Supplementary Material

What You See is What You Classify:
Black Box Attributions

A Compute Times

To evaluate the practicability of the methods, we have evaluated the time it takes for each method to generate all aggregated masks for the segmentation evaluation in Section 4.3. Note that this does not include any training time for our method and adapted RTIS (up to 12 hours on VOC-2007 and up to 2 days on COCO-2014), which only has to be performed once per model and dataset. Tab. 3 shows how significant the inference time differences between the methods are, even for a relatively small amount of images. RISE, EP and iGOS++ are orders of magnitude slower than GCAM, RTIS and our explainer. All experiments have been conducted using the same Tesla P100 PCIe 12GB GPU.

		Method					
Dataset	Classifier	GCAM	RISE	EP	iGOS++	RTIS	Explainer (ours)
VOC-2007	VGG-16	00:00:11	01:15:29	05:57:16	02:57:15	00:00:08	00:00:08
	Resnet-50	00:00:10	00:46:21	04:45:59	01:55:58	00:00:11	00:00:11
COCO-2014	VGG-16	00:01:23	06:00:47	33:54:33	28:55:22	00:00:34	00:00:34
	Resnet-50	00:01:30	03:41:35	26:40:42	19:09:10	00:00:47	00:00:47

Table 3: Time for computing segmentation masks with each method. The times are given in hh:mm:ss format. 210 and 1000 images have been segmented for VOC-2007 and COCO-2014, respectively.

B Classification accuracy

We demonstrate through Tab. 4 that our *Explainer* architecture does not adversely affect the performance of the (pre-)trained classification network. On occasions, a slight drop in accuracy can be expected, since each individual image changes quite drastically after being masked, not only visually, but also in terms of pixel intensity. Such a drop can be explained simply by the fact that classifiers are not trained to recognize objects on a black background, which can behave as adversarial information to the VGG-16 and ResNet-50 classifiers. A loss in performance may also be indicative of the pre-trained classifier not being able to capitalize on an unexpected cue or correlations in the data (such as presence of blue and green in images with cows).

C Multiclass attribution examples on VOC using VGG-16 and ResNet-50

In this section, we show examples for the attribution over multiple classes on the VOC-2007 test set, using VGG-16 and ResNet-50 classifiers. In Fig. 4 and Fig. 5 below, each row corresponds to a random image from the VOC-2007 dataset, while each column corresponds to: the original image, the aggregated mask (per-pixel maximum over the class-masks), and the top 5 strongest attributed classes (TAC). Fig. 4 shows the results for VGG-16, while Fig. 5 shows the ones for ResNet-50. Each mask is scored according to the average mask activation (AMA), then sorted in descending order of the first 5 classes, as shown below. As this process is completely independent from the classifier itself, we also add the multi-label scores (logits through the sigmoid) for the original classifier on the unmasked image (CLS), multiplied by 100. When a class from the ground truth is attributed, the TAC is accompanied by two asterisks (**). We do not report the classification scores for the masked images, because they are greatly perturbed since the classifier has never learned to classify heavily masked images. For example, the classifier will give relatively high scores for all classes when it is given a completely black (all zeros) image.

Data	Model	Precision	Recall	F-Score
VOC-2007	VGG-16 Cl.	88.4	63.9	74.2
	VGG-16 Expl.	86.4	62.9	72.8
	ResNet-50 Cl.	85.2	72.3	78.2
	ResNet-50 Expl.	77.3	69.6	73.2
COCO-2014	VGG-16 Cl.	73.5	45.4	56.1
	VGG-16 Expl.	66.9	45.0	53.8
	ResNet-50 Cl.	77.6	44.6	56.6
	ResNet-50 Expl.	69.1	48.9	57.3

Table 4: Comparison between base (pre-trained, frozen) classifiers (Cl.) and their *explained* counterparts (Expl.) on the respective test sets. All numbers are given as percentages (%). Since masked images result in overall lower logits, we have used slightly different thresholds for the logits to count as positive versus negative predictions, depending on whether they result from unmasked or masked images. The presented numbers demonstrate that the use of our *Explainer* masks on the inputs to the pre-trained networks does not significantly affect their classification scores.

One can see that for both VGG-16 (Fig. 4) and ResNet-50 (Fig. 5), masks are sharp and outline (parts of) the object(s) of interest. By looking at the classification scores and the average mask activation, one can see that they usually correlate, meaning that the attribution works well. In those rare cases where the *Explainer* masks show their highest activation for classes that are not the ones for which the classifier gives the highest probability, the *Explainer* might have learned to detect those objects better than the classifier itself. Remember that those predictions are made directly on out-of-sample data. Also, note that the average mask activation does of course not directly relate to a classification, since smaller objects will give a smaller AMA score than large ones.

An interesting example is at the second row in Fig. 4, where in the fourth column (second most attributed class) the *Explainer* shows a significant attribution for class 5 ("bottle") which is not present in the image. Only when looking at the classifier score on the unmasked image (CLS) this makes perfect sense since the classifier gives the "bottle" class a probability of over 60%. A similar example for this behavior is the fifth row in the same figure or the seventh row in Fig. 5. This indicates that the attributions by our *Explainer* are sensible even in failure cases of the classifier. In a few cases, the attributions for several classes are on the same image areas, as in the third and seventh rows of Fig. 5. This might happen for image areas that appear to be generally interesting for several classes at the same time but such cases are rare. In general, we see that ResNet-50 tends to provide masks with slightly more artifacts and spurious activations, when compared to the attributions of VGG-16.

Both experiments seem to agree that our *Explainer* is able to attribute classifiers to regions occupied by the correct class, or provide attributions caused by ambiguous biases in the classifier itself.

ID	Class name	ID	Class name
1	Aeroplane	2	Bicycle
3	Bird	4	Boat
5	Bottle	6	Bus
7	Car	8	Cat
9	Chair	10	Cow
11	Dining Table	12	Dog
13	Horse	14	Motorbike
15	Person	16	Potted Plant
17	Sheep	18	Sofa
19	Train	20	TV/Monitor

Table 5: Summary of the Pascal VOC-2007 Classes.

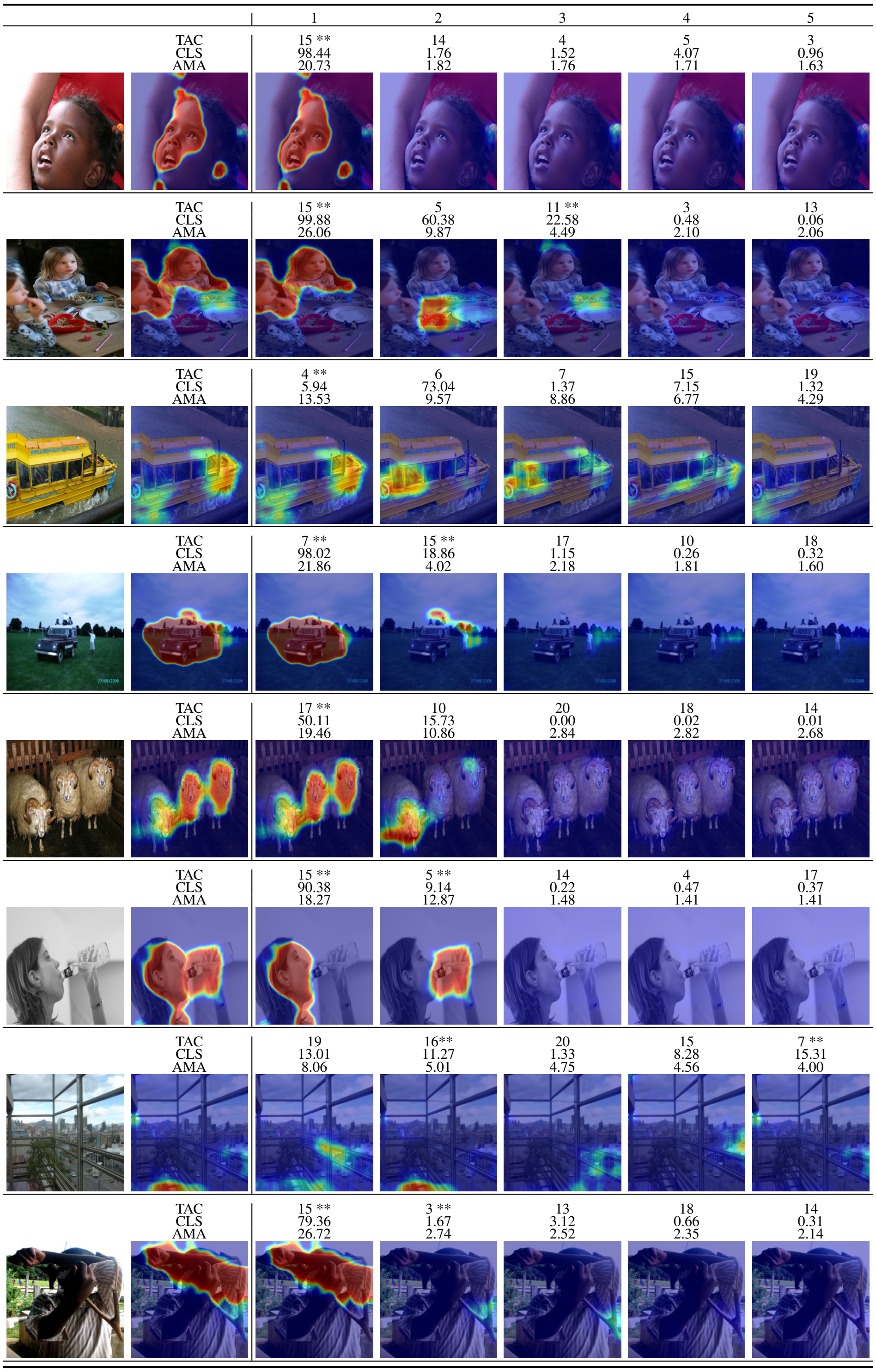
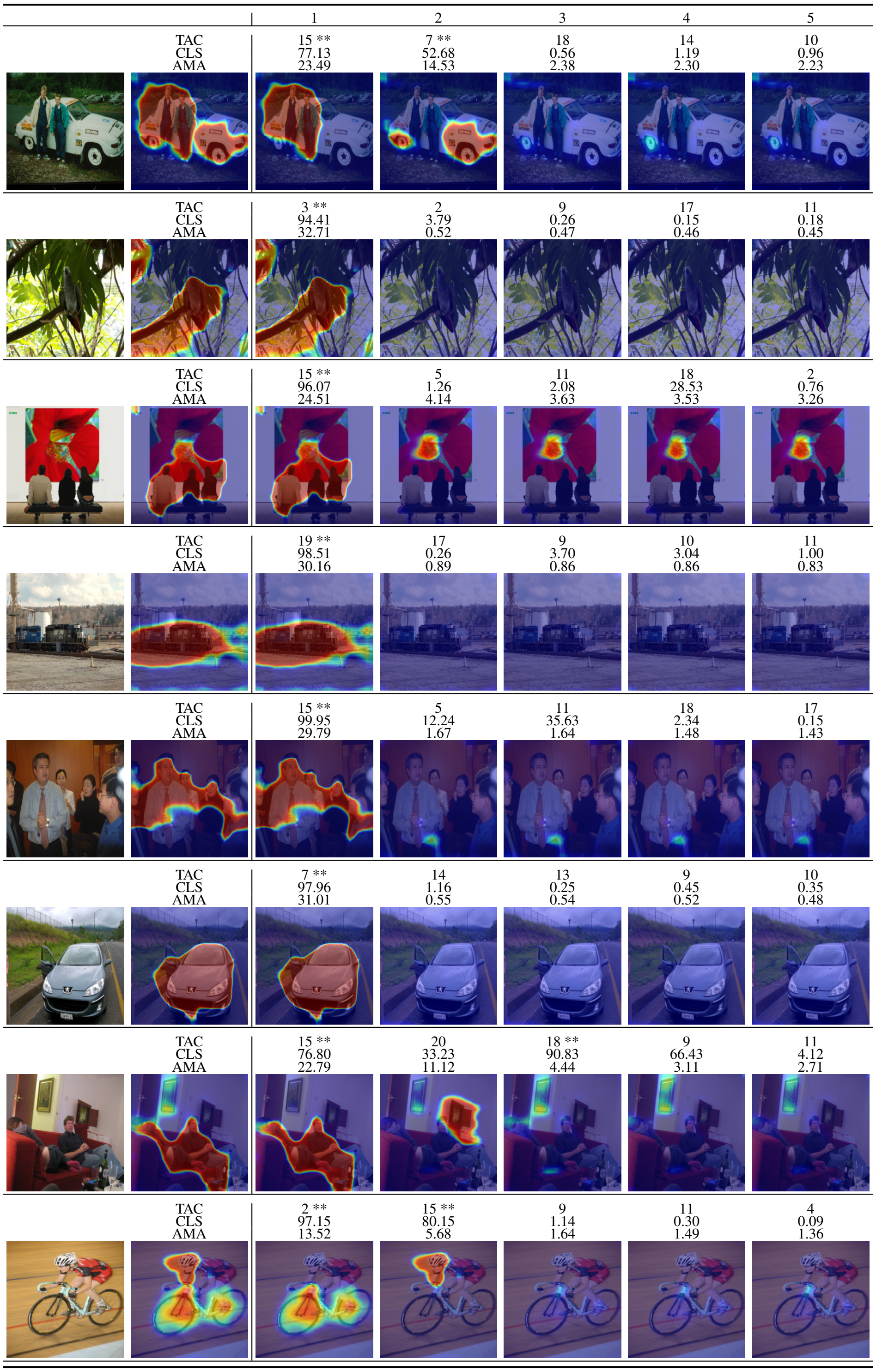


Image		1	2	3	4	5
1	TAC	15 **	14	4	5	3
	CLS	98.44	1.76	1.52	4.07	0.96
	AMA	20.73	1.82	1.76	1.71	1.63
2	TAC	15 **	5	11 **	3	13
	CLS	99.88	60.38	22.58	0.48	0.06
	AMA	26.06	9.87	4.49	2.10	2.06
3	TAC	4 **	6	7	15	19
	CLS	5.94	73.04	1.37	7.15	1.32
	AMA	13.53	9.57	8.86	6.77	4.29
4	TAC	7 **	15 **	17	10	18
	CLS	98.02	18.86	1.15	0.26	0.32
	AMA	21.86	4.02	2.18	1.81	1.60
5	TAC	17 **	10	20	18	14
	CLS	50.11	15.73	0.00	0.02	0.01
	AMA	19.46	10.86	2.84	2.82	2.68
6	TAC	15 **	5 **	14	4	17
	CLS	90.38	9.14	0.22	0.47	0.37
	AMA	18.27	12.87	1.48	1.41	1.41
7	TAC	19	16**	20	15	7 **
	CLS	13.01	11.27	1.33	8.28	15.31
	AMA	8.06	5.01	4.75	4.56	4.00
8	TAC	15 **	3 **	13	18	14
	CLS	79.36	1.67	3.12	0.66	0.31
	AMA	26.72	2.74	2.52	2.35	2.14

Figure 4: Top-5 class-wise attributions for the VGG-16 classifier, for 8 random images from the VOC-2007 test set. Class-wise masks are sorted according to their average activation on the image plane. TAC corresponds to the top activating class, please refer to the legend in Tab. 5. CLS shows the respective class probabilities by the classifier on the original (unmasked) images, multiplied by 100. AMA shows the average mask activations for the respective class, also multiplied by 100.



1 2 3 4 5
TAC 15 ** 7 ** 18 14 10
CLS 77.13 52.68 0.56 1.19 0.96
AMA 23.49 14.53 2.38 2.30 2.23
TAC 3 ** 2 9 17 11
CLS 94.41 3.79 0.26 0.15 0.18
AMA 32.71 0.52 0.47 0.46 0.45
TAC 15 ** 5 11 18 2
CLS 96.07 1.26 2.08 28.53 0.76
AMA 24.51 4.14 3.63 3.53 3.26
TAC 19 ** 17 9 10 11
CLS 98.51 0.26 3.70 3.04 1.00
AMA 30.16 0.89 0.86 0.86 0.83
TAC 15 ** 5 11 18 17
CLS 99.95 12.24 35.63 2.34 0.15
AMA 29.79 1.67 1.64 1.48 1.43
TAC 7 ** 14 13 9 10
CLS 97.96 1.16 0.25 0.45 0.35
AMA 31.01 0.55 0.54 0.52 0.48
TAC 15 ** 20 18 ** 9 11
CLS 76.80 33.23 90.83 66.43 4.12
AMA 22.79 11.12 4.44 3.11 2.71
TAC 2 ** 15 ** 9 11 4
CLS 97.15 80.15 1.14 0.30 0.09
AMA 13.52 5.68 1.64 1.49 1.36

Figure 5: Top-5 class-wise attributions for the ResNet-50 classifier, for 8 random images from the VOC-2007 test set. Class-wise masks are sorted according to their average activation on the image plane. TAC corresponds to the top activating class, please refer to the legend in Tab. 5. CLS shows the respective class probabilities by the classifier on the original (unmasked) images, multiplied by 100. AMA shows the average mask activations for the respective class, also multiplied by 100.

D Comparison of attribution methods on VOC-2007 using ResNet-50.

Figure 6 provides additional comparisons using ResNet-50 on VOC-2007. The comparison follows Fig. 3 of the manuscript. Overall, attribution methods for ResNet-50 are behaving better on VOC-2007 than on COCO-2014. Still, GCam, RISE and EP tend to provide smooth attributions, with the latter two sometimes scoring at multiple locations which are hard to relate to the classes present in the image. iGOS++ provides small activations that mostly seem to outline the most important object parts but sometimes also show artifacts that are difficult to interpret when it is less certain. RTIS provides masks that are often outlining objects, but fails in being sharp and providing clear attributions due to overshooting too much. Our *Explainer* method is mostly artifact-free and masks are very sharp on discriminative object parts.

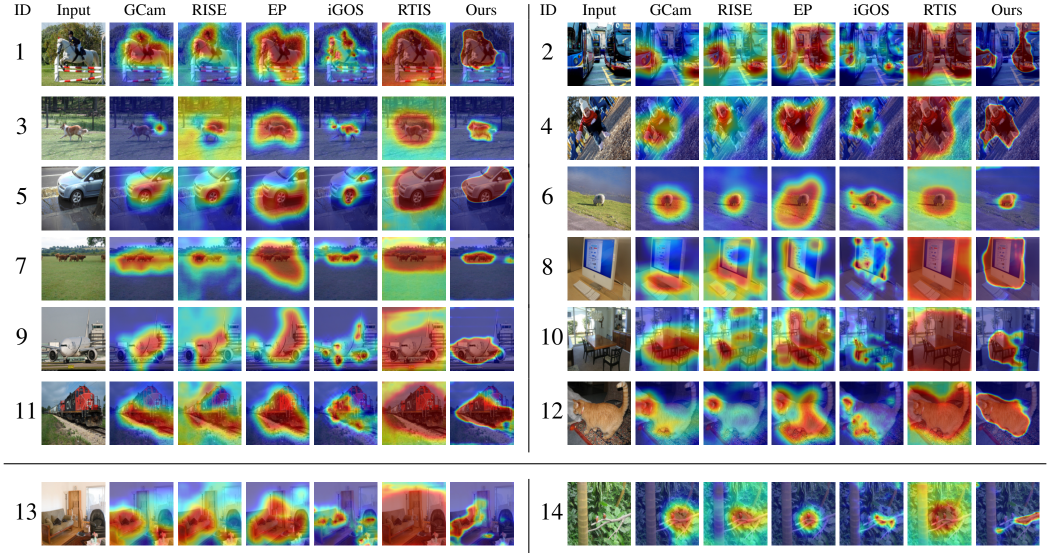


Figure 6: Comparison of attributions methods for a ResNet-50 network fine-tuned on images from the VOC-2007 dataset. Refer to Fig. 3 of the manuscript for more details.

E Comparison of attribution methods on COCO-2014 using VGG-16.

Figure 7 provides additional comparisons using VGG-16 on COCO-2014. The comparison follows Fig. 3 of the manuscript. Using VGG-16 on COCO-2014 results in good attributions, where masks are sharp and outline discriminative parts for the predicted classes precisely.

Not too differently than the results presented in the manuscript – attribution for the VGG-16 classifier on VOC-2007 – we can see that GCam [24] provides good attributions, if not for a fairly large smoothing. RISE [20] behaves similarly, and it often tends to provide several oversmoothed saliency locations, which are hard to interpret. EP [12] is much sharper, but also tends to overpredict attributions. On the other hand, iGOS++ shows small, localized attributions. However, it often also includes artifacts and sometimes does not capture certain objects at all. Finally, RTIS [5] does not show well delineated attribution on this combination of data and classifier as it overshoots too much to allow for a clear interpretation. Our *Explainer* provides localized attributions which are easy to interpret, with sharp boundaries, which in most cases coincide with the classes of interest or discriminative parts composing them. In general, the *Explainer* predicts compact masks with smaller areas, similarly to GCam, but much less smoothed overall.

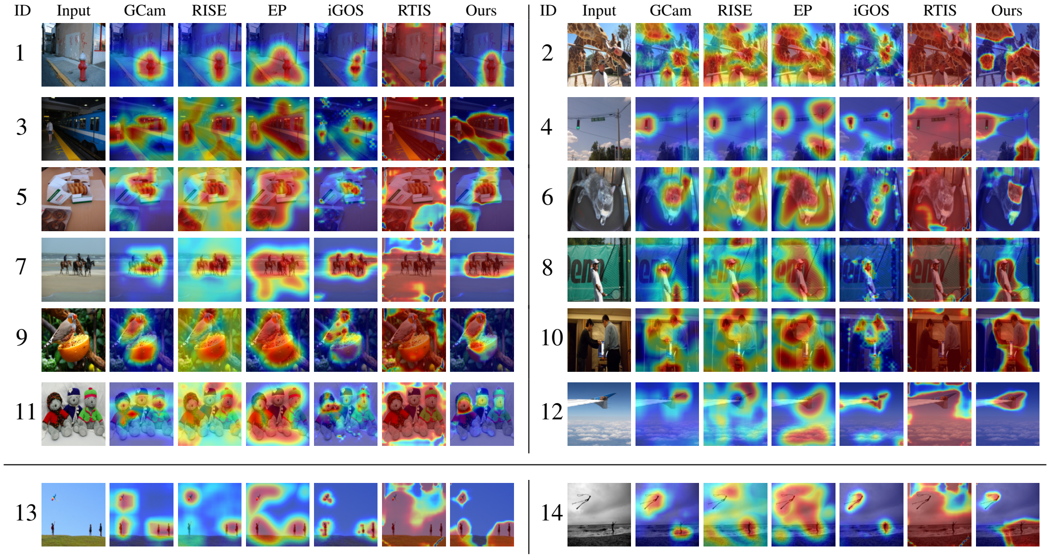


Figure 7: Comparison of attributions methods for a VGG-16 network fine-tuned on images from the COCO-2014 dataset. Refer to Fig. 3 of the manuscript for more details.

F Comparison of attribution methods on COCO-2014 using ResNet-50.

Figure 8 provides additional comparisons using ResNet-50 on COCO-2014. The comparison follows Fig. 3 of the manuscript. As discussed in the manuscript, this combination of data and classifier is the hardest to train the *Explainer* on.

As for previous comparisons, GCam shows accurate but very smooth attributions. RISE and EP behave similarly, by attributing the classification on the correct areas of the input image, but failing in focusing on precise image locations. iGOS++ once again shows small attributions, which in many cases lead to clear explanations but sometimes might also leave out parts where the other methods agree that they are important as well. In some cases it also includes artifacts, which make the interpretation harder. RTIS overpredicts the attributions, in particular when the objects of interest are small with respect to the input image plane. Our *Explainer* shows very clear attributions in most cases but sometimes also includes artifacts that occur as an active area in the top portion of the image. We have found that this specifically happens in images with the "person" class. Nonetheless, we can notice that the attribution masks are still precise and sharp in the object areas, suggesting that better results could probably be achieved by better hyperparameter and model selection.

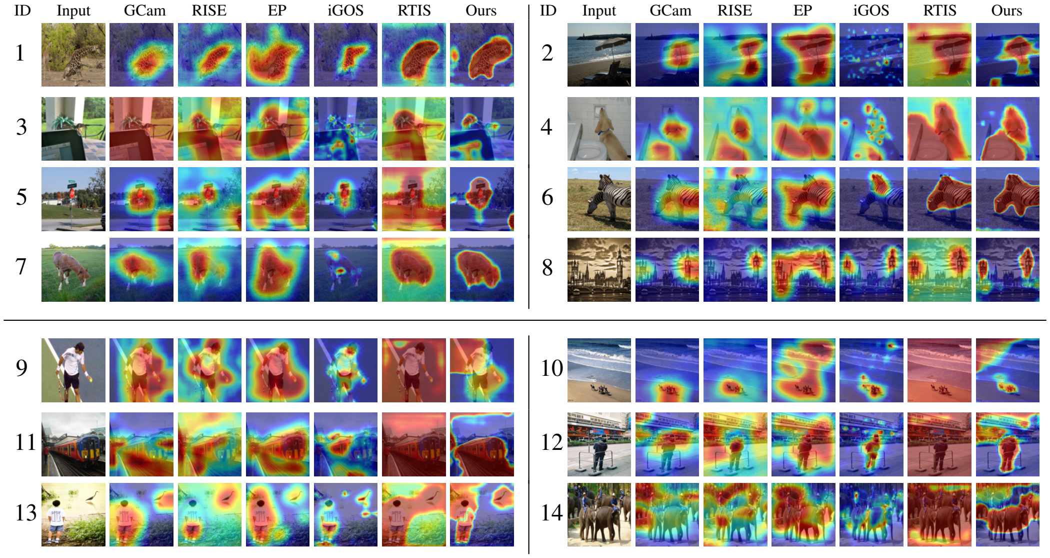


Figure 8: Comparison of attributions methods for a ResNet-50 network fine-tuned on images from the COCO-2014 dataset. Refer to Fig. 3 of the manuscript for more details.