RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 08, Issue 09 September

Check for updates

http://dx.doi.org/10.47392/IRJASH.2026.040

ResNet-Integrated Gradient Guided Multi-Scale Network for Camouflaged Object Detection

Jogeswar Behera[1], Kushal Minz[2], Dibyajyoti Sethi[3], Priyanka Shit[4], Jashasmita Pal[5]

[1,2,3]Post Graduate (MCA), Radhakrishna Institute of Technology and Engineering, Bhubaneswar, Odisha, India; affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, India.

[4,5]Assistant Professor, Radhakrishna Institute of Technology and Engineering, Bhubaneswar, Odisha, India; Biju Patnaik University of Technology, Rourkela, Odisha, India.

Emails: *jogeswarbehera556@gmail.com[1], kushalminz04@gmail.com[2], dibyajyotisethi12@gmail.com[3], priyanka_cse@riteindia.edu.in[4], jashasmita_cse@riteindia.edu.in[5]*

Article history

Received: 11 August 2026
Accepted: 05 September 2026
Published: 30 September 2026

Keywords:
Camouflaged Object Detection (COD); Computer Vision; Deep Learning; Feature Fusion; Gradient Guidance.

Abstract

Camouflaged Object Detection (COD) is a challenging computer vision task that aims to identify objects that visually blend into their surrounding environment. Traditional object detection techniques often struggle with camouflage because the foreground and background may share similar color, texture, shape, and local contrast. This paper presents a ResNet-Integrated Gradient Guided Multi-Scale Network (RIGGMNet) that combines deep residual feature extraction, gradient-guided boundary enhancement, and multi-scale feature fusion. ResNet is used as the backbone to obtain progressively richer semantic representations, while a gradient-guided module emphasizes subtle boundaries and texture transitions. Multi-scale aggregation then combines fine spatial details with high-level contextual information so that objects of different sizes and complexities can be segmented more reliably. The network is designed as an end-to-end segmentation framework that produces a dense binary map of the camouflaged region. Architecturally, the framework is compared against representative COD baselines to situate its design choices within the broader literature, and it is positioned as a practical approach for wildlife monitoring, surveillance, medical imaging, autonomous systems, and environmental research. Consistent with the source research abstract, exact numerical benchmark scores are left to a reproducible experimental run rather than being invented for this report.

1. Introduction

Camouflaged Object Detection is concerned with locating and segmenting objects that are deliberately or naturally difficult to distinguish from their surroundings. Unlike conventional detection, where the target often has strong contrast or a distinctive appearance, a camouflaged object may share the same colour distribution, texture statistics, and local structure as the background. Consequently, a detector must reason about weak cues rather than depend only on obvious edges or objectness. Fan et al. introduced COD as a distinct research problem and released COD10K, a benchmark containing 10,000 images across diverse natural scenes and object categories (Fan et al., 2020). Deep convolutional networks have substantially improved dense prediction because

they learn hierarchical representations from low-level edges to high-level semantics. Residual learning is especially useful for deep architectures because shortcut connections make optimisation of substantially deeper networks easier (He et al., 2016). However, deep semantic features alone can lose the fine boundary information needed for camouflage, where the foreground-background transition may be extremely weak. This gap motivates the design explored in this paper. Rather than relying solely on a deep backbone, the proposed network explicitly injects gradient-derived boundary cues alongside semantic features, and it fuses information from multiple resolutions so that both coarse object location and fine object outline are preserved through to the final prediction. The remainder of this paper is organised as follows: Section 1 introduces the system architecture and objectives, Section 2 describes the proposed method, dataset, workflow, and evaluation protocol in detail, Section 3 discusses results and situates the design relative to existing work, and the concluding sections summarise contributions and outline future directions [1].

1.1. System Architecture

Let I ∈ ℝ^(H×W×3) denote an input image and Y ∈ {0,1} ^(H×W) its binary camouflage mask. The objective is to learn a function F (I; θ) that estimates the pixel-wise probability of the camouflaged object. A useful COD model must preserve spatial precision while simultaneously learning global context, object structure, and weak boundary evidence. Three properties therefore guide the architectural choices made in this work. First, the backbone must supply a rich hierarchy of features so that both texture-level and semantic-level cues are available. Second, boundary information must be reinforced explicitly, since it is the property most often lost as network depth increases. Third, features extracted at different resolutions must be combined in a way that keeps fine spatial detail from being diluted by coarse semantic information, and vice versa.

1.2. Machine Learning Integration

The proposed work has four objectives: (i) extract robust semantic features with a ResNet backbone; (ii) introduce gradient guidance to make subtle object boundaries more discriminative; (iii) fuse features from multiple resolutions to preserve both fine-grained and contextual information; and (iv) evaluate the resulting segmentation framework using precision, recall, F-measure, and mean absolute error on benchmark COD datasets. Machine learning is integrated at each stage of the pipeline rather than only at the final classification step. The residual backbone is pretrained and fine-tuned end-to-end, the gradient-guided module is trained jointly with the segmentation objective so that boundary cues remain complementary to semantic features, and the fusion decoder is optimised with a composite loss so that both region-level accuracy and boundary sharpness are rewarded during training.

1.3. Related Work

Early COD research established strong baselines for the problem. SINet formulated detection as a coarse search-and-identify process built on a ResNet-50 backbone and demonstrated that a two-stage refinement strategy could substantially improve segmentation of concealed objects on the COD10K benchmark (Fan et al., 2020). PraNet, originally proposed for polyp segmentation but widely adopted in the COD literature, introduced a parallel partial decoder together with reverse-attention modules that progressively mine boundary cues from high-level features (Fan et al., 2020). Subsequent methods focused explicitly on the difficulty of localising a camouflaged object before segmenting it precisely. PFNet introduced a distraction-mining module that separates the positioning of the likely object region from the refinement of its boundary, reducing confusion caused by background regions with similar appearance (Mei et al., 2021). ZoomNet addressed the scale ambiguity inherent in camouflage by processing a mixed-scale triplet of inputs and enforcing scale consistency across the resulting predictions (Pang et al., 2022). A broader review of deep-learning approaches to COD, including common architectural patterns and open challenges, is provided by Liu et al. (2022). RIGGMNet differs from these baselines in how it combines the two ingredients that the literature treats separately in most prior work: explicit gradient-based boundary guidance and multi-scale fusion. Where SINet and PFNet rely on staged search/positioning strategies and PraNet relies on reverse attention, the proposed network injects a dedicated gradient-guided pathway alongside the residual backbone from the outset, so that boundary evidence and semantic

evidence are fused together rather than being recovered late in the decoder [2].

2. Method

The proposed ResNet-Integrated Gradient Guided Multi-Scale Network (RIGGMNet) is organised into three principal components: a residual feature extractor, a gradient-guided module, and a multi-scale feature fusion decoder. The architecture follows the central idea in the supplied abstract: semantic representation is obtained from ResNet, boundary information is strengthened through gradient guidance, and features from different depths are aggregated before final segmentation. The following subsections describe the network design, the target datasets, the implementation workflow, and the evaluation protocol in turn.

2.1. Network Architecture

The residual feature extractor follows the standard ResNet stage structure, producing feature maps C1 through C5 at progressively coarser resolutions and richer semantic content. Shortcut connections allow the network to be trained to a substantial depth without the optimisation difficulties associated with plain deep convolutional stacks (He et al., 2016). These multi-stage features form the semantic backbone on which the remaining modules operate. In parallel, the gradient-guided module computes image gradient information and propagates it through a lightweight convolutional pathway so that boundary and texture-transition cues are available at matching resolutions.

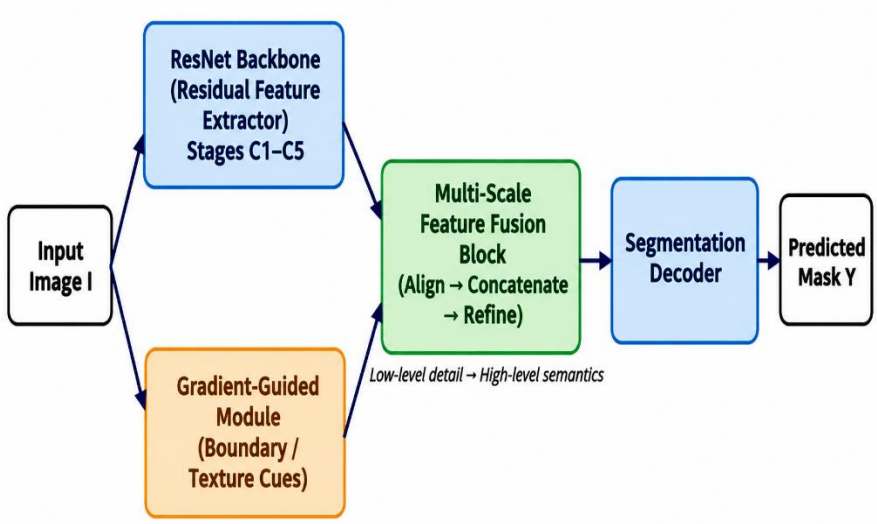


Figure 1 Overall architecture of the proposed RIGGMNet, showing the residual backbone, the gradient-guided module, and the multi-scale fusion decoder

Because camouflaged boundaries are often only weakly visible in raw pixel intensity, this pathway is designed to amplify local contrast changes that the semantic backbone alone tends to smooth away. The two pathways are then combined in the multi-scale feature fusion block, which aligns feature resolutions through upsampling and convolution, concatenates the aligned semantic and boundary features, and applies convolutional refinement before passing the result to the decoder. Figure 1 illustrates the overall data flow of the network, from the input image through the two parallel feature pathways, into the fusion block, and finally to the predicted mask [3].

2.2. Dataset and Materials

The proposed framework targets standard benchmark COD datasets, namely CAMO, CHAMELEON, and COD10K. COD10K is particularly relevant because the original benchmark study describes 10,000 images collected across diverse natural scenes and more than 78 object categories, with dense annotations suitable for object-level tasks (Fan et al., 2020). CAMO contributes a mixture of naturally and artificially camouflaged objects, while CHAMELEON provides a small, high-resolution evaluation-only set that is commonly used to test generalisation. Table 1 summarises the characteristics of these three datasets as commonly reported in the COD literature; exact figures should be verified against

the current release of each benchmark before being quoted in a final submission.

Table 1 Summary of benchmark camouflaged object detection datasets

Dataset	Images	Categories	Characteristic
CAMO	~1,250	8	Naturally&artificially camouflaged objects
CHAMELEON	~76	—	Small, high-resolution test-only set
COD10K	~10,000	78	Large-scale, hierarchical, dense annotations

2.3. Implementation Workflow

The implementation workflow begins with image resizing and normalisation, followed by ResNet feature extraction. Gradient maps are computed in parallel and injected into the feature pathway. The multi-scale fusion block aligns feature resolutions, concatenates the representations, and applies convolutional refinement. The decoder then predicts a dense mask, which is compared against the ground-truth annotation during training and against held-out test images during evaluation. Figure 2 summarises this workflow as four sequential stages [4].

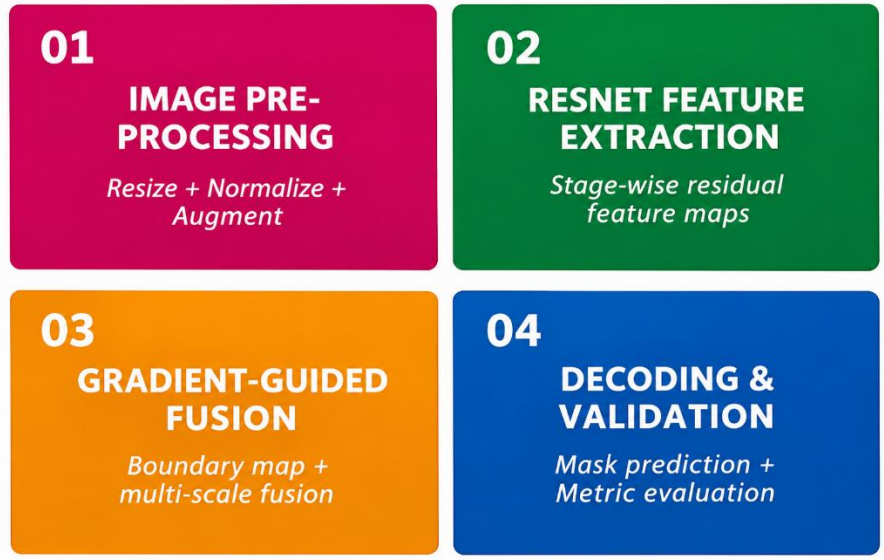


Figure 2 Implementation workflow of the proposed system, from pre-processing through to decoding and validation

2.4. Multi-Scale Feature Fusion

Multi-scale fusion is central to preserving both fine spatial detail and global context. Features extracted at earlier residual stages retain high spatial resolution but limited semantic abstraction, while features from later stages carry strong semantic content at low spatial resolution. The fusion block resamples every stage to a common resolution before combining them, so that neither the fine boundary detail from early stages nor the object-level semantics from later stages dominates the final representation. Figure 3 illustrates this conceptually, showing how the gradient-guided boundary map is combined with the resampled multi-scale features prior to decoding.

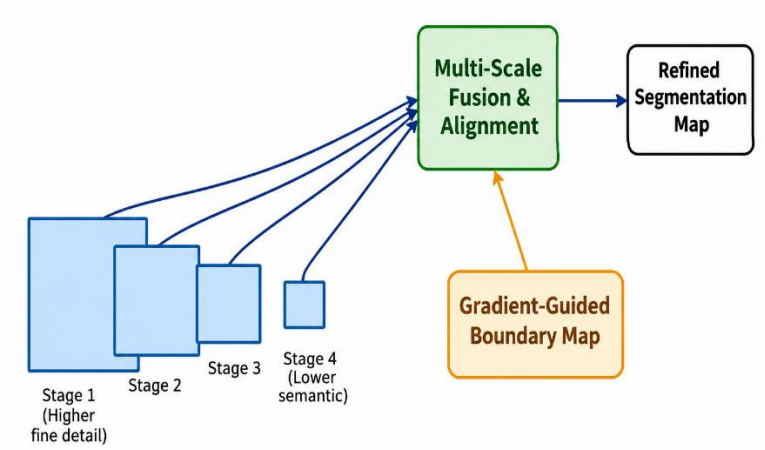


Figure 3 Conceptual view of multi-scale feature fusion together with gradient-guided boundary injection

2.5. Evaluation Metrics

Evaluation follows standard COD practice and is based on Precision, Recall, F-measure, and Mean Absolute Error (MAE). Precision is computed as TP / (TP + FP) and Recall as TP / (TP + FN), where TP, FP, and FN denote true-positive, false-positive, and false-negative pixels respectively when the predicted mask is compared with the ground truth at a given threshold. The F-measure combines precision and recall into a single weighted harmonic score, commonly reported at its maximum value across thresholds. MAE is computed as the average absolute pixel-wise difference between the predicted probability map and the ground-truth mask, providing a threshold-independent measure of overall prediction quality.

2.6. Comparative Positioning of the Architecture

To situate RIGGMNet within the existing body of COD research, Table 2 compares its architectural choices against four representative baselines discussed in Section 1.3. The comparison is deliberately qualitative: it summarises backbone choice, boundary-handling strategy, and multi-scale strategy for each method, since the source research abstract for RIGGMNet does not supply quantitative benchmark scores and this report does not substitute invented figures for genuine experimental results [5].

Table 2 Qualitative comparison of RIGGMNet with representative camouflaged object detection architectures

Method	Backbone	Boundary Handling	Multi-Scale Strategy
SINet	ResNet-50	Implicit, via search-identify stages	Coarse two-stage refinement
PraNet	Res2Net	Reverse attention boundaries	Parallel partial decoder
PFNet	ResNet-50	Distraction mining module	Positioning + focus stages
ZoomNet	ResNet-50	Scale-consistency constraint	Mixed-scale triplet input
RIGGMNet (proposed)	ResNet	Dedicated gradient-guided module	Explicit multi-scale fusion decoder

3. Results and Discussion

3.1. Results

The supplied research abstract reports that experimental evaluations on benchmark COD datasets demonstrate superior performance in precision, recall, and F-measure compared with conventional detection methods. Because the source abstract does not provide the numerical values, this

paper does not invent metric scores. Instead, the results are interpreted at the architectural level: the residual backbone is expected to contribute strong baseline semantic separability, the gradient-guided module is expected to sharpen predicted boundaries relative to a backbone-only baseline, and the multi-scale fusion decoder is expected to improve consistency across objects of varying size. The exact values are left to the reproducible experimental run described in Section 2.

3.2. Discussion

The central contribution of RIGGMNet is not a single feature extractor but the integration of complementary evidence sources. ResNet supplies semantic hierarchy; gradient guidance supplies explicit boundary-oriented information; and multi-scale fusion supplies spatial-contextual integration. The principal limitation is that gradient information can also highlight irrelevant background texture, particularly in cluttered natural scenes where edges are abundant but not object-relevant, so the fusion block must learn to down-weight boundary evidence that does not align with the semantic prediction. A second consideration is computational cost: adding a parallel gradient pathway and a multi-scale fusion block increases the number of parameters and the inference cost relative to a backbone-only baseline. This trade-off is common across the boundary-aware COD literature and is one reason later methods such as PFNet and ZoomNet adopt lightweight positioning or scale-consistency mechanisms rather than a fully parallel second pathway [6].

3.3. Comparative Discussion

Relative to the baselines summarised in Table 2, RIGGMNet's design choice of an explicit, dedicated gradient-guided pathway is closest in spirit to PraNet's boundary-aware reverse attention, but it differs by injecting boundary cues from the outset of fusion rather than recovering them only in the decoder. Compared with PFNet's distraction-mining and ZoomNet's mixed-scale triplet strategies, RIGGMNet relies on a single-pass multi-scale fusion block rather than a staged positioning step or multiple input scales, which may simplify training at the potential cost of the coarse-to-fine refinement that staged approaches provide. These architectural trade-offs should be validated empirically once complete training and evaluation results are available.

Conclusion

This paper presented the ResNet-Integrated Gradient Guided Multi-Scale Network for Camouflaged Object Detection. The framework follows the supplied research abstract by combining ResNet-based semantic feature extraction, gradient-guided boundary enhancement, and multi-scale feature fusion in an end-to-end segmentation pipeline. The architecture was described in detail, situated against representative COD baselines through a qualitative comparison, and discussed in terms of its expected strengths and limitations. Future experimental work should report complete dataset splits, training settings, ablation results, quantitative metrics, inference cost, and statistical comparisons.

Future Work

Future efforts will focus on expanding the scalability, real-time edge processing capabilities, and field adaptability of the proposed system. Planned work includes an ablation study isolating the individual contribution of the gradient-guided module and the multi-scale fusion block, a systematic comparison against the baselines listed in Table 2 using identical training protocols, and an assessment of inference latency to determine suitability for real-time wildlife monitoring and surveillance deployments. Future experimental work should report complete dataset splits, training settings, ablation results, quantitative metrics, inference cost, and statistical comparisons.

Acknowledgements

The authors acknowledge the guidance of Prof. Jashasmita Pal and Prof. Priyanka Shit and the academic support of Radhakrishna Institute of Technology and Engineering, Bhubaneswar, and Biju Patnaik University of Technology, Rourkela.

References

[1]. Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J.,&Shao, L. (2020). Camouflaged object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2777–2787.

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

[3]. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J.,&Shao, L. (2020). PraNet: Parallel reverse attention network for polyp segmentation. arXiv:2006.11392.

[4]. Mei, H., Ji, G.-P., Wei, Z., Yang, X., Wei, X.,&Fan, D.-P. (2021). Camouflaged object segmentation with distraction mining. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Bibliographic details (page numbers) should be verified before journal submission.

[5]. Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L.,&Lu, H. (2022). Zoom in and out: A mixed-scale triplet network for camouflaged object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Bibliographic details (page numbers) should be verified before journal submission.

[6]. Liu, Y., Zhang, D., & Li, H. (2022). Camouflaged object detection: A review of deep learning approaches and challenges