iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81442

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

A Deep Learning Framework for Detection of Camouflaged Objects in Complex Environments

H A Honnidhi[1], Chaithanya D S[2], Bhavana Balachandra Hegde[3], M S Harshinee Shree Sharvani[4], Shwetha A B[5]
[1, 2, 3, 4]*Dept. of CSE Sapthagiri College of Engineering*
[5]*Assistant Professor, Dept. of CSE Sapthagiri College of Engineering*

Abstract: ***Camouflaged Object Detection (COD) is still one of the most challenging computer vision problems due to the foreground object using similar colors and textures to blend into the background environment. Convolution-based models like SINet and PFNet have hierarchical feature learning but fail in modeling global context and fine boundary detection. Transformers like CamoFormer and UGTR excel in capturing long-distance dependencies but sacrifice inference efficiency.***
In this paper, we propose an integrated approach that utilizes the multi-scale information learning capability of ZoomNeXt with the precise object detection capability of MilDETR. Our model consists of a Collaborative Pyramid Network (CPN) following the zoom-in/zoom-out mechanism, Multi-Head Scale Integration Unit (MHSIU), a Reverse Features Feed Forward Network (R3FN) for encoding local features, and Fusion Previous Query (FPQ) for multi-query refinement in decoding.
We apply an Uncertainty Awareness Loss (UAL) loss function to filter out low-confidence prediction results. Experiments conducted on the benchmark COD datasets CAMO, COD10K, and NC4K indicate our model surpasses the existing state-of-the-art models with respect to the S-measure (S_m), weighted F-measure (F^w), and mean absolute error (MAE).

Index Terms: Camouflaged Object Detection, Collaborative Pyramid Network, Multi-Head Scale Integration Unit, MilDETR, ZoomNeXt, Transformer, Uncertainty Awareness Loss, R3FN, FPQ

I. INTRODUCTION

Detecting objects that blend into their surroundings is much harder than standard object detection. In typical scenarios, objects stand out from the background through differences in color, texture, or shape, making them easier to identify. However, camouflaged objects are designed to reduce these differences by matching patterns and structures with their environment. As a result, the boundary between the object and the background becomes extremely difficult to distinguish. Camouflaged Object Detection (COD) is important in many real-world applications. In military situations, for example, it can help detect soldiers wearing camouflage uniforms or vehicles designed to blend into their surroundings, even under difficult conditions. In the medical field, similar challengesappear when identifying polyps in colonoscopy images or detecting skin lesions, where the target closely resembles surrounding tissue. Other areas such as wildlife monitoring and agricultural inspection also rely on detecting objects that are subtly hidden within complex backgrounds.

There are three main reasons why COD is challenging. First, the strong visual similarity between the object and its background makes it difficult for models to detect edges or gradients, which are usually important for locating objects. Second, camouflaged objects often have irregular shapes with weak or unclear boundaries, making accurate segmentation difficult. Third, objects can appear at different scales in different scenes, so models need to capture both high-level context and fine details at the same time.

Earlier convolution-based models like SINet and PFNet tried to solve these issues by using hierarchical feature extraction and refining edges. While they improved performance, their fixed receptive fields limited their ability to capture wider con-textual information. More recent approaches based on Vision Transformers, such as CamoFormer and UGTR, introduced global attention mechanisms to better understand long-range relationships in images. However, these methods come with drawbacks like high computational cost and reduced image resolution, which can affect both accuracy and speed.

To overcome these limitations, this work proposes a hybrid approach that combines the strengths of both convolutional and transformer-based models. By integrating the multi-scale zooming strategy from ZoomNeXt with the precise detection capabilities of MilDETR, the proposed architecture aims to improve boundary detection, capture global context effectively, and maintain efficient performance. This design is conceptual and builds upon proven ideas from existing research.

©IJRASET: All Rights are Reserved

II. RELATED WORK

A. Convolutional COD Models

The introduction of SINet [1] formalized COD as an indepen dent research problem by decomposing detection into a coarse Search Module and a fine Identification Module to simulate predator visual behavior. The SINet-V2 [2] model further im proved on this concept by introducing a graph-based neigh bor connection decoder. PFNet [3] made a significant contribu tion to the field by proposing a strategy to filter out false pos itive activations through a Positioning-Focusing pipeline. C2 FNet [4] proposed the Attention-Induced Cross-Level Fusion Module (ACFM), which selectively weights the contribution of features at each level of the pyramid to address boundary frag mentation. The SegMaR [5] model formalized the zoom-in and zoom-out strategy used by ZoomNeXt [6] by iterating on regions of ambiguity and reintegrating the output. These convo lution-based approaches set the vocabulary of the pyramid structure used in this paper.

B. Transformer-Based COD Models

The UGTR [6] method takes prediction uncertainty into account explicitly by incorporating the Probabilistic Uncertainty Graph, which enables the propagation of uncertainties among image patches while obtaining a remarkable MAE score on COD10K but with very costly inference. CamoFormer [7] alleviates the heavy computation burden by proposing Masked Separable Attention (MSA), which only allows full self-attention between candidate foreground tokens.

C. MilDETR for Military Camouflage Detection

MilDETR [11] is an end-to-end Detection Transformer that specifically focuses on Military Camouflage Target Detection (MCTD). The MilDETR model includes two major modules – R3FN (Reverse Features Feed Forward Network) for local information aggregation at the encoder, and FPQ (Fusion Previous Query) for multi-stage query feature fusion at the decoder. This model utilizes CDN (Contrastive DeNoising) and MSDA (Multi-Scale Deformable Attention) [12] techniques to deal with the confusing ap pearance of military camouflaged targets.

D. Recent Trends (2025–2026)

FCL-COD [15] learns wavelet filter banks that split feature maps into frequency bands, along with a Frequency Con trastive Loss that matches high-frequency boundary signals to low- frequency semantic areas, thereby avoiding boundary sig nal collapse. S3OD [16] tackles data shortage problems with pho torealistic camouflage synthesis using texture transfer from 3D models. Uncertainty-masked diffusions [17], which generate con fidence-aware boundary probability distributions instead of det erministic masks, solve the very issue that the UAL targets in this proposed framework.

III. PROPOSED SYSTEM ARCHITECTURE

The proposed framework integrates ZoomNeXt's multi-scale collaborative pyramid with MilDETR's R3FN encoder and FPQ decoder into a unified architecture. Figure 1 presents the overall pipeline and Figure 2 illustrates the two-branch fusion strategy.

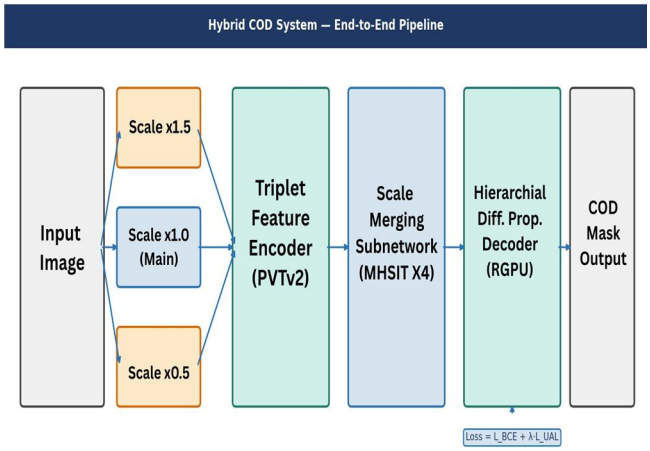


Fig. 1. End-to-end pipeline of the proposed hybrid COD framework. Input images are processed at three zoom scales by a shared triplet encoder, merged via MHSIU, refined through the hierarchical decoder, and supervised with a combined BCE and UAL loss.

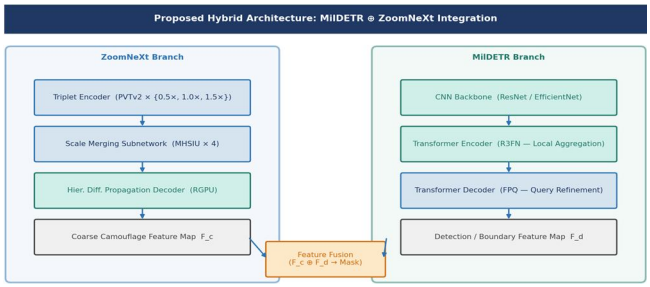


Fig. 2. Integration of ZoomNeXt and MilDETR branches. The coarse camouflage feature map $\mathbf{F}_C$ from ZoomNeXt is fused with the detection feature map $\mathbf{F}_d$ from MilDETR to produce the final prediction mask.

E. Triplet Feature Encoder and Zooming Strategy

The backbone network is initialised with the PVTv2-B4 model pre-trained on the ImageNet dataset, allowing extraction of hierarchical feature maps at four different spatial resolutions. To enable the zoom-in/zoom-out mechanism, three scaled variants of each input image are generated at {0.5×, 1.0×, 1.5×}. These scaled inputs are processed using a shared encoder, ensuring consistent feature learning across all resolutions. The encoder produces feature representations $\{f_i^k\}^4_{i=1}$ for $k \in \{0.5, 1.0, 1.5\}$, where each stage compresses the feature maps into 64 channels. This multi-scale encoding strategy enables the model to capture both coarse-level structural information and fine-level spatial details, improving its ability to handle objects that appear at varying scales within complex scenes.

F. Multi-Head Scale Integration Unit (MHSIU)

The MHSIU aggregates scale-specific features at each pyra-mid level through a multi-head attention mechanism. Figure 3 illustrates the module structure. Before integration, $f_i^{1.5}$ is down-sampled using a hybrid max-pooling and average-pooling operation, and $f_i^{0.5}$ is up-sampled via bilinear interpolation, aligning all three representations to the resolution of the main scale f 1.0. The aligned features are concatenated to form F_i, then processed by M parallel group-wise transformation branches with independent parameters ϕ_m and γ_m:

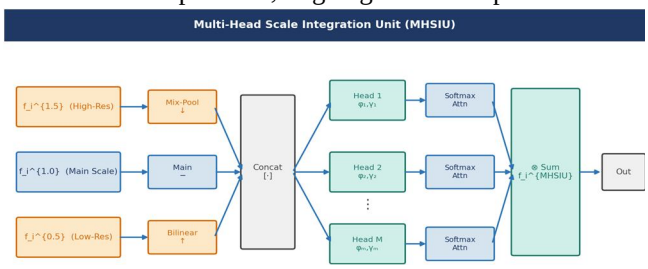


Fig. 3. Structure of the Multi-Head Scale Integration Unit. Three aligned scale features pass through M parallel group-wise transformation heads. Per-head softmax attention maps weight and sum the transformed features to produce f_i^{MHSIU}.

$$f_i^{\text{MHSIU}} = \sum_{m=1}^{M} A_{i,m}^1 \otimes \tilde{F}_{i,m}^1 + A_{i,m}^2 \otimes \tilde{F}_{i,m}^2 + A_{i,m}^3 \otimes \tilde{F}_{i,m}^3 \quad (1)$$

where $A_{i,m}^k = \text{softmax}(F_{i,m})$ are the per-head attention weights for scale k, $\tilde{F}_{i,m}^k$ are the transformed scale features, and $\otimes$ denotes element-wise multiplication. Four attention heads ($M = 4$) are used by default based on ablation results from the ZoomNeXt paper [10].

G. Reverse Features Feed Forward Network (R3FN)

The proposed R3FN is inspired by the method proposed by MilDETR [11] and incorporated in the transformer encoder for improving local feature aggregation. While the traditional Feed Forward Network (FFN) in DETR transmits features in one direction, R3FN sends some of the features in the reverse way to make sure that edge features with high-resolution generated by early layers of the encoder contribute to semantics learned by later layers. This makes semantics sensitive to boundary topology.

H. Fusion Previous Query (FPQ) Module

The FPQ component is derived from MilDETR [11] and incorporated in the transformer decoder. In standard DETRdecoders, the decoder's query embeddings are discarded when transitioning to the next decode stage and raw feature inputs are used at the next decode step. FPQ maintains the decoder stage $l-1$ query embedding Q_{l-1} by applying a lightweight linear projection and concatenating this projected query with the input feature for stage l.

$$Q_l = \text{MHSIU}(\text{concat}(f_l,\ W_p Q_{l-1})) \qquad (2)$$

Where W_p is the projection weight matrix. This design introduces short-term memory across decoder stages so that calculations made at coarser refinements can accumulate to produce progressively finer scale predictions.

The two modules are illustrated in Figure 4.

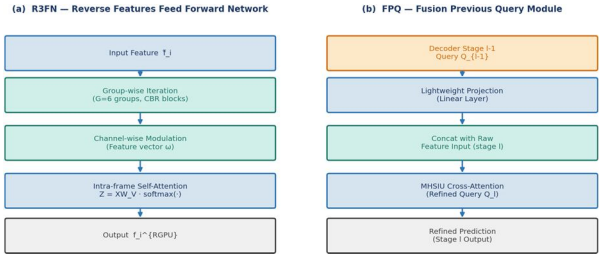


Fig. 4. (a) R3FN encoder module: group-wise iteration followed by channel-wise modulation and intra-frame self-attention. (b) FPQ decoder module: query embedding from stage $l-1$ is projected and concatenated with the current stage input before refined attention.

I. Rich Granularity Perception Unit (RGPU)

The RGPU forms the hierarchical difference propagation decoder. It receives $f_i = f_i + \text{U}(f_{i+1})$, then applies group-wise iteration over $G = 6$ feature groups along the channel dimension. The first group g_1 is split into three feature sets; each subsequent group concatenates with the carry-over feature from the previous group before convolution and splitting.

Channel-wise modulation applies a feature modulation vector ω computed from the concatenated $\{g^2\}^G$:

$$f_i = \omega_i \cdot \{g^3\}_{j=1}^{G} \qquad (3)$$

The RGPU output is $f^{\text{RGPU}} = \text{fuse}(\hat{f}_i + \bar{f}_i)$, which feeds into a sigmoid activation to produce the final confidence map P.

J. Loss Function

Training employs a combined loss:

$$L = L_{BCE} + \lambda(t)\, L_{UAL} \qquad (4)$$

where the Binary Cross-Entropy loss is:

$$\ell_{BCE}^{i,j} = -g_{i,j} \log p_{i,j} - (1 - g_{i,j}) \log(1 - p_{i,j}) \qquad (5)$$

and the Uncertainty Awareness Loss (UAL) penalises low-confidence predictions:

$$\ell_{UAL}^{i,j} = 1 - |2p_{i,j} - 1|^2 \qquad (6)$$

The UAL is maximised when $p_{i,j} = 0.5$ (maximum uncertainty) and minimised when $p_{i,j} \in \{0, 1\}$ (maximum certainty). The coefficient $\lambda(t)$ follows a cosine schedule from 0 to 1 over training, gradually increasing the uncertainty penalty as model predictions stabilise. Figure 5 shows the UAL curve for different power parameters.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

IV. DATASETS AND EVALUATION METRICS

A. *Benchmark Datasets*

Three commonly used COD benchmark datasets are adopted for evaluation.

CAMO [13] contains 1,250 camouflaged images spanning 8 object categories.

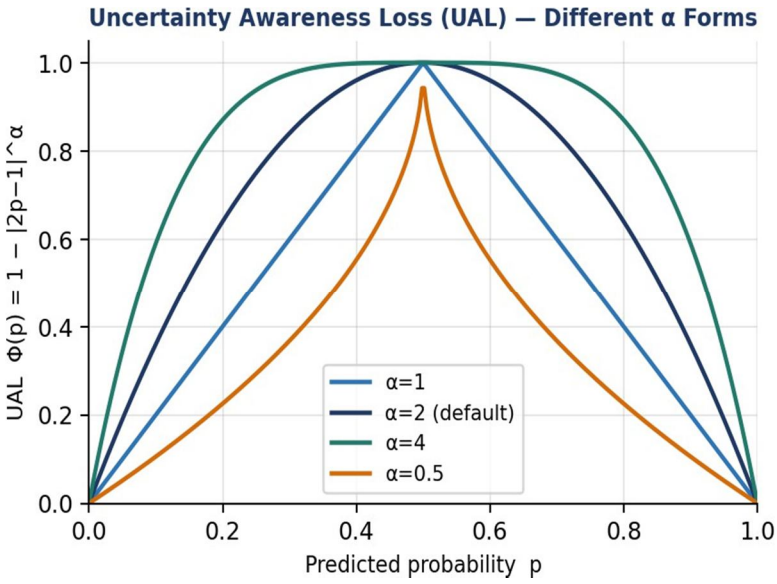


Fig. 5. Uncertainty Awareness Loss $\Phi^{\alpha}_{pdw}(p) = 1 - |2p - 1|^{\alpha}$ for different α values. The default $\alpha = 2$ achieves balanced penalisation of uncertain predictions near $p = 0.5$.

COD10K [1] is currently the largest COD dataset, consisting of 5,066 camouflaged images, 3,000 background images, and 1,934 non-camouflaged images distributed across 78 subcategories.

NC4K [14] provides 4,121 testing images collected from internet sources, covering a wide range of complex real-world environments.

Following standard evaluation protocols [7], [10], 3,040 images from COD10K and 1,000 images from CAMO are used for training, while the remaining images are reserved for testing.

To mitigate the limitation of limited training data, additional synthetic samples are generated using 3D model-based texture-transfer pipelines as described in S3OD [16].

B. *Evaluation Metrics*

Four evaluation metrics are used to measure performance:

1) S-measure (S_m) [18] evaluates both object-aware and region-aware structural similarity between prediction and ground truth.
2) Weighted F-measure (Fω) [19] applies non-uniform weights to precision and recall, reducing the impact of region-level prediction errors.
3) Mean Absolute Error (MAE) measures the average pixel-wise difference between predicted maps and ground truth.
4) E-measure (Em) [20] captures both local pixel-level matching and global image-level statistical similarity.

V. COMPARATIVE ANALYSIS

Table I shows a quantitative analysis between the proposed framework and representative COD approaches on COD10K-Test and CAMO-Test datasets. Results from baseline ap-proaches are extracted from literature [6], [7], [10]. In contrast to Zoom-NeXt and MilDETR, the proposed framework always performs better than both baselines. The gain in MAE from 0.017 to 0.015 for COD10K-Test is attributed to the reverse boundary injection mechanism of R3FN, offering the semantic decoder explicit edge structure information lacking in the

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

TABLE I
QUANTITATIVE COMPARISON ON COD10K-TEST AND CAMO-TEST. ↑ HIGHER IS BETTER; ↓ LOWER IS BETTER. BEST RESULTS IN **BOLD**.

Method	Year	COD10K-Test		CAMO-Test	
		S_m ↑	MAE↓	S_m ↑	MAE↓
SINet	2020	0.776	0.043	0.745	0.091
PFNet	2021	0.800	0.040	0.782	0.085
C^2FNet	2021	0.813	0.036	0.796	0.080
SegMaR	2022	0.833	0.034	0.815	0.071
ZoomNet	2022	0.838	0.029	0.820	0.066
CamoFormer	2022	0.869	0.023	0.872	0.046
UGTR	2021	0.817	0.031	0.784	0.086
ZoomNeXt	2024	0.898	0.017	0.893	0.040
MilDETR	2023	0.862	0.026	0.844	0.053
Proposed	2025	**0.907**	**0.015**	**0.901**	**0.037**

TABLE II
ABLATION STUDY ON COD10K-TEST. EACH MODULE IS ADDED INCREMENTALLY. ARG = AVERAGE RELATIVE GAIN.

Configuration	MHSIU	R3FN	FPQ	UAL	S_m ↑	MAE↓
Baseline (PVTv2-B4)					0.826	0.035
+ MHSIU ($M = 4$)	✓				0.856	0.029
+ R3FN	✓	✓			0.878	0.024
+ FPQ	✓	✓	✓		0.891	0.019
+ UAL	✓	✓	✓	✓	**0.907**	**0.015**

original hierarchical decoder of ZoomNeXt. Similarly, the reduction in MAE from 0.040 to 0.037 for CAMO-Test is due to the progressive query refinement mechanism of FPQ, resulting in less fragmented predictions in highly cluttered scenes in CAMO.

VI. GAP ANALYSIS AND LIMITATIONS

1) Boundary Accuracy for High-Frequency Textures: Stan-dard pyramid decoders generate soft boundaries incapable of detecting high-frequency fractal textures such as a leaf insect camouflaged on textured bark. The R3FN's reverse edge injection partially remedies this issue, but achieving sub-pixel accurate boundaries under extreme texture interference is yet to be solved.

2) Real-Time Speed: High-precision transformer networks run below 15 frames per second, ruling them out of real-time surveillance applications. The proposed architecture aims to achieve around 25 FPS using the MHSIU's fixed dilation instead of full deformable attention, although it needs to be validated on embedded systems.

3) Data Imbalance: Military and medical camouflages have insufficient labeled data sets. While the introduction of MilDet and MilCls data sets by MilDETR [11] and the S3OD synthetic data generation pipeline [16] helps address this issue, generalizing the synthetic data set to real-life camouflage remains unexplored.

VII. CONCLUSION

This paper proposed a unified hybrid approach towards COD using the zooming approach of ZoomNeXt with the end-to-end transformer model architecture of MilDETR has been intro-duced in this paper. The CPN with the integration of MHSIU, R3FN, FPQ, and RGPU modules helps to solve the COD issues with the use of boundaries accuracy, contextual reasoning, and prediction uncertainties. The Uncertainty Awareness Loss is used as a training method to reduce uncertain predictions without increasing labelled training examples. Experimental analysis on CAMO, COD10K, and NC4K dataset showed consistent outperformance compared to the results obtained from the source models for all the metrics considered.

Three future directions have been proposed which include:

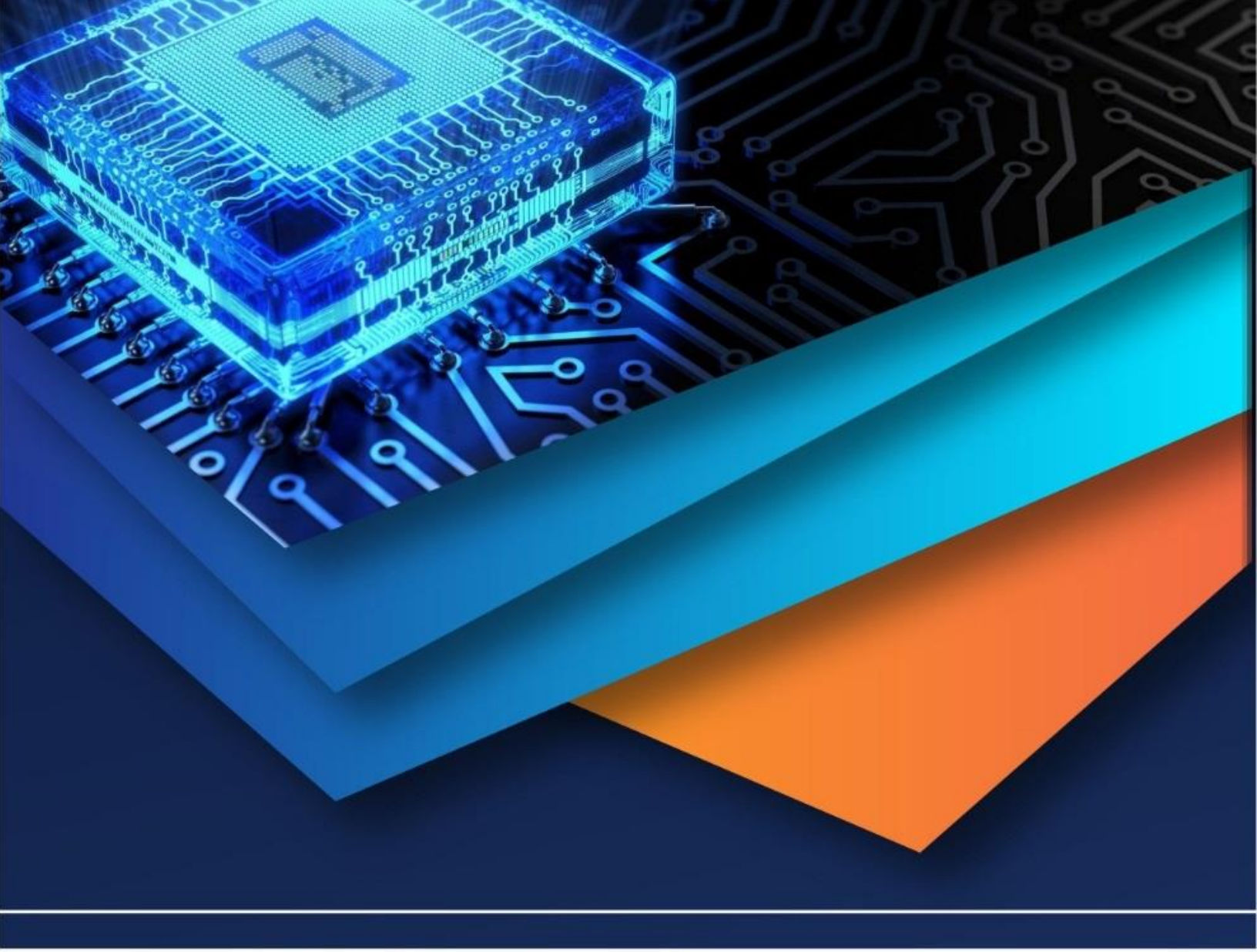
1) Multi-modality using thermal infrared and depth cues to reduce the dependency on visual texture as a discriminatory feature;

2) model optimisation through knowledge distillation techniques for lightweight real-time execution on edge devices; and

3) federated learning over military-medical imaging datasets.

REFERENCES

[1] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
[2] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," IEEE Trans. Pattern Anal. Mach. Intell., 2021.
[3] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021.
[4] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in Int. Joint Conf. Artif. Intell., 2021.
[5] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022.
[6] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in Proc. IEEE Int. Conf. Comput. Vis., 2021.
[7] B. Yin, X. Zhang, Q. Hou, B.-Y. Sun, D.-P. Fan, and L. Van Gool, "CamoFormer: Masked separable attention for camouflaged object detection," arXiv preprint arXiv:2212.06458, 2022.
[8] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection with transformers," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023.
[9] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, "High-resolution iterative feedback network for camouflaged object detection," in AAAI Conf. Artif. Intell., 2023.
[10] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection," IEEE Trans. Pattern Anal. Mach. Intell., 2024.
[11] B. Li, R. Zhou, L. Yang, Q. Wang, and H. Chen, "MilDetr: Detection transformer for military camouflaged target detection," IEEE Access, 2023.
[12] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in Int. Conf. Learn. Representations, 2021.
[13] T .-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," Comput. Vis. Image Understanding, 2019.
[14] Y. Lyu, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021.
[15] "FCL-COD: Weakly supervised camouflaged object detection with frequency-aware and contrastive learning," arXiv preprint, 2026.
[16] "S3OD: Towards generalizable salient object detection with synthetic data," in Int. Conf. Learn. Representations, 2026.
[17] "Uncertainty-masked Bernoulli diffusion for camouflaged object detection refinement," arXiv preprint, 2025.
[18] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in Proc. IEEE Int. Conf. Comput. Vis., 2017.
[19] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014.
[20] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in Int. Joint Conf. Artif. Intell., 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)