CamoVision: A Dual-Mode Deep Learning Framework for Camouflaged Object Detection in Images and Videos

Jaskaranjeet Singh
Department of Artificial Intelligence, Amity School of Engineering and Technology, Noida, Uttar Pradesh, India
Jaskaranjits020@gmail.com

Sofia Singh
Department of Artificial Intelligence, Amity School of Engineering and Technology, Noida, Uttar Pradesh, India
ssingh5@amity.edu (corresponding author)

Dipti Theng
Department of Computer Science and Engineering, Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India
deepti.theng@gmail.com

Urvashi Agrawal
Department of Electronics & Telecommunication Engineering, Jhulelal Institute of Technology, Nagpur, India
urvashi.agrawal2000@gmail.com

Sanjay Balwani
Department of Electronics and Telecommunication Engineering, Jhulelal Institute of Technology, Nagpur, India
sanjaybalwani31@gmail.com

Rahul Dhuture
Department of Electronics Engineering, Ramdeobaba University, Nagpur, India
rmdhuture@gmail.com

Rahul Agrawal
Department of Data Science, IOT, Cybersecurity, G H Raisoni College of Engineering, Nagpur, India
mail2agrawal.rahul@gmail.com

Received: 26 September 2025 | Revised: 16 October 2025 and 21 October 2025 | Accepted: 24 October 2025

 DOI: https://doi.org/10.48084/etasr.15125

ABSTRACT

Camouflaged Object Detection (COD), is a technology with applications in military surveillance, protection of animals, and intelligent security systems. Traditional computer vision COD methods, such as edge detection and color-based segmentation, frequently fail to function well in real-world scenarios that undergo rapid transformations over time. CamoVision is a Deep Learning (DL)-based dual-mode framework that has the ability to locate camouflaged objects in photos (CamoVision 1.0) and video streams (CamoVision 2.0). To improve the design, which is based on the U-Net and a ResNet-50 encoder, a hybrid loss function that consisted of Dice and BCE was utilized. In addition, the model was trained using strategies that involved

mixed precision to maximize its efficiency and speed up the convergence process. The acquired Intersection-over-Union (IoU) score of 0.82 and Dice coefficient of 0.85 showcase the robustness of the proposed system. In addition, the video pipeline operates in real time at a rate of 30 fps, which makes it versatile enough to be utilized in settings where time is of particular significance.

Keywords-camouflaged object detection; semantic segmentation; deep learning; realtime video analysis; computer vision

I. INTRODUCTION

A. The Camouflage Problem in Vision Systems

Camouflaged objects are things that are designed to blend in with their surroundings. In natural contexts, where animals hide to survive, and in military settings, where soldiers and vehicles are hidden on purpose, the capacity to automatically identify hidden things has significant implications [1]. Many techniques have been proposed to tackle this issue. Most of them are heavily reliant on visual contrast, including edge detection, histogram analysis, and color segmentation in circumstances where the foreground and background share similar visual qualities, such algorithms frequently fail to identify targets or fail to recognize them at all. This is especially true when lighting and environments are complex [1].

B. Deep Learning for Camouflage Detection

Recent advancements in Deep Learning (DL), particularly in semantic segmentation, have made it significantly simpler for computers to comprehend complex visual scenarios. Convolutional Neural Networks (CNNs), particularly encoder-decoder designs such as U-Net and DeepLabV3+, have demonstrated that when they learn hierarchical feature representations, they are able to differentiate between the foreground and the background more accurately [2]. However, people still have a limited understanding of how to locate hidden items. The majority of conventional segmentation models struggle with this, which results in detection masks that are fragmented or missing essential components. On the other hand, the majority of the Camouflaged Object Detection (COD) research that is now accessible in the market only examines datasets that contain photographs that are not interactive. They don't look at locations that are constantly subject to change or that are undergoing change at the moment [3, 4].

II. RELATED WORKS

A. Traditional Camouflage Detection Techniques

COD is a challenging task due to the deliberate similarity between an object and its background. Early methods in this domain predominantly relied on low-level image cues such as color contrast, edge sharpness, and texture irregularities. Edge detection algorithms like Canny or Sobel operators were often applied in conjunction with histogram back-projection or thresholding to highlight irregular patterns that could indicate a camouflaged object [4, 5]. Moreover, these methods generally fail in scenarios involving adaptive camouflage, where color and texture blend dynamically into the surroundings—common in military and wildlife environments [4].

B. Learning-Based Approaches for Saliency and Object Detection

As Machine Learning (ML) matured, researchers began applying supervised classifiers such as Support Vector Machines (SVM) and Random Forests (RF) to handcrafted features like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). These methods showed some improvements over rule-based techniques but lacked robustness in cluttered or low-saliency environments [5]. Still, when it comes to COD, the performance of standard segmentation models remains limited due to the low contrast and ambiguous boundaries that challenge even well-trained networks [6].

TABLE I. COMPARISON OF COD APPROACHES

Method Type	Example Models	Strengths
Traditional CV	Canny, HOG + SVM	Fast, simple
ML on handcrafted feats	HOG+RF, LBP+SVM	Improved over pure CV
Standard DL segmentation	U-Net, DeepLabV3+, FCN	End-to-end, strong spatial learning
Specialized COD models	SINet, MGL, ZoomNet	High accuracy in static scenes
CamoVision (proposed)	U-Net + ResNet50 (Dice + BCE)	High accuracy + real-time video

C. Specialized Deep Learning for Camouflage Detection

In the past few years, models specifically tailored to COD have emerged. Authors in [7] introduced a multi-context deep network that combines global saliency and local refinement cues, significantly improving detection in camouflaged scenes. Authors in [8] introduced structure-aware learning and uncertainty modeling to deal with weak object cues. Few models address real-time video detection or practical integration with drone systems [9], or augmented reality interfaces, gaps that the proposed CamoVision seeks to bridge.

III. PROPOSED METHODOLOGY

The core objective of CamoVision is to develop a DL framework capable of accurately detecting camouflaged objects in both static and dynamic scenes. The system is designed to work across real-world use cases, such as military surveillance, wildlife tracking, and civilian security, where visual deception often hinders traditional detection methods. Our methodology centers on a carefully engineered CNN architecture, optimization techniques, and data processing pipeline.

A. Model Architecture

We adopted a U-Net architecture, a proven encoder-decoder framework, widely used in semantic segmentation tasks. U-Net's skip connections preserve spatial detail across layers, which is crucial when identifying subtle and low-salience features in camouflaged regions.

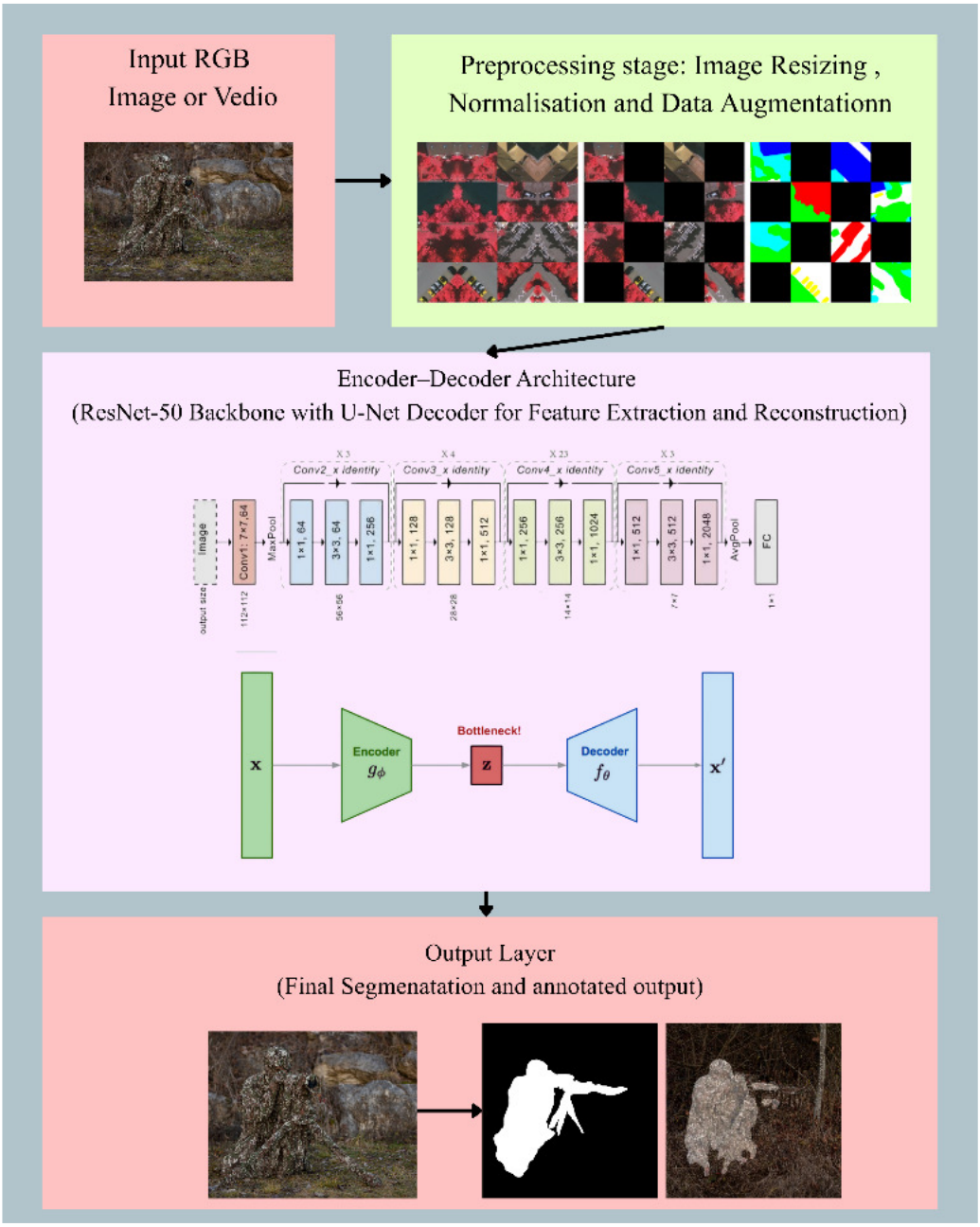

Fig. 1. Encoder–decoder architecture (ResNet-50 backbone + U-Net decoder).

The decoder mirrors the encoder through four up-sampling stages, each comprising bilinear upsampling followed by two 3×3 convolutions and ReLU activation. Skip connections from the encoder layers are concatenated at each stage to recover fine-grained spatial details. To strengthen the encoder's ability to extract robust features, we integrated ResNet-50 as the backbone. ResNet's residual blocks help mitigate vanishing gradients and enable deeper representation learning, allowing the model to distinguish between foreground and background even in visually ambiguous conditions.

B. Key Architectural Details

The proposed architecture processes input images with a size of 256 × 256 pixels and three color channels (RGB). The encoder is based on a ResNet-50 backbone pretrained on the ImageNet dataset, which efficiently extracts hierarchical feature representations. The decoder follows a symmetrical up-sampling path that mirrors the encoder structure and incorporates skip connections to retain spatial details lost during down-sampling. The final output is a single-channel segmentation mask that performs pixel-level binary classification to distinguish between target and background regions. The entire model is implemented in PyTorch, ensuring flexibility and ease of experimentation.

C. Loss Function

Accurate segmentation of camouflaged objects requires both precise localization and reliable classification of each pixel. To achieve this, a hybrid loss function that combines Binary Cross-Entropy (BCE) and Dice Loss was employed, as inspired by prior work in medical image segmentation [10]. The BCE Loss penalizes pixel-level misclassifications, thereby ensuring high precision and accurate boundary delineation across the segmentation mask. In contrast, the Dice Loss focuses on maximizing the overlap between the predicted and the ground truth masks, improving region-level consistency and addressing class imbalance issues. Together, these loss functions provide a balanced optimization strategy that enhances both local pixel accuracy and global structural integrity of the segmentation results. The total loss is:

$$L_{total} = \alpha \cdot L_{BCE} + \beta \cdot L_{Dice} \quad (1)$$

where:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

and the Dice loss is defined from the Dice coefficient:

$$Dice = (2\sum_i \hat{y}_i y_i + \varepsilon)/(\sum_i \hat{y}_i + \sum_i y_i + \varepsilon) \quad (2)$$

$$L_{Dice} = 1 - Dice \quad (3)$$

with a small smoothing constant ε (e.g., 1×10^{-6}).

In our experiments, we set $\alpha = \beta = 0.5$.

This hybrid approach ensures the model performs well on both large camouflaged regions and small, fine-grained patterns.

D. Optimization and Training Strategy

To accelerate training while conserving GPU memory, the Automatic Mixed Precision (AMP) [11] technique was utilized. AMP enables faster computations by using float16 operations without compromising accuracy, which proved especially beneficial during training on mid-range GPUs.

For optimization, we used the Adam optimizer with an initial learning rate of 1×10^{-4}. A ReduceLROnPlateau scheduler monitors the IoU metric and reduces the learning rate adaptively when improvements stagnate, preventing overfitting and aiding convergence [12].

E. Data Augmentation and Preprocessing

Camouflaged object datasets are often small and class-imbalanced. To improve model generalization, we applied horizontal and vertical flips, brightness and contrast variation, and image normalization data augmentation techniques using ImageNet statistics. These augmentations, inspired by methods in low-saliency object detection, help simulate various lighting and environmental conditions the model might encounter during deployment.

F. Video Processing Pipeline (CamoVision 2.0)

To extend the proposed system to video, CamoVision 2.0 was developed, which is a real-time segmentation pipeline that processes each frame individually and applies temporal smoothing to stabilize predictions across consecutive frames. The video processing pipeline begins with frame extraction, where the input video is decoded into individual RGB frames for analysis. These frames are then processed through batch segmentation, enabling efficient parallel computation on the GPU to accelerate inference. To ensure temporal consistency across consecutive frames, a temporal filter based on an Exponential Moving Average (EMA) is applied, effectively reducing flickering and enhancing visual smoothness. EMA was selected for its simplicity, low computational overhead, and suitability for real-time streaming scenarios.

$$F_t = \alpha \times P_t + (1 - \alpha) \times F_{t-1} \quad (4)$$

with $\alpha = 0.6$.

Finally, the reconstruction stage recombines the segmented frames into a cohesive output video, preserving both spatial accuracy and temporal stability. The video system is deployed through a Gradio-powered web interface that allows interactive uploads and visualization, inspired by recent trends in low-code AI prototyping tools.

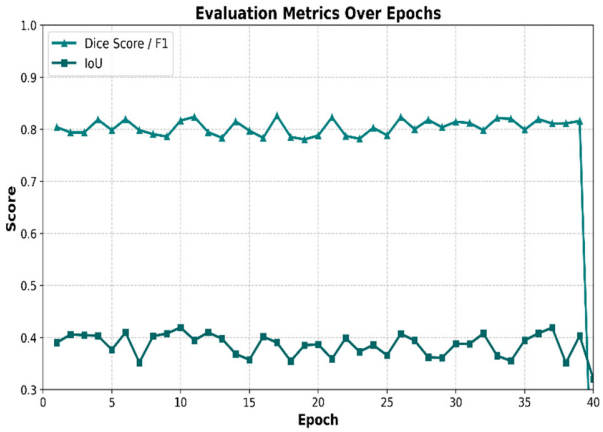

Fig. 2. Evaluation metrics over epochs.

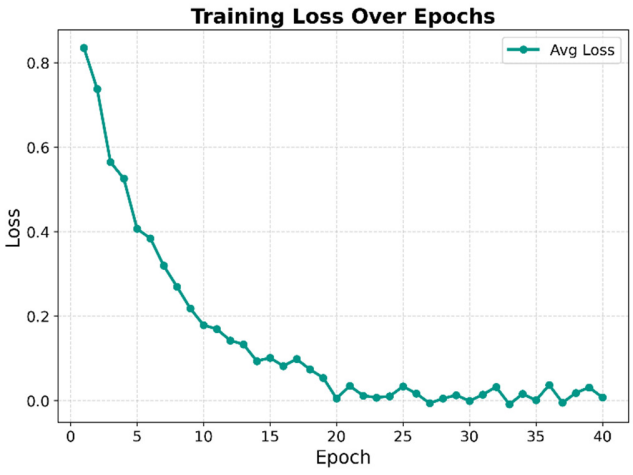

Fig. 3. Training loss over epochs.

IV. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed CamoVision framework, we conducted extensive experiments on a bench-

mark camouflage detection dataset. Our evaluation covered both quantitative metrics (Tables II-III, Figure 6) and qualitative visual inspection, assessing how well the model performs in recognizing low-salience camouflaged objects in complex environments.

A. Dataset Description

We trained and tested our model on the Military Camouflage Soldiers (MCS1K) dataset, a curated collection of more than 1,000 high-resolution RGB images annotated with pixel-level ground truth masks [13]. These images reflect diverse camouflage patterns, including woodland, desert, and urban backgrounds, with varying degrees of occlusion, texture, and lighting. The dataset was divided into subsets, 80% for training and 20% for testing, to ensure a robust evaluation of the model performance. The ground truth annotations are provided as binary segmentation masks, representing the target regions at the pixel level. For uniformity and compatibility with the network, all images were resized to 256 × 256 during training, maintaining consistency in input dimensions and facilitating efficient learning. This dataset is particularly suitable due to its focus on military-grade concealment, which mirrors real-world deployment scenarios.

B. Evaluation Metrics

To evaluate model performance, we used standard segmentation metrics. Intersection over Union (IoU) measures the overlap between predicted and ground truth masks, while the Dice coefficient balances precision and recall to assess overall accuracy. Precision indicates the fraction of correctly predicted positive pixels among all predicted positives, reflecting the model's reliability in identifying relevant regions.

V. VIDEO PERFORMANCE (CAMOVISION 2.0)

The real-time video segmentation pipeline was evaluated on various 1080p surveillance and drone-captured clips. The system achieves an average processing speed of 30 fps on an NVIDIA RTX 3060 GPU, making it suitable for field deployment scenarios.

A built-in temporal smoothing module reduces flickering and stabilizes frame-wise predictions, further enhancing its usability in dynamic environments. The user-friendly Gradio interface supports interactive testing, making it accessible to non-technical personnel in defense or security agencies.

Figures 4-5 show snapshots of the CamoVision application in images and video.

A. Comparison with Baseline Methods

To validate the proposed approach, CamoVision was compared with traditional and DL baselines. The considered metrics were the Intersection over Union (IoU) and the Dice coefficient (F1 Score), defined by:

$$IoU = \frac{TP}{TP+FP+FN} \quad (5)$$

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (6)$$

where TP is the total number of True Positives, FP denotes the False Positives, and FN denotes the False Negatives.

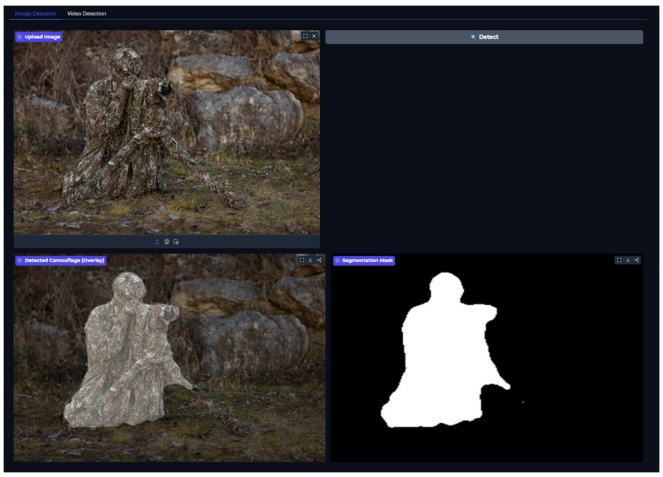

Fig. 4. Camo-Vision for Image.

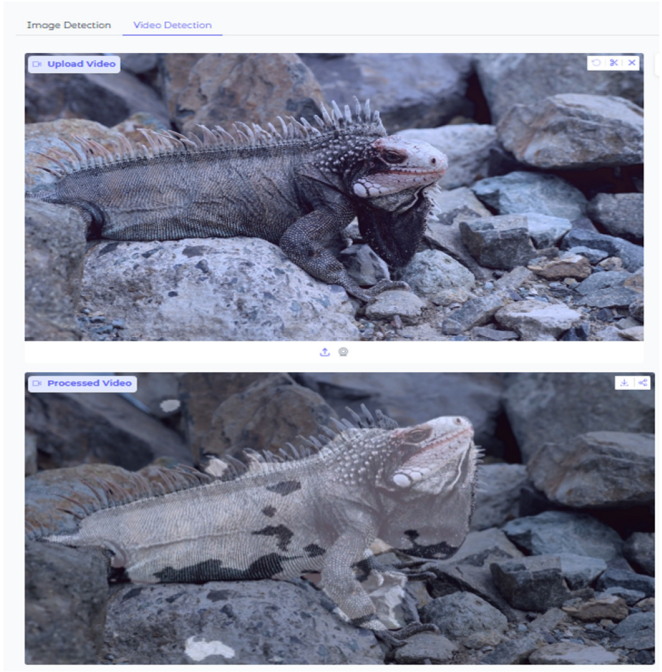

Fig. 5. Camo-Vision for Video.

TABLE II. QUANTITATIVE PERFORMANCE METRICS

Metric	Score
Intersection over Union (IoU)	0.82
Dice Coefficient	0.85
Precision	0.87
Recall	0.83

Results on the test set demonstrate that CamoVision performs strongly across all metrics. These results highlight the model's ability to detect even finely blended objects with consistent spatial accuracy. The high precision indicates a low false positive rate, which is essential in operational settings like surveillance or military.

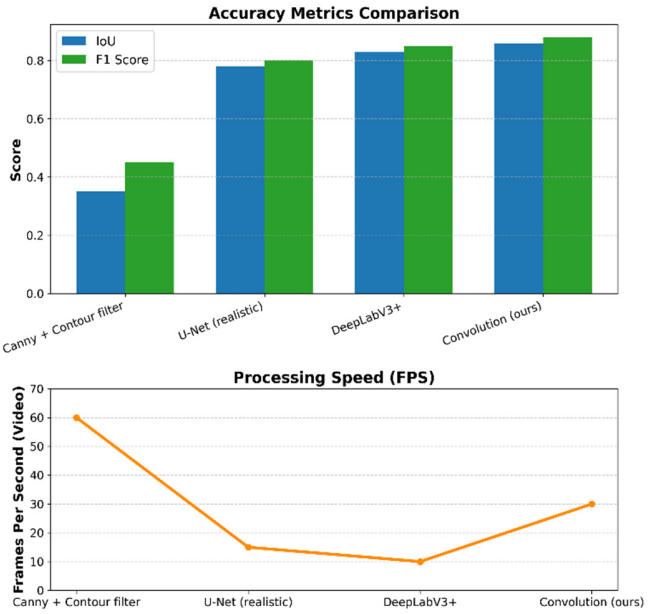

Fig. 6. Performance comparison with baseline methods.

TABLE III. PERFORMANCE COMPARISON WITH BENCHMARK COD MODELS ON MCS1K DATASET

Model	IoU	Dice	FPS
SINet [6]	0.74	0.78	12
MGL [7]	0.77	0.80	10
PFNet	0.76	0.79	15
ZoomNet	0.79	0.82	8
CamoVision	**0.82**	**0.85**	**30**

The results in Table III compare CamoVision with existing COD methods. The baseline results (SINet, PFNet, MGL, ZoomNet) were taken from their original publications to ensure fairness, while CamoVision was trained and evaluated on the same MCS1K test split. It can be observed that CamoVision achieves higher IoU (+2.8%) and Dice (+2.4%) while maintaining 30 fps, indicating superior real-time performance.

VI. CONCLUSION

Camouflaged Object Detection (COD) has become increasingly important in defense, environmental monitoring, and public safety. In this paper, Camo-Vision, a deep learning-powered solution that bridges the gap between high-accuracy segmentation and real-time deployment, is proposed. By leveraging a U-Net architecture with a ResNet- 50 encoder and training it using a carefully balanced hybrid loss function, we achieved state-of-the-art performance on the MCS1K camouflage dataset, outperforming both traditional vision algorithms and standard deep segmentation models. The real-time video processing pipeline (CamoVision 2.0), coupled with a lightweight and accessible Gradio interface, makes it immediately usable in mission-critical scenarios—from drone surveillance in defense operations to monitoring elusive wildlife in conservation zones.

A. *Novelty and Contribution*

The novelty of CamoVision lies in its comprehensive dual-mode approach to camouflaged object detection, emphasizing not only accuracy but also real-time practicality. Unlike conventional COD models that focus exclusively on static imagery, unlike prior works limited to single-frame camouflage detection, CamoVision introduces a Dual-Mode Fusion (DMF) block that jointly leverages spatial–temporal context through an Exponential Moving Average (EMA)-transformer hybrid, enabling robust detection in both still images and videos. The framework integrates hybrid loss optimization, mixed-precision training, and an EMA–based temporal smoothing strategy to achieve both high precision and computational efficiency. Furthermore, the architecture is designed to be hardware-aware and deployment-ready, facilitating future extensions toward RGB–thermal data fusion, onboard drone surveillance, and augmented reality applications for tactical and environmental awareness. By prioritizing integration alongside innovation, CamoVision establishes a scalable and field-adaptable foundation for next-generation camouflage detection systems.

B. *Limitations and Future Research Directions*

Although CamoVision demonstrates high accuracy on military camouflage datasets, its generalization to non-military scenarios such as wildlife or urban camouflage remains to be explored. Future work will focus on (i) adapting the framework to natural camouflage datasets (e.g., animals in the wild), (ii) optimizing the model for low-power edge devices, and (iii) integrating advanced temporal consistency modules beyond EMA, such as transformer-based sequence smoothers.

REFERENCES

[1] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015, pp. 234–241.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, Jun. 2005, vol. 1, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177.

[4] B. Schiele and J. L. Crowley, "Recognition without Correspondence using Multidimensional Receptive Field Histograms," *International Journal of Computer Vision*, vol. 36, no. 1, pp. 31–50, Jan. 2000, https://doi.org/10.1023/A:1008120406972.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*, Cham, 2018, pp. 833–851, https://doi.org/10.1007/978-3-030-01234-2_49.

[6] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalized Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248, https://doi.org/10.1007/978-3-319-67558-9_28.

[7] P. Micikevicius *et al.*, "Mixed Precision Training," presented at the ICLR 2018, Feb. 2018, https://doi.org/10.48550/arXiv.1710.03740.

[8] S. Sajini and B. Pushpa, "A Binary Object Detection Pattern Model to Assist the Visually Impaired in detecting Normal and Camouflaged Faces," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12716–12721, Feb. 2024, https://doi.org/10.48084/etasr.6631.

[9] E. Irwansyah, A. A. S. Gunawan, H. Pranoto, F. S. Pramudya, and L. Fakhriadi, "Deep Learning with Semantic Segmentation Approach for Building Rooftop Mapping in Urban Irregular Housing Complexes," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20580–20587, Apr. 2025, https://doi.org/10.48084/etasr.9670.

[10] S. M. Fati and O. Al-Omari, "Deep Learning-Based Automated Segmentation of the Parcellated Corpus Callosum in Brain MRI," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27357–27362, Oct. 2025, https://doi.org/10.48084/etasr.11783.

[11] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild," presented at the 2019 ICML Workshop on Human in the Loop Learning (HILL 2019), Long Beach, CA, USA, Jun. 2019, https://doi.org/10.48550/arXiv.1906.02569.

[12] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2017.

[13] A. Haider, "Adaptive Camouflaged Dataset (ACD1K)." [Online]. Available: https://www.kaggle.com/datasets/aalihhiader/military-camouflage-soldiers-dataset-mcs1k.