Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33518

Semantics-Aware Adaptive Erasing Augmentation for Long-Tail Recognition

Haoran Zhao

School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, Beijing, China.
zhaohaoran@bupt.edu.cn

Abstract: Existing erasing-based data augmentation methods suffer from two key limitations: mask generation is decoupled from image semantics, resulting in random and blind erasing locations; and a uniform erasing intensity is applied across all categories, which exacerbates training signal imbalance under long-tail distributions. To address these issues, this paper proposes the Semantic-aware Category-Adaptive Erasing augmentation method (SCAE): a Grad-CAM-based saliency estimation module is introduced to guide erasing locations via semantic activation distributions, combined with a curriculum learning strategy for progressive hard sample generation; meanwhile, erasing probabilities are adaptively adjusted according to class frequency, and mixup augmentation is incorporated for extreme tail classes. Experiments show that SCAE outperforms existing mainstream methods on CIFAR-100,

Keywords: Object Detection, Comprehensive Performance Evaluation

1. Introduction

Data augmentation is an important technique for mitigating overfitting in deep neural networks and improving generalization, and has been widely applied in scenarios such as autonomous driving [1] and remote sensing image interpretation [2]. Occlusion-based augmentation methods, represented by Random Erasing [3] and GridMask [4], simulate occlusion by randomly masking image regions. However, mask generation is completely decoupled from semantics, which may destroy critical discriminative regions or merely perturb the background. SaliencyMix [5] and KeepAugment [6] incorporate semantic information to some extent, but their exploitation remains limited. In addition, real-world datasets commonly exhibit long-tail distributions [7], and uniform-intensity erasing further degrades the training signal for tail classes, yet existing methods do not adjust augmentation strategies according to class frequency. To address these problems, this paper proposes the SCAE method. The main contributions are as follows: (1) a Grad-CAM [8]-based semantics-aware erasing strategy that guides mask sampling via a semantic activation probability distribution, actively generating occlusion in high-semantics regions; (2) a curriculum learning scheduling mechanism that linearly increases the semantic guidance strength throughout training, enabling progressive hard sample generation; (3) a class-frequency-adaptive erasing probability mechanism combined with a tail-class MixUp strategy to alleviate long-tail training imbalance at the augmentation level; and (4) systematic experiments on three benchmark datasets validating the effectiveness of the proposed method.

DOI: 10.54254/2755-2721/2026.33518

2. Related work

Data augmentation methods can be broadly categorized into three types: geometric transformations, region erasing, and mixing-based augmentation [9]. AutoAugment and RandAugment [10] improve augmentation effectiveness through automated policy search. Cutout, Random Erasing [3], and GridMask [4] simulate occlusion by masking rectangular regions without incorporating semantic information, resulting in mask generation that is independent of image content. Mixing-based augmentation methods such as CutMix [11] and SaliencyMix [5] leverage saliency maps to guide paste locations, but do not actively exploit semantic information during mask sampling. Building on these works, the proposed method introduces semantics-driven mask sampling and curriculum learning scheduling to generate more challenging training samples.

Long-tail recognition methods typically address the problem from three perspectives: re-sampling, re-weighting, and decoupled learning [7]. Class-balanced loss [12] weights per-class losses by the effective number of samples; decoupled learning [13] separates representation learning and classifier training into distinct stages; MiSLAS [14] combines mixup augmentation with label smoothing. Our method approaches the problem from the data augmentation perspective, adaptively adjusting erasing intensity according to class frequency, and can be orthogonally combined with the above methods as a plug-and-play component.

Grad-CAM [8] visualizes the spatial attention regions of a network via gradient-weighted class activation mapping, and has been widely used in model interpretation and augmentation guidance [15]. ScoreCAM [16] further improves the accuracy of activation maps. This paper deeply integrates Grad-CAM saliency maps with erasing augmentation, and dynamically adjusts the semantic guidance direction through curriculum learning, enhancing occlusion robustness while avoiding gradient instability in early training.

3. Method

This paper proposes a Semantic-aware Category-Adaptive Erasing augmentation method (SCAE), which builds upon traditional random erasing and grid masking by integrating a semantic guidance mechanism and a class-frequency-adaptive strategy, addressing two core challenges: insufficient occlusion robustness and sparse training signals for tail classes under long-tail distributions.

3.1. Problem formulation

Let the training set be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times C}$ is the input image and $y_i \in \{1, 2, \ldots, K\}$ is the class label. Let n_k denote the number of samples in class k, satisfying the long-tail distribution assumption $n_1 \geq n_2 \geq \cdots \geq n_K$, where head classes have sufficient samples and tail classes are sparsely represented.

Traditional erasing methods generate a binary mask $M \in \{0, 1\}^{H\times W}$ on image x to produce the augmented image:

$$\tilde{x} = x \odot M \tag{1}$$

where $\odot$ denotes element-wise multiplication. In existing methods, the generation of mask M is completely independent of image semantics and class distribution, resulting in random and blind erasing locations that may destroy critical discriminative regions or only perturb the background, potentially causing excessive information loss for tail classes.

3.2. Semantics-Aware adaptive erasing

DOI: 10.54254/2755-2721/2026.33518

3.2.1. Saliency map generation

To capture the semantic distribution of an image, we introduce a lightweight saliency estimation module based on Gradient-weighted Class Activation Mapping (Grad-CAM). Given input image x and its label y, let $\mathcal{F}^k \in \mathbb{R}^{H' \times W'}$ denote the feature map of the last convolutional layer (where k is the channel index). The importance weight of the k-th channel for class y is:

$$\alpha_k = \frac{1}{H'W'} \sum_i \sum_j \frac{\partial S^y}{\partial \mathcal{F}_{ij}^k} \tag{2}$$

where S^y is the classification score for class y. The semantic saliency map is obtained by a weighted sum over channels followed by ReLU activation:

$$A = \text{ReLU}\left(\sum_k \alpha_k \mathcal{F}^k \right) \tag{3}$$

After upsampling A to the original image size $H \times W$ and normalizing, we obtain the normalized saliency map $\hat{A} \in [0,1]^{H \times W}$, where higher values indicate stronger semantic importance at that location.

3.2.2. Semantics-Guided mask generation

Based on the saliency map $\hat{A}$, we design a semantics-aware mask sampling strategy. The sampling probability distribution for erasing locations is defined as:

$$P(i,j) = \frac{\exp\left(\lambda \cdot \hat{A}(i,j)\right)}{\sum_{i'} \sum_{j'} \exp\left(\lambda \cdot \hat{A}(i',j')\right)} \tag{4}$$

where λ is a temperature coefficient controlling the strength of semantic guidance: when $\lambda > 0$, erasing is biased toward high-saliency regions, forcing the network to exploit secondary discriminative features; when $\lambda < 0$, high-saliency regions are protected, which is suitable for scenarios with sparse training signals; when $\lambda = 0$, the distribution degenerates to uniform random sampling, equivalent to traditional methods.

The center of the erasing rectangle is determined by sampling from $P(i,j)$. Combined with an area ratio uniformly sampled from $[s_l, s_h]$ and an aspect ratio uniformly sampled from $[r_1, r_2]$, the final semantics-aware mask M_{sem} is generated.

3.2.3. Curriculum learning scheduling

To prevent gradient instability caused by overly strong semantic guidance in early training, we introduce a curriculum learning scheduling strategy based on training epochs. Let t denote the current epoch and T the total number of epochs; the semantic guidance strength increases linearly with training progress:

$$\lambda(t) = \lambda_{\max} \cdot \frac{t}{T} \tag{5}$$

In early training, $\lambda(t) \approx 0$ and the model builds foundational representations under random erasing. As training progresses, semantic guidance gradually strengthens, directing the model to extract more robust features on top of the established foundation, thereby achieving progressive hard sample generation.

DOI: 10.54254/2755-2721/2026.33518

3.3. Class-Aware adaptive erasing intensity

3.3.1. Class frequency weights

In long-tail datasets, applying the same erasing intensity to all classes further exacerbates the training signal imbalance between head and tail classes. To address this, we adaptively adjust erasing intensity based on per-class sample counts.

The relative frequency of class k is defined as:

$$f_k = \frac{n_k}{\max_j n_j} \in (0, 1] \tag{6}$$

where f_k closer to 1 indicates a head class, and f_k closer to 0 indicates a tail class.

3.3.2. Adaptive erasing probability

Based on the class relative frequency, the erasing probability for class k is defined as:

$$p_k = p_{\min} + (p_{\max} - p_{\min}) \cdot f_k^{\gamma} \tag{7}$$

where $p_{\min}$ and $p_{\max}$ are the lower and upper bounds of the erasing probability, and $\gamma > 0$ is a scaling coefficient controlling the nonlinearity of the frequency-to-probability mapping. Head classes ($f_k \approx 1$) receive stronger erasing near $p_{\max}$, while tail classes ($f_k \approx 0$) receive weaker erasing near $p_{\min}$, thereby alleviating training imbalance caused by long-tail distributions at the augmentation level.

3.3.3. Mixup augmentation strategy for tail classes

For extreme tail classes ($f_k < \tau$, where τ is a threshold), reducing the erasing probability alone is insufficient to compensate for the scarcity of samples. We further combine MixUp with class-aware erasing: with probability β, MixUp is applied to tail-class samples; with probability $1 - \beta$, weakly-intensity semantics-aware erasing is applied, generating augmented samples as:

$$\tilde{x} = \begin{cases} \mu \cdot x_i + (1 - \mu) \cdot x_j, & \text{with probability } \beta \\ x_i \odot M_{\text{sem}}^{(k)}, & \text{with probability } 1 - \beta \end{cases} \tag{8}$$

where $\mu \sim \text{Beta}(\alpha, \alpha)$, x_j is a sample randomly drawn from the same or a neighboring class, and $M_{\text{sem}}^{(k)}$ is the semantics-aware mask generated for class k.

4. Experiments

4.1. Experimental setup

Experiments are conducted on three benchmark datasets: CIFAR-100 [17], ImageNet-1K [18], and CIFAR-100-LT — a long-tail variant of CIFAR-100 with imbalance ratios $\rho = 100$ and $\rho = 50$. Comparison methods include: Baseline, Random Erasing [3], GridMask [4], CutMix [11], SaliencyMix [5], KeepAugment [6], and MiSLAS [14] (used only in long-tail experiments). All experiments are implemented in PyTorch on 4 NVIDIA RTX 3090 GPUs. For CIFAR-100 and CIFAR-100-LT, ResNet-32 [19] is used as the backbone, trained for 200 epochs with a batch size of 128, an initial learning rate of 0.1, and cosine annealing decay to 10^{-4}. For ImageNet-1K, ResNet-50 [19] is used, trained for 100 epochs with a batch size of 256, with the learning rate decayed by a factor of 0.1 at epochs 30, 60, and 90. Key hyperparameters for CIFAR-100-LT are: $p_{\min} = 0.1$, $p_{\max} = 0.5$, $\gamma = 0.5$, $\tau = 0.1$, $\alpha = 0.4$, $\beta = 0.5$, $\lambda_{\max} = 2.0$.

DOI: 10.54254/2755-2721/2026.33518

4.2. Results and analysis

4.2.1. Standard classification on CIFAR-100

Table 1 reports the Top-1 accuracy of each method on the CIFAR-100 test set.

Table1. Comparison of Top-1 accuracy on CIFAR-100 (ResNet-32, %)

Method	**Top-1 Acc (%)**	**Gain Δ**
Baseline	78.43	—
Random Erasing [3]	79.81	+1.38
GridMask [4]	80.17	+1.74
CutMix [11]	80.64	+2.21
SaliencyMix [5]	81.02	+2.59
KeepAugment [6]	81.35	+2.92
SCAE (Ours)	**82.61**	**+4.18**

As shown in Table 1, SCAE achieves 82.61% accuracy on CIFAR-100, outperforming all comparison methods with a gain of 4.18 percentage points over the baseline and 1.59 percentage points over SaliencyMix, which also incorporates saliency information. This indicates that the semantics-aware erasing strategy contributes more to occlusion robustness than saliency-guided mixing. Compared with KeepAugment, which only protects salient regions, SCAE dynamically adjusts the semantic guidance direction via curriculum learning, erasing high-activation regions in later training stages to encourage the model to utilize a broader set of discriminative features, thereby achieving higher accuracy.

4.2.2. Large-Scale classification on ImageNet-1K

Table 2 presents the performance comparison on the ImageNet-1K validation set.

Table2. Comparison of data augmentation methods on ImageNet-1K (ResNet-50, %)

Method	**Top-1 Acc (%)**	**Top-5 Acc (%)**	**Extra Overhead**
Baseline	76.13	92.86	—
Random Erasing [3]	77.02	93.41	Negligible
GridMask [4]	77.28	93.57	Negligible
CutMix [11]	77.91	93.88	Negligible
SaliencyMix [5]	78.12	94.03	∼5%
KeepAugment [6]	78.34	94.21	∼6%
SCAE (Ours)	**78.87**	**94.52**	∼8%

On ImageNet-1K, SCAE achieves a Top-1 accuracy of 78.87% and a Top-5 accuracy of 94.52%, surpassing all comparison methods. SCAE introduces approximately 8% additional training time

DOI: 10.54254/2755-2721/2026.33518

overhead; through a saliency map caching and reuse mechanism, this overhead is reduced to a level that remains proportionate to the observed performance gains.

4.2.3. Long-Tail recognition results

Table 3 reports the performance of each method on CIFAR-100-LT under different imbalance ratios.

Table3. Comparison of Top-1 accuracy on CIFAR-100-LT (ResNet-32, %)

Method	$\rho = 100$	$\rho = 50$
Baseline	38.32	43.87
Random Erasing [3]	39.14	44.61
GridMask [4]	39.53	45.02
CutMix [11]	40.28	45.79
MiSLAS [14]	41.15	47.02
SCAE w/o Class-Aware	40.87	46.38
SCAE (Ours)	**43.26**	**48.91**

On the long-tail recognition task, SCAE achieves 43.26% and 48.91% accuracy under the $\rho = 100$ and $\rho = 50$ settings, respectively, outperforming all comparison methods. Compared with the ablated variant without the class-aware module (SCAE w/o Class-Aware), the full SCAE achieves a gain of 2.39 percentage points under $\rho = 100$, indicating the contribution of the class-frequency-adaptive erasing strategy in extremely imbalanced scenarios. MiSLAS, a method designed for long-tail recognition, performs competitively under standard settings; however, SCAE achieves higher accuracy under both imbalance ratios by integrating semantics-aware erasing with the class-adaptive strategy.

4.2.4. Ablation study

To quantitatively analyze the contribution of each module, we conduct an ablation study on CIFAR-100; results are shown in Table 4.

Table4. Ablation study results on CIFAR-100 (ResNet-32, %)

Sem.-Aware Erasing	**Curriculum Sched.**	**Class-Aware**	**Top-1 Acc (%)**
			78.43
✓			80.89
✓	✓		81.74
✓		✓	81.52
✓	✓	✓	**82.61**

The ablation results indicate that each component contributes positively to the final performance. Semantics-aware erasing alone yields a gain of 2.46 percentage points over the baseline, representing the largest individual contribution among the three modules. The addition of curriculum learning

DOI: 10.54254/2755-2721/2026.33518

scheduling provides a further gain of 0.85 percentage points, suggesting that a gradual increase in semantic guidance strength benefits training stability in early stages. The class-aware strategy contributes a gain of 0.63 percentage points on the standard balanced dataset, showing that adaptive erasing intensity remains useful even without class imbalance. When all three modules are combined, the model achieves the highest accuracy, confirming that the components function in a complementary manner.

5. Conclusion

This paper proposes SCAE, a Semantic-aware Category-Adaptive Erasing augmentation method designed to address two limitations of existing occlusion-based augmentation methods: semantics-agnostic mask generation and fixed augmentation intensity under long-tail distributions. SCAE incorporates a lightweight Grad-CAM-based saliency estimation module that guides erasing location sampling via a semantic activation probability distribution, directing occlusion toward high-saliency regions so that the model is encouraged to rely on a broader set of discriminative features. A curriculum learning scheduling mechanism is further introduced, in which semantic guidance strength increases linearly with training progress, producing a gradual transition from random perturbation to semantically guided sample generation. For class adaptation, erasing probabilities are allocated according to per-class sample frequency, and a MixUp strategy is applied to tail classes to reduce training signal imbalance at the augmentation level. Experiments on CIFAR-100, ImageNet-1K, and CIFAR-100-LT show that SCAE outperforms existing comparison methods on both standard classification and long-tail recognition tasks, with an additional training overhead of approximately 8%. Ablation studies confirm the individual contribution and complementarity of each module.

References

[1] Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386. https://doi.org/10.1002/rob.21918

[2] Tuia, D., Persello, C., & Bruzzone, L. (2022). Domain adaptation for aerial image analysis: A survey of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, *10*(2), 107–135. https://doi.org/10.1109/MGRS.2021.3135379

[3] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(07), 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000

[4] Chen, P. (2020). GridMask data augmentation. *arXiv preprint arXiv:2001.04086*. https://arxiv.org/abs/2001.04086

[5] Uddin, A. F. M. S., Monira, M. S., Shin, W., Chung, T., & Bae, S.-H. (2021). SaliencyMix: A saliency guided data augmentation strategy for better regularization. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=XSYAVzbR3Bl

[6] Gong, C., Wang, D., Li, M., Chandra, V., & Liu, Q. (2021). KeepAugment: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1055–1064). https://doi.org/10.1109/CVPR46437.2021.00111

[7] Zhang, Y., Kang, B., Hooi, B., Yan, S., & Feng, J. (2023). Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10795–10816. https://doi.org/10.1109/TPAMI.2022.3236977

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[9] Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*. https://arxiv.org/abs/2204.08610

[10] Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, *33*, 18613–18624.

[11] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Training procedure regularizes and localizes features by cutting and mixing training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6023–6032). https://doi.org/10.1109/ICCV.2019.00612

DOI: 10.54254/2755-2721/2026.33518

[12] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9268–9277). https://doi.org/10.1109/CVPR.2019.00949

[13] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., & Kalantidis, Y. (2020). Decoupling representation and classifier for long-tailed recognition. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=r1gRTCVFvB

[14] Zhong, Z., Cui, J., Liu, S., & Jia, J. (2021). Improving calibration for long-tail recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16489–16498). https://doi.org/10.1109/CVPR46437.2021.01622

[15] Kim, J.-H., Choo, W., & Song, H. O. (2021). Flexible sample generation for long-tailed recognition with semantic-aware augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8954–8963). https://doi.org/10.1109/ICCV48922.2021.00883

[16] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 24–25). https://doi.org/10.1109/CVPRW50498.2020.00020

[17] Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* (Technical Report). University of Toronto.

[18] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

[19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90