JOURNAL OF MODERN AND VERSATILE RESEARCH APPROACH (JMVRA)
© 2026 JMVRA | Volume 4, Issue 10, October 2026 | ISSN: 2984-8911 | JMVRA.ORG

Saliency-Based Image Classification: A Survey of Methods, Applications, and Open Challenges

[1]Mehjabin Saiyad, [2]Meet Trivedi, [3]Preet Prajapati,[4]Jigar Dalvadi
[1]Assistant Professor, [2,3] Lecturer, [4]Assistant Professor
[1]Information Technology, [2,3,4] Computer Engineering
Sardar Patel College of Engineering and Technology, Bakrol, Anand, Gujarat, India.

Abstract: Saliency, the identification of the image regions that matter most for a visual decision, has evolved from a hand-crafted model of human attention into a general-purpose signal that is used to supervise, augment, compress, and explain modern image classifiers. This survey reviews recent work from 2024 to 2026 on saliency-based image classification and organises it into six categories: (A) saliency-guided data augmentation and mixup, (B) saliency-guided training objectives, (C) saliency-guided input and token selection for efficient inference, (D) saliency-guided feature mining, (E) saliency as explanation and its evaluation, and (F) robustness, trust, and security. For each category we summarise representative methods, the reported empirical gains, and the assumptions on which these gains rest. Across studies, we observe that (i) human or human-like saliency is a useful training signal when labelled data are scarce, (ii) Vision Transformer (ViT) attention has become the dominant saliency source for efficiency-oriented methods, and (iii) the faithfulness of saliency maps remains contested, which limits their use as clinical or safety evidence. We close with a discussion of open problems, including standardised evaluation, the cost of collecting human saliency, and the interaction between saliency and robustness. All quantitative results are reported as stated by the original authors and are not directly comparable across studies.

Keywords **-** saliency maps; image classification; vision transformers; explainable AI; saliency-guided training; data augmentation; token pruning; medical imaging

I. INTRODUCTION

Visual attention research has long held that biological vision selects a small subset of the scene for detailed processing (Itti et al., 1998). In computer vision, the resulting notion of a saliency map has taken several forms: bottom-up conspicuity maps, gradient-based attribution maps for trained networks (Simonyan et al., 2014), class-activation maps such as Grad-CAM (Selvaraju et al., 2017), and the self-attention maps of Vision Transformers (Caron et al., 2021; Dosovitskiy et al., 2021). Each of these can be used in a classification pipeline either as an output (an explanation) or as an input (a guidance signal).

Between 2024 and 2026 the second use has expanded considerably. Saliency now steers which patches are mixed during augmentation (Kumar et al., 2025; Qin et al., 2024; Wang et al., 2025a), regularises the training loss with human or pseudo-human regions of interest (Boyd et al., 2024; Webster & Czajka, 2025), decides which tokens a ViT or vision-language model should keep (Parikh et al., 2026; Su et al., 2026; Wang et al., 2025b), and supports medical imaging systems that must run on edge hardware (Lavaei et al., 2025). In parallel, a critical literature questions whether saliency maps deserve the trust placed in them (Elisha et al., 2026; Zhang et al., 2025).

Contributions. This paper (1) proposes a taxonomy that separates where saliency enters the classification pipeline (Figure 1 and Figure 2); (2) reviews 23 works published between 2024 and 2026, alongside a small set of foundational references; (3) compares reported results and their experimental conditions; and (4) identifies open challenges and research directions.

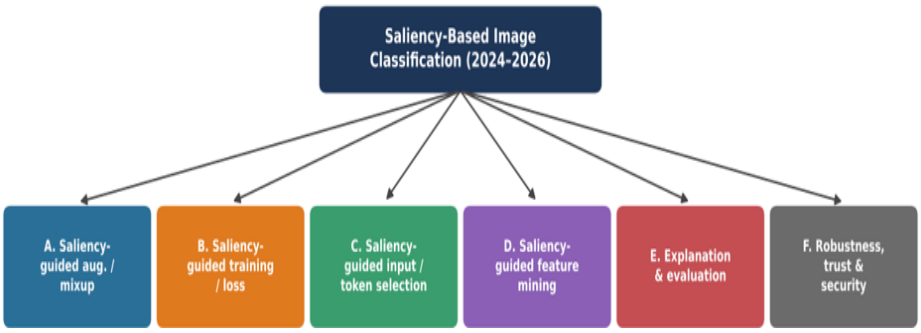


Figure 1. Taxonomy of saliency-based image classification used in this survey.

II. Review Methodology and Scope

We searched arXiv, publisher platforms (IEEE Xplore, Springer, MDPI, SCITEPRESS, PMLR), and Google Scholar-style web searches for combinations of the terms saliency, saliency-guided, attention-guided, token pruning, mixup, saliency map evaluation, and image classification, restricted to publications dated 2024–2026. Inclusion required that (a) saliency, attention-as-saliency, or an attribution map is a central component of an image classification or recognition method, or the paper directly evaluates such maps; and (b) the full text or abstract page could be accessed and the bibliographic details verified. Papers whose pages could not be accessed were excluded. This is a targeted, not exhaustive, review; no formal PRISMA protocol was followed, and the selection may under-represent venues that were not accessible during the search. Several included works are arXiv preprints that may not have been peer reviewed at the time of writing; these are flagged in the reference list by their preprint status.

III. Background

a. Sources of saliency

Saliency maps used in classification fall into four groups (Figure 3). Bottom-up models derive conspicuity from low-level features (Itti et al., 1998). Human-gaze models record or predict where people look; the CYBORG line of work uses such annotations as supervision (Boyd et al., 2024). Post-hoc attribution methods explain a trained network via gradients (Simonyan et al., 2014) or class-weighted activations (Selvaraju et al., 2017). Finally, ViT attention, particularly from self-supervised models such as DINO, yields saliency-like maps without any additional computation (Caron et al., 2021). Yamamoto et al. (2025) report that self-supervised ViTs develop attention that resembles human gaze patterns, which supports the use of ViT attention as a proxy for human saliency. A related caveat is that ViTs can allocate high-norm artefact tokens in background regions, which registers were proposed to correct (Darcet et al., 2024); later work suggests such registers can be obtained without retraining (Jiang et al., 2025). Salient object detection (SOD), surveyed for the RGB-D case by Chen et al. (2024), is an additional upstream source of region proposals.

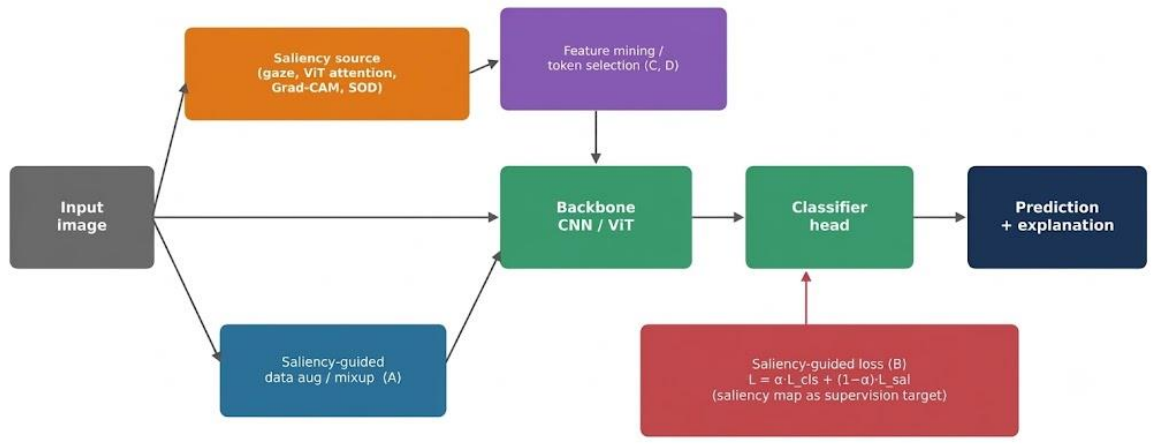


Figure 2. Illustrative comparison of saliency representations

b. Where saliency enters the pipeline

A saliency map can act on the input (augmentation, cropping, token selection), on the objective (a regularisation or auxiliary loss), on the representation (feature selection or re-weighting), or on the output (explanation and audit). Figure 3 shows these insertion points, which define categories A to F in our taxonomy.

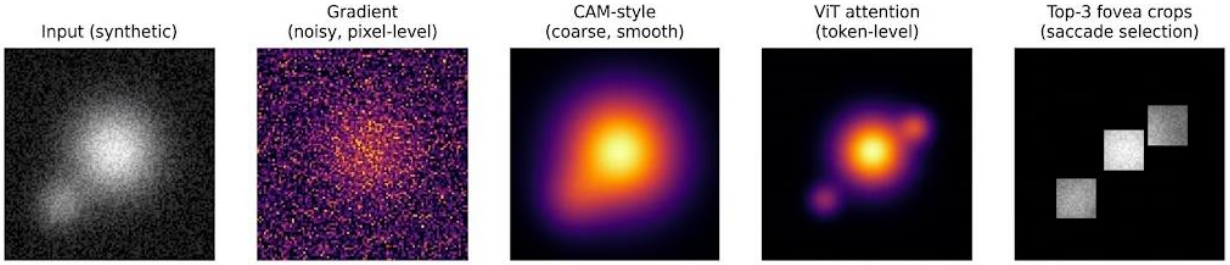


Figure 3. Insertion points of saliency in a classification pipeline.

IV. Saliency-Guided Data Augmentation

Mixing- and erasing-based augmentations are cheap regularisers, but random patch selection may remove the object or paste background, creating label noise. Saliency-guided variants address this by using a saliency map to decide what to cut or mix. RandSaliencyAug applies six saliency- based erasing strategies, combined using a weighted selection scheme, and reports improvements from 84.67% to 91.61% on CIFAR-10 with ResNet-18 and from 76.5% to 78.1% on ImageNet with ResNet-50 (Kumar et al., 2025). SUMix extends mixup by accounting for semantic and uncertainty information in the mixing ratio so that the label reflects the actual content of the mixed image (Qin et al., 2024). TdAttenMix uses top-down, task-driven attention to select the regions to mix, motivated by the observation that bottom-up saliency may not reflect what is relevant for the classification task (Wang et al., 2025a).

In a clinical setting, Choi et al. (2025) use saliency-guided augmentation to improve the generalisation of a YOLOv7-based pressure-ulcer stage classifier. External accuracy on 38 images rose from 75% to 89%, while internal mAP@0.5 fell from 0.96 to 0.85, a trade-off that illustrates how augmentation can improve cross-site behaviour at the expense of in-distribution fit. The small external sample (n = 38) limits the strength of this conclusion.

V. Saliency-Guided Training Objectives

A second family modifies the loss. The CYBORG formulation blends the standard classification loss with a term that penalises disagreement between the model's saliency and human saliency, $L = \alpha \cdot L_cls + (1 - \alpha) \cdot L_sal$ (Boyd et al., 2024). Webster and Czajka (2025) apply this idea to fingerprint presentation attack detection. The best configuration, which used a large training set, minutiae-based areas of interest, DenseNet, and $\alpha = 0.9$, reached 0.938 accuracy versus 0.911 for the baseline; with limited data the best result was 0.885 versus 0.862. Because collecting human saliency is expensive, the authors also test pseudosaliency from an autoencoder, minutiae locations, and image quality, which suggests that saliency need not always come from humans. The authors' own human-saliency set comprises 800 samples annotated by 50 participants.

Lavaei et al. (2025) combine saliency-guided training with quantisation for edge-deployable medical image classification. On an 8-class Kvasir endoscopy task they report 79.05% accuracy versus 77.48% for a standard saliency-guided training baseline (masking ratio 50%, 50 epochs, batch size 128), with sensitivity and specificity of 84.37%.

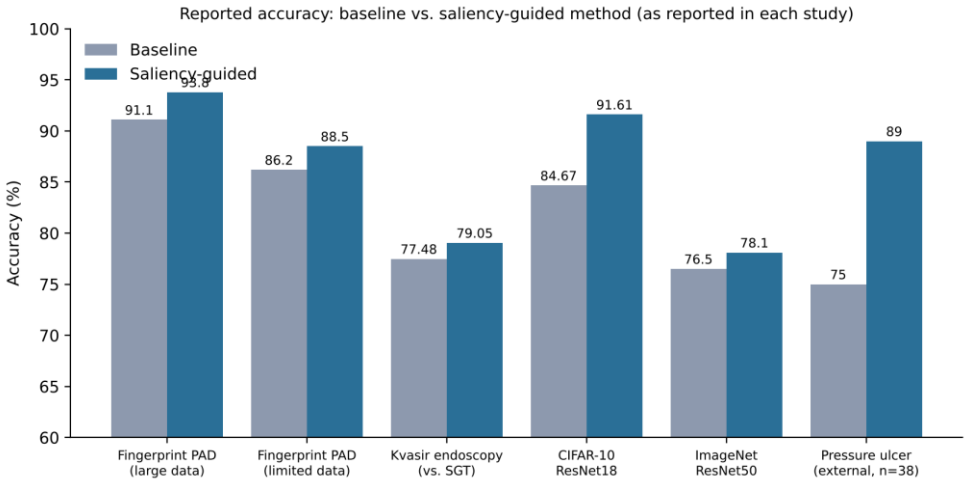


Figure 4. Reported accuracy of baseline versus saliency-guided methods. Values are taken from the individual studies (Choi et al., 2025; Kumar et al., 2025; Lavaei et al., 2025; Webster & Czajka, 2025) and use different datasets, models, and baselines; bars should not be compared across groups.

VI. Saliency-Guided Input and Token Selection

ViTs process every patch token with quadratic attention cost, so many methods use saliency to drop or select tokens. Dallain et al. (2026) propose a saccade-inspired classifier in which ViT attention maps select a small fovea (3×3 or 5×5 tokens) at each of up to ten saccades. The approach preserves most of the full-image accuracy while using fewer than half of the pixels. DINO attention outperformed the GBVS and UNISAL saliency models by about 5%, and the cumulative accuracy over saccades (a sample counts as correct if any saccade predicts correctly) exceeded full-image accuracy, which suggests complementary information across glimpses.

TinyDrop uses a small model to guide token dropping in a larger ViT, with up to 80% FLOPs reduction and negligible accuracy loss on ImageNet-1K (Wang et al., 2025b). Su et al. (2026) argue that [CLS]-attention is a weak pruning criterion and propose a Rényi-entropy-based token metric as an alternative. For large vision-language models, ConsensusDrop fuses visual and cross-modal saliency and retains 99.7%, 99.2%, 95.2%, and 90.5% of LLaVA-1.5-7B performance at 192, 128, 64, and 32 retained tokens (Parikh et al., 2026; Figure 5), while DeSAP uses decoupled similarity to prune tokens in a task-aware manner (Ma et al., 2026). The broader landscape of token compression for compact ViTs on edge devices is surveyed by Nguyen and Cheung (2025).

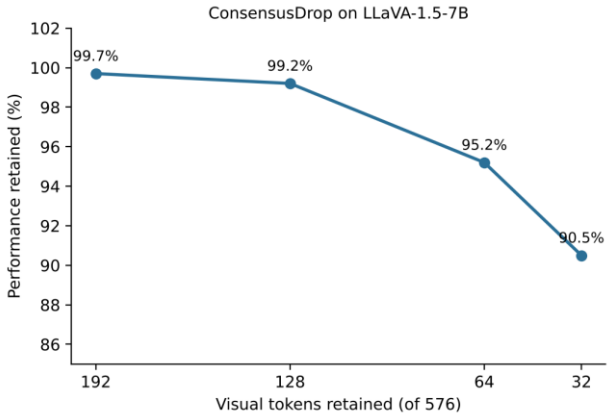


Figure 5. Performance retained by ConsensusDrop at different token budgets (values from Parikh et al., 2026).

VII. Saliency-Guided Feature Mining

Saliency can also select or weight features rather than pixels. Revikumar et al. (2026) propose a saliency-guided feature mining network for multi-scale, fine-grained scene classification in remote sensing image archives, where discriminative cues are small and scale varies widely. For RGB-D indoor scenes, Meena et al. (2024) summarise a volumetric image sequence using a saliency-guided procedure; they report an accuracy

of 0.82 against 0.62 for the best baseline in their comparison. RGB-D salient object detection, which supplies depth-aware saliency for such pipelines, is reviewed by Chen et al. (2024).

VIII. Saliency as Explanation and Its Evaluation

When saliency is the output, its reliability becomes the central question. Zhang et al. (2025) evaluate several saliency methods across metrics and argue that maps can give a false sense of explainability, since methods and metrics frequently disagree. Elisha et al. (2026) propose the RFxG taxonomy, which organises explanation quality along reference (pointwise versus contrastive) and granularity (class versus instance) axes, and introduce four metrics (CCS, CGC, PGS, CGS). Evaluating ten methods across four architectures and three datasets, they find that the best method depends on the chosen property; IIA performed best in their study. These findings reinforce earlier sanity-check concerns about saliency (Adebayo et al., 2018) and imply that conclusions drawn from a single metric should be treated with caution.

IX. Comparative Summary

Table 1. Summary of surveyed works. Results are as reported by the original authors; datasets and baselines differ.

Study	Category	Saliency source	Task / data	Headline result (as reported)
Kumar et al. (2025)	A	Saliency-based erasing	CIFAR-10, ImageNet	84.67→91.61% (ResNet-18); 76.5→78.1% (ResNet-50)
Qin et al. (2024)	A	Semantic/uncertainty-aware mixing	Image classification	Improved mixup (ECCV 2024)
Wang et al. (2025a)	A	Top-down attention	Mixup for classification	Attention-guided mixup
Choi et al. (2025)	A	Saliency-guided augmentation	Pressure-ulcer staging	External acc. 75→89% (n=38)
Webster & Czajka (2025)	B	Human / pseudo saliency	Fingerprint PAD	0.911→0.938 accuracy
Boyd et al. (2024)	B	Human saliency	Classification	Better training with human saliency
Lavaei et al. (2025)	B	Saliency + quantisation	Kvasir (8 classes)	79.05% vs 77.48% (SGT)
Dallain et al. (2026)	C	ViT/DINO attention	Image classification	Most accuracy with <50% pixels
Wang et al. (2025b)	C	Tiny-model guidance	ImageNet-1K	Up to 80% FLOPs reduction
Su et al. (2026)	C	Rényi entropy	ViT token pruning	Alternative to [CLS] attention
Parikh et al. (2026)	C	Visual + cross-modal	LLaVA-1.5-7B	99.7%→90.5% retained (192→32 tokens)
Ma et al. (2026)	C	Decoupled similarity	Large VLMs	Task-aware token pruning
Revikumar et al. (2026)	D	Saliency-guided features	Remote-sensing scenes	Fine-grained scene classification
Meena et al. (2024)	D	Volumetric saliency	RGB-D scenes	0.82 vs 0.62 best baseline
Zhang et al. (2025)	E	Multiple methods	Classifier explanations	Metrics disagree
Elisha et al. (2026)	E	10 methods	3 datasets, 4 architectures	RFxG metrics; IIA best
Najafi et al. (2025)	F	Gradient saliency	Bone-fracture X-ray	98.87% vs 92.76% coverage at 15%

Kulkarni et al. (2026)	F	EigenCAM	CheXpert	Watermark success rate 100%

X. Discussion

Three themes emerge. First, saliency works best as a training or selection signal when the downstream constraint is explicit: limited labels (Webster & Czajka, 2025), limited compute (Dallain et al., 2026; Parikh et al., 2026; Wang et al., 2025b), or limited cross-site generalisation (Choi et al., 2025). Second, ViT attention has largely replaced external saliency models as the guidance source, because it is free at inference and aligns with human gaze (Dallain et al., 2026; Yamamoto et al., 2025), though artefact tokens may need correcting (Darcet et al., 2024; Jiang et al., 2025). Third, the explanatory use of saliency is under scrutiny: methods and metrics disagree (Elisha et al., 2026; Zhang et al., 2025), so saliency-based claims should be accompanied by multiple faithfulness metrics.

Limitations of the literature. Reported gains are measured on different datasets, backbones, and baselines, so no single ranking is justified. Several studies use small external test sets (e.g., n = 38; Choi et al., 2025) or a single clinical task (Najafi et al., 2025). Many of the works surveyed are preprints.

XI. Open Challenges and Future Directions

a. Standardized benchmarks that jointly report accuracy, efficiency, and faithfulness for saliency-guided classifiers.
b. Cheaper human-saliency supervision, including pseudo saliency, eye-tracking proxies, and self-supervised ViT attention (Webster & Czajka, 2025; Yamamoto et al., 2025).
c. Task-aware saliency: bottom-up saliency can differ from task-relevant regions, motivating top-down guidance (Wang et al., 2025a).
d. Joint treatment of robustness, interpretability, and efficiency in safety-critical domains such as medical imaging (Lavaei et al., 2025; Najafi et al., 2025).
e. Extending saliency-guided compression from classifiers to multimodal models (Ma et al., 2026; Parikh et al., 2026) and to edge deployment (Nguyen & Cheung, 2025).

XII. Conclusion

Saliency-based image classification has matured from a collection of attention-inspired heuristics into a set of practical techniques for augmentation, supervision, efficient inference, and auditing. The 2024–2026 literature shows consistent gains in data-scarce and compute-constrained settings, but also exposes weaknesses in the faithfulness and comparability of saliency-based claims. Progress will depend on shared evaluation protocols and on understanding when saliency guidance helps and when it merely reflects the biases of the saliency source.

XIII. Declarations

Funding: none declared. Conflicts of interest: none declared. Figures 1–5 were generated for this survey; Figure 3 uses synthetic data and Figure 4 and Figure 5 re-plot values reported in the cited studies.

Acknowledgement

The author would like to express sincere gratitude to Prof. Jigar Dalvadi, Assistant Professor, SPCET, Bakrol, Anand, for his valuable guidance, continuous support, and constructive suggestions throughout for carry this work. His encouragement and insightful feedback have been instrumental in the successful completion of this survey. The author is also thankful to the management of SPCET, Bakrol, Anand, for providing the necessary facilities, equipment, infrastructure, and institutional support required to carry out this work. The support and resources provided by the institute are gratefully acknowledged. Finally, the author extends sincere thanks to everyone who directly or indirectly contributed to the successful completion of this work.

REFERENCES

[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (Vol. 31, pp. 9505–9515).

[2] Boyd, A., Tinsley, P., Bowyer, K. W., & Czajka, A. (2024). Training better deep learning models using human saliency. arXiv. https://arxiv.org/abs/2410.16190

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650–9660).

[4] Chen, A., Li, X., He, T., Zhou, J., & Chen, D. (2024). Advancing in RGB-D salient object detection: A survey. Applied Sciences, 14(17), Article 8078. https://doi.org/10.3390/app14178078

[5] Choi, J.-W., Rhee, W. L., Han, D.-H., & Kang, M. (2025). Improving clinical generalization of pressure ulcer stage classification through saliency-guided data augmentation. Diagnostics, 15(23), Article 2951. https://doi.org/10.3390/diagnostics15232951

[6] Dallain, M., Rodriguez, L., Perrinet, L. U., & Miramond, B. (2026). A saccade-inspired approach to image classification using Vision Transformer attention maps. arXiv. https://arxiv.org/abs/2603.09613

[7] Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2024). Vision transformers need registers. In The Twelfth International Conference on Learning Representations. https://arxiv.org/abs/2309.16588

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

[9] Elisha, Y., Cohen, S., Barkan, O., & Koenigstein, N. (2026). Rethinking saliency maps: A cognitive human aligned taxonomy and evaluation framework for explanations. In Proceedings of the AAAI Conference on Artificial Intelligence. https://arxiv.org/abs/2511.13081

[10] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259. https://doi.org/10.1109/34.730558

[11] Jiang, N., Dravid, A., Efros, A., & Gandelsman, Y. (2025). Vision transformers don't need trained registers. arXiv. https://arxiv.org/abs/2506.08010

[12] Kulkarni, P., Guo, J., & Huang, H. (2026). X-Mark: Saliency-guided robust dataset ownership verification for medical imaging. arXiv. https://arxiv.org/abs/2602.09284

[13] Kumar, T., Mileo, A., & Bendechache, M. (2025). RandSaliencyAug: Balancing saliency-based data augmentation for enhanced generalization. In Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Vol. 3: VISAPP, pp. 283–290). SCITEPRESS. https://doi.org/10.5220/0013098700003912

[14] Lavaei, M., Beigzad, S., Abadi, Z., & Maleki, A. (2025). Resource-efficient medical image classification for edge devices. arXiv. https://arxiv.org/abs/2512.17515

[15] Ma, K., Xiao, J., Chen, C., Min, G., Zhu, G., Wang, J., & Liao, L. (2026). Decoupled similarity for task-aware token pruning in large vision-language models. arXiv. https://arxiv.org/abs/2604.11240

[16] Meena, P., Kumar, H., & Yadav, S. K. (2024). A volumetric saliency guided image summarization for RGB-D indoor scene classification. arXiv. https://arxiv.org/abs/2401.16227

[17] Najafi, M. H., Pashanejad, M., Norouzi, M., Morsali, M., Soleimani Roudi, S., & Bagheri Shouraki, S. (2025). Secure diagnostics: Adversarial robustness meets clinical interpretability. arXiv. https://arxiv.org/abs/2504.05483

[18] Nguyen, P., & Cheung, N.-M. (2025). Token compression meets compact vision transformers: A survey and comparative evaluation for edge AI. arXiv. https://arxiv.org/abs/2507.09702

[19] Parikh, D., Fan, H., Kannan, R., & Prasanna, V. (2026). ConsensusDrop: Fusing visual and cross-modal saliency for efficient vision language models. arXiv. https://arxiv.org/abs/2602.00946

[20] Qin, H., Jin, X., Zhu, H., Liao, H., El-Yacoubi, M. A., & Gao, X. (2024). SUMix: Mixup with semantic and uncertain information. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), Computer Vision – ECCV 2024 (pp. 70–88). Springer. https://doi.org/10.1007/978-3-031-73021-4_5

[21] Revikumar, D., Byju, A. P., & Bruzzone, L. (2026). Saliency-guided feature mining network for multi-scale fine-grained scene classification in remote sensing image archives. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2025.3631832

[22] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

[23] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Workshop at the International Conference on Learning Representations.

[24] Su, W.-Y., Zhang, R., & Zhang, Z. (2026). Rényi entropy: A new token pruning metric for vision transformers. arXiv. https://arxiv.org/abs/2603.27900

[25] Wang, G., Wang, Q., Huang, B., Chen, S., & John, D. (2025b). TinyDrop: Tiny model guided token dropping for vision transformers. arXiv. https://arxiv.org/abs/2509.03379

[26] Wang, Z., Gu, L., & Lu, F. (2025a). TdAttenMix: Top-down attention guided mixup. arXiv. https://arxiv.org/abs/2501.15409

[27] Webster, S., & Czajka, A. (2025). Saliency-guided training for fingerprint presentation attack detection. arXiv. https://arxiv.org/abs/2505.02176

[28] Yamamoto, T., Akahoshi, H., & Kitazawa, S. (2025). Emergence of human-like attention in self-supervised vision transformers: An eye-tracking study. Neural Networks, 189, Article 107595. https://arxiv.org/abs/2410.22768

[29] Zhang, H., Figueroa, F. T., & Hermanns, H. (2025). Saliency maps give a false sense of explanability to image classifiers: An empirical evaluation across methods and metrics. In V. Nguyen & H.-T. Lin (Eds.), Proceedings of the 16th Asian Conference on Machine Learning (Vol. 260, pp. 479–494). PMLR. https://proceedings.mlr.press/v260/zhang25a.html

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.