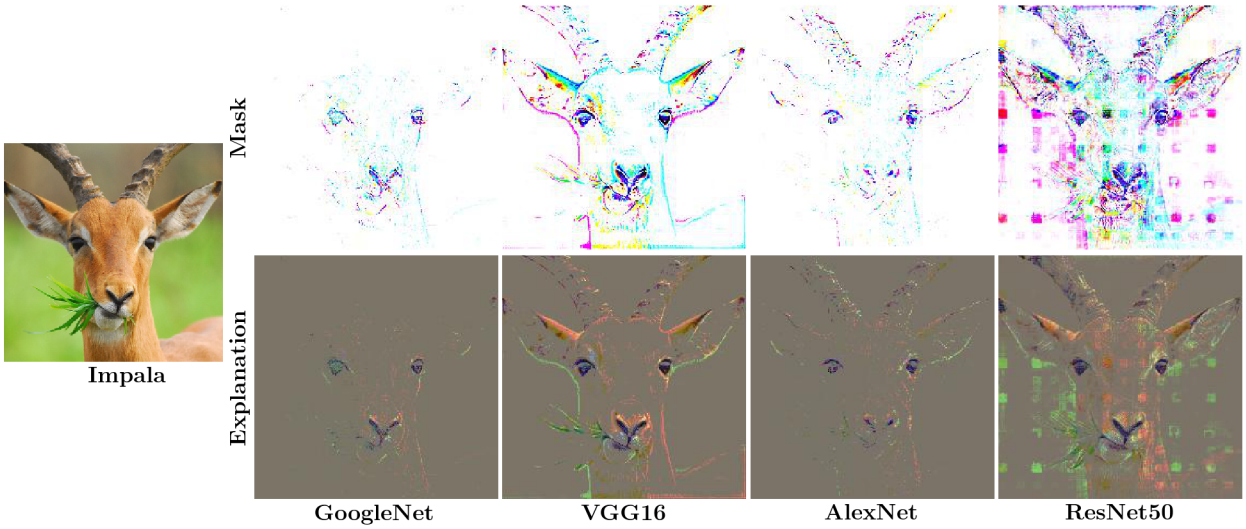


Figure A11: Masks and explanations computed using the *deletion game* for different networks.

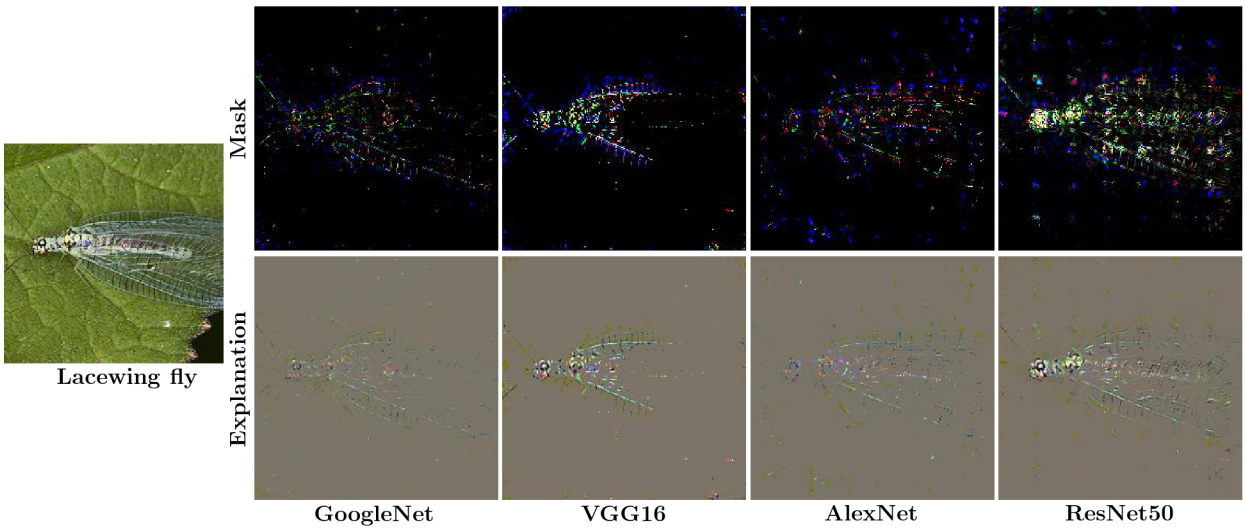


Figure A12: Masks and explanations computed using the *preservation game* for different networks.

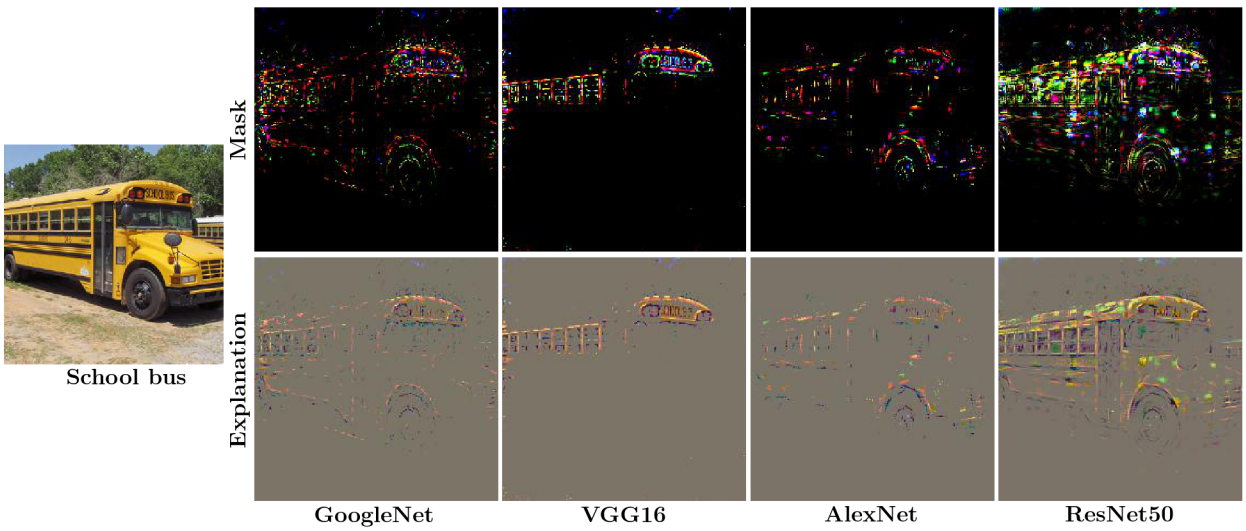


Figure A13: Masks and explanations computed using the *preservation game* for different networks.

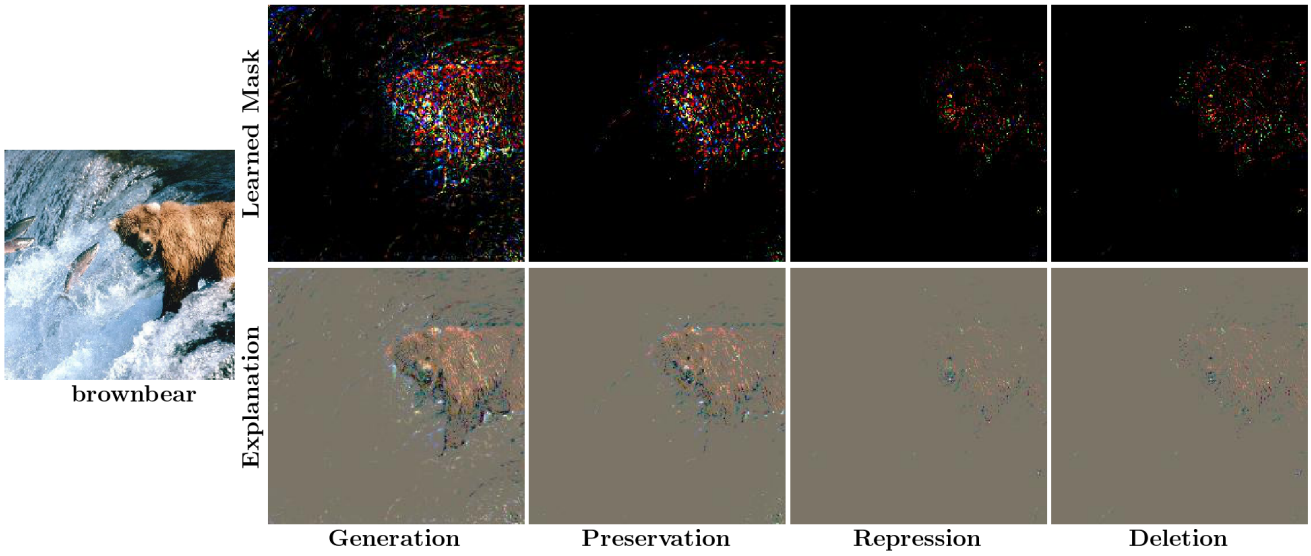


Figure A14: Explanations and masks computed using the different games for *GoogleNet*. For the *repression* and *deletion game* the complementary masks $(1 - \mathbf{m})$ are plotted to have true-color representations (see Sec. A3.4).

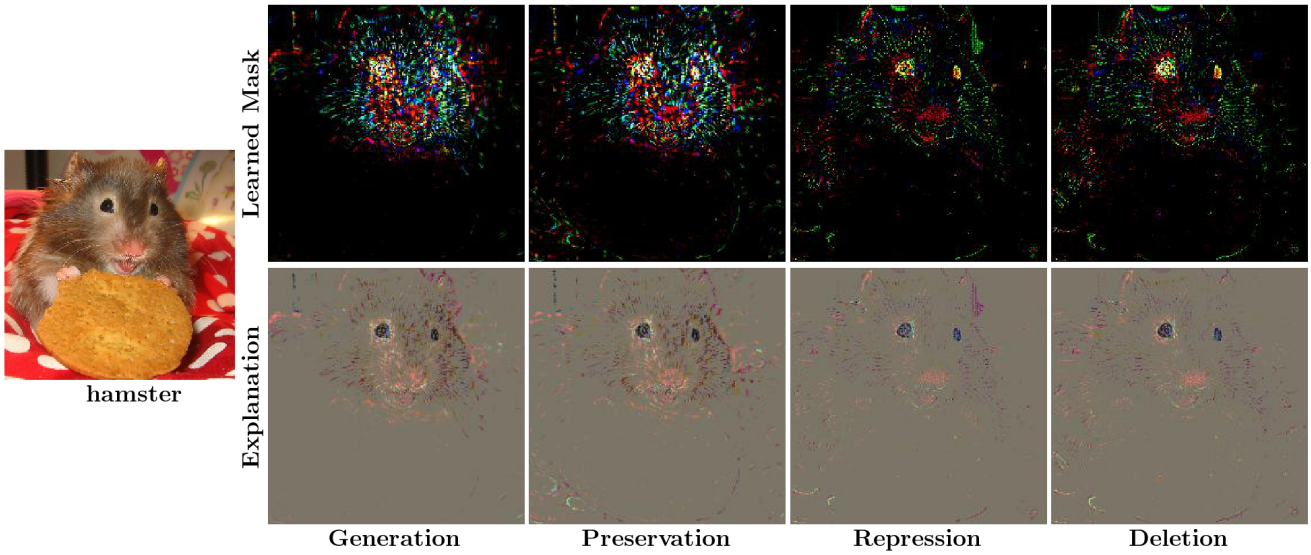


Figure A15: Explanations and masks computed using the different games for *GoogleNet*. For the *repression* and *deletion game* the complementary masks $(1 - \mathbf{m})$ are plotted to have true-color representations (see Sec. A3.4).

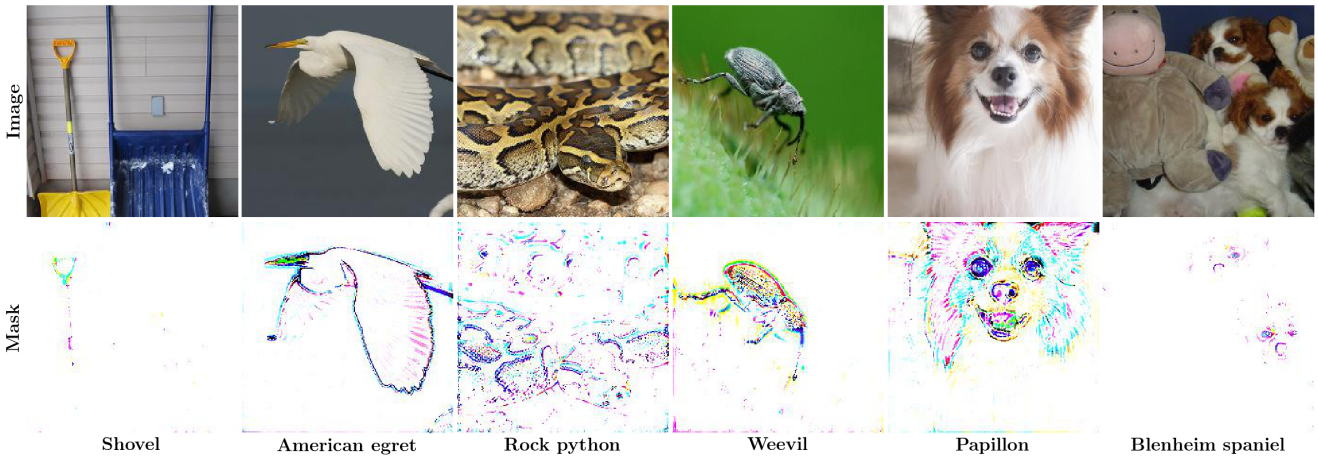


Figure A16: Explanation masks computed using the *repression game* for *VGG16*.

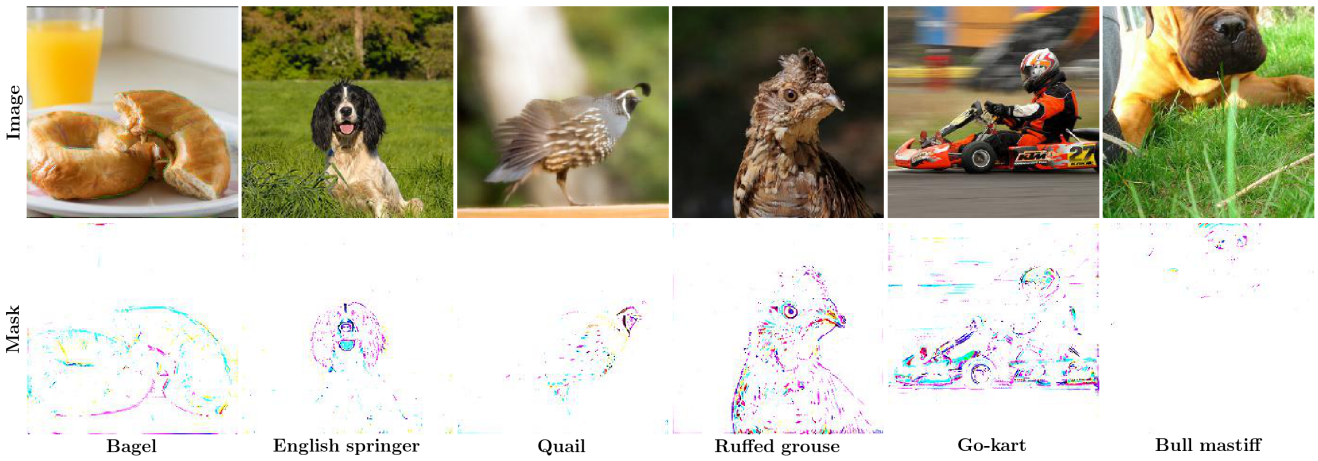


Figure A17: Explanation masks computed using the *repression game* for *VGG16*.

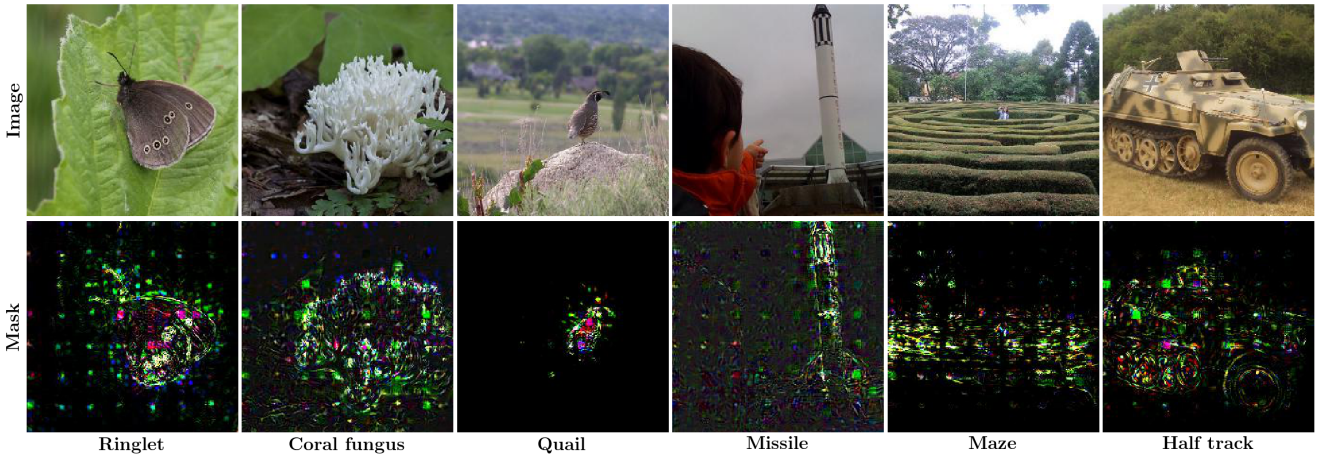


Figure A18: Explanation masks computed using the *preservation game* for *ResNet50*.

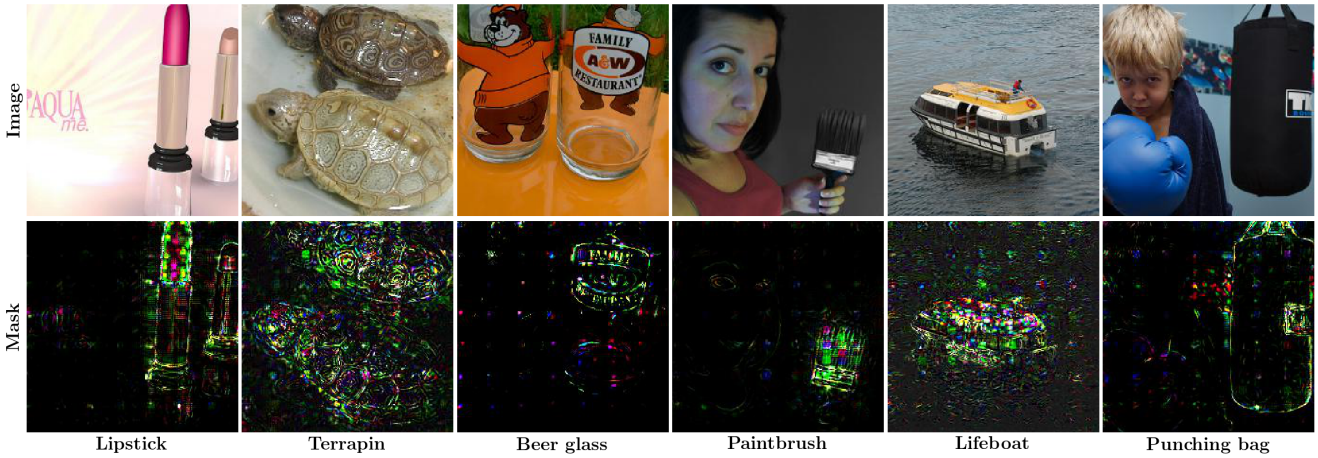
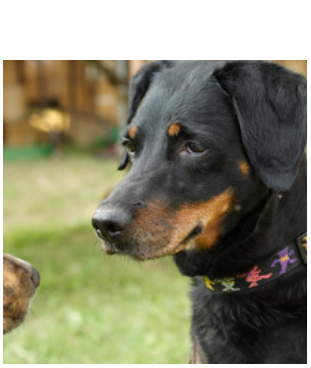
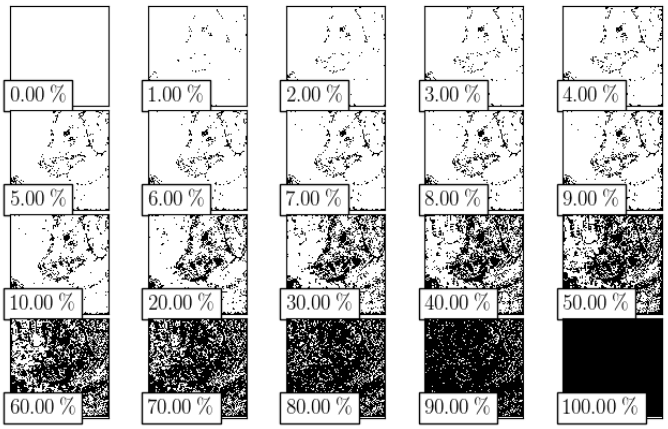


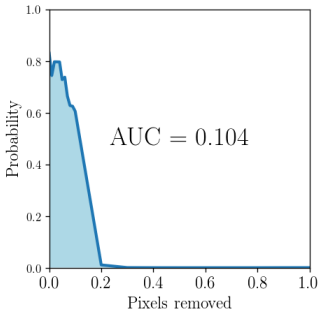
Figure A19: Explanation masks computed using the *preservation game* for *ResNet50*.



(a) Image



(b) Binary deletion masks with fraction of removed pixels



(c) Deletion curve

Figure A20: The deletion curve (c) is computed by successively deleting pixels (b) from the image according to their importance and measuring the resulting probability of the class c_{ml}.

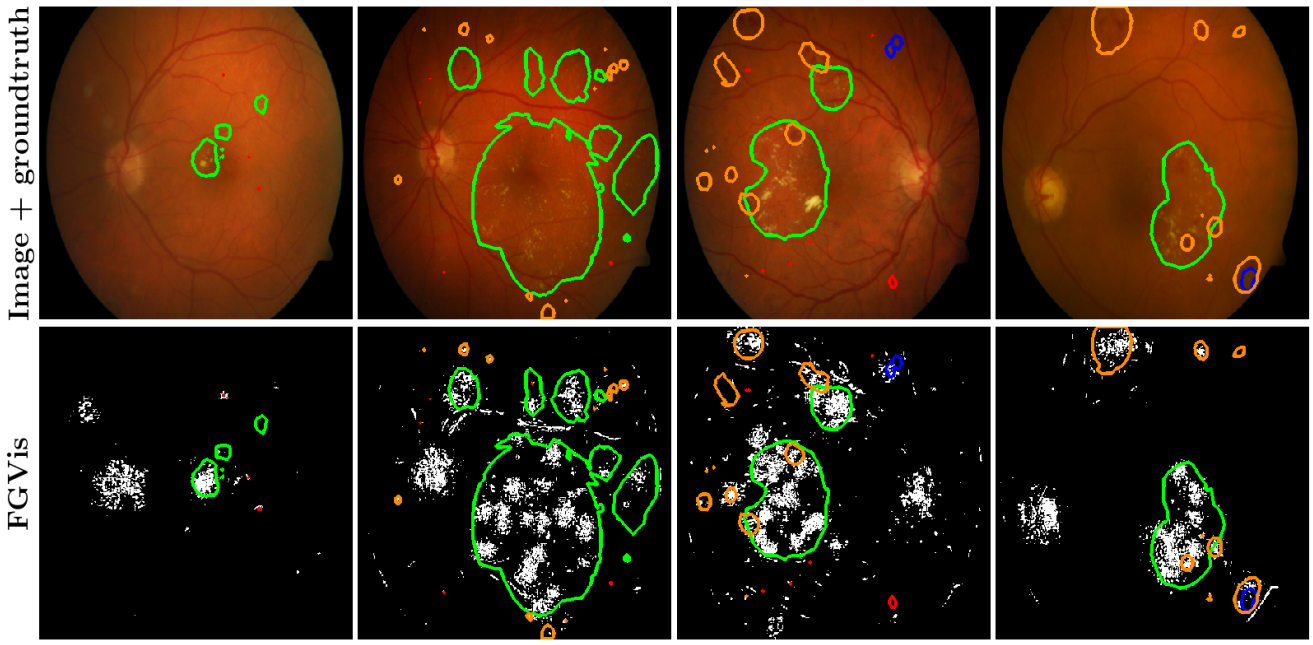


Figure A21: Weakly-supervised localization results on DiaretDB1 images. The top row shows fundus images, the bottom row our detection. All images are overlaid with ground truth markings in green (hard exudates), blue (soft exudates), orange (hemorrhages), red (red small dots). Though the network was trained in a weakly-supervised way given only the image label, most of the regions highlighted by FGVis fall within the ground truth markings. Note that mutations in the optic disk or blood vessels are an indicator for the disease [41] but these are not covered by the ground truth markings leading visually to false positives. FGVis highlights part of the blood vessels and optic disks.