Optimizing Explanations by Network Canonization and Hyperparameter Search - Supplementary Material -

Frederik Pahde[1] Galip Ümit Yolcu[1,2] Alexander Binder[3,4]
Wojciech Samek[1,2,5,*] Sebastian Lapuschkin[1,*]

[1]Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, Germany

[2]Technische Universität Berlin, Germany

[3]ICT Cluster, Singapore Institute of Technology, Singapore

[4]University of Oslo, Norway

[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Germany

$*$ corresponding authors {wojciech.samek|sebastian.lapuschkin}@hhi.fraunhofer.de

A. Canonization Details

Linear $\rightarrow$ BN: In Eqs. (A.1) - (A.5), we show more detailed steps required to fuse Linear $\rightarrow$ BN components into a single affine transformation, as outlined in Eq. (4) in the main paper:

$$f(\mathbf{x}) = \text{BN}(\text{Linear}(\mathbf{x})) \tag{A.1}$$
$$= \text{BN}(w_L^\top \mathbf{x} + b_L) \tag{A.2}$$
$$= w_{BN}\Big(\frac{w_L^\top \mathbf{x} + b_L - \mu}{\sqrt{\sigma+\epsilon}}\Big) + b_{BN} \tag{A.3}$$
$$= \frac{w_{BN}}{\sqrt{\sigma+\epsilon}}(w_L^\top \mathbf{x} + b_L - \mu) + b_{BN} \tag{A.4}$$
$$= \underbrace{(\frac{w_{BN}}{\sqrt{\sigma+\epsilon}} w_L)^\top}_{w_{\text{new}}} \mathbf{x} + \underbrace{\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}(b_L - \mu) + b_{BN}}_{b_{\text{new}}} \tag{A.5}$$

BN $\rightarrow$ Linear: In Eqs. (A.6) - (A.9), we show more detailed steps required to fuse BN $\rightarrow$ Linear component chains into a single affine transformation, as outlined in Eq. (6) in the main paper:

$$f(\mathbf{x}) = \text{Linear}(\text{BN}(\mathbf{x})) \tag{A.6}$$
$$= w_L^\top \left(w_{BN}\Big(\frac{\mathbf{x}-\mu}{\sqrt{\sigma+\epsilon}}\Big) + b_{BN} \right) + b_L \tag{A.7}$$
$$= w_L^\top \Big(\frac{w_{BN}\mathbf{x} - w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN}\Big) + b_L \tag{A.8}$$
$$= \underbrace{\frac{w_L^\top w_{BN}}{\sqrt{\sigma+\epsilon}}}_{w_{\text{new}}} \mathbf{x} \underbrace{- \frac{w_L^\top w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + w_L^\top b_{BN} + b_L}_{b_{\text{new}}} \tag{A.9}$$

Padding Issue in BN $\rightarrow$ Linear Canonization: If the linear layer is a Convolutional layer with constant valued padding, the bias of the linear layer after canonization can no longer be shown as a scalar:

$$f(\mathbf{x}) = \text{Conv}(\text{Pad}(\text{BN}(\mathbf{x}))) \tag{A.10}$$
$$= \text{Conv}(\text{Pad}(\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}\mathbf{x} - \frac{w_{BN}\mu}{\sqrt{\mu+\epsilon}} + b_{BN})) \tag{A.11}$$
$$= \text{Conv}(\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}\text{Pad}(\mathbf{x}) + \text{Pad}(-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN})) \tag{A.12}$$
$$= w_L * \Big(\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}\text{Pad}(\mathbf{x}) + \text{Pad}(-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN})\Big) + b_L \tag{A.13}$$
$$= \Big(w_L^\top \frac{w_{BN}}{\sqrt{\sigma+\epsilon}}\Big) * \text{Pad}(\mathbf{x}) + w_L * \text{Pad}(-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN}) + b_L \tag{A.14}$$
$$= \underbrace{\Big(w_L^\top \frac{w_{BN}}{\sqrt{\sigma+\epsilon}}\Big)}_{w_{\text{new}}} * \text{Pad}(\mathbf{x}) + \underbrace{\text{Conv}(\text{Pad}(-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN}))}_{b_{\text{new}}} \tag{A.15}$$

In the equations above, $*$ stands for convolution. The new bias term is a full feature map, as opposed to a scalar as in linear layers without padding. The feature map does not depend on input $\mathbf{x}$ and is computed by putting a feature map (of the same size as $\mathbf{x}$) with all features equal to $-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN}$ through the original linear layer. Notice that if the padding value is nonzero, then the padding value of

the canonized layer must be scaled by $\frac{\sqrt{\sigma+\epsilon}}{w_{BN}}$

We now give a simple example to help illustrate the problem and the proposed solution. Specifically, we set BN parameters $\mu = 0, \sigma = 1, \epsilon = 0, w_{\text{BN}} = 1, b_{\text{BN}} = 1$. Furthermore we define a single Convolutional filter with zero padding of width 1, $w_L = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and no bias, $b_L = 0$. Finally, we choose to show the case of a simple 3×3 input feature map

$$\mathbf{x} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix} \tag{A.16}$$

$$Pad(BN(\mathbf{x})) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 3 & 4 & 0 \\ 0 & 3 & 4 & 5 & 0 \\ 0 & 4 & 5 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{A.17}$$

$$w_L * Pad(BN(\mathbf{x})) = \begin{bmatrix} 2 & 5 & 7 & 4 \\ 5 & 12 & 16 & 9 \\ 7 & 16 & 20 & 11 \\ 4 & 9 & 11 & 6 \end{bmatrix} \tag{A.18}$$

$$= \begin{bmatrix} 1 & 3 & 5 & 3 \\ 3 & 8 & 12 & 7 \\ 5 & 12 & 16 & 9 \\ 3 & 7 & 9 & 5 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 2 & 1 \\ 2 & 4 & 4 & 2 \\ 2 & 4 & 4 & 2 \\ 1 & 2 & 2 & 1 \end{bmatrix} \tag{A.19}$$

$$= w_{\text{new}} * Pad(\mathbf{x}) + w_L * Pad(BNBias) + b_L \tag{A.20}$$

where and $BNBias = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ is the feature map composed of values equal to $-\frac{w_{BN}\mu}{\sqrt{\sigma+\epsilon}} + b_{BN}$

B. Canonization of Relation Networks

Architecture: Relation Networks [4] are the state-of-the-art model architecture for the CLEVR dataset. It uses two separate encoders for image and text input. For the image, a simple convolutional neural network is used with 4 blocks, each containing a Convolutional layer, followed by a ReLU activation function and a BN layer. The text input is processed by a LSTM. The pixels from the feature map from the last Convolutional block from the image encoder are pair-wise concatenated along with their coordinates and the text encoding. This representation is then passed to a 4-layer fully connected network, summed up and then processed by a 3-layer fully connected network with ReLU activation.

Canonization: Relation Networks, as implemented by the authors of [1], use Batch Norm (BN) layers at the end of each block *not* directly after Convolutional layers. Therefore, we suggest to merge the BN layers with the Convolutional layers at the beginning of the following block. The BN layer of the last block of the image encoder can be merged into the following fully connected layer. However, attention as to be paid to make sure only weights operating on activations coming from the image encoder are updated, as outlined below. The proposed canonization of Relation Networks is visualized in Fig. A.1

Challenge: In relation networks, the last BN layer of the image encoder has to be merged into a linear layer of the succeeding block, which takes as input a concatenation of image pairs, text and indices. Therefore, only the weights responsible for the activations coming from the image encoder have to be updated.

Incoming Activations:

$$\mathbf{x} = \text{concat}\overbrace{\Big[\underbrace{[\mathbf{x1}]}_{24}, \underbrace{[\text{coord1}]}_{2}, \underbrace{[\mathbf{x2}]}_{24}, \underbrace{[\text{coord2}]}_{2}, \underbrace{[\text{question}]}_{128}\Big]}^{180} \tag{A.21}$$

Only **x1** and **x2** pass the BN layer, i.e., indices $0:24$ and $26:50$ have to be updated. Here, the indexing $i:j$ signifies the elements with indices from i to $j-1$, where the first element is indexed with 0. In order to update only the relevant part of the weights of the linear layer w_L, we have to split them into:

$$w_L = \text{concat}\Big[\underbrace{[w_L^{0:24}]}_{\mathbf{x1}}, \underbrace{[w_L^{24:26}]}_{\text{coord1}}, \underbrace{[w_L^{26:50}]}_{\mathbf{x2}}, \underbrace{[w_L^{50:52}]}_{\text{coord2}}, \underbrace{[w_L^{52:180}]}_{\text{text}}\Big] \tag{A.22}$$

Using Eq. (A.9), each relevant weight part can than be updated as follows:

$$w_{L_{\text{new}}}^{0:24} = \frac{{w_L^{0:24}}^\top w_{BN}}{\sqrt{\sigma+\epsilon}} \tag{A.23}$$

$$w_{L_{\text{new}}}^{26:50} = \frac{{w_L^{26:50}}^\top w_{BN}}{\sqrt{\sigma+\epsilon}} \tag{A.24}$$

This gives a new weight matrix:

$$w_{L_{\text{new}}} = \text{concat}\Big[\underbrace{[w_{L_{\text{new}}}^{0:24}]}_{\mathbf{x1}}, \underbrace{[w_L^{24:26}]}_{\text{coord1}}, \underbrace{[w_{L_{\text{new}}}^{26:50}]}_{\mathbf{x2}}, \underbrace{[w_L^{50:52}]}_{\text{coord2}}, \underbrace{[w_L^{52:180}]}_{\text{text}}\Big] \tag{A.25}$$

Similarly, the new bias can be calculated as:

$$b_{L_{\text{new}}} = w_L^{0:24} b_{\text{linBN}} + w_L^{26:50} b_{\text{linBN}} + b_C \tag{A.26}$$

with $b_{\text{linBN}} = b_{BN} - \frac{w_B N \cdot \mu}{\sqrt{\sigma+\epsilon}}$

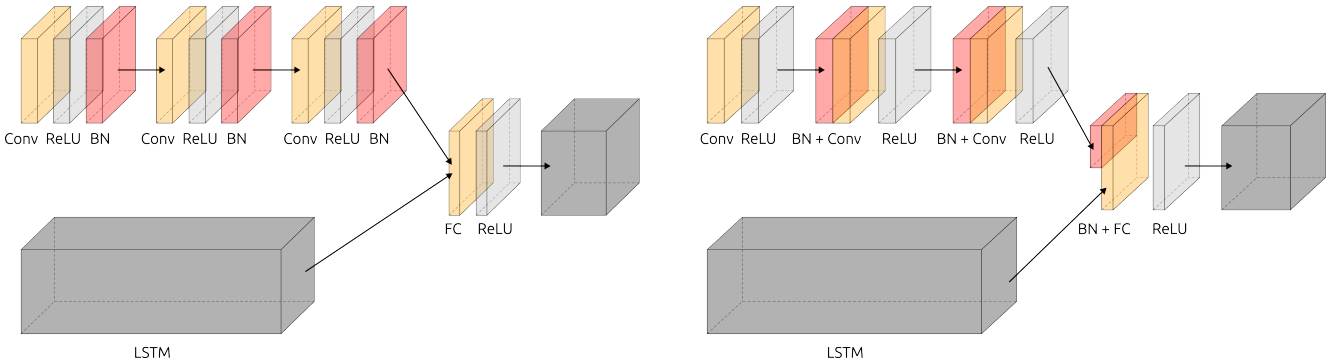

Figure A.1. Canonization of Relation Network. *(Left):* Part from the original Relation Network. *(Right):* Suggested canonization for the corresponding part of Relation Networks. BN layers are merged into the Convolutional layer at the beginning of the following block. The BN layer of the last block is merged into the following fully connected layer. However, only weights operating upon activations coming from the image encoder are updated, as outlined in Eq. (A.25).

C. Composites

The composites, i.e., pre-defined layer-to-rule assignments as suggested in the literature, that we used in the paper, are described in Tab. A.1.

Table A.1. Details for composites used in our experiments.

Composite	Layer Type	Rule
LRP-$\epsilon+$	Convolutional Fully Connected	$\alpha1\beta0$-rule ϵ-rule
LRP-$\alpha2\beta1$	Convolutional Fully Connected	$\alpha2\beta1$-rule ϵ-rule
LRP-Custom (RN)	First Convolutional Other Convolutionals Fully Connected	box-rule $\alpha1\beta0$-rule $\alpha1\beta0$-rule

D. Pascal VOC 2012 Experiments

D.1. Dataset Description

Pascal Visual Object Classes (VOC) 2012 dataset has images from 20 categories, including 5717 training samples and 5823 validation samples with bounding box annotations, along with a private test set. From those, 1464 training samples and 1449 validation samples are annotated with binary segmentation masks. As opposed to ILSVRC2017, the images are much more diverse in composition. Many images contain multiple instances of several categories. The dataset does not suffer from the center-bias mentioned for ILSVRC2017. In the experiments, we use the validation samples with segmentation masks. Due to the robustness of models to input perturbations, the faithfulness correlation scores are very low, even entirely zero for some models. In order to obtain more meaningful results, we report faithfulness correlation scores with bigger perturbations compared to the ILSVRC2017 experiments.

D.2. Models

We evaluate explanations on VGG-16, ResNet-18, ResNet-50, EfficientNet-B0, EfficientNet-B4, DenseNet-121 and DenseNet-161. We fine tune models using the full training set, using pre-trained models from the PyTorch model zoo. We use stochastic gradient descent as the learning algorithm, with a cosine annealing learning rate scheduler [3]. We use sum of binary cross entropy losses as the loss, and train the networks until convergence. We opted to use a smaller learning rate for pretrained parameters.

D.3. Results

The results are shown in Tables A.2 - A.8. The results suggest that canonization increases performance in the localization metrics. The randomization metric is generally improved for the LRP-$\epsilon+$ and LRP-$\alpha2\beta1$ rules, however it lowers the randomization scores of Excitation Backprop explainers (equivalent to LRP-$\alpha1\beta0$). For the robustness metrics, canonization helps for Excitation Backpropagation. However, it makes robustness scores improve by a bigger margin for all methods for DenseNet models. Similar to the results for ILSVRC2017, DenseNet models seem to be affected negatively in their complexity scores when canonized. For other model architectures, complexity measures are also uniformly improved by canonization.

E. MS-Coco 2014 Experiments

E.1. Classes

In our experiments we considered the following 10 classes that were picked randomly:

"person", "bear", "umbrella", "suitcase", "kite", "surf board", "wine glass", "carrot", "toilet", "tv".

Table A.2. Results for Pascal VOC XAI evaluation with VGG-16. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.59	**0.60**	0.04	0.04	0.21	**0.22**	0.33	**0.34**	0.40	**0.41**	1.00	1.00	0.24	**0.20**	0.26	**0.21**
LRP-$\alpha 2\beta 1$	0.75	**0.86**	**0.04**	0.03	0.27	0.27	0.38	**0.44**	0.40	**0.41**	**0.53**	0.64	0.53	0.53	0.80	**0.78**
LRP-ε+	0.59	**0.68**	0.05	0.05	0.25	**0.26**	0.36	**0.40**	0.46	**0.47**	0.49	0.49	0.51	0.51	0.80	**0.78**

Table A.3. Results for Pascal VOC XAI evaluation with ResNet-18. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.53	**0.57**	0.03	**0.04**	0.22	**0.23**	0.31	**0.33**	0.40	**0.41**	**0.96**	0.97	0.17	**0.15**	0.19	**0.16**
LRP-$\alpha 2\beta 1$	0.70	**0.77**	0.03	0.03	0.23	0.23	0.34	**0.39**	0.36	**0.39**	0.65	**0.64**	0.48	0.48	0.77	0.77
LRP-ε+	0.56	**0.63**	0.03	0.03	0.22	0.22	0.31	**0.34**	0.39	**0.41**	0.68	0.68	0.46	**0.45**	0.77	**0.76**

E.2. Dataset Description

The MS Coco 2014 dataset has images from 80 categories, including about 83000 training and 4100 validation samples with segmentation mask annotations, along with a private test set. This dataset is also much more diverse compared to ILSVRC2017 and does not suffer from the center-bias.

E.3. Models

We evaluate explanations on VGG-16, ResNet-18, EfficientNet-B0 and DenseNet-121. Similar to the Pascal VOC dataset, we fine tuned models using pre-trained models from the PyTorch model zoo. We use stochastic gradient descent as the learning algorithm. We use sum of binary cross entropy losses as the loss, and train the networks for until convergence. We opted to use a smaller learning rate for pretrained parameters. We excluded hard examples while training and during the evaluations.

E.4. Results

The results are shown in Tables A.9 - A.12. The results suggest that for MS COCO, canonization helps explanations in all metrics, except for Faithfulness and Randomization metrics where results are not consistent across model architectures.

F. ILSVRC2017 Experiments

F.1. Classes

In our experiments we considered the following 50 classes that were picked randomly:

"Bernese_mountain_dog", "Christmas_stocking", "Gila_monster", "Shetland_sheepdog", "Windsor_tie", "amphibian", "ant", "bubble", "cassette", "cicada", "collie", "crossword_puzzle", "dalmatian", "eft", "file", "flute", "goldfish", "gorilla", "gown", "grasshopper", "green_snake", "gyromitra", "hammer", "hen_of_the_woods", "indigo_bunting", "kimono", "magnetic_compass", "mongoose", "mountain_tent", "otterhound", "palace", "patio", "pencil_sharpener", "platypus", "pomegranate", "pool_table", "redshank", "refrigerator", "rhinoceros_beetle", "screw", "screwdriver", "shoe_shop", "shopping_basket", "stage", "standard_poodle", "stethoscope", "toaster", "tree_frog", "vase", "wolf_spider".

F.2. Additional Results

In Tables A.13 - A.19 we show additional results for our experiments with the ILSVRC2017 dataset. Specifically, in addition to the architectures evaluated in the main paper, we present results for ResNet50, EfficientNet-B4 and DenseNet-161. Moreover, we include results for Faithfulness Correlation [2] and Max Sensitivity [6].

Table A.4. Results for Pascal VOC XAI evaluation with ResNet-50. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.55	**0.63**	0.04	0.04	0.20	**0.23**	0.30	**0.35**	0.38	**0.42**	0.96	**0.93**	0.20	**0.17**	0.21	**0.18**
LRP-$\alpha 2\beta 1$	0.73	**0.81**	0.03	0.03	0.24	0.24	0.38	**0.41**	0.39	0.39	**0.60**	0.62	**0.48**	0.49	0.74	**0.73**
LRP-ε+	0.60	**0.69**	0.04	0.04	0.24	**0.25**	0.35	**0.39**	0.44	0.44	0.60	**0.57**	0.46	0.46	0.75	**0.74**

Table A.5. Results for Pascal VOC XAI evaluation with EfficientNet-B0. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.55	**0.69**	0.01	**0.02**	0.13	**0.18**	0.26	**0.37**	0.30	**0.41**	1.00	**0.99**	0.43	**0.39**	0.50	**0.43**
LRP-$\alpha 2\beta 1$	0.73	**0.76**	0.00	0.00	**0.18**	0.11	**0.37**	0.34	**0.39**	0.34	0.63	**0.51**	0.58	0.58	0.80	**0.79**
LRP-ε+	0.51	**0.72**	0.02	0.02	0.17	**0.19**	0.30	**0.39**	0.41	**0.43**	**0.59**	0.67	**0.46**	0.49	0.77	0.77

G. Additional CLEVR-XAI Results

In Table A.20, we show additional results for our CLEVR-XAI experiments. Specifically, in additional to *pos-l2-norm-sq* pooling, we also present results for *max-norm* pooling.

H. Attribution Heatmaps

In Figures A.2 - A.8 we show attribution heatmaps for three samples using various XAI methods, both with and without model canonization using the ILSVRC2017 dataset for different model architectures. Similarly, in Figures A.9 - A.10 we show attribution heatmaps for different XAI methods with and without canonization for Relation Networks using *pos-l2-norm-sq* pooling and *max-norm* pooling.

Table A.6. Results for Pascal VOC XAI evaluation with EfficientNet-B4. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.68	**0.77**	0.0	0.0	**0.14**	0.13	0.33	**0.36**	0.37	0.37	**0.97**	1.00	**0.32**	0.36	**0.37**	0.43
LRP-$\alpha 2\beta 1$	0.71	**0.75**	0.0	0.0	**0.12**	0.08	0.27	**0.32**	0.31	**0.33**	0.46	**0.42**	0.62	**0.61**	0.93	**0.91**
LRP-ε+	0.54	**0.77**	0.0	0.0	0.15	**0.16**	0.28	**0.40**	0.36	**0.42**	**0.51**	0.57	0.51	0.51	0.88	**0.85**

Table A.7. Results for Pascal VOC XAI evaluation with DenseNet-121. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.63**	0.59	0.02	**0.04**	0.11	**0.25**	0.25	**0.37**	0.28	**0.48**	0.98	**0.75**	0.60	**0.20**	1.07	**0.22**
LRP-ε+	**0.70**	0.63	0.02	**0.04**	0.18	**0.24**	0.34	**0.36**	0.39	**0.44**	**0.35**	0.44	0.63	**0.50**	1.08	**0.77**
LRP-$\alpha 2\beta 1$	**0.83**	0.73	0.01	**0.03**	0.16	**0.24**	0.31	**0.39**	0.32	**0.41**	0.38	**0.36**	0.64	**0.51**	1.11	**0.76**

Table A.8. Results for Pascal VOC XAI evaluation with DenseNet-161. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.80**	0.58	0.01	**0.05**	0.05	**0.23**	0.11	**0.35**	0.17	**0.47**	0.99	**0.77**	0.57	**0.19**	0.84	**0.21**
LRP-ε+	**0.67**	0.63	0.02	**0.05**	0.17	**0.22**	0.34	**0.36**	0.40	**0.44**	**0.32**	0.49	0.62	**0.48**	1.06	**0.74**
LRP-$\alpha 2\beta 1$	**0.82**	0.74	0.01	**0.03**	0.16	**0.21**	0.31	**0.39**	0.33	**0.41**	0.52	**0.45**	0.64	**0.50**	1.13	**0.74**

Table A.9. Results for MS Coco XAI evaluation with VGG-16. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.58	0.58	0.03	0.03	0.25	0.25	0.26	0.26	0.32	**0.33**	1.00	1.00	0.23	**0.19**	0.25	**0.21**
LRP-$\alpha 2\beta 1$	0.74	**0.86**	**0.03**	0.01	0.32	**0.35**	0.32	**0.39**	0.35	**0.37**	**0.53**	0.61	0.52	0.52	0.79	0.79
LRP-ε+	0.58	**0.66**	0.05	0.05	0.29	**0.32**	0.29	**0.34**	0.41	**0.42**	0.50	**0.47**	0.50	0.50	0.79	**0.78**

Table A.10. Results for MS Coco XAI evaluation with ResNet-18. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.51	**0.55**	0.03	0.03	**0.23**	0.22	0.25	**0.26**	0.33	0.33	**0.97**	0.98	0.16	**0.15**	0.17	**0.16**
LRP-$\alpha 2\beta 1$	0.69	**0.76**	0.02	0.02	0.25	0.25	0.28	**0.32**	0.32	**0.34**	0.69	**0.67**	**0.45**	0.46	0.73	0.73
LRP-ε+	0.54	**0.60**	0.03	0.03	0.23	**0.24**	0.25	**0.27**	0.33	**0.34**	0.73	0.73	0.42	**0.41**	0.72	**0.71**

Table A.11. Results for MS Coco XAI evaluation with EfficientNet-B0. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.56	**0.70**	0.00	0.0	0.17	**0.20**	0.23	**0.30**	0.28	**0.35**	1.00	**0.99**	0.40	**0.35**	0.47	**0.39**
LRP-$\alpha 2\beta 1$	0.73	**0.77**	0.00	0.0	**0.21**	0.11	**0.31**	0.30	**0.34**	0.31	0.64	**0.54**	0.56	0.56	0.77	0.77
LRP-ε+	0.53	**0.73**	**0.01**	0.0	0.22	0.22	0.27	**0.34**	0.38	**0.39**	**0.61**	0.67	**0.40**	0.43	**0.70**	0.71

Table A.12. Results for MS Coco XAI evaluation with DenseNet-121. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.65**	0.58	0.01	**0.04**	0.13	**0.29**	0.22	**0.29**	0.25	**0.39**	0.85	**0.78**	0.57	**0.20**	1.02	**0.23**
LRP-ε+	0.64	0.64	0.01	**0.03**	0.17	**0.31**	0.25	**0.32**	0.29	**0.40**	0.61	**0.48**	0.62	**0.47**	1.14	**0.78**
LRP-$\alpha 2\beta 1$	**0.82**	0.74	0.00	**0.01**	0.16	**0.31**	0.24	**0.34**	0.25	**0.37**	0.48	**0.38**	0.62	**0.50**	1.14	**0.77**

Table A.13. Results for ILSVRC2017 XAI evaluation with VGG-16. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.57	**0.59**	0.06	0.06	0.35	**0.36**	0.68	**0.70**	0.70	**0.71**	1.00	1.00	0.22	**0.18**	0.23	**0.20**
LRP-$\alpha 2\beta 1$	0.70	**0.84**	**0.05**	0.03	0.38	**0.39**	0.65	**0.77**	0.63	**0.67**	**0.59**	0.66	**0.31**	0.34	**0.34**	0.37
LRP-ε+	0.51	**0.62**	**0.09**	0.08	0.36	**0.39**	0.64	**0.71**	0.69	**0.71**	0.57	**0.54**	**0.19**	0.21	**0.21**	0.24

Table A.14. Results for ILSVRC2017 XAI evaluation with ResNet-18. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	0.55	**0.57**	0.03	**0.04**	0.29	0.29	0.66	**0.67**	0.68	**0.69**	0.97	0.97	0.16	**0.14**	0.18	**0.15**
LRP-$\alpha 2\beta 1$	0.67	**0.76**	**0.04**	0.03	0.32	0.32	0.69	**0.75**	0.65	**0.67**	**0.65**	0.61	**0.21**	0.26	**0.22**	0.28
LRP-ε+	0.51	**0.58**	0.04	0.04	0.30	0.30	0.65	**0.69**	0.69	**0.70**	0.70	0.70	**0.14**	0.15	**0.15**	0.16

Table A.15. Results for ILSVRC2017 XAI evaluation with ResNet-50. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.72**	0.64	0.02	**0.04**	0.24	**0.36**	0.65	**0.71**	0.66	**0.69**	0.95	**0.93**	0.36	**0.17**	0.42	**0.18**
LRP-$\alpha 2\beta 1$	0.71	**0.81**	**0.04**	0.01	0.37	0.37	0.72	**0.77**	0.66	**0.67**	**0.59**	0.61	**0.25**	0.30	**0.27**	0.33
LRP-ε+	0.57	**0.67**	**0.05**	0.04	0.37	0.37	0.70	**0.74**	**0.72**	0.71	0.61	**0.60**	**0.15**	0.18	**0.16**	0.19

Table A.16. Results for ILSVRC2017 XAI evaluation with EfficientNet-B0. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold. We analyzed the low faithfulness correlation scores and found that the model was very robust towards input perturbation, with output values remaining unaffected.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.85**	0.70	0.00	0.0	0.24	**0.27**	**0.79**	0.72	**0.73**	0.67	**0.99**	1.00	0.42	**0.33**	0.48	**0.37**
LRP-$\alpha 2\beta 1$	0.75	**0.77**	0.00	0.0	**0.29**	0.20	**0.79**	0.73	**0.72**	0.65	0.57	**0.51**	**0.48**	0.49	**0.52**	0.54
LRP-ε+	0.50	**0.73**	**0.01**	0.0	0.28	**0.30**	0.69	**0.79**	0.75	0.75	**0.61**	0.65	**0.12**	0.21	**0.13**	0.23

Table A.17. Results for ILSVRC2017 XAI evaluation with EfficientNet-B4. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold. We analyzed the low faithfulness correlation scores and found that the model was very robust towards input perturbation, with output values remaining unaffected.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.84**	0.77	0.0	0.0	**0.20**	0.19	0.64	**0.69**	**0.68**	0.66	**0.90**	1.0	**0.30**	0.35	**0.33**	0.40
LRP-$\alpha 2\beta 1$	0.77	**0.79**	0.0	0.0	**0.15**	0.13	0.53	**0.67**	0.61	**0.64**	0.43	**0.4**	0.61	**0.53**	0.68	**0.59**
LRP-ε+	0.56	**0.77**	0.0	0.0	0.13	**0.24**	0.56	**0.76**	0.62	**0.70**	0.54	**0.5**	**0.14**	0.23	**0.15**	0.26

Table A.18. Results for ILSVRC2017 XAI evaluation with DenseNet-121. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.66**	0.62	0.01	**0.03**	0.15	**0.31**	0.53	**0.73**	0.58	**0.72**	0.75	**0.89**	0.57	**0.17**	1.05	**0.19**
LRP-$\alpha 2\beta 1$	**0.82**	0.81	0.01	**0.02**	0.25	**0.33**	0.68	**0.81**	0.64	**0.71**	0.40	**0.44**	0.65	**0.28**	1.30	**0.31**
LRP-ε+	**0.67**	0.66	0.01	**0.03**	0.26	**0.33**	0.71	**0.77**	0.70	**0.74**	0.39	**0.48**	0.63	**0.19**	1.23	**0.21**

Table A.19. Results for ILSVRC2017 XAI evaluation with DenseNet-161. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

	Complexity		Faithfulness				Localization				Random.		Robustness			
	↑ Spars.		↑ Corr.		↑ AoPC		↑ RMA		↑ RRA		↓ Logit		↓ Avg. Sens.		↓ Max Sens.	
canonized	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
EB	**0.86**	0.61	0.00	**0.03**	0.05	**0.30**	0.25	**0.71**	0.46	**0.71**	**0.86**	0.90	0.56	**0.17**	0.80	**0.18**
LRP-$\alpha 2\beta 1$	0.81	**0.82**	0.01	0.01	0.25	**0.32**	0.67	**0.82**	0.65	**0.71**	**0.34**	0.45	0.65	**0.29**	1.29	**0.33**
LRP-ε+	0.64	**0.66**	0.02	**0.03**	0.25	**0.32**	0.70	**0.76**	0.70	**0.74**	**0.36**	0.47	0.62	**0.18**	1.19	**0.19**

Table A.20. Results for CLEVR-XAI with Relation Network using *max-norm* pooling. Arrows indicate whether high (↑) or low (↓) are better. Best results are shown in bold.

Questions	canonized	↑ Complexity		↑ Faithfulness		↑ Local. (RRA)		↑ Local. (RMA)		↓ Robustness		↓ Random.	
		no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
Simple	EB	**0.92**	0.79	0.50	0.50	**0.66**	0.63	**0.56**	0.38	1.34	**1.29**	1.00	1.00
	LRP-Custom	0.69	**0.82**	0.52	0.52	0.71	0.71	0.34	**0.46**	**1.19**	1.24	0.99	0.99
Complex	EB	**0.91**	0.81	**0.45**	0.44	**0.67**	0.64	**0.74**	0.60	1.31	**1.22**	0.99	0.99
	LRP-Custom	0.70	**0.82**	0.45	0.45	0.55	**0.64**	0.48	**0.63**	**1.16**	1.20	**0.98**	0.99

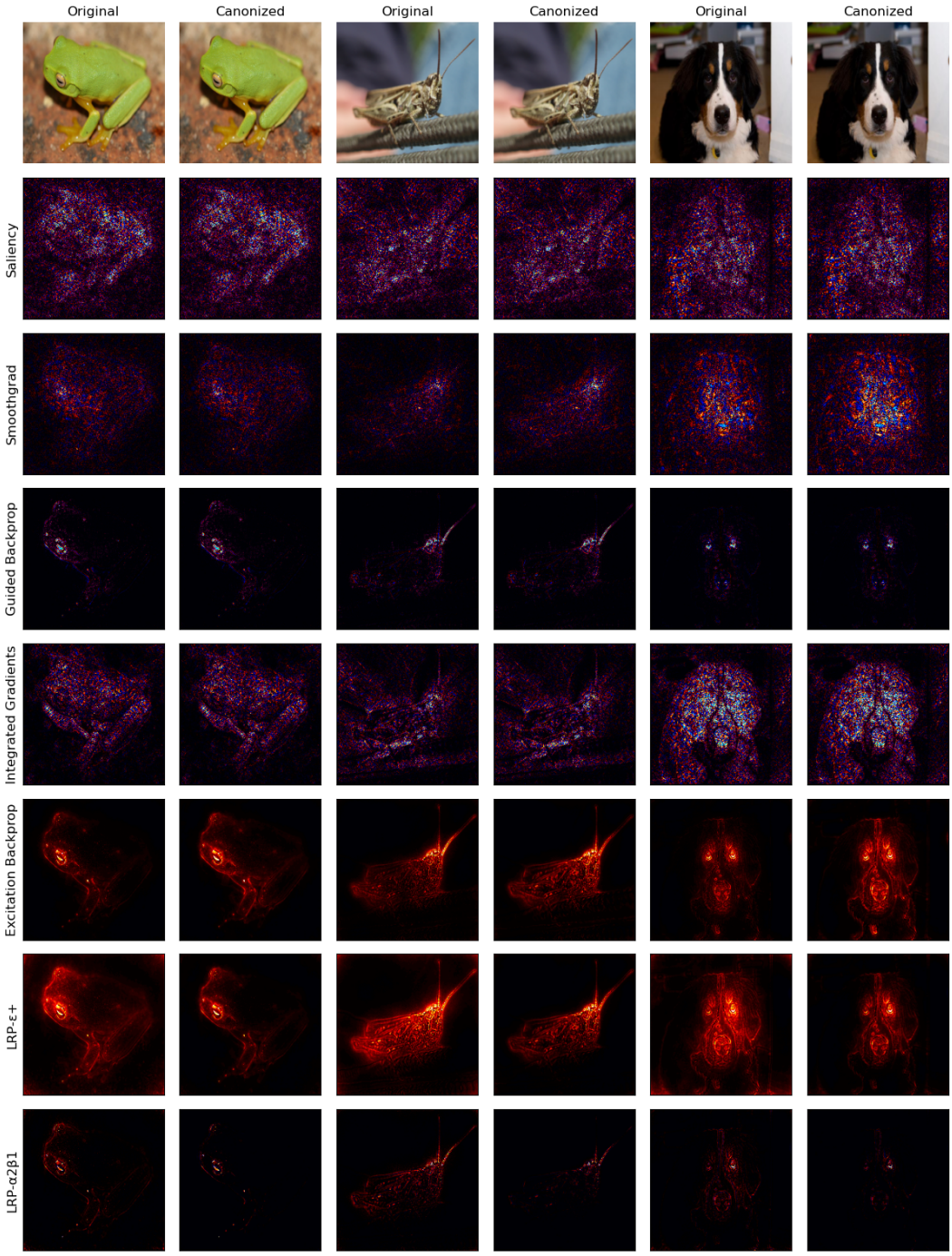

Figure A.2. Attribution heatmaps with different XAI methods for VGG-16 model on ILSVRC2017 dataset.

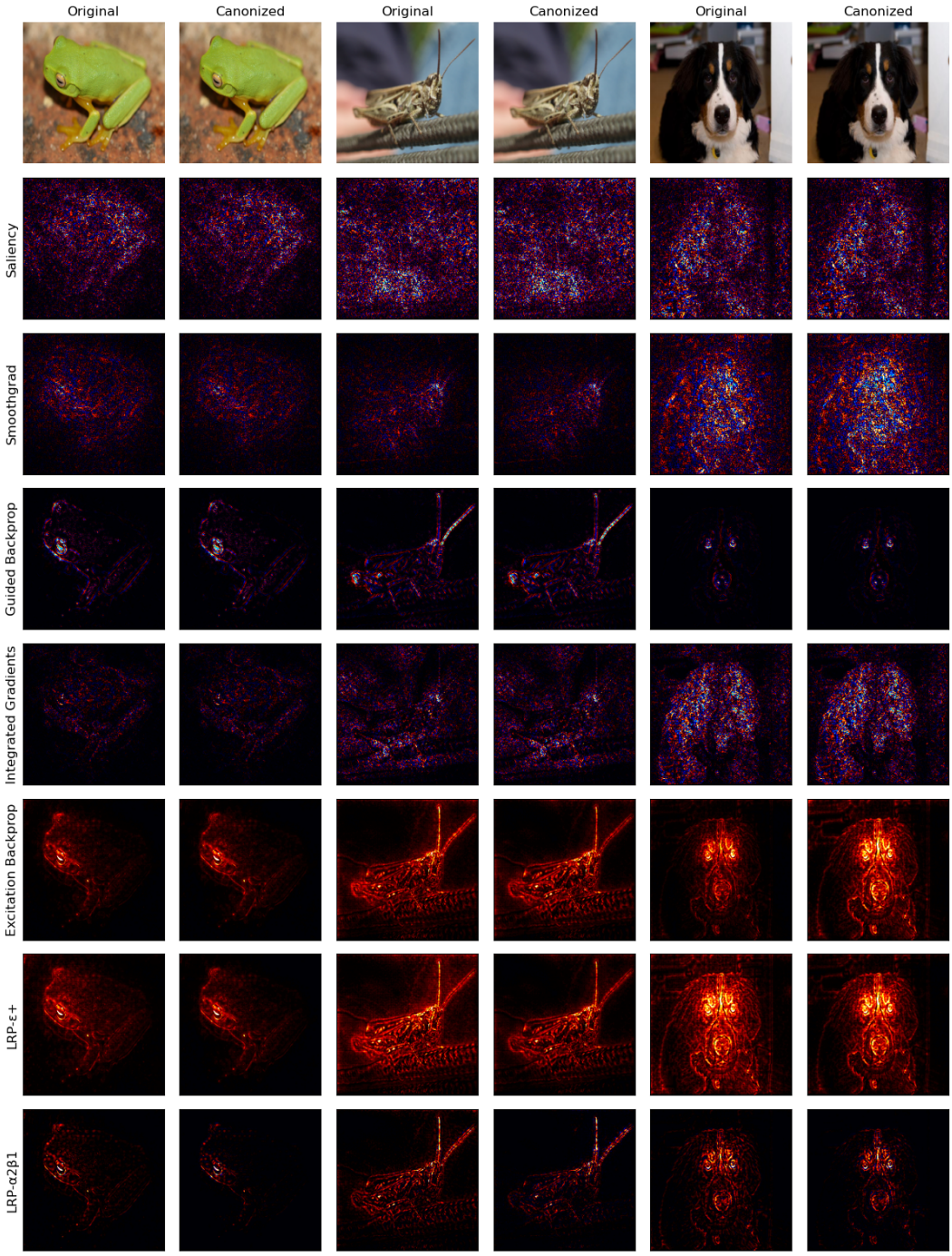

Figure A.3. Attribution heatmaps with different XAI methods for ResNet-18 model on ILSVRC2017 dataset.

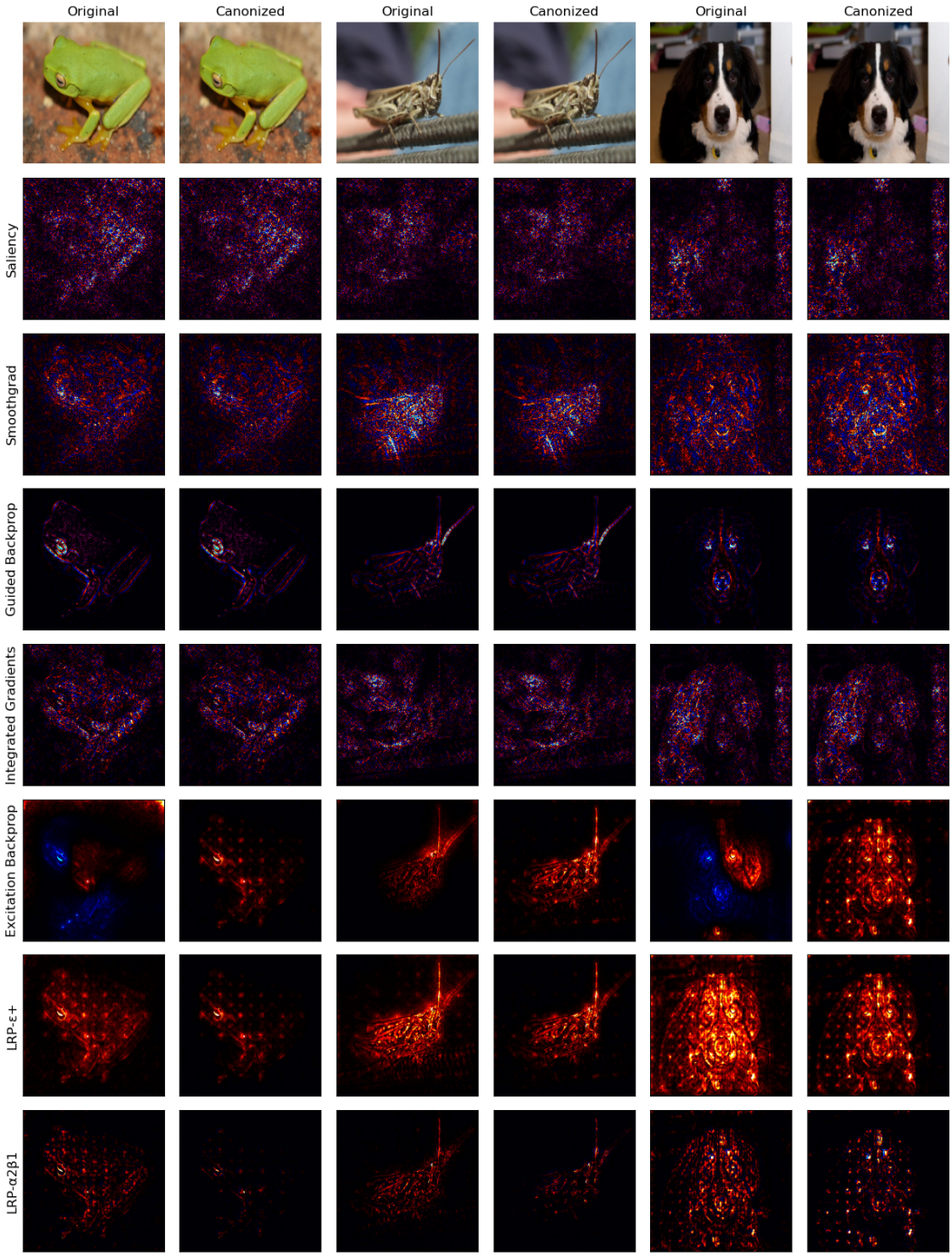

Figure A.4. Attribution heatmaps with different XAI methods for ResNet-50 model on ILSVRC2017 dataset. The checkerboard pattern is due to the downsampling shortcuts in the network. We refer the reader to [5] for details.

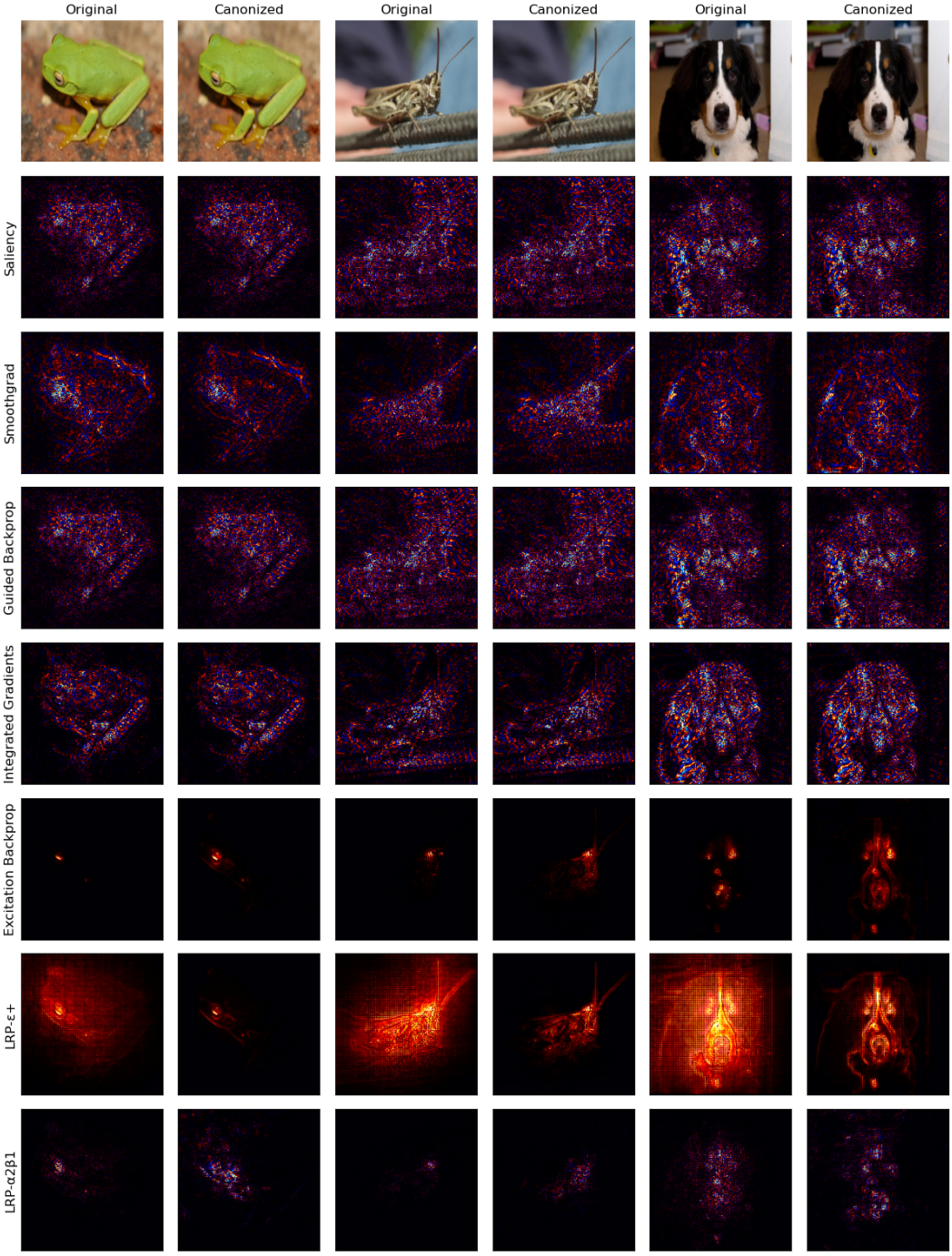

Figure A.5. Attribution heatmaps with different XAI methods for EfficientNet-B0 model on ILSVRC2017 dataset.

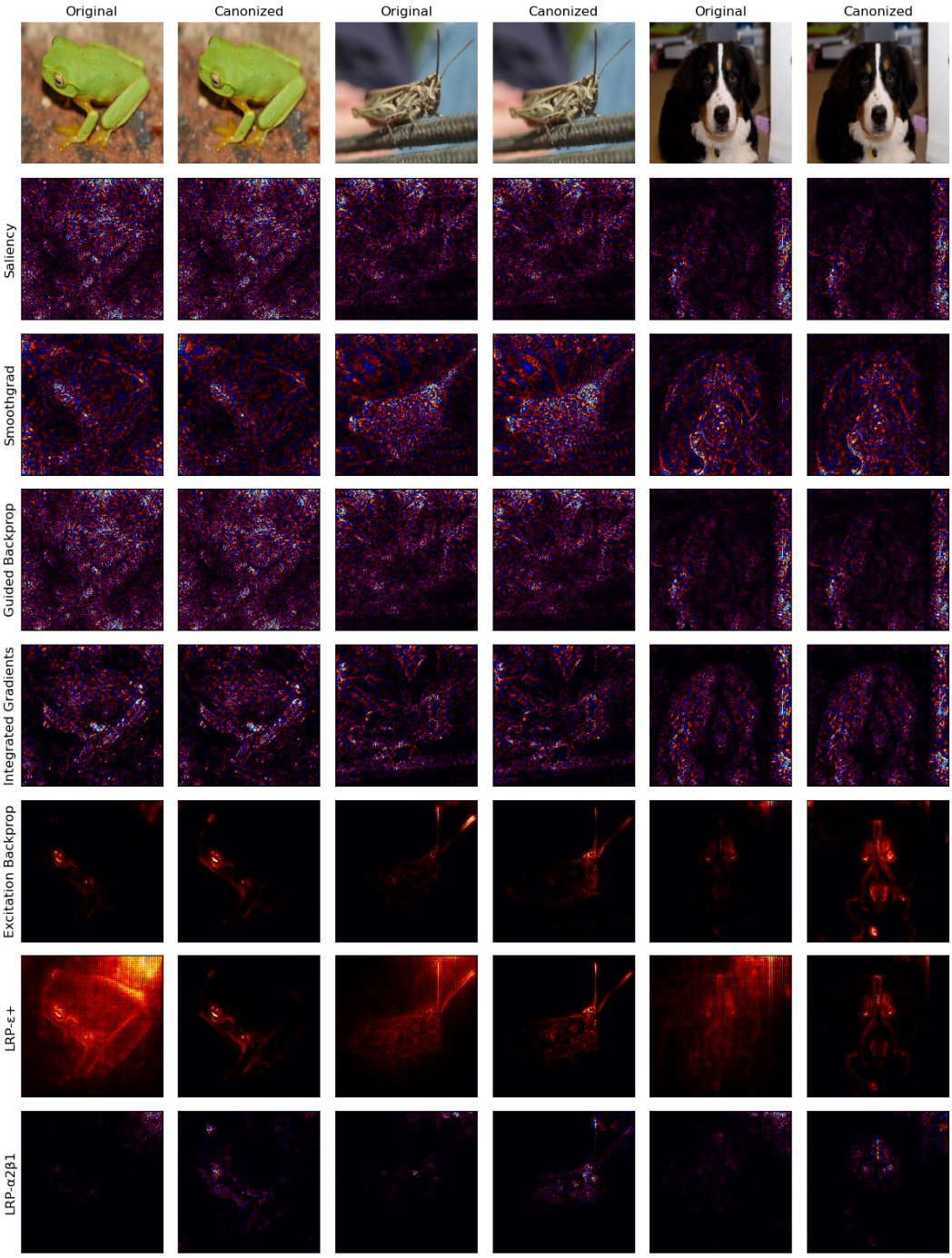

Figure A.6. Attribution heatmaps with different XAI methods for EfficientNet-B4 model on ILSVRC2017 dataset.

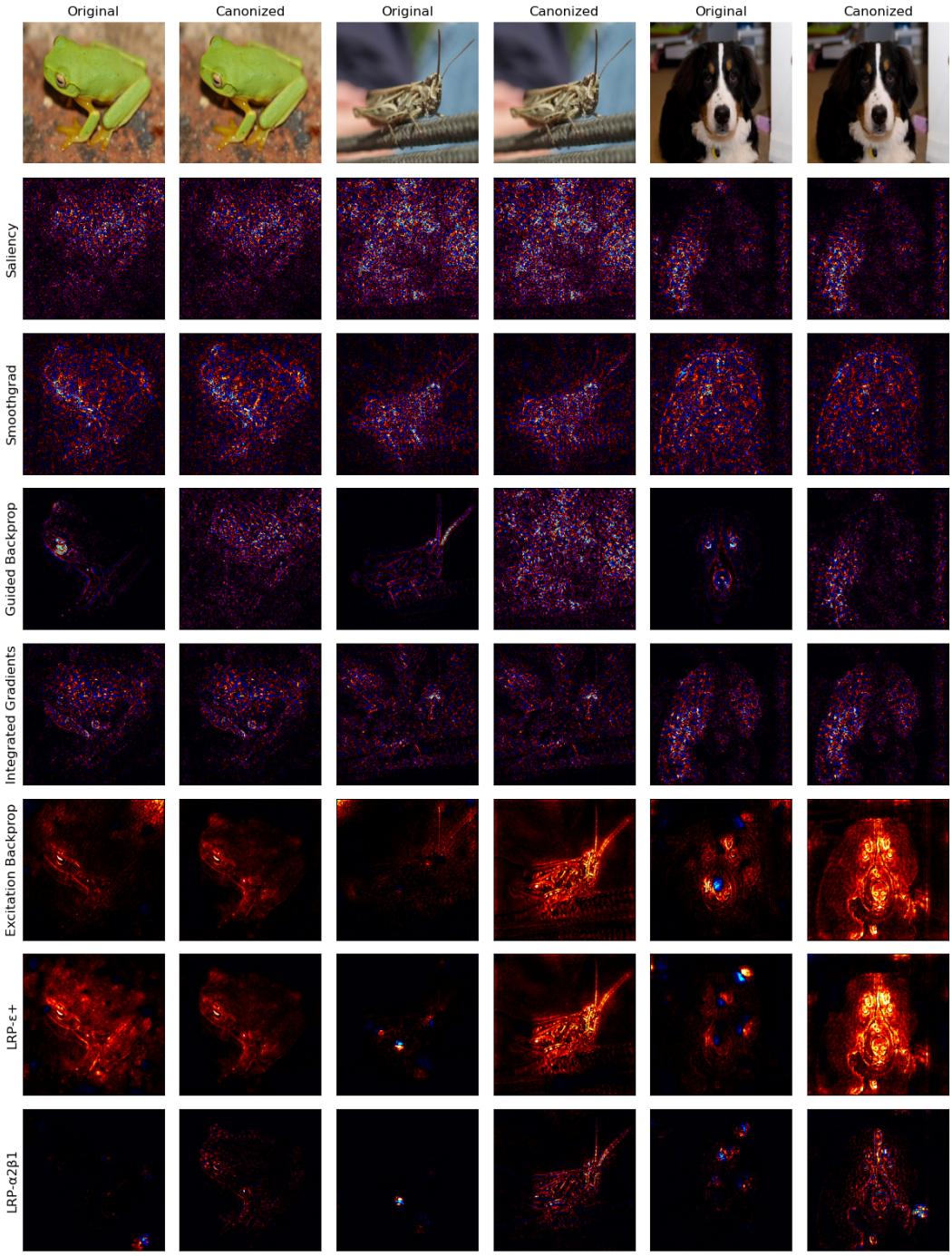

Figure A.7. Attribution heatmaps with different XAI methods for Densenet-121 model on ILSVRC2017 dataset.

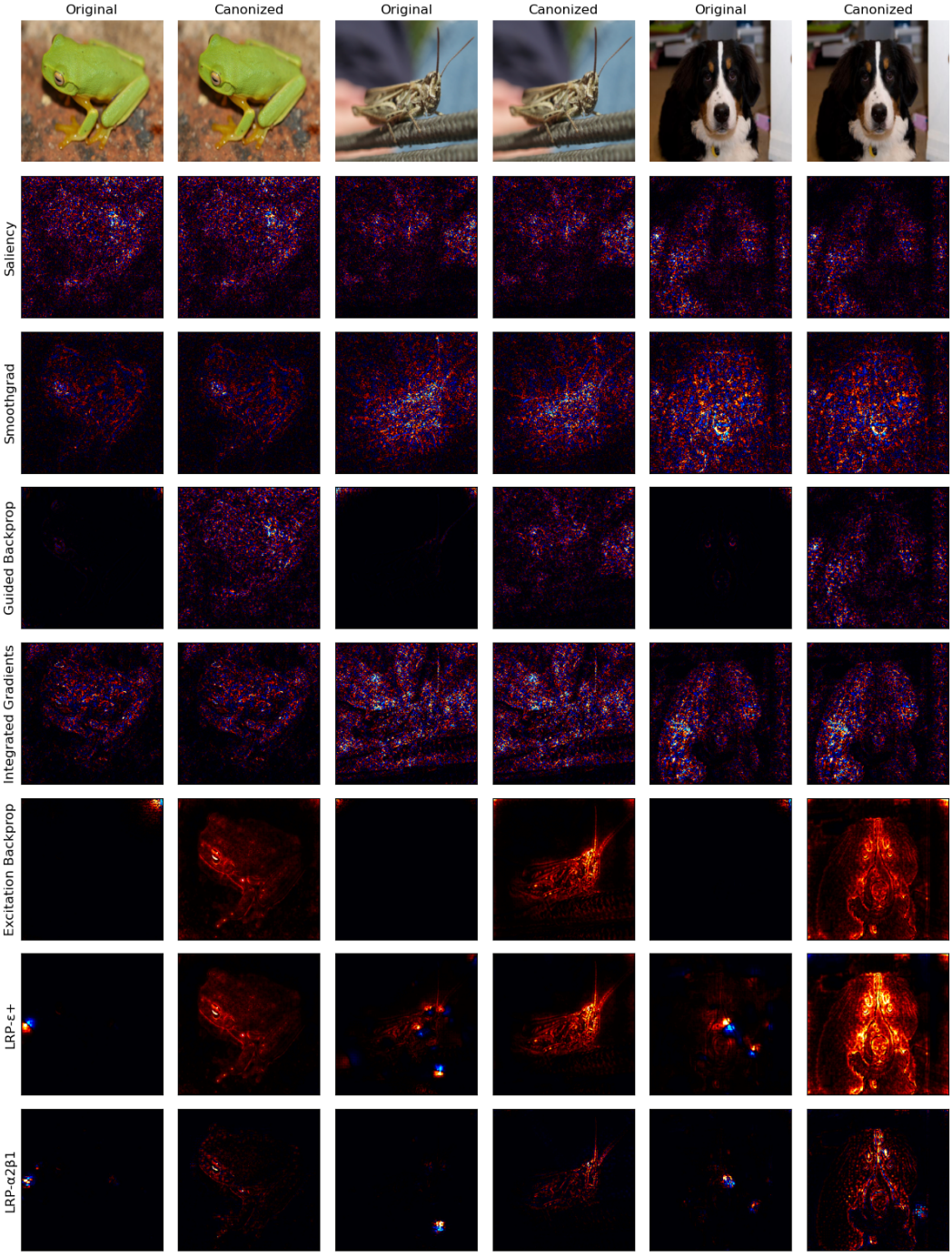

Figure A.8. Attribution heatmaps with different XAI methods for Densenet-161 model on ILSVRC2017 dataset.

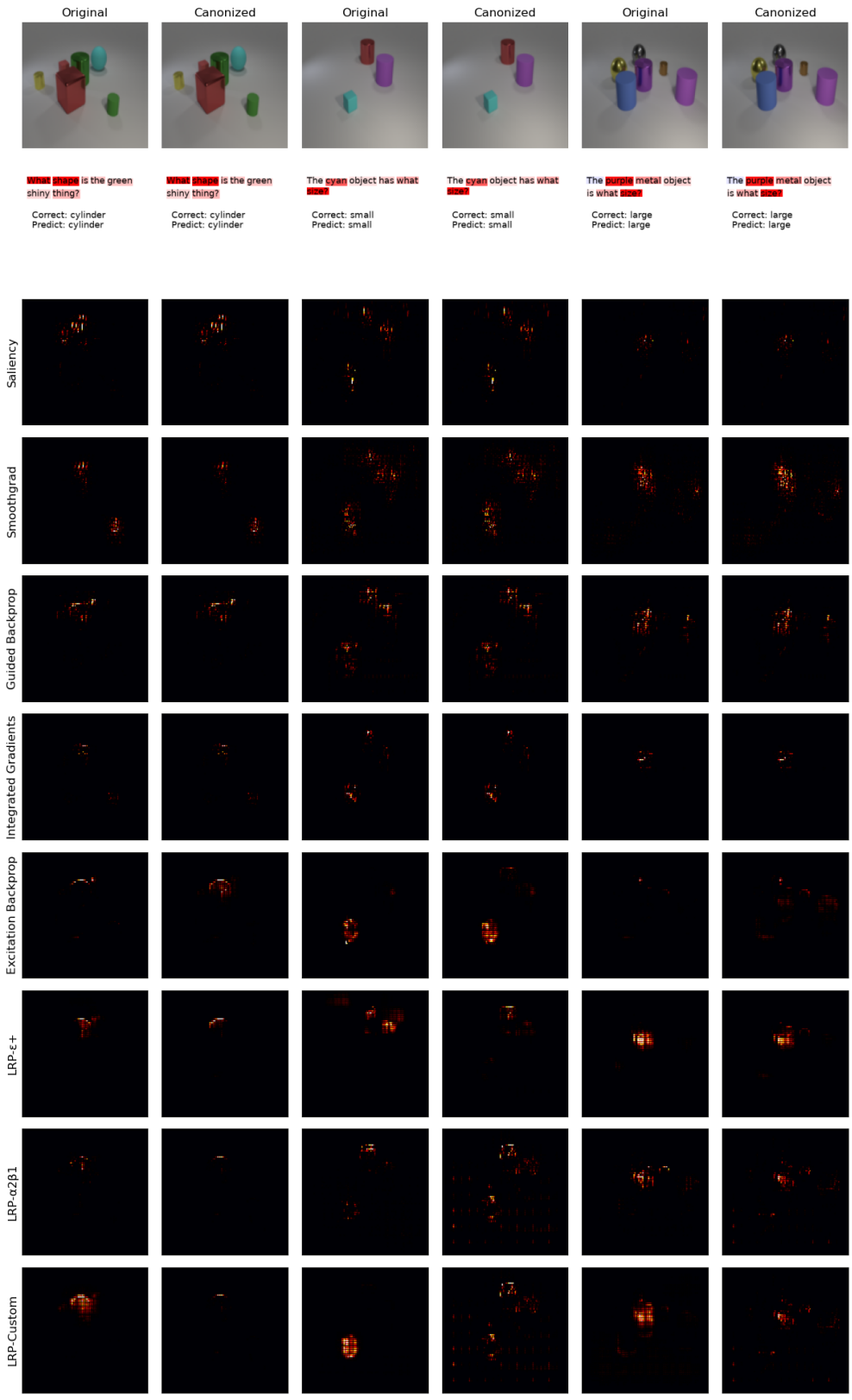

Figure A.9. Attribution heatmaps for Relation Network on CLEVR-XAI dataset using *pos-l2-norm-sq* pooling.

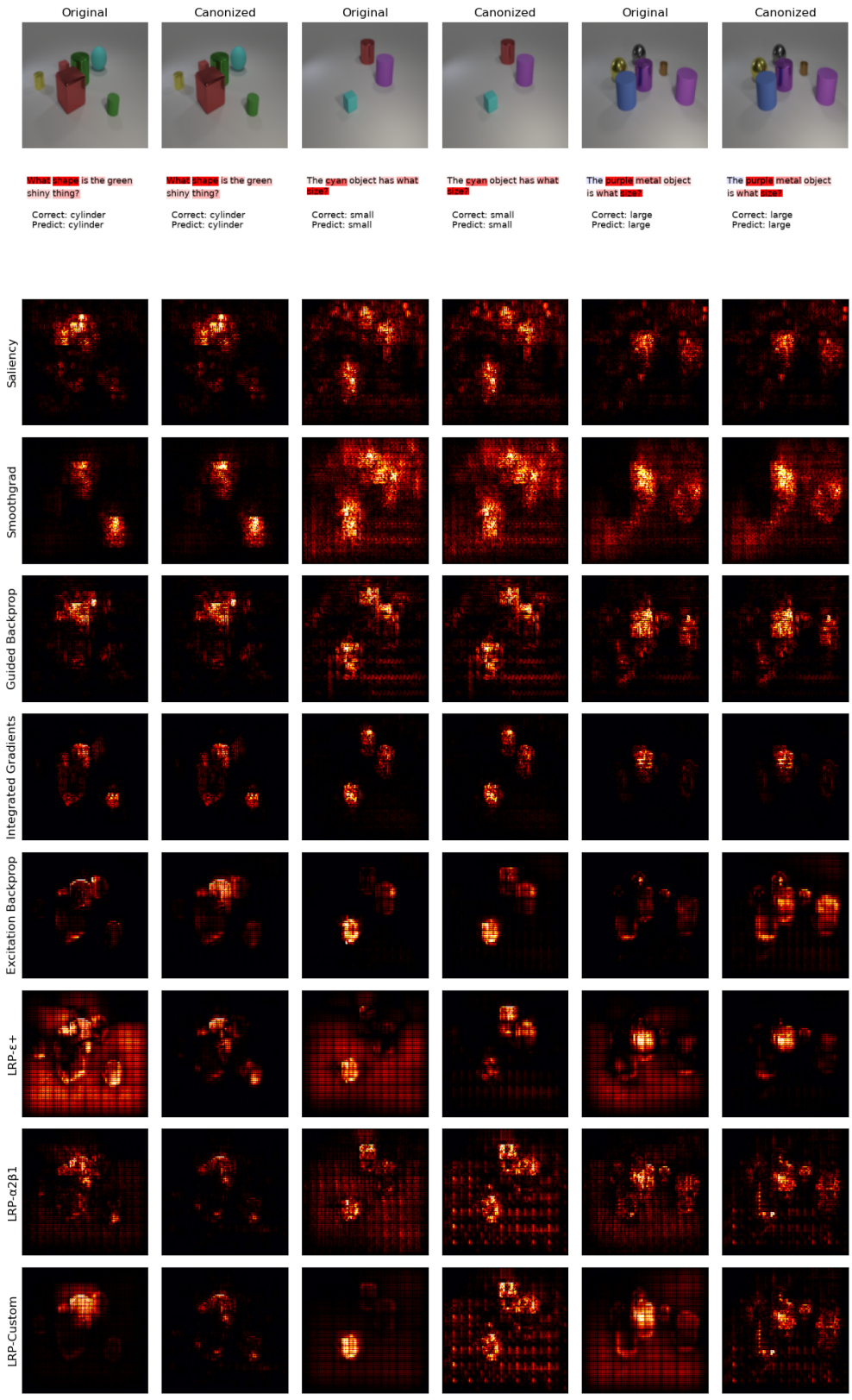

Figure A.10. Attribution heatmaps for Relation Network on CLEVR-XAI dataset using *max-norm* pooling.

References

[1] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. Publisher: Elsevier. 2

[2] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3016–3022. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 4

[3] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. 3

[4] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017. 2

[5] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement, 2022. 12

[6] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 4