WGRNet: Wavelet and Gradient Guided Refinement Network for Multi-Class Military Camouflage Object Segmentation

Vishal Srivastava[a], Monica Bhutani[ab]
[a]School of AI Computing and Multimedia, Lincoln University College, Petaling Jaya, Selangor, Malaysia
[b]Department of Electronics and Communications, Bharati Vidyapeeth's College of Engineering, New Delhi, India

Abstract

Camouflaged object detection is a critical problem in military applications where soldiers, tanks, and trucks are deliberately designed to blend into their surroundings. Traditional computer-vision methods fail when the contrast between the object and background is extremely low. This work addresses the problem of multi-class semantic segmentation of camouflaged military targets across three foreground classes (Soldier, Tank, Truck) using deep learning. Five deep-learning segmentation architectures were implemented and compared: SCPNet, SINetV2, DGNet, EW-CASCADE, and our proposed WGRNet. All models were trained on a three-class subset of the self-constructed MilCamSeg dataset using a common loss function (Focal Loss + Dice Loss) to ensure a fair comparison. WGRNet is a custom architecture that combines the PVTv2-B2 transformer backbone, wavelet-based frequency decomposition, and gradient-weighted attention, specifically designed to handle the low-contrast patterns found in camouflage. Experimental results show that EW-CASCADE achieves the best mean IoU (0.821) and pixel accuracy (0.962), while WGRNet achieves the best Dice overlap (0.896). Both models significantly outperform the other three architectures on all foreground classes, demonstrating that transformer backbones combined with wavelet- and attention-based decoding are highly effective for camouflaged object segmentation.

Keywords: Camouflaged Object Detection, Semantic Segmentation, Wavelet and Gradient Guided Refinement Network, Pyramid Vision Transformer v2

1. Introduction

Military camouflage is made to hide objects in plain sight. Soldiers, tanks, and trucks are painted and textured so that they merge with forests, deserts, rocky land, or mixed backgrounds. In many images, the tar-get does not stand out clearly, and this is exactly what makes the detection task difficult.

The problem has practical importance for defence-related workflows. Current procedures use drone imagery, satellite data, and automated detection, which means that the inability to spot a hidden target carries real operational cost. As camouflage methods continue to improve, detection methods must keep pace.

From a computer-vision standpoint, camouflaged objects are difficult to handle because there is little to no visual separation between the foreground and back-ground. Colour differences are minimal, texture differences are subtle, and lighting makes it hard to tell where the object starts and ends. In such situations, classical solutions such as edge detection or hand-designed features do not generalise well.

Bounding boxes are not sufficient for this kind of ambiguity, so pixel-level approaches are necessary. Se-mantic segmentation satisfies this requirement by as-signing a class label to every pixel, rather than a coarse box around the object.

In this work, we study a multi-class segmentation configuration and make the following contributions:

- We evaluate four classes for a multi-class segmentation problem: Background, Soldier, Tank, and Truck, drawn from a three-foreground-class subset of our self-constructed MilCamSeg dataset.
- We implement and evaluate five segmentation architectures relevant to camouflage research, comparing them under identical training conditions.
- We adopt a common training loss function – Focal Loss combined with Dice Loss – across all five

models to ensure fair comparison.

- We propose WGRNet, a wavelet and gradient-guided refinement network, and show that it achieves the best region-overlap (Dice) performance among all evaluated models.

2. Related Work

Camouflaged object detection (COD) has progressed from hand-crafted, texture- and edge-based feature methods toward deep segmentation networks capable of learning discriminative cues directly from data. Re-cent COD literature has converged on a small number of recurring design ideas: multi-scale context aggregation, attention mechanisms that suppress background noise, gradient- or edge-guided decoding, and, more recently, transformer backbones for long-range context modelling. We review four representative architectures below, each of which we re-implement as a baseline for direct comparison against our proposed WGRNet, followed by a discussion of general multi-class semantic segmentation approaches that inform the encoder-decoder and multi-class prediction design choices made in this work.

SCPNet [1] targets computational efficiency using an EfficientNet-B1 encoder (8.2M parameters). It separates the segmentation problem into two sub-tasks – se-mantic understanding of object existence and spatial localisation of boundaries – solved through a Twin Positioning Module (TPM) and a Spatial Awareness Module (SAM), respectively, and merges the two through a Context Propagation Module (CPM) with a cascaded decoder. Its main strength is a strong accuracy-to-parameter ratio, making it a reasonable choice for resource-constrained deployment. SINetV2 [2] instead adopts a two-stage Search-then-Identification strategy on a Res2Net-50 backbone (25.8M parameters). Receptive Field (RF) blocks gather multi-scale context, a Similarity Aggregation (SA) module clusters regions by feature similarity, and Group-Reversal Attention (GRA) re-cursively refines weak boundaries by suppressing dominant background activations. This staged search-and-refine design is effective at localising ambiguous, low contrast objects, though it relies on a purely convolutional backbone.

DGNet [3] explicitly leverages image gradients as a cue for object boundaries, using an EfficientNet-B4backbone (17.6M parameters) with a dual-stream de-sign that processes semantic and texture/gradient in-formation separately. A Gradient-Induced Transition (GIT) module guides feature exchange between the two streams before a feature-aggregation decoder produces the final mask. Because it treats gradient in-formation as a primary signal rather than an auxiliary one, DGNet is comparatively robust to low colour contrast between object and background, although, like SCPNet and SINetV2, it does not use a trans-former backbone and therefore has a more limited receptive field. EW-CASCADE [4] moves in a different direction, combining a PVTv2-B2 transformer encoder [5] with a wavelet-aware cascaded decoder. An Attention Gate (AG) filters noisy skip-connection features, an Enhanced Efficient Multi-scale Aggregation (EEMA) module fuses spatial and semantic representations, and Wavelet Transform Convolution (WT-Conv) decomposes decoder features into low- and high-frequency sub-bands to preserve edges. In our evaluation, EW-CASCADE achieves the strongest mean IoU (0.821) and pixel accuracy (0.962) among all baselines, motivating the wavelet-based direction we pursue in WGRNet. However, EW-CASCADE does not explicitly model boundary gradients, and conversely DGNet demonstrates that gradient-based cues are effective but does not exploit transformer backbone or frequency decomposition. WGRNet is positioned to combine both ideas – frequency-domain decomposition and gradient-guided attention – on top of a transformer backbone, which we detail in Section 3.

Complementing these encoder-decoder and transformer-based approaches, our own earlier work [6] explored a metric-learning perspective on the same problem, extending a U-Net backbone with a Pixel-Wise Metric Learning (PWML) branch that learns dense per-pixel embeddings jointly with the segmentation mask via a pixel-wise contrastive loss. On the single-class Adaptive Camouflaged Dataset (ACD1K), this approach improved structural and boundary-aware metrics over both U-Net and U-Net++ baselines, with S-measure rising from 0.5130 to 0.5501, E-measure from 0.5036 to 0.5352, and weighted F-measure from 0.3054 to 0.3550. While that work established that discriminative pixel-level

embeddings are beneficial for camouflage segmentation, it remained a single-class, binary-mask formulation without frequency-domain or gradient-based refinement, both of which are central to the multi-class WGRNet design introduced in Section 3.

Beyond COD specifically, this work also builds on the broader multi-class semantic segmentation literature, since our task requires assigning each pixel to one of four classes (Background, Soldier, Tank, Truck) rather than the binary foreground/background masks common in most COD benchmarks. Encoder-decoder architectures with skip connections, established by U-Net [7], remain the dominant template for dense pixel-wise prediction and underlie the cascaded-decoder de-signs used by all five models evaluated in this pa-per. DeepLabv3+ [8] introduced atrous spatial pyramid pooling to aggregate multi-scale context without losing spatial resolution, a concern that is directly analogous to the receptive-field and multi-scale aggregation modules used in SINetV2 and EW-CASCADE. PSPNet [9] similarly demonstrated that pyramid pooling over multiple sub-regions improves scene parsing accuracy in cluttered, multi-object scenes, which is relevant to our set-ting since Soldier, Tank, and Truck instances frequently co-occur with cluttered natural backgrounds. More recently, transformer-based segmentation models such as SegFormer [10] have shown that a lightweight all-MLP decoder paired with a hierarchical transformer encoder can outperform heavier CNN decoders on multi-class benchmarks, an architectural pattern that directly in-formed the design of our companion MilCamSeg five-class segmentation model. Mask2Former [11] further unifies semantic, instance, and panoptic segmentation under a single mask-classification framework using masked attention, illustrating an alternative, query-based route to multi-class prediction that does not rely on per-pixel argmax decoding as used in this paper. Collectively, this body of work confirms that multi-scale context aggregation and transformer-based en-coders – the same ingredients used in EW-CASCADE and WGRNet – are effective well beyond the specific case of camouflage, while also highlighting that most multi-class segmentation research targets natural scenes rather than the extreme low-contrast conditions characteristic of military camouflage, which is the gap this pa-per addresses.

3. Proposed Method: WGRNet

We propose WGRNet, the Wavelet and Gradient Guided Refinement Network, which integrates frequency separation, gradient guidance, and transformer-based global context to improve camouflaged object segmentation. WGRNet contains 32.1M parameters and achieves the highest Dice score (0.896) among all evaluated models. The design is motivated by two complementary observations. In the frequency domain, camouflage patterns mimic low-frequency background structures, but high-frequency edge and texture variations often remain visible at finer scales; wavelet decomposition isolates these components to enhance foreground-background separation. In the gradient domain, intensity transitions near boundaries yield subtle gradient cues that aid contour localisation in low-contrast regions where colour and texture cues alone are insufficient.

Building on these two observations, WGRNet is composed of four components. A PVTv2-B2 backbone [5] captures multi-scale contextual features across four resolution levels (f_1 to f_4). These features are then passed through wavelet decomposition and WT-Conv stage, which splits scale, features into subbands (LL, LH, HL, HH) so that fine textures and overall structures are processed independently. In parallel, a Gradient-Weighted Attention (GWA) module converts intermediate features into gradient maps to guide attention toward object boundaries. Finally, a cascaded decoder with Enhanced Attention Gates (EAG) applies gating across top-down fusion levels to suppress background noise while progressively refining the segmentation map. Figure 1 shows the full pipeline: multi-scale features from the transformer backbone are passed into the WT-Conv and GWA modules before being progressively decoded through EAG-guided cascaded connections to produce the final four-class segmentation mask.

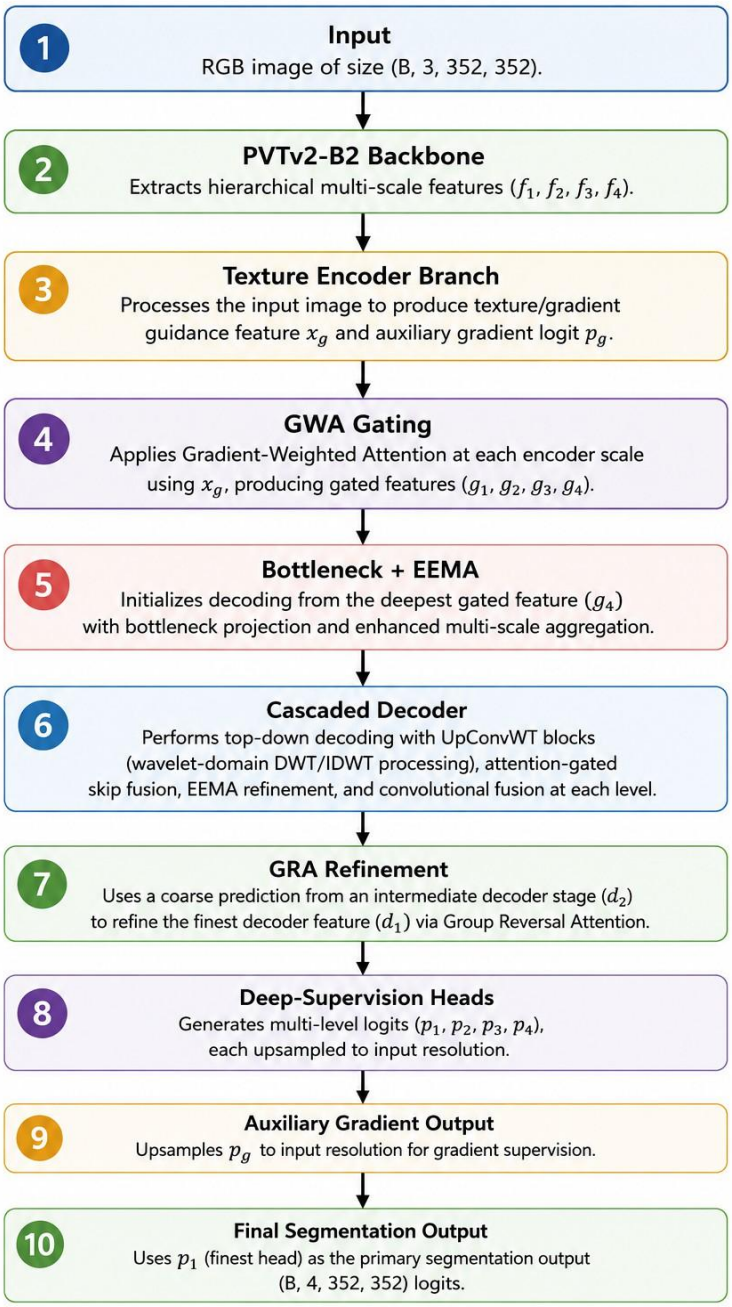


Figure 1: WGRNet Pipeline.

4. Dataset

All five models are trained and evaluated on a three-foreground-class subset of MilCamSeg, our self-constructed military camouflage dataset. The full Mil-CamSeg dataset defines five foreground classes (Sol-dier, Military Truck, Military Tank, Rocket Launcher, Landmine); for this work, only the **Soldier**, **Tank**, and **Truck** classes are used, together with the Background class, giving the four-class (K+1=4) segmentation setting used throughout this paper. The dataset is organised into camouflaged, non-camouflaged, and background-only subsets, each containing paired image and mask folders (the background-only subset has no masks). Ground-truth annotations are provided as per-pixel, class-colour-encoded PNG masks rather than COCO-style JSON annotations, which directly supports the semantic segmentation formulation used here.



Figure 2: Qualitative comparison on Soldier sample.

5. Common Loss Function

To ensure fair comparison, all five architectures were trained using an identical objective function combining Focal Loss and Dice Loss with deep supervision. Focal Loss [12] addresses class imbalance by focusing training on hard, uncertain pixels near object boundaries rather than easy background pixels:

$$L_{focal} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \qquad \gamma = 2.0 \tag{1}$$

Where p_t represents predicted probability and α_t balances the class weights. Complementing this, Dice Loss [13] measures region-level overlap, driving the model to accurately capture complete object shapes:

$$L_{dice} = 1 - \frac{2\sum p_i g_i + \epsilon}{\sum p_i + \sum g_i + \epsilon} \tag{2}$$

Where p_i and g_i denote predicted and ground-truth pixel values, with $\epsilon = 10^{-5}$. The total loss combines both terms equally:

$$L_{total} = L_{focal} + L_{dice} \tag{3}$$

Deep supervision [14] applies this total loss at intermediate decoder stages as well as the final output to stabilise gradient flow.

6. Implementation Details

All five models were trained on a single NVIDIA Tesla P100 GPU provided via the Kaggle notebook environment. Input images are resized to 352 × 352 and normalised prior to being passed through each network; pixel wise argmax over the output logits yields the final multi-class prediction map.

7. Results and Evaluation

Figures 2, 3, and 4 illustrate visual comparisons across test samples for each foreground target. Turning to quantitative results, models are evaluated using Mean Intersection over Union (mIoU), Dice Similarity Coefficient, and Pixel Accuracy. Table 1 re-ports performance across all models. EW-CASCADE achieves the top mIoU (0.821) and pixel accuracy (0.962), WGRNet attains the highest Dice score (0.896), showing strong region-overlap performance, and SCP-Net offers the most lightweight footprint at 8.2M parameters. Table 2 breaks these results down further, reporting per-class IoU scores across the four categories.



Figure 3: Qualitative comparison on Tank sample.



Figure 4: Qualitative comparison on Truck sample.

Table 1: Model comparison: overall metrics

Model	mIoU	Dice	Pixel Acc.	Params (M)
SCPNet	0.752	0.797	0.943	8.2
SINetV2	0.726	0.751	0.934	25.8
DGNet	0.709	0.714	0.929	17.6
EW-CASCADE	0.821	0.848	0.962	–
WGRNet	0.816	0.896	0.961	32.1

Table 2: Per-class IoU comparison

Model	BG	Soldier	Tank	Truck	Mean
SCPNet	0.945	0.663	0.637	0.763	0.752
SINetV2	0.937	0.642	0.595	0.729	0.726
DGNet	0.932	0.627	0.573	0.706	0.709
EW-CASCADE	0.962	0.712	0.759	0.851	0.821
WGRNet	0.963	0.731	0.731	0.841	0.816

8. Limitations

Several limitations should be considered when interpreting these results. First, all models are evaluated at a single fixed input resolution (352 × 352); performance at higher resolutions, which may better preserve fine

camouflage textures, has not been assessed. Second, the evaluation is restricted to three foreground classes (Soldier, Tank, Truck) drawn from MilCamSeg; the full five-class setting (including Rocket Launcher and Land-mine) is addressed separately and is not directly comparable to the results reported here. Third, inference speed and computational cost (e.g. FLOPs, latency) were not benchmarked, so the practical deployability of WGR-Net relative to the lighter-weight SCPNet remains an open question. Fourth, WGRNet's improvement is concentrated in Dice overlap rather than mIoU or pixel ac-curacy, where EW-CASCADE remains stronger, indicating that the wavelet-plus-gradient design trades some boundary precision for region-level completeness; the underlying cause of this trade-off is not yet fully characterised. Finally, all models were trained and evaluated on a single self-constructed dataset; cross-dataset generalisation to unseen terrains, lighting conditions, or camouflage patterns has not been tested.

9. Conclusion

This paper evaluated deep learning models for multi-class segmentation of camouflaged military targets (Background, Soldier, Tank, and Truck) on a three-class subset of the MilCamSeg dataset. Five architectures were evaluated under a unified training framework using combined Focal and Dice losses.

Key takeaways include:

- Transformer backbones (PVTv2-B2) better capture global context than standard CNN backbones in low-contrast environments.
- Wavelet decomposition offers valuable frequency-domain representations that assist in isolating camouflage textures from backgrounds.
- Gradient guidance improves edge boundary definition.

EW-CASCADE yielded the highest overall mIoU (0.821), while WGRNet achieved the highest Dice overlap (0.896). Future work will address the limitations noted in Section 8, including inference-cost benchmarking and evaluation on the full five-class MilCamSeg set-ting.

References

[1] J. Ren, Q. Zhang, B. Kang, Y. Zhong, M. He, Y. Ge, H. Bi, Semantic-spatial guided con-text propagation network for camouflaged object detection, Applied Intelligence DOI: 10.1007/s10489-025-06264-0 (2025).

[2] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (10) (2022) 6024–6042.

[3] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, L. V. Gool, Deep gradient learning for efficient camouflaged object detection, Machine Intelligence Research 20 (2023) 92–108.

[4] J. Li, Y. Liu, J. Wu, J. Yang, S. Mei, Multiscale feature attention and edge refinement for improved camouflaged locust segmentation, Applied Intelligence DOI: 10.1007/s10489-026-07152-x (2026).

[5] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[6] A. Kumar, V. Srivastava, Enhancing military camouflaged object detection using convolutional neural networks and deep metric learning, Procedia Computer Science 283 (2026) 4972–4981.

[7] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, ECCV (2018).

[9] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: CVPR, 2017.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, in: NeurIPS, 2021.

[11] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: CVPR, 2022.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[13] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: International Conference on 3D Vision (3DV), 2016, pp. 565–571.

[14] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: International Conference on Artificial Intelligence and Statistics (AIS-TATS), 2015, pp. 562–570.