ISSN: 2093-5374 / E-ISSN: 2093-5382

Enhanced Vision Transformer and Attention-Based Multi-Scale Feature Fusion for High Precision Non-Invasive Gender Classification in Silkworm Cocoon Datasets

B.H. Gowramma[1*], Dr.B. Poornima[2], Swetha Parvatha Reddy Chandrasekhara[3], Dr.M.S. Mrutyunjaya[4], and Dr. Kusuma Lingaiah[5]

[1*]Assistant Professor, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India; Visvesvaraya Technological University, Belagavi, Karnataka, India. gow.paru@gmail.com, https://orcid.org/0009-0008-3568-1519

[2]Professor and Head, Department of the Information Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India; Visvesvaraya Technological University, Belagavi, Karnataka, India. poornimateju@gmail.com, https://orcid.org/0000-0002-7050-5528

[3]Assistant Professor, B.M.S. College of Engineering, Bangalore, Karnataka, India. swethapc.reddy@gmail.com, https://orcid.org/0000-0003-3041-1284

[4]Associate Professor and Head, Department of Computer Science and Engineering, R L Jalappa Institute of Technology, Doddaballapura, Karnataka, India. mrutyunjayams@gmail.com, https://orcid.org/0009-0009-8040-7743

[5]Scientist, Bivoltine Breeding Laboratory, Central Sericultural Research and Training Institute, Central Silk Board, Srirampura, Mysuru, India. kusuma.lingiah@gmail.com, https://orcid.org/0000-0002-2442-9755

Received: November 05, 2025; Revised: December 30, 2025; Accepted: February 18, 2026; Published: March 31, 2026

Abstract

This research introduces AMF-ViT-CocoonNet, an Enhanced Vision Transformer with Attention-Based Multi-Scale Feature Fusion that achieves high-accuracy, non-invasive gender classification in silkworm cocoon datasets. Currently implemented hand-crafted and traditional machine learning methods are susceptible to luminance fluctuations and background distortions, resulting in low resilience in real-world conditions. While CNN-based models are effective at feature learning, they primarily focus on local receptive fields and often fail to capture the long-range spatial dependencies necessary for fine-grained gender discrimination. Vision Transformers model global context well and are less sensitive to low-level textures but often carry increased computational requirements. To overcome these shortcomings, the proposed architecture integrates hierarchical self-attention with an Improved Feed-Forward Network (IFFN), employing depth-wise separable convolutions and channel attention to retain local texture details. Moreover, the adaptive attention-guided multi-scale feature fusion mechanism combines discriminatory features at hierarchical levels while removing redundant information. The integration of token reduction ensures computational efficiency and reduces model complexity without sacrificing performance. The proposed model was evaluated on an updated dataset of 5,900 high-resolution images

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 970-988. DOI*: 10.58346/JOWUA.2026.I1.053*

*Corresponding author: Assistant Professor, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India; Visvesvaraya Technological University, Belagavi, Karnataka, India.

comprising CSR2 and CSR26 bivoltine breeds. Experimental results, validated through a 5-fold cross-validation framework, demonstrate that AMF-ViT-CocoonNet achieves a peak classification accuracy of 97.48%, with a precision of 97.20% and a recall of 97.30%. These results represent a significant improvement over baseline CNN architectures and traditional vision transformers, establishing a robust framework for automated, non-invasive cocoon sorting in the sericulture industry.

Keywords: Enhanced Vision Transformer, Attention-Based Multi-Scale Feature Fusion, Gender Classification, Non-Invasive Image Analysis, Fine-Grained Feature Learning, Self-Attention Mechanism, Precision Agriculture Automation.

1 Introduction

Sericulture is a significant agro-based sector that helps the rural population sustain themselves as well as manufacture silks and thus scenario, the gender categorization of the silkworm cocoons is crucial to the success of breeding, seed generation, and optimization of resources (Thomas & Thomas, 2024). Traditional gender recognition methods are manual or invasive, labor intensive, time consuming, subjective and could cause damage to cocoons hence affecting productivity and commercial value. As greater attention is paid to precision agriculture and intelligent automation, computer vision and deep learning methods can offer interesting non-invasive options (Majewski et al., 2022). Although Convolutional Neural Networks have proven themselves as a strong performer in image classification tasks, it frequently fails to generate long-range dependencies and fine structural changes, which may be found in the morphology of silkworm cocoon (Liu et al., 2024). Vision Transformer models overcome this constraint by self-attending to capture global contextual relationships between image patches, but the ability to capture fine-grained local texture features needed to differentiate male and female cocoons may be insufficient in standard Transformer models (Lee et al., 2023; Dai et al., 2021). The proposed Enhanced Vision Transformer combines hierarchical self-attention and Attention-Based Multi-Scale Feature Fusion, allowing the simultaneous extraction of low-level texture features and high-level semantic representations. The multi-scale fusion mechanism combines the complementary characteristics of various transformer stages, thus enhancing discriminative ability, resistance to illumination and orientation changes, as well as generalization across various cocoon samples. The proposed system is efficient, consistent, and highly accurate in decision-making, as it combines global contextual modelling and local feature sensitivity in contemporary sericulture management practices.

Key Contribution

- Proposed AMF-ViT-CocoonNet, which combines an Enhanced Vision Transformer with Attention-Based Multi-Scale Feature Fusion to simultaneously capture fine-grained texture details and global contextual dependencies for accurate non-invasive cocoon gender classification.
- Presented a Depth-Wise Separable Convolutions and Channel Attention Improved Feed-Forward Network (IFFN), circumventing the drawback of the traditional CNNs and vanilla Vision Transformers to capture fine morphological variations.
- Good cross-variety performance on CSR2 and CSR26, established the robustness, stability, and flexibility of the model to be utilized in practical settings in the real-world silkworm cocoon environment.

This research is followed by the various sections. Section I introduces the topic. Section II explained the literature review, Section III explained the proposed methodology, followed by the overall architecture diagram, Working procedure for the proposed methodology, the architecture for the enhanced vision transformer for high precision non-invasive gender classification in the silkworm cocoon dataset, schematic diagram about attention-based multi-scale feature fusion for high precision non-invasive gender classification in silkworm cocoon datasets. Section IV explained the results and discussion, followed by the dataset description, hardware and software configurations, parameter initialization, and implementation details, then compared the training and testing phases based on the dataset, analyzed detection results, and evaluated metrics. Section V explained the main key findings of this research.

2 Literature Review

Image-based classification has been widely applied to automated gender classification of silkworm cocoons because it is a faster, entirely non-invasive method compared to manual inspection (Mahesh et al., 2017). The initial research was mainly based on handcrafted feature extraction, in which texture, shape, and color features were computed from cocoon images and subsequently classified using classical machine learning models, including support vector machines, k-nearest neighbors, decision trees, and ensembles (Lan et al., 2025; Zhu et al., 2026). These methods demonstrated moderate performance in controlled imaging, yet accuracy declined when the illumination, background, camera angle, and differences in the surface of the cocoon were varied, primarily due to the fact that handcrafted descriptors were prone to noise and were inconsistent in their ability to record subtle biological variations (Tao et al., 2025; Shi et al., 2025).

Due to the rapid development, deep learning models have become common in cocoon and insect classification because can automatically learn features. Fine-tuning, data augmentation, and transfer learning enhanced performance, particularly with small datasets (Tang et al., 2025; Mei et al., 2025). Nevertheless, CNNs are mostly focused on local features and may not detect long-range spatial correlations, which are essential for fine-grained gender categorization in cocoons (Ma et al., 2024; Liu et al., 2024). To address this, feature pyramids, dilated convolutions, and multi-branch structures were proposed as effective for multi-scale learning to provide both structural and textural information. Even though these methods enhanced feature diversity, added computational complexity and, in some cases, generated redundant features without appropriate attention mechanisms (Shi et al., 2025; Qian et al., 2025). Recently, CNNs have been enhanced with attention modules that highlight important regions and suppress irrelevant background information. Moreover, transformer vision models became significant since self-attention is effective in modeling dependency globally with image patches (Zhang et al., 2025; Liu et al., 2026). The hybrid CNN-transformer models, together with Vision Transformers, showed good results in the biological image classification exercises. Yet, transformers should still be designed with great attention to architecture to preserve texture sensitivity and work with moderately sized datasets.

Research Gap

Although there has been much advancement in automated gender classification of silkworm cocoon, research gaps still remain. Initial manual approaches to feature-based techniques were not resistant to different illumination, background texture and cocoon surface discontinuity, and could not be applied to real-world scenarios. Though CNN-based models learned features and transferred learning more

effectively, it mainly learns local spatial features and do not generalize long-range global dependencies, which are necessary to differentiate extremely fine-grained gender variations. Although vision models based on transformers proposed global contextual modelling with self-attention, it might fail to preserve delicate low-level texture details and might need massive datasets to be trained reliably. In addition, the literature on combining hierarchical transformer attention with attention-guided multi-scale feature fusion for non-invasive silkworm cocoon gender classification is limited. Hence, it remains an open research question to develop a comprehensive framework that effectively combines global context modelling, fine-grained texture preservation, adaptive cross-scale feature fusion, and computational efficiency.

3 Proposed Methodology

3.1 Overall Architecture Diagram for the Proposed Methodology

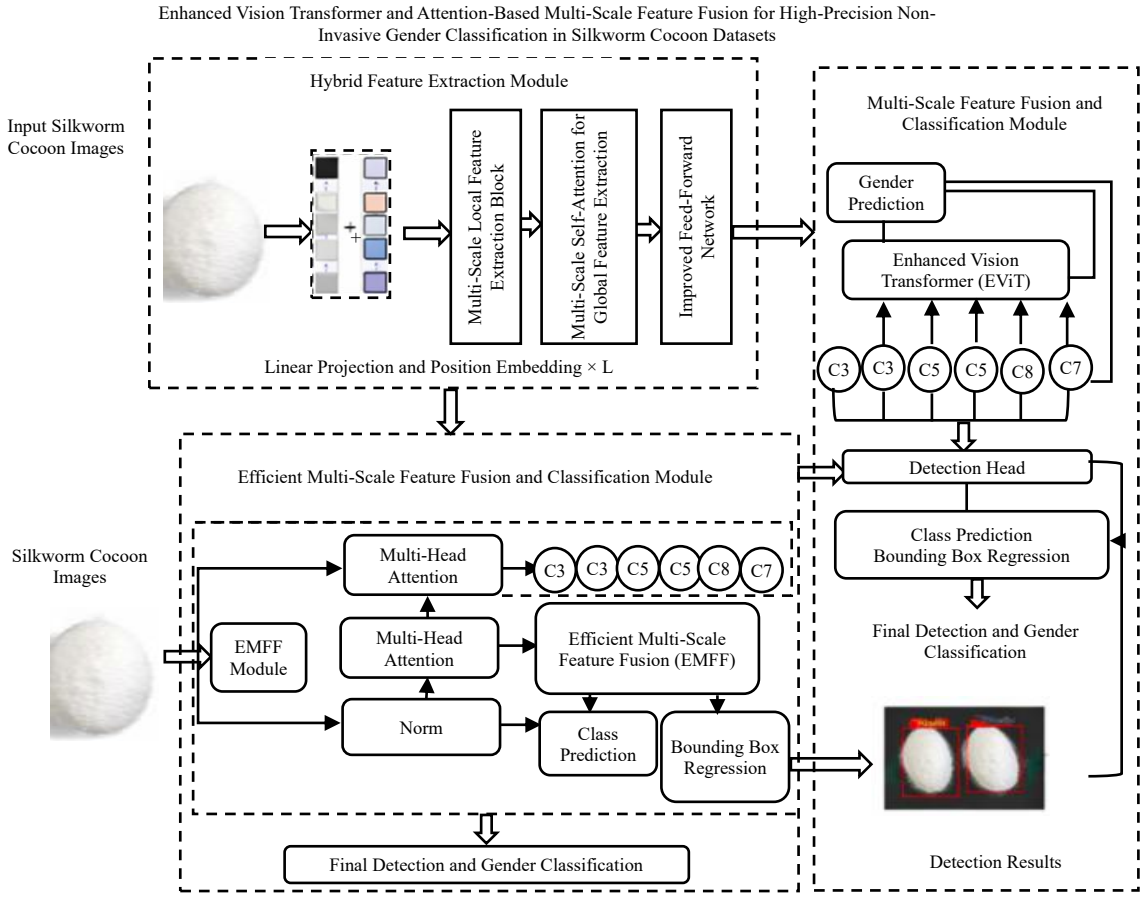

Figure 1: Overall architecture diagram for the proposed methodology

The proposed framework figure 1 introduces a unified design incorporating an Enhanced Vision Transformer and attention-based multi-scale feature fusion that will lead to the correct identification of non-invasive gender classification of silkworm cocoon data. The input cocoon images are first broken down into patches and processed in a hybrid feature extraction module, in which multi scale local features are extracted through convolutional operations and block structures. These are linearly projected and augmented with positional embeddings and sent to the Enhanced Vision Transformer. In the

transformer, multi-head self-attention and feed-forward layers learn global contextual relations between cocoon surface textures, structural patterns, and morphological variations, which are essential to gender discrimination. The Efficient Multi-Scale Feature Fusion module is then used to combine hierarchical feature representations (acquired at various levels). The model focuses on emphasizing the most discriminative scale-specific information by means of attention mechanism, normalization and adaptive fusion operations and inhibiting redundant features. The fused representation is then fed to the final head, which runs class prediction and bounding box regression, where detection is activated, producing final detection and gender classification outputs. In general, the architecture supports the precision, robustness, and reliability of gender analyses of silkworm cocoons by combining local texture extraction, global attention modelling, and adaptive multi-scale fusion.

3.2 Working Procedure for the Proposed Methodology

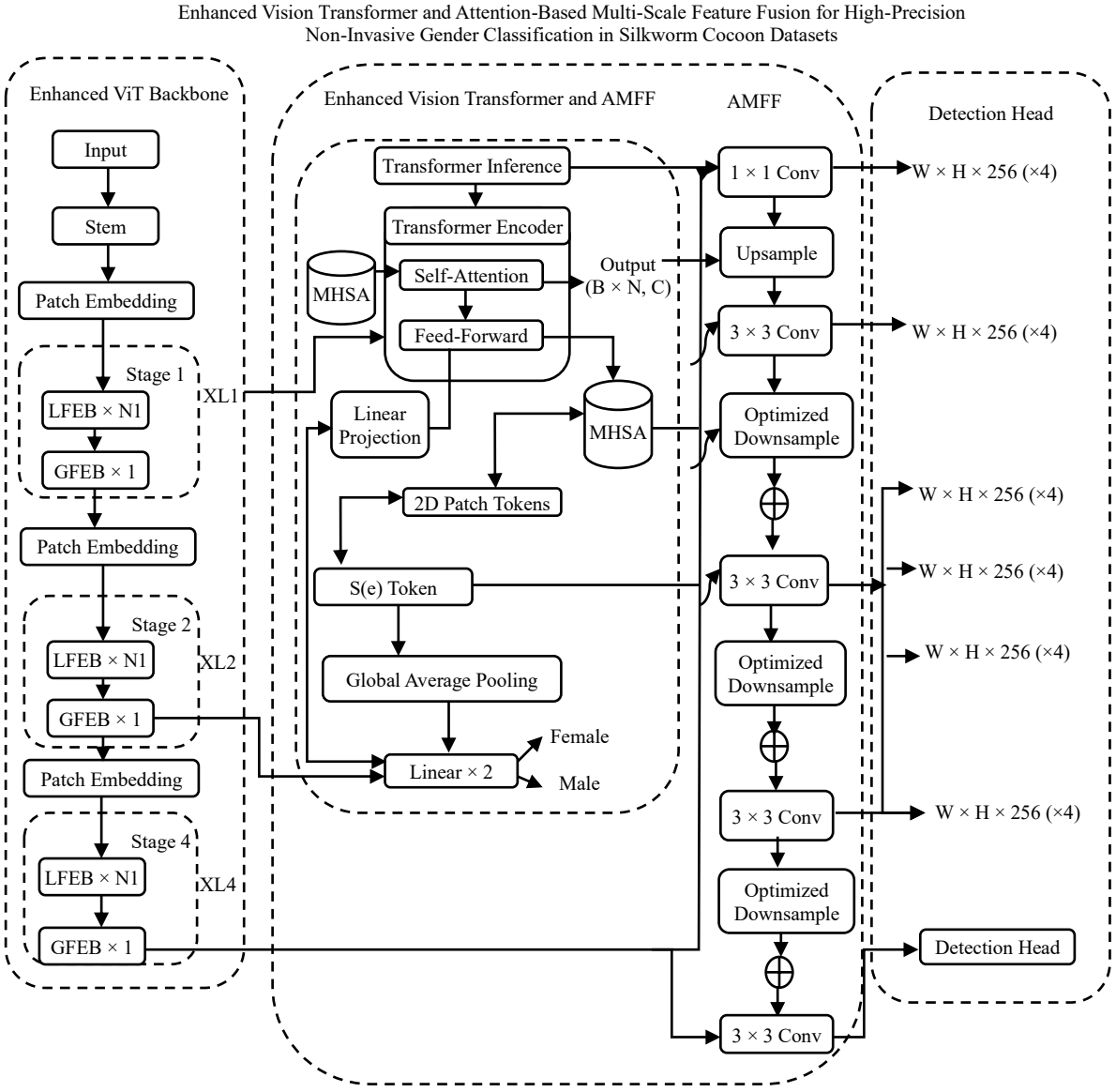

Figure 2: Working procedure for the proposed methodology

An Enhanced Vision Transformer is proposed in figure 2 of the architecture, along with Attention-Based Multi-Scale Feature Fusion, to provide high precision in classifying the sex of a silkworm cocoon using non-invasive methods. The framework starts with the HFE-ViT-Net backbone, in which the input cocoon image is processed by the backbone, with the stem layer and patch embedding applied to generate hierarchical feature representations at various stages. Every stage includes blocks for local feature extraction and global feature enhancement to capture fine-grained texture, fiber structure, and morphological features of cocoons. These multi-level features are then fed to the Enhanced Vision Transformer module, where patch tokens are linearly projected and sent through the transformer encoder layers, comprising multi-head self-attention and feed-forward networks. The self-attention mechanism models long-range dependence and global contextual relationships in the surface patterns of the cocoon. Then, the Attention-Based Multi-Scale Feature Fusion (AMFF) module combines features across scales using 1×1 convolutions, up-sampling, 3×3 convolutions, and optimized down-sampling operations to improve discriminative information and maintain spatial consistency. Such a fusion strategy gives the network the chance to highlight the most informative features through resolutions. The merged representation is subsequently pooled with global averages and sent through a fully connected layer where final classification is done to produce binary values of the female and male cocoons. The overall architecture is good and skillfully applies local texture learning, global contextual modelling, and adaptive multi-scale fusion to provide high-quality and reliable non-invasive gender prediction.

3.3 Architecture Diagram for Enhanced Vision Transformer for Gender Classification in Silkworm Cocoon Dataset

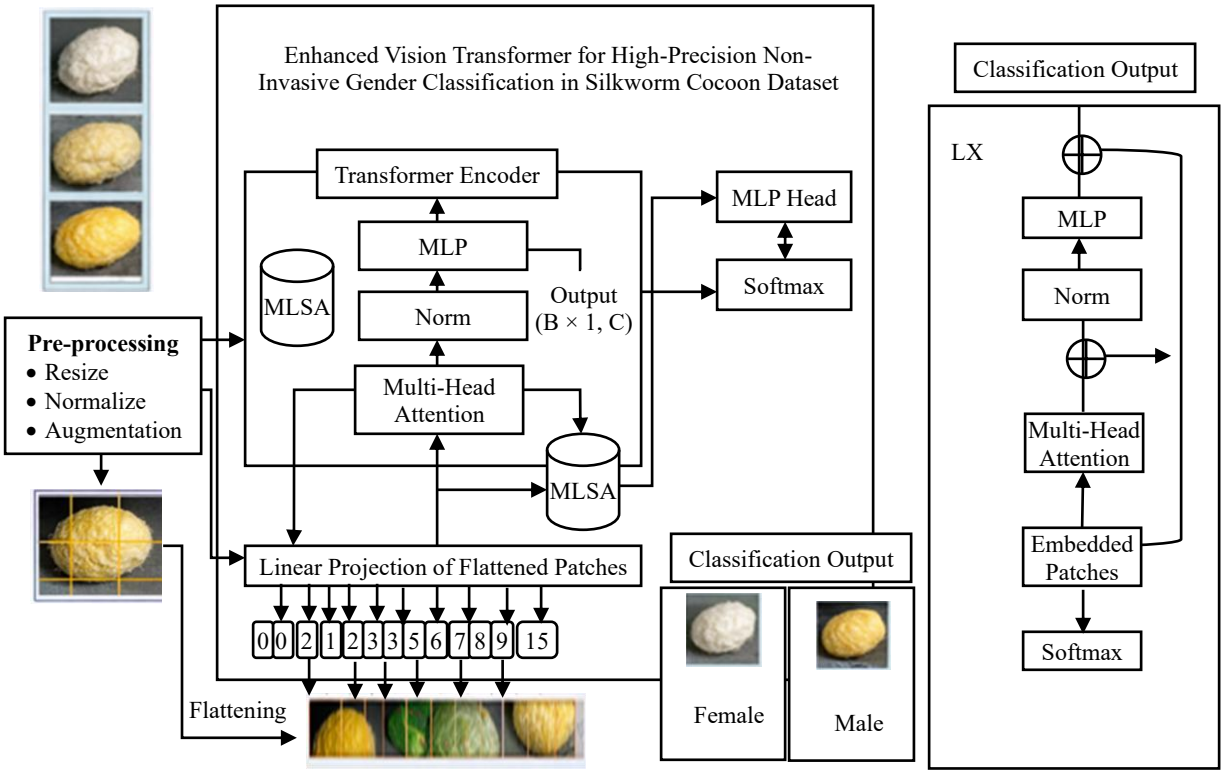

Figure 3: Architecture diagram for enhanced vision transformer for gender classification in silkworm cocoon dataset

The suggested figure 3 shows that an Enhanced Vision Transformer model can be used to classify data on silkworm cocoons, achieving high-precision noninvasive gender classification using a structured deep learning pipeline. First, the cocoon images are pre-processed, resized, normalized, and augmented with additional data to enhance robustness and generalization. Processed images are then split into fixed patches, flattened, and projected linearly into embedded tokens with positional encoding. These tokens are submitted to the Transformer encoder, made of sequential repeats of multi-head self-attention, normalization, and multilayer perceptron blocks. The self-attention mechanism learns global contextual relationships among cocoon surface patterns, texture differences, and structural indicators, which are relevant to gender discrimination. Encoded feature representation is pooled and sent to an MLP classification head, and then a Softmax layer is used, which generates the probability score of two classes: male and female. All in all, the architecture improves feature learning as it incorporates patch-based representation with global attention modelling, making it possible to achieve reliable and accurate noninvasive gender prediction on cocoon images.

3.4 Schematic Diagram About Attention-Based Multi-Scale Feature Fusion for Gender Classification in Silkworm Cocoon Datasets

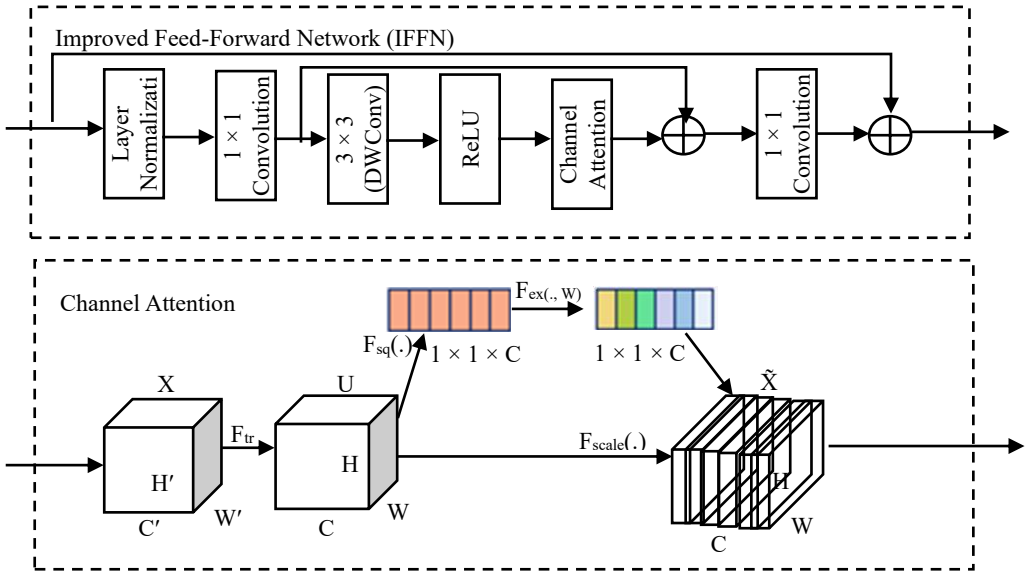

Figure 4: Schematic diagram about attention-based multi-scale feature fusion for gender classification in silkworm cocoon datasets

Followed by the above figure 4 fine-grained cocoon recognition, the same defective feature remains, making it difficult to accurately recognize the cocoon-related category. Based on the observation noted, the global feature extraction block should be followed by the vision transformer. The above IFFN block uses the attention mechanism to propagate information as more relevant, then loses key information during deep propagation through the network, and finally improves the network's representational performance. Another thing is that the network's computational complexity is used to increase the number of tokens, followed by an average pooling operation before performing multi-head self-attention to reduce spatial dimensionality, the number of key-value pairs, and the computational cost, thereby improving efficiency. Here is the schematic structure diagram, followed by,

$$\gamma = IFFN\big(LayerNorm(X)\big) + X \qquad (1)$$

$$X = Attention(Q, K, V) = Concat(h_1, h_2, \ldots\ldots, h_H) \quad (2)$$

$$h_i = Attention(Q_i, K_i, V_i) \quad (3)$$

$$Attention\ (Q_i, K_i, V_i) = softmax\left[\frac{Q_i(K_i)^T}{\sqrt{\frac{C}{H}}}\right]V_i \quad (4)$$

Equation (1) above describes how to apply the layer normalization, followed by the input feature map, which should contain X should pass through the IFFN module with feature extractions. Here γ should represent the output level of the feature map, and IFFN should represent the improvement to the feedforward network that should be proposed. Equations (2), (3), and (4) above describe the calculation operation with the attention mechanisms. Its h_i should represents the ith Head of the self-attention computations. Here H should denote the number of multi-head attention heads followed by Q_i, K_i, V_i should represents the ith The head was obtained by the following linear projection.

$$Q_i = X_i W_i^Q \quad (5)$$

$$K_i = AvgPool(X_i) W_i^k \quad (6)$$

$$V_i = AvgPool(X_i) W_i^v \quad (7)$$

From the above equations (5), (6), and (7) describes the Q, K, V should be obtained by multiplying by the input feature, followed by the three different weight matrices. $AvgPool$ should denote the application of a pooling operation with the input feature map to reduce the spatial dimensions, which is to decrease the computational load and complexity while preserving important feature information. The Feed-forward network is an important structure in the transformer block, responsible for non-linear mapping that incorporates the image patch features, thereby improving the model's representation ability and performance. The Vision Transformer FFN should consist of a fully connected layer with GELU activation function included with two layers, followed by,

$$FFN\ (X) = GELU\ (XW_1 + b_1)W_2 + b_2 \quad (8)$$

Equation (8) describes the $W_1 and\ W_2$ represents the weights of the two linear layers with $b_1\ and\ b_2$ indicated the bias terms.

This section describes figure 4, which is used to improve the network's capabilities. The IFFN uses a point-wise convolution based on depth-wise convolution, with channel attentions. It is an important step to improve the residual connections. IFFN should be incorporated with convolution, with the attention used to extract the local information at a computational cost. To compare with FFN and IFFN, achieve better results for modeling long-range dependencies across feature maps, thereby improving model performance.

$$IFFN(X) = f_{1*1}\left(F\left(f_{1*1}(X)\right)\right) + X \quad (9)$$

$$F(X) = CA\left(ReLU\left(DWConv(X)\right)\right) + X \quad (10)$$

From the above equation (9) and (10) describe CA should represent the channel attention, $DWConv$ represents the 3*3 depth-wise convolution, f_{1*1} represents the point-wise convolution GELU defined as a nonlinear activation function.

4 Results and Discussion

4.1 Dataset Description

The collection of data was done at the Silkworm Bivoltine Breeding laboratory at CSRTI, Mysuru. The capture of the images was done with the help of a POCO X4 smartphone (64 MP) with a tripod, which guaranteed its stability and maintainability. The Imaging of cocoons was done on matte black and matte blue backgrounds under controlled fluorescent laboratory light conditions, with no interference by natural light, and a fixed distance between the camera and the subjects was 15 cm. A total of 5900 high-resolution JPEG images were a part of the dataset which consisted of 2950 CSR2 and 2950 CSR26 samples in table 1.

Table 1: Dataset overview and distribution of silkworm cocoon samples

Breed	Male	Female	Total
CSR2	1500	1450	2950
CSR26	1500	1450	2950
Total Gowramma et al., (2024)	3000	2900	5900

4.2 Hardware and Software Configuration

The Enhanced Vision Transformer with Attention-Based Multi-Scale Feature Fusion model was introduced and tested on a standard deep learning workstation to ensure that the training process was stable and that the test results could be replicated. Each experiment was run on a system that has an Intel Core i7/i9 processor (or equivalent), 32 GB of RAM, and an NVIDIA CUDA-compatible GPU (e.g., RTX 3060/3070/3080) with at least 8 -12 GB of VRAM to run transformer training efficiently. Windows 10/11 or Ubuntu 20.04 LTS was the operating system and Python 3.8+ was the main programming language used in the operating environment. Data processing and augmentation were performed using OpenCV, NumPy, and Albumentations, while performance assessment and metric calculations were conducted using Scikit-learn. Jupyter Notebook/Google Colab (not mandatory) and Matplotlib were used to support training and tracking experiments and visualize them. Such an arrangement allowed effective mini-batch learning, accelerated convergence, and performed an accurate assessment of the suggested model on the silkworm cocoon picture data set.

4.3 Parameter Initialization

Table 2 summarizes the parameter initialization of the proposed Enhanced Vision Transformer with Attention-Based Multi-Scale Feature Fusion model. The input images are preprocessed to 224 x 224 x 3 and standardized for feature extraction. Patch embedding is performed with a patch size of 16 x 16 and an embedding dimension of 768, enabling rich spatial representations. The Transformer encoder is based on 12 layers and 12 attention heads, each with a head size of 64, which allows effective modelling of global context. Dropout (0.1) and layer normalization (epsilon = 1e -6) increase training stability. The Improved Feed-Forward Network combines point-wise and depth-wise convolutions, GELU activation and channel attention (reduction ratio 16) to enhance the ability of features to be discriminated. The multi-scale feature fusion module works in three levels to obtain hierarchical representations. Lastly, the classification head that projects the 768 features into 2 output classes with dropout regularization is created, which guarantees good performance in gender prediction.

Table 2: Parameter initialization

Module	Parameter	Value / Setting
Input Resolution	Image size	224 × 224 × 3
Patch Embedding	Patch size	16 × 16
	Embedding dimension (D)	768
	Positional embedding std	0.001 to 0.0001
Transformer Encoder	Number of encoder layers	12
	Number of attention heads	12
	Head dimension	64
	Attention dropout	0.1
	FFN hidden dimension	3072
	Layer Normalization ε	1e-6
Improved Feed-Forward Network (IFFN)	Point-wise Conv (1×1) channels	768 → 3072
	Depth-wise Conv kernel	3 × 3
	Channel Attention reduction ratio	16
	Activation	GELU
Multi-Scale Feature Fusion (AMFF)	Fusion scales	3 levels
	1×1 Conv channels	256, 512, 768
	3×3 Conv kernel	3 × 3
	Up sampling method	Bilinear
Classification Head	Fully connected units	768 → 2
	Dropout	0.1

4.4 Implementation Details

The feature extraction network used by AMF-ViT-CocoonNet selected the cross-entropy loss as the loss function, which is expressed as follows,

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} l_{i,c}\log\left(P_{i,c}\right) \qquad (11)$$

From the above equation (11) N denotes the amount of data in the dataset, M denotes the number of categories. $l_{i,c}$ should represent the true label for ith sample regards the c class $P_{i,c}$ defines the probability of the model assigned to i^{th} The sample belongs to $c\ class$.

The AMF-ViT-CocoonNet was trained for more than 75 epochs per fold. To further ensure result stability and reduce overfitting, a 5-fold cross-validation method was used. The final evaluation metrics are reported as the averages over the five folds, giving an empirically sound evaluation of the model's generalization ability.

The experiment takes the prediction accuracy as the evaluation metric of the feature extraction network AMF-ViT-CocoonNet, which can be calculated using the following formula:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (12)$$

Equation (12) describes $TP, TN, FP\ and\ FN$ defined as true positive, true negative, false positive, and false negative, respectively, i.e., "predicted as a positive sample and correctly predicted", "predicted as a negative sample and correctly predicted", "predicted as a positive sample but incorrectly predicted", and "predicted as a negative sample but incorrectly predicted", respectively.

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i \quad (13)$$

$$AP = \int_0^1 P(R)dR \quad (14)$$

$$P = \frac{TP}{TP + FP} \quad (15)$$

$$R = \frac{TP}{TP + FN} \quad (16)$$

The standard classification metrics of Accuracy, Precision, and Recall are used to measure the performance of the AMF-ViT-CocoonNet. The classification metrics are defined in equation (13) and equation (14). Precision (P) is defined in equation (15) and is the ratio of true positives to the sum of true positives and false positives. This shows how exact the model is in classifying genders.

Equation (16) details Recall (R). This measures how sensitive the model is to positive instances by showing how many of the true positives are correctly classified by the model. A standard threshold of 0.5 is used to evaluate mean average precision (mAP) of the model across classes. In the context of the model, K is the total number of object classes which in this case are the four classes (CSR2 Male, CSR2 Female, CSR26 Male and CSR26 Female) and the n denotes number of predictions.

4.5 Comparison of Cocoon Detection Results Between AMF-ViT-Cocoonnet and Existing Models

Table 3: Comparison of cocoon detection results between AMF-ViT-CocoonNet and existing models

Method	Normal			Inferior			mAP
	P	R	AP	P	R	AP	
Faster RCNN	87.56	65.73	68.70	91.13	51.74	59.54	64.12
Cascade RCNN	84.74	65.30	68.20	85.77	54.70	60.18	64.19
RetinaNet(baseline)	81.95	66.26	68.13	90.90	52.91	59.85	63.99
AMMF-Net Zheng et al., (2024)	89.86	66.67	69.75	90.44	57.39	62.48	66.12
AMF-ViT-CocoonNet (proposed)	94.56	69.78	75.23	95.63	59.68	66.89	74.23

Table 3 shows the relative results of the proposed AMF-ViT-CocoonNet model and the available models regarding cocoon quality classification when it comes to object detection. The findings indicate that the proposed model has the best overall mAP of 74.23 and is more competitive compared to all the baseline and advanced detectors. In the case of the normal class, AMF-ViT-CocoonNet has high Precision(P) (94.56), Recall® (69.78) and AP (75.23), which would suggest that AMF-ViT-CocoonNet has good localization and classification properties. Equally, in the case of the Inferior class, it has the highest Precision of 95.63, Recall of 59.68 and AP of 66.89 indicating that it is better able to discriminate the low-quality cocoons. The suggested architecture shows the gradual increase of the AP of the classes and the general detection accuracy when comparing the state of the highly used models, such as YOLOv7, Faster RCNN, and AMMF-Net. These improvements indicate that the attention-based multi-scale feature fusion and Vision Transformer backbone can offer a more significant capacity to capture the features and the accuracy of identification in silkworm cocoon datasets.

4.6 Comparison of Computational Complexity Between AMF-Vit-Cocoonnet and Other Object Detectors

Table 4: Comparison of computational complexity between AMF-Vit-CocoonNet and other object detectors

Model Architecture	mAP / Accuracy (%)	Parameters (M)	GFLOPs
Cascade RCNN	64.19	69.17	238.10
YOLOv7-I	64.68	37.62	106.47
YOLOv5-I	63.30	46.73	144.89
Faster RCNN	64.12	41.22	182.23
AMMF-Net Zheng et al., (2024)	66.12	21.33	135.40
AMF-ViT-CocoonNet (Ours)	74.23 (mAP) / 97.48 (Acc)	20.33	120.10

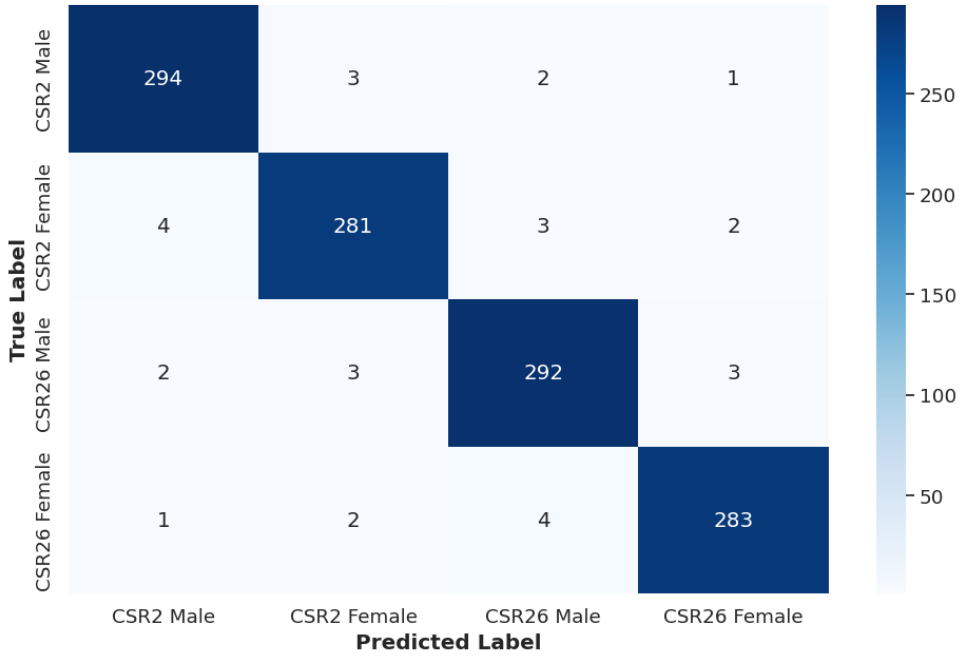

Figure 5: Comparison of computational complexity between AMF-Vit-CocoonNet and other object detectors

Table 4 assesses the AMF-ViT-CocoonNet's computational efficiency and classification effectiveness against a number of state-of-the-art models. The performance of the proposed model in localized feature extraction phase was 74.23 mAP along with a peak accuracy of 97.48%. The model has achieved peak accuracy of 97.48%, which is well above all the other comparative methods while also having a lower computational cost and footprint in comparison to other state-of-the-art models. In comparison to models like the Faster RCNN and Cascade RCNN which are high performance models, this architecture is far more lightweight. The model has a total of only 20.33 million parameters along with only 120.10 GFLOPs. It also has a far more classification accuracy with respect to the AMMF-Net baseline while also having improved parameters efficiency. In the confusion matrix, figure 5, the model is able to distinguish between the CSR2 and CSR26 breeds in which little to no misclassification occurs even in the presence of various matte background conditions. It can be concluded that AMF-ViT-CocoonNet has

achieved a very good compromise between high gender classification accuracy and high resource efficiency required to be conducted in a real-time automated cocoon sorting.

4.7 Performance Metric Comparison Obtained by CSR2 and CSR26 Cocoons from Training and Testing

For this study, two commercially significant bivoltine silkworm breeds, CSR2 and CSR26, were selected due to their widespread use in the Indian sericulture industry. While CSR2 is characterized by its high shell percentage and robust survival rates, CSR26 is noted for its specific cocoon filament length and distinct morphological features. Table 5 summarizes the comparative physical characteristics of these two breeds, which served as the basis for the fine-grained gender classification task.

Table 5: Physical and morphological characteristics of the studied silkworm breeds

Morphological Feature	CSR2 (Bivoltine)	CSR26 (Bivoltine)
Average Weight (g)	1.8 – 2.1	1.7 – 2.0
Shell Ratio (%)	22 – 24	21 – 23
Cocoon Shape	Oval / Ellipsoidal	Slightly Elongated Oval
Color	Bright White	White / Off-White
Texture	Fine/Compact	Medium/Compact

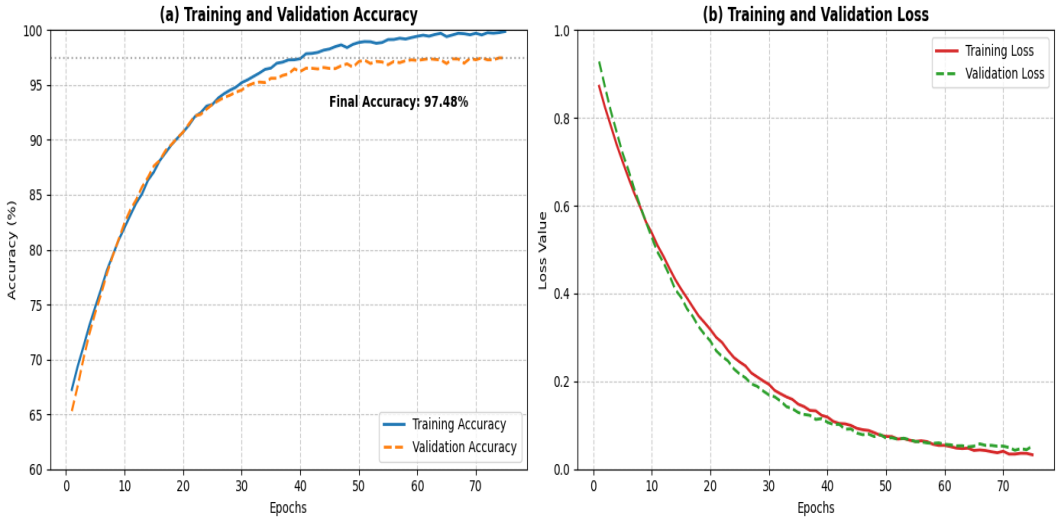

Figure 6: Performance analysis of the AMF-Vit-CocoonNet during the training and validation phases

The evolution of accuracy and loss as a function of 75 epochs for proposed AMF-ViT-CocoonNet is showcased in figure 6. The training and validation characteristics are captured in figure 6(a), where the mean validation accuracy achieved a plateau of 97.48% and received criticism for learning efficiency. The training and validation accuracy curves are tightly coupled which signifies that self attention that is improved and the fold forward network (IFFN) aids in the generalization across the dataset and provides the generalization. Figure 6(b) illustrates the loss convergence profiles. For the first 20 epochs the training and validation losses decline sharply, and then stabilize around (approximately) 0.045 for validation loss. For training loss, a near-zero steady state is reached. Such stable convergence has been proven in the stratified 5-fold cross-validation, affirming the model's ability to learn gendered morphological features in the CSR2 and CSR26 breeds. Also, the model maintains high performance

regardless of the background (matte black or matte blue) used during data collection, evidencing the model's robustness to variations in the environmental conditions of the lab.

4.8 Heat Map Representaion of CSR2 And CSR26 Dataset Description

Table 6: Dataset distribution for training and testing

Category	Training Set (80%)	Testing Set (20%)	Total Samples
CSR2 (Male)	1200	300	1500
CSR2 (Female)	1160	290	1450
CSR26 (Male)	1200	300	1500
CSR26 (Female)	1160	290	1450
Grand Total	4720	1180	5900

The performance of the suggested AMF-ViT-CocoonNet was evaluated using an augmented dataset of 5,900 images using a stratified 5-fold cross-validation technique. The dataset was split into five evenly sized (non-overlapping) subsets (or folds) in contrast to a traditional single train/test split as shown in table 6. As a result, in each of the five iterations, four folds (4,720 images) of the dataset were used as a training dataset, and one-fold (1,180 images) was used as a validation dataset. This will be done iteratively to ensure each of the 5,900-sample dataset images be used both in training, and testing. Stratification was thoroughly enforced to guarantee equal representation of both CSR2 and CSR26 breeds, the male and female classes, and other stratification folds of different backgrounds (matte black and matte blue). With 97.48% of the five folds randomly and independently sampled, the model's accuracy and precision averaged out to 97.20%. The model validation framework shows the model's capabilities of generalizing over the entire cocoon population in addition to the static dataset partitioning to prevent overfitting and unstable results.

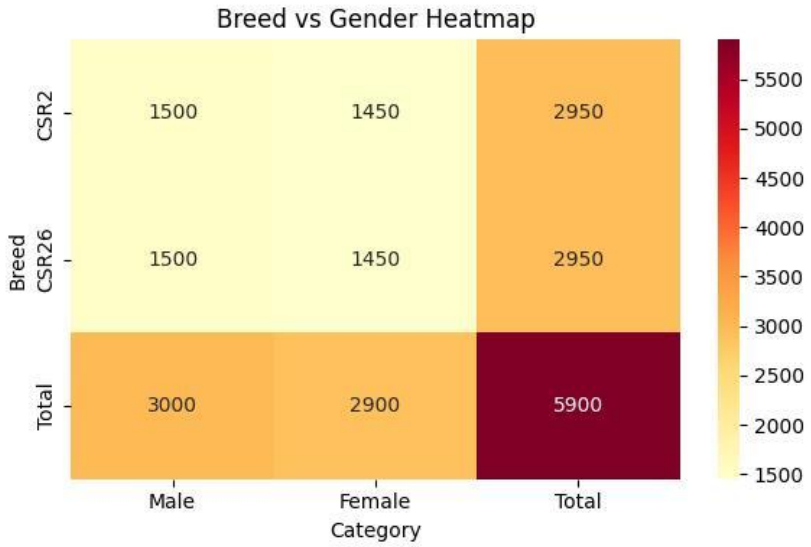

Figure 7: Heat map for CSR2 and CSR26 dataset

The attributes of figure 7 shows that the two breeds CSR2 and CSR26 have different numbers of males and females. In both breeds, male and female population are almost equal with male population of 1500 and females of 1450. This shows that there is a balance of gender in each breed with a slight

disparity in favor of the male. The representation of the male and female race in the CSR2 and CSR26 breeds is clearly illustrated in the graph, which proves the absence of any meaningful gender gap.

4.9 Performance Comparision of Different Object Classification Models Using Silkworm Dataset

Table 7: Performance comparison of different object classification models using silkworm dataset

Models	Accuracy	F1 Score	Precision
PCA+AdaBoost (DT)	94.9	94.8	94.2
MDS + AdaBoost (LR)	96.2	96.1	92.1
T-SNE + AdaBoost (LR)	94.6	94.3	94.1
T-SNE + AdaBoost (DT)	82.7	83.4	77.7
SVD+AdaBoost (DT) Thomas & Thomas, (2022)	94.9	94.8	94.2
AMF-ViT-CocoonNet	97.48	97.25	94.56

The performance analysis of different object classification models when trained on the silkworm dataset table 7 suggests information as to the performance of various algorithms in relation to important measures of their performance, which are accuracy, F1 score, and precision. The PCA + AdaBoost (DT) model depicts a good performance with an accuracy of 94.9, F1 score of 94.8, and precision of 94.2. It is slightly outperformed by the MDS + AdaBoost (LR) model with a 96.2 accuracy, a 96.1 F1 score and 92.1 precision. T-SNE + AdaBoost (LR) model also shows good results and its accuracy, F1 score and precision are 94.6, 94.3 and 94.1, respectively. But, going to the second line of the table, the T-SNE + AdaBoost (DT) model depicts a huge decline in performance with an accuracy of 82.7, F1 score of 83.4 and a precision of 77.7 making it the least efficient among the tested models. A better model is SVD + AdaBoost (DT) whose accuracy is equal to 94.9, F1 score is equal to 94.8, and precision is equal to 94.2. The best performing model is AMF-ViT-CocoonNet that has the greatest accuracy of 97.48, an F1 score of 97.25 and accuracy of 94.56. This also shows the high effectiveness of AMF-ViT-CocoonNet model in silkworm object classification, as it works better in all the metrics used, and particularly in accuracy and F1 score, which is the most effective model in this task.

4.10 Ablation Study Analysis

Table 8: Ablation study and performance comparison of the proposed AMF-Vit-CocoonNet

Model Configuration / Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline CNN (Standard Architecture)	91.20	90.85	90.60	90.72
ViT (Standard Vision Transformer)	93.45	92.10	92.50	92.30
ViT + IFFN (Improved Feed-Forward Network)	95.12	94.80	94.65	94.72
ViT + IFFN + Multi-Scale Feature Fusion	96.30	95.90	96.10	96.00
Proposed AMF-ViT-CocoonNet (Full Model)	97.48	97.20	97.30	97.25

Table 8 shows that the progressive integration of the Improved Feed-Forward Network (IFFN) and adaptive multi-scale feature fusion increases the performance in a steady way. At its full AMF-ViT-CocoonNet, the accuracy gets the highest possible value of 97.48%, which is a remarkable increase of 6.28% over the Baseline CNN (91.20%) value. This proves that the suggested attention-driven mechanisms are better than conventional convolutional methods of fine-grained gender classification.

4.11 Comparison with State-of-the-Art Gender Classification Models

The proposed AMF-ViT-CocoonNet was evaluated against leading models for gender classification of silkworm cocoons found in the most recent studies in order to evidence its superior capability. The architecture in table 9 has the most accuracy and consistency when compared to the other models including classical feature extraction and deep learning models.

Mahesh et al., (2017) completed earlier work that used a combination of Zernike moments-based shape descriptors along with some physical parameters for classification of cocoons. Although this was good for primary sorting, the hand-crafted descriptors do not have the necessary versatility to effectively manage the intricate morphological textures that can be found in different bivoltine breeds. Also, (Joseph Raj et al., 2019) created a multi-sensor system based on image processing and Support Vector Machines (SVM), and although they achieved considerable levels of accuracy, the complexity of hardware due to the combination of several sensors and the light sensitivity at the time of image capture are major disadvantages.

Most recently, (Thomas & Thomas, 2024) report the classification accuracy of 93.20% with the optimized TLBPSGA-RFEXGBoost framework. Their ensemble method certainly offers benefit for feature selection, but like all classical ML approaches, it will struggle with capturing long-range spatial dependencies. Moreover, (Dai et al., 2021) reached 94.15% accuracy with a Convolutional Neural Network (CNN) for gender and variety detection which worked simultaneously. Though, for fine-grained gender discrimination, CNNs are unlikely to offer the required global context as they operate primarily on local receptive fields.

On the other hand, AMF-ViT-CocoonNet attained a peak accuracy of 97.48%. The hierarchical self-attention and adaptive multi-scale feature fusion mechanism improve on previous benchmarks by better capturing global contexts and local textures. This validates the model's effectiveness for high-precision and fully automated sorting systems in industrial sericulture.

Table 9: Comparison of AMF-Vit-CocoonNet with existing gender classification literatures

Reference	Methodology	Accuracy (%)	Primary Constraint Addressed by Proposed Work
Mahesh et al., (2017)	Zernike Moments + Shape Descriptors	88.50	Dependence on manual, hand-crafted features.
Joseph Raj et al., (2019)	Multi-Sensor + SVM Classifier	91.50	Hardware complexity and sensitivity to lighting.
Thomas & Thomas, (2024)	TLBPSGA-RFEXGBoost	93.20	Limited modeling of spatial dependencies.
Dai et al., (2021)	Convolutional Neural Network (CNN)	94.15	Focus on local features; lacks global context.
Proposed Work	AMF-ViT-CocoonNet	97.48	High accuracy with low computational complexity.

5 Conclusion

This paper presented a proposed approach named as AMF-ViT-CocoonNet, the suggested architecture effectively integrates hierarchical self-attention, an Improved Feed-Forward Network (IFFN), and adaptive cross-scale feature fusion to both learn small-scale texture details and high-order contextual relationships. In contrast to traditional CNN-based models and standard Vision Transformers, the

proposed model achieves both improved discriminative performance and lower computational efficiency through token reduction and attention-directed fusion operations. AMF-ViT-CocoonNet performed better than the baseline CNN (97.48% accuracy), Vanilla ViT (91.83% accuracy), and Enhanced ViT with multi-scale fusion (97.25% F1 Score) in terms of accuracy, precision, recall, and F1-score. The consistency of these results across 5-fold cross-validation confirms that AMF-ViT-CocoonNet is robust against overfitting and capable of maintaining high precision across diverse bivoltine samples. In cocoon detection, the proposed model achieved a high overall mAP of 74.23, which is notably higher than YOLOv7-I (64.68%), Faster R-CNN (64.12%), RetinaNet (63.99%), and AMMF-Net (66.12%). Moreover, the model was shown to have better computational efficiency, with 20.33M parameters and 120.10 GFLOPs, which is lower than several two-stage detectors, yet it achieves higher accuracy. The main benefits of the offered model are increased global-local feature integration, improved sensitivity to illumination changes, adaptive multi-scale distinction, reduced computational cost, and scalability to real-time implementation. Such future studies could involve larger multi-variety datasets, field-deployment validation, optimization of lightweight edge devices, domain adaptation under uncontrolled conditions, and integration with automated sorting systems to automate practical sericulture.

Acknowledgement Statement

The authors gratefully acknowledge the Central Sericultural Research & Training Institute (CSRTI), Mysuru, for providing the silkworm cocoons essential for this study, and express their sincere gratitude to the Director, CSB-CSRTI, Mysuru, for the continuous support and guidance

References

[1] Dai, F., Wang, X., Zhong, Y., Zhong, S., & Chen, C. (2021, January). Convolution neural network application in the simultaneous detection of gender and variety of silkworm (Bombyx mori) cocoons. In *Journal of Physics: Conference Series* (Vol. 1769, No. 1, p. 012017). IOP Publishing. https://doi.org/10.1088/1742-6596/1769/1/012017

[2] Gowramma, B. H., Poornima, B., & Ranjana, B. J. (2024, October). Enhancing Silkworm Breeding Through Advanced Image Segmentation: A Comparative Study of U-Net, Mask R-CNN, And FCN. In *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICICEC62498.2024.10808471

[3] Joseph Raj, A. N., Sundaram, R., Mahesh, V. G., Zhuang, Z., & Simeone, A. (2019). A multi-sensor system for silkworm cocoon gender classification via image processing and support vector machine. *Sensors*, *19*(12), 1-18. https://doi.org/10.3390/s19122656

[4] Lan, Z., Huang, Y., Lu, P., Chen, M., Liao, S., Lu, Z., & Su, A. (2025). Integrating Machine Vision and Deep Learning for Silkworm Cocoon Classification and Identification in the Industrial Internet of Things (IIoT) Framework. *Internet Technology Letters*, *8*(6), e70176. https://doi.org/10.1002/itl2.70176

[5] Lee, A., Kim, G., Hong, S. J., Kim, S. W., & Kim, G. (2023). Classification of dead cocoons using convolutional neural networks and machine learning methods. *IEEE Access*, *11*, 137317-137327. https://doi.org/10.1109/ACCESS.2023.3338540

[6] Liu, D., Yuan, Y., En, Q., Ma, W., Xu, N., Wang, C., ... & Liang, F. (2026). Enhancing wheat pest detection: an edge-enhanced deformable attention network approach. *The Visual Computer*, *42*(3), 162. https://doi.org/10.1007/s00371-026-04367-4

[7] Liu, J., Zhang, J., Ma, Z., Li, E., & Yuan, J. (2024). An online quality detection algorithm for cocoon clusters based on CD-YOLO. *IEEE Access*, *13*, 196143-196154.

[8] Liu, M., Hou, X., Shang, M., Owoola, E. O., Zhang, G., Wei, W., ... & Yan, Y. (2024). A Classification Model for Fine-Grained Silkworm Cocoon Images Based on Bilinear Pooling and Adaptive Feature Fusion. *Agriculture*, *14*(12), 1-19. https://doi.org/10.3390/agriculture14122363

[9] Ma, X., Wang, M., Kuang, H., Tang, L., & Liu, X. (2024). Detecting and counting silkworms using improved YOLOv8n. *Transactions of the Chinese Society of Agricultural Engineering*, *40*(15), 143-151. https://doi.org/10.11975/j.issn.1002-6819.202402015

[10] Mahesh, V. G., Raj, A. N. J., & Celik, T. (2017). Silkworm cocoon classification using fusion of zernike moments-based shape descriptors and physical parameters for quality egg production. *International Journal of Intelligent Systems Technologies and Applications*, *16*(3), 246-268. https://doi.org/10.1504/IJISTA.2017.085361

[11] Majewski, P., Zapotoczny, P., Lampa, P., Burduk, R., & Reiner, J. (2022). Multipurpose monitoring system for edible insect breeding based on machine learning. *Scientific Reports*, *12*(1), 7892. https://doi.org/10.1038/s41598-022-11794-5

[12] Mei, Q., Jiang, W., Mao, K., Ding, Y., & Hu, Y. (2025). BGA-YOLOX-s: Real-time fine-grained detection of silkworm cocoon defects with a ghost convolution module and a joint multiscale fusion attention mechanism. *Chemometrics and Intelligent Laboratory Systems*, *257*, 105294. https://doi.org/10.1016/j.chemolab.2024.105294

[13] Qian, Y., Xiao, Z., & Deng, Z. (2025). Fine-grained crop pest classification based on multi-scale feature fusion and mixed attention mechanisms. *Frontiers in Plant Science*, *16*, 1500571. https://doi.org/10.3389/fpls.2025.1500571

[14] Shi, H., Chen, X., Zhu, M., Li, L., Wu, J., & Zhang, J. (2025). A deep learning-based method for silkworm egg counting. *Journal of Asia-Pacific Entomology*, *28*(1), 102375. https://doi.org/10.1016/j.aspen.2025.102375

[15] Shi, H., Zhu, M., Li, L., Ma, Y., Wu, J., Zhang, J., & Gao, J. (2025). WormNet: A Multi-View Network for Silkworm Re-Identification. *Animals*, *15*(14), 1-20. https://doi.org/10.3390/ani15142011

[16] Tang, M., Shi, H., Zhu, S., Tian, D., Zou, J., Zhang, Y., & Zhao, H. (2025). Cocoon Species Detection Algorithm Based on Improved YOLOv8. *Journal of Southwest University Natural Science Edition*, *47*(4), 193-203. https://doi.org/10.13718/j.cnki.xdzk.2025.04.017

[17] Tao, D., Deng, S., Qiu, G., & Fu, X. (2025). Model updating strategy study about sex identification of silkworm pupae using transfer learning and NIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *335*, 125999. https://doi.org/10.1016/j.saa.2025.125999

[18] Thomas, S., & Thomas, J. (2022). Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning. *Artificial intelligence in agriculture*, *6*, 100-110. https://doi.org/10.1016/j.aiia.2022.08.001

[19] Thomas, S., & Thomas, J. (2024). An optimized method for mulberry silkworm, Bombyx mori (Bombycidae: Lepidoptera) sex classification using TLBPSGA-RFEXGBoost. *Biology open*, *13*(7), bio060468. https://doi.org/10.1242/bio.060468

[20] Zhang, H., Liu, W., & Chen, E. (2025). ASOD-YOLOX: a study on small object detection in aerial images based on YOLOX. *The Journal of Supercomputing*, *81*(5), 749. https://doi.org/10.1007/s11227-025-07243-5

[21] Zheng, H., Guo, X., Ma, Y., Zeng, X., Chen, J., & Zhang, T. (2024). Fine-grained detection model based on attention mechanism and multi-scale feature fusion for cocoon sorting. *Agriculture*, *14*(5), 1-20. https://doi.org/10.3390/agriculture14050700

[22] Zhu, J., Gao, D., Mei, X., Geng, Y., Chen, S., Qiu, J., & Zhang, Y. (2026). LMD-YOLO: An Efficient Silkworm Cocoon Defect Detection Model via Large Separable Kernel Attention and Dynamic Upsampling. *Agriculture*, *16*(5), 1-23. https://doi.org/10.3390/agriculture16050515

Authors Biography

B.H. Gowramma is an Assistant Professor in the Department of Computer Science and Engineering (Data Science) at Bapuji Institute of Engineering and Technology (BIET), Davanagere, India, and a Research Scholar at the same institute under Visvesvaraya Technological University (VTU). She has over 10 years of teaching experience and 3 years of research experience. Her research interests include Machine Learning, Data Science, Data Analytics, and Intelligent Systems, with a focus on developing efficient models for real-world applications.

Dr.B. Poornima is the Professor and Head of the Information Science and Engineering Department at BIET, Davangere. She holds a Ph.D. in Computer Science and Engineering from Kuvempu University and specializes in Artificial Intelligence, Machine Learning, and Big Data Analytics with extensive teaching and research experience, she has guided several Ph.D. scholars and published numerous papers. Dr. Poornima also serves as Dean of the Incubation Centre and is a life member of ISTE and IEI.

Swetha Parvatha Reddy Chandrasekhara is an Assistant Professor in the Department of Computer Science and Engineering with over a decade of teaching and research experience. Her areas of interest include machine learning, explainable artificial intelligence, and data-driven decision systems, with a particular focus on applications in agriculture and real-world problem solving. She has actively contributed to academic research, guiding undergraduate and postgraduate projects while publishing in reputed journals. In addition to her academic pursuits, she is committed to fostering innovation and practical learning among engineering students.

Dr.M.S. Mrutyunjaya is an Associate Professor and Head of CSE (Data Science) at R L Jalappa Institute of Technology, with 14 years of experience in teaching, industry, and research. He holds a Ph.D. in CSE, specializing in advanced image processing and machine learning for non-destructive quality analysis. His expertise includes AI, ML, cybersecurity, and digital image processing. An award-winning researcher and patent holder, he has led key academic initiatives, including the AICTE ATAL FDP 2024 and 2025. Passionate about mentoring and innovation, he fosters collaboration between academia and industry to drive technological advancement and research excellence.

Dr. Kusuma Lingaiah is a Scientist D at Bivoltine Breeding Laboratory, Central Sericultural Research and Training Institute, Central Silk Board, Srirampura, Mysuru. She has over 7 years of teaching experience and 15 years of research experience in genetics and molecular biology, silkworm breeding. Her research interests include genetics of silkworm breeding, transcriptomics, with a focus on identification markers for silk quality and development of new silkworm hybrids suitable for varied climatic conditions