Employing Explanations for Effective Image Classification

1st Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

2nd Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

3rd Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

4th Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

5th Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

6th Anonymous
Anonymous
Anonymous
Anonymous
Anonymous

Abstract*—The accurate classification of objects is essential for numerous real-world applications, including autonomous navigation, environmental monitoring, and urban planning. However, standard loss functions for neural network-based image classification may cause the learning process to focus on confounding features that exhibit spurious correlation with the class label of an image. Incorporating ground-truth segmentations into the loss functions can address this issue by ensuring that models learn to recognize and accurately classify images based on the objects they contain, by ensuring a focus on the features of the actual objects. This paper introduces segment overlap loss (SO-Loss), a simple yet novel method to incorporate strong supervision provided by ground-truth segmentations to improve classification accuracy. This loss is based on the idea that the explanation of an image classification, provided by a method such as GradCAM, should be consistent with the corresponding ground-truth segmentation. We evaluate our approach on four image classification datasets for which ground-truth segmentation masks are available, combining the new segment overlap loss with standard cross-entropy loss. We also consider a refinement of our approach where the ground-truth segmentation masks are blurred before the classification model is trained using this hybrid loss. Our findings indicate that significant improvements in accuracy can be obtained by using SO-Loss together with traditional cross-entropy loss, highlighting the importance of loss function selection in classification tasks.

***Index Terms*—image classification, convolutional neural networks, vision transformers, explainability**

I. Introduction

Deep neural networks, particularly convolutional neural networks (CNNs) and vision transformers (ViTs), are able to learn complex patterns and concepts from training data and have become the standard approach to tackle image classification tasks. However, recent studies have highlighted a critical issue: CNNs and ViTs often learn to discriminate image categories based on correlated features rather than features of the salient objects in the images. This can lead to poor generalization, especially in crowded scenes where multiple objects are present. Moreover, relying on proxy or confounding features that do not actually represent the objects of interest can compromise the models' robustness and reliability.

Fig. 1 shows some examples of GradCAM [1] explanations of classifications obtained from a deep image classification model illustrating cases the model's (correct) classifications are based on spurious correlations.

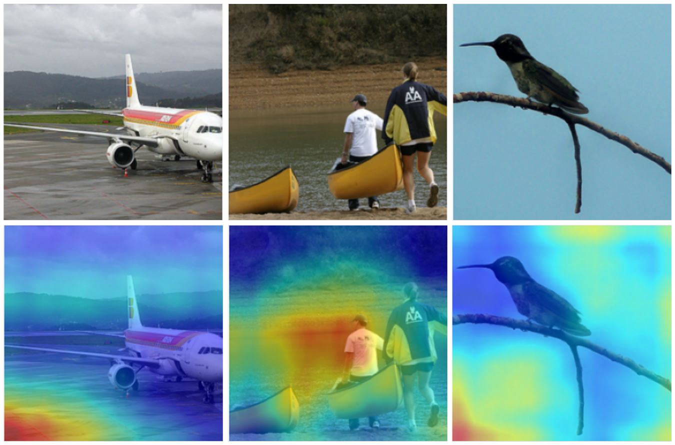

Fig. 1. Examples of GradCAM explanations on common datasets showing spurious correlations. For classifying an airplane (left), the network focuses on the runway; for classifying a boat (center), it focuses on the water; for classifying a bird (right), it focuses on the sky.

Providing explanations for the predictions of deep neural networks has become a key research focus, aiming to enhance the transparency and trust in these models [2]. Explainability methods such as GradCAM help visualize the regions of an image that a model considers important for classification. These methods can help identify spurious correlations that the network may have learned from biased data. In this paper, we investigate whether we can profitably use such methods during *training*: we propose a novel loss function, segment overlap loss (SO-Loss), designed to increase the alignment between GradCAM explanations and the actual object masks, yielding a strongly supervised training approach that leverages segmentation masks to improve image classification.

The objective of our approach is to enhance the reliability and robustness of neural network decision-making processes, thereby increasing the trustworthiness of these models. It is important to acknowledge that the outputs of interpretability methods such as GradCAM can exhibit sensitivity to minor perturbations in the input image (see Fig. 2). Our method explicitly mitigates this issue by incorporating explanations derived from multiple views of an image, ensuring a more stable and consistent result.

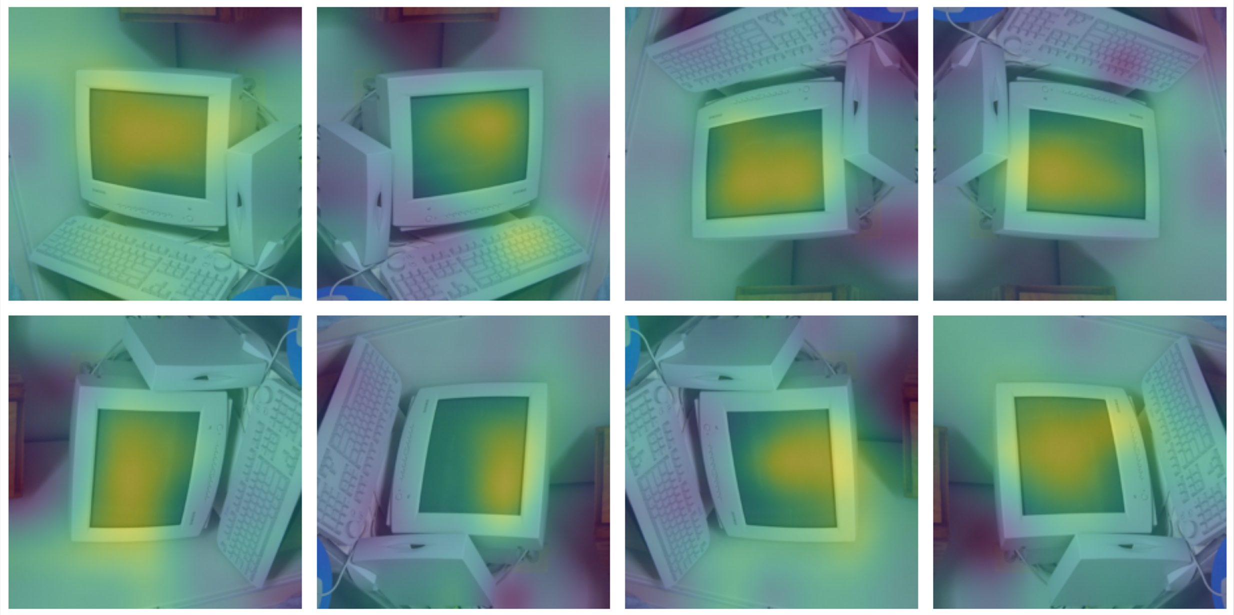

Fig. 2. Changes in the saliency map of the GradCAM explanation resulting from rotations and flips of the input image.

We evaluate the proposed SO-Loss on a variety of image classification datasets, demonstrating its efficacy in enhancing both classification accuracy and the alignment between GradCAM explanations and ground-truth masks. Our principal contributions in this paper are summarized as follows:

1) We introduce a novel loss function, SO-Loss, designed to provide strong supervision in image classification tasks.
2) We demonstrate that our method significantly improves classification accuracy across multiple datasets.
3) We demonstrate that SO-Loss better aligns the GradCAM explanations with the salient objects in the images.
4) We conduct comprehensive ablation studies to investigate the impact of various hyperparameter choices on the performance of our approach.

By enabling the use of explanations in conjunction with ground-truth segmentations when training deep image classification models, our work aims to support the applicability of such models in tasks where correct explanations of classifications are critical.

The rest of the paper is organized as follows: Section II presents the related work, providing an overview of existing methods in saliency-based interpretation and explanation-guided learning. Section III details our proposed methods, including the formulation of the SO-Loss. Section IV describes the datasets, model architectures, and training procedures. Section V presents the results. Section VI conducts an ablative study, analyzing the impact of various hyperparameters on the effectiveness of our approach. Finally, Section VII provides a discussion of the findings, and Section VIII concludes the paper summarizing the contributions.

II. Related Work

A. *Saliency-Based Interpretation Methods*

Saliency-based interpretation methods aim to identify and highlight the regions of an image that are most relevant to a neural network's decision-making process. These methods include Layer-wise Relevance Propagation (LRP) [3], DeepLIFT [4], Integrated Gradients [5], Occlusion Sensitivity [6], and Guided Backpropagation [7]. Among these methods, GradCAM is one of the most popular approaches for visualizing the salient features used for classification by generating a heatmap of the relevant pixels.

GradCAM was introduced by Selvaraju et al. [1] which was built upon earlier work by Zhou et al. [8]. By backpropagating the gradients of a target class through the network, GradCAM produces a coarse localization map of the important regions in the image. Some shortcomings of GradCAM include its dependency on noisy gradients, especially in very deep neural networks. This issue is addressed by variants of GradCAM, such as Score-CAM [9], GradCAM++ [10], and Smooth GradCAM [11].

- **Score-CAM**: Eliminates the dependency on gradients by using the model's output scores to generate saliency maps, leading to more robust and less noisy interpretations.
- **GradCAM++**: Improves upon GradCAM by providing better visual explanations for images containing multiple instances of the target object and producing higher resolution maps.
- **Smooth GradCAM**: Combines GradCAM with SmoothGrad to reduce noise by averaging the saliency maps obtained from multiple noisy versions of the input image.

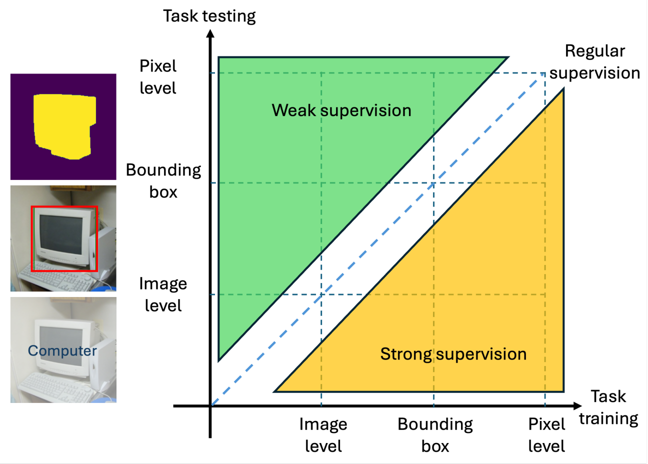

Fig. 3. Weak and Strong supervision in terms of task vs. annotation level.

While GradCAM and its variants have been applied in various domains, including medical imaging [12], autonomous driving [13], and remote sensing [14], these methods still have limitations. One significant limitation of GradCAM is the sensitivity of the saliency maps to input perturbations, and while the variants of GradCAM such as GradCAM++, Score-CAM and Smooth GradCAM tends to be more robust, the improved robustness comes at the cost of increased computation cost. Fig. 2 illustrates the changes in the saliency map of GradCAM interpretations due to rotations and flips in the input image.

B. Explanation-guided learning

Explanation-guided learning leverages explainability methods to improve model performance by incorporating human-understandable explanations into the training process. Work by Jia et al. [15] shows that the correlation between model explanations and expert human annotations can be used for model selection, resulting in models with improved generalizability. This approach enhances trust in the reliability of the model and aids in debugging and refining neural networks.

Strong supervision utilizes detailed annotations, such as segmentation masks, to guide the learning process. Although less prevalent than weakly-supervised learning, strong supervision offers precise guidance that can substantially enhance model performance. Fig. 3 illustrates the distinctions between strong supervision and weak supervision.

Other explainability-integrated learning methods have been explored to improve model performance. Rieger et al. [16] investigated the use of contextual decomposition as the explanation function and achieved improved AUC-ROC and test accuracy on the ISIC [17] and DecoyMNIST [18] datasets. By aligning model predictions with human-interpretable features, these methods enhance both the accuracy and transparency of neural networks. For example, using loss functions that penalize deviations from expected saliency maps can help models learn more robust and interpretable representations.

GradCAM explanations have also been used in explainability-integrated learning. Research by Caforio et al. [19] demonstrates that the GradCAM explanations of a LeNet-5 CNN can be used to improve the accuracy for network intrusion detection. In their approach, they used K-Nearest Neighbour classification of the k-means clustered output of the GradCAM activations to achieve improved F1 scores on three network intrusion datasets. A more related work by Ahmadi et al. [20] introduces the Region of Interest Activation Loss (RIA Loss), which incorporates GradCAM explanations to enhance model training. While Ahmadi et al. uses a similar idea of incorporating GradCAM as a loss, our approach to formulating the loss is significantly different and was developed independently. As readers might note in Section III, we applied a Gaussian blur to the target mask, and derived the GradCAM explanations using the pooled maximum of multiple views. Furthermore, we use cosine similarity as opposed to using the IoU on a binarized thresholded GradCAM as done by Ahmadi et al.

III. Methods

Our approach focuses on defining a differentiable loss that improves the alignment of model explanations with segmentation masks, enhancing the coincidence of the explanation with the segmentation annotations. To this end, we propose Segment Overlap Loss (SO-Loss), which is defined as the cosine similarity between the GradCAM explanations of the penultimate convolutional layer (or encoder layer in the case of ViTs) and the ground-truth segmentations. Section VI presents results from our ablative study, discussing the rationale behind the choices made in defining SO-Loss and our intuition on why these choices improve the performance of neural networks trained with SO-Loss.

Let t be the target label for the classification, X be the input to the network, Y^t be the output logits for class t, $\mathcal{M}^t$ be the annotated pixel mask for class t, and i, j be the spatial coordinates with respect to the input.

As defined by Selvaraju et al. [1], the GradCAM outputs are weighted by the spatially averaged partial derivatives of the predicted output with respect to the activation of convolutional layer channel k. Mathematically, the weights are defined as

$$\alpha_k^t = \frac{1}{Z} \sum_{i,j} \frac{\partial Y^t}{\partial A_{i,j}^k}$$

where $A_{i,j}^k$ is the activation of the convolutional layer for channel k at pixel locations i, j, and Z a normalizing constant.

The GradCAM output is then defined as

$$\text{GradCAM}_{i,j}^t = \text{ReLU}(\sum_k \alpha_k^t A_{i,j}^k),$$

where ReLU is the Rectified Linear Unit function.

We leverage the sensitivity of GradCAM explanations to input perturbations by presenting the neural network with multiple views of the image. In this work, we consider only Euclidean transformations for these multiple views, though other affine transformations could also improve the neural network's performance. We define GC_m^t as the GradCAM explanation for view m applied under transformation $\mathcal{T}_m$.

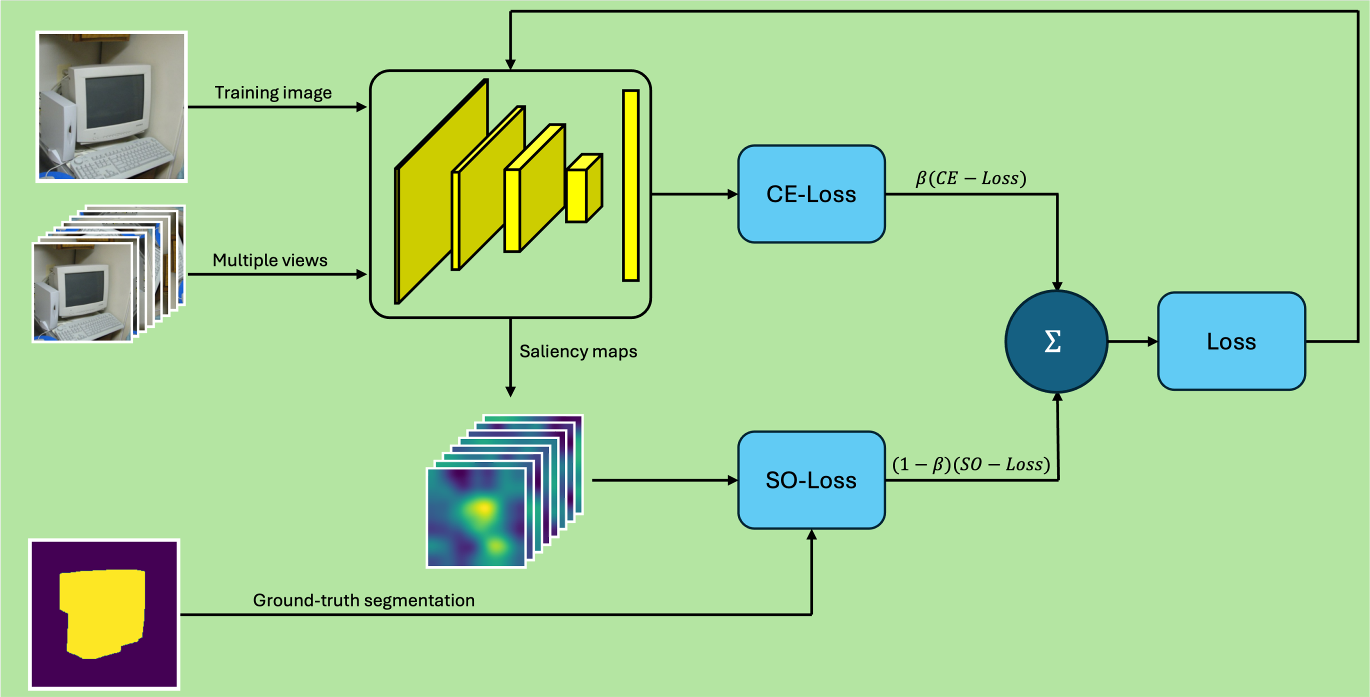

Fig. 4. The flow of the training process.

Specifically, our family of $\mathcal{T}_m$ consists of the eight combinations of horizontal flips, vertical flips, and 90° clockwise and counterclockwise rotations of the input image. Let us define

$$GC^{t,m}_{i,j} = \text{GradCAM}(\mathcal{T}_m(X_{i,j})).$$

Inspired by max-pooling, we calculate the pixel-wise softmax of the GradCAM explanations over these multiple views, returning the maximum GradCAM explanation for each pixel while maintaining the differentiability of the loss function.

$$\text{PooledGC}^t_{i,j} = \text{SoftMax}(\mathcal{T}_0^{-1}(GC^{t,0}_{i,j}), \ldots, \mathcal{T}_m^{-1}(GC^{t,m}_{i,j}))$$

We found that applying a Gaussian blur to the target segmentation mask improves the performance of the network trained with SO-Loss. Our intuition suggests that this may be due to the regularizing effect of blurring the target mask.In our approach, we propose applying a Gaussian blur with a kernel size of 23 roughly $1/10$-th of the input dimension of the neural networks. We thereby define

$$\mathcal{M}^{*t} = \text{GaussianBlur}(\mathcal{M}^t, 23)$$

Finally, we calculate cosine similarity between pooled GradCAM explanations and blurred target segmentation masks:

$$L_{\text{SO}} = \frac{\sum_{i,j}\left(\text{PooledGC}^t_{i,j} \cdot \mathcal{M}^{*t}_{i,j}\right)}{\sqrt{\sum_{i,j}\left(\text{PooledGC}^t_{i,j}\right)^2}\sqrt{\sum_{i,j}\left(\mathcal{M}^{*t}_{i,j}\right)^2}}.$$

We define the hyperparameter β as the mixing factor between the cross-entropy loss and SO-Loss:

$$L_{\text{total}} = \beta L_{\text{CE}} + (1-\beta)L_{\text{SO}}.$$

Algorithm 1 and Fig. 4 provide a high-level view of the training process and the definition of SO-Loss.

IV. Experiments

A. Dataset

For evaluation, we use four different datasets focusing on different tasks to verify the performance of the deep learning models trained using the proposed loss functions under different settings.

The first dataset is the Oxford-IIIT Pet (OPET) dataset. It is a single-label dataset and consists of 7,349 images of cats and dogs from 37 different breeds. Each image is annotated with both a class label and a pixel-level segmentation mask. This dataset was introduced by Parkhi et al. [21] and it provides a rich variety of poses, backgrounds, and lighting conditions, presenting a robust challenge for both classification and segmentation tasks.

The second dataset we employ is the Caltech-UCSD Birds-200-2011 (CUB) dataset [22]. It is a prominent benchmark in fine-grained visual categorization. This dataset comprises 11,788 images of 200 bird species, each accompanied by detailed annotations including species labels, bounding boxes, and segmentations. The CUB-200-2011 dataset is designed to facilitate the development and evaluation of algorithms that can distinguish between visually similar categories, a task that is inherently challenging due to the subtle differences in plumage patterns, colors, and shapes among bird species. The annotations and high intra-class variability make this dataset a challenging resource for training and testing models on fine-grained recognition tasks, which is useful for testing the robustness and versatility of the proposed deep learning models.

Algorithm 1 Training with Segment Overlap Loss (SO-Loss)

```
Require: Neural network model $F$, training dataset $\{(X_n, \mathcal{M}_n, t_n)\}$, number of epochs $E$, learning rate $\eta$, mixing factor $\beta$,
    and Gaussian blur kernel size $o$.
Ensure: Trained model $F$
Initialize model parameters
for epoch $= 1$ to $E$ do
    for each training example $(X, \mathcal{M}, t)$ do
        $Y^t \leftarrow f(X)$                                            ▷ Forward pass to get logits
        Compute GradCAM for multiple views:
        for each transformation $\mathcal{T}_m$ in $\{\mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_m\}$ do
            $X_m \leftarrow \mathcal{T}_m(X)$                             ▷ Transform the input
            $Y^t_m \leftarrow F(X^m)$                                     ▷ Forward pass on transformed input
            $\alpha^{t,m}_k \leftarrow \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^t_m}{\partial A^{k,m}_{i,j}}$
            $\text{GradCAM}^{t,m}_{i,j} \leftarrow \text{ReLU}(\sum_k \alpha^{t,m}_k A^{k,m}_{i,j})$     ▷ GradCAM on each view of the input
        end for
        $\text{PooledGradCAM}^t_{i,j} \leftarrow \text{SoftMax}(\mathcal{T}_0^{-1}(\text{GradCAM}^{t,0}_{i,j}), \dots, \mathcal{T}_m^{-1}(\text{GradCAM}^{t,m}_{i,j}))$     ▷ Compute Pooled GradCAM
        $\mathcal{M}^{*t}_{i,j} \leftarrow \text{GaussianBlur}(\mathcal{M}^t_{i,j}, o)$     ▷ Apply Gaussian blur to target segmentation mask
        $L_{\text{SO}} \leftarrow \frac{\text{PooledGradCAM}^t \cdot \mathcal{M}^{*t}}{\|\text{PooledGradCAM}^t\| \|\mathcal{M}^{*t}\|}$     ▷ Calculate SO Loss using cosine similarity
        $L_{\text{CE}} \leftarrow \text{CrossEntropy}(Y^t, t)$           ▷ Calculate crossentropy loss
        $L_{\text{total}} \leftarrow \beta L_{\text{CE}} + (1 - \beta) L_{\text{SO}}$     ▷ Convex combination of CE Loss and SO Loss
        $\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{total}}$     ▷ Update model parameter using gradient descent
    end for
end for
return Trained model $F$
```

The third dataset is the PASCAL Visual Object Classes (VOC) 2012 segmentation dataset [23], which is widely recognized for its role in advancing object detection and segmentation research. Since we focus on single-label classification tasks, in this dataset, only the images with a single label are used for training and evaluation in our experiments. The VOC dataset contains 20 object classes and comprises of 931 single-label training images and 925 single-label test images.

The last dataset is an aerial imagery dataset of a region in New Zealand.[1] The dataset is also a single-label classification problem with 12 classes. The classes in this dataset are balanced, with 8,000 examples of each class in the training set and 2,000 examples of each class in the test set.

B. Implementation Details

We employ three neural network architectures: ResNet18, ResNet50 [24], and ViT-B_16 [25], each pre-trained on the ImageNet dataset [26].

The training procedure involves a two-stage fine-tuning strategy. Initially, we freeze the pre-trained weights of all layers except the top layers and fine-tune the newly added fully connected layers. Subsequently, we unfreeze the entire network and continue training with a reduced learning rate, allowing for fine-tuning of both high-level and low-level features while preventing large updates that could destabilize the pre-trained weights.

We utilize the Adam optimizer due to its adaptive learning rate properties, which help stabilize training. During the initial fine-tuning stage, the learning rate is set to 1×10^3, and it is reduced to 3×10^4 during the full network training stage. A batch size of 32 is employed to balance computational efficiency and training stability. The loss functions used include cross-entropy loss (CE-Loss), SO-Loss, and a combination of CE-Loss and SO-Loss in different experiment conditions.

As discussed in Section III, we propose two novel loss functions: SO-Loss-Blur and SO-Loss. In our experiments, we evaluate the performance of models trained on each of the selected datasets using the following configurations: i) CE-Loss alone, ii) SO-Loss-Blur alone, iii) SO-Loss alone, iv) a combination of CE-Loss and SO-Loss-Blur, and v) a combination of CE-Loss and SO-Loss. This approach allows us to determine how well the new loss functions perform on their own and when combined with the traditional CE-Loss. The code can be found here[1].

V. Results

We present the classification accuracy for the trained ResNet18, ResNet50, and ViT models on the CUB, OPET, VOC, and Aerial Imagery datasets, each evaluated using all five different configurations specified above.

Table I shows the mean accuracy (in %) and the standard deviations for all the cases. The mean accuracy and the standard deviations reported are over five runs for all the models trained on CUB, OPET, and VOC. Due to the extensive

[1] https://anonymous.4open.science/r/aerial_dataset_for_eval-20E1/

TABLE I
ACCURACY OF THE TRAINED MODELS WITH THE PROPOSED LOSS FUNCTIONS ON DIFFERENT DATASETS

Model	Loss Function	CUB		OPET		VOC		Aerial
		Mean	SD	Mean	SD	Mean	SD	Accuracy
ResNet18	CE-Loss	63.27	0.40	81.52	0.56	75.97	0.85	71.22
	SO-Loss	50.97	0.43	70.87	1.59	41.62	0.86	48.31
	CE-Loss + SO-Loss (β = 0.5)	64.92	0.60	82.68	0.56	75.91	0.69	71.85
	SO-Loss-Blur	61.95	0.43	74.93	0.27	50.89	1.10	49.44
	CE-Loss + SO-Loss-Blur (β = 0.5)	**66.17**	0.28	**84.84**	0.23	**77.51**	0.38	**76.17**
ResNet50	CE-Loss	62.15	0.48	86.20	0.51	86.92	0.43	73.25
	SO-Loss	40.99	0.39	83.04	0.55	83.75	2.12	50.76
	CE-Loss + SO-Loss (β = 0.5)	63.91	0.25	88.21	0.28	88.00	0.60	74.99
	SO-Loss-Blur	42.90	0.30	84.98	0.40	87.59	0.35	49.97
	CE-Loss + SO-Loss-Blur (β = 0.5)	**65.37**	0.33	**90.14**	0.25	**90.30**	0.52	**77.58**
ViT	CE-Loss	56.18	0.56	92.59	0.51	92.81	0.25	72.51
	SO-Loss	63.02	0.44	90.62	0.64	87.73	0.46	61.34
	CE-Loss + SO-Loss (β = 0.5)	59.79	0.38	93.73	0.49	93.01	0.70	73.32
	SO-Loss-Blur	**63.82**	0.36	92.61	0.37	89.12	0.81	61.36
	CE-Loss + SO-Loss-Blur (β = 0.5)	60.94	0.33	**96.05**	0.36	**94.72**	0.43	**74.85**

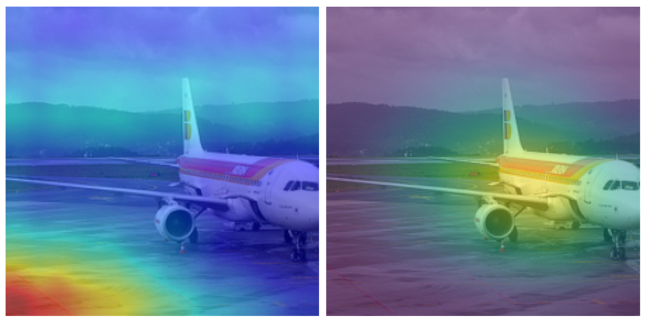

Fig. 5. GradCam explanations from a model trained using CE-Loss (left) vs GradCam explanations from a model trained using a mixture (β = 0.5) of CE-Loss and SO-Loss-Blur.

training time required, the accuracy for the aerial imagery dataset is reported from a single run. From Table I, it is evident that the models trained using a mixture of CE-Loss and SO-Loss-Blur (β = 0.5) outperform those trained with other loss functions in almost all cases. Additionally, models trained with a combination of CE-Loss and SO-Loss ($\beta = 0.5$) show improved performance compared to those trained with CE-Loss alone. However, these trends do not hold for the vision transformer (ViTs) trained on the CUB dataset, likely due to overfitting. Note that to maintain consistency across all models and datasets, the training regime was not altered.

Fig. 5 shows an example of the explanations from ResNet18 trained using CE-Loss and explanations from a ResNet18 trained using a mixture (β = 0.5) of CE-Loss and SO-Loss-Blur. It can be seen that the ResNet18 model trained with the CE-Loss tends to focus more on the confounding features exhibiting spurious correlations. Incorporating ground-truth segmentations into the training process, as facilitated by the SO-Loss-Blur, effectively redirects the model's focus towards the object of interest, thereby enhancing its interpretability and classification robustness.

VI. Ablation Study

To thoroughly evaluate the efficacy of the proposed loss functions, we conduct a series of ablative studies focusing on key hyperparameters. These studies aim to isolate the impact of specific factors on model performance, ensuring a comprehensive understanding of each component's contribution to the overall effectiveness of the training process.

A. *Impact of Mixing Factor*

One critical parameter in our approach is the mixing factor β, which determines the balance between the Cross-Entropy Loss (CE-Loss) and the Segment Overlap Loss (SO-Loss). We perform experiments with different values of β to observe its influence on model performance. The values tested range from 0.25 (25% contribution of CE-Loss) to 1 (solely CE-Loss), with increments of 0.25. As shown in Fig. 6, the performance metrics across various datasets indicate that an optimal mixture (β = 0.5) consistently yields superior accuracy and robustness compared to extreme values.

B. *Number of Views for Explanations*

Another parameter we explore is the number of views used for generating GradCAM explanations in SO-Loss and SO-Loss-Blur. We experiment with different numbers of views to determine the optimal number. We evaluated 1, 2, 4, and 8 views, applying transformations such as horizontal and vertical flips, and 90-degree rotations. Fig. 7 demonstrates that increasing the number of views generally enhances model performance, likely by providing more reliable explanations.

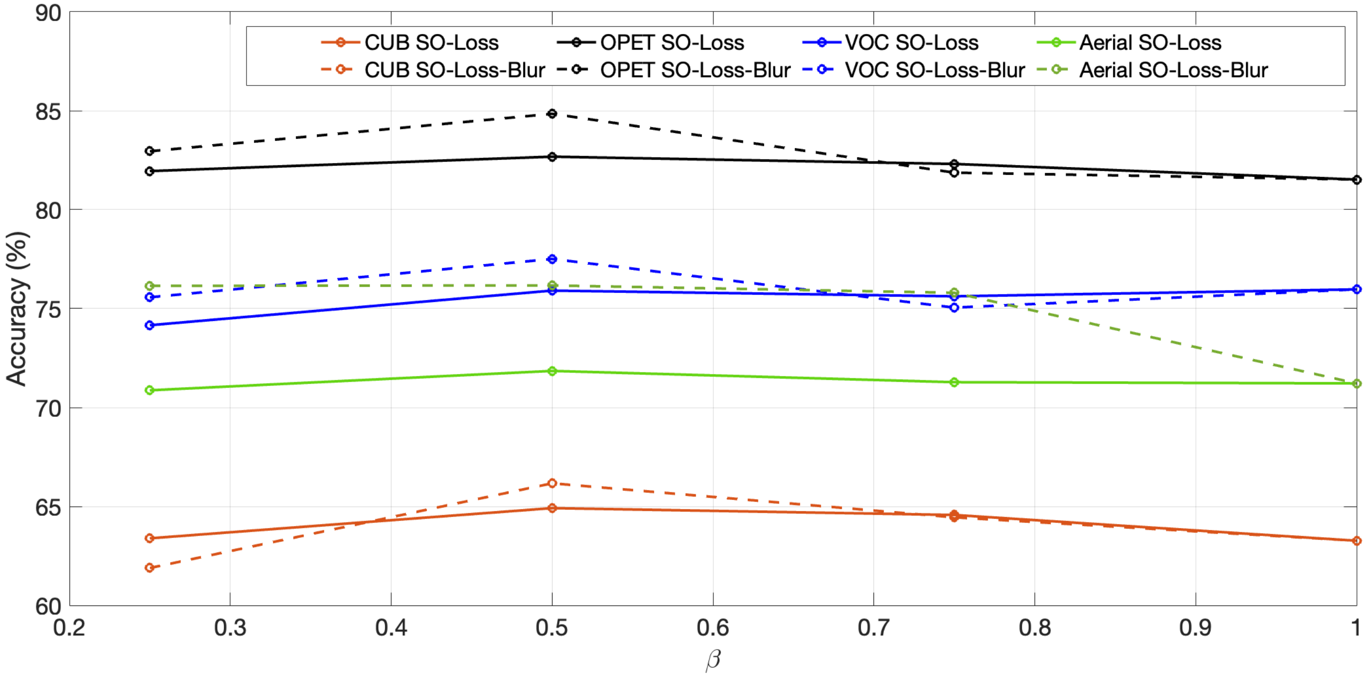

Fig. 6. Test accuracy with differing β for ResNet18; observe that the best test accuracy is obtained when mixing the SO-loss and the cross entropy loss.

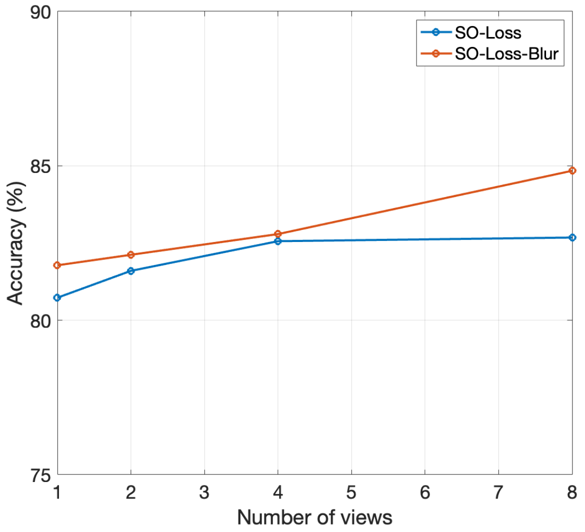

Fig. 7. Test accuracy with differing number of views for the SO-Loss.

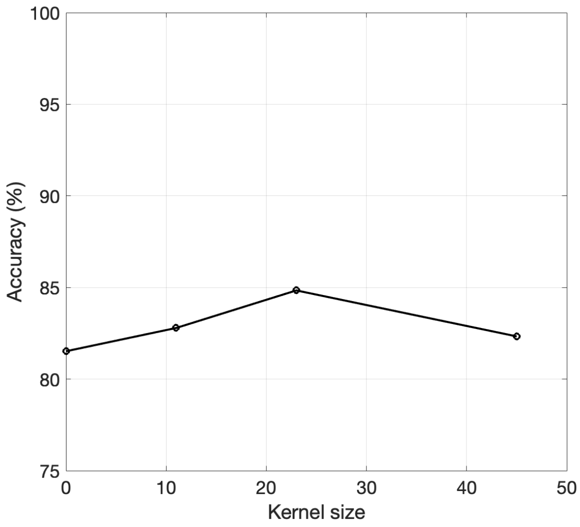

Fig. 8. Test accuracy with different levels of Gaussian blur.

C. Gaussian Blur Kernel Size

For the SO-Loss-Blur configuration, the size of the Gaussian blur kernel applied to the segmentation masks is a significant factor. We examine the impact of different kernel sizes to identify the most effective configuration. More specifically, we considering kernel sizes of 0 (no bluring of the gound-truth segmentation masks), 11, 23, and 45, corresponding to 1/20th, 1/10th and 1/5th of the input dimensions respectively. As shown in Fig. 8, a kernel size of 23 provides the best trade-off between regularization and maintaining segmentation details, thereby improving model accuracy.

VII. Discussion

Our results show that a convex combination of SO-Loss and CE-Loss improves classification accuracy over using only CE-Loss or SO-Loss alone as training loss. In particular, a mixing ratio (β) of 0.5 generally yields the best performance. The only exception was observed on the CUB200 dataset for the VIT-B-16 model, where overfitting to the training data may have occurred; in this case, using only SO-Loss as training loss yields the more accurate classifier. A propensity for overfitting on this dataset is evidenced by the training accuracy reaching 100% when using CE-Loss and a noticeable generalization gap between the training accuracy and the test accuracy. This also suggests that SO-Loss helps mitigate overfitting and enhances generalization, especially in complex datasets or model configurations.

Introducing a Gaussian blur to the target segmentation mask results in a larger accuracy increase. Our intuition suggests that blurring the target mask acts similar to a regularization technique, which helps in smoothing the target distribution, preventing the model from overfitting to exact pixel-wise annotations. On the datasets we evaluated, a kernel size of 23 (approximately 1/10th of the input dimension) provided the best accuracy. This indicates that an optimal level of smoothing can enhance the performance of SO-Loss.

Our method also improves the alignment of GradCAM with the salient objects in the images. By ensuring that the GradCAM explanations are more consistent with the ground truth segmentation masks, the model's predictions are better grounded in causal relationships. The better alignment suggests that the model is focusing on more relevant features during training, which likely contributes to the observed accuracy improvements.

We found that our method performs better with an increased number of views for GradCAM. In our experiments, we utilized Euclidean transformations (horizontal flips, vertical flips, and 90-degree rotations) to generate these multiple views. The improvement in performance indicates that incorporating diverse perspectives helps the model learn more robust and comprehensive features. In future work, we plan to explore the effect of other Euclidean transformations (e.g., scaling) and affine transformations, which may provide additional performance benefits by further diversifying the views. Preliminary results suggest that including a wider range of transformations could further enhance model performance by providing more varied training data. We are also planning to evaluate other CAM-based explainability methods such as GradCAM++, ScoreCAM, and AblationCAM. However, we note that unlike using multiple views of GradCAM, for which the computation can be paralellized at the cost of increased VRAM usage, the serial nature and computational overhead of these method may hinder their viability as replacements.

VIII. Conclusion

In this paper, we introduced two novel loss functions, SO-Loss and SO-Loss-Blur, designed to improve the classification accuracy of a model by exploiting the ground-truth segmentations of the images in a strongly supervised training process. The new loss functions encourage alignment of the explanations of the model with the salient features in the objects. Our experimental results across multiple datasets, including CUB, OPET, VOC, and an aerial imagery dataset, and considering multiple neural network architectures, demonstrate that incorporating these loss functions enhances classification accuracy in most cases. Additional results obtained in ablation studies show the effect of the hyperparameters in our losses, such as the mixing factor β, the number of views for GradCAM explanations, and the Gaussian blur kernel size.

The current work focuses on single-label classification tasks. In future work, we plan to extend the application of SO-Loss and SO-Loss-Blur to multi-label classification tasks, thereby broadening the scope and applicability of our approach. Additionally, we aim to develop methods for generating automated segmentation masks in scenarios where only approximate ground-truth segmentations are available or where segmentations may evolve over time, such as in the case of aerial imagery, where the landcover keeps changing with time. To achieve this, we intend to leverage the model explanations that align with the salient features of objects to extract keypoints within the images. These keypoints can then be used in conjunction with segmentation models, such as the Segment Anything model [27], to enhance the accuracy and quality of ground-truth segmentation masks.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[2] M. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, J. DeNero, M. Finlayson, and S. Reddy, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101. [Online]. Available: https://aclanthology.org/N16-3020

[3] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 2016, pp. 63–71.

[4] J. Li, C. Zhang, J. T. Zhou, H. Fu, S. Xia, and Q. Hu, "Deep-lift: Deep label-specific feature learning for image annotation," *IEEE transactions on Cybernetics*, vol. 52, no. 8, pp. 7732–7741, 2021.

[5] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[6] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.

[7] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6806

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[9] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[10] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[11] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[12] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons &*

Fractals, vol. 140, p. 110190, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077920305865

[13] S. Paniego, E. Shinohara, and J. Cañas, "Autonomous driving in traffic with end-to-end vision-based deep learning," *Neurocomputing*, vol. 594, p. 127874, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224006453

[14] W. Song, S. Dai, J. Wang, D. Huang, A. Liotta, and G. Di Fatta, "Bi-gradient verification for grad-cam towards accurate visual explanation for remote sensing images," in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 473–479.

[15] Y. Jia, E. Frank, B. Pfahringer, A. Bifet, and N. Lim, "Studying and exploiting the relationship between model accuracy and explanation quality," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 2021, pp. 699–714.

[16] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, "Interpretations are useful: penalizing explanations to align neural networks with prior knowledge," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[17] D. A. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. K. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1605.01397, 2016. [Online]. Available: http://arxiv.org/abs/1605.01397

[18] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2662–2670. [Online]. Available: https://doi.org/10.24963/ijcai.2017/371

[19] F. P. Caforio, G. Andresini, G. Vessio, A. Appice, and D. Malerba, *Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems*. Springer International Publishing, 10 2021, pp. 385–400.

[20] R. Ahmadi, M. J. Rajabi, M. Khalooiem, and M. Sabokrou, "Mitigating bias: Enhancing image classification by improving model explanations," in *ACML*, 2023.

[21] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Jul 2011.

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.