Rebuttal for the ICML25 Submission: Pixel-level Certified Explanations via Randomized Smoothing

In the following document, we share figures and images that address the comments of the reviewers **s6h4**, **WAuf**, **itGy**, and **iTN6**. Note that this is only a supplementary document that we use to include figures which we reference in the rebuttal comment on OpenReview: https:/openreview.netforumid=NngoETL9IK&r.

Contents

1. Evaluation of certified attributions on 5 models: ResNet-18, W-ResNet-50-2, ResNet-152, VGG-11 and ViT-B/16

1.1. Certified Robustness (%certified)

We assess attribution robustness on five different models using the %certified metric evaluated at the input and final layers across certified radii ($R = 0.10, 0.17, 0.22$) in Figure 1 and sparsification ($K = 50$, 30 and 10) in Figure 2. Note that ViT-B/16 is not evaluated on CAM, LRP and Occlusion. The general trend of reduced certification rates by increasing the certified radius R, as well as reduced certified top $K\%$ pixels by decreasing K holds across all five models.

Input layer LRP and RISE exhibit the highest robustness, as they also maintain a relatively high %certified scores across all three radii values in Figure 1. Interestingly, they also maintain a balance in certified top $K\%$ pixels by decreasing K in Figure 2.

Final layer Though Grad and GB still seem the most robust on CNNs in Figure 1, they localize very poorly, this is because they produce grid-like almost constant attributions at the final layer of the CNN architecture, due to max-pooling.

ViT-B/16 Gradient-based methods exhibit low robustness on ViT-B/16 comparable to their CNNs performance on both the input and final layers in Figure 1. Interesting, RISE maintains the same performance of having high robustness across both architecture types, with the highest at radius $R = 0.10$ at the final layer on ViT-B/16. Activation-based methods perform poorly on the transformer model compared to CNNs on the final layer. This is likely due to the lack of spatially structured convolutional features in ViTs, which affects the quality and stability of activation-based attributions when applied to token-based representations. We discuss this implementation detail in Section 5.

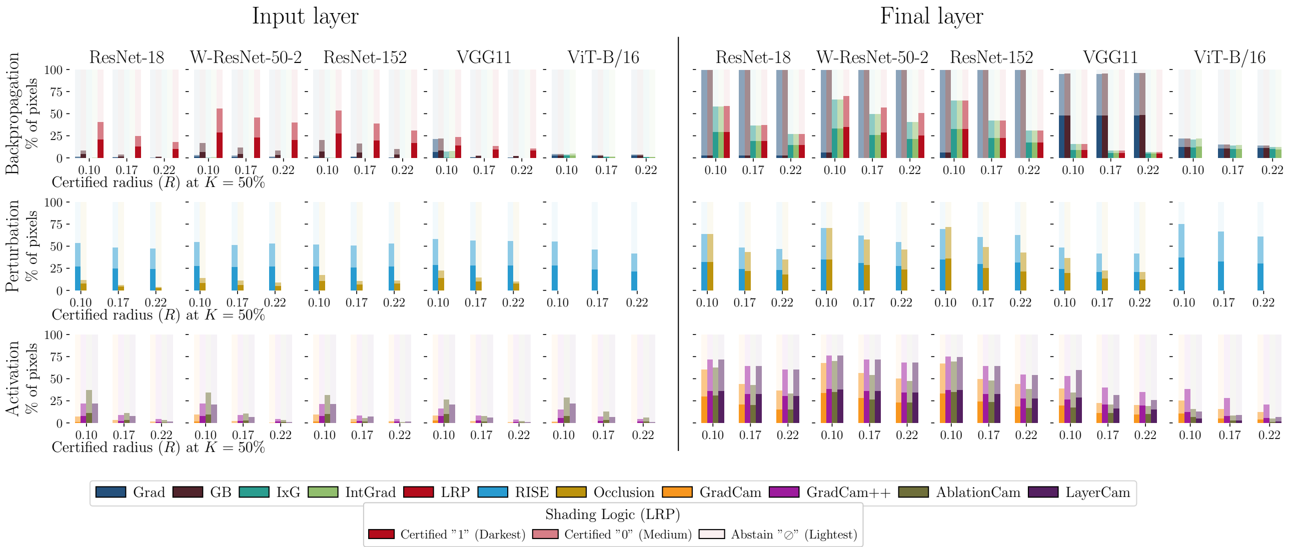

Figure 1: **Comparison of the per-pixel certification rate (%certified) against the certified radius R on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer. The darkest shades denote %certified pixels, while brightest denotes abstain $\oslash$.

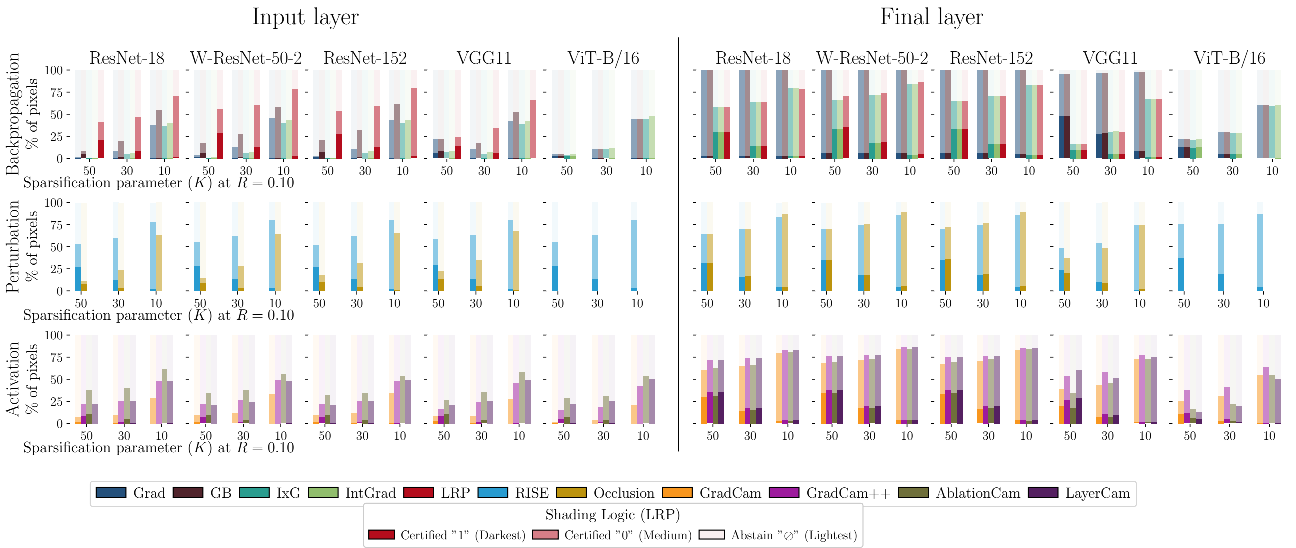

Figure 2: **Comparison of the per-pixel certification rate (%certified) against sparsification values K on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

1.2. Certified Localization (Certified GridPG)

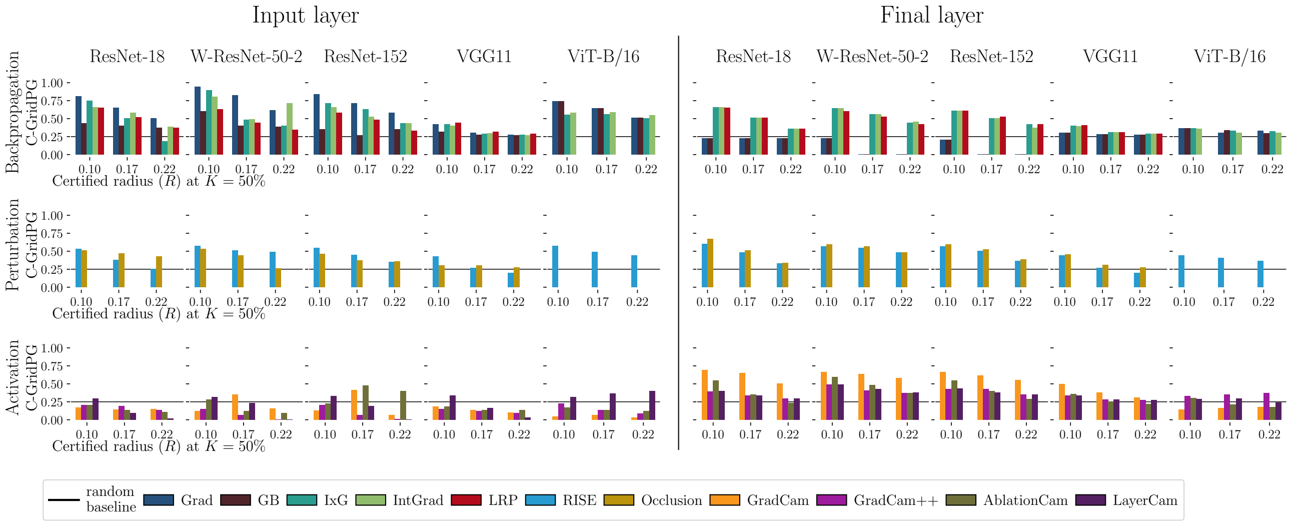

Figure 3: **Comparison of the certified localization (Certified GridPG) against the certified radius R on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

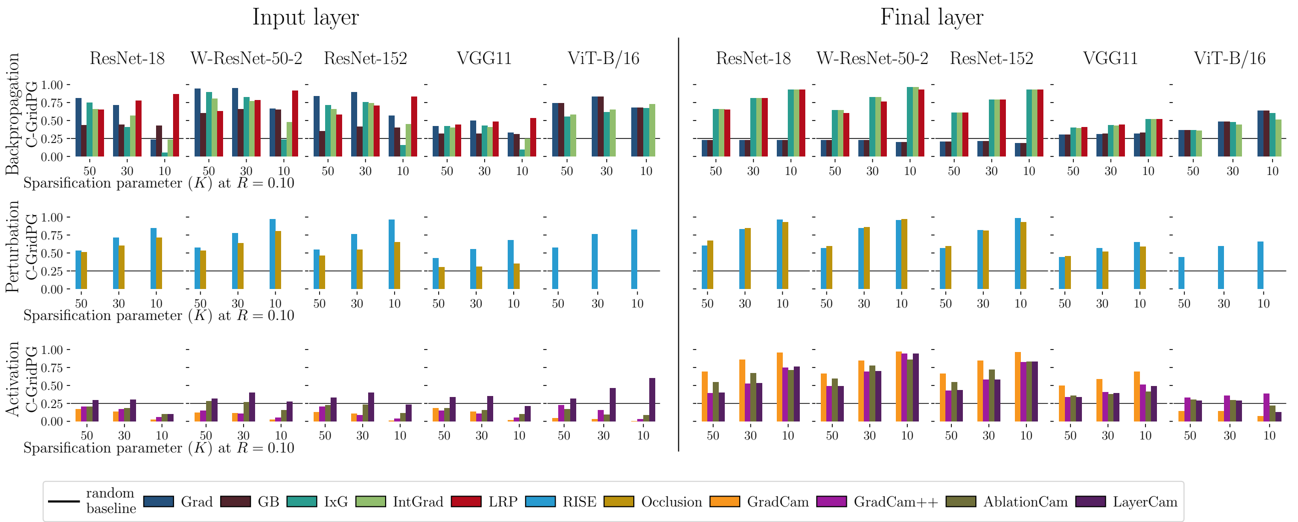

Figure 4: **Comparison of the certified localization (Certified GridPG) against the sparsification values K on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

1.3. Robustness vs. Localization Tradeoff

In Figure 5, we show the certified robustness and localization tradeoff across all five models. Similar to our original results previously on ResNet-18 (He et al., 2016), RISE strikes the best balance at both layers. Interestingly, LRP comes second, which outperforms the rest of backpropagation methods by a big margin.

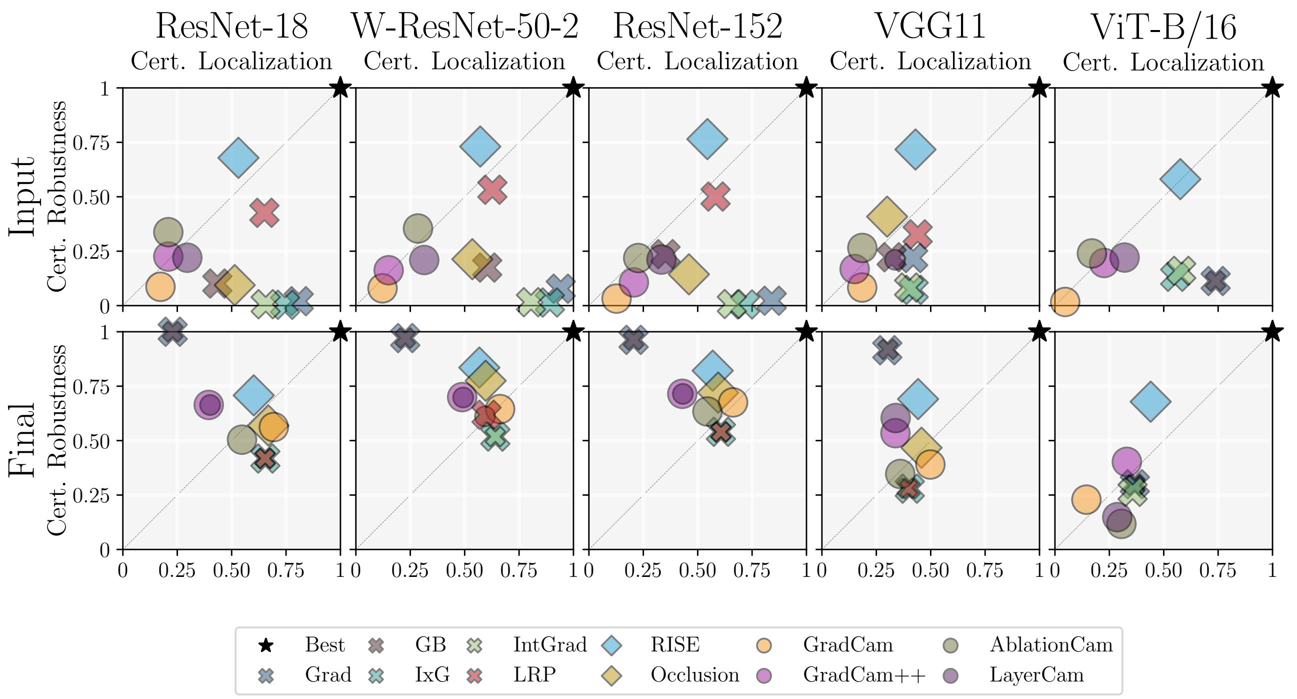

Figure 5: **Robustness-localization tradeoff of attribution methods on all 5 models using the %certified and Certified GridPG metrics.** (*Left*) evaluation is on input and (*Right*) final layers.

1.4. Certified attribution visuals on ViT-B/16

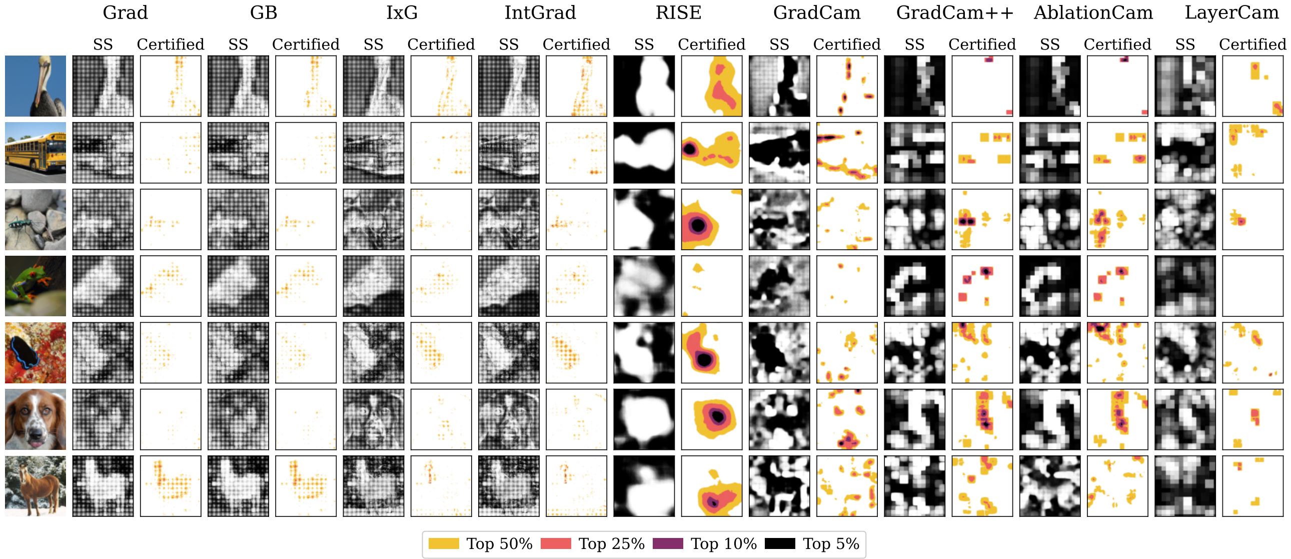

Figure 6: **Overlayed certified attributions on ViT-B/16 at different K values across methods .** SS (Smoothed Sparsified), which refers to the average of the sparsified attributions, is evaluated on $n = 100$ noisy input samples per image and at $K = 50\%$.

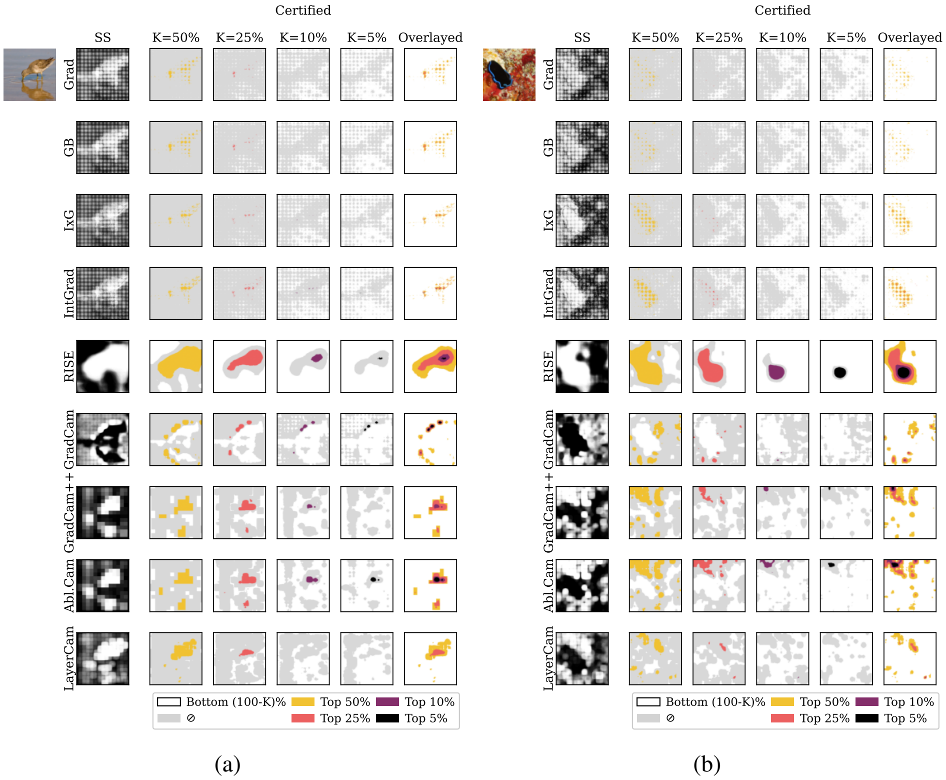

Figure 7: **Certified attribution maps of ViT-B/16 of all methods at different sparsification parameter (K) values**. SS (Smoothed Sparsified) is evaluated on $n = 100$ noisy input samples per image and at $K = 50\%$. The “Overlayed” column shows certified top $K\%$ pixels per row, with lower K taking precedence.

2. Faithfulness analysis of certified attributions

2.1. Deletion metric

We include a deletion-based (Petsiuk et al., 2018) faithfulness analysis on ResNet-18 and VGG-11 in Figure 8. Given a certified attribution map at different sparisifcation K values, we define a deletion step to remove the most important corresponding certified top $K\%$ pixels in the input image (i.e., the ones with the lowest K). We progressively remove the certified top $K\%$ pixels (lowest to highest) and measure the change in the ground truth (GT) class confidence by the model. If the drop in confidence is steep at ealier deletion steps, then this likely suggests that the certified attribution method does highlight class-discriminative features.

Input layer LRP and RISE cause the steepest confidence drop in the first removal step on ResNet-18 and VGG-11, indicating their high faithfulness. GradCam is the least faithful on both ResNet-18 and VGG-11, which causes an increase in the confidence in VGG-11, suggesting that it does not highlight any class-descriminative features at the input layer. This highly aligns with its near-random certified localization performance on the input layer in Figure 3 and 4. Interestingly, there is an alignment between certified localization and faithfulness, as LRP and RISE also have the highest Certified GridPG scores on the input layer across all five models in Figure 3.

Final layer At the final layer, most methods cause a prediction flip after the first deletion step, showing relatively high faithfulness compared to the input layer.

Some methods are faithful, but localize poorly (e.g., GB and Grad have random Certified GridPG in Figure 3), but show high faithfulness in Figure 8 on VGG-11 at the final layer. This suggests that the gird-like constant pattern produced at the final layer by both has a deletion effect that distorts the image, but not because it highlights class-discriminative features. This way, all three metrics: %certified, Certified GridPG and deletion-based faithfulness complement each other to understand different aspects of certified attributions, which is essential for producing trustworthy attributions in safety-critical domains.

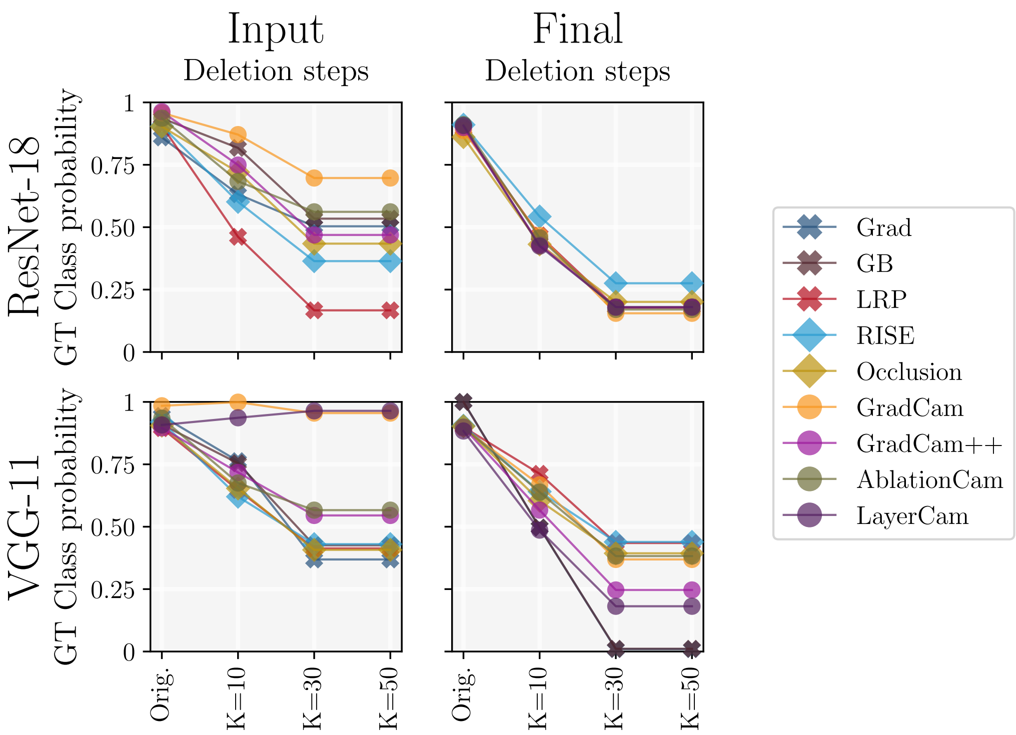

Figure 8: Faithful comparison of certified attribution methods using the ground truth (GT) class confidence against the deletion steps in which top K certified pixels are removed in descending order of importance.

3. Additional attribution methods: CAM and LRP

3.1. Robustness and Localization

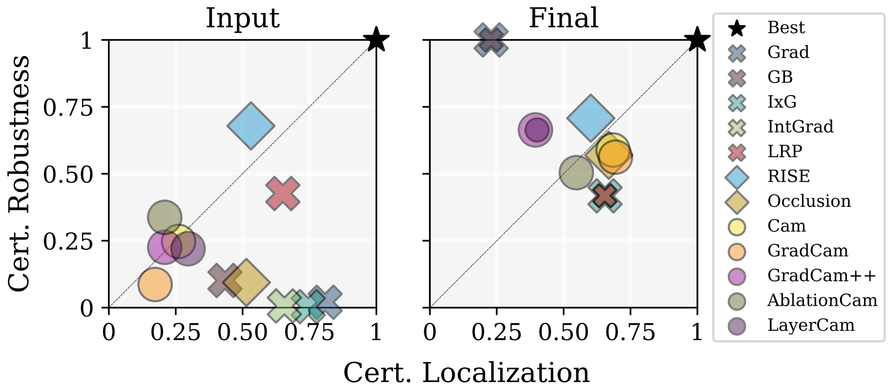

Figure 9: **Robustness-localization tradeoff of attribution methods on ResNet-18 including additionally LRP and CAM using the %certified and Certified GridPG metrics.** (*Left*) evaluation is on input and (*Right*) final layers.

3.2. Certified visuals on ResNet-18

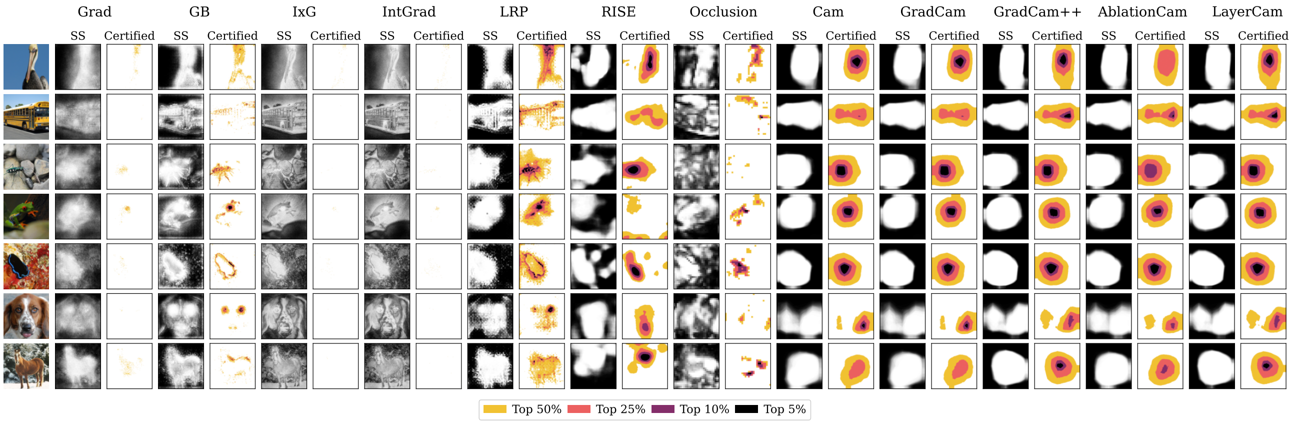

Figure 10: **Overlayed certified attributions on ResNet-18 at different K values across methods including LRP and CAM.** SS (Smoothed Sparsified), which refers to the average of the sparsified attributions, is evaluated on $n = 100$ noisy input samples per image and at $K = 50\%$.

4. Qualitative examples of Certified GridPG

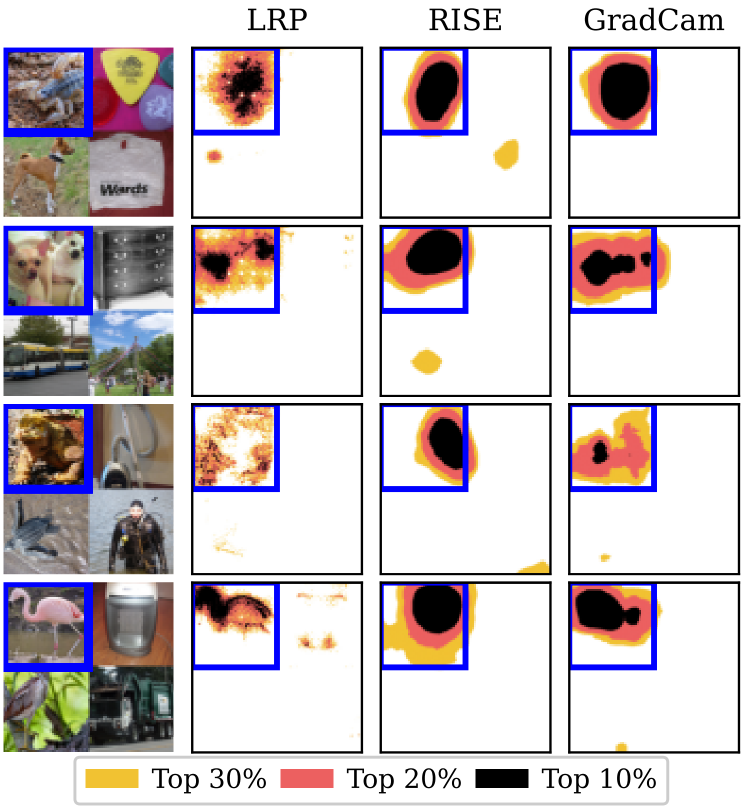

Figure 11: **Examples from overlayed certified grids at different sparsification K values**. The blue square denotes the correct subimage where top K% attributions should be concentrated.

5. Experimental setup

The setup includes default parameters from the submission, unless stated otherwise.

5.1. Attribution methods

In addition to the ten attribution methods in the submission, we extend our results to two new attribution methods: Layer-wise Relevance Propagation (Bach et al., 2015) and Class Activation Map (CAM) (Zhou et al., 2016).

5.2. Architecture

We evaluate all twelve attributions on five models, ResNet-18 (He et al., 2016), VGG-11, Wide ResNet-50-2 (Zagoruyko & Komodakis, 2016), ResNet-152 (He et al., 2016) and a transformer-based model ViT-B/16 (Dosovitskiy et al., 2020).

ViT-B/16 Implementation Details evaluating attribution methods at the final layer of a transformer-based model requires changes in the attribution logic. In CNNs, the final layer contains activations in the form of a 3D tensor of shape (C, W, H), where C is the number of channels and (H, W) corresponds to spatial locations at a lower dimension compared to the input image. This structure aligns naturally with the input image, and is usually interpolated (i.e., resized) to its dimensions.

However, Vistion Transformers (e.g., ViT-B/16) process the image as a sequences of patches. Each patch is considered a token. The final feature representation prior to classification is typically of the shape (N, D), where N is the number of tokens (including [CLS] token) and D is the hidden dimension (e.g., 768 for ViT-B/16). For ViT-B/16 and a 224×224 input image, this gives $N = 197$ tokens: 1 [CLS] token and 196 patch tokens, where each patch corresponds to a 14×14 grid in the original image (a result of diving 224 into 16 patches).

To enable attribution on ViTs, we discard the [CLS] token and reshape the remaining $(B, 14, 14)$, analogous to CNN activations. This allows us to treat patch embeddings as spatial features and apply any attribution method at the final layer similarly to how they are applied in CNNs. We adapted our attribution logic to extract these reshaped features, which are then interpolated to the original image size.

Note We have extended all attributions to be ViT-compatible with the exception of LRP (Bach et al., 2015), CAM (Zhou et al., 2016) and Occlusion (Zeiler & Fergus, 2014) due to limited rebuttal time.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference for Learning Representations (ICLR)*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint*, 2016.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.