Two-Stage Holistic and Contrastive Explanation of Image Classification (Supplementary Material)

Weiyan Xie [1] **Xiao-Hui Li** [2] **Zhi Lin** [1] **Leonard K. M. Poon** [3] **Caleb Chen Cao** [1] **Nevin L. Zhang** [1]

[1]The Hong Kong University of Science and Technology, Hong Kong, China
[2]Huawei Technologies Co., Ltd, Shenzhen, China
[3]The Education University of Hong Kong,, Hong Kong, China

A COMPARISON OF BASE EXPLAINERS

We have evaluated CWOX-2s against SWOX and CWOX-1s in terms of contrastive faithfulness in the main paper, where Grad-CAM and RISE were used as the base explainer. Next, we compare Grad-CAM and RISE, and two other base explainers MWP Zhang et al. [2018] and LIME Ribeiro et al. [2016], in the framework of CWOX-2s. We show how the choice of base explainer influences the contrastive faithfulness of CWOX-2s to the target model. Like Grad-CAM, MWP is a model-dependent explanation method. It backpropagates from the probability score of the target class to compute the marginal winning probability (MWP), over the pixels on a target layer, of a random walk defined using positive network weights and forward activations. Contrastive MWP (c-MWP) provides a contrast to the target class by negating the network weights of the last layer. To avoid double contrasting, we only consider the use of MWP as the base explainer for CWOX-2s, but not c-MWP.

As discussed in Section 4.2 of the main paper, we use different layers as the pivot layer in the two stages of CWOX-2s in order to allow more fine-grained explanations in the second stage when applying back-propagation explanation methods like Grad-CAM as the base explainer. We use the same settings for the Grad-CAM and MWP:

Grad-CAM/MWP	CWOX-2s Stage 1	CWOX-2s Stage 2
ResNet50 Pivot Layer	ReLU of Conv5_3	ReLU layer of Conv4_6
GoogleNet Pivot Layer	Inception5b	Inception4e

In the forward-propagation method - RISE, we specify the pixel mask probability to have different settings in the two stages:

RISE	CWOX-2s Stage 1	CWOX-2s Stage 2
Number of Masks	5,000	3,000
Pixel Mask Probability	0.3	0.15

Like RISE, LIME is a model-agnostic explanation method. It learns a surrogate linear regression model, from superpixels to class score, in the neighborhoods of the input image. We use the Quickshift algorithm to compute the superpixels. LIME has a hyperparameter that determines the number of samples to use for the regression model. Quickshift has a hyperparameter called kernel size, where the larger the size, the larger the neighborhoods of pixels considered. The two hyperparameters are set for CWOX-2s as follows:

LIME	CWOX-2s Stage 1	CWOX-2s Stage 2
Number of Samples	3000	1000
Quickshift Kernel Size	8	4

We have conducted experiments on 10,000 randomly selected images from the ImageNet validation set. Each example is fed

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

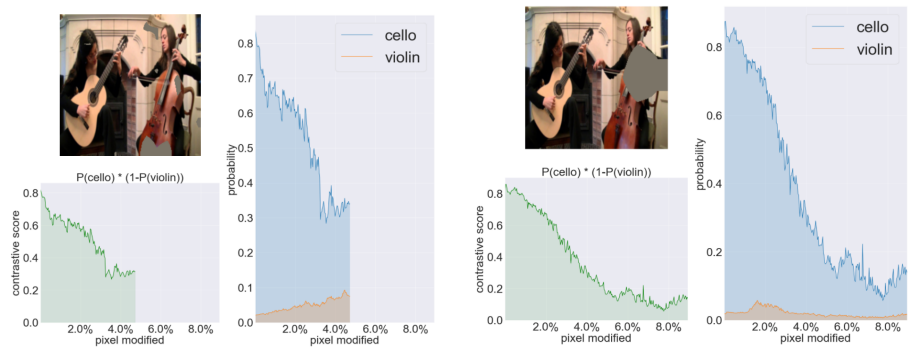
to a target model and the top classes in the outputs are explained using CWOX-2s. The four base explainers are tested in turns, resulting four heatmaps for each case.

Table A.1 displays the performance statistics of CWOX-2s when using each of the four base explainers. Regarding contrastive faithfulness metrics, RISE performs the best, having the lowest CAUC score and the highest CDROP scores. Additionally, RISE identifies the fewest salient pixels, indicating that it can help CWOX-2s accurately pinpoint essential evidence. Grad-CAM is the next best performer, followed by MWP. MWP results in a large number of salient pixels, making it less precise in identifying important pixels. Among the four base explainers, LIME leads to the worst performance for CWOX-2s.

We present in Figure C.9 the results of CWOX-2s explanation using four different base explainers for the `cello-guitar` image. We now focus on the contrastive faithfulness of CWOX-2s heatmaps for `cello` against `violin` generated by the four different base explainers. Figure A.1 demonstrates how the probabilities $P(\texttt{cello})$ and $P(\texttt{violin})$, as well as the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$, change as pixels are removed based on the orderings determined by each of the four heatmaps.

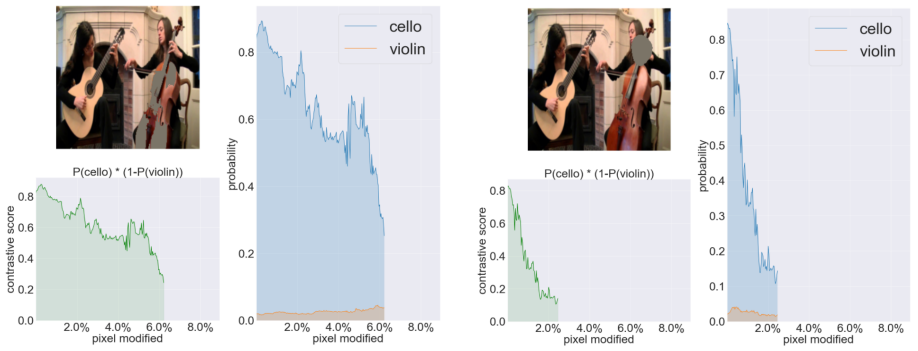
	ResNet50			GoogleNet		
Base Explainer	$\bar{n}_\delta$	CAUC ↓	CDROP ↑	$\bar{n}_\delta$	CAUC ↓	CDROP ↑
Grad-CAM	2,029	3.11×10^{-3}	8.01×10^{-2}	2,181	1.80×10^{-3}	7.46×10^{-2}
MWP	4,194	3.12×10^{-3}	7.37×10^{-2}	3,026	1.77×10^{-3}	5.85×10^{-2}
LIME	2,464	3.40×10^{-3}	4.89×10^{-2}	2,351	1.92×10^{-3}	3.64×10^{-2}
RISE	1,282	3.07×10^{-3}	8.97×10^{-2}	1,105	1.72×10^{-3}	8.32×10^{-2}

Table A.1: Performances of CWOX-2s with four base explainer. Here, $\bar{n}_\delta$ stands for average number of δ-salient pixels.



(a) Grad-CAM: $n_\delta = 2,377$
CAUC=0.0153, CDROP=0.505

(b) MWP: $n_\delta = 4,495$
CAUC=0.0178, CDROP=0.461



(c) LIME: $n_\delta = 3,131$
CAUC=0.0184, CDROP=0.470

(d) RISE: $n_\delta = 1,223$
CAUC=0.0902, CDROP=0.694

Figure A.1: Changes in the probabilities $P(\texttt{cello})$ and $P(\texttt{violin})$ and the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ as δ-salient pixels are deleted according to the order induced the CWOX-2s heatmap that is obtained with the base explainer: (a) Grad-CAM, (b) MWP, (c) LIME, and (d) RISE. Note that the CAUC score for Grad-CAM are lower than that in Figure 6 because the number of pixels deleted is 1,223 — the number of δ-salient pixels in the RISE heatmap. It is smaller than the number of δ-salient pixels in the Grad-CAM heatmap (2,337).

B MORE DETAILS OF THE USER STUDY

The overall procedures and results of the user study have been discussed in Section 5.2 of the main paper. We provide some additional details in this appendix.

B.1 STUDY FORM AND PARTICIPANTS

The user study was carried out using a web-based survey via the Qualtrics Survey Tool. Two groups of participants were invited to take part in the survey through email invitations. The first group, known as the expert group, consisted of postgraduate students who were enrolled in a machine learning course at the time of the study. These students had experience with deep computer vision models, including training CNN models. However, they had not yet been exposed to XAI. The second group, the non-expert group, was made up of first-year undergraduate students who had no prior experience or knowledge in training deep learning models.

B.2 DETAILED PROCEDURES OF THE STUDY

The procedure for the user study is detailed as follows.

1. Tutorial: Participants first received a tutorial on the basics of image classification and the explanations.
2. Training phase: Participants were shown a set of examples and explanations for a pair of confusing class labels (e.g. `cello` and `violin`). For each training example, the explanation results and confusing class labels were shown to the participants, but the matching relationship between the explanations and labels was initially unknown to them. They were asked to guess the features the model uses to distinguish the two confusing classes. They then received verification of their guesses on the next page. Screenshots of the user interface during the training phase can be seen in Figure B.1.
3. Evaluation phase: An evaluation phase followed the training phase which was set up similarly to the guessing step in training phase. The participants' understanding of the discriminative features was evaluated by testing how well they can tell the matching relationship between the explanations and the confusing class labels on new unseen examples.

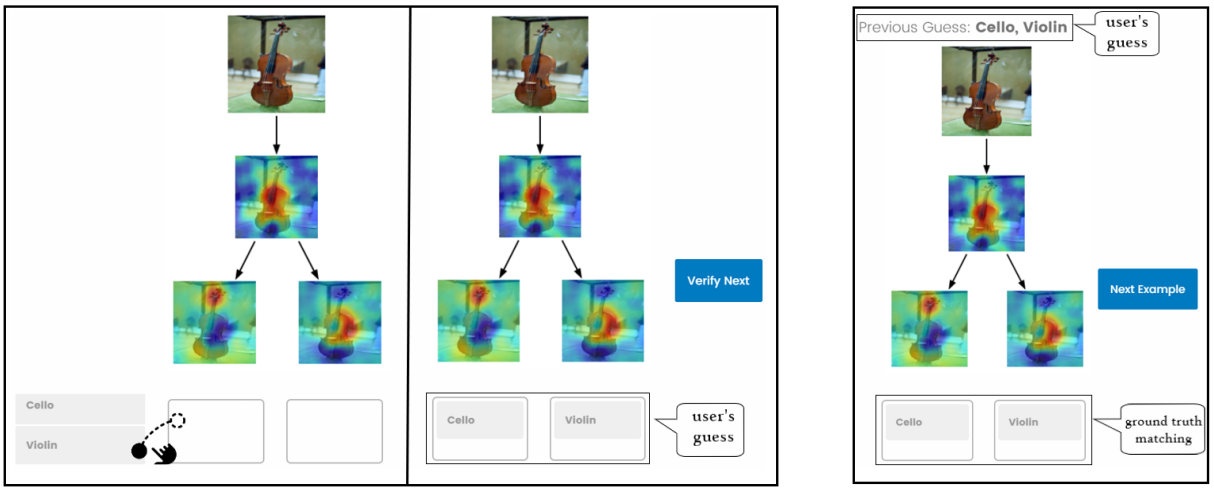
It is noted that the "guess first, verify next" setting in the training phase of the user study is similar to the training process of a deep neural network model. At the beginning of the model training, the model is initialized with random weights, which may result in random guesses and high training losses. However, if there are distinct features for different class labels in the training set, the model can learn these features and gradually reduce the training loss. Similarly, if the explanations provided can reveal the discriminative features used by the model, human subjects are expected to gain a better understanding of these features as they see more examples.

B.3 THE CONFUSING CLASS LABELS AND EXAMPLES USED

Since there are a variety of images in the ImageNet, to reduce the burden on users, we showed examples with the same pair of confusing class labels in one round of the training and evaluation. Each participant completed two rounds of the study, each with examples from a randomly assigned label pair.

To make the study manageable, we limited the choices of input images and confusing class labels. We selected the confusing class labels from the latent tree model built from the outputs of ResNet50. In the latent tree model, the classes are grouped under the same node because they frequently appear together in the top prediction classes of ResNet50. Therefore, many pairs of confusing classes, can be obtained from the level-1 latent nodes of the model. In order to determine which pairs to be used in the human study, we first followed the three criteria [1] introduced in Zhang et al. [2019] to filter out the unfamiliar, ambiguous, and expert-specific classes. After filtering those classes, we invited ten AI researchers to nominate ten pairs of confusing classes based on their interests from the remaining part of the latent tree model. The final ten confusing class pairs were selected through voting among the ten researchers and included {`cello`, `violin`}, {`ambulance`, `police van`}, {`harvester`, `tractor`}, {`folding chair`, `rocking chair`}, {`basketball`, `volleyball`}, {`acoustic guitar`, `electric guitar`}, and so on.

[1] (1) Familiar, the class should be familiar to all the human subjects; (2) Unambiguous, the class should have only one clear connotation for the given object; (3) Non-specific, the class should not be a specialization or a potential sub-class of another class in ImageNet.



(a) The guessing step of the training phase

(b) The verification step of the training phase

Figure B.1: An example of the training phase of the user study: (a) The user is presented with an example along with the associated CWOX-2s explanation and the two confusing class labels. They are instructed to match the class labels with the second-stage CWOX-2s heatmaps by dragging and dropping the labels into boxes below the heatmaps. (b) Once the user has made their guess, they are then able to verify their answer on the next page. The evaluation phase is structured in a similar manner, with participants performing the matching task on new, unseen examples.

For each pair of confusing class labels, we collected examples from the ImageNet validation set with ground-truth labels as one of the two confusing class labels. Examples were used in the study if both classes in the confusing class pair appeared in the model's top prediction labels, regardless of whether the classification was correct or incorrect. The images were randomly divided into training and evaluation examples.

B.4 GUIDELINE TO THE PARTICIPANTS

The following is the guideline provided to the participants at the beginning of the study. We also made an introductory video based on the guideline.

Guideline to the Study Participants:

1. General Background.

Hi there, thank you for joining our study. We are a research team from [hidden institution], aiming to build better explanation tools for users to understand the behaviors of AI models. AI models are generally non-intuitive and difficult for humans to understand. They are considered black-box models. Explainable AI, or XAI, aims to provide explanations for model predictions to help users understand how the predictions are reached.

The purpose of this study is to evaluate how well different XAI methods can help YOU, as a model user, understand model behaviors. Your understanding will be assessed by asking you to predict model's prediction on new inputs. The study will carry out in the context of image classification.

2. Tutorial: Image classification & Saliency Map & Confusion Classes.

Image classification is a supervised learning problem: with a set of target classes, to train a model to recognize the class on the labeled example images. The output of image classification is a probability distribution over multiple classes.

An explanation of the image classification reveals what regions of the input image that the model relies on to predict the specific label. An explanation is typically given as a saliency map. The saliency map is a heatmap, where the high-temperature regions are what a model considers important for a prediction. For instance, in the example of Figure B.2, this image is classified as goldfish, and the bodies of goldfishes are highlighted.

Although usually we only pick the class with highest prediction probability as the output label for the image, the other classes with high prediction probabilities also reflect some important aspects of the model behaviors.

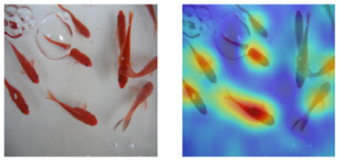


Figure B.2: An example of saliency map.

For instance, if two classes are always with high prediction probabilities at the same time and co-occur as top classes in classification outputs, they are considered confusion classes where the model always has trouble determining which of them to use as the output class label. Cello and Violin is one pair of such classes for ResNet50. Understanding what features the model relies on to distinguish those confusion classes can help us understand the model behaviors better.

In this study, we will test how well saliency maps created by different methods enable YOU, as a model user, to understand the evidences that the model uses to distinguish between confusion classes.

3. Study Procedure: Training Phase.

The study is divided into two phases, the training phase and the evaluation phase.

In the training phase, a series of examples will be presented to you. Each example involves two class labels and a tree of saliency maps. The task is to match the class labels with the saliency maps at the leave nodes. The guess-verify strategy is adopted so that you can complete the training phase efficiently. Your first make a guess about the matching, and drag/drop the class labels to the appropriate boxes below the saliency maps. Next, you click the "Next" button to see the ground-truth match.

Note that the model might not be necessary to think like humans. In this human study, the task is **NOT** about how you feel the class labels should be matched with saliency maps based on your previous life experience. Rather, it is about *learning how saliency maps created by XAI method match predictions of a model*.

4. Study Procedure: Evaluation Phase.

In the evaluation phase, you are asked to do the matching for a series of new examples, just as in the guess step of the training phase. Note that the matching must be one-to-one. Otherwise, you won't be able to proceed. In most cases, you will not be 100% sure. Just pick the one that you feel more likely according to what you have learned in the training phase. Pick randomly if necessary.

5. Overall Flow of the Study.

You will experience two rounds of the study. Each round contains a training and evaluation phase with examples from a particular pair of confusion classes. The two pairs will be shown to you and you can pick one to start first and work on another pair later.

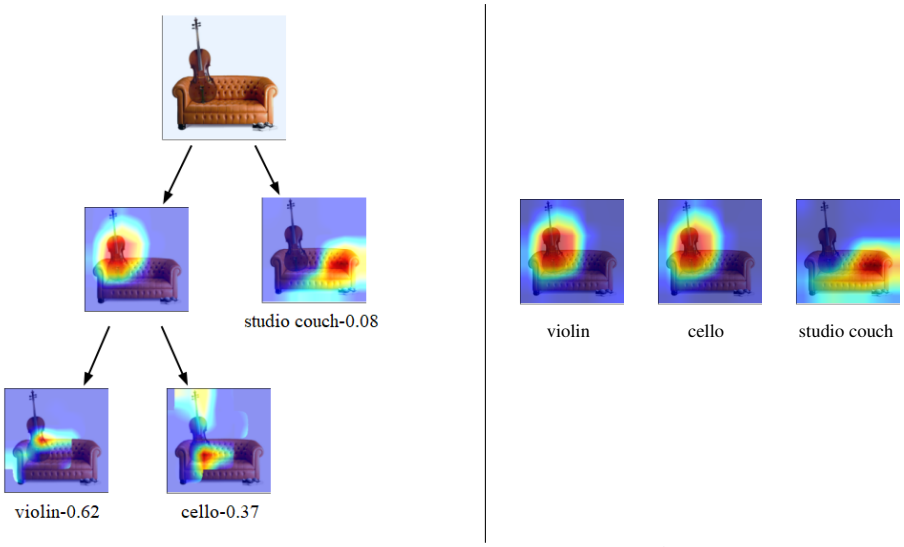
Although you are encouraged to complete both rounds of study in one sitting, you can leave in the middle and go back to study later. This web-based system will use cookies to keep track of your progress and you can return to the point where you left off, as soon as you access the survey link again with the same device.

If you have any questions, do not hesitate to contact us. Thanks again for joining our study.

C MORE VISUAL EXAMPLES OF CWOX-2S

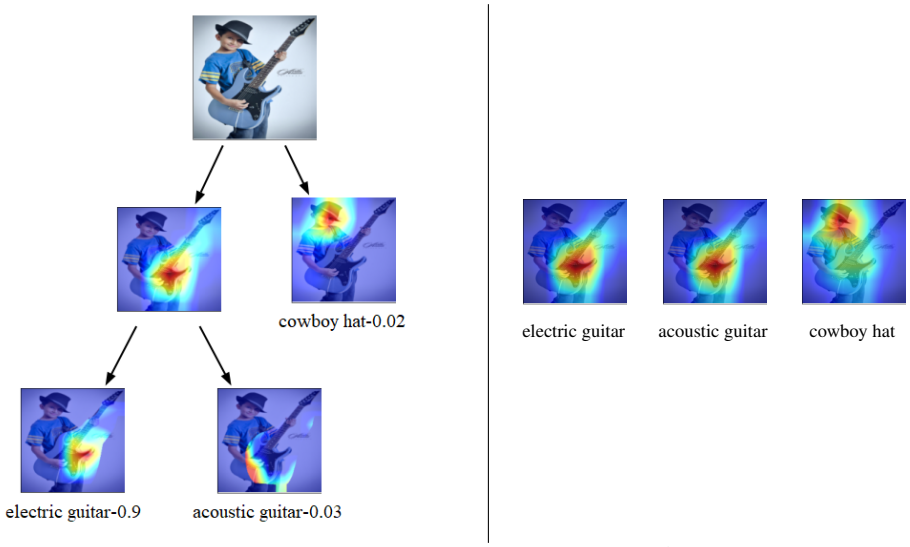
C.1 EXAMPLES OF VISUAL SIMILAR CLASS LABELS

In the Figure 1 of the main paper, we have shown that CWOX-2s can provide more meaningful and discriminative explanations for the visual similar class labels - `screwdriver` and `syringe`. We here provide more examples (Figure C.1, C.2 and C.3) to show that how CWOX-2s can provide users understanding of discriminative features model used to distinguish the visual similar classes (`violin` and `cello` in Figure C.1; `electric guitar` and `acoustic guitar` in Figure C.2; `pitcher` and `water jug` in Figure C.2).



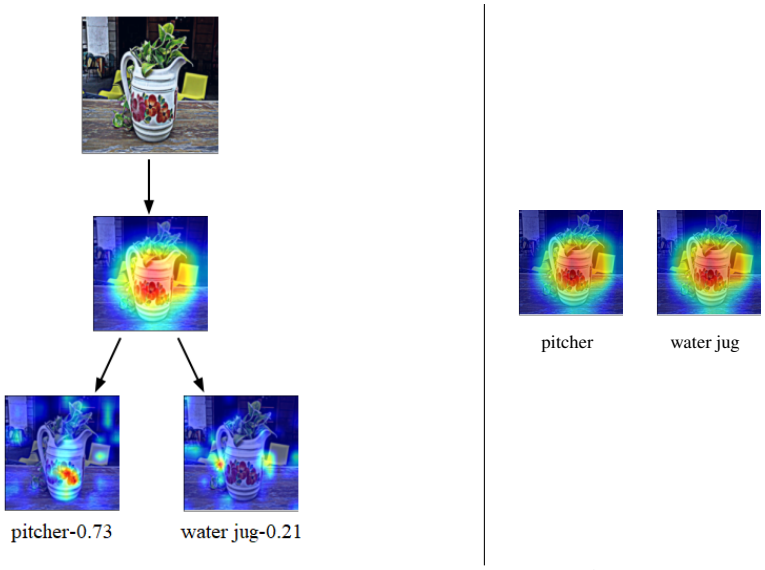
(a) CWOX-2s (b) SWOX

Figure C.1: Results of SWOX and CWOX-2s: Input image with ground-truth label `violin`. The ResNet50 output on the input consists of three top classes `violin` (0.62), `cello` (0.37) and `studio couch` (0.08). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.



(a) CWOX-2s (b) SWOX

Figure C.2: Results of SWOX and CWOX-2s: Input image with ground-truth label `electric guitar`. The ResNet50 output on the input consists of three top classes `electric guitar` (0.90), `acoustic guitar` (0.03) and `cowboy hat` (0.02). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.

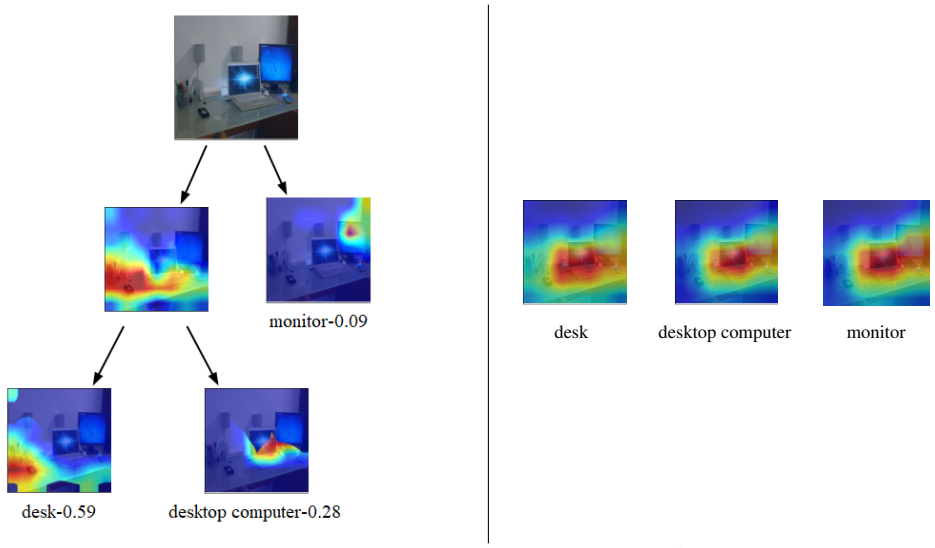


(a) CWOX-2s (b) SWOX

Figure C.3: Results of SWOX and CWOX-2s: Input image with ground-truth label `pitcher`. The ResNet50 output on the input consists of two top classes `pitcher` (0.73) and `water jug` (0.21). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.

C.2 EXAMPLES OF COMPOSITE OBJECT

In the Figure 8 of the main paper, we have shown that CWOX-2s can well identify the composite object `mouse+computer-keyboard` . We here provide more such examples (Figure C.4, C.5 and C.6) where the composite object is `desk+desktop-computer` in Figure C.4, `street-sign+traffic-light` in Figure C.5 and `drumstick+drum` in Figure C.6.

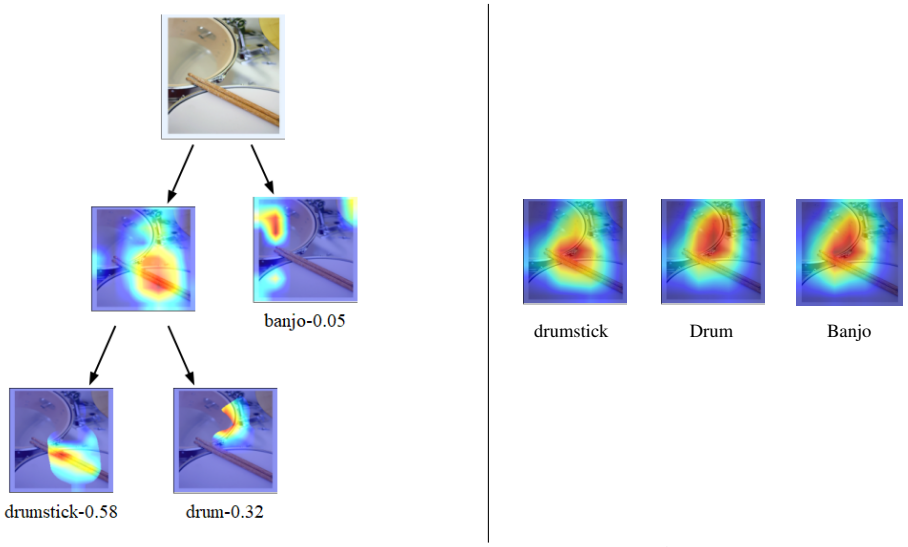


(a) CWOX-2s (b) SWOX

Figure C.4: Results of SWOX and CWOX-2s: Input image with ground-truth label `desk`. The ResNet50 output on the input consists of three top classes `desk` (0.59), `desktop computer` (0.28) and `monitor` (0.09). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.

(a) CWOX-2s (b) SWOX

Figure C.5: Results of SWOX and CWOX-2s: Input image with ground-truth label `traffic light`. The ResNet50 output on the input consists of two top classes `street sign` (0.73) and `traffic light` (0.27). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.



(a) CWOX-2s (b) SWOX

Figure C.6: Results of SWOX and CWOX-2s: Input image with ground-truth label `drumstick`. The GoogleNet output on the input consists of three top classes `drumstick` (0.58), `drum` (0.32) and `banjo` (0.05). (a) CWOX-2s heatmaps, (b) SWOX heatmaps.

C.3 EXAMPLES OF EXPLAINING MISCLASSIFICATION

Figures C.7 (a) - (d) show the contrastive heatmaps produced by CWOX-2s for the output of ResNet50 on the image with ground-truth label `padlock`. ResNet50 is completely wrong in this case. Both SWOX (e.5) and CWOX-2s (d) reveal the apparently reasonable evidence for `wall clock` — the two keys look like the hands on a clock. However, CWOX-2s does a better job than SWOX at identifying the evidence for `necklace` (a) and `whistle` (b). Furthermore, heatmap (c) suggests that a part of the ring is evidence for `magnet-compass` and `stopwatch`. The contrastive evidence for distinguishing those two classes includes an area where one would expect a hanging ring for a compass (c.1), and an area where one would expect a button for a stopwatch (c.2).

Figure C.8 depicts an example with a ground-truth label of `gown`. ResNet50 misclassifies it as `hoopskirt` (0.54). Other

Figure C.7: Results of SWOX and CWOX-2s: (a) Input image with ground-truth label `padlock`. The ResNet50 output on the input consists of five top classes `necklace` (0.130), `whistle` (0.127), `magnetic-compass` (0.114), `stopwatch` (0.089), and `wall-clock` (0.069). (a-d) CWOX-2s heatmaps; (e) SWOX heatmaps.

top predicted classes for this example include `gown` (0.22), `groom` (0.14), and `lakeside` (0.09). When Grad-CAM is applied to explain these top classes, the SWOX heatmaps for `hoopskirt`, `gown`, and `groom` (d.1 - d.3) are essentially the same. In contrast, CWOX-2s provides clearer explanations for them, where (a) highlights the large skirt as evidence for `hoopskirt`, (b.1) highlights the female body as evidence for `gown`, and (b.2) highlights the male face as evidence for `groom`.

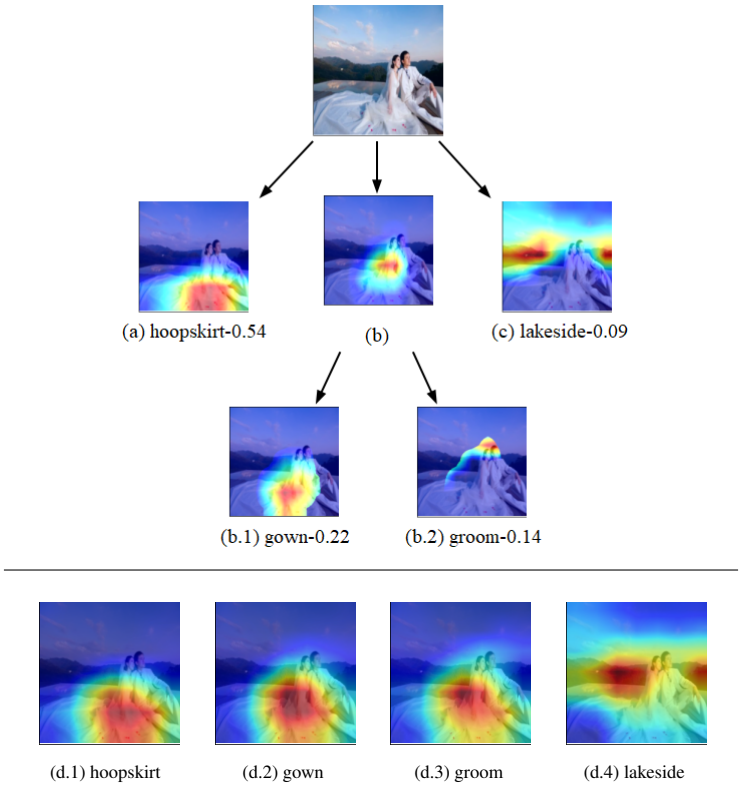


Figure C.8: Results of SWOX and CWOX-2s: Input image with ground-truth label `gown`. The ResNet50 output on the input consists of four top classes `hoopskirt` (0.54), `gown` (0.22), `groom` (0.14), and `lakeside` (0.09). (a-c) CWOX-2s heatmaps; (d) SWOX heatmaps.

C.4 EXAMPLES WITH DIFFERENT BASE EXPLAINERS

While we used Grad-CAM as the CWOX-2s base explainer in the main paper and Appendices C.1, C.2, C.3, we now include some examples in Figure C.9,C.10 and C.11 with three other base explainers, namely MWP, LIME and RISE.

References

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

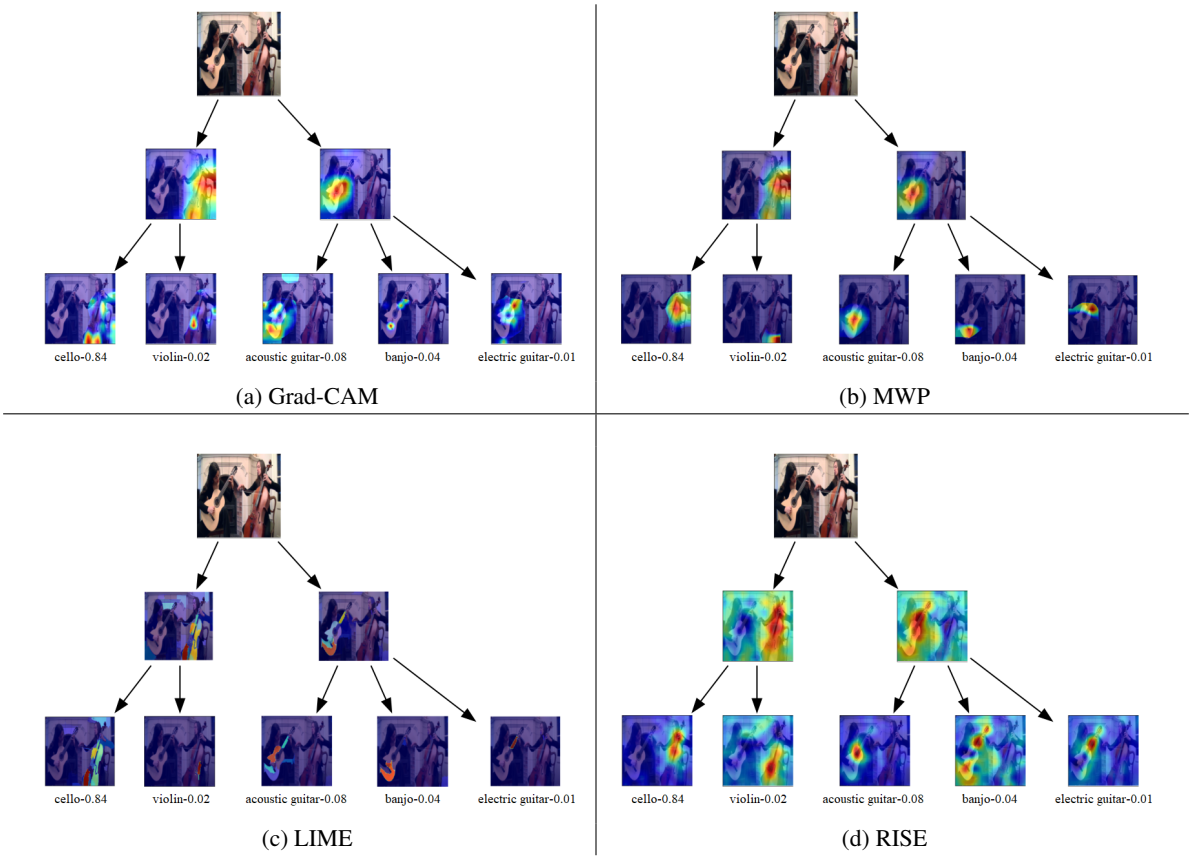


Figure C.9: CWOX-2s explanations with the four different base explainers on the Figure 2 & 3 `cello-guitar` example.

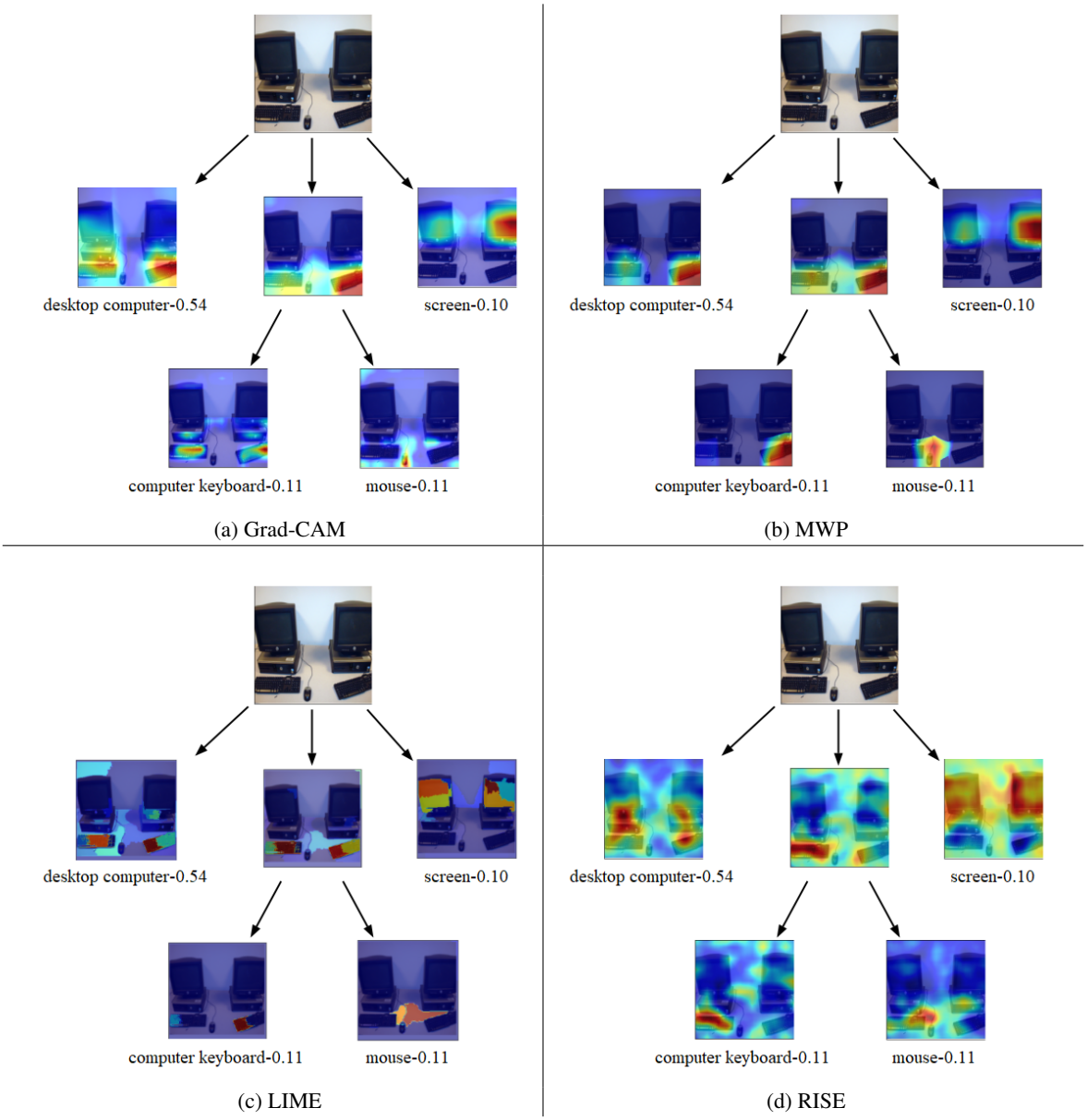


Figure C.10: CWOX-2s explanations with the four different base explainers on the Figure 8 example.

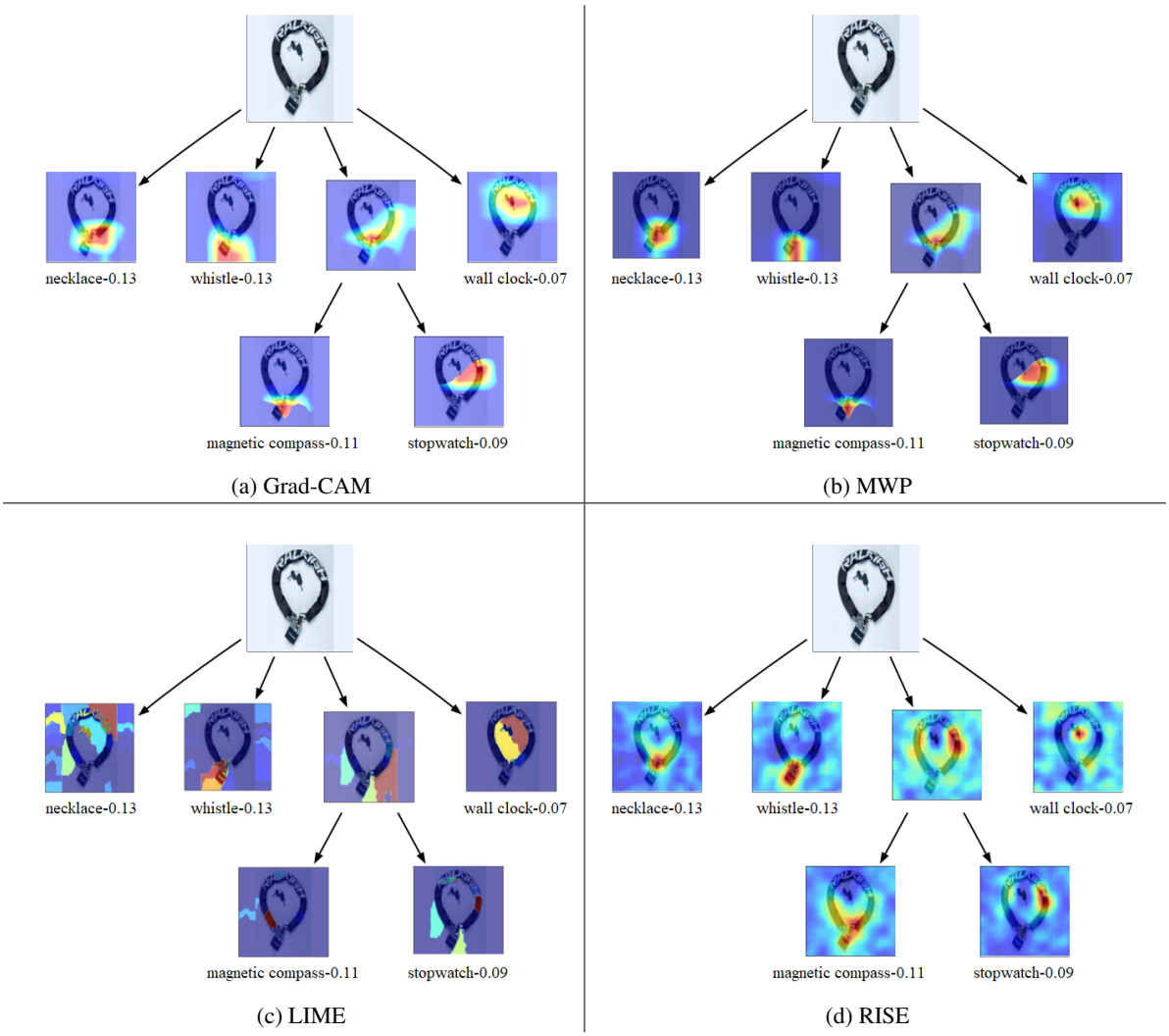


Figure C.11: CWOX-2s explanations with the four different base explainers on the Figure C.7 example.