568 ## Supplementary Material

569 In the appendices, we provide more details on the experimental settings as well as results.

570 ## Appendix A. Visualization of Best/Worst Case w.r.t. Fidelity Metrics

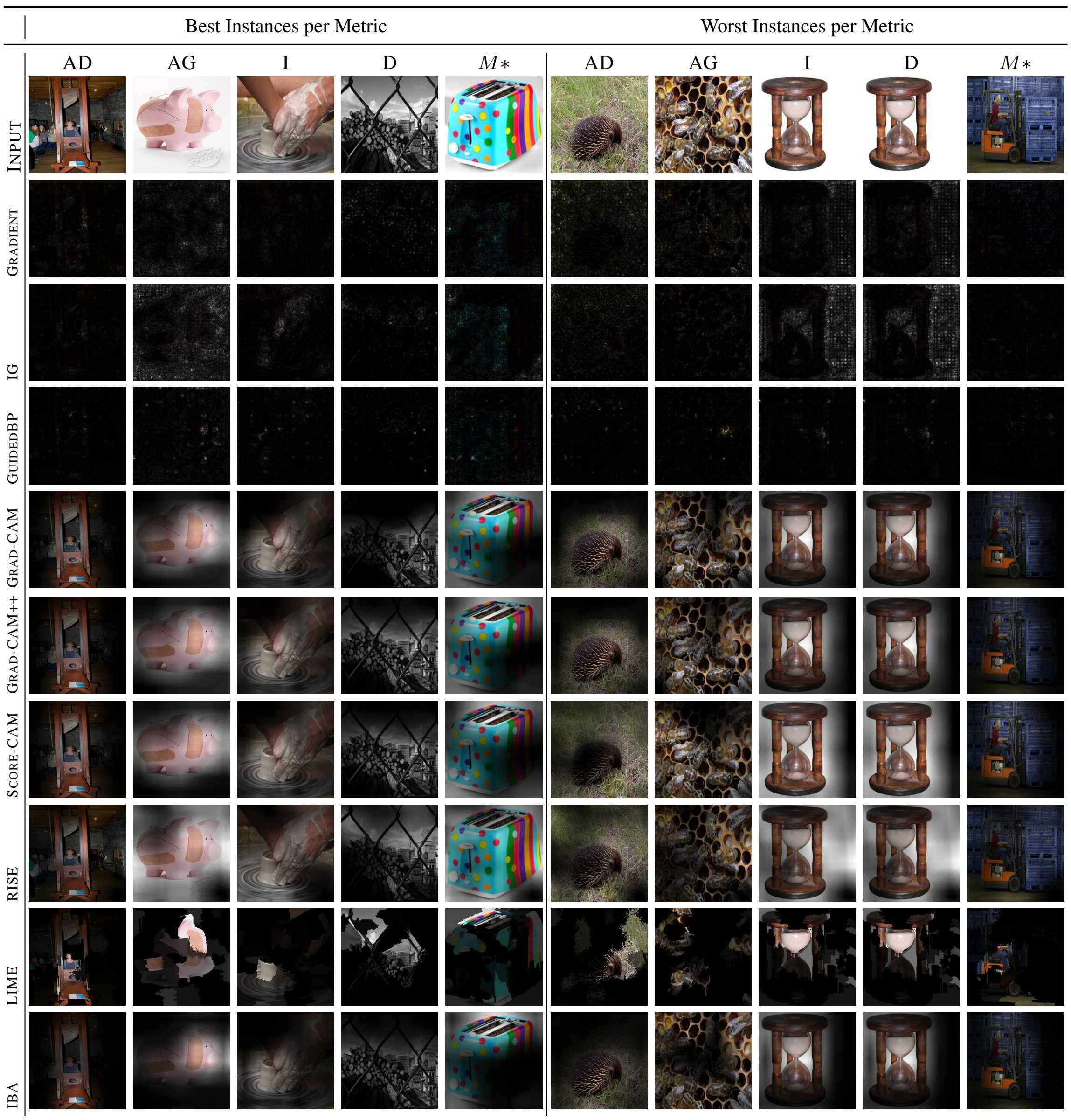

Figure A3: Images achieving best/worst performance w.r.t. each fidelity metric. $M*$ stands for Monotonicity. Our visualization of IG appears worse than those in the original paper due to different normalization techniques. Our primary focus, however, is on fair comparison on mathematical evaluation rather than visualization.

Appendix B. Experimentation with a Different Model

We extend the evaluation procedure to another popular model on the domain, namely VGG16 (Simonyan and Zisserman, 2015). Following evaluation procedure described previously, we replicate our experimentation for *Fidelity*, *Fidelity* ***vs.*** *Prediction* and *Robustness and Complexity*.

Fidelity Similar to the evaluation encountered on Table 1, VGG16 highlights similar occurrences. On Table A5 we observe that CAM-based methods consistently outperform other attributions regarding these metrics. Additionally, the shortcomings of fidelity metrics for image classification are maintained as Fake-CAM still ranks amongst the top for AI/AD/Monotonicity.

Methods		AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	Monotonicity
Uninformed	Fake-CAM	43.2	99.5	0.6	55.5	64.1	0.18
Gradient	Gradient	2.6	5.0	0	46.6	86.2	-0.19
	IG	2.8	5.7	0.1	49.3	90.2	0.28
	GuidedBP	2.6	4.9	0	46.9	85.4	-0.17
CAM	Grad-CAM	40.5	85.4	14.7	64.9	89.0	0.50
	Grad-CAM++	33.5	82.5	10.3	62.5	87.8	0.57
	Score-CAM	37.7	84.0	13.4	62.9	88.2	0.55
Occlusion	RISE	32.8	83.9	9.4	60.4	78.1	0.32
	LIME	10.1	34.2	2.5	60.8	84.9	0.30
Learning	IBA	28.2	76.0	8.6	63.3	88.5	0.57

Table A5: Evaluation of selected saliency mapping methods for different fidelity metrics w.r.t. the respective ground truth classes, where $\overline{\text{AD}} = 100 - \text{AD}$ and $\overline{\text{D}} = 100 - \text{D}$. This adjustment aligns all metrics so that higher values correspond to better performance.

Fidelity vs Prediction Extending the study case for multiple instances of class-specific behavior, in Table A6 we observe the consistency in our findings. On one hand, the performance of attributions generated for instances where we consider the ground truth class is as expected, optimal. On another hand contrasting experiments using ResNet50 in Table 2, we observe that while Score-CAM attains a higher performance in the first case mentioned; this is not the case on VGG16. With this in mind, and the small performance difference between this approach and Grad-CAM, we argue that the latter still maintains usefulness given its simplicity and competitive results.

Robustness and Complexity Lastly, on Table A7 we highlight robustness and complexity for VGG16. Our findings remain consistent with the observations made for ResNet50. In particular, we remark that while this family of metrics highlights mathematical properties, they do not describe adequately explainability.

Appendix C. Sensitivity to Transformation

Resize, Rotation, and Crop. When testing on Resize, Rotation, and Crop transformations, we use several parameter settings for each and report the average results in Table 4. For Resize, we resized from the original size, *i.e.*, 224×224, to 32×32, 64×64, 128×128, and 448×448. For Rotation, we chose angles of $45°$, $135°$, $255°$, and $315°$. For Crop, we performed random cropping with seeds 32, 44, 55, and 93. In this section, We provide the mean values with standard deviations

Methods		Ground Truth					Predicted Class					Least Probable				
		AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$
Uninformed	Fake-CAM	43.2	99.5	0.6	55.5	64.1	42.2	99.6	0.7	62.0	59.8	63.8	98.8	0	0	100
Gradient	Gradient	2.6	5.0	0	46.6	86.2	0	1	0	50.7	84.4	100	100	0	0	100
	IG	2.8	5.7	0.1	49.3	90.2	0.1	1.6	0	54.0	89.1	99.9	100	0	0	100
	GuidedBP	2.6	4.9	0	46.9	85.4	0	1.0	0	51.7	84.1	100	100	0	0	100
CAM	Grad-CAM	40.5	85.4	14.7	64.9	89.0	33.7	83.0	15.5	73.4	87.9	99.9	100	0	0	100
	Grad-CAM++	33.5	82.5	10.3	62.5	87.8	27.8	80.8	11.2	70.9	86.6	87.7	93.3	0	0	100
	Score-CAM	37.7	84.0	13.4	62.9	88.2	31.4	81.8	14.2	71.2	87.0	96.6	98.2	0	0	100
Occlusion	IBA	28.2	76.0	8.6	63.3	88.5	20.3	72.0	8.5	71.7	87.3	94.7	97.1	0	0	100
	RISE	32.8	83.9	9.4	60.4	78.1	30.8	84.0	11.7	68.5	79.8	85.8	93.0	0	0	100
Learning	LIME	10.1	34.2	2.5	60.8	84.9	4.7	29.2	2.1	68.3	83.3	98.5	99.1	0	0	100

Table A6: Evaluation of fidelity metrics with respect to different classes. Experimentation with VGG 16.

Methods		MS	AS	Sparseness	Complexity	EC
Uninformed	Fake-CAM	0.96	0.95	0	10.82	50175.0
Gradient	Gradient	0.97	0.94	42.4	10.5	50175.0
	IG	1.17	1.11	50.9	10.3	50174.4
	GuidedBP	1.60	1.50	42.5	10.5	50174.9
CAM	Grad-CAM	0.87	0.86	41.9	10.5	49307.2
	Grad-CAM++	0.87	0.86	39.7	10.5	50162.1
	Score-CAM	0.88	0.87	46.5	10.4	50120.1
Occlusion	RISE	0.91	0.90	30.9	10.67	50175.0
	LIME	0.92	0.91	72.5	9.73	26599.1
Learning	IBA	0.85	0.84	52.0	10.3	50173.1

Table A7: Evaluation of robustness and complexity metrics.

in Table A8, Table A9 and Table A10. The larger the standard deviation over the results, the more sensitive the results are. Comparing the standard deviation values between gradient-based methods and others, as well as between AG and others, confirms our key observations. Additionally, we find that the standard deviation is largest for resize, followed by rotation, and smallest for crop, indicating that fidelity is most sensitive to resizing and least sensitive to cropping.

Methods		AI	$\overline{AD}$	AG	I	$\overline{D}$
Basis	Fake-CAM	47.4 (4.7)	2.0 (1.8)	90.5 (9.0)	24.8 (21.2)	86.2 (12.3)
Gradient	Gradient	34.7 (34.4)	37.8 (35.5)	0.1 (0.1)	22.0 (18.2)	95.2 (4.6)
	IG	34.2 (34.3)	37.3 (35.3)	0.0 (0.0)	22.2 (18.3)	95.1 (4.9)
	GuidedBP	34.8 (34.8)	38.2 (35.5)	0.2 (0.2)	21.7 (17.8)	95.2 (4.9)
CAM	Grad-CAM	47.7 (20.3)	71.7 (10.7)	6.3 (6.3)	31.3 (18.8)	93.8 (5.0)
	Grad-CAM++	46.2 (21.2)	71.0 (10.4)	5.8 (5.6)	30.9 (18.6)	91.4 (3.8)
	Score-CAM	42.9 (10.9)	75.7 (10.3)	11.5 (7.4)	36.7 (19.7)	91.5 (5.4)
Occlusion	RISE	39.6 (16.7)	66.0 (6.6)	4.6 (3.6)	26.4 (21.7)	92.2 (6.8)
	LIME	29.9 (25.6)	40.6 (26.6)	2.7 (2.9)	27.2 (22.7)	93.1 (5.1)
Learning	IBA	33.6 (12.1)	57.4 (12.0)	4.9 (6.0)	26.9 (23.5)	93.1 (6.3)

Table A8: Report the average (standard deviation) of four different **Resize** settings.

Methods		AI	$\overline{AD}$	AG	I	$\overline{D}$
Basis	Fake-CAM	65.6 (1.7)	97.0 (0.4)	4.0 (0.9)	24.3 (6.8)	82.6 (4.7)
Gradient	Gradient	8.5 (4.0)	13.2 (5.2)	0.0 (0.0)	18.9 (6.4)	95.8 (1.3)
	IG	7.7 (2.8)	11.8 (4.1)	0.0 (0.0)	21.8 (6.7)	96.7 (0.9)
	GuidedBP	9.1 (3.9)	14.1 (5.3)	0.1 (0.1)	19.4 (6.1)	96.9 (0.9)
CAM	Grad-CAM	54.2 (1.7)	79.2 (1.7)	9.7 (5.6)	29.5 (8.1)	96.9 (1.0)
	Grad-CAM++	51.7 (1.9)	78.8 (1.5)	11.5 (2.2)	29.1 (8.1)	96.8 (1.0)
	Score-CAM	57.4 (1.5)	81.8 (2.1)	14.9 (3.1)	29.5 (8.1)	96.5 (1.1)
Occlusion	RISE	45.4 (3.4)	73.6 (1.5)	8.6 (1.6)	27.7 (7.7)	94.5 (2.1)
	LIME	15.3 (2.8)	26.2 (2.6)	1.5 (0.2)	27.8 (7.9)	95.1 (1.2)
Learning	IBA	45.3 (3.9)	72.4 (1.0)	9.2 (1.4)	29.3 (8.1)	96.1 (1.1)

Table A9: Report the average (standard deviation) of four different **Rotation** settings.

Methods		AI	$\overline{AD}$	AG	I	$\overline{D}$
Basis	Fake-CAM	47.5 (1.4)	98.5 (0.2)	1.7 (0.1)	49.9 (0.3)	71.2 (0.7)
Gradient	Gradient	3.0 (0.1)	5.2 (0.3)	0.0 (0.0)	43.3 (0.3)	92.5 (0.2)
	IG	3.1 (0.2)	5.5 (0.2)	0.0 (0.0)	43.5 (1.0)	92.8 (0.1)
	GuidedBP	3.1 (0.2)	5.7 (0.2)	0.0 (0.0)	42.9 (0.4)	93.3 (0.2)
CAM	Grad-CAM	37.3 (1.4)	78.2 (1.4)	11.4 (1.2)	55.3 (1.3)	78.8 (22.7)
	Grad-CAM++	35.5 (1.1)	76.0 (1.3)	10.3 (0.7)	52.4 (1.5)	90.4 (0.4)
	Score-CAM	41.7 (1.0)	80.4 (1.6)	14.5 (0.4)	53.8 (1.1)	89.5 (0.2)
Occlusion	RISE	27.8 (1.3)	69.4 (0.9)	7.5 (0.5)	51.9 (1.2)	84.2 (2.7)
	LIME	7.2 (0.5)	18.4 (0.9)	1.5 (0.1)	53.8 (0.7)	88.6 (0.2)
Learning	IBA	26.3 (0.7)	65.0 (0.8)	7.2 (0.3)	54.0 (1.1)	89.4 (0.1)

Table A10: Report the average (standard deviation) of four different **Crop** settings.

Mixup. Since Mixup generates synthetic images by interpolating two images from different classes, we refer to these as the first class and second class. Given that we assign equal weights to both classes, the statistics of the evaluation metrics on saliency maps generated from either class should be from the same distribution. Therefore, we only evaluate the results generated based on the first class. We assess the performance using both the first and second classes in Table A11. The saliency map generated for the first class should not highlight regions contributing to the prediction of the second class. However, as shown in Table A11, the performance of AI/D is better for the second class, and AD also shows better performance when evaluated with respect to the second class. This again highlights the failure of AI/AD/I. Conversely, AG and I perform as expected. However, the observation that the numbers for AG and I are quite low indicates a failure of the saliency map method. When visualizing the saliency maps generated by Grad-CAM for the first and second classes in Figure A4, we find that many images highlight the same region for different classes.

Methods		First					Second				
		AI	$\overline{AD}$	AG	I	$\overline{D}$	AI	$\overline{AD}$	AG	I	$\overline{D}$
Basis	Fake-CAM	49.6	97.8	0.6	20.9	86.6	53.7	97.0	0.1	2.0	98.3
Gradient	Gradient	40.8	44.6	0.1	17.6	95.6	73.6	77.7	0.1	1.7	99.7
	IG	39.7	43.7	0.1	18.8	95.3	74.0	78.5	0.1	1.9	99.6
	GuidedBP	43.1	47.4	0.1	17.2	95.8	76.6	80.6	0.1	1.8	99.6
CAM	Grad-CAM	64.8	84.8	7.0	25.1	96.3	66.7	75.7	1.0	3.0	99.1
	Grad-CAM++	54.1	79.1	6.6	24.9	96.0	57.3	71.1	0.7	2.6	99.3
	Score-CAM	61.5	83.4	8.3	24.1	96.0	66.7	75.7	1.0	3.0	99.1
Occlusion	RISE	55.7	79.6	5.3	22.1	93.7	62.6	75.0	0.5	2.1	99.2
	LIME	41.9	51.1	0.9	23.1	95.3	69.8	75.2	0.3	2.5	99.4
Learning	IBA	54.6	77.1	5.2	19.3	93.6	62.9	74.0	0.5	2.7	99.1

Table A11: Report the fidelity performance of the saliency maps generated according to the first class and evaluate their performance with respect to both the first and second class.

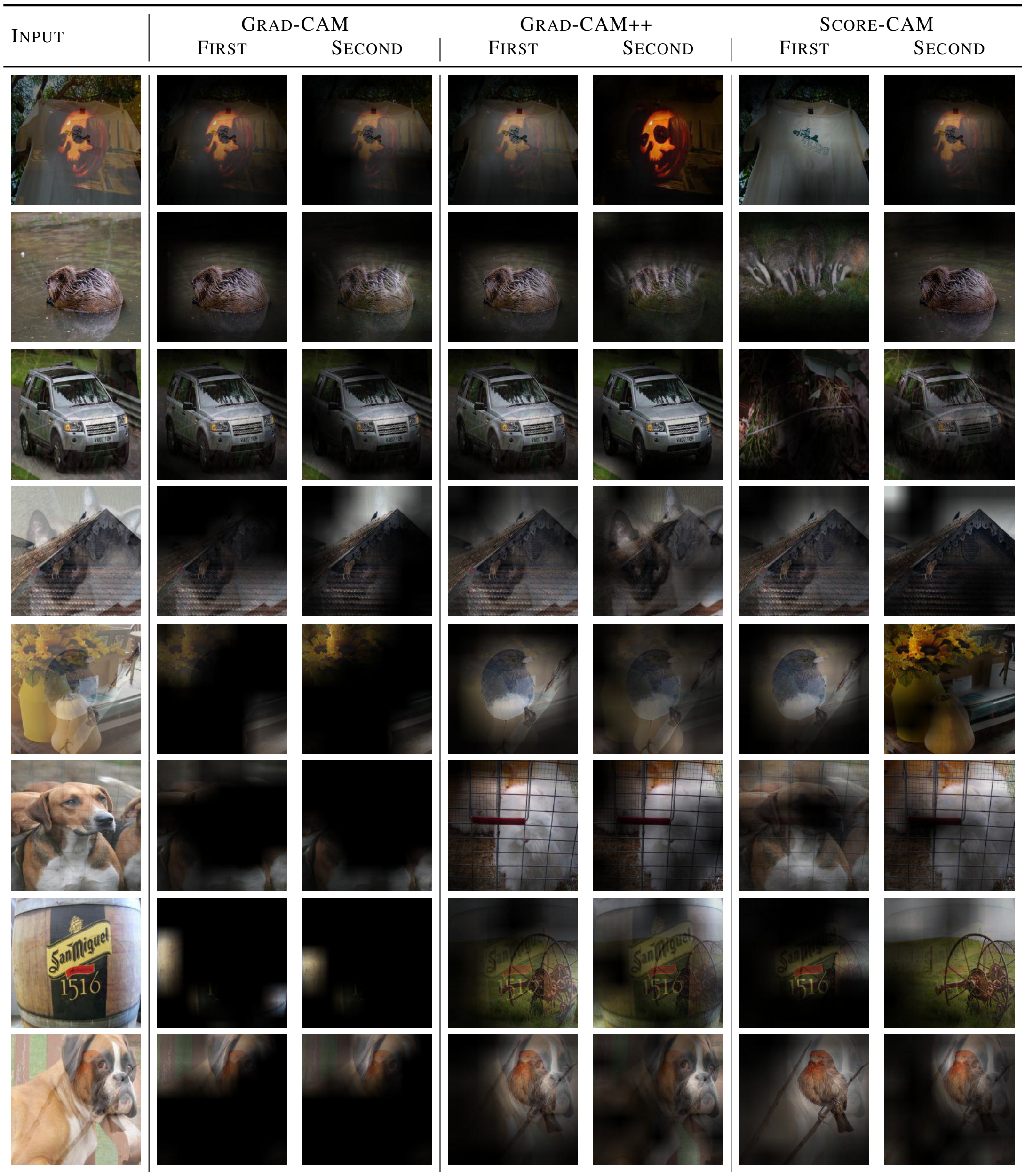

Figure A4: Saliency maps generated by Grad-CAM w.r.t. first and second classes.