RESEARCH ARTICLE

SemanticBlur: Semantic-Aware Attention Network with Multi-Scale Feature Refinement for Defocus Blur Detection

Asad Ullah Haider[1], **Alexandros Gazis**[2,3] **and Faryal Zahoor**[4,*]

[1] Ilmenau University of Technology, Ilmenau 98693, Germany
[2] Democritus University of Thrace, Xanthi 67100, Greece
[3] Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom
[4] BRAINS Institute Peshawar, Peshawar 25000, Pakistan

Abstract

Defocus blur detection is essential for computational photography applications, but existing methods struggle with accurate blur localization and boundary preservation. We propose SemanticBlur, a deep learning framework which integrates semantic understanding with attention mechanisms for robust defocus blur detection. Our semantic-aware attention module combines channel attention, spatial attention, and semantic enhancement to leverage high-level features for low-level feature refinement. The architecture employs a modified ResNet-50 backbone with dilated convolutions that preserves spatial resolution while expanding receptive fields, coupled with a feature pyramid decoder using learnable fusion weights for adaptive multi-scale integration. A combined loss function balancing binary cross-entropy and structural similarity achieves both pixel-wise accuracy and structural coherence. Extensive experiments on four benchmark datasets (CUHK, DUT, CTCUG, EBD) demonstrate state-of-the-art (SOTA) performance, with ablation studies confirming that semantic enhancement provides the most significant gains while maintaining computational efficiency. SemanticBlur generates visually coherent detection maps with sharp boundaries, validating its practical applicability for real-world deployment.

Keywords: blur detection, semantic segmentation, attention mechanisms, deep neural networks, image analysis, multi-scale processing.

1 Introduction

Defocus blur detection (DBD) aims to identify objects or regions with blurred pixels in videos or images. It is often caused by limitations in the imaging system when scene areas fall outside the camera's focal plane. It serves as a crucial preprocessing step for numerous computer vision applications, including image refocusing [1], blur reconstruction [2], depth estimation [3], image deblurring [4, 5], and object detection [6–8]. Traditional DBD methods

Academic Editor:
Jing Na

Submitted: 05 September 2025
Accepted: 23 October 2025
Published: 19 February 2026

Vol. 3, **No.** 1, 2026.
10.62762/TIS.2025.879161

***Corresponding author:**
✉ Faryal Zahoor
faryal.zahoorjan@gmail.com

Citation
Haider, A. U., Gazis, A., & Zahoor, F. (2026). SemanticBlur: Semantic-Aware Attention Network with Multi-Scale Feature Refinement for Defocus Blur Detection. *ICCK Transactions on Intelligent Systematics*, 3(1), 21–31.

relied heavily on hand-crafted image priors based on empirical observations [9–11], utilizing manually extracted low-level features such as gradients [24], local binary patterns [27], and frequency domain analysis [26]. However, these approaches face significant limitations as they are challenging to design, often lack generalizability, and can only succeed in simple scenes. Most critically, they cannot capture high-level semantic information necessary to distinguish foreground objects from complex backgrounds and to accurately detect boundary details, especially in gradient regions.

To address these limitations, current state-of-the-art methods [12–15] utilize deep learning networks to implicitly learn more general priors by extracting defocused blur image semantics and texture information from large-scale datasets. The enhanced efficacy of Convolutional Neural Networks (CNNs) in DBD primarily hinges upon the complexity of their model architecture. Various network modules have been devised for DBD, including residual learning [19, 20], generative adversarial learning [18], attentional mechanisms [21], dense connections [22], and depth distillation [14]. In semantic segmentation tasks [16, 17], deep learning approaches consistently leverage semantic information of target objects to achieve distinctive feature segmentation in prominent regions. Similarly, in DBD, these methods effectively mitigate detection challenges in simple homogeneous regions. Jonna et al. [18] further address limitations by employing adversarial mechanisms to exploit weak correlations in semantic content of defocused blurry regions, thereby realizing a self-supervised DBD methodology. Some methods [22, 32] employ integrated learning approaches that combine results from multiple network branches to improve feature diversity, while others [19, 23] utilize different layers of information to complement each other and enhance feature representation. However, due to the susceptibility of low-level information to background noise interference, especially in transition edge regions where similarity discrimination ability is weak, it becomes essential to utilize high-level semantic information to guide refinement of low-level features.

Our work makes the following main contributions:

- We propose SemanticBlur, a novel semantic-aware attention module integrating channel attention, spatial attention, and semantic enhancement to effectively leverage high-level semantic features for guiding low-level feature refinement in defocus blur detection.
- We introduce a hierarchical feature extraction and fusion architecture consisting of: (i) a modified ResNet-50 backbone with strategically placed dilated convolutions (rates: 2, 4, 8) in the final stages to preserve spatial resolution while capturing multi-scale contextual information, and (ii) a novel feature pyramid decoder with learnable fusion weights $(\beta_1, \beta_2, \beta_3, \beta_4)$ that adaptively combines features across scales, enabling effective capture of both fine-grained boundary details and global semantic context.
- We introduce a combined loss function balancing binary cross-entropy and structural similarity loss $(\lambda = 0.1)$ to achieve both pixel-wise accuracy and structural preservation, with optimal weighting validated through comprehensive ablation studies across multiple benchmark datasets.

The remainder of this paper is organized as follows: Section 2 reviews related work in defocus blur detection, covering hand-crafted feature-based methods, deep learning approaches, and attention mechanisms. Section 3 presents our proposed methodology, detailing the backbone feature extraction, semantic-aware attention module, and decoder architecture. Section 4 describes our experimental setup, including datasets, implementation details, evaluation metrics, and presents comprehensive results including ablation studies and comparative analysis with SOTA methods. Section 5 concludes the paper with a summary of contributions and discussion of future research directions.

2 Related Work

Defocus blur is a common phenomenon in natural images, mainly caused by the limited depth of field of cameras or by unfavorable imaging conditions. Research on defocus blur detection (DBD) can be broadly categorized into two major directions: approaches relying on hand-crafted features and those leveraging deep learning.

2.1 Hand-crafted Feature-based Methods

Early studies predominantly relied on low-level image cues to distinguish blurred from sharp regions. A common observation was that defocus blur weakens object boundaries, leading to smoother edge transitions. To exploit this, Shi et al. [24] integrated multiple blur indicators such as gradients, frequency information, and local filters into a multi-scale framework. Su et al. [25] instead analyzed the singular value decomposition (SVD) of images and introduced an α-channel constraint to discriminate between defocus and motion blur. Tang et al. [26] proposed a blur metric derived from logarithmically averaged spectral residuals and refined the resulting blur map by exploiting correlations among neighboring regions. Yi and Eramian [27] presented a sharpness measure based on localized binary patterns (LBP), which enabled separation of in-focus and blurred regions. More recently, Golestaneh et al. [9] developed HiFST, a method that combines high-frequency multi-scale fusion with gradient-based transforms for spatially adaptive blur detection. Although these approaches demonstrated effectiveness in relatively simple cases, their reliance on local cues limited their generalization to complex natural scenes where semantic context is crucial.

2.2 Deep Learning Based Methods

In recent years, deep learning based methods have achieved superior performance in the field of DBD due to the powerful learning capability of CNN. Studies, such as STNet [28], demonstrate that combining transformer-based multi-scale feature extraction with effective attention mechanisms can achieve high performance while maintaining a lightweight architecture. Park et al. [29] pioneered a unified framework where deep CNN features were fused with conventional descriptors and processed through an FCN for blur segmentation. Zhao et al. [30] introduced an end-to-end CNN that integrates both semantic features and low-level details, progressively refining blur maps in a scale-sensitive manner. Zeng et al. [31] further combined CNN-extracted features with principal component analysis on image superpixels, followed by iterative refinement to improve spatial consistency.

To address the limited diversity of CNN feature representations, Zhao et al. [22] introduced a cross-ensemble strategy with a cross negative correlation loss to train multiple detectors jointly, promoting diversity in feature representations. Building on this, an encoder-feature integration network was later introduced [32], generating multiple sets of convolutional features from a single encoder to enhance representation richness. Beyond CNNs, recent studies highlight the benefits of transformer architectures and attention mechanisms. Similarly, approaches such as GPRNet [33] integrate multi-level attention, edge-awareness, and uncertainty modeling to better capture fine boundaries and handle challenging homogeneous regions.

2.3 Attention Mechanisms in DBD

Attention mechanisms have become a cornerstone in computer vision by focusing on critical information while suppressing redundant data. The prevalent attention modules propagate channel dimensions [34], spatial dimensions, or the coexistence of both [35]. These mechanisms, including Convolutional Block Attention Module (CBAM), effectively combine channel and spatial attention to enhance feature representation, as demonstrated in various domains such as deepfake detection [36]. For DBD, most existing approaches only consider the attention mechanism as a module for auxiliary performance enhancement [37]. Similarly, attention-driven strategies have shown remarkable success in SOD by refining spatial and contextual details through multi-stage processing and progressive feature fusion [38]. Both Zhao et al. [39] and Chai et al. [40] integrated transformers into backbone networks for DBD. However, the weak correlation between DBE and the semantic information of images [41] makes it challenging to achieve satisfactory results when dealing with challenging images using purely attention-based approaches.

3 Proposed Methodology

In this section, we present the main components of our architecture ref: a backbone feature extraction phase for learning hierarchical representations, an attention module for emphasizing semantically relevant features, and a final decoder for generating precise blur maps. The overall architecture is designed to effectively capture both low-level textural details and high-level semantic information essential for accurate defocus blur detection.

3.1 Backbone Feature Extraction

The backbone network serves as the foundation of our approach, responsible for extracting multi-level feature representations from input images. We adopt a ResNet-50 architecture as our feature extractor due to its proven effectiveness in capturing hierarchical features while maintaining computational efficiency. The input image $I \in \mathbb{R}^{H \times W \times 3}$ is processed through a series of convolutional blocks to generate feature maps at different scales:

$$F_i = \mathcal{B}_i(F_{i-1}), \quad i = 1, 2, 3, 4 \tag{1}$$

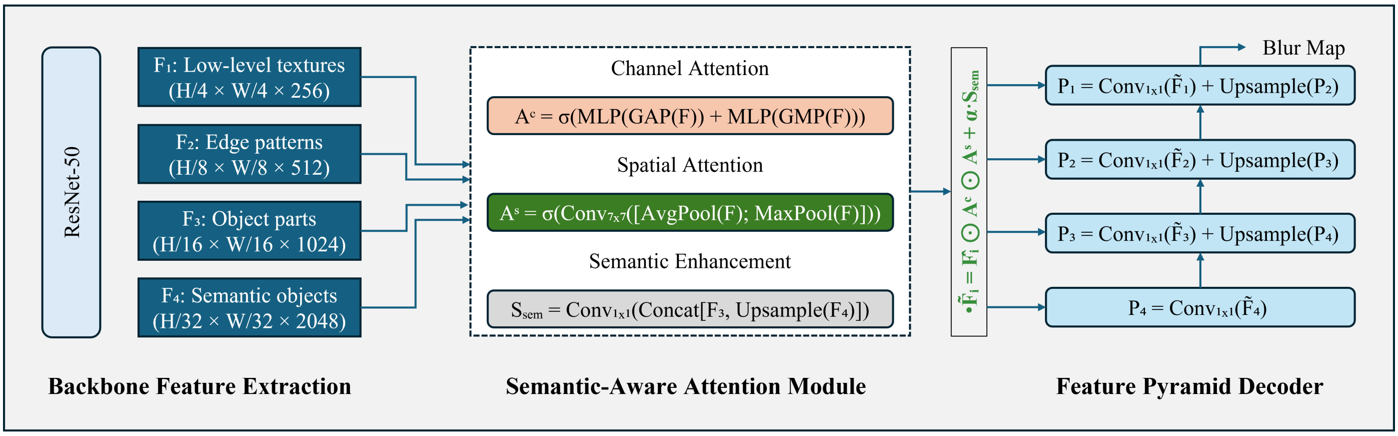
ResNet-50
F₁: Low-level textures (H/4 × W/4 × 256)
F₂: Edge patterns (H/8 × W/8 × 512)
F₃: Object parts (H/16 × W/16 × 1024)
F₄: Semantic objects (H/32 × W/32 × 2048)
Channel Attention
$A^c = \sigma(\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F)))$
Spatial Attention
$A^s = \sigma(\mathrm{Conv}_{7x7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)]))$
Semantic Enhancement
$S_{sem} = \mathrm{Conv}_{1x1}(\mathrm{Concat}[F_3, \mathrm{Upsample}(F_4)])$
$\bullet\tilde{F}_i = F_i \odot A^c \odot A^s + \alpha \cdot S_{sem}$
Blur Map
$P_1 = \mathrm{Conv}_{1x1}(\tilde{F}_1) + \mathrm{Upsample}(P_2)$
$P_2 = \mathrm{Conv}_{1x1}(\tilde{F}_2) + \mathrm{Upsample}(P_3)$
$P_3 = \mathrm{Conv}_{1x1}(\tilde{F}_3) + \mathrm{Upsample}(P_4)$
$P_4 = \mathrm{Conv}_{1x1}(\tilde{F}_4)$
Backbone Feature Extraction
Semantic-Aware Attention Module
Feature Pyramid Decoder

Figure 1. Overview of the proposed **SemanticBlur** architecture. A modified ResNet-50 backbone extracts multi-scale features (F_1–F_4) using dilated convolutions. The *Semantic-Aware Attention Module* refines these features through channel attention ($\mathbf{A}^c$), spatial attention ($\mathbf{A}^s$), and semantic enhancement (Ssem), which fuses high-level features F_3 and F_4. The *Feature Pyramid Decoder* progressively upsamples and merges the attended features with learnable fusion weights (β_i) to produce the final blur map at the original resolution.

where $\mathcal{B}_i$ represents the i-th residual block, and $F_0 = I$ is the input image. The extracted features $\{F_1, F_2, F_3, F_4\}$ have spatial resolutions of $\{\frac{H}{4}, \frac{H}{8}, \frac{H}{16}, \frac{H}{32}\}$ respectively, with corresponding channel dimensions of $\{256, 512, 1024, 2048\}$. To enhance the representational power of the backbone features, we incorporate dilated convolutions in the final stages to maintain spatial resolution while expanding the receptive field:

$$F_i^{dilated} = \text{DilatedConv}(F_i, d_i) \tag{2}$$

where d_i represents the dilation rate for the i-th feature level. This design enables the network to capture both fine-grained local patterns indicative of blur boundaries and broader contextual information necessary for semantic understanding. Furthermore, we apply feature normalization and activation functions to stabilize training and improve convergence:

$$\hat{F}_i = \text{ReLU}(\text{BatchNorm}(F_i^{dilated})) \tag{3}$$

The normalized features $\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\}$ are then passed to the attention module for further refinement.

3.2 Semantic-Aware Attention Module

This module is designed to selectively emphasize features that are most relevant for defocus blur detection while suppressing noise and irrelevant information. Our approach incorporates both channel attention and spatial attention mechanisms to capture complementary aspects of the feature representations.

3.2.1 *Channel Attention Mechanism*

The channel attention mechanism learns to weight different feature channels based on their importance for blur detection. Given a feature map $\hat{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, we first apply global average pooling and global max pooling to capture channel-wise statistics:

$$\mathbf{f}_{avg} = \text{GAP}(\hat{F}_i) = \frac{1}{H_i \times W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} \hat{F}_i(h, w, :) \tag{4}$$

$$\mathbf{f}_{max} = \text{GMP}(\hat{F}_i) = \max_{h,w} \hat{F}_i(h, w, :) \tag{5}$$

The channel attention weights are computed through a shared multi-layer perceptron (MLP) followed by element-wise addition and sigmoid activation:

$$\mathbf{A}_c = \sigma(\text{MLP}(\mathbf{f}_{avg}) + \text{MLP}(\mathbf{f}_{max})) \tag{6}$$

where $\mathbf{A}_c \in \mathbb{R}^{C_i}$ represents the channel attention weights.

3.2.2 *Spatial Attention Mechanism*

The spatial attention mechanism identifies regions that require more focus during the blur detection process. We concatenate the channel-wise average and maximum pooled features along the channel dimension:

$$\mathbf{S} = \text{Concat}[\text{AvgPool}_c(\hat{F}_i), \text{MaxPool}_c(\hat{F}_i)] \tag{7}$$

where AvgPool_c and MaxPool_c denote average and max pooling operations along the channel dimension, resulting in $\mathbf{S} \in \mathbb{R}^{H_i \times W_i \times 2}$. The spatial attention map is generated through a convolutional layer followed by sigmoid activation:

$$\mathbf{A}_s = \sigma(\text{Conv}_{7\times7}(\mathbf{S})) \tag{8}$$

where $\mathbf{A}_s \in \mathbb{R}^{H_i \times W_i \times 1}$ represents the spatial attention weights.

3.2.3 *Feature Refinement*

The final attended features are obtained by applying both channel and spatial attention weights:

$$\tilde{F}_i = \hat{F}_i \odot \mathbf{A}_c \odot \mathbf{A}_s \tag{9}$$

where $\odot$ denotes element-wise multiplication with appropriate broadcasting. Additionally, we incorporate a semantic enhancement module that leverages high-level semantic information to guide the attention mechanism:

$$\mathbf{S}_{sem} = \text{Conv}_{1\times1}(\text{Concat}[\tilde{F}_3, \tilde{F}_4]) \tag{10}$$

$$\tilde{F}_i^{final} = \tilde{F}_i + \alpha \cdot \text{Upsample}(\mathbf{S}_{sem}) \tag{11}$$

where α is a learnable parameter controlling the contribution of semantic information.

3.3 Final Decoder

The decoder network is responsible for integrating multi-level attended features and generating the final defocus blur detection map. Our decoder employs a feature pyramid network (FPN) structure with lateral connections to effectively combine features from different scales.

3.3.1 *Feature Pyramid Construction*

We construct the feature pyramid by progressively upsampling higher-level features and combining them with lower-level features through lateral connections:

$$\mathbf{P}_4 = \text{Conv}_{1\times1}(\tilde{F}_4^{final}) \tag{12}$$

$$\mathbf{P}_i = \text{Conv}_{1\times1}(\tilde{F}_i^{final}) + \text{Upsample}(\mathbf{P}_{i+1}), \quad i = 3, 2, 1 \tag{13}$$

where $\{\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\}$ represent the pyramid features at different scales.

3.3.2 *Multi-Scale Feature Fusion*

To leverage information from all scales, we apply a multi-scale fusion mechanism:

$$\mathbf{P}_{fused} = \sum_{i=1}^{4} \beta_i \cdot \text{Resize}(\mathbf{P}_i, H \times W) \tag{14}$$

where β_i are learnable fusion weights and Resize operation resizes all features to the original input resolution.

3.3.3 *Final Prediction*

The final blur detection map is generated through a series of convolutional layers with progressive channel reduction:

$$\mathbf{M}_{blur} = \text{Conv}_{3\times3}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(\mathbf{P}_{fused}))) \tag{15}$$

where the final output $\mathbf{M}_{blur} \in \mathbb{R}^{H \times W \times 1}$ represents the probability map of defocus blur regions.

3.3.4 *Loss Function*

We employ a combination of binary cross-entropy loss and structural similarity loss to train the network:

$$\mathcal{L}_{total} = \mathcal{L}_{BCE}(\mathbf{M}_{blur}, \mathbf{G}) + \lambda \mathcal{L}_{SSIM}(\mathbf{M}_{blur}, \mathbf{G}) \tag{16}$$

where $\mathbf{G}$ is the ground truth blur mask, $\mathcal{L}_{BCE}$ is the binary cross-entropy loss, $\mathcal{L}_{SSIM}$ is the structural similarity loss, and λ is a balancing parameter.

4 Experiments

4.1 Experimental Setup

To comprehensively evaluate the performance of our proposed SemanticBlur network, we conduct extensive experiments on multiple benchmark datasets following standard evaluation protocols [20, 42]. Our experimental design employs a single dataset for training and multiple datasets for testing to assess the generalization capability of our method across diverse scenarios.

Training Dataset: CUHK Dataset [24]: We utilize the CUHK dataset for training our SemanticBlur network. This dataset comprises 704 high-quality images with pixel-wise defocus blur annotations. Following standard practice, we employ 604 images for training (CUHK-TR) and reserve the remaining 100 images for testing (CUHK-TE).

Testing Datasets: We evaluate our model on four benchmark datasets to assess generalization and robustness. CUHK-TE [24] provides 100 test images while the DUT dataset [30] contains 500 images with varied blur patterns and scene types. CTCUG [20], contains 150 images, presenting particularly difficult cases with complex blur transitions. EBD [42] is the largest benchmark with 1,605 high-resolution images covering diverse scenarios including macro shots, portraits, indoor and outdoor scenes with pixel-level annotations.

4.1.1 *Implementation Details*

All experiments are conducted using PyTorch framework on NVIDIA RTX 3090 GPUs. The model

Table 1. Ablation study comparing different backbone architectures for SemanticBlur. Best results are **bolded**. MAE↓ indicates lower is better; F$_{\beta}$↑ and IoU↑ indicate higher is better.

Backbone	**CUHK-TE**			**DUT**			**CTCUG**		
	MAE↓	F$_{\beta}$↑	IoU↑	MAE↓	F$_{\beta}$↑	IoU↑	MAE↓	F$_{\beta}$↑	IoU↑
VGG-16	0.052	0.921	0.891	0.089	0.903	0.862	0.098	0.841	0.823
ResNet-34	0.043	0.957	0.932	0.078	0.925	0.884	0.087	0.863	0.851
ResNet-50	**0.039**	**0.969**	**0.951**	**0.067**	**0.942**	**0.909**	**0.086**	**0.887**	**0.894**
EfficientNet-B4	0.041	0.961	0.938	0.073	0.934	0.895	0.081	0.872	0.867

is trained for 100 epochs with a batch size of 6 images. We employ the Adam optimizer with an initial learning rate of 1×10^{-4}. The learning rate is adjusted using polynomial decay strategy with a decay factor $\gamma = 0.9$. Gradient clipping is applied with a maximum norm of 0.5 to ensure stable training. For data augmentation and preprocessing, images are resized to 320×320 pixels during training.

4.1.2 Evaluation Metrics

Following standard evaluation protocols in defocus blur detection, we employ multiple metrics to comprehensively assess model performance. We use F-measure (F$_{\beta}$) as the harmonic mean of precision and recall with $\beta^2 = 0.3$ to evaluate the overall detection accuracy. Mean Absolute Error (MAE) measures the average pixel-wise absolute difference between prediction and ground truth, providing insight into pixel-level accuracy. Intersection over Union (IoU) quantifies the overlap between predicted and ground truth blur regions, indicating how well the model captures the spatial extent of defocused areas.

4.2 Ablation Studies

To validate the effectiveness of our proposed SemanticBlur network and understand the contribution of each component, we conduct comprehensive ablation studies analyzing backbone architectures, progressive module integration, and loss function combinations.

4.2.1 Backbone Architecture Analysis

We evaluate different backbone networks to determine the optimal feature extraction architecture for defocus blur detection. Table 1 presents the comparative performance of four popular backbone architectures across three benchmark datasets. The results demonstrate clear performance differences across architectures. VGG-16, despite its simplicity, achieves reasonable performance with MAE values of 0.052, 0.089, and 0.098 on CUHK-TE, DUT, and CTCUG respectively. However, its lack of skip connections and limited representational capacity becomes evident when handling complex blur patterns. ResNet-34 shows substantial improvements over VGG-16, with MAE reductions on CUHK-TE and DUT, highlighting the importance of residual connections for gradient flow and feature learning. ResNet-50 emerges as the optimal backbone, achieving the best performance across all datasets with MAE values of 0.039, 0.067, and 0.086 for CUHK-TE, DUT, and CTCUG respectively. The deeper architecture with bottleneck blocks provides superior feature representation while maintaining computational efficiency. Interestingly, EfficientNet-B4, despite its modern compound scaling approach, performs slightly worse than ResNet-50 with MAE values of 0.041, 0.073, and 0.081. While EfficientNet demonstrates competitive performance, particularly on the challenging CTCUG dataset, ResNet-50's established architecture and proven effectiveness in dense prediction tasks make it the preferred choice for our semantic-aware attention

Table 2. Progressive ablation study of attention components in SemanticBlur. Components are added sequentially to ResNet-50 baseline. Best results are **bolded**.

Configuration	**CUHK-TE**			**DUT**			**CTCUG**			**Params (M)**
	MAE↓	F$_{\beta}$↑	IoU↑	MAE↓	F$_{\beta}$↑	IoU↑	MAE↓	F$_{\beta}$↑	IoU↑	
Baseline (ResNet-50 only)	0.058	0.891	0.863	0.095	0.876	0.834	0.124	0.798	0.756	23.5
+ Channel Attention	0.049	0.924	0.902	0.081	0.912	0.871	0.103	0.837	0.802	24.2
+ Spatial Attention	0.043	0.951	0.928	0.074	0.928	0.889	0.089	0.859	0.831	24.8
+ Semantic Enhancement	**0.039**	**0.969**	**0.951**	**0.067**	**0.942**	**0.909**	**0.086**	**0.887**	**0.894**	25.3

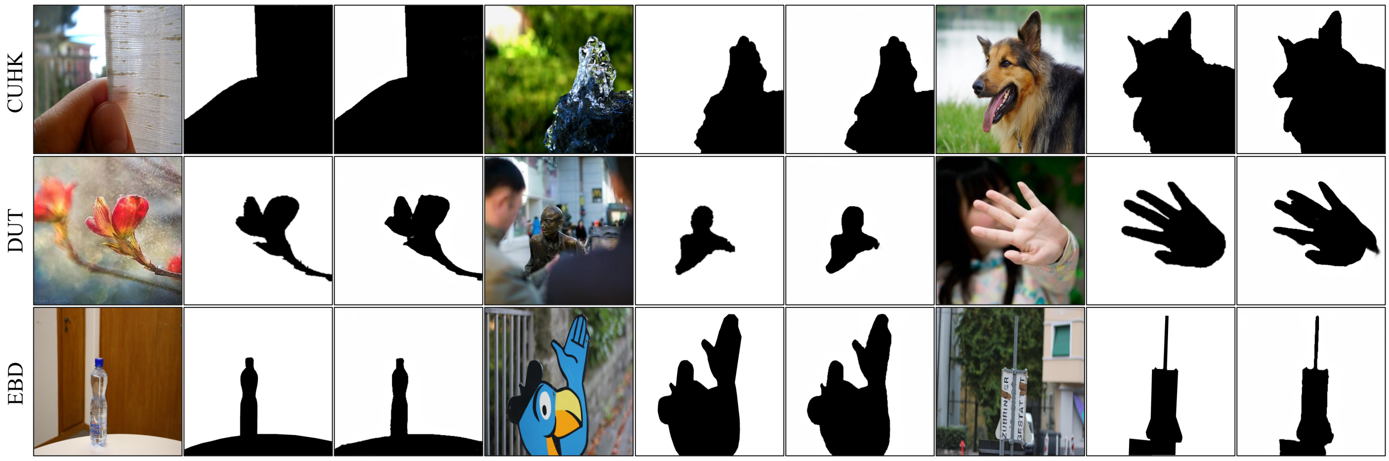

Figure 2. Qualitative results of **SemanticBlur** on representative samples from the CUHK-TE, DUT, and EBD datasets. For each row, from left to right: input image, ground truth annotation, and the predicted defocus blur map. The proposed method effectively localizes defocused regions while maintaining sharp boundary details across diverse photographic scenarios.

framework.

4.2.2 *Progressive Module Integration*

We systematically add attention components to a baseline ResNet-50 to isolate each module's contribution. Table 2 shows results across three datasets with parameter counts. The baseline (23.5M parameters) achieves MAE values of 0.058, 0.095, and 0.124 on CUHK-TE, DUT, and CTCUG. Adding channel attention cuts MAE by 15.5% and 14.7% on CUHK-TE and DUT with just 0.7M additional parameters, showing that global pooling operations effectively weight channel importance. Spatial attention further reduces MAE by 12.2% and 8.6% on these datasets while adding 0.6M parameters, successfully localizing blur boundaries and transition zones.

Semantic enhancement yields the largest gains, reaching MAE of 0.039, 0.067, and 0.086 with improvements of 16.3%, 9.5%, and 3.4% over spatial attention alone. This module uses high-level features F3 and F4 to refine lower layers, confirming that multi-level interaction matters. Total parameters reach 25.3M (7.7% over baseline), showing efficiency alongside accuracy gains. Results hold across all datasets, validating our design.

4.2.3 *Loss Function Analysis*

Table 3 examines how SSIM weight λ affects performance on DUT. BCE alone gives MAE 0.078, F$_\beta$ 0.915, IoU 0.871 reasonable but prone to fragmented outputs since it ignores spatial structure. Adding SSIM at $\lambda = 0.05$ improves to MAE 0.071 and F$_\beta$ 0.928, showing that structural constraints help. Peak performance occurs at $\lambda = 0.1$ (MAE 0.067, F$_\beta$ 0.942, IoU 0.909), balancing pixel accuracy with regional coherence. Higher values ($\lambda = 0.2, 0.3$) degrade results, MAE rises to 0.069 and 0.073 because excessive smoothness blurs boundaries. The sweet spot at $\lambda = 0.1$ preserves both sharp edges and consistent regions.

Table 3. Effect of SSIM loss weight λ on SemanticBlur performance (DUT dataset). Optimal value $\lambda = 0.1$ is **bolded**.

Loss Function	MAE↓	F$_\beta$↑	IoU↑
BCE only	0.078	0.915	0.871
BCE + SSIM ($\lambda = 0.05$)	0.071	0.928	0.889
BCE + SSIM ($\lambda = 0.1$)	**0.067**	**0.942**	**0.909**
BCE + SSIM ($\lambda = 0.2$)	0.069	0.938	0.904
BCE + SSIM ($\lambda = 0.3$)	0.073	0.931	0.896

4.3 Comparative Analysis

In this section, we evaluate the performance of our proposed network in comparison to several methodologies, focusing on both quantitative metrics and qualitative assessments.

4.3.1 *Quantitative Analysis*

Table 4 presents a comprehensive comparison of our proposed SemanticBlur network with eleven SOTA defocus blur detection methods. These results demonstrate the superior performance of our approach across multiple evaluation metrics. On the CUHK-TE dataset, SemanticBlur achieves competitive performance with MAE of 0.039, matching DFFNet. Our method significantly outperforms traditional approaches, showing MAE improvement over DBDF

Table 4. Quantitative comparison with state-of-the-art methods on benchmark datasets. MAE↓ indicates lower is better, while F_β↑ and IoU↑ indicate higher is better.

Method	CUHK-TE			DUT			CTCUG			EBD		
	MAE↓	F_β↑	IoU↑	MAE↓	F_β↑	IoU↑	MAE↓	F_β↑	IoU↑	MAE↓	F_β↑	IoU↑
DBDF [24]	0.292	0.806	0.549	0.384	0.754	0.519	0.360	0.668	0.438	0.389	0.603	0.435
LBP [27]	0.154	0.917	0.755	0.199	0.873	0.783	0.291	0.732	0.638	0.334	0.705	0.611
HiFST [9]	0.223	0.803	0.743	0.313	0.839	0.640	0.274	0.775	0.618	0.376	0.603	0.498
BTBNet [30]	0.110	0.949	0.914	0.196	0.861	0.803	0.177	0.809	0.762	–	–	–
BTBNet2 [21]	0.085	0.935	0.908	0.145	0.873	0.837	–	–	–	–	–	–
CENet [22]	0.061	0.945	0.932	0.141	0.869	0.833	0.117	0.845	0.820	0.072	0.810	0.899
BR2Net [19]	0.059	0.966	0.927	0.083	0.943	0.893	0.140	0.834	0.788	0.087	0.813	0.880
DD [14]	0.045	0.966	0.941	0.074	0.935	0.903	0.155	0.797	0.775	0.118	0.812	0.842
DefuNet [20]	–	–	–	0.085	0.952	0.889	0.132	0.828	0.798	–	–	–
IS2CNet [12]	0.049	0.964	0.937	0.142	0.868	0.831	0.112	0.858	0.826	0.070	0.809	0.901
DFFNet [42]	0.039	0.971	0.947	0.072	0.938	0.903	0.082	0.879	0.868	0.091	0.823	0.872
SemanticBlur (Ours)	0.039	0.969	0.951	0.067	0.942	0.909	0.086	0.887	0.894	0.084	0.835	0.886

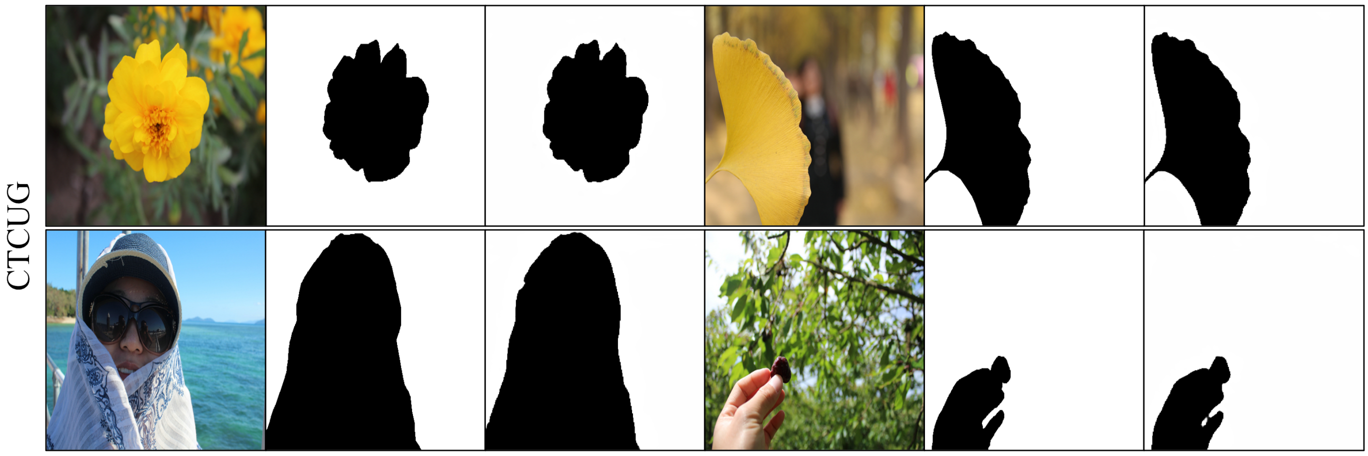

Figure 3. Additional qualitative results on challenging examples from CTCUG dataset.

(0.292 → 0.039) and improvement over LBP (0.154 → 0.039). The DUT dataset reveals the most significant performance gains for our method. SemanticBlur achieves the best MAE of 0.067, representing 9.5% improvement over the previous best result from DD (0.074 → 0.067).

On the challenging CTCUG dataset, known for complex blur patterns and subtle focus transitions, SemanticBlur maintains robust performance with MAE of 0.086, closely following DFFNet's 0.082. Despite the marginal difference in MAE, our method achieves superior F_β of 0.887 compared to DFFNet's 0.879, and significantly better IoU of 0.894 versus 0.868. The EBD dataset results show SemanticBlur achieving MAE of 0.084, outperforming most competing methods while maintaining competitive F_β of 0.835. Although IS2CNet achieves the highest IoU of 0.901 on this dataset, our method's IoU of 0.886 represents strong performance while maintaining superior MAE accuracy. The balanced performance across precision and recall metrics indicates robust generalization.

4.3.2 *Qualitative Analysis*

Visual comparisons in Figures 2 and 3 display results from CUHK, DUT, EBD, and CTCUG datasets, showing input images alongside ground truth annotations and our model's predictions. CUHK examples illustrate the method's precision, the dog portrait achieves clear separation between in-focus facial features and out-of-focus background regions without introducing noise. DUT samples reveal strong performance on varied scenes: the flower image successfully distinguishes focused foreground elements from defocused surroundings, and the hand gesture example preserves boundary accuracy despite varying blur intensities.

EBD samples cover multiple photographic contexts including indoor selective focus shots, architectural images, and outdoor scenes. Generated detection maps represent blur patterns accurately, maintaining edge sharpness and regional consistency. Semantic enhancement particularly benefits these cases, applying contextual information to separate deliberately blurred backgrounds from focused subjects. CTCUG presents difficult scenarios with gradual focus transitions and intricate blur variations, yet our approach maintains consistency, generating precise maps with limited erroneous detections. Predictions match ground truth closely throughout all test sets, demonstrating that attention mechanisms guided by semantic information retain fine details while correctly identifying defocused areas.

5 Conclusion

SemanticBlur presents a framework integrating semantic information with attention mechanisms operating across multiple scales for defocus blur detection. The architecture employs three attention types: channel, spatial, and semantic enhancement, to progressively refine feature representations. ResNet-50 modified with dilated convolutions retains spatial detail while expanding receptive fields, paired with a feature pyramid decoder using trainable fusion weights to combine multi-scale information. Loss function design weighs BCE against SSIM ($\lambda = 0.1$), optimizing for localization accuracy and structural consistency. Evaluation across four datasets (CUHK-TE, DUT, CTCUG, EBD) demonstrates strong results, supported by ablation experiments validating individual components. Potential extensions include incorporating transformer mechanisms for modeling longer-range relationships and examining training approaches that improve performance across varied image domains.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Zhang, W. & Cham, W. K. (2011). Single-image refocusing and defocusing. *IEEE Transactions on Image Processing, 21*(2), 873-882. [CrossRef]

[2] Zhang, X., Wang, R., Jiang, X., Wang, W. & Gao, W. (2016). Spatially variant defocus blur map estimation and deblurring from a single image. *Journal of Visual Communication and Image Representation, 35*, 257-264. [CrossRef]

[3] Gur, S., & Wolf, L. (2019, June). Single Image Depth Estimation Trained via Depth From Defocus Cues. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7675-7684). IEEE. [CrossRef]

[4] Liu, Y. Q., Du, X., Shen, H. L., & Chen, S. J. (2020). Estimating generalized gaussian blur kernels for out-of-focus image deblurring. *IEEE Transactions on circuits and systems for video technology, 31*(3), 829-843. [CrossRef]

[5] Abuolaim, A., & Brown, M. S. (2020, August). Defocus deblurring using dual-pixel data. In *European conference on computer vision* (pp. 111-126). Cham: Springer International Publishing. [CrossRef]

[6] Jiang, N., Sheng, B., Li, P. & Lee, T. Y. (2022). Photohelper: portrait photographing guidance via deep feature retrieval and fusion. *IEEE Transactions on Multimedia, 25*, 2226-2238. [CrossRef]

[7] Liu, Z. Y. & Liu, J. W. (2023). Hypergraph attentional convolutional neural network for salient object detection. *The Visual Computer, 39*(7), 2881-2907. [CrossRef]

[8] Chen, S., Sun, P., Song, Y., & Luo, P. (2023, October). DiffusionDet: Diffusion Model for Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 19773-19786). IEEE. [CrossRef]

[9] Golestaneh, S. A., & Karam, L. J. (2017, July). Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 596-605). IEEE. [CrossRef]

[10] Xu, G., Quan, Y., & Ji, H. (2017, October). Estimating Defocus Blur via Rank of Local Patches. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 5381-5389). IEEE. [CrossRef]

[11] Sakurikar, P., & Narayanan, P. J. (2017, October). Composite Focus Measure for High Quality Depth Maps. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1623-1631). IEEE. [CrossRef]

[12] Zhao, F., Lu, H., Zhao, W., & Yao, L. (2021). Image-scale-symmetric cooperative network for defocus blur detection. *IEEE Transactions on Circuits*

and Systems for Video Technology, 32(5), 2719-2731. [CrossRef]

[13] Zhao, Z., Yang, H., Liu, P., Nie, H., Zhang, Z., & Li, C. (2024). Defocus blur detection via adaptive cross-level feature fusion and refinement. *The Visual Computer, 40*(11), 8141-8153. [CrossRef]

[14] Cun, X., & Pun, C. M. (2020, August). Defocus blur detection via depth distillation. In *European conference on computer vision* (pp. 747-763). Cham: Springer International Publishing. [CrossRef]

[15] Lee, J., Lee, S., Cho, S., & Lee, S. (2019, June). Deep Defocus Map Estimation Using Domain Adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12214-12222). IEEE. [CrossRef]

[16] Li, E., Casas, S., & Urtasun, R. (2023, October). MemorySeg: Online LiDAR Semantic Segmentation with a Latent Memory. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 745-754). IEEE. [CrossRef]

[17] Jain, J., Singh, A., Orlov, N., Huang, Z., Li, J., Walton, S., & Shi, H. (2023, October). SeMask: Semantically Masked Transformers for Semantic Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 752-761). IEEE. [CrossRef]

[18] Jonna, S., Medhi, M., & Sahay, R. R. (2023). Distill-dbdgan: Knowledge distillation and adversarial learning framework for defocus blur detection. *ACM Transactions on Multimedia Computing, Communications and Applications, 19*(2s), 1-26. [CrossRef]

[19] Tang, C., Liu, X., An, S. & Wang, P. (2020). BR^{2}Net: Defocus blur detection via a bidirectional channel attention residual refining network. *IEEE Transactions on Multimedia, 23*, 624-635. [CrossRef]

[20] Tang, C., Liu, X., Zheng, X., Li, W., Xiong, J., Wang, L., Zomaya, A. Y. & Longo, A. (2020). DeFusionNET: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(2), 955-968. [CrossRef]

[21] Zhao, W., Zhao, F., Wang, D. & Lu, H. (2019). Defocus blur detection via multi-stream bottom-top-bottom network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(8), 1884-1897. [CrossRef]

[22] Zhao, W., Zheng, B., Lin, Q., & Lu, H. (2019). Enhancing diversity of defocus blur detectors via cross-ensemble network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8905-8913). [CrossRef]

[23] Tang, C., Liu, X., Zhu, X., Zhu, E., Sun, K., Wang, P., ... & Zomaya, A. (2020, April). R^{2}MRF: Defocus Blur Detection via Recurrently Refining Multi-Scale Residual Features. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(07), 12063-12070. [CrossRef]

[24] Shi, J., Xu, L., & Jia, J. (2014, June). Discriminative Blur Detection Features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2965-2972). IEEE. [CrossRef]

[25] Su, B., Lu, S. & Tan, C. L. (2011). Blurred image region detection and classification. In *Proceedings of the 19th ACM International Conference on Multimedia* (pp. 1397-1400). [CrossRef]

[26] Tang, C., Wu, J., Hou, Y., Wang, P., & Li, W. (2016). A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Processing Letters, 23*(11), 1652-1656. [CrossRef]

[27] Yi, X. & Eramian, M. (2016). LBP-based segmentation of defocus blur. *IEEE Transactions on Image Processing, 25*(4), 1626-1638. [CrossRef]

[28] Wang, B., Yang, M., Cao, P., Shen, A., & Liu, Y. (2025). STNet: a lightweight spectral transform framework for salient object detection. *Complex & Intelligent Systems, 11*(9), 381. [CrossRef]

[29] Park, J., Tai, Y. W., Cho, D., & Kweon, I. S. (2017, July). A Unified Approach of Multi-scale Deep and Hand-Crafted Features for Defocus Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2760-2769). IEEE. [CrossRef]

[30] Zhao, W., Zhao, F., Wang, D. & Lu, H. (2018). Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3080-3088). [CrossRef]

[31] Zeng, K., Wang, Y., Mao, J., Liu, J., Peng, W., & Chen, N. (2018). A local metric for defocus blur detection based on CNN feature learning. *IEEE Transactions on Image Processing, 28*(5), 2107-2115. [CrossRef]

[32] Jiang, Z., Xu, X., Zhang, C., & Zhu, C. (2020, September). Multianet: a multi-attention network for defocus blur detection. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)* (pp. 1-6). IEEE. [CrossRef]

[33] Wijayasingha, L., Alemzadeh, H., & Stankovic, J. A. (2024). Camera-independent single image depth estimation from defocus blur. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3749-3758). [CrossRef]

[34] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(8), 2011-2023. [CrossRef]

[35] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[36] Jiang, Z., Xu, X., Zhang, L., Zhang, C., Foo, C. S., & Zhu, C. (2022). MA-GANet: A multi-attention

generative adversarial network for defocus blur detection. *IEEE Transactions on Image Processing, 31*, 3494-3508. [CrossRef]

[37] Li, J., Fan, D., Yang, L., Gu, S., Lu, G., Xu, Y. & Zhang, D. (2021). Layer-output guided complementary attention learning for image defocus blur detection. *IEEE Transactions on Image Processing, 30*, 3748-3763. [CrossRef]

[38] Habib, K., Talha, U. M., & Imad, R. (2024). Attention enhanced machine instinctive vision with human-inspired saliency detection. [CrossRef]

[39] Zhao, Z., Yang, H. & Luo, H. (2022). Defocus Blur detection via transformer encoder and edge guidance. *Applied Intelligence, 52*(12), 14426-14439. [CrossRef]

[40] Chai, S., Zhao, X., Zhang, J. & Kan, J. (2024). Defocus blur detection based on transformer and complementary residual learning. *Multimedia Tools and Applications, 83*(17), 53095-53118. [CrossRef]

[41] Zhang, N., & Yan, J. (2020, August). Rethinking the defocus blur detection problem and a real-time deep DBD model. In *European Conference on Computer Vision* (pp. 617-632). Cham: Springer International Publishing. [CrossRef]

[42] Jin, Y., Qian, M., Xiong, J., Xue, N., & Xia, G. S. (2023, July). Depth and DOF cues make a better defocus blur detector. In *2023 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 882-887). IEEE. [CrossRef]

Asad Ullah Haider is a master's student at Ilmenau University of Technology, Germany, specializing in the fields of Machine Learning, Deep Learning, and Artificial Intelligence. His academic work focuses on exploring intelligent algorithms, data-driven models, and advanced computational techniques to solve real-world problems. He is actively engaged in research and projects that involve the application of AI to various domains, aiming to contribute to the development of efficient and innovative solutions.

Alexandros Gazis received his diploma in Electronic and Computer Engineering and his MSc in Microelectronics and Computer Systems from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2016 and 2018, respectively. Since 2018, he has been a PhD candidate in the field of computer science at the same university, where he is a member of the "Operating Systems and Middleware for Pervasive Computing and Wireless Sensor Networks" research group. He is also currently pursuing an MBA at Heriot-Watt University since February 2023. Moreover, he is a Teaching Assistant and Lab Demonstrator, supervised by Assistant Professor Eleftheria Katsiri. Mr. Gazis is a member of the Technical Chamber of Greece and works in the private sector as a Software Engineer for Piraeus Bank S.A., specializing in banking systems. He has published articles on Artificial Intelligence, game engines, web data analytics, remote sensing, and neural networks. His research focuses on the Internet of Things via wireless sensor networks, cloud computing, and middleware development for pervasive computing. (Email: agazis@ee.duth.gr)

Faryal Zahoor received her BS degree in Computer Science from the University of Agriculture, Peshawar, Pakistan, and MS degree in Computer Science from Islamia College University, Peshawar. Her research interests include computer vision, machine learning, deep learning, medical image processing, and pattern recognition. She is also affiliated with the BRAINS Institute, Peshawar, where she serves as a Lecturer. (Email: faryal.zahoorjan@gmail.com)