INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Special Issue: IIoT Sensors in Energy Management | DOI: 10.2478/ijssis-2026-0014 Issue 1 | Vol. 19 (2026)

Lightweight inception-UNet with attention mechanisms for semantic segmentation

Twinkle Tiwari* and
Mukesh Saraswat

Jaypee Institute of Information Technology, Noida, India

*E-mail: 18403018@mail.jiit.ac.in

Received for publication
July 11, 2025.

Abstract

Semantic segmentation is a pivotal step in extracting regions of interest from images to enhance scene understanding. However, this task is challenging when dealing with complex images, where factors such as occlusions, varying lighting conditions, diverse viewpoints, and dynamic human movement introduce substantial obstacles. For efficient segmentation, this paper presents a lightweight inception-UNet with attention mechanism to enhance the model ability to discern crucial information from the input image. Specifically, the inception module in the presented UNet captures the features at multiple levels in the encoder phase. To identify the spatial features of an image effectively, decoder phases incorporate the attention module for dense prediction. Both qualitative and quantitative results demonstrate the superiority of our proposed model by achieving highest Intersection over Union (IoU) scores of 0.8768, 0.9283, 0.8768, 0.9768, and 0.9650 across datasets. An ablation study of the proposed model is conducted along with statistical analysis and computational complexity.

Keywords

semantic segmentation, UNet, attention module

I. Introduction

Semantic segmentation is a fundamental but challenging task in the field of computer vision. It dissects the image at the pixel level, providing categorical information that is beneficial in many real-world applications such as autonomous driving vehicles [1], automated medical imaging [2], computational photography [3], augmented reality [4], human pose detection [5], and web-search engines [6]. Due to its importance in various fields of computer vision, many researchers have implemented the task and tried to address the numerous challenges such as occlusions, unlabeled mask, illumination, and unseen annotation issues that arise during segmentation of the images accurately in a complex environment.

With the evolution of hardware and technology, deep learning allows neural networks to be used in a much larger range of semantic segmentation problems than ever before. Deep neural networks have proven to be extremely effective for semantic segmentation, which essentially means classifying what is in each and every pixel or region of an image. This is arguably the most important piece of functionality for any image-understanding application and is used in a wide range of applications in computer vision as well as in artificial intelligence.

The applications span diverse domains, including but not limited to autonomous vehicle driving [1], color imaging [7], data classification [8], data mining [9], cognitive and computational sciences with applications such as salient object detection and classification, agriculture sciences, speech and text

recognition [10], and intelligent healthcare systems [11], particularly emphasizing on intelligent medical imaging. In contrast to earlier approaches like text-on-forest and random-forest (RF) based classifiers [12], deep learning techniques have enabled more precise and significantly faster semantic segmentation. The overall contribution of the proposed model is given below:

- The proposed lightweight architecture enhances the network's ability to discern crucial information from the input images, enabling improved accuracy and minimal computational resources.
- A multiscale convolution is integrated in inception-based encoder to extract multiscale features that allow the network to capture both fine-grained details and high-level context.
- The integration of attention unit in the decoder side leverages in effective identification of region of interest (ROI) from an image.
- The proposed model can attain superior performance in terms of accuracy and sensitivity without implementing complicated heuristics.

The proposed model underwent rigorous validation against five state-of-the-art segmentation models, namely, UNeT [13], ENet [14], SegNet [15], UNeT with ResNet-18 encoder (UNet- ResNet18) [16], and UNeT with ResNet-34 encoder (UNet-ResNet34) [17] encoder. This comprehensive comparative analysis was conducted by employing five publicly accessible datasets, namely, Autorickshaw Detection Dataset [18], Indian Driving Dataset-Lite (IDD-Lite) [19], Computed Tomography Liver (CT-Liver) [20], martial arts, dancing and sports (MADS) [21], and Oxford-IIIT Pets (PETS) [22]. The chosen datasets exemplify diverse scenarios in complex surroundings by their inherent irregularity and unpredictability. The experimental process encompassed a thorough evaluation, combining both qualitative and quantitative assessments. Seven pivotal parameters, namely, intersection over union (IoU) score, error (Err), specificity (Sp), sensitivity (Ss), accuracy (Acc), F-score, and correlation coefficient (Cc) were employed for the evaluation.

The remaining of this paper is structured as follows: Section II discusses the literature review of the semantic segmentation. Section III offers a detailed explanation of the proposed lightweight deep learning-based segmentation model. Section IV provides a comprehensive overview of the experimental setup, followed by a presentation of the experimental analysis and a detailed discussion of the results in Section V. Finally, Section VI summarizes the work drawn from this study. The following section represents the overview of the existing study related to semantic segmentation.

II. Literature Review

Generally, deep learning-based models can be categorized into four categories, namely, contextual information-based methods, feature enhancement-based methods, recurrent neural network (RNN)-based, and deconvolution-based methods. Context-based methods deal with the crucial information in a scene, which accelerates the segmentation process. For dense prediction, Yu and Koltun [23] proposed a dilated model known as DilatedNet, which aggregates the contextual information at multiple scales without losing resolution or analyzing the rescaled images. However, the model lacks end-to-end simplification and unification. To overcome this, Liu et al. [24] designed ParseNet based on FCN32s. The model is based on enlarging the receptive field to extract global features directly. The model outputs smooth segmentation and accurate results on the PASCAL VOC 2012 test dataset. Lin et al. [25] designed Piecewise based on VGG-16, extracting the patch-wise context by integrating convolutional neural network (CNN) with CRF. In EncNet, proposed by Zhang et al. [26], an additional fully connected layer is added, incorporating a sigmoid activation function. This additional layer is used to generate individual predictions for the presence of object categories. The authors introduced the concept of object context, which enhances object information by leveraging semantic relationships between pixels. In this approach, a dense relation matrix is employed as a substitute for the binary relation matrix. Dense relation network (DRN) and context-restricted loss (CRL), as proposed by Zhang et al. [27], combined both global and local context information, resulting in improved segmentation accuracy. Surpassing the global and local context information, Zhang et al. [28] proposed a model based on the detection of co-occurrent features, which harnesses co-occurrent features to provide finely detailed representations. Furthermore, Ding et al. [29] noticed that the unique shapes and intricate arrangement of objects in images can hinder the effectiveness and efficiency of context aggregation. To overcome the same, the author suggested a model that utilizes semantic masks that vary in scale and shape for individual pixels. These semantic masks are created using paired and shape-variant convolution techniques and generate the contextual region for each pixel accordingly.

In a traditional CNN pipeline, feature-enhancement-based deep layers extract features that are rich in semantics but lack spatial features due to the process of pooling and strided convolution. Conversely, shallow layers yield features with a heightened awareness of finer details like stronger edges, and so on. The effective combination of deep and shallow layers enhances the performance of semantic segmentation. Long et al. [30] proposed a fully convolutional network (FCN) that enhances the features through skip connections between prediction features and mid-layer features and significantly increases the final resolution resulting in improvement in mean-intersection over union accuracy. Ronneberger et al. [13] emphasized the importance of skip connection and proposed UNet gaining considerable attention in the field of medical image analysis [31, 32]. Furthermore, to address the spatial information limitations in 2D-based models along the z-dimension, Li et al. [33] extended UNet with 2D-DenseUNet and 3D-DenseUNet, and incorporated a fusion block to combine the intraslice and interslice features. UNet also found extensions in other applications, including natural image segmentation with stacked UNets [34], introducing residual blocks [35] and proposing the hyper-column representation [36], which concatenates features from various CNN pipeline layers for final inferences. RefineNet, introduced by Lin et al. [37], used the long-range residual connections aiming to yield the high-level resolution prediction. The model employs a multifold refinement process that progressively enhances spatial details in feature maps. However, it is worth noting that RefineNet is relatively computationally expensive [38]. Pohlen et al. [39] proposed a full-resolution residual network that is based on the amalgamation of pixelwise accuracy with a mutiscale-based context. The coupling is achieved by separating the network into two sub-streams: the pooling preprocessing stream for high-level semantics (low resolution) and the residual preprocessing stream for detailed preservation (high resolution), which collaborates with features obtained from the latter. The models mentioned above primarily involve designing networks to connect and fuse different feature types to achieve feature enhancement. Collectively, these methods predominantly revolve around the design of networks aimed at connecting and amalgamating different feature types to realize feature enhancement.

RNNs have demonstrated promising capabilities in processing sequential data, including text and speech [40, 41], and have found applications in semantic segmentation. Pinheiro and Collobert [42] introduced the recurrent convolutional neural network (RCNN) aimed to train the model end-to-end on raw pixel data, which enables the system to effectively capture intricate spatial relationships at minimal computational cost. Poudel et al. [43] proposed recurrent fully convolutional networks (RFCNs) for segmenting objects in MRI sequences, showcasing the synergy between FCN and RNN. Byeon et al. [44] presented a model foccusing on the challenge of pixel-level segmentation and classification of scene images, employing a purely learning-based methodology that harnesses long short-term memory (LSTM) RNN. Visin et al. [45] introduced an efficient, flexible, architecture ReSeg, built upon the ReNet model. Each layer comprises four RNNs that traverse the image both horizontally and vertically in both directions. Extending the concept of a two-dimensional-RNN, Shuai et al. [46] developed the DAG-RNN method, which exhibits denser connections compared with 2D-RNN, addressing the issue of fading dependencies along long-range paths. To mitigate the problem of fading dependencies and enhancement of long-range connections, Fan and Ling [47] proposed a dense recurrent neural network (DRNN). Dense RNNs, characterized by their interconnectedness with every pair of image elements, can encapsulate more intricate contextual dependencies for each image unit. This model employs an attention mechanism [48] to assign greater weight to beneficial dependencies while diminishing the significance of irrelevant ones.

The inception of deconvolution-based semantic segmentation marked a significant milestone in computer vision research, with its origin attributed to Noh et al. [49], under the moniker “DeconvNet.” This method is hinged upon the symmetrical encoder–decoder architectural paradigm. Within the encoder component, DeconvNet methodically extracts semantic features, albeit at a reduced resolution due to the application of max pooling. In the decoder module, an unpooling operator effectively employs the saved indices to upsample the low-resolution feature map, thereby transforming it into a high-resolution counterpart. For performing pixelwise semantic segmentation, Badrinarayan et al. [50] designed an efficient model known as SegNet. Furthermore, the decoder employs the pooling indices to execute the nonlinear upsampling resulting in dense feature maps. Fourure et al. [51] proposed “GridNet” which adopted a grid pattern, facilitating the operation of multiple interconnected streams operating at various spatial resolutions. Kendall et al. [52] expanded the SegNet framework into a probabilistic-based pixelwise segmentation model known as “Bayesian-SegNet.”

The model can predict class labels for individual pixels while also quantifying the model's uncertainty. Furthermore, to facilitate the smooth flow of information and gradient propagation across the network, Fu et al. [53] introduced the Stacked Deconvolution Network (SDN). The proposed model ensured the precise localization information recovery due to its sequential layer structure.

From the literature, it has been witnessed that feature-enhancement-based deep learning models can indeed yield benefits in terms of improving feature representations and ultimately boosting the performance of downstream tasks. However, it is crucial to remain mindful of the accompanying drawbacks and factors to consider. Therefore, a meticulous evaluation of these methods is imperative, considering the trade-offs and their appropriateness for specific datasets and problem contexts. In a similar vein, context-based deep learning methods have showcased impressive capabilities across a spectrum of applications, ranging from natural language understanding to computer vision. A mindful approach can be followed while employing context-based methods in real-world applications as they struggle to deal with sequential data and have high inference time. Furthermore, deconvolution-based methods output high-resolution masks but suffer from slow convergence rate and low interpretability where the segmentation of the object depends upon the global contextual information, hence, making them trivial to be used in the decision-making process.

To overcome the various limitations of deep learning-based models, this paper proposed a new attention-inception-based lightweight modified UNet model (LWA-MoDUNet), specifically developed for semantic segmentation of objects within diverse images of scenes characterized by congested and unstructured patterns.

III. Proposed Model

In this paper, the proposed model draws inspiration from the family of UNet architectures. UNet is a fully CNN architecture meticulously designed for the specific purpose of semantic segmentation. The architecture consists of an encoding phase and a decoding phase that are symmetrically connected. The central part of UNet is the bottleneck, which serves as a bridge between the encoder and the decoder. UNet incorporates skip connections that concatenate feature maps from the contracting path to the corresponding layers in the expansive path. These connections help to preserve spatial information and enable precise segmentation. However, UNet struggles with extremely large images or cases where objects have complex shapes and interactions. Furthermore, UNet may encounter challenges in capturing intricate image features that could potentially enhance segmentation accuracy, particularly when dealing with images portraying road traffic scenarios marked by congestion and unstructured traffic patterns. As a result, this research introduces the inception and attention modules in the U-Net architecture and presents a new variant termed "lightweight attention-based modified-UNet" (LWA-ModUNet). Figure 1 represents the architecture of the proposed LWA-ModUNet model. The inception module at the encoder side helps in extracting multiscale features and captures fine-grained details with high-level context while the attention module captures the contextual information by considering relationships between pixels or regions within an image. The objective is to enhance the network's efficacy in accurately segmenting images, particularly in scenarios involving complex patterns. Both the encoder and decoder parts of the proposed model are described in the following sections.

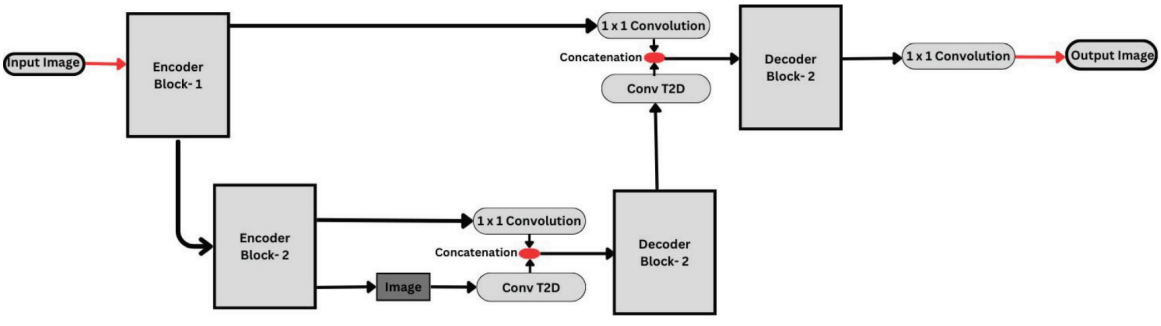


Figure 1: Framework of the proposed LWA-MoDUNet.

a. Encoding phase

The encoding phase extracts features and reduces spatial dimensions. For this, the proposed LWA-MoDUNet employs only two encoding blocks for the encoding path. Each encoding block performs a comprehensive set of operations by incorporating multiple convolution layers, and max-pooling layers, with an inception block, and a batch normalization. Figure 2 illustrates the structure of an encoder block.

From Figure 2, it is seen that the input data undergoes processing through three paths. The first path includes data processing through a 3 × 3 convolution layer and then through the inception module. The inception block consists of multiple kernels that capture features at different receptive fields. This ensures that the model has a better context than the standard convolution layer. The complete block of inception block is depicted in Figure 3 The features extracted from the inception module are passed to the decoding phase as well as within the same encoder block which enables the network to efficiently differentiate between the multiple-object boundaries and context, prevents information loss and mitigates the noise and occlusions in an image. In the second path, the 1 × 1 convolution layer, batch normalization, and 2 × 2 max-pooling process the input data. The processed data are included with extracted features from the inception module. This encourages the model to ensure that the learned features are more enriched than the original input in varied conditions for better scene understanding. Finally, the third path employs 2 × 2 max-pooling to downsample the input data and concatenate with the features from the second path. This acts as an input reinforcement to the next block and also aids the further layers to always have a comparative sense

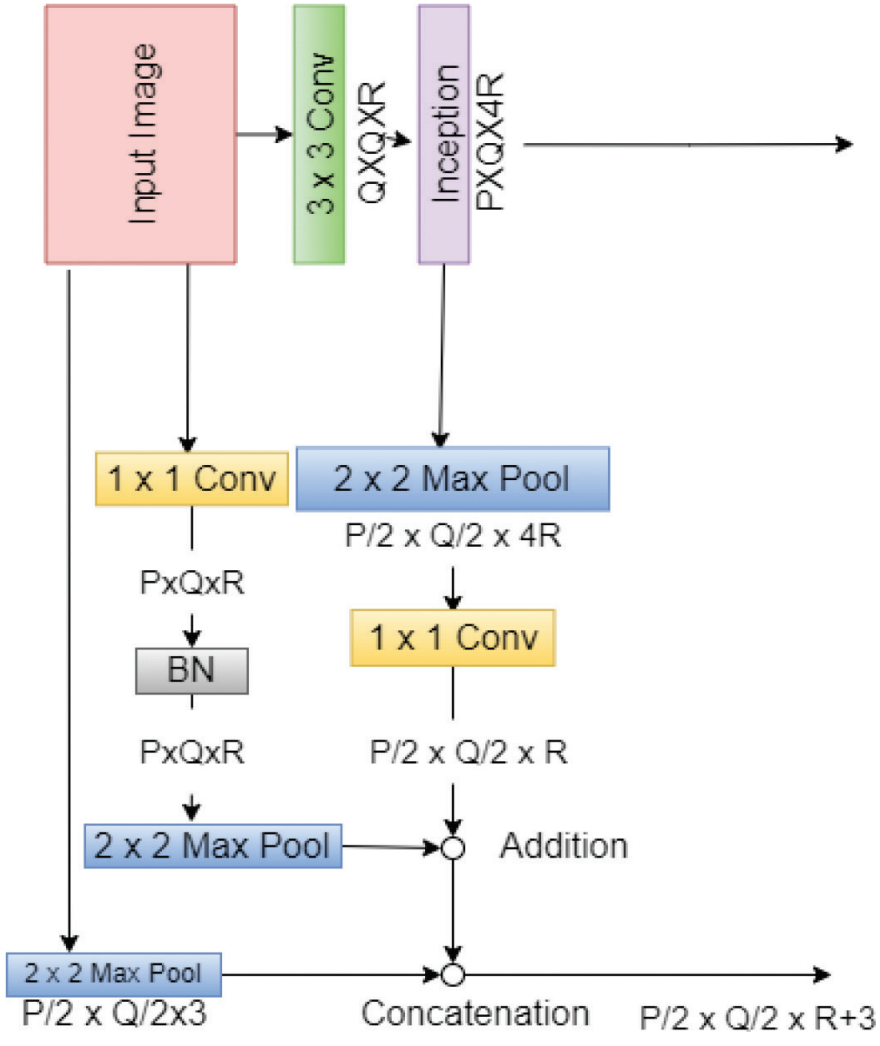


Figure 2: Architecture of the encoder block of the proposed LWA-ModUNet.

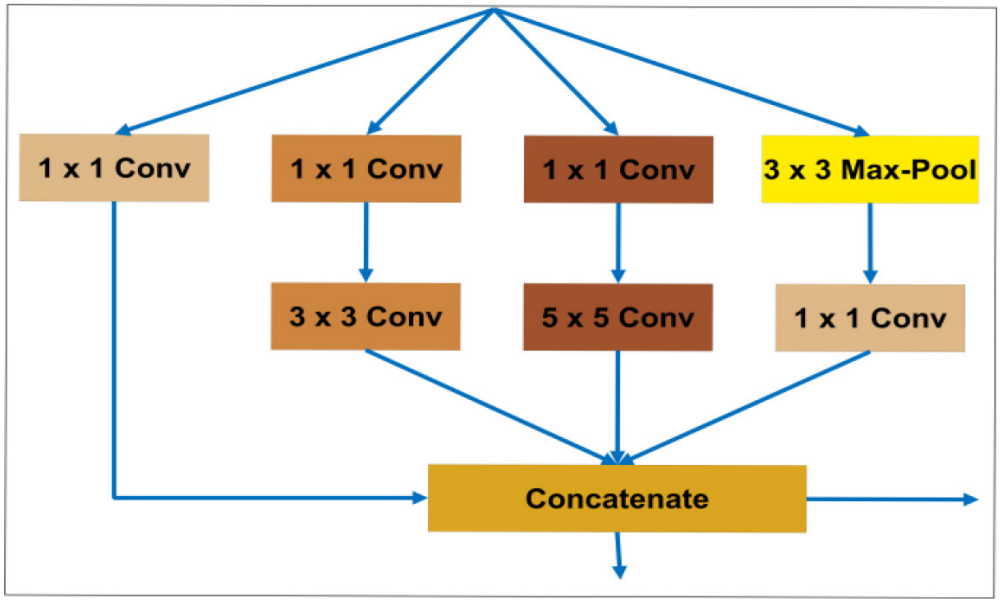


Figure 3: Architecture of the inception block of the proposed LWA-ModUNet [54].

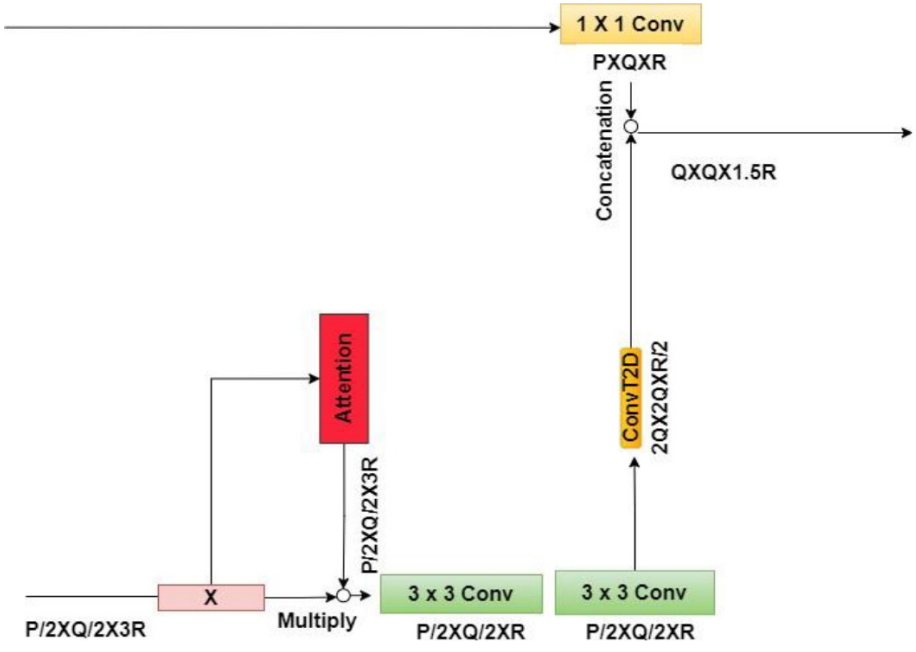


Figure 4: Decoder architecture of the proposed LWA-MoDUnet.

between the learned and the original features. This architecture of the encoder block enhances the network's ability to discern crucial information from input images and distinguish between different objects in better scene understanding.

b. Decoding phase

The decoding phase upsamples the feature space to produce a high-resolution segmentation mask. In the proposed architecture, only two decoder blocks are incorporated in the decoding phase. Each decoder

block employs an attention module, a deconvolution layer, and convolution layers to perform the feature upsampling. The illustrative structure of the decoder block is depicted in Figure 4. In the decoder block, the features from the encoding phase are passed to the attention module to excite the important features at the channel level. The skip connections used in UNet provides spatial information from the decoder to the corresponding encoder but they bring along the poor feature representation. Hence, the modified inception has been integrated to enhance UNet with attention module as depicted in Figure 4 to reduce the complexity during feature extraction in an image.

The attention block [55], as depicted in Figure 5, focuses on attention coefficients $\beta_i \in [0, 1]$, which identifies the significant features and then assign the weights to them. The attention coefficient is element wise multiplied with generated input feature maps, that is, $m^l_i,\ c = m^l_i,\ c \cdot = \beta^l_i$, to produce the result from attention gates. In a default configuration, the value of single scalar attention is determined for each pixel vector $m^l_i \in R^{Nl}$ in L layer for N number of feature maps. The gating vector $v_i \in R^{Nv}$ that contains the contextual information significantly reduces the influence of less relevant features. The gating coefficients are obtained by using additive attention, which has been equated in Eqs (1) and (2):

$$ql_{att} = \Psi^T \sigma_1(W^T m^l_i + W^T v_i + b_v) + b_\psi \quad (1)$$

$m \qquad v$

$$\beta^l,i = \sigma_2(q^l_{att}(m^l_i, v_i; \Theta_{att})) \quad (2)$$

where $\sigma_2(m_{i,c})$ denotes the sigmoid activation, Θ_{att} depicts the parameters used to characterize attention gates that include $W_m \in R^{Nl} \times N\text{int}$ and $W_v \in R^{Nl \times N\text{int}}$. For input tenors, a channel-wise convolution of $1 \times 1 \times 1$ has been performed to calculate the linear transformations. $b_\psi \in R$ and $b_v \in R^{N\text{int}}$ denote the bias used.

The sequential use of softmax activation [56], used in attention block, is to normalize the attention coefficients. Due to the implementation of sigmoid activation, the proposed model has shown better convergence during training. Within the architecture, skip connections are utilized to emphasize important features. The updating rule for convolution parameters from the lower layers (l – 1) is shown in Eq. (3):

$$\partial(\hat{m}^l)$$

$$\partial(\beta^l f\ (m^{l-1}; \phi^{l-1}))$$

$$\partial(f\ (m^{l-1};\ \phi^{l-1}))$$

$$\partial(\beta^l)$$

$$i = i = \beta^l + i\ m^l \quad (3)$$

$$\partial(\phi^{l-1})$$

$$\partial(\phi^{l-1})$$

$$i\ \partial(\phi^{l-1})$$

$$\partial(\phi^{l-1})^{\ i}$$

This focuses on capturing the relevant contextual information between the regions of an image. Furthermore, the attention module can effectively fuse features from multiple scales within an image. This benefits in capturing both local and global contexts, making it suitable for objects of varying sizes, which is particularly valuable for distinguishing objects from their surroundings and handling complex scenes. Consecutively, a couple of standard convolution layers are followed by a deconvolution layer in a decoder block. This advantage in suppressing the noise in the predicted segmentation mask and pixelwise prediction is for accurate segmentation of complex images. Finally, the decoder block has a 1×1 convolution layer that processes the features from the skip connections and performs concatenation with the upsampled features by the deconvolution layer. This

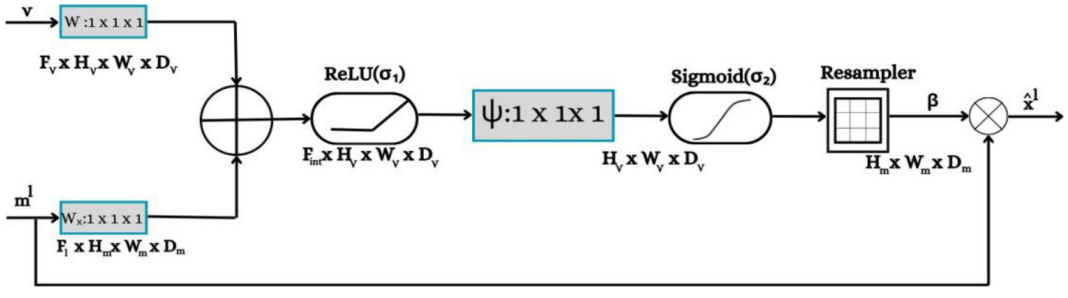


Figure 5: Diagram of additive attention gate [55].

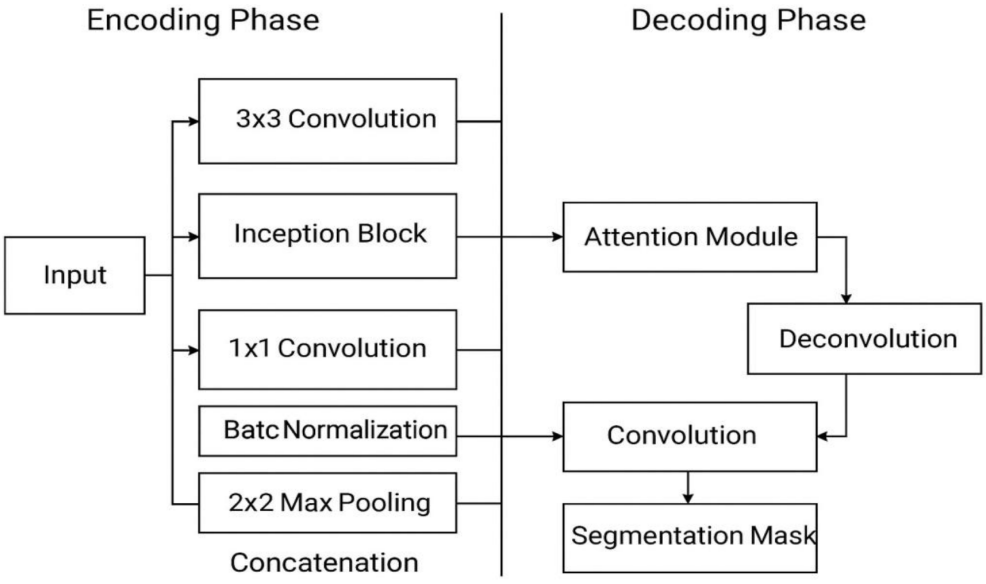


Figure 6: Flowchart of the proposed LWA-MoDUNet depicting the encoding and decoding phases with inception and attention modules.

helps the model to retain the coarse features from the encoder side. At the end of the decoding phase, a 1×1 convolution layer is applied to generate the output segmentation mask. The flowchart of the proposed model is shown in Figure 6, depicting the generation of the segmentation mask.

IV. Experimental Setup

a. Considered datasets

To assess the performance of the proposed semantic segmentation method, five publicly accessible datasets, namely, Autorickshaw Detection Dataset [18], IDD-Lite [19], CT-Liver [20], MADS [21], and PETS [22], are used. The used datasets contain objects within diverse images of scenes characterized by congested and unstructured patterns, which make them ideal for the assessment. The Autorickshaw Detection Dataset [18] consists of 1,000 images and has been publicly released by the Indian Institute of Information Technology, Hyderabad, India, in conjunction with the Autorickshaw Detection Challenge. The IDD-Lite [19] comprises 1,404, 204, and 408 training, validation, and testing images, respectively. This dataset is organized into seven distinct classes, namely, drivable regions, non-drivable areas, living entities, vehicles, roadside objects, distant objects, and the sky. The PETS [22] dataset consists of 37 sets of various breeds of cats and dogs in varied lighting, poses, and background conditions. Each set of breeds contains approximately 200 images, resulting in a total of 7,349 images that have been divided into training, testing, and validation datasets. From 7,349 images, 3,312 have been taken for training, 368 for validation, and 3,669 for testing purposes. The CT-Liver [20] dataset consists of 232 images in which 152 images have been considered for training, 35 for testing, and 45 for validation purposes. The MADS [21] dataset is a collection of human movement images of martial arts, dance, and sports movements in varied angles for pose detection. For training, validation, and testing of segmentation, 861, 152, and 179, respectively, images have been taken. Figure 7 depicts the representative images from each dataset along with the respective ground-truth (GT) images.

The quality of the generated segmentation masks for the abovementioned datasets has been comprehensively evaluated against seven performance parameters, namely, IoU score, error (Err), specificity (Sp), sensitivity (Ss), accuracy (Acc), correlation coefficient (Cc), and F-score.

b. Implementation details

During training, each image is first resized to $p \times q$, where $p = 256$, $q = 256$, and randomly flipped horizontally, and croppped $m \times n$, where $m = 128$ and $n = 128$. The input images were subjected to different augmentations to avoid distortions and over-fitting

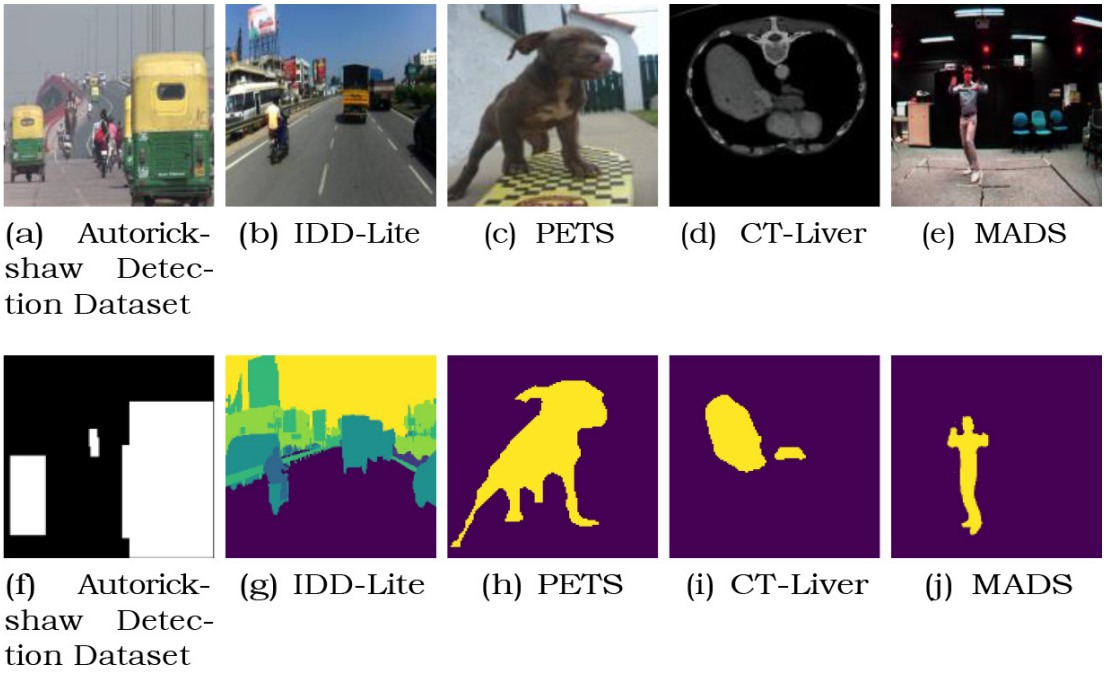


Figure 7: Sample images of considered datasets (Row 1) along with their corresponding GT (Row 2). GT, ground truth; IDD-Lite, Indian driving dataset-lite; MADS, martial arts, dancing and sports.

such as cropping, zoom-in/out, rotation, lightning, and affine (scaling and translation). For the data normalization, each image is subtracted from the mean and divided by standard deviation. Each input image is resized to $128 \times 128 \times 3$ dimensions, where 3 represents the RGB channel. For the autorickshaw, PETS, MADS, and CT-Liver datasets, the loss function employed is the flattened binary cross-entropy loss [57], also known as flattened sigmoid cross-entropy loss. As the IDD-Lite is a multi-class dataset, cross-entropy loss [58] is utilized. In LWA-MoDUNet, since both the instances of feature and target networks are convolutional layers, inspired from the idea mentioned above to rectify the gradient vanishing issue, we have used Xavier initialization [59] for convolutional layers to speed up the training process. Surprisingly, no existing backbone is used as all of the model is trained from scratch by training over 200 epochs. The Adam optimizer [60] with early stopping and step learning rate scheduler with initial learning rate as $1E-3$ is used. All other hyperparameters are considered as default. At the output, bilinear interpolation is employed to resize the predicted saliency maps. The model is implemented in PyTorch 2.0.0 and trained on an Intel Xeon 2.2 GHz CPU (24GB RAM) and an NVIDIA A100-SXM4 (40 GB memory), which are a Hexa-core, 12-thread PC.

c. Ablation study

To explore the robustness of the proposed model, an ablation study [61] is conducted through a quantitative analysis of the proposed model with and without inception and attention modules. The hyperparameter settings of the proposed model are detailed in Section "Implementation details". Experimental evaluation is performed on the autorickshaw dataset by considering four metrics, that is, IoU score, Error, Accuracy, and F-score. The experimental results are depicted in Table 1.

Table 1: Ablation study of LWA-MoDUNet on autorickshaw dataset

Measure	LWA-MoDUNet	W/o inception module	W/o attention
IoU score	0.8768	0.6729	0.8173
Error	0.0663	0.1452	0.0927
Accuracy	93.3691	78.7966	86.4217
F-score	0.9344	0.7913	0.8786

IoU, intersection over union.

It can be observed from Table 1 that the inclusion of the inception module has significantly enhanced the performance of the LWA-MoDUNet (proposed model) as compared with its absence. Furthermore, the model has attained superior performance with the use of the attention module as shown in Table 1. Therefore, this study supports the architectural advantages of the proposed model and is evident from the efficient results.

V. Experimental Analysis

In this paper, a comprehensive evaluation of the proposed LWA-MoDUNet is conducted both qualitatively and quantitatively by performing a comparison with SOTA models, including U-Net, UNet-ResNet18, UNet-ResNet34, SegNet, and ENet.

For qualitative assessment, Figures 8–12 showcase the segmentation outcomes of these methods on images sampled from PETS, MADS, autorickshaw, IDD-Lite, and CT-Liver datasets, respectively. In the visual representations, the first two columns correspond to an unaltered image and the respective GT. For a thorough quantitative assessment, seven crucial parameters, namely, IoU score, sensitivity, accuracy, error, specificity, correlation coefficient, and F-score, have been considered, and generated values are outlined in Tables 2–6 corresponding to the autorickshaw, IDD Lite, PETS, MADS, and CT-Liver datasets, respectively. In Tables 2–6, the numerals highlighted in bold signify the most optimal results. The IoU score served as the primary metric for evaluating segmentation accuracy. Additionally, to visualize the variations in IoU scores during the 200 training and validation epochs for the models under consideration, line plots are presented in Figures 13–17 for the autorickshaw, IDD-Lite, PETS, MADS, CT-Liver datasets, respectively. Furthermore, this paper conducts a nonparametric Wilcoxon signed-rank test [62] to statistically validate the segmentation results of the LWA-MoDUNet with the compared SOTA models. Extending the analysis further, we include Figure 18 showing the trade-off between type 1 and type 2 errors for the proposed and compared state-of-the-art models. Moreover, to evaluate the complexity of LWA-ModUNet, its parameter count and inference time were analyzed in comparison to five state-of-the-art deep learning models.

a. Discussion of results

In this section, detailed analysis of the results obtained are presented. Examining Figure 8, it becomes apparent that the comparative SOTA, especially UNet

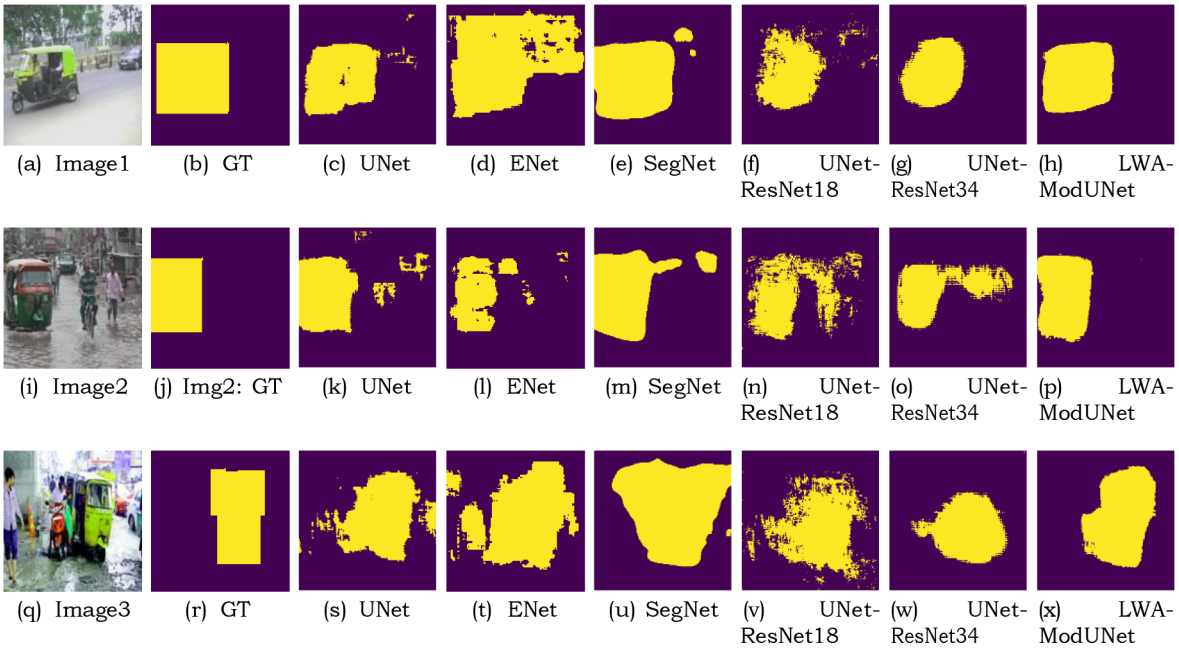


Figure 8: Generated output masks by UNet, E-Net, SegNet, UNet with ResNet-18 encoder, UNet with ResNet-34 encoder, and LWA-ModUNet models on exemplary images of autorickshaw dataset. GT, ground truth.

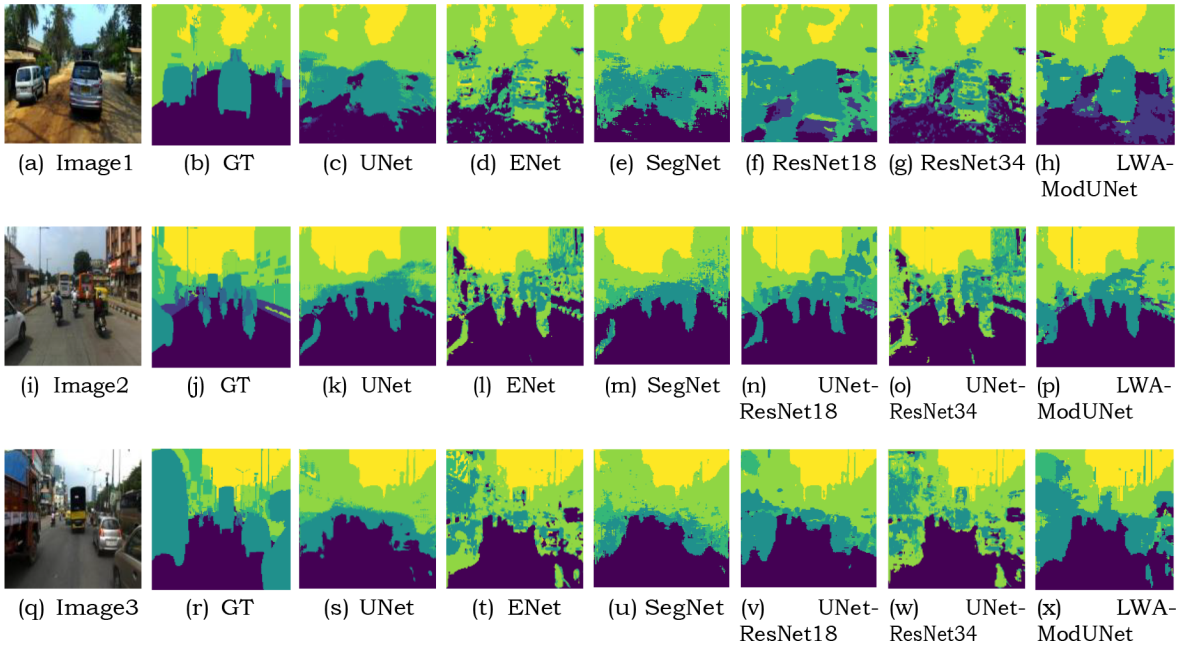


Figure 9: Generated output masks by UNet, E-Net, SegNet, UNet with ResNet-18 encoder, UNet with ResNet-34 encoder, and LWA-ModUNet on exemplary images of IDD Lite dataset. GT, ground truth; IDD-Lite, Indian driving dataset-lite.

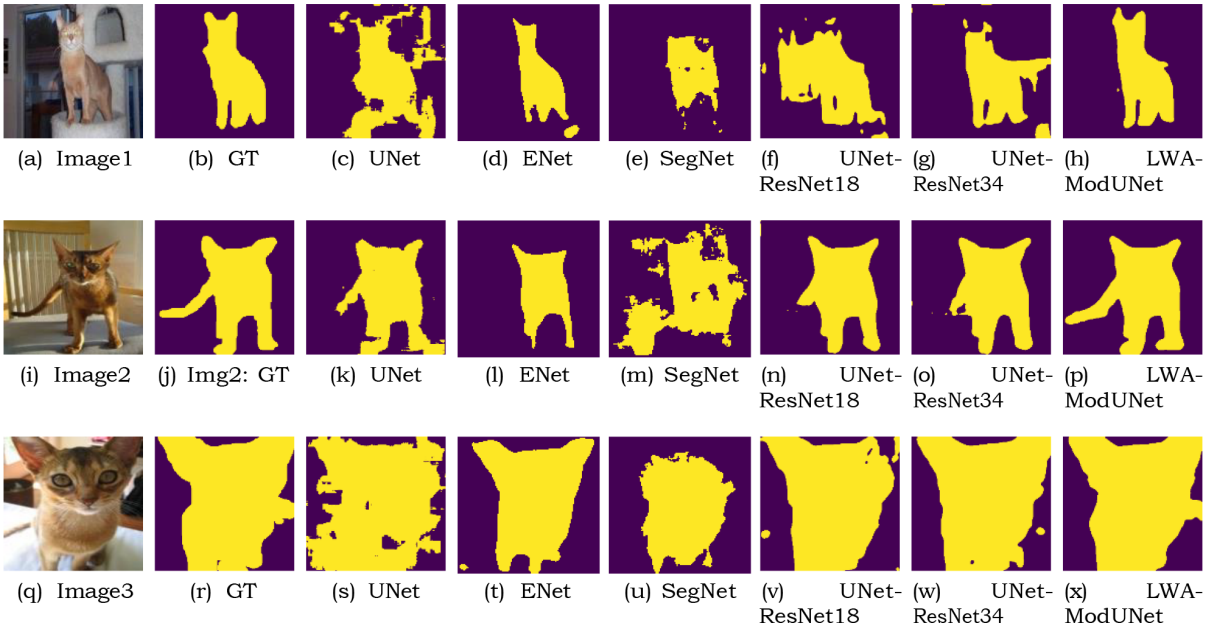


Figure 10: Generated output masks by UNet, E-Net, SegNet, UNet with ResNet-18 encoder, UNet with ResNet-34 encoder, and LWA-ModUNet models on exemplary images of PETS dataset. GT, ground truth.

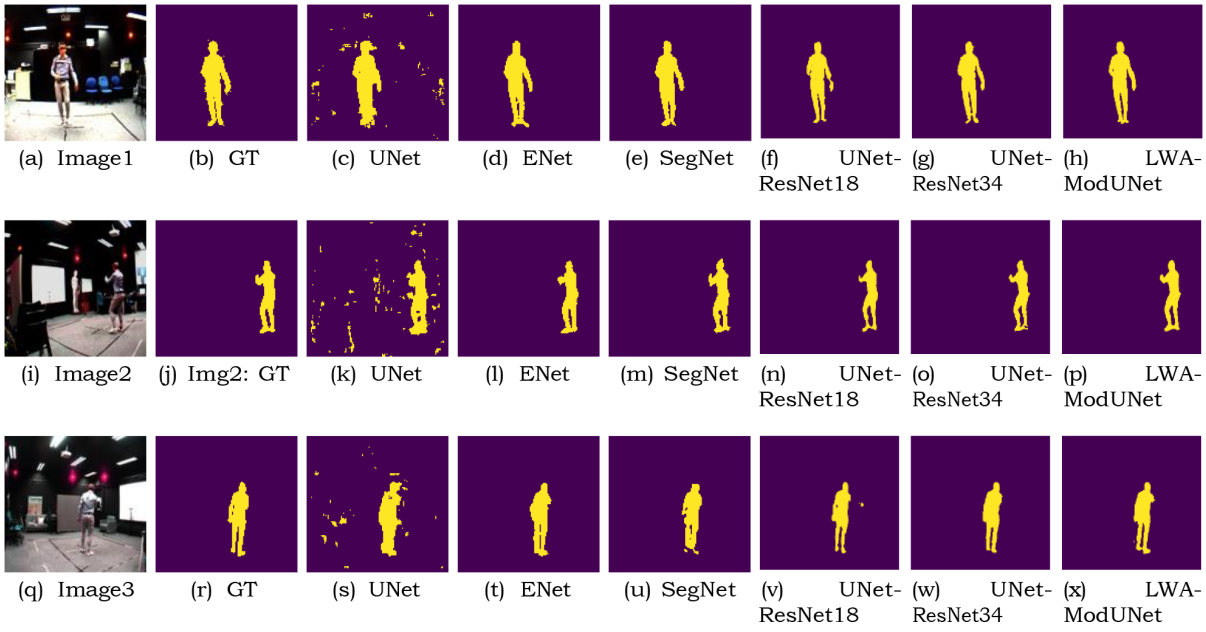


Figure 11: Generated output masks by UNet, E-Net, SegNet, UNet with ResNet-18 encoder, UNet with ResNet-34 encoder, and LWA-ModUNet models on exemplary images of MADS dataset. GT, ground truth; MADS, martial arts, dancing and sports.

Figure 12: Generated output masks by UNet, E-Net, SegNet, UNet with ResNet-18 encoder, UNet with ResNet-34 encoder, and LWA-ModUNet models on exemplary images of CT-Liver dataset. GT, ground truth.

Table 2: Comparison of key performance indicators of LWA-MoDUNet and SOTA models on autorickshaw dataset

Measures	UNet	UNet-ResNet34	UNet-ResNet18	E-Net	SegNet	LWA-MoDUNet
IoU score	0.7665	0.8459	0.8356	0.8480	0.7592	**0.8768**
Err	0.1326	0.0849	0.0912	0.0857	0.1403	**0.0663**
Acc	86.7412	91.5061	90.8837	91.4349	85.9677	**93.3691**
Sp	0.8697	0.9303	0.9241	0.9514	0.8789	**0.9429**
Ss	0.8651	0.9009	0.8946	0.8829	0.8423	**0.9248**
F-score	0.8678	0.9165	0.9104	0.9177	0.8631	**0.9344**
Cc	0.7348	0.8306	0.8182	0.8315	0.7203	**0.8676**

Acc, accuracy; Cc, correlation coefficient; Err, error; IoU, intersection over union; Sp, specificity; Ss, sensitivity.

Table 3: Comparison of key performance indicators of LWA-MoDUNet and compared models on IDD-Lite dataset

Measures	UNet	UNet-ResNet34	UNet-ResNet18	E-Net	SegNet	LWA-MoDUNet
IoU score	0.6031	0.6174	0.5981	0.566	0.3076	**0.9283**
Acc	0.9203	0.9398	0.9356	0.9321	0.8971	**0.9616**
Err	0.0716	0.0601	0.0643	0.0679	0.1028	**0.00435**
Sp	0.9500	0.9484	0.9469	0.9395	0.8975	**0.9328**
Ss	0.8534	0.8617	0.8472	0.8669	0.8896	**0.9436**
F-score	0.7056	0.7635	0.7485	0.7229	0.4705	**0.8909**
Cc	0.6794	0.7371	0.7186	0.6979	0.4965	**0.8153**

Acc, accuracy; Cc, correlation coefficient; Err, error; IDD-Lite, Indian driving dataset-lite; IoU, intersection over union; Sp, specificity; Ss, sensitivity.

Table 4: Comparison of key performance indicators of LWA-MoDUNet and compared models on PETS dataset

Measures	UNet	UNet-ResNet34	UNet-ResNet18	E-Net	SegNet	LWA-MoDUNet
IoU score	0.7665	0.8459	0.8356	0.8480	0.7592	**0.8768**
Err	0.1326	0.0849	0.0912	0.0857	0.1403	**0.0663**
Acc	86.7412	91.5061	90.8837	91.4349	85.9677	**93.3691**
Sp	0.8697	0.9303	0.9241	0.9514	0.8789	**0.9429**
Ss	0.8651	0.9009	0.8946	0.8829	0.8423	**0.9248**
F-score	0.8678	0.9165	0.9104	0.9177	0.8631	**0.9344**
Cc	0.7348	0.8306	0.8182	0.8315	0.7203	**0.8676**

Acc, accuracy; Cc, correlation coefficient; Err, error; IoU, intersection over union; Sp, specificity; Ss, sensitivity.

with ResNet-18 encoder and E-Net, struggled in the segmentation of the input images, exhibiting noise, poor-boundary definitions, and gaps. While LWA-MoDUNet can generate similar segmentation masks that best match the GT. The resultant segmentation masks have sharp and smooth boundaries with minimal noise. In Figure 9, noticeable discrepancies were observed in the segmented results of E-Net and Seg- Net. Conversely, LWA-MoDUNet consistently generated outputs closely resembling their respective GT images for the examined figures. Furthermore, the output images of the LWA-MoDU-Net demonstrated

Table 5: Comparison of key performance indicators of LWA-MoDUNet and compared models on MADS dataset

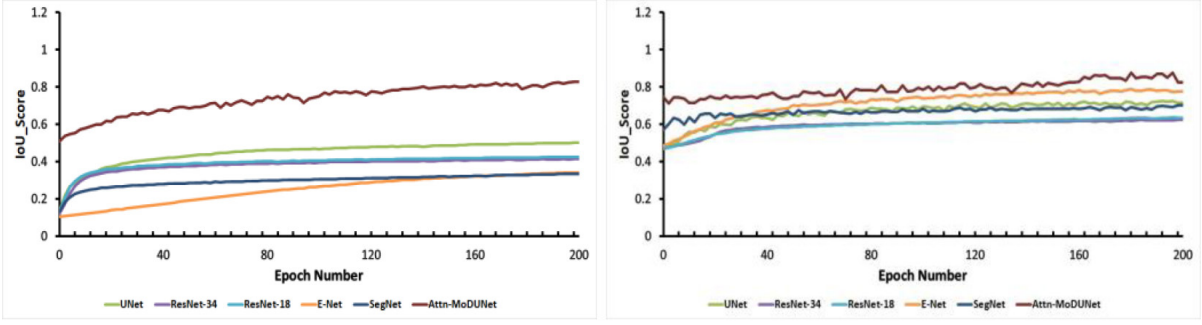
Measures	UNet	UNet-ResNet34	UNet-ResNet18	E-Net	SegNet	LWA-MoDUNet
IoU score	0.8968	0.9448	0.9084	0.9609	0.9721	**0.9768**
Err	0.0545	0.0290	0.0496	0.0201	0.0142	**0.0118**
Acc	94.5520	97.1042	95.0428	97.9919	98.5781	**98.8225**
Sp	0.9464	0.9905	0.9818	0.9872	0.9916	**0.9934**
Ss	0.9447	0.9531	0.9229	0.9728	0.9801	**0.9832**
F-score	0.9456	0.9716	0.9520	0.9801	0.9859	**0.9883**
Cc	0.8910	0.9428	0.9028	0.9599	0.9716	**0.9765**

Acc, accuracy; Cc, correlation coefficient; Err, error; IoU, intersection over union; MADS, martial arts, dancing and sports; Sp, specificity; Ss, sensitivity.

Table 6: Comparison of key performance indicators of LWA-MoDUNet and compared models on CT-Liver dataset

Measures	UNet	UNet-ResNet34	UNet-ResNet18	E-Net	SegNet	LWA-MoDUNet
IoU score	0.8839	0.9445	0.9565	0.9349	0.9495	**0.9650**
Err	0.0616	0.0287	0.0223	0.0338	0.0262	**0.0178**
Acc	93.8400	97.1308	97.7686	96.6167	97.3808	**98.2162**
Sp	0.9384	0.9758	0.9811	0.9719	0.9845	**0.9855**
Ss	0.9384	0.9669	0.9743	0.9605	0.9635	**0.9788**
F-score	0.9384	0.9714	0.9778	0.9664	0.9741	**0.9822**
Cc	0.8768	0.9427	0.9554	0.9324	0.9478	**0.9643**

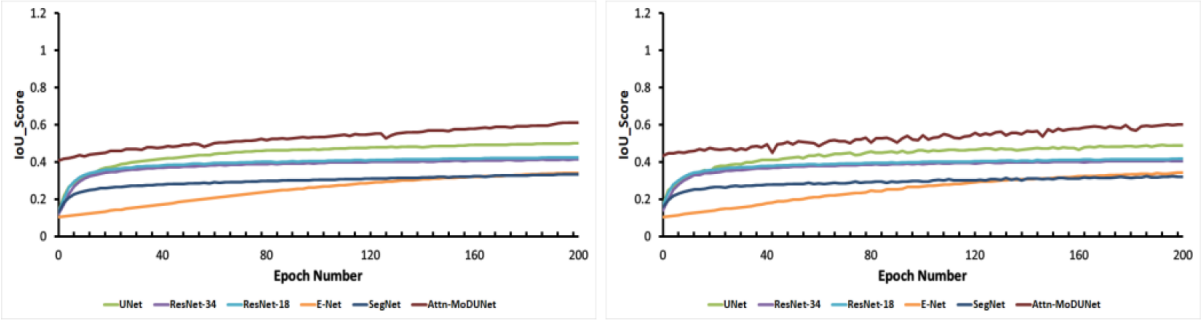
Acc, accuracy; Cc, correlation coefficient; Err, error; IoU, intersection over union; Sp, specificity; Ss, sensitivity.



(a) Training IoU Score (b) Validation IoU Score

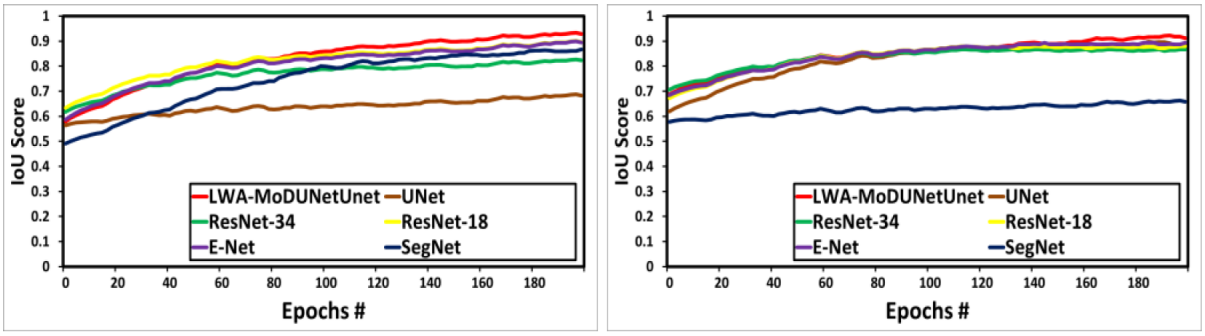
Figure 13: Comparative analysis of IoU score on training and validation images taken from autorickshaw dataset. IoU, intersection over union.

comparatively lower levels of distortion. Similarly, from Figure 10, it can be observed that the results obtained from the LWA-MoDUNet are approximately similar to the GT of the respective images under consideration. The segmentation masks have accurate boundaries even in the cluttered backgrounds, while ResNet18 and ENet models generated segmentation masks that were partially matching with



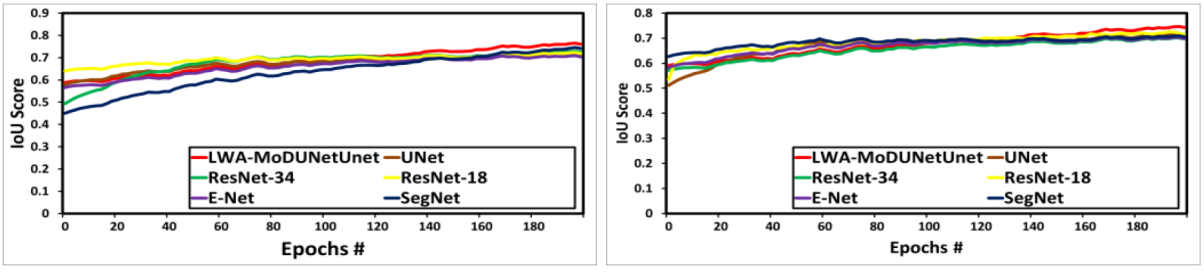
(a) Training IoU Score (b) Validation IoU Score

Figure 14: Comparative analysis of IoU score on training and validation images taken from IDD-Lite dataset. IDD-Lite, Indian driving dataset-lite; IoU, intersection over union.



(a) Training IoU Score (b) Validation IoU Score

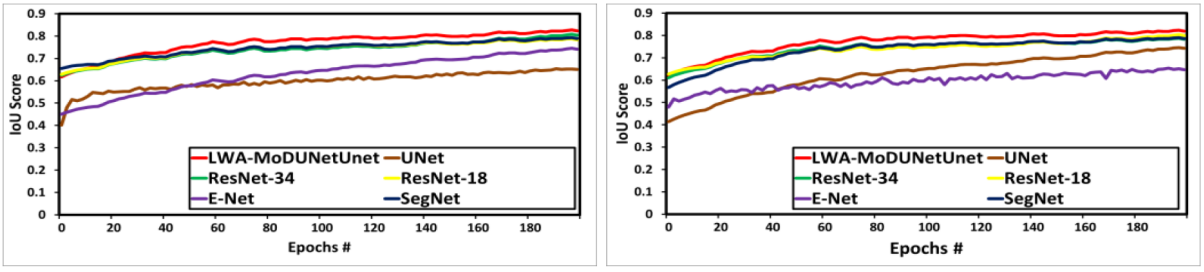
Figure 15: Comparative analysis of IoU score on training and validation images taken from PETS dataset. IoU, intersection over union.



(a) Training IoU Score (b) Validation IoU Score

Figure 16: Comparative analysis of IoU score on training and validation images taken from MADS dataset. IoU, intersection over union; MADS, martial arts, dancing and sports.

the GT of the original image. Conversely, UNet and SegNet underperformed by generating distorted output masks. On the contrary, considering Figure 11, it can be witnessed that LWA-ModUNet and UNet with ResNet34 have achieved competitive results. The segmented outputs of the earlier one are better than the later one. While UNet underperformed and outputs the distorted masks. Moreover, Figure 12 shows that the considered models have shown average results. The result suggests that the masks produced



(a) Training IoU Score (b) Validation IoU Score

Figure 17: Comparative analysis of IoU score on training and validation images taken from CT-Liver dataset. IoU, intersection over union.

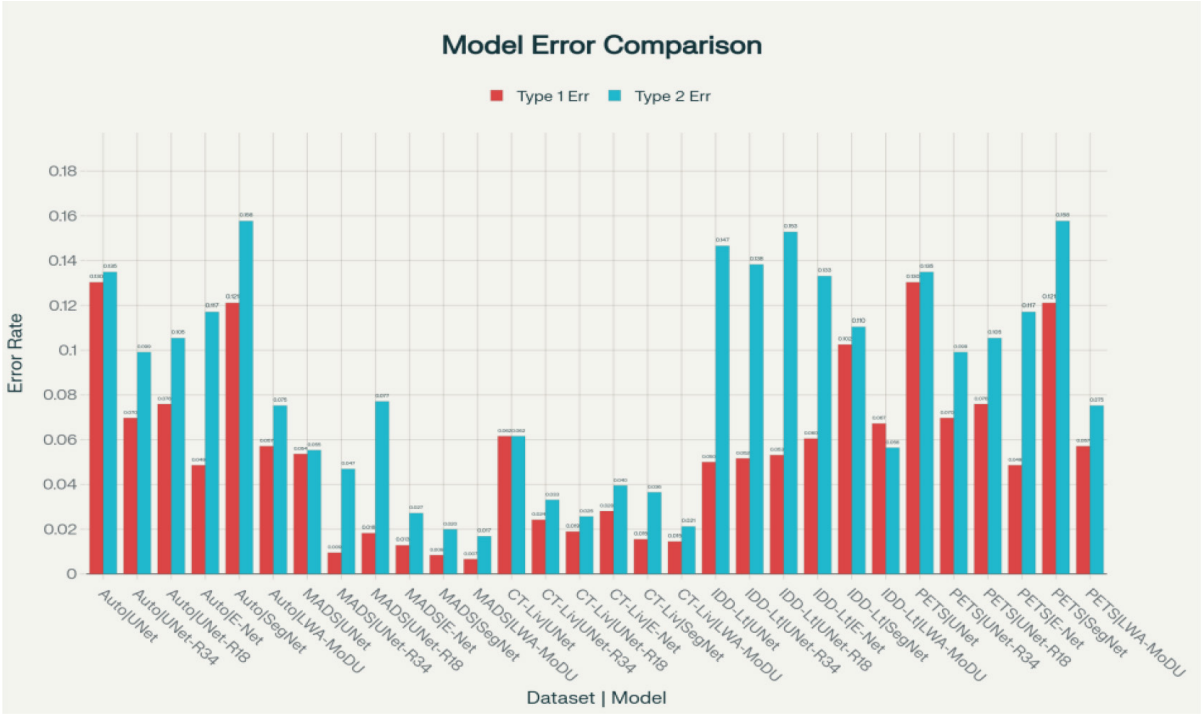


Figure 18: Model type 1 and type 2 errors comparison across datasets. IDD-Lite, Indian driving dataset-lite; MADS, martial arts, dancing and sports.

by LWA-ModUNet are considerably similar to their original GT images. The compared models have generated distorted images and hence cannot segment the image effectively.

Table 2 presents the experimental results on the autorickshaw dataset. UNet with ResNet18/ ResNet34 encoder demonstrate similar performance across all the parameters. However, their performance falls behind that of the other models under consideration, as well as the proposed methods for the autorickshaw dataset. Notably, LWA-MoDUNet outperformed comparative models, achieving the highest IoU-score, that is, 0.8768. Furthermore, LWA-MoDUNet demonstrated superiority across all considered parameters, that is, error as 0.0663, accuracy as 0.9336, specificity as 0.9429, sensitivity as 0.9248, F-score as 0.9344, and correlation coefficient as 0.8676. Similarly, upon analyzing the results for the IDD Lite dataset in Table 3, it is evident that our proposed model exhibited exceptional performance, surpassing the state-of-the-art comparative models, achieving the highest IoU score of 0.9283. The second highest IoU of 0.6174 is achieved by UNet-ResNet34, which shows that the proposed model segments the complex images efficiently.

Similarly, the performance of the compared models and proposed model on the PETS dataset has been illustrated in Table 4. The proposed LWA-ModUNet performed better than the compared SOTA models, indicating the highest IoU score of 0.8768. On other parameters, LWA-ModUNet performed commendably achieving the highest values for accuracy, sensitivity, specificity, F-score, and correlation-coefficiency. According to the result, LWA-MoDUNet is most suitable for accurately segmenting objects in complex images, achieving and maintaining a better balance between precision and recall. Contrarily, SegNet consistently delivers lower performances on all parameters across the metrics, thus making the model a crucial choice for the semantic segmentation of complex images. By contrast, UNet with ResNet34 and UNet with ReSNet18 performed better but remained behind LWA-MoDUNet reporting an IoU score of 0.8459 and 0.8356, respectively.

The performance evaluation of the models on the MADS dataset is illustrated in Table 5. The results depict that the performance of SegNet is comparable to the LWA-MoDUNet in terms of all the considered parameters. In terms of IoU score, LWA-MoDUNet reported 0.9768, which is the highest while SegNet achieved 0.9721. UNet underperformed by achieving the lowest IoU score value of 0.8968. Table 6 illustrates the segmentation results of the compared models on the CT-Liver dataset. The table suggests that the proposed model marked the highest parametric values across the metrics, achieving an IoU score of 0.9651, accuracy of 0.9821, correlation coefficient of 0.9643, specificity of 0.9855, F-score of 0.9822, and sensitivity of 0.9788, respectively, which are the highest. UNet with ResNet34 and UNet with ResNet-18 fall behind LWA-MoDUNet with an IoU score of 0.9445 and 0.9565, respectively. On the contrary, SegNet, and Enet fall slightly behind the ResNet encoders resulting in IoU score of 0.9495 and 0.9349, respectively. Moreover, UNet underperformed in the segmentation of image reporting the lowest IoU score of 0.8839. Hence, based on the segmentation accuracy (IoU score), it can be confidently affirmed that our proposed model represents a substantial improvement, attributable to its enhanced capacity to extract deeper features.

Figures 13–17 demonstrate that LWA-MoDUNet attains the best IoU score, exhibiting a substantial margin of superiority over the compared models for all five compared datasets. Furthermore, for visual illustrations, the graphs for the models' performance at the validation phase provide unequivocal evidence that the proposed method consistently outperforms throughout the entire span of epochs considered. It is worth noting that the proposed model maintains its supremacy in terms of IoU score and outperforms the compared state-of-the-art-methods.

To validate, the same Wilcoxon rank sum test has been conducted by using the IoU scores of the two models across varied sample sizes, that is, 30 and 50. If the *p*-value is lower than the significance level ($\alpha = 0.05$), then the null hypothesis (i.e., both models are statistically similar) is rejected, and the alternative hypothesis (i.e., the proposed model is statistically superior to the compared model) is accepted. Table 7 depicts the *p*-values of the proposed model in comparison to other considered models. Notably, it is clearly evident from the Table 7 that the *p*-value of the compared models is considerably less than the α. Furthermore, it can be observed in Table 7 that the *p*-value is indirectly proportional to the sample size. Therefore, it is pertinent to state that the proposed model performs statistically better segmentation in comparison to the compared models.

Additionally, the results in Figure 18 clearly show that LWA-ModUNet yields fewer Type 1 errors than the compared state-of-the-art models. On the contrary, the proposed model depicted comparable Type 2 errors on PETS and IDD datasets but lower errors on the remaining datasets as compared with the state-of-the-art models. As compared -with the UNet and ResNet-based models LWA-MoDUNet raises less false alarms.

Furthermore, this study analyzed the complexity of these models in terms of parameters and inference time. As illustrated in Table 8, LWA-ModUNet has significantly fewer parameters among the considered

Table 7: Statistical performance analysis of LWA-MoDUNeT against other compared SOTA models

Models	Samples size = 30 *p*-value	Samples size = 50 *p*-value
UNet and LWA-MoDUNet	0.0029 × E–05	3.54 × E–09
ResNet34 and LWA-ModUNet	8.49 × E–04	7.28 × E–06
ENet and LWA-MoDUNet	3.04 × E–02	9.51 × E–03
SegNet and LWA-ModUNet	4.85 × E–02	6.39 × E–04

LWA-MoDUNet, lightweight modified UNet model.

Table 8: Comparative analysis of computational complexity

Model	Measures	Inference time (ms/image)
FCN-8s [28]	134.27	86.1
SegNet [29]	29.46	60.7
DeconvNet [31]	251.84	214.3
U-Net [30]	31.03	47.2
ResUNet [51]	8.10	55.8
LWA-ModUNet	**5.17**	**36.4**

models. On the contrary, DeconvNet exhibits the highest number of parameters, along with the training and inference times. The inference time for ResUNet is longer than that for LWA-ModUNet, indicating the drawbacks of deep residual learning. Furthermore, U-Net has nearly six times more parameters than LWA-ModUNet despite being based on the UNet architecture. When compared with FCN-8s and SegNet, the inference time of LWA-ModUNet is comparatively less than both. The simpler structure of LWA-ModUNet has contributed to reduced execution time. With improved accuracy and minimal computational resources, LWA-ModUNet demonstrates superior performance. Hence, LWA-ModUNet is a computationally lightweight model, making it desirable for real-time applications.

VI. Conclusion

In this paper, a lightweight attention-based modified UNet (LWA-MoDUNet) has been proposed for semantic segmentation. The semantic segmentation process strongly relies on the attention model, integrating it with the inception module and deconvolution. The attention module is a component through which we were able to dynamically highlight crucial regions in complex environments such as unstructured driving scenarios, medical images, various human poses, and pets, hence allowing the model to attend the useful features and suppress redundant or noisy information. The purpose of this selective attention is to allow the model to have more fine-grained details and complex patterns essential for the precise segmentation of the scene. The object or region to be segmented in the image is delineated optimally (with improvements) by the proposed model. To sum up, the inception-based encoder integrated with UNet in conjunction with attention mechanism at the decoder side is a lightweight and effective architecture for semantic segmentation in sophisticated diverse images. The experimental results presented here establish that the model can attain the state-of-the-art performance while making a substantial leap toward generalizing the object of interest in cluttered scenes in a more robust manner. In the future, it can be proposed to design a network with a larger and denser receptive field in LWA-MoDUNet. The model is implemented for semantic segmentation, though it can be further improved and modified for multi-class segmentation. In addition, the inclusion of some effective postprocessing techniques to analyze the performance of the model.

Declarations

Funding information

None.

Author contribution

All authors contributed to conception and design of this study. Material preparation, data collection, and analysis were performed by Twinkle Tiwari. The first draft of the manuscript was written by Twinkle Tiwari, and it was reviewed by Mukesh Saraswat.

Conflict of interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability

On user request.

Research involving human and/or animals

None.

Informed consent

None.

References

1. S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321.

2. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Applied Soft Computing 70 (2018) 41–65.

3. X. Yuan, J. Shi, L. Gu, A review of deep learning methods for semantic segmentation of remote sensing imagery, Expert Systems with Applications 169 (2021) 114417.

4. G. Lampropoulos, E. Keramopoulos, K. Diamantaras, Enhancing the functionality of augmented reality using deep learning, semantic web and knowledge graphs: A review, Visual Informatics 4 (1) (2020) 32–42.

5. H. T. Nguyen, N. N. Truong, L. T. T. Pham, N. H. Pham, An approach using skeleton-based representations and neural networks for yoga pose recognition, Applied Computer Systems 30 (1) (2025) 75–84.

6. K. Thyagharajan, G. Kalaiarasi, A review on near-duplicate detection of images using computer vision tech- niques, Archives of Computational Methods in Engineering 28 (2021) 897–916.

7. R. Dhir, M. Ashok, S. Gite, et al., An overview of advances in image colorization using computer vision and deep learning techniques, Rev. Comput. Eng. Res 7 (2) (2020) 86–95.

8. N. T. K. Son, N. H. Quynh, B. T. Minh, Refining graduation classification accuracy with synergistic deep learning models, Cybernetics and Information Technologies 25 (2) (2025).

9. M. Grupac, G. Lăzăroiu, Image processing computational algorithms, sensory data mining techniques, and predictive customer analytics in the metaverse economy, Review of Contemporary Philosophy 21 (2022) 205– 222.

10. Y. Zhu, C. Yao, X. Bai, Scene text detection and recognition: Recent advances and future trends, Frontiers of Computer Science 10 (2016) 19–36.

11. H. Greenspan, B. Van Ginneken, R. M. Summers, Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique, IEEE transactions on medical imaging 35 (5) (2016) 1153– 1159.

12. V. Naosekpam, N. Sahu, Text detection, recognition, and script identification in natural scene images: A review, International Journal of Multimedia Information Retrieval 11 (3) (2022) 291–314.

13. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Confer- ence, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

14. A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147 (2016).

15. V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (12) (2017) 2481– 2495.

16. C. Peng, T. Tian, C. Chen, X. Guo, J. Ma, Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation, Neural Networks 137 (2021) 188–199.

17. A. Abedalla, M. Abdullah, M. Al-Ayyoub, E. Benkhelifa, The 2st-unet for pneumothorax segmentation in chest x-rays using resnet34 as a backbone for u-net, arXiv preprint arXiv:2009.02805 (2020).

18. Autorikshaw detection challenge, https://cvit.iiit.ac.in/autorickshaw_detection/, (Accessed on 01/10/2024).

19. Autonue challenge 2019, https://cvit.iiit.ac.in/autonue2019/challenge/overview.php, (Accessed on 12/21/2023).

20. Ct liver, https://www.kaggle.com/datasets/zxcv2022/digital-medical-images-for-download-resource/, (Accessed on 12/21/2023).

21. W. Zhang, Z. Liu, L. Zhou, H. Leung, A. B. Chan, Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation, Image and Vision Computing 61 (2017) 22–39.

22. Visual geometry group - university of oxford, https://www.robots.ox.ac.uk/~vgg/data/pets/, (Accessed on 12/21/2023).

23. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2015).

24. W. Liu, A. Rabinovich, A. C. Berg, Parsenet: Looking wider to see better, arXiv preprint arXiv:1506.04579 (2015).

25. G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3194–3203.

26. H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151– 7160. doi:10.1109/CVPR.2018.00747.

27. Y. Zhuang, F. Yang, L. Tao, C. Ma, Z. Zhang, Y. Li, H. Jia, X. Xie, W. Gao, Dense relation network: Learning consistent and context-aware representation for semantic image segmentation, in: 2018 25th IEEE international conference on image processing (ICIP), IEEE, 2018, pp. 3698–3702.

28. H. Zhang, H. Zhang, C. Wang, J. Xie, Co-occurrent features in semantic segmentation, in:

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 548–557.

29. H. Ding, X. Jiang, B. Shuai, A. Q. Liu, G. Wang, Semantic correlation promoted shape-variant context for segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8885–8894.

30. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

31. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

32. O¨. C¸ i¸cek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

33. X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE transactions on medical imaging 37 (12) (2018) 2663–2674.

34. S. Shah, P. Ghosh, L. S. Davis, T. Goldstein, Stacked u-nets: a no-frills approach to natural image segmen- tation, arXiv preprint arXiv:1804.10343 (2018).

35. T. M. Quan, D. G. C. Hildebrand, W.-K. Jeong, Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics, Frontiers in Computer Science 3 (2021) 613981.

36. B. Hariharan, P. Arbel´aez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.

37. G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1925–1934.

38. V. Nekrasov, C. Shen, I. Reid, Light-weight refinenet for real-time semantic segmentation, arXiv preprint arXiv:1810.03272 (2018).

39. T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4151–4160.

40. H. Sak, A. Senior, F. Beaufays, Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, arXiv preprint arXiv:1402.1128 (2014).

41. R. Messina, J. Louradour, Segmentation-free handwritten chinese text recognition with LSTM-RNN, in: 2015 13th International conference on document analysis and recognition (icdar), IEEE, 2015, pp. 171–175.

42. P. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, in: International conference on machine learning, PMLR, 2014, pp. 82–90.

43. R. P. Poudel, P. Lamata, G. Montana, Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: Reconstruction, Segmentation, and Analysis of Medical Images: First International Workshops, RAMBO 2016 and HVSMR 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 1, Springer, 2017, pp. 83–94.

44. W. Byeon, T. M. Breuel, F. Raue, M. Liwicki, Scene labeling with LSTM recurrent neural networks, in: Pro- ceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3547–3555.

45. F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 41–48.

46. B. Shuai, Z. Zuo, B. Wang, G. Wang, Dag-recurrent neural networks for scene labeling, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3620–3629.

47. H. Fan, H. Ling, Dense recurrent neural networks for scene labeling, arXiv preprint arXiv:1801.06831 (2018).

48. Q. Zhao, J. Liu, Y. Li, H. Zhang, Semantic segmentation with attention mechanism for remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–13.

49. H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.

50. V. Badrinarayanan, A. Kendall, R. C. SegNet, A deep convolutional encoder-decoder architecture for image segmentation, arXiv preprint arXiv:1511.00561 5 (2015).

51. D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, C. Wolf, Residual conv-deconv grid network for semantic segmentation, arXiv preprint arXiv:1707.07958 (2017).

52. A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder- decoder architectures for scene understanding, arXiv preprint arXiv:1511.02680 (2015).

53. J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, H. Lu, Stacked deconvolutional network for semantic segmentation, IEEE Transactions on Image Processing (2019).

54. S. M. Sam, K. Kamardin, N. N. A. Sjarif, N. Mohamed, et al., Offline signature verification using deep learning convolutional neural network (CNN) architectures googlenet inception-v1 and inception-v3, Procedia Computer Science 161 (2019) 475–483.

55. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018).

56. J. Lee, J. Choi, J. Mok, S. Yoon, Reducing information bottleneck for weakly supervised semantic segmenta- tion, Advances in Neural Information Processing Systems 34 (2021) 27408–27421.

57. Y. Kim, Y. Lee, M. Jeon, Imbalanced image classification with complement cross entropy, Pattern Recognition Letters 151 (2021) 33–40.

58. Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, Advances in neural information processing systems 31 (2018).

59. C. K. Dewa, et al., Suitable CNN weight initialization and activation function for Javanese vowels classification, Procedia computer science 144 (2018) 124–132.

60. U. M. Khaire, R. Dhanalakshmi, High-dimensional microarray dataset classification using an improved adam optimizer (iadam), Journal of Ambient Intelligence and Humanized Computing 11 (11) (2020) 5187–5204.

61. R. Meyes, M. Lu, C. W. de Puiseau, T. Meisen, Ablation studies in artificial neural networks, arXiv preprint arXiv:1901.08644 (2019).

62. R. F. Woolson, Wilcoxon signed-rank test, Wiley encyclopedia of clinical trials (2007) 1–3.