CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

V-UNet: Medical Image Segmentation Based on Variational Attention Mechanism

Yang Zhang[1] | Qiang Yang[1] | Tian Li[2,3] | Fanghong Zhang[4] | Yu Ren[5] | Yinhao Li[5] | Chuanyun Xu[1]

[1]College of Computer and Information Sciences, Chongqing Normal University, Chongqing, China | [2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia | [3]BPM Group, RWTH Aachen University, Aachen, Germany | [4]National Center for Applied Mathematics, Chongqing Normal University, Chongqing, China | [5]School of Artificial Intelligence, Chongqing University of Technology, Chongqing, China

Correspondence: Chuanyun Xu (xcy@cqnu.edu.cn)

Received: 15 February 2025 | **Revised:** 13 July 2025 | **Accepted:** 29 July 2025

Funding: This research was supported by the China Chongqing Municipal Education Commission (Grant KJZDM202500505); China Chongqing Municipal Science and Technology Bureau (Grants CSTB2024TIADCYKJCXX0009, CSTB2024NSCQ-LZX0043); Chongqing University of Technology graduate education high-quality development project (Grants gzlsz202304, gzlkc202401, gzltd202502); Chongqing University of Technology—Chongqing LINGLUE Technology Co. Ltd. Electronic Information (artificial intelligence) graduate joint training base.

Keywords: image segmentation | medical image processing | variational techniques

ABSTRACT

Accurate medical image segmentation plays a crucial role in improving the precision of computer-aided diagnosis. However, complex boundary shapes, low contrast and blurred anatomical structures make fine-grained segmentation a challenging task. Variational Bayesian inference quantifies uncertainty through probability distributions and can construct robust probabilistic models for the boundaries of ambiguous organs and tissues. In this paper, we apply variational Bayesian inference to medical image segmentation and propose variational attention to model the uncertainty of low-contrast and blurry tissue and organ boundaries. This enhances the model's ability to perceive segmentation boundaries, improving robustness and segmentation accuracy. Variational attention first estimates the parameters of the probability distribution of latent representations based on input features. Then, it samples latent representations from the learnt distribution to generate attention weights that optimise the interaction between global features and ambiguous boundaries. We integrate variational attention into the U-Net model by replacing its skip connections, constructing a multi-scale variational attention segmentation model (V-UNet). Experiments on the ISBI 2012 and MoNuSeg 2018 datasets show that our method achieves Dice scores of 95.89% and 82.18%, respectively. Moreover, we integrate V-UNet into the Mask R-CNN framework by replacing the FPN feature extraction head and propose a two-stage segmentation method. Compared to the original Mask R-CNN, our method improves the Dice score by 0.81%, mAP by 8.06% and F1 score by 0.51%.

1 | Introduction

Medical image segmentation plays a critical role in healthcare image analysis, directly impacting diagnostic and therapeutic outcomes in applications ranging from cancer screening and tumour detection to organ segmentation and lesion identification. The primary objective is to achieve precise boundary annotation and regional segmentation of anatomical structures and pathological areas, which forms the foundation for subsequent qualitative and quantitative analysis. However, manual segmentation of medical images remains labour-intensive, time-consuming and susceptible to inter-observer and intra-observer variability [1]. With rapid advancements in computer technology, particularly breakthroughs in artificial intelligence, computer-aided

Yang Zhang and Qiang Yang contributed equally and are joint first authors.

segmentation techniques have emerged. These techniques enable automated and accurate processing of large-scale medical images, significantly improving diagnostic efficiency and precision.

Among existing segmentation methods, deep learning techniques, especially convolutional neural networks (CNNs), have demonstrated remarkable capabilities in various medical image segmentation tasks [2]. Compared to traditional feature-engineered approaches, deep learning models possess the ability to automatically learn effective feature representations from data, thereby substantially enhancing segmentation performance [3]. Nevertheless, medical image segmentation still faces several challenges:

- *Complex Structures and Blurred Boundaries:* Cellular images often contain target structures such as nuclei and cytoplasm with considerable variability in morphology, scale and spatial distribution. These structures frequently exhibit adhesion, overlap or uneven staining, resulting in ambiguous and poorly defined boundaries. For example, in haematoxylin and eosin (H&E) stained histopathological slides or datasets like MoNuSeg, cell boundaries are often indistinct due to low contrast, posing significant challenges to precise segmentation.
- *Insufficient Modelling of Uncertainty:* Cell images are susceptible to various sources of uncertainty, including staining artefacts, imaging noise and section thickness variation. However, most existing segmentation approaches rely on deterministic frameworks and lack mechanisms to explicitly model predictive uncertainty. This limitation reduces robustness, particularly in the presence of ambiguous regions or distributional shifts between training and testing data.
- *Limited Capability in Capturing Global Contextual Features:* Accurate segmentation in histological images often requires modelling long-range dependencies, as the contextual relationship among distant regions can be critical for disambiguating local structures. Conventional convolutional networks are constrained by their local receptive fields, leading to suboptimal global context representation and inconsistent segmentation performance in densely populated or structurally complex scenarios.

As shown in Figure 1, cell images may exhibit issues such as low tissue clarity, contamination, uneven staining, blurred boundaries and indistinct features due to the methods used in tissue preparation and image acquisition. These factors lead to ambiguous criteria and poor recognisability, making it difficult to determine whether a specific region corresponds to a particular tissue structure based solely on simple observations. Therefore, it is essential to explore contextual relationships from their underlying structural distributions in order to eliminate ambiguities in these uncertain areas.

Various segmentation techniques have been proposed to tackle these challenges. Recent deep learning-based methods demonstrate particular advantages in handling complex scenarios and feature representation. Attention, as a technique to enhance model focus, has been widely adopted in image processing tasks. By dynamically weighting features across spatial regions,

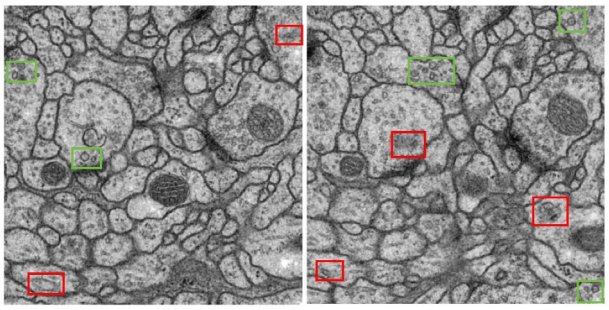

FIGURE 1 | The inherent uncertainty in the data is evident. The area within the green box, characterised by distinct and clear contours, is likely to be misclassified as a cellular structure, although it is not. Conversely, the structure within the red box that appears to lack cellular characteristics is actually a cellular structure.

attention enables models to better concentrate on critical image areas, thereby improving performance. However, existing attention inadequately accounts for image uncertainties, potentially compromising robustness and accuracy when dealing with complex backgrounds and ambiguous boundaries. Bayesian inference addresses this limitation by updating model posterior distributions through data observation to capture uncertainties. Nevertheless, direct computation of posterior distributions in Bayesian inference typically involves intractable high-dimensional integrals. Variational Bayesian inference overcomes this challenge by introducing a tractable variational distribution and minimising the Kullback–Leibler (KL) divergence between the variational and true posterior distributions. Integrating variational Bayesian inference with attention enables effective uncertainty modelling in images, thereby enhancing global feature capture capabilities and demonstrating superior robustness and generalisability, Figure 2 illustrates the computation process of variational attention.

This paper proposes V-UNet, which combines a variational attention with the U-Net architecture [4]. The variational attention addresses computational overhead in high-resolution medical images through multi-scale pooling operations. By performing pooling at multiple scales (e.g., 3×3, 6×6, 9×9), the network extracts multi-scale contextual information while reducing computational complexity and preserving semantic richness. For latent variable modelling, we employ the reparameterisation trick [5] to enable gradient backpropagation through sampling operations. Specifically, instead of direct sampling from the latent distribution, the model generates latent representations through differentiable affine transformations of noise variables sampled from a standard normal distribution $\epsilon \sim \mathcal{N}(0, I)$, combined with learnt mean (μ) and variance (σ) parameters. This approach combines stochasticity with differentiability, supporting gradient propagation through computational graphs.

Experimental results demonstrate that the proposed method achieves excellent performance on both ISBI 2012 and MoNuSeg 2018 datasets, particularly outperforming most existing mainstream methods in boundary recognition and segmentation accuracy. To enhance instance segmentation performance, this paper further optimises Mask R-CNN [6] by replacing the

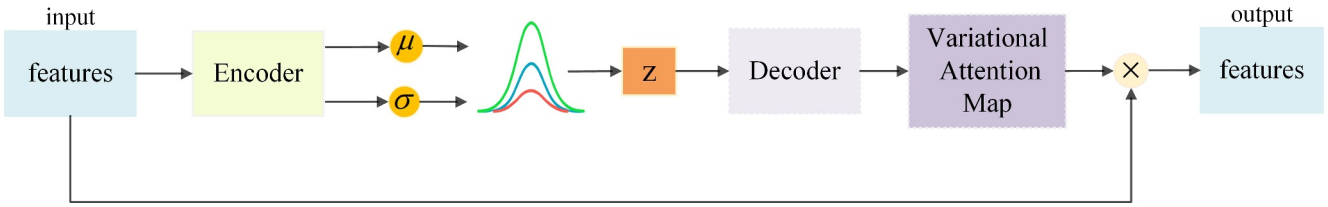

FIGURE 2 | The principle of variational attention. The input features are first processed by a variational encoder to compute the mean and variance, from which a latent representation z is sampled. This latent code is then used by a decoder to generate an attention map, which is subsequently combined with the feature map to produce the final output.

original FPN-extracted multi-scale feature maps with those generated during the upsampling process of V-UNet. This modification removes the FPN layer and constructs an efficient two-stage segmentation framework. This design significantly improves the model's precision in handling complex instances, especially in scenarios with ambiguous boundaries, demonstrating superior generalisation capability and robustness.

To address challenges in medical image segmentation, this paper proposes V-UNet, an innovative segmentation method that integrates a variational attention into U-Net. The method models uncertainty in the latent space of images through variational inference, thereby enhancing the model's perception of ambiguous boundaries and global semantic information. While maintaining segmentation accuracy, the network introduces multi-scale pooling operations to reduce computational overhead caused by high-resolution images, effectively decreasing parameter count and computational complexity while efficiently capturing multi-scale image information. Based on the above design, the main contributions of this paper are as follows:

1. Application of variational attention to medical image segmentation, improving the accuracy of complex boundary detection and global feature capture.
2. Proposal of V-UNet, a variational attention-based image segmentation network.
3. The implementation of a novel two-stage segmentation approach combining V-UNet with Mask R-CNN to optimise instance segmentation performance.

The remainder of this paper is organised as follows: Section 2 reviews related work in the field of image segmentation. Section 3 presents the proposed V-UNet architecture based on variational attention, detailing its implementation and discussing a two-stage segmentation framework that integrates V-UNet with Mask R-CNN. Section 4 describes the experimental setup, dataset configurations and evaluation metrics and reports comparative results on several public medical image segmentation datasets and summarises the segmentation performance of V-UNet. Finally, Section 5 discusses the limitations of the proposed method and outlines future research directions.

2 | Related Works

Traditional medical image segmentation methods primarily rely on morphological operations, thresholding and image filtering. However, they have limitations when dealing with images that feature complex shapes and blurred boundaries. To address these issues, model-based segmentation methods have emerged, including random walks, conditional random fields and active contour models [7–9]. These methods perform well in certain scenarios, but their generalisation ability is limited, and they struggle to handle complex shapes and noise interference [10]. With the rise of deep learning, convolutional neural networks (CNNs) have demonstrated significant advantages in medical image segmentation. CNNs automatically learn features, making it easier to handle complex shapes and blurred boundaries. Among these, U-Net is a fully convolutional network based on an encoder–decoder architecture. By utilising symmetric structures and skip connections, it combines multi-scale features with local details, achieving remarkable results. Building upon this, Oktay introduced Attention U-Net, which incorporates a self-attention to enhance feature representation, significantly improving segmentation accuracy, especially when dealing with complex boundaries and blurred regions [11]. Zhang proposed Res-UNet, which integrates a residual network architecture into U-Net, enhancing the model's segmentation performance and training efficiency [12]. Furthermore, Ibtehaz and Rahman proposed MultiResUNet, which adopts a multi-resolution feature fusion strategy to enhance the recognition capability of structures at different scales, though it also increases training time [13]. Although U-Net and its variants have made substantial progress in medical image segmentation, they still face limitations in capturing global contextual information and handling scenarios with complex shapes. To overcome these limitations, researchers have introduced transformer architectures to compensate for CNNs' shortcomings in capturing global features [14, 15]. Cao et al.'s Swin-Unet combines the hierarchical structure of the Swin Transformer, enhancing the ability to capture boundaries and fine details, although it incurs significant training overhead [16]. Chen et al.'s TransUNet combines the strengths of transformer and CNN, achieving outstanding performance on small sample datasets by integrating global semantic information with local features, but it also introduces high computational costs [17].

Besides classical convolutional and transformer-based methods, Mask R-CNN is a network framework based on instance segmentation that combines object detection with semantic segmentation, enabling precise organ and lesion area segmentation at the instance level. Recent research has further optimised Mask R-CNN. Wang et al. improved Mask R-CNN through context fusion and deconvolution pyramid modules, enhancing the detection and segmentation performance of overlapping regions [18]. However, even the optimised Mask R-CNN still has limitations in capturing boundary details. Despite the progress

made by existing improvements in enhancing segmentation performance, it remains challenging to accurately capture subtle boundary information, especially in scenarios with blurry or complex shapes. In other areas of deep learning, technologies for image analysis tasks are also continuously advancing. Zhang et al. proposed MHKD, which achieved effective recognition of glomerular structures in low-resolution whole-slide images, significantly improving the detection of small targets through a multi-stage distillation strategy [19]. Yao et al. further proposed a position-based anchor optimisation method, combined with a point supervision mechanism, which effectively enhanced the accuracy of high-density cell nucleus detection, particularly demonstrating strong application values under weak supervision [20].

Despite the significant progress made by traditional methods and deep learning-based segmentation models in medical image segmentation, they still face certain challenges when dealing with blurred boundaries, complex shapes or multi-modal images. To address these issues, methods based on attention have gradually become a research hotspot in recent years. Attention-based methods dynamically adjust the focus on key regions, effectively improving the model's segmentation accuracy under complex structures and blurred boundaries. SMFNet, proposed by Li et al. [21], utilises a sub-pixel-level multi-scale fusion attention, effectively enhancing the segmentation accuracy of boundary regions. Vman, proposed by Song et al. [15], introduces a visual-modified attention, significantly improving the semantic fusion ability in multi-modal images, but its performance drops noticeably when handling single-modal images. ATTransUNet, designed by Li et al. [22], combines spatial and channel attention for joint modelling, utilising transformers to enhance contextual understanding, effectively addressing the shortcomings of traditional CNNs in modelling long-range dependencies. Furthermore, feature fusion-based methods have also shown superior performance in medical image segmentation, particularly in multi-source information fusion and complex region recognition. Zheng et al. [23] combined the advantages of asymmetric adaptive heterogeneous networks and graph neural networks to improve the recognition ability of complex structures and boundary regions in multi-modal medical image segmentation, but the model structure is complex, and the high computational resource consumption makes it less suitable for lightweight deployment. SwinURNet, proposed by Wang et al. [24], combines transformer and CNN architectures, suitable for irregular road segmentation tasks, and performs excellently in real-time applications, especially in autonomous driving. However, it is highly dependent on the diversity of training data and has limited generalisation ability. CISA-UNet, proposed by Lu et al. [25], improves the segmentation accuracy of structural details in CBCT dental images by fusing structural priors and contextual information in dual auxiliary paths, but its robustness is weak under different resolutions and scanning conditions. Combining attention and feature fusion, Yu et al. [26] proposed a serial-parallel network structure that combines convolutional neural networks and transformers for 3D medical image segmentation. This method effectively promotes the interaction between global and local features through cross-window self-attention and multi-scale local enhancement modules. Xie et al. [27] proposed a network structure called U-shaped deformable transformer (UDT) for subarachnoid haemorrhage (SAH) image segmentation. This method improves the feature modelling of SAH lesions through multi-scale deformable attention and cross-deformable attention modules, enhancing segmentation performance.

In recent years, generative models and probabilistic graphical models have made significant progress in the field of cell image segmentation. As one of the important generative models, the variational autoencoder (VAE), first proposed by Kingma and Welling [5], aims to efficiently learn the latent distribution of complex data through variational inference. In the field of image processing, VAE not only generates images with a continuous latent space structure but is also widely used in tasks such as image reconstruction, denoising, style transfer and data augmentation, providing rich latent feature representations for subsequent image segmentation and analysis. Kebaili et al. [28] proposed a hybrid architecture called the Discriminative Hamiltonian Variational Autoencoder (HVAE) for tumour segmentation in medical images, particularly suitable for data-scarce scenarios. This method combines discriminative regularisation with Hamiltonian dynamics, effectively improving the generation quality of images and masks, reducing artefacts and abnormal distributions. Labbihi et al. [29] proposed a hybrid network for 3D medical image segmentation, combining CNN, frequency transformers and VAE. The VAE branch in the encoder–decoder structure is used to learn the latent distribution of 3D spatial features, enhancing feature representation and improving segmentation accuracy. Ichou et al. [30] proposed the VAE-AL-UNet model, which combines VAE with active learning for lung segmentation in chest X-ray images. The VAE branch effectively extracts latent feature distributions, reducing the dependency on a large amount of labelled data and improving segmentation accuracy under small sample conditions. Therefore, VAE can effectively learn latent feature distributions when dealing with small datasets, data scarcity or complex image tasks, enhancing the model's ability to capture details, thus playing a crucial role in fields such as medical image segmentation.

3 | Methodology

Medical image segmentation faces challenges when processing images with complex morphologies, ambiguous boundaries and noise interference. Although the symmetric encoder–decoder structure of U-Net and its variants is effective to some extent, its performance remains limited for images with complex morphologies, blurred boundaries and low contrast, making it difficult to meet the requirements of high-precision segmentation tasks. At the same time, existing methods also fall short in capturing global contextual information and detailed features. To solve these problems, this paper proposes an improved segmentation network based on the U-Net structure, introducing a variational attention module (VAEM) to enhance segmentation performance. The overall network architecture includes an encoder, decoder, bottleneck layer and VAEM, as shown in Figure 3.

The encoder stage is designed to gradually extract image feature information through multi-layer convolutional modules and pooling operations. In implementation, each convolutional

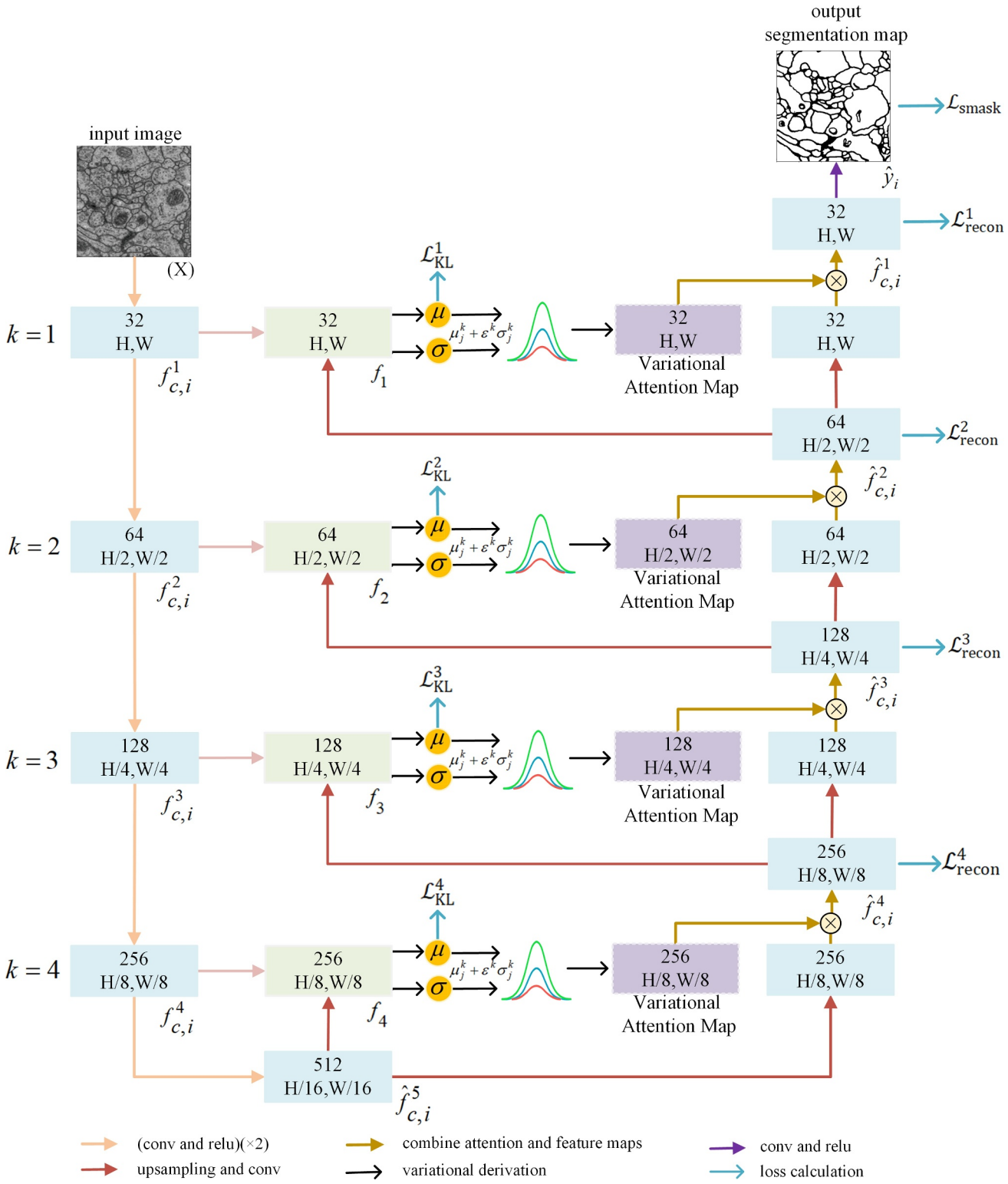

FIGURE 3 | Schematic diagram of the V-UNet architecture. $f^k_{c,i}$ is the down-sampled feature map of the k-th layer, $\hat{f}^k_{c,i}$ is the top-level down-sampled feature map and $\hat{f}^k_{c,i}$ represents the fusion of the k-th layer up-sampled feature map after variational attention, where c represents the number of channels. f_k is the fusion of $f^k_{c,i}$ and $\hat{f}^{k+1}_{c,i}$, with varying attention. The attention map (variational attention map) is obtained by sampling and reconstructing f_k.

module adopts a double convolution design, and we first apply a standard 3 × 3 convolution to extract spatial features, followed by a 1 × 1 convolution to further adjust the feature dimensions. This convolution combination can not only fully capture spatial contextual information but also effectively reduce computational overhead. After each convolution, a ReLU

activation function is used for nonlinear transformation, thereby effectively capturing local details and complex boundary information. To further enhance feature stability and representation ability, normalisation is added after convolution. In the decoder stage, the spatial resolution of the image is gradually restored through upsampling modules, using a combination of transposed convolution and bilinear interpolation to ensure the gradual reconstruction of fine target structures. At each upsampling step, VAEM is introduced to dynamically reconstruct the features.

3.1 | Variational Attention

The core of the VAEM lies in projecting feature representations to latent variables that are presumed to follow a multivariate Gaussian distribution. By constructing a multivariate Gaussian model, latent variables are sampled and used to generate refined attention maps. Specifically, $\hat{f}_{c,i}^{k+1}$ and $f_{c,i}^{k}$ are processed through two convolution layers, batch normalisation and activation functions. $\hat{f}_{c,i}^{k+1}$ is also resized to match the dimensionality of $f_{c,i}^{k}$, and then element-wise addition is performed. This is followed by multi-scale pooling (PL), where features are extracted at multiple scales, and the processed feature maps are concatenated to form a fused feature map f_k. Pooling results at different scales are flattened and concatenated into a one-dimensional vector to provide global contextual information. The fused feature map is then passed through two fully connected branches to calculate the mean μ_n^k and variance $\log \sigma_n^{k^2}$, where the standard deviation is computed as follows:

$$\sigma_n^k = \exp\left(0.5 \cdot \log \sigma_n^{k^2}\right). \tag{1}$$

Based on the probabilistic model, the latent variable is sampled for the change variable, and the latent variable in the representation is obtained as follows:

$$z_n^k = \mu_n^k + \epsilon^k \cdot \sigma_n^k, \tag{2}$$

where $\boldsymbol{\epsilon}^k$ is sampled from a standard normal distribution $N(0, I)$. After resizing, the optimisation is carried out using the reparameterisation (RE) method, and the advantages of obtaining the attention map are obtained. The VAEM combines multi-dimensional feature representation and latent variable models, which improve the accuracy of the model and enhance the sparsity and accuracy of the representation, and its complete structure is shown in Figure 4.

3.2 | Loss Function of V-UNet

The network loss $\mathcal{L}_{\text{total\_loss}}$ consists of the segmentation loss and the loss induced by variational learning, including reconstruction and KL divergence:

$$\mathcal{L}_{\text{total\_loss}} = \alpha \mathcal{L}_{\text{smask}} + \mathcal{L}_{\text{var}}, \tag{3}$$

where $\mathcal{L}_{\text{smask}}$ is the segmentation loss term, and binary cross-entropy loss (BCE Loss) is used to compute the difference between the network prediction output $\hat{y}_i$ and the ground truth label y_i. This loss can effectively address the foreground-background imbalance issue in pixel-level segmentation tasks by minimising the binary cross-entropy loss and optimising the model's performance on the segmentation task. The formula is as follows:

$$\mathcal{L}_{\text{smask}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right], \tag{4}$$

where N denotes the total number of pixels, y_i is the ground truth label for the i-th pixel (value is 0 or 1) and $\hat{y}_i$ is the predicted probability for the i-th pixel. $\mathcal{L}_{\text{var}}$ is the variational loss, composed of reconstruction loss and KL divergence loss:

$$\mathcal{L}_{\text{var}} = \beta \mathcal{L}_{\text{recon}} + \gamma \mathcal{L}_{\text{KL}}, \tag{5}$$

where α is a hyperparameter for weighting the segmentation loss, with a value set to 0.74; β and γ control the weights of the reconstruction loss and KL divergence loss, respectively, and are set to $\beta = 0.15$ and $\gamma = 0.11$. These values are obtained through hyperparameter tuning using the Bayesian optimisation method on the validation set. In the VAE model, given data X and latent variable z, the goal is to compute the posterior distribution. The derivation process can be referred to in Equation (5):

$$P(z|X) = \frac{P(X|z)P(z)}{P(X)}. \tag{6}$$

However, $P(X)$ involves high-dimensional integration and is difficult to compute. Therefore, a simpler distribution $q(z|X)$ is introduced to approximate the true posterior $P(z|X)$ through optimisation. We formulate an optimization problem where the objective is to minimize the KL divergence between a variational distribution $q(z|X)$ and $P(z|X)$. This divergence minimization ensures $q(z|X)$ closely approximates $P(z|X)$ when direct computation of the latter is infeasible:

$$\text{KL}(q(z|X)||P(z|X)) = \mathbb{E}_{q(z|X)}\left[\log \frac{q(z|X)}{P(z|X)}\right]. \tag{7}$$

Through KL divergence decomposition, the intractable marginal likelihood $\log P(X)$ can be approximated and optimised as follows:

$$\log P(X) = \mathbb{E}_{q(z|X)}\left[\log \frac{P(X,z)}{q(z|X)}\right] + \text{KL}(q(z|X)||P(z|X)). \tag{8}$$

The first term is the evidence lower bound (ELBO), since $\log P(X)$ is constant and the second term involves the true posterior, the computation becomes intractable due to high-dimensional integration. Therefore, maximising the ELBO is equivalent to minimising $KL(q(z|X)||P(z|X))$. ELBO can be further written as follows:

$$\text{ELBO}_{(q)} = \mathbb{E}_{q(z|X)}[\log P(X|z)] - \text{KL}(q(z|X)||p(z)). \tag{9}$$

The first term is the reconstruction term, which measures the reconstruction ability of latent variable z for the input data X, and the second term is the KL divergence, which regularises the latent

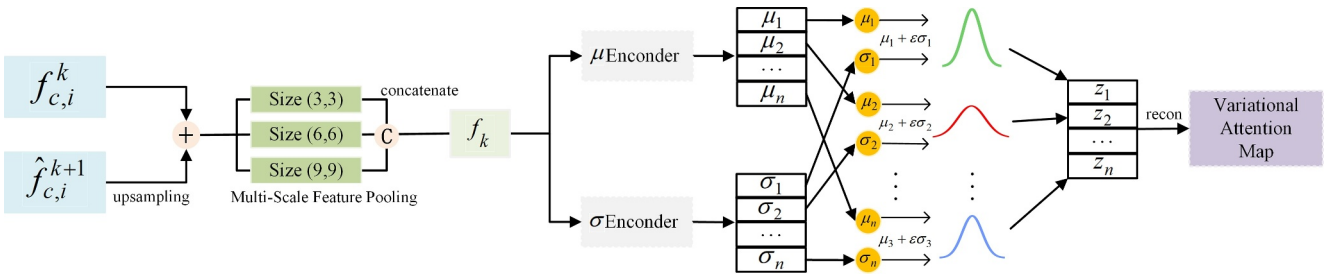

FIGURE 4 | The multi-scale pooling module PL generates feature maps of three scales: 3 × 3, 6 × 6 and 9 × 9. These feature maps are processed by PL and concatenated to form a fused feature map. The fused feature map is then passed to the mean and variance encoders to extract n samples of means and variances. Based on these samples, a variational attention map is generated through the reconstruction process.

space. The reconstruction term can be transformed into the reconstruction loss $\mathcal{L}_{\text{recon}}$, which is computed by averaging the reconstruction losses $\mathcal{L}_{\text{recon}}^{k}$ at each layer. $\mathcal{L}_{\text{recon}}$ measures reconstruction quality using the mean squared error (MSE loss) between $\hat{f}_{c,i}^{k+1}$ and $f_{c,i}^{k}$:

$$\mathcal{L}_{\text{recon}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{recon}}^{k}, \tag{10}$$

where

$$\mathcal{L}_{\text{recon}}^{k} = \frac{1}{N_c}\sum_{c=1}^{N_c}\frac{1}{N_f^k}\sum_{i=1}^{N_f^k}\left(f_{c,i}^{k} - \hat{f}_{c,i}^{k}\right)^2. \tag{11}$$

Among them, $f_{c,i}^{k}$ is the i-th element of the c-th feature map at the k-th layer, and $\hat{f}_{c,i}^{k+1}$ is the i-th element of the c-th feature map at the k-th layer. K represents the total number of layers, N_c denotes the number of channels in each layer and N_f^k represents the number of feature elements in the k-th layer. The experimental section of this paper introduces a simple distribution as a Gaussian distribution $q(z|X) = \mathcal{N}(z|\mu, \sigma^2)$, where the prior is set as a multivariate standard normal distribution $p(z) = \mathcal{N}(0, I)$. The KL divergence term in the ELBO can be derived. The detailed derivation process can be found in ref. [5].

$$\mathcal{L}_{\text{KL}} = \frac{1}{K}\sum_{k=1}^{K}N_f^k\left(\frac{1}{2}\left(1 + \log\left(\sigma_n^{k^2}\right) - \mu_n^{k^2} - \sigma_n^{k^2}\right)\right), \tag{12}$$

where μ_n^k and σ_n^k are the mean and standard deviations of the n-th feature element of the k-th layer, obtained by sampling from the fusion of features f_k. This term is used to minimise the variational loss $\mathcal{L}_{\text{var}}$, which optimises the ELBO.

3.3 | Two-Stage Segmentation Method

Through Mask R-CNN using FPN for feature extraction segmentation results analysis, although FPN can enhance target detection and segmentation performance by fusing multi-scale features, it still has shortcomings in handling complex boundaries and ambiguous regions in medical images. Specifically, FPN mainly adopts a top-down feature fusion strategy, combining features of different resolutions, but this approach is prone to boundary detail processing errors and information loss, negatively impacting segmentation performance, especially in medical images that require precise boundary delineation. To improve Mask R-CNN performance during feature extraction and considering that multi-stage segmentation strategies inevitably incur additional computational overhead, this work utilises the multi-scale feature maps $\hat{f}_{c,i}^{k}$ (with $k \leq 4$) generated during the upsampling process of V-UNet as substitutes for the FPN feature maps extracted in Mask R-CNN. These are fed into the subsequent Mask R-CNN branches as inputs to the region proposal network (RPN), directly pruning the FPN layers in Mask R-CNN to control computational cost. Subsequently, the regions proposed by RPN are precisely aligned using RoIAlign, and through the remaining network, classification, bounding box regression and binary mask prediction are performed to complete the instance segmentation task. Figure 5 represents the pipeline of the proposed model.

The loss function of the network model $\mathcal{L}_{\text{vm\_loss}}$ combines the variational loss of V-UNet and the loss in Mask R-CNN, expressed as follows:

$$\mathcal{L}_{\text{vm\_loss}} = \mathcal{L}_{\text{var}} + \mathcal{L}_{m}, \tag{13}$$

where $\mathcal{L}_{\text{var}}$ denotes the variational loss of V-UNet, including the reconstruction loss and KL divergence loss. $\mathcal{L}_{m}$ is the loss from Mask R-CNN, mainly consisting of classification loss $\mathcal{L}_{\text{cls}}$, bounding box regression loss $\mathcal{L}_{\text{bbox}}$ and segmentation loss $\mathcal{L}_{\text{mask}}$, formulated as follows:

$$\mathcal{L}_{m} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{mask}}. \tag{14}$$

The classification loss uses cross-entropy to measure the discrepancy between predicted class probabilities and ground truth labels, expressed as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{j=1}^{K_m} y_j \log\left(\hat{p}_j\right), \tag{15}$$

where K_m is the total number of classes, y_j is the ground truth label for class j and $\hat{p}_j$ is the predicted probability for class j. The bounding box regression loss $\mathcal{L}_{\text{bbox}}$ employs the smooth L_1 loss function to measure the difference between the predicted bounding box and the ground truth bounding box, calculated as follows:

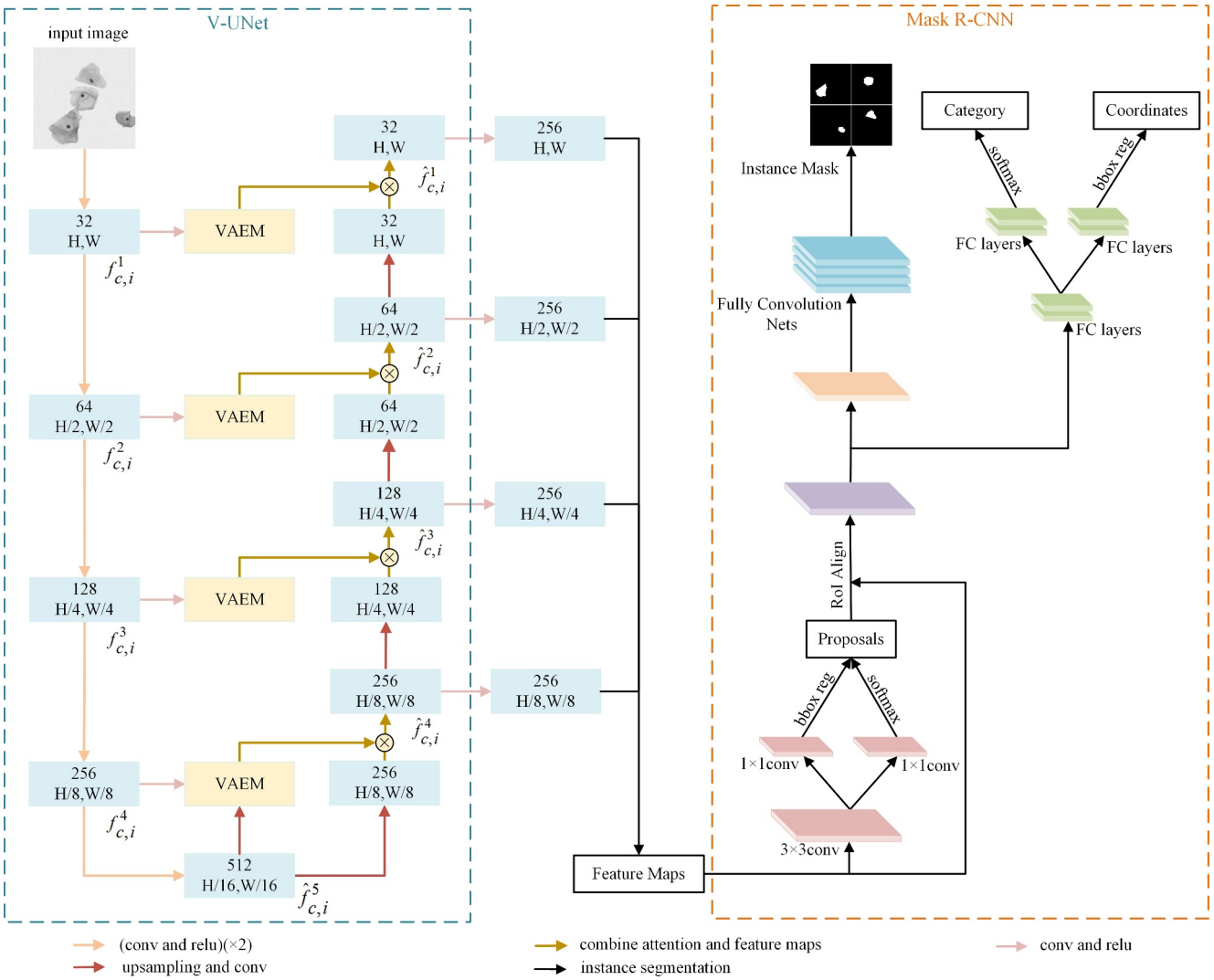

FIGURE 5 | Two-stage segmentation method. The feature map $\hat{f}_{c,i}^{k}$ from the process of V-Net is taken as the input to Mask R-CNN, and then passed through Mask R-CNN to generate region proposals. These region proposals are generated through the region proposal network (RPN) and refined by RoI Align, where the features are aligned and used to produce the final segmentation and classification.

$$\mathcal{L}_{\text{bbox}} = -\sum_{e=1}^{M} \text{smooth}_{L_1}(t_e - \hat{t}_e), \tag{16}$$

where M is the number of bounding boxes, and t_e and $\hat{t}_e$ denote the ground truth and predicted bounding box parameters, respectively. The segmentation loss $\mathcal{L}_{\text{mask}}$ uses binary cross-entropy loss to measure the discrepancy between the predicted mask and the ground truth mask, calculated similarly to Equation (4).

4 | Experimental Results

4.1 | Experimental Setup

The proposed method was evaluated on three publicly available image datasets, including tasks of semantic segmentation and instance segmentation.

ISBI2012: This dataset consists of 2D image stacks acquired from electron microscopy (EM) and is primarily used for neuron structure segmentation. It includes images extracted from EM slices for pixel-level segmentation and boundary detection of cellular components. The dataset is widely used in the biomedical image analysis field and has achieved notable performance in segmentation precision and recall. In our experiments, we used 30 2D EM images with a resolution of 512 × 512, splitting the dataset into training, validation and testing sets in a 4:1:1 ratio. These images were utilised for training models and evaluating generalisation ability.

MoNuSeg2018: This dataset is specifically designed for nuclear segmentation in histopathology images and focuses on automating the detection of nuclear boundaries. The dataset includes images from multiple medical centres, capturing a wide variety of image resolutions and complex morphologies. The training set consists of 30 histopathology images, each with dimensions of 1000 × 1000, and contains over 21,000 annotated

nuclei across diverse tissue types (e.g., kidney, liver, breast, prostate, bladder). The test set includes 14 unannotated images of the same size for evaluating the robustness of segmentation models.

ISBI2014: This dataset is tailored for mitotic figure segmentation under conditions of heavy cellular crowding. The primary task is to accurately segment mitotic cells in densely packed regions. The dataset includes 8-bit extended depth-of-field (EDF) images generated from high-resolution microscopy, with image dimensions of 512 × 512. To introduce additional challenges, we adopted a setting with 45, 90 and 180 images for training, validation and testing, respectively.

Evaluation Metrics: To comprehensively evaluate the overall performance of the networks, different performance metrics were adopted for comparison across datasets, including the Jaccard index (AJI), average Dice coefficient (Dice), precision (AP), mean average precision (mAP), F1 score (F1) and Intersection over Union (IoU). Additionally, the number of parameters and computational complexity (GFLOPs) of each network were compared on the ISBI2012 dataset Table 1.

4.2 | Comparative Analysis

Comparative experiments were conducted on the partitioned ISBI2012 dataset involving several U-Net-based networks. Figure 6 illustrates the segmentation performance of V-UNet compared to other methods on the ISBI2012 dataset. In the areas marked by the red box, V-UNet demonstrates superior segmentation results, accurately identifying fine structures within the cell images. However, there are still a few ambiguous cell structures that have not been accurately segmented, as indicated by the green box in Figure 6. The advantage of V-UNet in segmentation performance is attributed to the introduced variational attention mechanism, which significantly enhances attention to key regions and successfully captures more global features Table 2.

The proposed V-UNet outperforms other methods across multiple evaluation metrics. Specifically, V-UNet achieves a Dice coefficient of 0.9589, significantly higher than other models, indicating superior overall segmentation accuracy. Meanwhile, V-UNet attains an AJI score of 0.9186, representing a notable improvement over DC-UNet's 0.9117 and other networks, demonstrating better segmentation performance on complex boundaries. Additionally, V-UNet exhibits the best performance in average precision (AP), reaching 0.9526, which indicates higher robustness and stability in detecting and accurately segmenting target regions. In this experiment, the V-UNet model contains 6.91 million parameters, whereas the Attention U-Net model has 2.03 million. Moreover, the computational complexity of Attention U-Net is 14.01 GFLOPs, whereas V-UNet requires 19.71 GFLOPs. This difference is mainly attributed to the introduction of the variational attention. In the variational attention module, fully connected layers during the sampling process are inevitable, and their incorporation directly increases the parameter count. In particular, during the variational sampling process, large-scale weight matrices are required to ensure accurate latent space modelling, which increases both the model parameters and computational complexity to some extent. Furthermore, the multi-scale pooling layers (3 × 3, 6 × 6, 9 × 9) in the V-UNet model also have a significant impact on the total number of parameters. Increasing the pooling sizes results in the capture and integration of more extensive feature maps, consequently elevating computational requirements. Adjusting the pooling sizes can effectively control the parameter count.

The experimental results on the MoNuSeg2018 dataset further validate the superior segmentation performance of V-UNet. V-UNet achieved Dice and IoU scores of 0.8218 and 0.7053, respectively, significantly outperforming state-of-the-art methods such as FSA-Net (0.8032) and MBUTransUNet (0.8160), indicating its superior overall segmentation accuracy and more precise pixel-level classification. Furthermore, V-UNet's IoU metric surpasses all comparative models, especially compared to the traditional U-Net (0.6927) and the recent UCTransnet (0.6668), demonstrating better coverage of target regions in nucleus segmentation tasks.

The two-stage segmentation method combining V-UNet with Mask R-CNN achieved a Dice score of 0.9189, outperforming Mask R-CNN (0.9115), Cascade R-CNN (0.9129) and HTC (0.9139), indicating a significant advantage in overall segmentation accuracy. Additionally, in terms of mean average precision (mAP), V-UNet + Mask R-CNN achieved 0.6385, surpassing Mask Scoring R-CNN's 0.6356 and Occlusion R-CNN's 0.6235, validating its excellent performance in object detection and precise segmentation region localisation. In terms of the F1 score, our method achieved 0.9301, comparable to Occlusion R-CNN's 0.9318 and substantially higher than all other evaluated models Table 3.

4.3 | Ablation Analysis

To investigate the impact of network components, the variational module was divided into the variational computation module

TABLE 1 | Segmentation results of V-UNet and other advanced U-Net architectures on the ISBI2012 dataset.

Method	Dice ↑	AJI ↑	AP ↑	Params (M)	GFLOPs
U-Net [4]	0.9433	0.8931	0.9283	1.94	13.74
Attention U-Net [11]	0.9492	0.9030	0.9325	2.03	14.01
DC UNet [31]	0.9498	0.9117	0.9489	2.69	23.04
V-UNet (ours)	**0.9589**	**0.9186**	**0.9526**	6.91	19.71

Note: Bold values denote the optimal results within each comparison group.

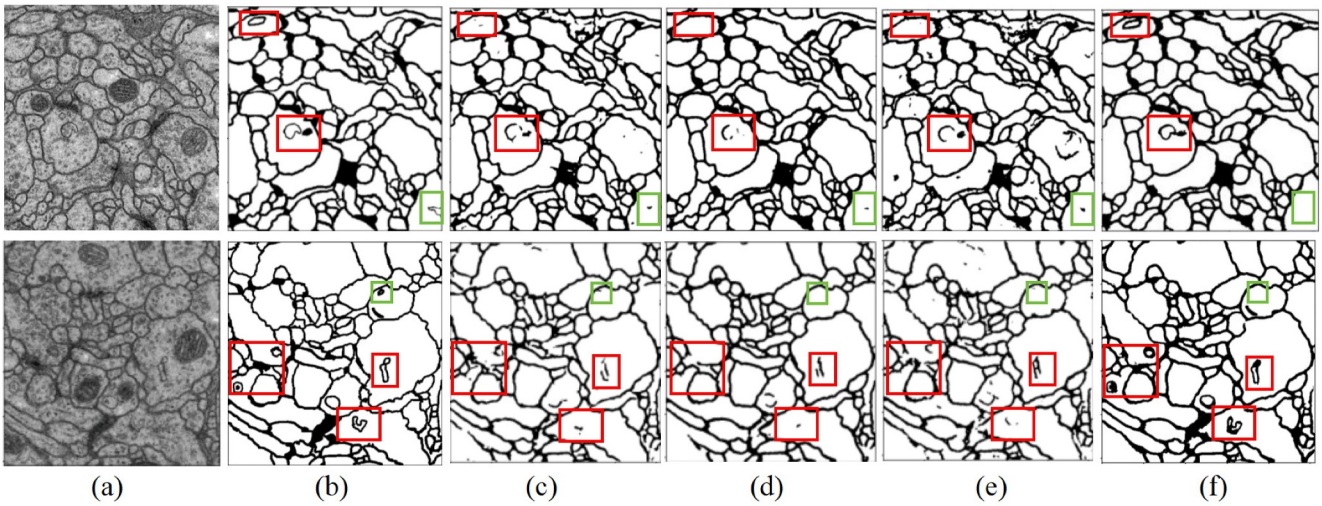

FIGURE 6 | Comparison of segmentation results on ISBI2012 between the proposed network and other methods. (a) Input image; (b) ground truth; (c) U-Net [4]; (d) Attention U-Net [11]; (e) DC UNet [31]; (f) V-UNet.

TABLE 2 | Segmentation results of V-UNet and other networks on the MoNuSeg2018 dataset.

Method	Dice ↑	IoU ↑
U-Net [4]	0.8185	0.6927
U-Net++ [32]	0.7528	0.6089
Attention U-Net [11]	0.7620	0.6264
MultiResUNet [13]	0.7754	0.6380
Bio-net [33]	0.7655	0.6252
TransUNet [17]	0.7920	0.6568
ATTransUNet [22]	0.7916	0.6551
UCTransNet [34]	0.7987	0.6668
MBUTransUNet [35]	0.8160	0.6902
Swin-Unet [16]	0.7956	0.6471
V-UNet (ours)	**0.8218**	**0.7053**

Note: Bold values denote the optimal results within each comparison group.

TABLE 3 | Comparative experiments of V-UNet combined with Mask R-CNN on the ISBI2014 dataset.

Method	Dice ↑	mAP ↑	F1 ↓
Mask R-CNN [6]	0.9115	0.5909	0.9254
Cascade R-CNN [36]	0.9129	0.6245	0.9251
Mask scoring R-CNN [37]	0.9128	0.6356	0.9187
HTC [38]	0.9139	0.5962	0.8808
Occlusion R-CNN [39]	0.9175	0.6235	**0.9318**
Xiao et al. [40]	0.9170	0.5734	0.9275
V-UNet + Mask R-CNN (ours)	**0.9189**	**0.6385**	0.9301

Note: Bold values denote the optimal results within each comparison group.

(VAE), pooling module (PL) and reconstruction module (RE). The comparison results are shown in Table 4. After incorporating the variational attention mechanism, improvements of 1.56%, 2.55% and 2.43% were observed in Dice, AJI and AP, respectively, on the ISBI2012 dataset.

This paper proposes a variational attention-based image segmentation network (V-UNet), aiming to address issues of boundary ambiguity and complex morphology recognition in medical image segmentation. V-UNet incorporates a variational attention into the U-Net architecture, modelling the latent probabilities of multi-scale features during the upsampling stage by combining shallow and deep features. To handle the non-differentiability of the sampling process, the reparameterisation trick is employed to generate new feature representations from the probabilistic model, which are then used to produce refined attention maps. This design significantly enhances the model's sensitivity to boundary details and its ability to capture global semantic information. It demonstrates outstanding segmentation performance, particularly in handling complex shape variations and expressing global features. Extensive experiments show that V-UNet achieves superior performance in medical image segmentation tasks, excelling in boundary recognition and exhibiting strong potential in other complex visual application scenarios.

5 | Limitations and Future Work

V-UNet has achieved outstanding segmentation performance across multiple public datasets and has significantly improved accuracy. However, there are still some limitations that require further optimisation. Although V-UNet demonstrates notable improvements in handling complex boundaries and capturing global contextual information, it may struggle with the accurate segmentation of fine-grained structures, particularly in regions with highly irregular or morphologically complex shapes. This limitation is primarily attributed to the stochastic nature of the sampling process in V-UNet, which may lead to suboptimal modelling of detailed structures, thus affecting the overall segmentation integrity and accuracy. To address this issue, a conventional attention branch could be introduced during the variational inference stage and fused with the variational

TABLE 4 | The impact of each module on segmentation results for ISBI2012 and MoNuSeg2018.

				ISBI2012			MoNuSeg2018	
Base	**VAE**	**PL**	**RE**	**Dice**	**AJI**	**AP**	**Dice**	**IoU**
✓				0.9433	0.8931	0.9283	0.8185	0.6927
✓	✓			0.9501	0.9053	0.9358	0.8135	0.6989
✓	✓	✓		0.9541	0.9127	0.9484	0.8198	0.7031
✓	✓	✓	✓	**0.9589**	**0.9186**	**0.9526**	**0.8218**	**0.7053**

Note: The ✓symbol indicates the inclusion of the corresponding module. Bold values denote the optimal results within each comparison group.

attention maps to compensate for the shortcomings of variational modelling in fine detail reconstruction. In future work, we plan to integrate the variational attention with graph reasoning methods [41], constructing a multi-layer sampling graph neural network [42] to effectively capture complex relationships among nodes. Additionally, the variational attention can be extended to multi-modal interaction [43], capsule networks [44] and saliency detection [45], where its incorporation could further enhance the model's robustness and accuracy. Finally, considering the limited availability of annotated data in practical medical applications, a self-paced semi-supervised learning strategy [46] may serve as an effective approach to achieving more robust segmentation performance under low-resource settings. Through these improvements, V-UNet is expected to play a broader and more impactful role in the field of medical image analysis.

Acknowledgements

We sincerely thank all reviewers and editors for their careful guidance and dedicated efforts, which have provided important support for the improvement and publication of this paper. We also thank the National Advanced Computing Taiyuan Centre and Taihang Laboratory in Shanxi Province (Advanced Computing Laboratory in Shanxi Province) for providing computational resources for this paper.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets used in this study are publicly available for download. The ISBI 2012 Neuronal Cell Dataset can be accessed at https://github.com/decouples/Unet, the ISBI 2014 Cervical Overlap Cell Dataset is available at https://cs.adelaide.edu.au/~carneiro/isbi14_challenge/dataset.html and the MoNuSeg 2018 dataset can be found at https://github.com/monuseg/2018-dataset.

References

1. H. Boulehmi and R. L. Filali, "Medical Image Segmentation Techniques: Advances and Challenges," in *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, (IEEE Press, 2025), 1–7, https://doi.org/10.1109/IPAS63548.2025.10924519.

2. X. Liu, L. Song, S. Liu, and Y. Zhang, "A Review of Deep-Learning-Based Medical Image Segmentation," *Sustainability* 13, no. 3 (2021): 1224, https://doi.org/10.3390/su13031224.

3. X. Zhang, Y. Wang, J. Wei, X. Yuan, and M. Wu, "A Review of Non-Fully Supervised Deep Learning for Medical Image Segmentation," *Information* 16, no. 6 (2025): 433, https://doi.org/10.3390/info16060433.

4. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science 9351*, (2015), 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

5. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114* (2013), https://doi.org/10.48550/arxiv.1312.6114.

6. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (IEEE Press, 2017), 2980–2988, https://doi.org/10.1109/ICCV.2017.322.

7. L. Grady, "Random Walks for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, no. 11 (2006): 1768–1783, https://doi.org/10.1109/TPAMI.2006.233.

8. A. Quattoni, M. Collins, and T. Darrell, "Conditional Random Fields for Object Recognition," *Advances in Neural Information Processing Systems* (2004): 1097–1104.

9. C. Xu and J. L. Prince, "Snakes, Shapes, and Gradient Vector Flow," *IEEE Transactions on Image Processing* 7, no. 3 (1998): 359–369, https://doi.org/10.1109/83.661186.

10. Z. Marinov, P. F. Jager, J. Egger, J. Kleesiek, and R. Stiefelhagen, "Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 10998–11018, https://doi.org/10.1109/TPAMI.2024.3452629.

11. O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv preprint arXiv:1804.03999* (2018), https://doi.org/10.48550/arXiv.1804.03999.

12. Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters* 15, no. 5 (2018): 749–753, https://doi.org/10.1109/LGRS.2018.2802944.

13. N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," *Neural Networks* 121 (2020): 74–87, https://doi.org/10.1016/j.neunet.2019.08.025.

14. M. M. Rahman and R. Marculescu, "Medical Image Segmentation Via Cascaded Attention Decoding," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (IEEE Press, 2023), 6211–6220, https://doi.org/10.1109/WACV56688.2023.00616.

15. X. Song, D. Han, C. Chen, X. Shen, and H. Wu, "Vman: Visual-Modified Attention Network for Multimodal Paradigms," *Visual Computer* 41, no. 4 (2025): 2737–2754, https://doi.org/10.1007/s00371-024-03563-4.

16. H. Cao, Y. Wang, J. Chen, et al., "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," in *Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science 13803* (2023), 205–218, https://doi.org/10.1007/978-3-031-25066-8_9.

17. J. Chen, J. Mei, X. Li, et al., "Transunet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers," *Medical Image Analysis* 97 (2024): 103280: Epub

2024 Jul 22. PMID: 39096845, https://doi.org/10.1016/j.media.2024.103280.

18. T. Wang, W. Feng, and M. Zhao, "Cervical Cell Image Segmentation Based on Improved the Mask R-CNN Model," in *Proceedings of the 10th International Conference on Advanced Intelligent Systems and Informatics 2024, Lecture Notes on Data Engineering and Communications Technologies* 220 (2024), 259–269, https://doi.org/10.1007/978-3-031-71619-5_22.

19. X. Zhang, L. Han, C. Xu, et al., "MHKD: Multi-Step Hybrid Knowledge Distillation for Low-Resolution Whole Slide Images Glomerulus Detection," *IEEE Journal of Biomedical and Health Informatics* 29, no. 2 (2025): 767–774, https://doi.org/10.1109/JBHI.2024.3513716.

20. J. Yao, L. Han, G. Guo, et al., "Position-Based Anchor Optimization for Point Supervised Dense Nuclei Detection," *Neural Networks* 171 (2024): 159–170, https://doi.org/10.1016/j.neunet.2023.12.006.

21. J. Li, Q. Chen, and X. Fang, "Sub-Pixel Multi-Scale Fusion Network for Medical Image Segmentation," *Multimedia Tools and Applications* 83, no. 41 (2024): 89355–89373, https://doi.org/10.1007/s11042-024-20338-0.

22. X. Li, S. Pang, R. Zhang, et al., "Attransunet: An Enhanced Hybrid Transformer Architecture for Ultrasound and Histopathology Image Segmentation," *Computers in Biology and Medicine* 152 (2023): 106365, https://doi.org/10.1016/j.compbiomed.2022.106365.

23. S. Zheng Yl, Z. Liu, C. Yang, et al., "Asymmetric Adaptive Heterogeneous Network for Multi-Modality Medical Image Segmentation," *IEEE Transactions on Medical Imaging* 44, no. 4 (2025): 1836–1852, https://doi.org/10.1109/TMI.2025.3526604.

24. Z. Wang, Z. Liao, B. Zhou, G. Yu, and W. Luo, "Swinurnet: Hybrid Transformer-CNN Architecture for Real-Time Unstructured Road Segmentation," in *IEEE Transactions on Instrumentation and Measurement* 73, (IEEE Press, 2024), 1–16: Art no. 5035816, https://doi.org/10.1109/TIM.2024.3470042.

25. J. Lu, X. Huang, C. Song, et al., "CISA-UNet: Dual Auxiliary Information for Tooth Segmentation From CBCT Images," *Alexandria Engineering Journal* 114 (2025): 103–118, https://doi.org/10.1016/j.aej.2024.11.103.

26. B. Yu, Q. Zhou, L. Yuan, H. Liang, P. Shcherbakov, and X. Zhang, "3D Medical Image Segmentation Using the Serial–Parallel Convolutional Neural Network and Transformer Based on Cross-Window Self-Attention," *CAAI Transactions on Intelligence Technology* 10, no. 2 (2025): 337–348, https://doi.org/10.1049/cit2.12411.

27. W. Xie, L. Jin, S. Hua, et al., "UDT: U-Shaped Deformable Transformer for Subarachnoid Haemorrhage Image Segmentation," *CAAI Transactions on Intelligence Technology* 9, no. 3 (2024): 756–768, https://doi.org/10.1049/cit2.12302.

28. A. Kebaili, J. Lapuyade-Lahorgue, P. Vera, and S. Ruan, "Discriminative Hamiltonian Variational Autoencoder for Accurate Tumor Segmentation in Data-Scarce Regimes," *Neurocomputing* 606 (2024): 128360, https://doi.org/10.1016/j.neucom.2024.128360.

29. I. Labbihi, O. El Meslouhi, Z. Elmrani Abou Elassad, M. Benaddy, M. Kardouchi, and M. Akhloufi, "Hybrid 3D Medical Image Segmentation Using CNN and Frequency Transformer Fusion," *Arabian Journal for Science and Engineering* 50, no. 15 (2024): 11713–11726, https://doi.org/10.1007/s13369-024-09602-5.

30. M. Ichou, M. Abik, and O. Naggar, "VAE-AL-UNet: Efficient Lung Segmentation in Chest X-ray Images Using Variational Autoencoder and Active Learning," in *International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023), Lecture Notes in Networks and Systems* 904 (2024), 136–153, https://doi.org/10.1007/978-3-031-52388-5_15.

31. A. Lou, S. Guan, and M. Loew, "DC-UNet: Rethinking the U-Net Architecture With Dual Channel Efficient CNN for Medical Images Segmentation," *arXiv preprint arXiv:2006.00414* (2020), https://doi.org/10.48550/arXiv.2006.00414.

32. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, DLMIA ML-CDS 2018 2018. Lecture Notes in Computer Science*, ed. D. Stoyanov, Z. Taylor, G. Carneiro, et al., (2018): 11045, https://doi.org/10.1007/978-3-030-00889-5_1.

33. Q. Zhu, B. Du, and P. Yan, "Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation," *IEEE Transactions on Medical Imaging* 39, no. 3 (2020): 753–763, https://doi.org/10.1109/TMI.2019.2935018.

34. H. Wang, P. Cao, J. Wang, O. R. Zaiane, "Uctransnet: Rethinking the Skip Connections in U-Net From a Channel-Wise Perspective With Transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence, Proceedings of the AAAI conference on artificial intelligence* 36, no. 3 2022, 2441–2449, https://doi.org/10.1609/aaai.v36i3.20144

35. J. Qiao, X. Wang, J. Chen, and M. Liu, "Mbutransnet: Multi-Branch U-shped Network Fusion Transformer Architecture for Medical Image Segmentation," *International Journal of Computer Assisted Radiology and Surgery* 18, no. 10 (2023): 1895–1902, https://doi.org/10.1007/s11548-023-02879-1.

36. Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE Press, 2018), 6154–6162, https://doi.org/10.1109/CVPR.2018.00644.

37. Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask Scoring R-CNN," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE Press, 2019), 6402–6411, https://doi.org/10.1109/CVPR.2019.00657.

38. K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, and S. Sun, "Hybrid Task Cascade for Instance Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE Press, 2019), 4974–4983, https://doi.org/10.1109/CVPR.2019.00511.

39. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd," in *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, ed. V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, 11207, (2018): 657–674, https://doi.org/10.1007/978-3-030-01219-9_39.

40. Y. Xiao, Y. Xu, Z. Zhong, W. Luo, J. Li, S. Gao, "Amodal Segmentation Based on Visible Region Segmentation and Shape Prior," in *Proceedings of the AAAI Conference on Artificial Intelligence, Proceedings of the AAAI Conference on Artificial Intelligence* 35 no. 4 2021, 2995–3003, https://doi.org/10.1609/aaai.v35i4.16407

41. M. Feng, C. Yan, Z. Wu, W. Dong, Y. Wang, and A. Mian, "History-Enhanced 3D Scene Graph Reasoning From RGB-D Sequences," *IEEE Transactions on Circuits and Systems for Video Technology* 35, no. 8 (2025): 7667–7682, https://doi.org/10.1109/TCSVT.2025.3548308.

42. X. Shi, Y. Zhang, A. Pujahari, and S. K. Mishra, "When Latent Features Meet Side Information: A Preference Relation Based Graph Neural Network for Collaborative Filtering," *Expert Systems with Applications* 260 (2025): 125423, https://doi.org/10.1016/j.eswa.2024.125423.

43. Y. Cai, X. Sui, G. Gu, and Q. Chen, "Multi-Modal Interaction With Token Division Strategy for RGB-T Tracking," *Pattern Recognition* 155 (2024): 110626, https://doi.org/10.1016/j.patcog.2024.110626.

44. Y. Liu, D. Cheng, D. Zhang, S. Xu, and J. Han, "Capsule Networks With Residual Pose Routing," *IEEE Transactions on Neural Networks and Learning Systems* 36, no. 2 (2025): 2648–2661, https://doi.org/10.1109/TNNLS.2023.3347722.

45. Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-Object Relational Visual Saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 7 (2022): 3688–3704, https://doi.org/10.1109/TPAMI.2021.3053577.

46. D. Guan, Y. Xing, J. Huang, A. Xiao, A. El Saddik, and S. Lu, "S2Match: Self-Paced Sampling for Data-Limited Semi-Supervised Learning," *Pattern Recognition* 159 (2025): 111121, https://doi.org/10.1016/j.patcog.2024.111121.