Toward Faithful Segmentation Attribution via Benchmarking and Dual-Evidence Fusion

Abu Noman Md Sakib, OFM Riaz Rahman Aranya, Kevin Desai, Zijie Zhang
The University of Texas at San Antonio
{abunomanmd.sakib, ofmriazrahman.aranya, kevin.desai, zijie.zhang}@utsa.edu

Abstract

Attribution maps for semantic segmentation are almost always judged by visual plausibility. Yet looking convincing does not guarantee that the highlighted pixels actually drive the model's prediction, nor that attribution credit stays within the target region. These questions require a dedicated evaluation protocol. We introduce a reproducible benchmark that tests intervention-based faithfulness, off-target leakage, perturbation robustness, and runtime on Pascal VOC and SBD across three pretrained backbones. To further demonstrate the benchmark, we propose Dual-Evidence Attribution (DEA), a lightweight correction that fuses gradient evidence with region-level intervention signals through agreement-weighted fusion. DEA increases emphasis where both sources agree and retains causal support when gradient responses are unstable. Across all completed runs, DEA consistently improves deletion-based faithfulness over gradient-only baselines and preserves strong robustness, at the cost of additional compute from intervention passes. The benchmark exposes a faithfulness–stability tradeoff among attribution families that is entirely hidden under visual evaluation, providing a foundation for principled method selection in segmentation explainability. Code is available at https://github.com/anmspro/DEA

1. Introduction

Post-hoc attribution methods for deep neural networks identify which input regions are causally responsible for a given prediction. In image classification, gradient-based [5, 12, 30, 35, 37, 41, 49] and perturbation-based [16, 17, 29, 45] methods offer well-characterised trade-offs between spatial resolution, computational cost, and faithfulness. Semantic segmentation poses a distinct challenge: an attribution map must explain not only the presence of a target class but whether the model's evidence is correctly localised within a predicted region without assigning importance to spatially disjoint areas. Despite this, segmentation attribution methods remain predominantly gradient-based CAM variants [19, 35, 39, 49], evaluated almost exclusively by visual plausibility. This criterion does not test causal faithfulness. Two questions remain unexamined: does occluding the highest-attributed pixels within the target region reduce the model's confidence in that region, and do attribution maps assign substantial importance to pixels outside the target mask? A method that fails either test may still produce convincing heatmaps, as gradient activations can reflect feature co-occurrence rather than causal evidence [1].

We introduce a reproducible benchmark for segmentation attribution faithfulness, formalising two evaluation axes absent from prior work: *target deletion faithfulness*, measuring the causal dependence of region-level confidence on highest-attributed pixels, and *absolute off-target leakage*, quantifying attribution credit outside the target mask. Together with perturbation robustness and runtime, these axes enable principled multi-criteria comparison of segmentation attribution methods. We demonstrate the benchmark with Dual-Evidence Attribution (DEA), showing that it surfaces faithfulness differences invisible to visual inspection. Our contributions are as follows:

- A reproducible segmentation attribution benchmark comprising intervention-based faithfulness tests, off-target leakage, perturbation robustness, and runtime profiling across three backbones on Pascal VOC and SBD.
- Dual-Evidence Attribution (DEA), a lightweight dual-evidence correction fusing gradient and intervention signals, improving deletion faithfulness over gradient baselines.
- All per-sample outputs and aggregation scripts are released for independent verification.

2. Related Work

Gradient-based attribution for dense prediction. CAM [49] and Grad-CAM [35] produce class-discriminative maps by weighting activations with globally pooled gradients, but discard spatial information through average pooling. Grad-CAM++ [5], Score-CAM [41],

Ablation-CAM [30], and HiResCAM [12] address different limitations of this pooling step. For segmentation, Vinogradova *et al*. [39] restricted the gradient signal to a masked target region (Seg-Grad-CAM), and Hasany *et al*. [19] further improved spatial specificity with elementwise weighting (Seg-XRes-CAM, our EGA baseline). Neither work evaluates causal faithfulness: both validate explanations by visual comparison rather than direct intervention tests.

Perturbation and intervention-based attribution. Occlusion-based methods [45] estimate pixel importance by masking regions and measuring the change in model output, providing a direct causal signal at higher computational cost. RISE [29], meaningful perturbations [17], and extremal perturbations [16] extend this with randomised and optimised masking schemes. These methods are model-agnostic but their evaluation in segmentation remains limited to visual plausibility. Our RIA baseline and the intervention component of DEA adopt the same causal viewpoint within the segmentation evaluation loop.

Broader attribution context. Beyond CAM-style heatmaps, integrated gradients [37], concept-based testing [24], and Shapley-value approximations [28] have shaped widely used desiderata for explanation quality, including sensitivity, implementation invariance, and completeness. As vision models expanded beyond standard CNNs [11, 22, 26, 42, 43], dedicated attention-based and propagation-based explanation methods followed [4, 6, 7, 15, 32, 38, 40], often revealing that attribution behaviour varies substantially across architectures. However, the majority of these methods and their evaluation protocols target image classification, leaving dense prediction tasks without comparable evaluation standards. These lines of work motivate the need for faithfulness evaluation that goes beyond visual plausibility and accounts for the spatial structure specific to segmentation.

Faithfulness evaluation. Adebayo *et al*. [1] showed that many saliency methods produce outputs largely independent of learned weights. Hooker *et al*. [20] proposed ROAR, measuring faithfulness by accuracy degradation after retraining on data with top-attributed pixels removed. Yeh *et al*. [44] introduced infidelity and sensitivity criteria. These works establish that visual plausibility is insufficient [2, 3, 8, 14, 23, 25, 31, 33, 34, 36, 47], but operate in the classification setting. ROAR requires retraining, and none address off-target leakage specific to dense prediction. Our benchmark instantiates inference-time deletion and leakage tests directly within the segmentation loop, requiring no retraining and explicitly accounting for the spatial structure of dense prediction targets.

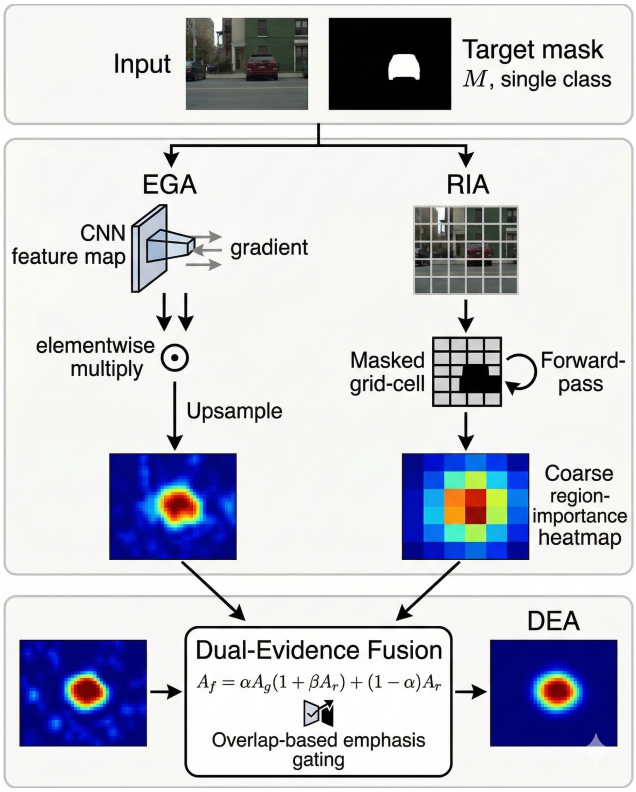

Figure 1. Overview of DEA. Elementwise gradient evidence (EGA) and region intervention evidence (RIA) are combined through multiplicative agreement and residual intervention support.

3. Method

We first define the region-level attribution objective for semantic segmentation, then introduce the dual-evidence correction used in DEA, and finally specify the evaluation metrics. Figure 1 illustrates the pipeline.

3.1. Problem Setup

Given an input image $x \in \mathbb{R}^{3\times H\times W}$ and a pretrained segmentation model f, we study explanations for a target class c and target mask $M \in \{0,1\}^{H\times W}$. Let $z = f(x)$ be per-pixel logits and let $p_c(x)$ denote the softmax probability map for class c. We evaluate evidence at the region level through

$$s_c(x, M) = \frac{\sum_{u,v} M_{uv}\, p_c(x)_{uv}}{\sum_{u,v} M_{uv} + \epsilon}, \tag{1}$$

which is the masked mean class probability inside the target region. An attribution method outputs a heatmap $A \in [0,1]^{H\times W}$, and we test whether high-valued pixels in A are causally important for $s_c(x, M)$.

3.2. Dual-Evidence Attribution

We compare three base attributions that expose complementary behavior. GPA uses gradient pooling at the selected feature layer, EGA uses elementwise gradient-activation products at the same layer, and RIA computes intervention deltas by masking fixed grid regions and measuring the corresponding drop in (1). All maps are min-max normalized to $[0, 1]$.

Let A_g be the EGA map and A_r the RIA map. Our corrected attribution is

$$A_f = \alpha A_g \odot (1 + \beta A_r) + (1 - \alpha) A_r, \tag{2}$$

where (α, β) control the balance between fine gradient structure and intervention support, and $\odot$ denotes elementwise multiplication. The multiplicative term increases weight on pixels where both sources agree, while the residual A_r term keeps coarse but causally supported evidence when gradient responses are unstable.

3.3. Metrics

For a heatmap A, let $S_t(A, M, k)$ be the top-k fraction of pixels within M, and let $S_o(A, M, k)$ be the top-k fraction outside M. Using mean-value occlusion operator $\mathcal{O}(x, S)$ and extending RISE's deletion principle [29] to mask-conditioned scoring, we define target deletion drop

$$\mathrm{TDD} = \frac{s_c(x, M) - s_c(\mathcal{O}(x, S_t), M)}{|s_c(x, M)| + \epsilon}, \tag{3}$$

and off-target deletion drop

$$\mathrm{ODD} = \frac{s_c(x, M) - s_c(\mathcal{O}(x, S_o), M)}{|s_c(x, M)| + \epsilon}. \tag{4}$$

Large TDD indicates that attributed target pixels are causally important. Large |ODD| indicates undesired sensitivity to high-valued pixels outside the target mask. We summarize this tradeoff with

$$\mathrm{LeakAbs} = \frac{|\mathrm{ODD}|}{|\mathrm{TDD}| + \epsilon}. \tag{5}$$

We also report target insertion gain by starting from a mean-value baseline image and reinserting $S_t(A, M, k)$, perturbation robustness as the average correlation between original and perturbed heatmaps (noise, brightness, contrast, blur, horizontal flip), and wall-clock runtime per explanation. We report absolute off-target metrics for primary ranking and keep the signed leakage ratio as a diagnostic statistic.

4. Experimental Setup

4.1. Datasets and Models

We evaluate on Pascal VOC 2012 and SBD [13, 18]. Images and masks are resized to 224×224 and processed with the TorchVision segmentation pipeline. We use pretrained DeepLabV3-ResNet50 [9, 10], FCN-ResNet50 [27], and LRASPP-MobileNetV3 [21]. These choices cover canonical dense-prediction design families used in modern semantic segmentation pipelines [22, 26, 46, 48]. For each image, we choose a single evaluation target class as the most frequent foreground label in the ground-truth mask (excluding background and ignore label), then define M as that class mask. This protocol avoids degenerate empty targets and makes per-sample comparisons consistent across methods.

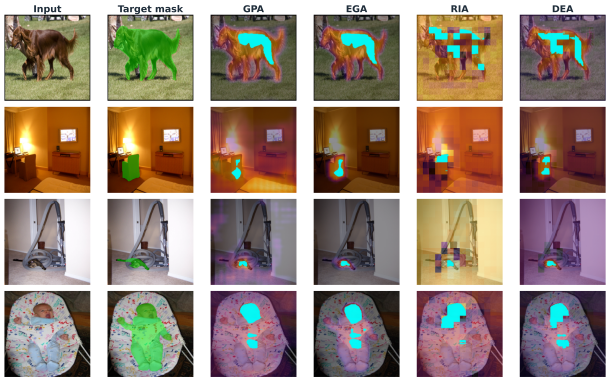

Figure 2. Representative success cases where DEA improves target-region faithfulness while preserving spatial focus.

4.2. Settings

The compared methods are Gradient-Pooled Attribution (GPA, corresponding to Seg-Grad-CAM [39]), Elementwise Gradient Attribution (EGA, corresponding to Seg-XRes-CAM [19]), Region Intervention Attribution (RIA), and DEA. The top-k fraction used in deletion and insertion tests is $k = 0.2$. Region-intervention methods (RIA and DEA) use grid size 14 by default. For DEA, unless noted otherwise, we use $\alpha = 0.65$ and $\beta = 0.35$ from the benchmark implementation. Ablation runs with varied α and β confirm that deletion faithfulness remains above EGA for $\alpha \in [0.5, 0.8]$. Robustness is evaluated with perturbation strength 0.03 over additive noise, brightness shift, contrast change, Gaussian blur, and horizontal flip. Runtime is measured as wall-clock time per explanation call in the same evaluation loop. For SBD, we aggregate all completed runs: six core runs (three backbones, two seeds each), three extra runs, and two DeepLab ablation runs with altered evidence-mixing and grid settings. For VOC, we aggregate six core runs (three backbones, two seeds each).

5. Results

5.1. Quantitative Results

Across completed runs, RIA reaches the strongest deletion faithfulness and the lowest absolute off-target drop, while

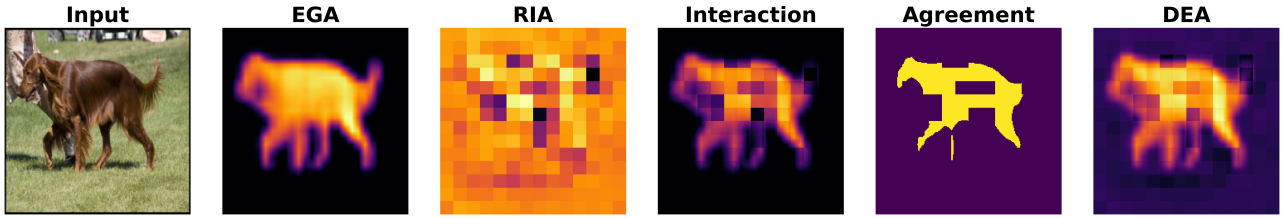

Figure 3. Mechanistic decomposition of DEA: elementwise gradient map, region intervention map, interaction, and corrected output (single case).

EGA remains best in stability and latency. DEA lies between these two ends of the tradeoff curve and consistently improves deletion faithfulness over both gradient baselines in every completed run (over EGA: SBD 11/11, VOC 6/6; over GPA: SBD 9/9, VOC 6/6). Detailed mean and standard deviation values are provided in the appendix in Sec. A.2 and Tab. 1. The tradeoffs are easiest to read relative to EGA, which is the strongest gradient-only baseline in robustness. DEA increases target-region deletion faithfulness and reduces off-target absolute drop versus EGA on both datasets, while remaining slower because intervention passes dominate compute. Relative to RIA, DEA gives up some absolute faithfulness but recovers substantial robustness. This supports a scoped claim: DEA is a practical correction when intervention-aligned faithfulness is needed without fully adopting the least stable intervention baseline. Target insertion gain follows a different ordering and tends to favor broader maps, with GPA highest on both datasets in our aggregates, and full insertion values are reported in the appendix in Sec. A.2 and Tab. 1. We therefore treat insertion as a complementary diagnostic and avoid using it as a single ranking criterion.

5.2. Qualitative Results

Fig. 2 shows the qualitative behavior behind the aggregate metrics and makes the deletion and leakage tradeoff visually explicit. Across success cases, DEA suppresses broad low-confidence context activation that appears in gradient-only maps, while preserving contiguous object structure inside the target region. Compared with RIA, the corrected map typically keeps sharper intra-object detail and avoids the block-like over-smoothing introduced by coarse intervention regions. Fig. 3 explains this behavior at the mechanism level. The interaction term acts as a gate that promotes pixels supported by both evidence streams and attenuates pixels favored by only one stream. In the selected mechanistic cases generated by our pipeline, agreement mass is concentrated on target pixels with near-zero off-target overlap, and the final map is consistently tighter than the raw intervention map while remaining less noisy than the gradient map. Additional failure-mode analysis, including thin-boundary under-coverage and clutter-driven residual leakage, is provided in the appendix in Sec. A.1 and Fig. 4.

6. Discussion

The central finding is not that DEA dominates every axis; it does not. The robust conclusion is that DEA reliably shifts gradient-based attribution toward intervention-validated faithfulness, at the predictable cost of intervention-level compute. This is useful in evaluation-heavy settings where explanation quality is more important than millisecond latency, and less attractive for strict real-time constraints where EGA is still preferable. Two limitations remain important. First, our uncertainty reporting currently uses run-level variability, not formal hypothesis tests, so claims should be interpreted as strong empirical trends rather than definitive significance statements. Second, intervention maps are built from fixed grids, which can miss thin structures and contribute to residual leakage in difficult scenes.

7. Conclusion

We introduced a reproducible benchmark for segmentation attribution and a dual-evidence correction that combines high-resolution gradient structure with intervention support. Across completed SBD and VOC runs, DEA consistently improves deletion-based faithfulness over gradient baselines while preserving high robustness, though it remains much slower than pure gradient methods because intervention passes dominate compute. The empirical picture is intentionally scoped. DEA is best viewed as a correction that moves gradient attribution toward intervention-aligned behavior, not as a universal best method across all axes. In our aggregates, pure intervention attribution still leads on absolute faithfulness metrics, while EGA remains strongest on speed and stability. Future work should add formal significance testing, denser and adaptive intervention schemes for boundary-sensitive regions, and faster region-level evaluation so intervention-grounded attribution becomes practical in larger-scale and lower-latency settings.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for

saliency maps. *Advances in neural information processing systems*, 31, 2018. 1, 2

[2] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8670–8680, 2020. 2

[3] Hamed Behzadi-Khormouji and Jose Oramas. A protocol for evaluating model interpretation methods from visual explanations. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1421–1429, 2023. 2

[4] Itay Benou and Tammy Riklin Raviv. Show and tell: Visually explainable deep neural nets via spatially-aware concept bottleneck models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30063–30072, 2025. 2

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 1

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 2

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, 2021. 2

[8] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892, 2018. 2

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[12] Rachel Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. 1, 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[14] Thomas Fel, David Vigouroux, Remi Cadene, and Thomas Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1565–1575, 2022. 2

[15] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Remi Cadene, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, 2023. 2

[16] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019. 1, 2

[17] 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. 1, 2

[18] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 3

[19] Syed Nouman Hasany, Caroline Petitjean, and Fabrice Mériaudeau. Seg-xres-cam: Explaining spatially local regions in image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3738, 2023. 1, 2, 3

[20] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019. 2

[21] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[22] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2989–2998, 2023. 2, 3

[23] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. Xrai: Better attributions through regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, 2019. 2

[24] 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. PMLR, 2018. 2

[25] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the

human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022. 2

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 2, 3

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2

[29] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 1, 2, 3

[30] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 983–991, 2020. 1, 2

[31] Sukrut Rao, Moritz Bohle, and Bernt Schiele. Towards better understanding attribution methods. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222, 2022. 2

[32] Sukrut Rao, Sweta Mahajan, Moritz Bohle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision (ECCV)*, pages 444–461, 2024. 2

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 2

[34] Andrew Slavin Ross, Michael C. Hughes, and Finale 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*, pages 2662–2670, 2017. 2

[35] 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

[36] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019. 2

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 2

[38] Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision (ECCV)*, pages 123–138, 2024. 2

[39] Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, pages 13943–13944, 2020. 1, 2, 3

[40] Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Learning bottleneck concepts in image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10962–10971, 2023. 2

[41] 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/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1

[42] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2023. 2

[43] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023. 2

[44] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019. 2

[45] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 2

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 3

[47] Wenqi Zhao, Satoshi Oyama, and Masahito Kurihara. Generating natural counterfactual visual explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 5204–5205, 2020. 2

[48] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[49] 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 (CVPR)*, pages 2921–2929, 2016. 1

A. Appendix

A.1. Additional Qualitative Cases

Failure cases are shown in Figure 4. Two recurring patterns appear. First, thin structures and weak boundaries produce under-coverage even when the target object is partially highlighted. Second, highly textured co-occurring regions can attract non-trivial activation when repeatedly reinforced by intervention responses. These failure modes explain the residual off-target emphasis in difficult scenes and motivate future work on adaptive region partitioning.

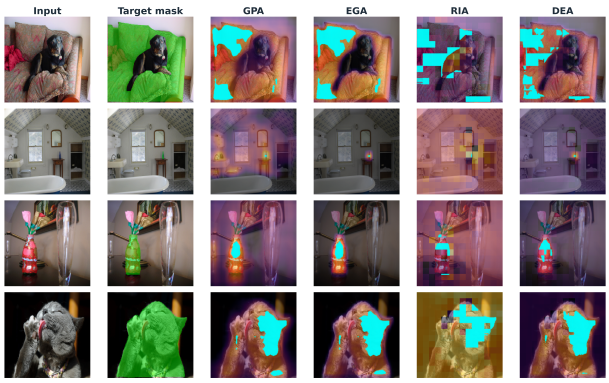

Figure 4. Representative SBD failure cases under the same comparison pipeline used for main-text figures. Residual off-target activation remains in cluttered contexts, and fine boundary detail can be missed on thin target structures.

A.2. Quantitative Details

Table 1 reports method performance as mean and standard deviation across run-level means, so each completed run contributes equally to the aggregate (11 SBD runs, 6 VOC runs). The insertion-gain column clarifies the complementary behaviour noted in the main text: broader maps score higher on insertion even when less selective under deletion and leakage diagnostics.

The VOC EGA outlier in off-target absolute drop (0.982 ± 1.228) is driven by the DeepLabV3 backbone in both seeds, where per-run values reach approximately 2.565, while FCN and LRASPP runs remain much lower (~ 0.265 and ~ 0.116 respectively). This reflects a backbone-specific concentration effect rather than a single-seed anomaly.

Table 1. Completed aggregates reported as mean $\pm$ std across runs. Higher is better for TDD, insertion gain, and stability; lower is better for off-target absolute drop.

Data	Method	TDD ↑	OT abs ↓	Ins. ↑	Stab. ↑
SBD	GPA	$.259_{\pm.021}$	$.748_{\pm.736}$	$.281_{\pm.196}$	$.883_{\pm.006}$
	EGA	$.223_{\pm.027}$	$.268_{\pm.302}$	$.226_{\pm.199}$	$.978_{\pm.004}$
	RIA	$.453_{\pm.021}$	$.177_{\pm.125}$	$.097_{\pm.037}$	$.827_{\pm.010}$
	DEA	$.381_{\pm.030}$	$.235_{\pm.134}$	$.114_{\pm.047}$	$.959_{\pm.009}$
VOC	GPA	$.270_{\pm.044}$	$.281_{\pm.133}$	$.123_{\pm.094}$	$.894_{\pm.004}$
	EGA	$.287_{\pm.048}$	$.982_{\pm1.23}$	$.060_{\pm.009}$	$.978_{\pm.003}$
	RIA	$.522_{\pm.050}$	$.249_{\pm.184}$	$.031_{\pm.013}$	$.821_{\pm.001}$
	DEA	$.449_{\pm.043}$	$.398_{\pm.313}$	$.074_{\pm.017}$	$.961_{\pm.004}$