

PAMI: partition input and aggregate outputs for model interpretation

Wei Shi
Sun Yat-sen University
shiw58@mail2.sysu.edu.cn

Wentao Zhang
Sun Yat-sen University
zhangwt65@mail2.sysu.edu.cn

Weishi Zheng
Sun Yat-sen University
wszheng@ieee.org

Ruixuan Wang
Sun Yat-sen University
wangruix5@mail.sysu.edu.cn

Abstract

There is an increasing demand for interpretation of model predictions especially in high-risk applications. Various visualization approaches have been proposed to estimate the part of input which is relevant to a specific model prediction. However, most approaches require model structure and parameter details in order to obtain the visualization results, and in general much effort is required to adapt each approach to multiple types of tasks particularly when model backbone and input format change over tasks. In this study, a simple yet effective visualization framework called PAMI is proposed based on the observation that deep learning models often aggregate features from local regions for model predictions. The basic idea is to mask majority of the input and use the corresponding model output as the relative contribution of the preserved input part to the original model prediction. For each input, since only a set of model outputs are collected and aggregated, PAMI does not require any model details and can be applied to various prediction tasks with different model backbones and input formats. Extensive experiments on multiple tasks confirm the proposed method performs better than existing visualization approaches in more precisely finding class-specific input regions, and when applied to different model backbones and input formats. The source code will be released publicly.

1. Introduction

Deep learning models have shown human-level performance in various machine learning tasks and started to be applied in real scenarios, such as face identification [26, 33, 77], medical image analysis [10, 37, 61], and language translation [70, 75, 76]. However, current deep learning models often lack interpretations for their decision making, which hinders the massive deployment of intelligent systems particularly in high-risk applications like med-

ical diagnosis and autonomous driving.

To improve interpretation of model predictions, multiple visualization approaches have been proposed to localize input regions or components which are more relevant to the model prediction given any specific input to the model. For image classification task as an example, the class activation map (CAM) and its variants utilize the output (i.e., feature maps) of certain convolutional layer in the convolutional neural networks (CNNs) and their contribution weights to find the image regions which are responsible for the specific model prediction given any input image [8, 41, 62, 73, 82], and the back-propagation approaches propagate the CNN output layer-by-layer to the input image space either based on gradient information (or its modified versions) at each layer [8, 62, 66–68] or based on the relevance between input elements and output at each layer [2, 6, 30, 48]. While the CAM-like approaches can only roughly localize relevant regions due to the lower resolution of feature maps, the back-propagation approaches often just find sparse and incomplete object regions relevant to the model prediction. Moreover, both types of approaches need either part of or the whole model structure and parameter details, which may be unavailable in some applications due to privacy or security concerns. When the model details are not available, the occlusion method may be utilized to roughly localize image regions relevant to the model prediction by occluding each local patch and checking the change in model output [54, 79]. However, the occlusion method often only localizes the most discriminative object part and misses the other parts which actually also contribute to the model prediction. LIME [58] is another method without requiring model details for model interpretation by locally approximating the model decision surface for any specific input, but it can often estimate the contributions of a subset of input parts and require an optimization process for interpretation of a specific model prediction. Furthermore, most existing visualization approaches for interpretation of model

predictions are developed for specific type of tasks (e.g., just for image classification), model backbone (e.g., for CNNs), and input format (e.g., just for image data). Substantial efforts are often required to adapt one visualization approach to various tasks (e.g., image caption) with different model backbones (e.g., Transformer backbone) or input formats (e.g., sequence of items).

Different from existing visualization approaches, a simple yet effective visualization framework for interpretation of model predictions is proposed in this study. The proposed framework, called PAMI (‘Partition input and Aggregate outputs for Model Interpretation’), is inspired by the observation that both humans and popular deep learning models extract and aggregate features of local regions for image understanding and decision making. Suppose a well-trained image classifier predicts an input image as a specific class. To find the relevant image regions and their contributions to the model prediction, the proposed framework first partitions the input into multiple parts, and then feeds only one part (with the remaining regions masked) to the model to obtain the corresponding output probability of the specific class. Aggregating the output probabilities over all the individual parts would result in an importance map representing the contribution of each input part to the original model prediction. In contrast to existing visualization approaches, the proposed PAMI framework does not require model structure and parameter details, can more likely find all possible input parts which are relevant to the model prediction, more precisely localize relevant parts, and work for various model backbones with different input formats. Such merits of the proposed PAMI framework has been confirmed by extensive experiments on multiple tasks with different model backbones and input formats.

2. Related Work

In computer vision, post-hoc interpretation of deep learning models focuses on either understanding of model neurons (e.g., convolutional kernels, output elements) which is independent of input information [4, 16, 32, 45, 78], or understanding of a specific model prediction given an input image [8, 15, 39, 54, 62]. This study belongs to the latter one, i.e., trying to understand what information in the input causes the specific model prediction.

Multiple approaches have been proposed for understanding of model predictions, including the activation map approach [8, 31, 62, 82], the back-propagation approach [24, 30, 79, 79], the perturbation approach [22, 54], the local approximation approach [58, 81] and the optimization-based approach [15, 39]. The activation map approach often obtains a class-specific activation map with the weighted sum of all feature maps often at the last convolutional layer, and considers the regions with stronger activation relevant to the specific model output [8, 62, 73, 82]. Since the activa-

tion map is often much smaller than the input image, only approximate image regions corresponding to the stronger activation regions can be localized for interpretation of the specific model prediction. Different from the activation approach which often works at higher layer of the deep learning model, the back-propagation approach tries to estimate the importance of each input pixel by propagating the specific model output layer-by-layer back to the input space. This can be obtained by calculating the gradient of the specific model output with respect to input elements at each layer [8, 62, 66–68], the input-relevant contribution of each kernel at each layer [34], or the relevance between output and each input element at each layer [2, 6, 30, 48]. The back-propagation approach considers each input pixel as an independent component and often only a subset of disconnected pixels in the relevant regions are estimated to be relevant to the model prediction.

To find local image regions rather than disconnected pixels relevant to the model prediction, the perturbation approach has been proposed by perturbing local image regions somehow and checking the change in model output of the predicted class [22, 79]. If certain perturbed local region causes large drop in the output, the local region in the input image is considered crucial to the original model prediction. Perturbation can be in the form of simply masking a local region by a constant pixel intensity [79], by neighboring image patches [83], or by blurring the original region information with a constant value or smoothing operator [22, 54]. This approach often can find only the most discriminative part of the relevant regions which are responsible for the model prediction, because perturbing less-discriminative part of the relevant regions often does not cause much drop in model output. Besides the perturbation approach, the local approximation approach provides another way to estimate the contribution of each meaningful image region (e.g., object parts, background region) to the model prediction [22, 58, 73]. This approach assumes that the model decision surface is locally linear in the region-based feature space for any specific input, and therefore can be approximated with a linear model in the feature space. The weight parameters in the linear model can directly indicate the contribution of each meaningful image region to the original model output, thus obtaining image regions most relevant to the model prediction.

While most of these visualization approaches to interpretation of model predictions were originally developed for image classification models, they have been extended or modified for other tasks [11, 72] or other deep learning models [9, 38]. Besides these approaches, prototype-based [49, 60] and attention-based [50] approaches have also been proposed for model interpretation. Note that except the local approximation approach and part of the perturbation approach (e.g., the occlusion method [22], most

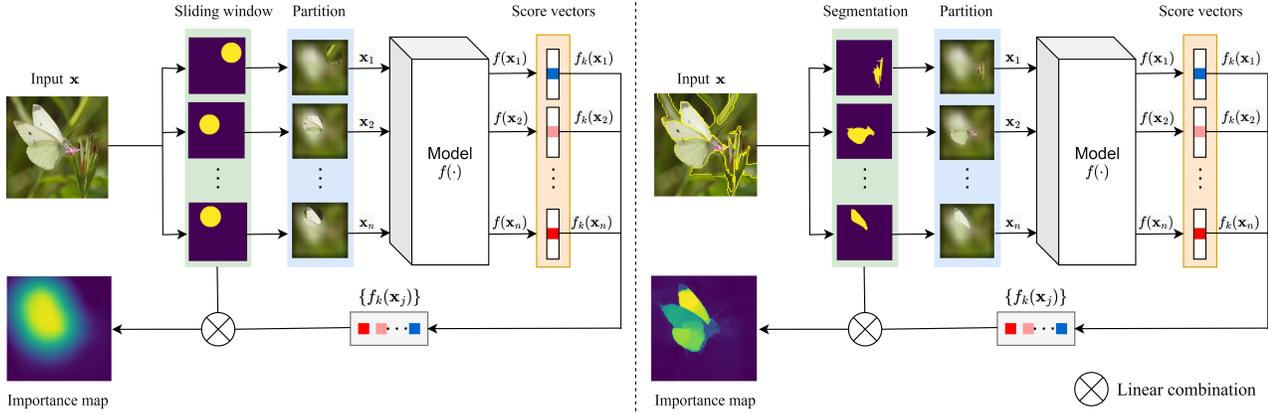


Figure 1. The proposed PAMI framework with the sliding window based (left half) or the pre-segmentation based (right half) input partition strategy. Each time only one local part of the input is preserved and the remaining parts are masked (blurred here).

approaches require at least part of the model structure and parameter details in order to find input parts which are relevant to the model prediction.

3. Method

In this study, we aim to provide interpretation for model prediction given any specific input to a well-trained and fixed deep learning model. The interpretation is demonstrated by estimating relative contribution of each input part to the specific model prediction and correspondingly localizing input regions or elements which are relevant to the model prediction. It is worth noting that no model structure and parameter details are assumed to be known during the model interpretation process.

3.1. Motivation

Although humans can often instantly recognize objects in images, certain attention mechanism in human brain is likely involved in the process of object recognition [29, 53]. In other words, humans often need to implicitly or explicitly attend to local regions for image understanding and object recognition. While the detailed human attention mechanism is yet to be further explored, initial studies [13] suggest that most local parts of an object in an image help humans recognize the object, and appearance of only an individual object part could help humans recall the corresponding class of the object. Consistent with the visual attention studies, recent exploration of convolutional neural networks (CNNs) shows that convolutional kernels even at higher convolutional layers (i.e., closer to the CNN output) often have smaller receptive fields than expected [44]. Considering that a global pooling is performed at the last convolutional layer in most CNN classifier models, it is widely accepted that CNN models largely depend on the collection of local image region features for image classification. For the

other type of deep learning model backbone Transformer and its variants (e.g., ViT [14], Swin Transformer [42]), since most items in the input sequence at each model layer correspond to components (e.g., words for a sentence input, image patches for an image input) of the original input, the final model prediction also largely depends on the collection of local features of the original input. With the above observation, we hypothesize that the model output response to each single component of the original input may directly imply the importance of the single input component for the specific model prediction.

3.2. The proposed PAMI framework

The proposed interpretation framework is demonstrated in Figure 1. For image classification task as an example, given a well-trained classifier model $f(\cdot)$ and any input image \mathbf{x} , denote by $f_c(\mathbf{x})$ the output prediction probability of the input image belonging to the c -th class. Suppose the model predicts the input as the k -th class, i.e., $f_k(\mathbf{x})$ is the maximum over all the output probabilities. To interpret the classifier's prediction, the proposed framework first partitions the input into multiple either overlapped or non-overlapped parts (see the following subsections), and then with the j -th individual part preserved and all the remaining image regions masked, the output probability $f_k(\mathbf{x}_j)$ of the k -th class for the majority-masked input \mathbf{x}_j is used to estimate the relative contribution of the preserved j -th image part to the original model prediction $f_k(\mathbf{x})$. By collecting and aggregating the output responses $f_k(\mathbf{x}_j)$'s over all the partitioned image parts, an importance map with the same spatial size as that of the input image \mathbf{x} can be generated to represent the contribution of each input element (i.e., pixel here). Local image regions with correspondingly higher response values in the importance map are supposed to contribute more to the model prediction $f_k(\mathbf{x})$, thus providing

visual evidence for the model prediction. It is worth noting that the interpretation framework can be applied to different tasks (e.g., image caption and sentiment analysis) with various input formats.

3.2.1 Input partition strategy I: sliding window

One simple way to partition input is to apply the sliding window strategy with a pre-defined window size and sliding step size, where the window is in certain regular shape (e.g., circular or rectangular). In this way, the original input can be easily partitioned into multiple parts, and each part can be more or less overlapped by its neighboring parts determined by window size and sliding step size. For each partitioned part, all the remaining image regions will be masked somehow (e.g., by black pixels or blurred version of the original regions), and the output probability of the originally predicted class can be directly obtained with the majority-masked image as input. Since each input element (e.g., pixel of an input image) could be covered by multiple partitioned parts, the contribution of each input element in the final importance map can be obtained by averaging the output probabilities of the originally predicted class over all the partitioned parts covering the input element.

Note that the window size would affect the resolution level of the final importance map. Although smaller window would result in desired higher resolution, an image part with much smaller size (i.e., too small window) could contain little semantic information such that it becomes challenging for the framework to estimate the contribution of the smaller image part to a specific model prediction. In practice, users can choose one appropriate window size for interpretation of model prediction, or multiple window sizes for interpretation at multiple scales of image parts.

3.2.2 Input partition strategy II: pre-segmentation

Another way to partition input is to pre-segment the input into multiple parts with certain segmentation strategy. When the input is an image, various unsupervised segmentation algorithms can be adopted for pre-segmentation of the input. In this study, super-pixel segmentation algorithms are used for input image partition [5, 19, 46, 57]. With a particular super-pixel segmentation method, an input image can be partitioned into multiple non-overlapped parts (i.e., super-pixels), with each part often having irregular form of region boundary and likely containing homogeneous visual information. As introduced above, the contribution of each super-pixel to the model prediction can be obtained by preserving the single super-pixel and masking the other super-pixels as the input and collecting the model output response of the predicted class to the majority-masked input.

In practice, due to the imperfect performance of any single super-pixel segmentation method, some super-pixels

may contain parts of both object region and background region, resulting in the importance map where part of background regions also have relatively higher responses. To alleviate such an issue, multiple super-pixel segmentation methods are employed, and multiple importance maps based on these segmentation methods are then averaged to estimate the contribution of each input element (e.g., pixels) to the original model prediction. The average importance map may be further improved by running the above process once more (i.e., second run), in which the super-pixel segmentation methods are performed on the average importance map rather than the original input image. In addition, when generating each majority-masked input, the highly smoothed version of the original input image is used to fill the corresponding masked regions.

Compared to the sliding window strategy, the partitioned parts by the pre-segmentation strategy have more precise and reasonable region boundaries particularly for image data. This in turn often leads to the final importance map with clear boundaries between object regions and background regions, thus more precisely locating the image regions which contribute to the model prediction. Note that both input partition strategies also work when input is a sequence of items. For example, when input is a sentence as in the sentiment analysis task, any input can be partitioned into words or phrases with either the sliding window strategy or appropriate pre-segmentation strategy.

3.3. Comparison with relevant studies

The proposed PAMI framework can provide interpretation of model prediction without requiring to know model structure and parameter information. In contrast, most existing interpretation methods requires either part of or the whole model details. For example, CAM and its variants need the feature maps from certain convolutional layers and part of model parameters in order to obtain the final class activation map [8, 41, 62, 73, 82], and the gradient-based methods need all the model details to calculate gradient information over model layers [8, 62, 66–68]. One exception is the occlusion method [54, 79] which does not require model details as the proposed PAMI framework. PAMI can be considered as an opposite version of the occlusion method, i.e., only preserving an input part versus only removing or occluding an input part for estimating the contribution of the input part to the model prediction. In image classification, occluding part of the foreground object in the image may not significantly affect the model prediction because the model can use the other object parts in the image for confident prediction. As a result, occlusion method may neglect contribution of certain object parts to the original model prediction, and often performs worse than the proposed PAMI.

Because the proposed PAMI framework can consider the

Table 1. Models and datasets used in experiments.

Task	Model source	Dataset
Classification	VGG19bn from PyTorch [52]	50000 images of ImageNet-2012 [59] validation set with 1000 classes
	VGG16 from TorchRay [21]	Ffirst 1000 images of Pascal VOC 2007 [18] test set with 20 classes
	VGG16 from TorchRay [21]	First 1000 images of COCO 2014 [40] instances validation set with 80 classes
Image caption	ClipCap [47]	COCO 2014 [40]
Sentiment analysis	Transformers library [74]	Sentiment140 [23]

well-trained model as a black-box, it can potentially work for various backbone structures (e.g., both CNN and Transformer backbones). In contrast, the majority of interpretation methods were proposed for the CNN backbone, and specific modifications are often required when applying existing interpretation methods (e.g., CAM or Grad-CAM) to other backbones like Transformer [9, 36] and graph neural networks [3, 12, 55]. Another merit of the proposed PAMI framework is its potential usage in multiple tasks with different input formats. While this study mainly use the image classification task for evaluation of the PAMI framework, PAMI in principle can be applied to various model prediction tasks, such as image caption and sentiment analysis, where the input data can be in the format of sentence or image. In contrast, most existing interpretation methods do not work across tasks without further modifications or extensions.

The most relevant interpretation methods are RISE [54] and ScoreCAM [73] which also estimate the importance map based on linear combination of input masks with weights from model outputs. However, RISE is based on a large set of randomly generated masks and ScoreCAM is based on the feature maps at last convolutional layer of the (CNN) model, both leading to low-resolution and often inappropriate importance maps.

4. Experiments

4.1. Experimental setup

In this study, three image classification datasets ImageNet-2012 [59], Pascal VOC 2007 [18], and COCO 2014 [40] were mainly used for evaluation of the proposed PAMI method. In addition, an image caption dataset COCO [40] and sentiment analysis dataset Sentiment140 [23] were also employed to show the wide applications of the proposed method. All the models were from the publicly released resources and evaluated on the corresponding validation or test set (see Table 1 for more details).

By default, for the sliding window strategy of the proposed PAMI, circular window with radius 40 pixels and step size 6 pixels was used to generate local image regions. For the pre-segmentation strategy, four super-pixel segmentation algorithms, i.e., felzenszwalb [19], SLIC [57], water-

shed [46] from scikit-image library [71], and SEEDS [5] from the OpenCV library [7] were utilized respectively for pre-segmentation of each image into multiple regions. Considering region of interest (i.e., relevant region to the model prediction) could vary a lot over images, each segmentation algorithm was run multiple times with different hyper-parameter settings to generate sets of local regions at different scales (see the supplementary A for detailed settings). The importance maps over all hyper-parameter settings and all the four pre-segmentation algorithms were averaged as the estimated importance map. A Gaussian kernel with size 49×49 pixels and standard deviation 100 was used to generate the smoothed (blurred) image for region masking.

The proposed PAMI was compared with widely used visualization methods for interpretation of model predictions, including Gradient [63], GradCAM [62], ScoreCAM [73], RISE [54], FullGrad [67], MASK [22], Occlusion [79], GuidedBP [66], SmoothGrad [65] and LRP [48]. The default hyper-parameter setting for each method was adopted (see supplementary B for details). Besides qualitative evaluation, quantitative evaluation was also performed using the pointing game [80] and the insertion metric [54]. In the pointing game, it measures whether the pixel with the highest activation in the importance map is successfully within the image region of the object corresponding to the interpreted class, with ‘hit’ for success and ‘miss’ for failure. The average hit rate over all classes is used to measure performance of each method. For the insertion metric, it gradually restores the original pixels in the blurred version of the original image, with pixels having higher activation in the importance map restored earlier. Higher insertion score would indicate a better performance of interpretation.

4.2. Qualitative evaluation

The efficacy of the proposed PAMI method was extensively evaluated on the ImageNet-2012 data. Figure 2 demonstrates the visualization results on multiple representative images with the VGG19bn model. The visualization results were generated with respect to the model output of the ground-truth class for each image. It can be observed that, the proposed PAMI with the pre-segmentation strategy for input partition (last column) can often precisely and largely completely localize the object regions which are actually relevant to the specific model prediction, while existing methods roughly localize either object regions at low resolution, sparse part of object regions, disconnected pixels within object regions, or even irrelevant background regions. Multiple colors within relevant region in the importance map from the proposed PAMI method suggests that different object regions may have different degrees of contributions to the model prediction. On the other hand, the proposed PAMI simply with the sliding window strategy can often obtain similar performance as GradCAM but

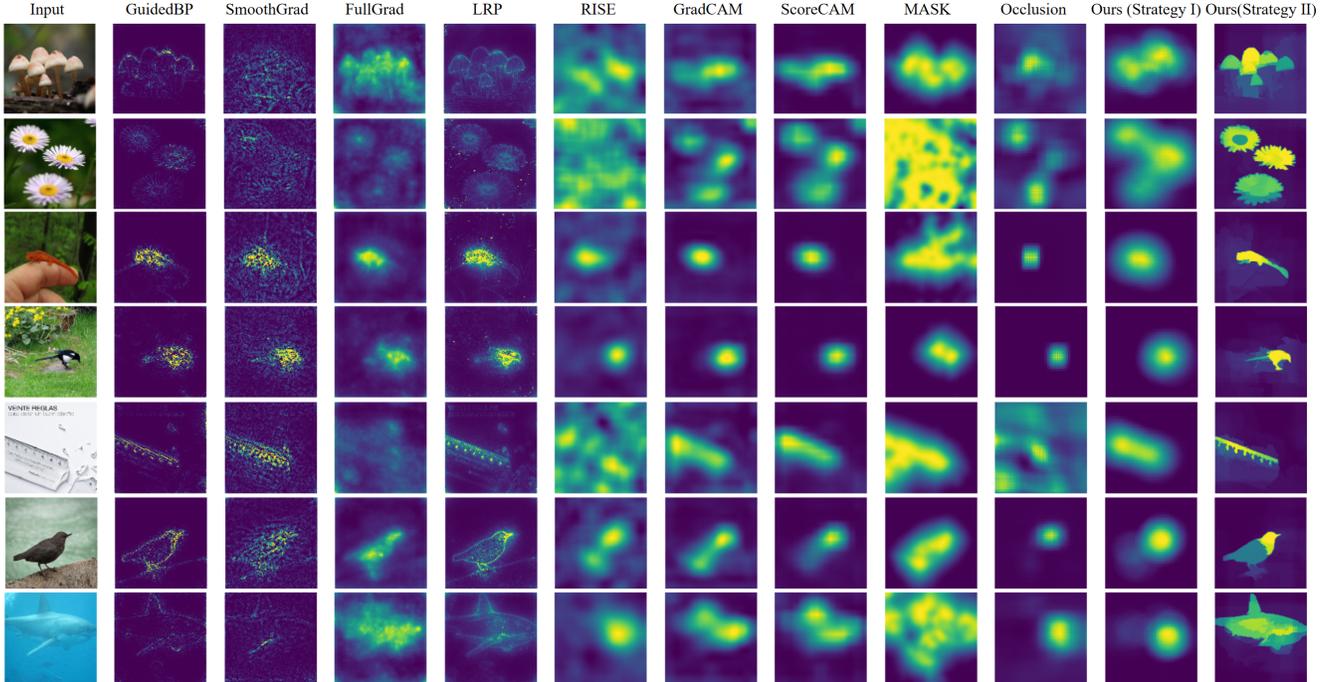


Figure 2. Qualitative evaluation of the proposed PAMI method on the ImageNet-2012 dataset. The first column list the input images to the classifier. The last two columns are the importance maps from the proposed PAMI method with the two strategies respectively, and all the other columns are from the representative strong baseline methods. In each importance map or heatmap, higher activation is in yellow and lower activation is in blue.

without requiring model structure and parameter details.

Another observation is that the proposed PAMI method can work more stably than existing methods under challenging conditions. In particular, the PAMI method can well localize small-scale objects (rows 3 & 4) and relatively large-scale objects (row 7) in images, and also can precisely localize the regions of multiple object instances of the same class (rows 1 & 2). In comparison, most existing methods often perform worse under at least some of these challenging conditions. Similar observations were also obtained on the PASCAL-VOC dataset and the COCO dataset (see supplementary C), consistently supporting that the proposed PAMI method is effective in providing visual evidence for interpretation of model predictions.

4.3. Quantitative evaluation

Although the non-existence of ground-truth or ideal interpretation for any specific model prediction makes it challenging to quantitatively evaluate any interpretation method, the pointing game [80] and the insertion metric [54] have been proposed to roughly evaluate the performance on correctly localizing regions relevant to model predictions. With the pointing game, Table 2 (columns 2, 4, and 6) shows that the proposed PAMI method with the pre-segmentation partition strategy (last row) has similar hitting

rate on the ImageNet and COCO datasets compared with the best baseline GradCAM and higher hitting rate than all the baselines on the VOC dataset, suggesting that the local region which is considered most relevant to the model prediction by PAMI is often actually part of the object region. Similarly with the insertion metric (Table 2, columns 3, 5, and 7), PAMI has the best performance on ImageNet and VOC datasets, and is close to the best baseline RISE on COCO dataset, again supporting that PAMI can well localize image regions belonging to the interpreted class. More experimental details for quantitative evaluation can be found in the supplementary D.

4.4. Generality of the proposed PAMI method

To evaluate the generality of the proposed method, well-trained deep learning classifiers with multiple different backbones were employed, including VGG16 [64], ResNet50 [25], SE-ResNet [27], InceptionV3 [69], DenseNet121 [28], RegNet-X-16GF [56], ConvNext-Tiny [43], ViT-L-16 [14] and SwinT-Tiny [42]. From Figure 3, we can see that the proposed PAMI method can robustly and precisely localize the object regions which are relevant to the model prediction for each input image, regardless of the classifier backbones. In contrast, for each representative baseline method, the importance maps often

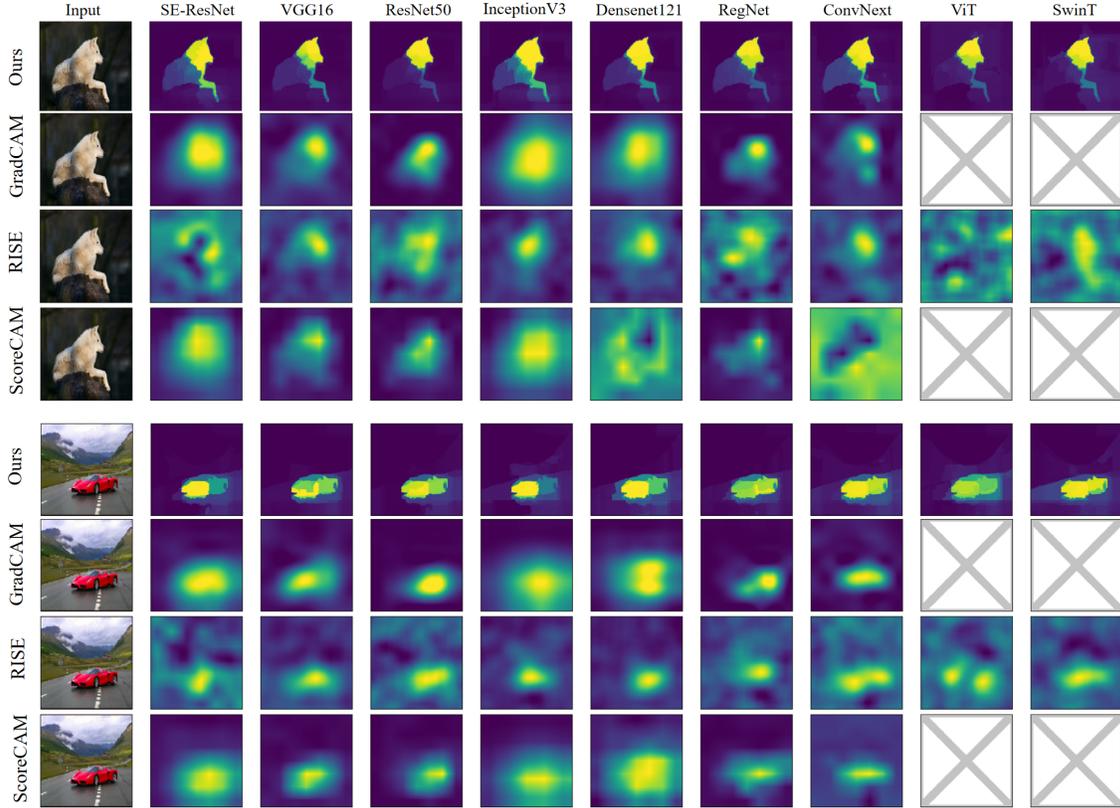


Figure 3. Representative visualization results from the proposed method and representative baselines with different model backbones. Cross sign means the relevant baseline is not working on the corresponding model backbone.

Table 2. Quantitative evaluation of the proposed PAMI method on the three image datasets. ‘Random’: randomly generating a heatmap for each input image. ‘Center’: taking the fixed image center position as the highest activation point for each input image.

Method	ImageNet		VOC		COCO	
	Pointing	Insertion	Pointing	Insertion	Pointing	Insertion
Random	47.89	-	33.39	-	11.27	-
Center	81.96	-	70.79	-	25.97	-
Gradient [63]	83.14	0.1928	72.61	0.3321	34.65	0.1585
GuidedBP [66]	83.95	0.2632	71.14	0.4737	32.83	0.1935
Occlusion [79]	84.53	<u>0.5741</u>	84.49	0.6753	54.83	0.3229
MASK [22]	84.49	0.4867	76.30	0.5616	49.78	0.2664
RISE [54]	91.58	0.5460	82.43	<u>0.6885</u>	<u>56.95</u>	0.3305
SmoothGrad [65]	86.51	0.2494	75.38	0.3824	39.45	0.1753
GradCAM [62]	93.22	0.5154	<u>87.45</u>	0.5720	57.95	0.2660
ScoreCAM [73]	92.01	0.5191	86.51	0.6030	55.01	0.2656
FullGrad [67]	87.01	0.5045	77.58	0.5049	44.52	0.2362
Ours (Strategy I)	89.17	0.5566	74.95	0.6133	48.19	0.2688
Ours (Strategy II)	<u>92.32</u>	0.5965	87.87	0.7213	56.85	<u>0.3291</u>

change over model backbones and even may not work for the Transformer backbone ViT and SwinT. This confirms that the proposed PAMI method is more stable and can be applied to interpretation of model predictions with various model backbones. Please see more results with consistent observations from the supplementary E.

4.5. Sensitivity and ablation study

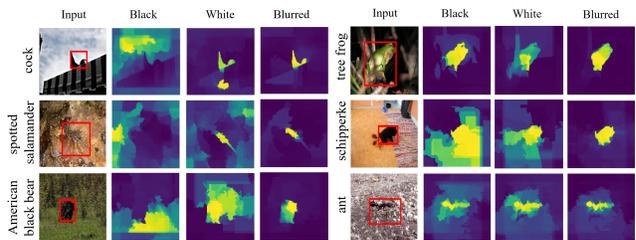


Figure 4. Importance map of an representative input based on different masking operators. ‘Black/White/Blurred’: masking types. Red boxes in images: object regions relevant to model predictions.

In the proposed method, masking majority of the input is one necessary step to estimate the contribution of each single region or part of the input. There are multiple choices for the masking operator. When input is an image, the to-be-masked region could be replaced by constant intensity value such as 0 (i.e., becoming black) or 255 (i.e., becoming white), or by the blurred version of the input image. Figure 4 demonstrates exemplar results with different masking operators, which shows that masking by blurred region

(‘Blurred’) results in better separation between the background regions and the object regions of interest at the first run, in turn leading to better importance maps with clearer boundaries between background and object regions at the second (i.e., final) run. When the majority of the input is replaced with extreme black or white pixels, the modified input image becomes further from the original class distribution in the feature space compared to the modified image with blurred region, which makes the model prediction unstable and therefore may not faithfully represent the importance of the preserved local region.

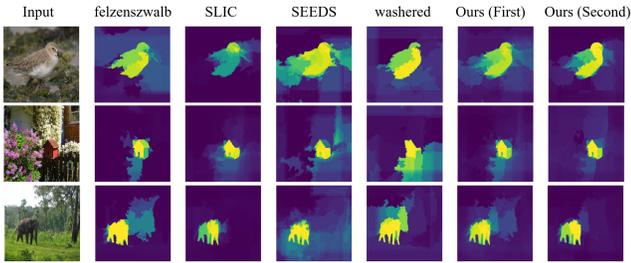


Figure 5. Effect of applying multiple pre-segmentation algorithms. From left to right: input, importance maps from four individual pre-segmentation algorithms, and average importance maps at the first and the second run respectively.

In addition, the average of multiple importance maps from multiple pre-segmentation algorithms often results in better visualization than that of using a single pre-segmentation algorithm, as demonstrated in Figure 5 (columns 2-5 vs. column 6). Figure 5 also shows that the second run (last column) can often refine the importance map from the first run (column 6). This is probably because sometimes the adopted pre-segmentation algorithms cannot well separate background regions from object regions at the first run, but the initially estimated importance map from the first run provides alternative information for pre-segmentation algorithms to well separate background regions from object regions. Note that the proposed PAMI is independent of the adopted pre-segmentation algorithms, and better pre-segmentation algorithms can be adopted to replace the current ones in the future. Please see more ablation study results with consistent observations from the supplementary F.

4.6. Extensive applications of the proposed PAMI

The proposed PAMI method is expected to work for multiple types of prediction tasks. For example, based on a well-trained image caption model [47], the PAMI method can well localize the image regions relevant to the predicted words (e.g., ‘dog’, ‘laying’, ‘sidewalk’, ‘bicycle’) which refer to any object or behaviour in the image (Figure 6, rows 1, 3), while the representative baseline method RISE often cannot precisely localize relevant regions (Figure 6, rows

2, 4). Another example is for the sentiment analysis task, where the model tries to evaluate whether the viewpoint in an input sentence is positive or negative. With an input partition strategy (see supplementary G) similar to the pre-segmentation for images, the proposed PAMI method can directly and correctly estimate the contribution of each word to the final model prediction (Figure 7). These results (also see more results in the supplementary G) confirm that the proposed PAMI method can work for various tasks with different input modalities.



Figure 6. Two exemplar visualization results from the proposed method and the strong baseline RISE for the image caption task.

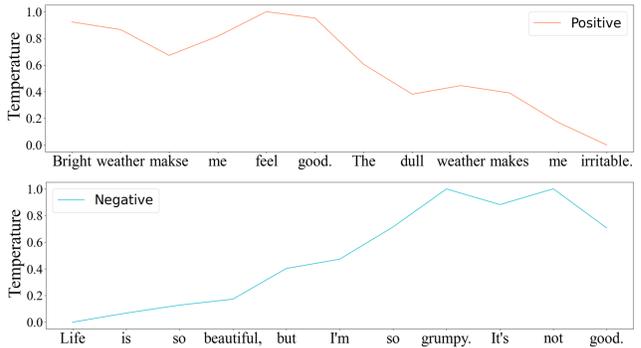


Figure 7. Two exemplar visualization results from the proposed method for the sentiment analysis task. The first row shows the contribution of each word to a positive emotion prediction, and the second row for a negative emotion prediction.

5. Conclusion

In this study, we propose a novel visualization method PAMI for interpretation of model predictions. PAMI does not require any model parameter details and works stably across model backbones and input formats. Compared to

existing visualization approaches, PAMI can more likely and precisely find the possible local input regions which contribute to the specific model prediction to some degree. It can be used as a plug-in component and applied to multiple types of prediction tasks, which has been partly confirmed by image classification, image caption, and sentiment analysis tasks. Its utility in more tasks including various natural language processing tasks will be evaluated in future work.

References

- [1] Visual explanation methods for pytorch. <https://github.com/vlue-c/Visual-Explanation-Methods-PyTorch>, 2020. 13
- [2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7):e0130140, 2015. 1, 2
- [3] Pablo Barceló, Egor V Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan-Pablo Silva. The Logical Expressiveness of Graph Neural Networks. In *International Conference on Learning Representations*, 2020. 5
- [4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. 2
- [5] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26, 2012. 4, 5, 13
- [6] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In *International Conference on Artificial Neural Networks*, pages 63–71, 2016. 1, 2
- [7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5
- [8] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *IEEE Winter conference on Applications of Computer Vision*, pages 839–847, 2018. 1, 2, 4
- [9] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 2, 5
- [10] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis. *arXiv preprint arXiv:1904.00625*, 2019. 1
- [11] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, 2020. 2
- [12] Nima Dehmamy, Albert-László Barabási, and Rose Yu. Understanding the Representation Power of Graph Neural Networks in Learning Graph Topology. *Advances in Neural Information Processing Systems*, 32, 2019. 5
- [13] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995. 3
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020. 3, 6
- [15] Andrew Elliott, Stephen Law, and Chris Russell. Explaining Classifiers using Adversarial Perturbations on the Perceptual Ball. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10693–10702, 2021. 2
- [16] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing Higher-Layer Features of a Deep Network. *University of Montreal*, 1341(3):1–13, 2009. 2
- [17] Mark Everingham. The PASCAL Visual Object Classes Challenge 2007. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 13
- [18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 5
- [19] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 4, 5, 13
- [20] François-Guillaume Fernandez. Torchcam: class activation explorer. <https://github.com/frgfm/torchcam>, March 2020. 13
- [21] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Torchray. <https://github.com/facebookresearch/TorchRay>, 2019. 5, 13
- [22] Ruth C Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 2, 5, 7, 13
- [23] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. <http://help.sentiment140.com/home>, 2009. 5, 13
- [24] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding Individual Decisions of CNNs via Contrastive Backpropagation. In *Asian Conference on Computer Vision*, pages 119–134, 2018. 2
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

- [26] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1761–1773, 2018. **1**
- [27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. **6**
- [28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. **6**
- [29] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. **3**
- [30] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pages 4176–4185, 2019. **1, 2**
- [31] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. **2**
- [32] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677, 2018. **2**
- [33] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. GroupFace: Learning Latent Groups and Constructing Group-based Representations for Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020. **1**
- [34] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: PatternNet and PatternAttribution. In *International Conference on Learning Representations*, 2018. **2**
- [35] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020. **13**
- [36] Ruiwen Li, Zheda Mai, Chiheb Trabelsi, Zhibo Zhang, Jongseong Jang, and Scott Sanner. TransCAM: Transformer Attention-based CAM Refinement for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2203.07239*, 2022. **5**
- [37] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. Medical Image Segmentation using Squeeze-and-Expansion Transformers. In *International Joint Conferences on Artificial Intelligence*, 2021. **1**
- [38] Xiaoxiao Li, Nicha C Dvornek, Yuan Zhou, Juntang Zhuang, Pamela Ventola, and James S Duncan. Graph Neural Network for Interpreting Task-fMRI Biomarkers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–493, 2019. **2**
- [39] Dohun Lim, Hyeonseok Lee, and Sungchan Kim. Building Reliable Explanations of Unreliable Neural Networks: Locally Smoothing Perspective of Model Interpretation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6468–6477, 2021. **2**
- [40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pages 740–755, 2014. **5, 13**
- [41] Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, and Qi Tian. Partial Class Activation Attention for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16836–16845, 2022. **1, 4**
- [42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. **3, 6**
- [43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. **6**
- [44] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 29, 2016. **3**
- [45] Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015. **2**
- [46] Fernand Meyer. Color Image Segmentation. In *International Conference on Image Processing and its Applications*, pages 303–306, 1992. **4, 5, 13**
- [47] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, 2021. **5, 8**
- [48] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. **1, 2, 5, 13**
- [49] Meike Nauta, Ron van Bree, and Christin Seifert. Neural Prototype Trees for Interpretable Fine-grained Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021. **2**
- [50] Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Sarafí García Vázquez, and Alejandro Álvaro Ramírez Acosta. Visual vs internal attention mechanisms in deep neural networks for image classification and object detection. *Pattern Recognition*, 123:108411, 2022. **2**
- [51] Utku Ozbek. Pytorch cnn visualizations. <https://github.com/utkuozbulak/pytorch-cnn-visualizations>, 2019. **13**

- [52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. [5](#)
- [53] Steven E Petersen and Michael I Posner. The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, 35:73, 2012. [3](#)
- [54] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference*, 2018. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [13](#)
- [55] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability Methods for Graph Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10772–10781, 2019. [5](#)
- [56] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. [6](#)
- [57] Xiaofeng Ren and Jitendra Malik. Learning a Classification Model for Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10–10, 2003. [4](#), [5](#), [13](#)
- [58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. [1](#), [2](#)
- [59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. [5](#), [13](#)
- [60] Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1420–1430, 2021. [2](#)
- [61] Edward Sanderson and Bogdan J Matuszewski. FCN-Transformer Feature Fusion for Polyp Segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 892–907, 2022. [1](#)
- [62] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. [1](#), [2](#), [4](#), [5](#), [7](#)
- [63] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014. [5](#), [7](#), [13](#)
- [64] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. [6](#)
- [65] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. [5](#), [7](#)
- [66] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for Simplicity: The All Convolutional Net. In *International Conference on Learning Representations Workshop*, 2015. [1](#), [2](#), [4](#), [5](#), [7](#)
- [67] Suraj Srinivas and François Fleuret. Full-Gradient Representation for Neural Network Visualization. *Advances in Neural Information Processing Systems*, 32, 2019. [1](#), [2](#), [4](#), [5](#), [7](#), [13](#)
- [68] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017. [1](#), [2](#), [4](#)
- [69] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. [6](#)
- [70] Sho Takase and Shun Kiyono. Rethinking Perturbations in Encoder-Decoders for Fast Training. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5767–5780, 2021. [1](#)
- [71] Stéfán van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. [5](#)
- [72] Eric Wallace, Matt Gardner, and Sameer Singh. Interpreting Predictions of NLP models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, 2020. [2](#)
- [73] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 24–25, 2020. [1](#), [2](#), [4](#), [5](#), [7](#), [13](#)
- [74] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: system demonstrations*, pages 38–45, 2020. [5](#)
- [75] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-Drop: Regularized Dropout for Neural Networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021. [1](#)
- [76] Haoran Xu, Benjamin Van Durme, and Kenton Murray. BERT, mBERT, or BiBERT? A Study on Contextualized

- Embeddings for Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675, 2021. [1](#)
- [77] Mengjia Yan, Mengao Zhao, Zining Xu, Qian Zhang, Guoli Wang, and Zhizhong Su. VarGFaceNet: An Efficient Variable Group Convolutional Neural Network for Lightweight Face Recognition. In *International Conference on Computer Vision Workshop*, 2019. [1](#)
- [78] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. In *International Conference on Machine Learning Workshop*, 2015. [2](#)
- [79] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*, pages 818–833, 2014. [1](#), [2](#), [4](#), [5](#), [7](#), [13](#)
- [80] 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. [5](#), [6](#)
- [81] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object Detectors Emerge in Deep Scene CNNs. In *International Conference on Learning Representations*, 2015. [2](#)
- [82] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. [1](#), [2](#), [4](#)
- [83] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In *International Conference on Learning Representations*, 2017. [2](#)

A. Hyper-parameters for pre-segmentation

The hyper-parameters of the four pre-segmentation algorithms used in both the first and the second run are summarized in table 3. The hyper-parameter names used in the methods correspond to those in the skimage and cv2 packages.

Table 3. Hyper-parameter configuration for each segmentation algorithm.

Method	Hyper-parameter
felzenszwalb [19]	scale=250, 200, 150, 100, 70, 50 sigma=0.8 min_size=784
SLIC [57]	n_segments=10, 20, 30, 40, 50, 60, 70, 80 compactness=20
SEEDS [5]	num_superpixels=10, 20,30 num_levels=5 n_iter=10
watershed [46]	markers=10, 20, 30 compactness=0.0001

B. Hyper-parameters in baseline methods

Details of each baseline method and reference source code were provided in Table 4. For gradient-based methods, following the related work [63], the maximum importance value among the three channels at each spatial position was used as the final importance value for the spatial position in the importance map. For LRP [48], the values from three channels were averaged as the final result for each spatial position.

Table 4. Hyper-parameter configuration for each baseline method.

Method	Hyper-parameter	Code Source
GradCAM [19]	The last layer of feature extractor	PyTorch CNN Visualizations [51]
ScoreCAM [73]	The last layer of feature extractor	TorchCam [20]
Occlusion [79]	strides=6, shapes=(3, 40, 40)	Captum [35]
RISE [54]	num_mask=4000, cell_size=7 probability=0.5	TorchVex [1]
MASK [22]	TV_beta=3, lr=0.1, max_iterations=500 l1_coeff=0.01, tv_coeff=0.2	TorchVex [1]
FullGrad [5]	-	original implementation [67]

C. More qualitative evaluation

C.1. More results on ImageNet-2012

More qualitative evaluation of the proposed PAMI method on the ImageNet-2012 dataset [59] can be seen in Figure 8, supporting the effectiveness of the method.

C.2. More results on Pascal VOC 2007

The effectiveness of the proposed PAMI method was also validated on the Pascal VOC 2007 dataset [17], as shown in Figure 9.

C.3. More results on COCO 2014

The effectiveness of the proposed PAMI method was also validated on the COCO 2014 dataset [40], as shown in Figure 10.

D. Details of quantitative evaluation

For the pointing game, the process provided by TorchRay [21] was followed. On the Pascal VOC and the COCO datasets, the evaluation code provided by TorchRay was directly used, and on the ImageNet dataset, the same process was performed by marking points that fall within the bounding box of the object as hits and the rest as misses. For the insertion metric, a Gaussian kernel with size 49×49 pixels and standard deviation 100 was used to generate the blurred image where we gradually restore the original pixels from. The importance map for quantitative evaluation were generated with respect to the model output of the ground-truth class for each image. Two exemplar results were shown in Figure 11.

E. Generality of the PAMI

More results with different model backbones were shown in Figure 12, supporting the generality of the proposed method.

F. More ablation study results

F.1. Effect of different masking operators

More visualization results based on different masking operators were provided in Figure 13, which shows that the blurred version results in better results especially when image background is complex.

F.2. Effect of multiple pre-segmentation algorithms

More results in Figure 14 shows that using multiple pre-segmentations and two runs result in better visualizations.

G. Extensive applications of PAMI

G.1. Image caption task

More experimental results for the image caption task on the COCO dataset can be seen in Figure 15.

G.2. Sentiment analysis task

More test sentences in Sentiment140 [23] were randomly selected for effectiveness evaluation of the proposed method in sentiment analysis tasks. The experimental results can be seen in Figure 16.

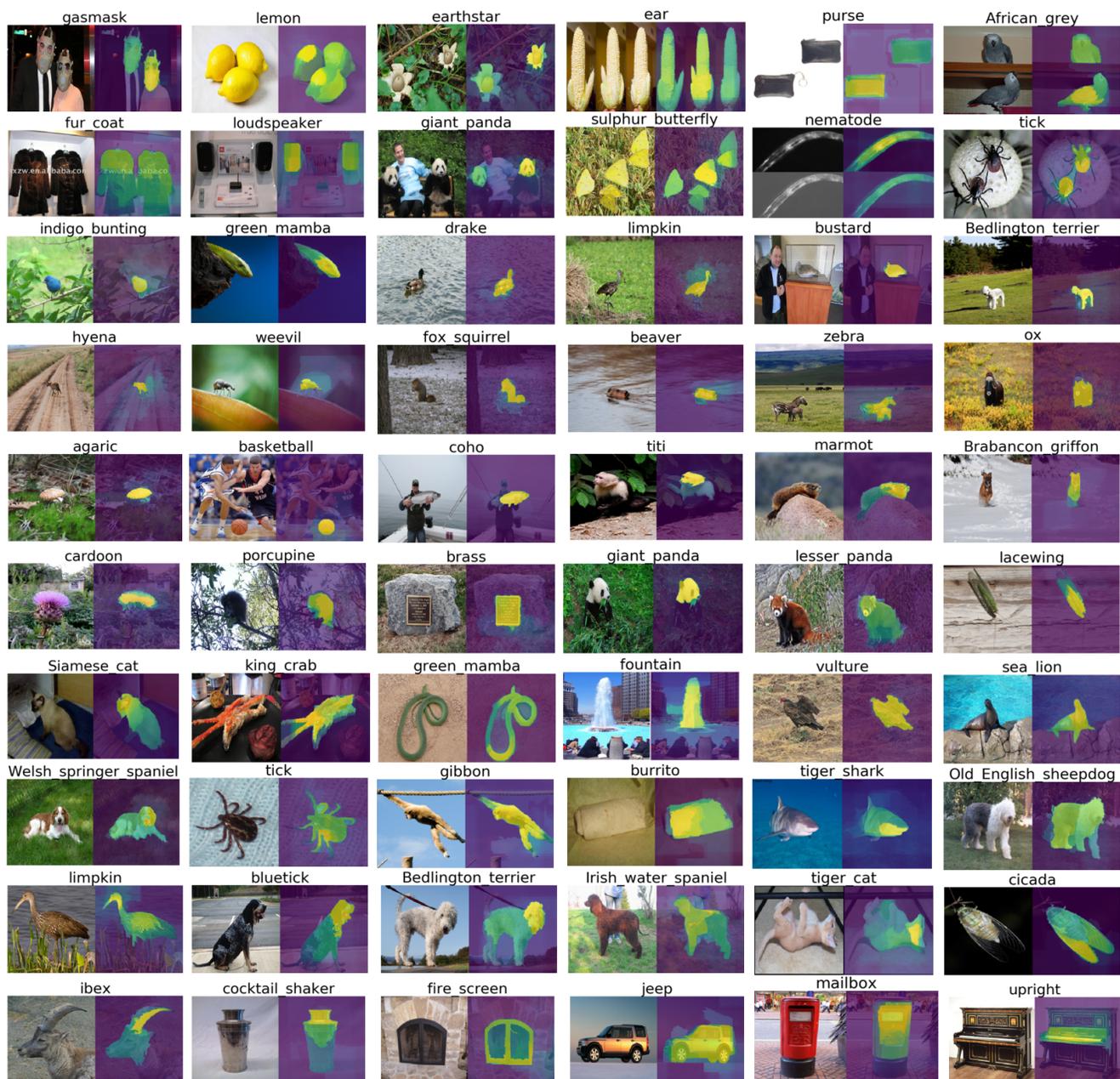


Figure 8. More qualitative evaluation of the proposed PAMI method on the ImageNet-2012 dataset. For each pair: input image is on the left and the visualization result from the proposed PAMI is on the right.

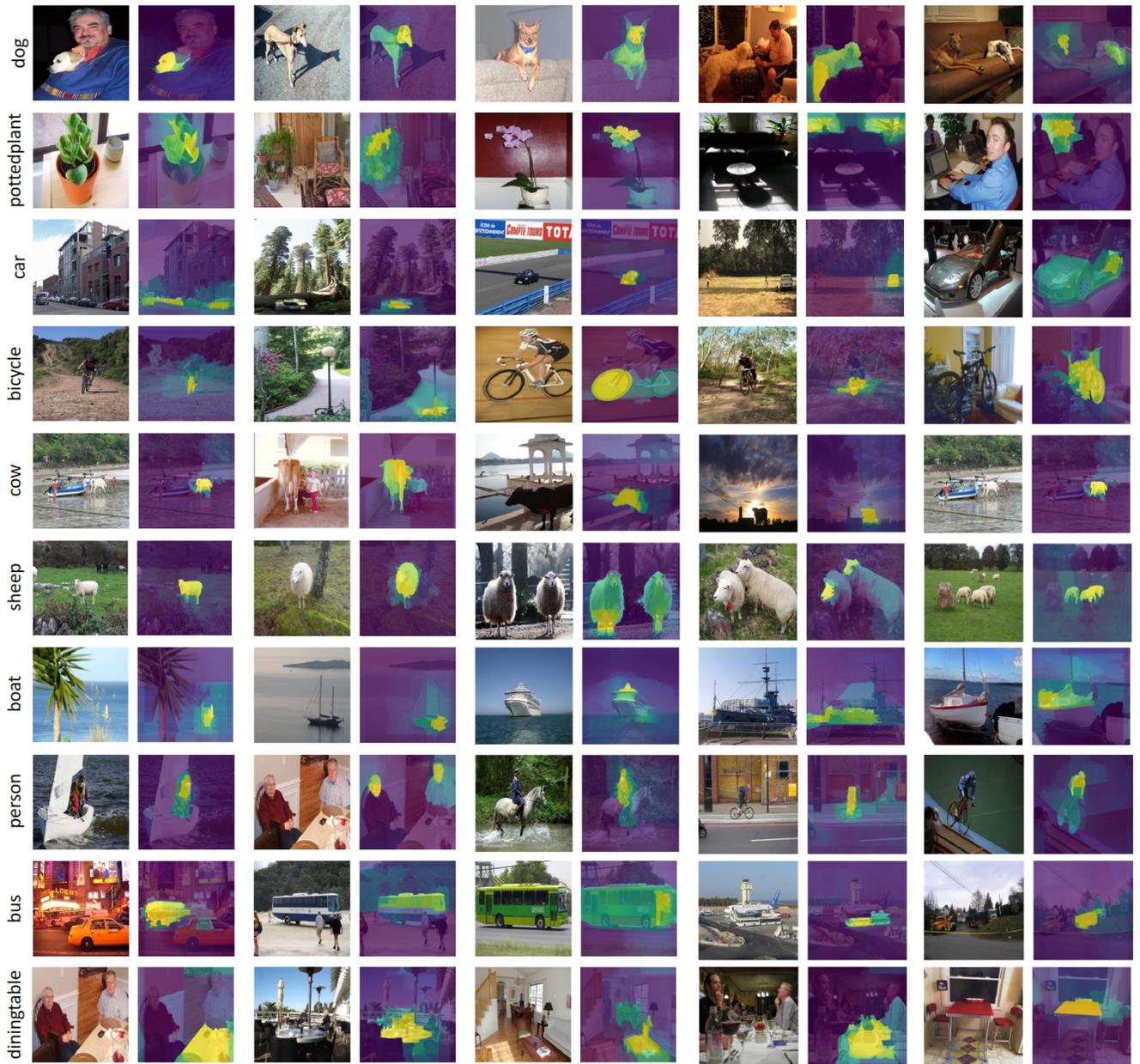


Figure 9. Qualitative evaluation of the proposed PAMI method on Pascal VOC 2007 dataset. Note that each input was resized to be square for demonstration.

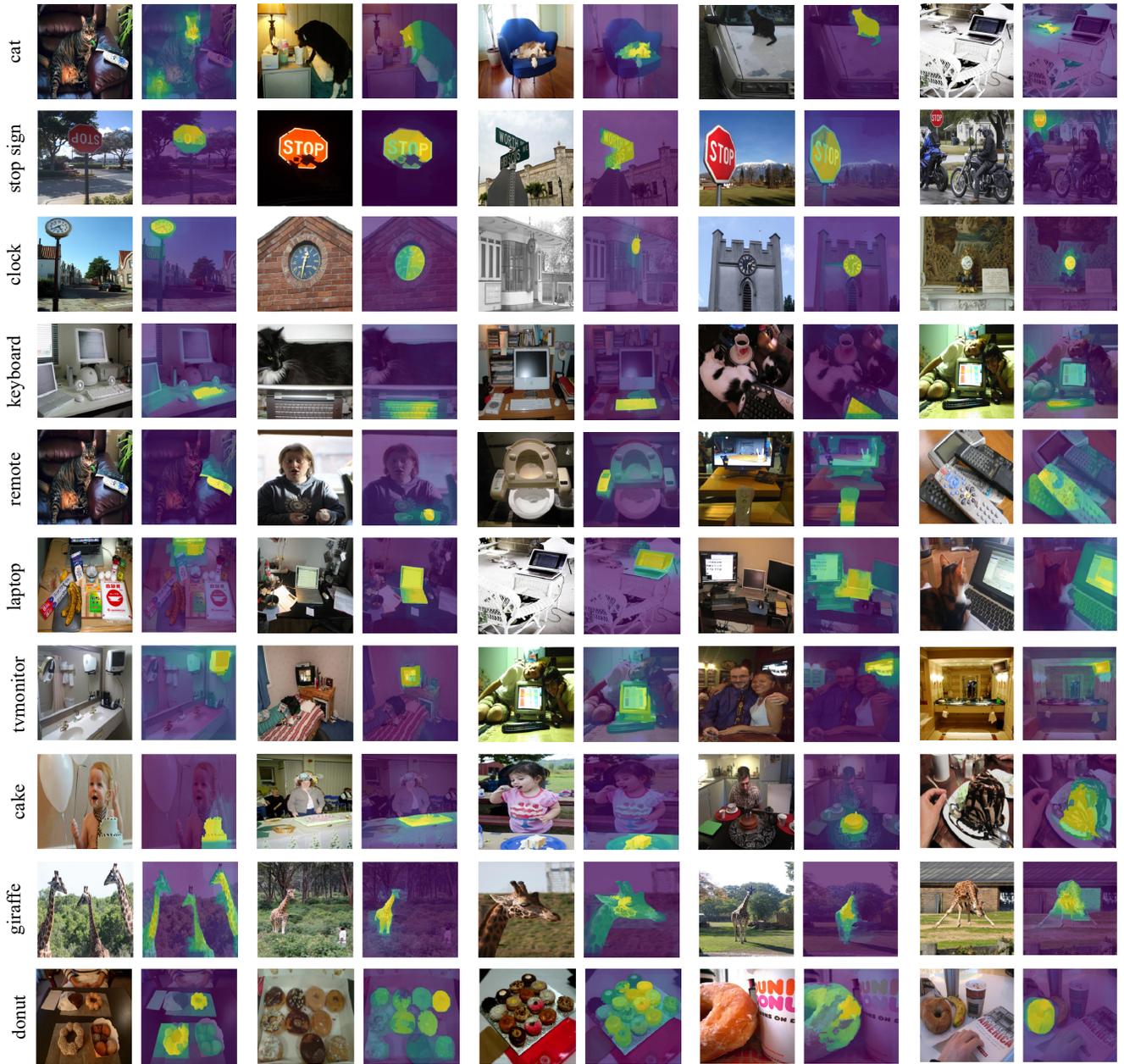


Figure 10. Qualitative evaluation of the proposed PAMI method on COCO 2014 dataset. Note that each input was resized to be square for demonstration.

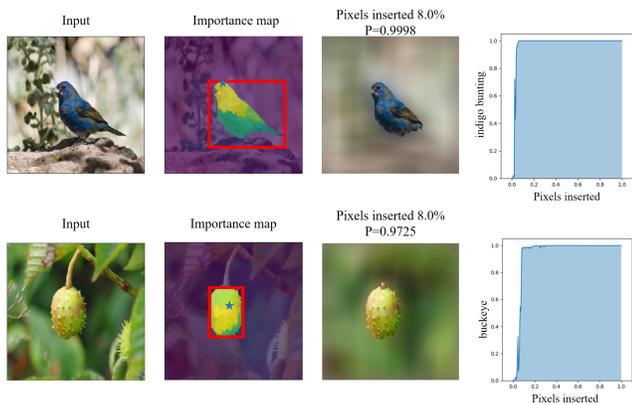


Figure 11. Two examples of quantitative evaluation. First column: original input image. Second column: the importance map and the most important pixel marked with asterisks. Third column: image after restoring a certain percentage of pixels and the prediction probability of the image being the ground-truth class by the model. Fourth column: the insertion curve and the area under the curve as the insertion score.

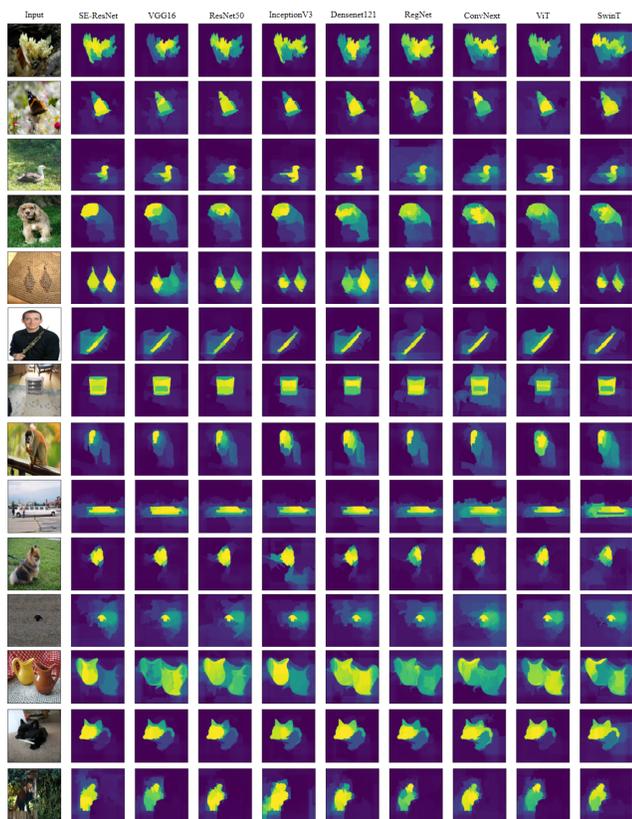


Figure 12. More representative visualization results from the proposed method with different model backbones.

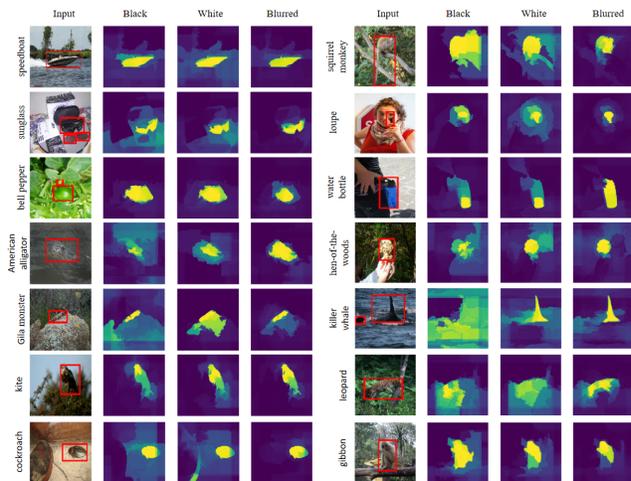


Figure 13. Importance maps of more representative inputs based on different masking operators.

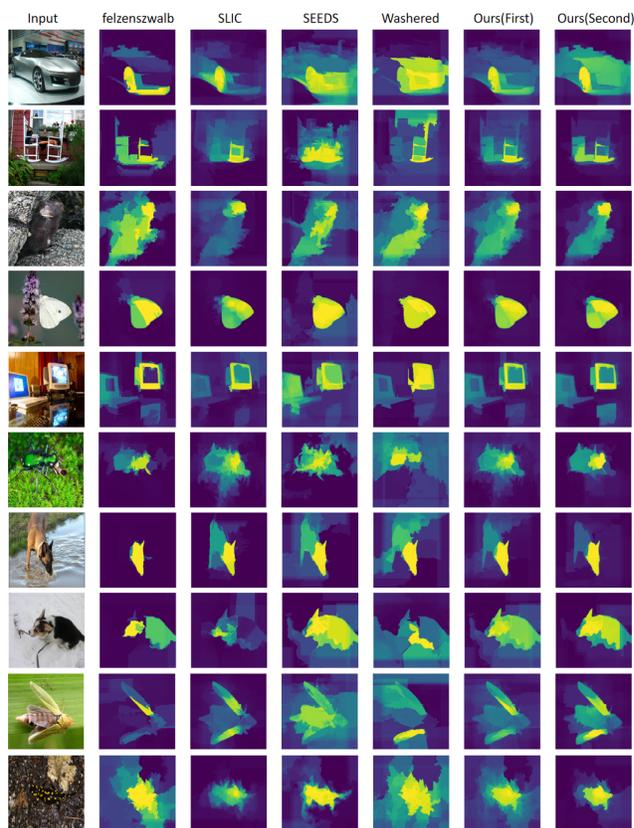


Figure 14. Effect of applying multiple pre-segmentation algorithms. From left to right: input, importance maps from four individual pre-segmentation algorithms, and average importance maps at the first and the second run respectively.

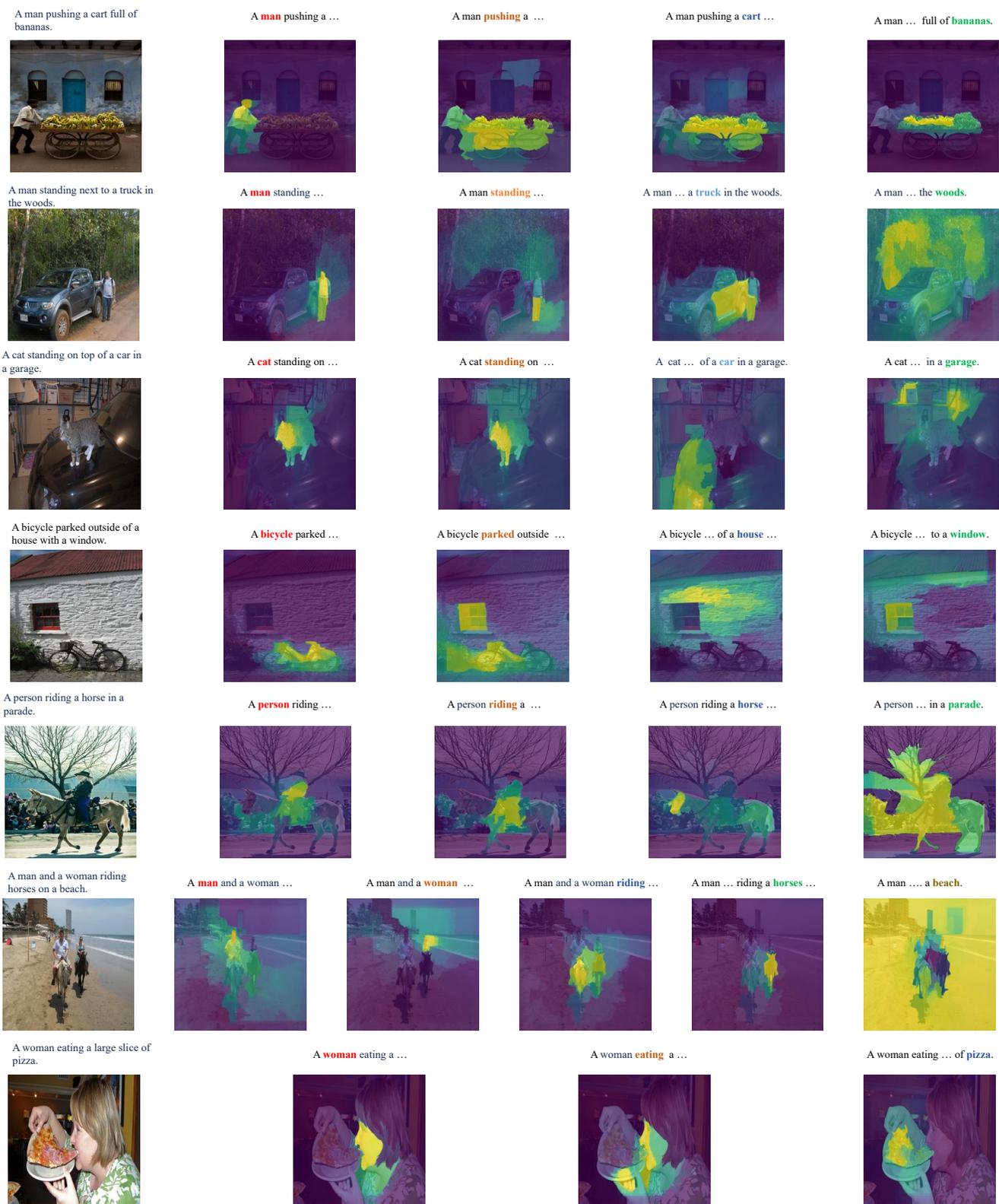


Figure 15. More visualization results from the proposed method for the image caption task.

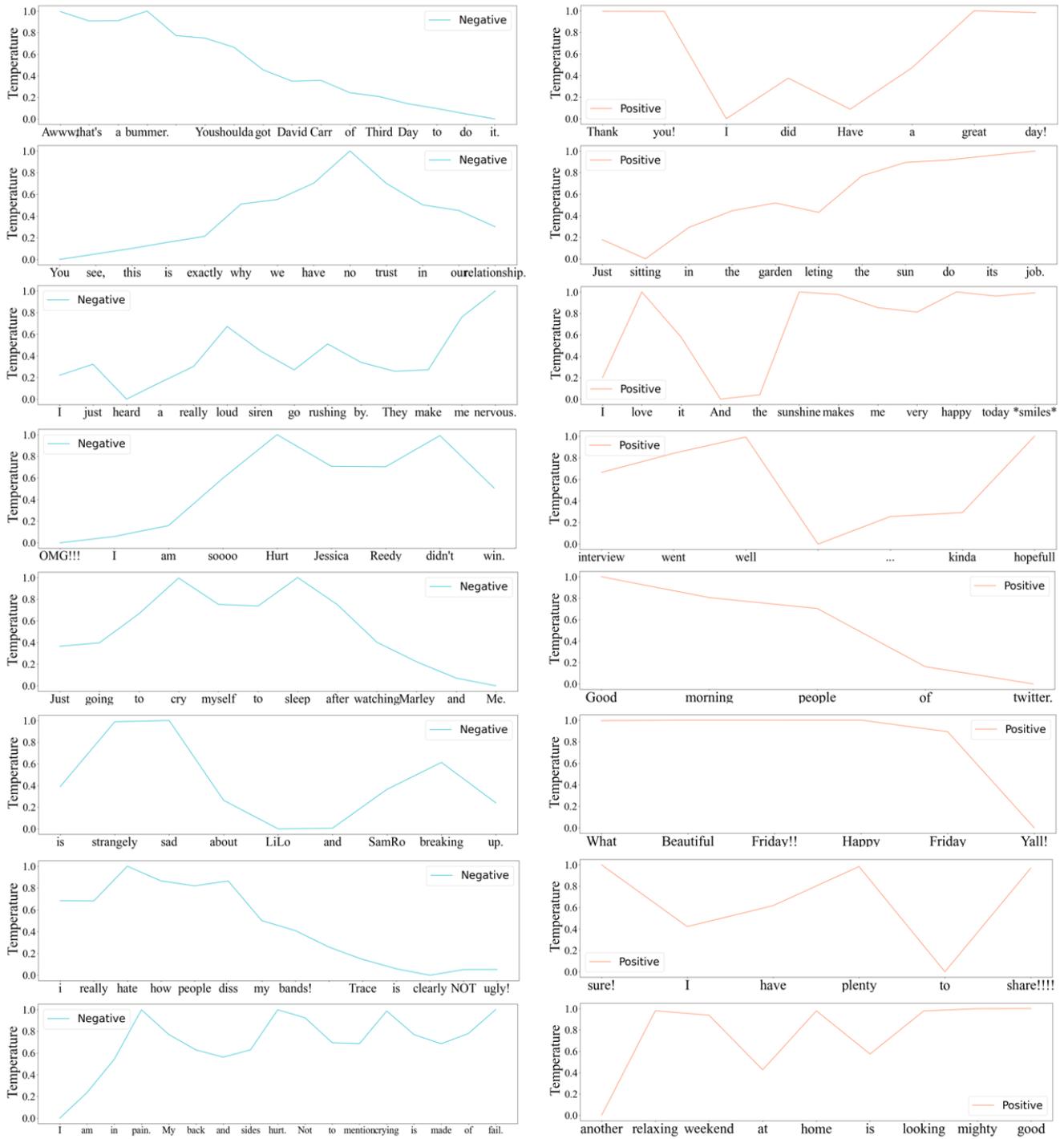


Figure 16. More visualization results from the proposed method for the sentiment analysis task.