Enhanced Square Fiducial Marker Recognition under Challenging Visual Environments Using Multi-Scale CNN-Transformer Fusion

Liliek Triyono
Doctoral Program of Information System, Diponegoro University, Semarang, Indonesia | Electrical Engineering Department, Politeknik Negeri Semarang, Semarang, Indonesia
liliek.triyono@polines.ac.id (corresponding author)

Rahmat Gernowo
Doctoral Program of Information System, Diponegoro University, Semarang, Indonesia
rahmatgernowo@lecturer.undip.ac.id

Prayitno
Electrical Engineering Department, Politeknik Negeri Semarang, Semarang, Indonesia
prayitno@polines.ac.id

Eko Harry Pratisto
Diploma Program of Informatics Engineering, Universitas Sebelas Maret, Surakarta, Indonesia
ekoharry@staff.uns.ac.id

Received: 11 June 2025 | Revised: 11 July 2025, 23 July 2025, and 4 August 2025 | Accepted: 11 August 2025

 DOI: https://doi.org/10.48084/etasr.12693

ABSTRACT

Recent methods using deep learning have demonstrated promising outcomes in tackling the issue of object recognition in low-light images. However, existing techniques often face challenges related to distortion and occlusions, and many strategies rely on neural networks with convolutional neural network (CNN) structures, which are limited in their ability to capture long-term dependencies. This frequently leads to inadequate recovery of very dark areas in low-light images. This work introduces a unique Transformer-based method for ArUco marker recognition in low-light environments, termed Extreme ArUco Vision Transformer (XAViT). We present a Transformer–CNN hybrid block that utilizes mixed attention to effectively capture both global and local information. This method integrates the Transformer's capacity to model long-range dependencies with the CNN's proficiency in extracting detailed features, facilitating the reliable detection of ArUco markers even in extreme lighting conditions. Additionally, we employ a Swin-Transformer discriminator to selectively improve various areas of low-light images, alleviating problems of overexposure, underexposure, and noise. Comprehensive experiments show that XAViT achieves 99.16% accuracy, 97.86% recall, 97.95% precision, and 97.89% F1-score on a realistic low-light dataset, outperforming state-of-the-art CNN and Transformer models. Moreover, its utilization in additional vision-based tasks underscores its potential for wider implementation in advanced vision applications.

Keywords-low-light image; indoor navigation; computer vision; markers; assistive technology

I. INTRODUCTION

Low-light images often suffer from issues such as reduced contrast, blurred features, and color distortion, which can hinder tasks such as object detection and semantic segmentation [1-3]. To address these issues, various low-light enhancement techniques, including Retinex-based algorithms, deep learning, and hybrid models, have been developed to improve brightness, contrast, and clarity [4-6]. Recent deep learning advancements, particularly using Convolutional Neural Networks (CNNs), have significantly improved low-light image quality. Approaches are mainly end-to-end or Retinex-based, with methods like EnlightenGAN and DSLR enhancing texture and illumination [7, 8]. However, they often struggle with long-range dependencies, noise amplification, and artifacts in dark areas [9]. Retinex-based methods, such as Retinex-Net and SCI, aim to adjust lighting while preserving natural appearance, but they still face challenges in improving darker areas and rely on manually defined parameters [10-12].

Deep learning models have also been used for fiducial marker detection in low-light settings. For example, authors in [13] employed a deep CNN for ArUco marker detection, whereas authors in [14] and authors in [15] proposed real-time detection methods for low-power devices. Additionally, hybrid CNN-Transformer models have been explored to enhance marker stability and accuracy [16, 17].

The proposed Transformer–CNN hybrid block combines global feature modeling with local feature extraction to improve low-light image enhancement. It enhances dark regions, preserves local details, and reduces noise. The model uses a mixed attention mechanism, a U-Net discriminator for per-pixel feedback, and perceptual loss to ensure visually coherent results. By combining CNNs and Transformers, it provides robust feature extraction, adaptive enhancement, and effective artifact suppression.

Despite numerous deep learning models for low-light enhancement and marker detection, most have not been tested in diverse low-light conditions. This paper introduces and tests a hybrid Transformer–CNN model, termed Extreme ArUco Vision Transformer (XAViT), using both synthetic and real low-light datasets. Our contribution is a Transformer–CNN hybrid block with a mixed attention mechanism that improves dark region recovery and local detail enhancement, along with a Swin-Transformer-based discriminator to reduce exposure issues, noise, and artifacts. Experimental results show that XAViT outperforms existing methods, setting a new benchmark in the field.

II. MATERIALS AND METHODS

A. *Overall Network Architecture*

The overall architecture of XAViT, shown in Figure 1, consists of three main components: the Parallel Convolutional Encoder (PCE), Aggregate Transposed-Convolutional Decoder (ATD), and Windowed Attention Classification (WAC). The PCE includes three sub-encoders (SE_1, SE_2, SE_3) based on VGGNet, ResNet, and MobileNet. VGG captures low-level features, ResNet enhances deeper feature extraction with residual connections, and MobileNet offers an efficient, lightweight structure for fast inference. This combination balances accuracy, depth, and speed, ensuring robustness across various inputs. To maintain consistent scaling with a global feature resolution of 224×224, SE_1, SE_2, and SE_3 output dimensions are 28×28, 56×56, and 112×112, respectively, enabling multi-resolution local feature extraction.

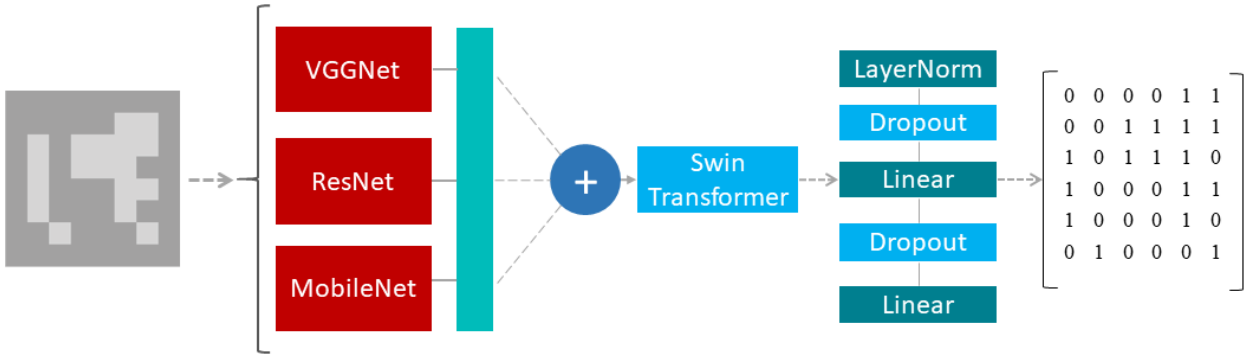

Fig. 1. Overview workflow of the proposed XAViT model framework.

The ATD module, consisting of sub-decoders SD_1, SD_2, and SD_3, produces a final resolution of 224×224, aligning with the outputs of SD_1, SD_2, and SD_3. Features are integrated using combined coefficients. The WAC, built on Swin-Transformer, includes the Encoding Window Attention (EWA) and Mixing Window Attention (MWA) blocks and processes the 224×224 aggregated feature maps from ATD to extract global features. The combination of PCE, ATD, and WAC allows XAViT to effectively capture multi-scale global and local features, from 28×28 to 224×224, improving marker detection and model generalizability across diverse data.

For the PCE module, SE_{1-3} are derived from the upper layers of MobileNet, ResNet, and VGGNet. SE_1 was constructed using the upper 17 segments of VGGNet. Meanwhile, SE_2 and SE_3 are produced using the uppermost five and two child sections of ResNet and MobileNet, respectively. In the ATD module, SD_1, SD_2 and SD_3 incorporate an upsampling layer along with varying numbers of 4×4 transposed convolutional layers. The ReLU activation function is employed throughout this structure.

Feature map-level aggregation within ATD is governed by a weighted summation defined by the coefficients α, β, and γ, constrained such that $\alpha, \beta, \gamma \in [0, 1]$ with $\alpha + \beta + \gamma = 1$. These coefficients control the relative contributions of SD_1, SD_2, and SD_3, managing feature integration across different resolutions. A rule-based search strategy optimizes the coefficient set, allowing XAViT to adjust feature extraction based on dataset attributes. This adjustment balances the impact of multiple scales and ensures effective generalization. SD_1, SD_2, and SD_3 represent the output feature maps, and the final result y of the ATD is expressed as:

$$y = \alpha \mathcal{F}_{SD_1} + \beta \mathcal{F}_{SD_2} + \gamma \mathcal{F}_{SD_3} \quad (1)$$

The WAC module consists of four stages, primarily featuring EWA and MWA modules, which capture spatial hierarchies using local and global attention mechanisms. Both modules utilize multi-head self-attention, combining Window-

based Self-Attention (W-MSA) and Shifting Window-based Self-Attention (SW-MSA) to enhance feature representation by adjusting window sizes for better locality and global context. The EWA block employs a patch embedding layer for rich feature extraction, whereas the MWA block uses a patch merging layer to efficiently combine features from multiple scales.

Each block has a Multilayer Perceptron (MLP) module consisting of two completely linked layers interspersed with a GELU activation function [18]. The predictions head consists of a single symmetrical layer featuring two nodes for output. Let $\hat{z}^l$ with z^l be the results of features generated by the W-MSA and MLP components, respectively, in the l-th block. The computation of attention, using next spatial bias in the EWA as well as the MWA blocks, is delineated as follows:

$$\hat{z}^l = z^{l-1} + \text{W-MSA}\big(\text{LN}(z^{l-1})\big) \quad (2)$$

$$z^l = \hat{z}^l + \text{MLP}\big(\text{LN}(\hat{z}^l)\big) \quad (3)$$

$$\hat{z}^{l+1} = z^l + \text{SW-MSA}\big(\text{LN}(z^l)\big) \quad (4)$$

$$\hat{z}^{l+1} = \hat{z}^{l+1} + \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) \quad (5)$$

B. Loss Functions

Focal loss was used as the main loss function during training to handle class imbalance in real-world image classification tasks. Unlike cross-entropy loss, focal loss adds a factor that changes how much each sample affects the total loss based on the prediction's confidence. This reduces the loss for correctly classified samples and gives more weight to difficult, misclassified ones. This helps the model focus on harder cases. The formula for focal loss is:

$$\mathcal{L}_{focal} = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t) \quad (6)$$

Here, $p_t \in [0,1]$ denotes the predicted probability associated with the ground truth class among the five marker classes, whereas $\alpha_t \in [0,1]$ serves as a class-specific weighting factor used to balance the contribution of each class during training. Additionally, $\gamma \geq 0$ is the concentrating variable that controls the degree to which simple illustrations are down-weighted. In this study, we empirically established the parameters at $\alpha = 0.8$ and $\gamma = 2$, consistent with configurations shown beneficial in previous research. This setup enables the model to maintain strong performance across both minority and majority classes, improving generalization, especially in cases with highly imbalanced class distributions.

The use of focal loss is key to enhancing convergence and increasing the robustness of the proposed deep learning architecture. By emphasizing difficult-to-classify samples, focal loss reduces the influence of well-classified examples, promoting more balanced training dynamics.

III. EXPERIMENTS AND DISCUSSION

A. Implementation Details

Each dataset is randomly divided into subsets based on a defined configuration. The Synthetic-ArUco dataset is split into training, validation, and testing sets with a 7:1.5:1.5 ratio. For the Realistic-ArUco dataset, the training and validation sets are in an 85:15 ratio, with the test set provided separately. To maintain class balance, the ratio of positive to negative cases is consistent across all subsets. The dataset partitions are shown in Table I. All images are resized and center-cropped to 224×224 pixels and then normalized.

TABLE I. DATASET COMPOSITION FOR SYNTHETIC-ARUCO AND REALISTIC-ARUCO DATASETS

Dataset	Category	Class					Total
		1	2	3	4	5	
Synthetic-ArUco	Train	6,631	6,442	6,247	6,762	6,579	32,661
	Val	1,363	1,429	1,332	1,483	1,391	6,998
	Test	1,384	1,422	1,354	1,392	1,448	7,000
	Total	9,378	9,293	8,933	9,637	9,418	46,659
Realistic-ArUco	Train	556	563	547	563	543	2,772
	Val	127	120	113	111	123	594
	Test	116	117	120	121	120	594
	Total	799	800	780	795	786	3,960

To improve the model's generalization, data augmentation techniques like random scaling, cropping, and horizontal flipping are applied only to the training data. Each model is trained on two datasets: Synthetic-ArUco and Realistic-ArUco. The Synthetic-ArUco dataset, with over 40,000 images, is trained for 15 epochs. The smaller Realistic-ArUco dataset is trained for 30 epochs to allow more learning from the limited data. The initial learning rate is set to 0.003 for both datasets, with the learning rate halved if performance does not improve after five epochs. The batch size is 32, focal loss is used as the loss function, and the Adam optimizer is applied.

For the WAC and PCE components, transfer learning is used, and the WAC forecasting head is removed. These components are initialized with weights pre-trained on the ImageNet dataset [19].

Network components using transfer learning are shown with solid lines in Figure 2, whereas those trained from scratch are shown with dotted lines. The ensemble coefficients α, β, and γ are set to 0.3 by default. All experiments are run on an NVIDIA Tesla T4 GPU with 12 GB memory and an Intel Xeon CPU, using PyTorch version 2.3.0.

B. Dataset and Evaluation Metrics

We employed the subsequent datasets to formulate and evaluate our methodologies, and to compare them with existing ones.

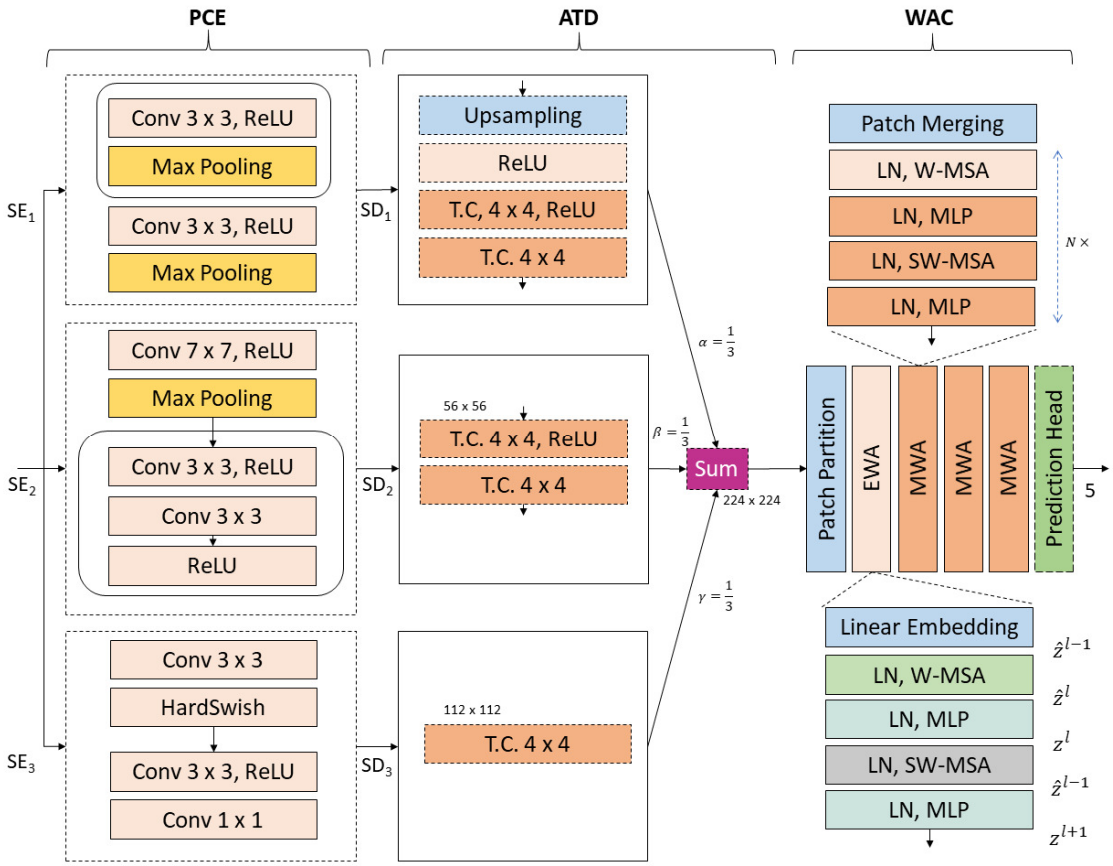

Fig. 2. Detailed XAViT architecture comprising three main blocks: PCE with sub-encoders $SE_1 - SE_3$, ATD with sub-decoders $SD_1 - SD_3$, and WAC. WAC integrates EWA and MWA with linear embedding or patch merging layers, followed by sequential attention modules (W-MSA and SW-MSA). MWA uses $N = 1, 3, 1$ attention blocks. MLP and LN refer to the multilayer perceptron and layer normalization. Solid lines represent modules trained from scratch, whereas dotted lines indicate pretrained (transfer learning) components.

1) Synthetic-ArUco Dataset

We created the Flying-ArUco dataset for training, which includes images of ArUco markers placed on backgrounds from the MSCOCO [20] 2017 dataset. We selected 2,500 images, rotated them for a wide view, and cropped them to 640×360. Backgrounds were collected by converting images to grayscale and calculating the median luma [21]. The images were grouped into five brightness bins, with an equal number of samples taken from each. The selected luma values ranged from 0.0039 to 0.9608.

We overlaid up to 20 markers from the ArUco DICT_6x6_250 dictionary on each background. To include negative examples, we added counterfeit markers such as black, color-inverted, and randomly designed markers. To improve realism, the marker luminance was adjusted to match the background lighting. The dataset includes various lighting and shadow conditions, making it suitable for training marker detection algorithms. Example images are shown in Figure 3.

The dataset, shown in Table I, consists of 46,659 images, split into 32,661 training images and 6,998 validation images. Markers from both sets were cropped for corner regression and marker decoding.

2) Realistic-ArUco Dataset

For the purpose of evaluating our proposed methods and conducting comparisons with other existing techniques, we used an additional dataset captured in real-world environments, called the Realistic-ArUco dataset [22].

To build this dataset, several markers were positioned on the walls of a dimly lit room, illuminated using realistic marker lights to simulate challenging lighting conditions. A video featuring dynamic shadows and lighting effects was then projected onto the corner area, and the scene was captured from various viewpoints, yielding numerous video segments. We employed a semi-manual annotation technique to determine the ground-truth IDs of the markers in the scene. Rather than relying on automatic detection of marker coordinates, we manually annotated the marker positions using a dedicated annotation tool.

Due to the camera's static position during each video sequence, the positioning of markers was uniform across frames. If a marker was not originally visible in a specific frame, its corner dimensions and ID were explicitly designated in an adjacent frame where it was clearly visible, and this information was subsequently disseminated to the remaining frames in the sequence. We recorded the scenario from six distinct viewpoints, resulting in five video sequences comprising 4,000 frames at a resolution of 640×640. Figure 4 displays representative frames from this collection.

(a) (b) (c) (d) (e)

Fig. 3. Examples from the Synthetic-ArUco dataset. Challenging training images are created by superimposing markers of diverse sizes, positions, and spatial placements onto background images sourced from the COCO dataset. Selected instances from the detection dataset: (a-e) correspond to classes 1-5.

(a) (b) (c) (d) (e)

Fig. 4. Examples from the Realistic-ArUco dataset. Challenging training images are created by superimposing markers of diverse poses, sizes, and orientations. Selected instances from the detection dataset: (a-e) correspond to classes 1-5.

3) Assessment Criteria

The performance of the proposed model is evaluated using several metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and F1-score (FOS). These metrics are based on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) from the confusion matrix. TP and TN represent correct predictions, whereas FP and FN represent incorrect predictions. All metrics are calculated from these four values as follows:

$$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (7)$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}} \quad (8)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (9)$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (10)$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (11)$$

$$\text{FOS} = \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}} \quad (12)$$

C. ArUco Marker Detection in Low-Light Conditions

To identify the most effective deep learning model for multi-class image classification, we conducted a comprehensive evaluation of five widely-used architectures: MobileNet, VGGNet, ResNet, Swin-Transformer, and the proposed XAViT hybrid model. The evaluation was based on the average area under the ROC curve (AUC) across five classes, measured on training, validation, and testing datasets.

The XAViT model outperforms baseline CNN models (VGGNet, ResNet, MobileNet) and the Transformer model (SwinT) in both controlled and challenging environments for ArUco marker detection. Table II compares the models on the Synthetic- and Realistic-ArUco datasets. While all models perform well on the synthetic dataset, XAViT consistently achieves high accuracy (99.16%), sensitivity (97.86%), and segmentation quality (Dice: 97.89%, IoU: 95.88%) in realistic conditions, such as low light and rotational distortion. Other models, such as ResNet and SwinT, showed performance drops under domain shifts, indicating less robustness to visual noise and lighting changes. XAViT's hybrid architecture, which combines CNN encoding (VGG, ResNet, MobileNet) with Swin-Transformer attention, captures both local and global features, improving generalization across domains. These results demonstrate XAViT's potential for real-world applications, especially in autonomous systems and robotics operating in challenging environments.

D. Learning Rate Evaluation and Comparative Analysis

To assess the effect of different learning rate schedulers on model performance, three methods were evaluated: ReduceLROnPlateau (RLRP), CosineAnnealingLR (CALR), and StepLR (SLR). These schedulers were tested across training, validation, and testing datasets. Their performance is illustrated in Figure 5 (accuracy) and Figure 6 (loss), whereas Table III summarizes their impact on convergence, stability, and generalization. As shown in Table III, RLRP outperformed the others, achieving 98.59% accuracy, 0.9656 precision, 0.9654 recall, and 0.9321 IoU, strong generalization and reduced overfitting. CALR performed slightly worse, with 97.58% accuracy, 0.9390 precision, 0.9390 recall, and 0.8850 IoU. SLR performed the worst, with 83.43% accuracy, 0.6434

precision, 0.5906 recall, and 0.4248 IoU, indicating poor generalization and overfitting. Overall, RLRP was the most effective scheduler, providing better model performance, especially in spatial accuracy (IoU), as shown in Figure 7.

TABLE II. COMPARISON OF MODEL PERFORMANCE ON THE SYNTHETIC-ARUCO AND REALISTIC-ARUCO DATASETS

Dataset	Model	ACC (%)	NPV (%)	PPV (%)	SEN (%)	SPE (%)	FOS (%)	IOU (%)	DICE (%)	AUC (%)		
										Train	Val	Test
Synthetic -ArUco	VGGNet	99.06	99.41	97.73	97.67	99.41	97.69	95.49	97.69	96.12	99.64	99.54
	ResNet	94.55	96.59	86.60	86.77	96.59	86.54	76.50	86.54	88.21	88.95	90.15
	MobileNet	97.24	98.28	93.28	93.05	98.27	93.10	87.14	93.10	94.79	96.24	96.27
	SwinT	94.61	96.62	87.10	86.62	96.62	86.77	76.87	86.77	84.15	95.07	93.24
	XAViT (proposed)	98.59	99.12	96.56	96.54	99.12	96.48	93.21	96.48	94.78	98.44	98.58
Realistic-ArUco	VGGNet	99.71	99.82	99.28	99.27	99.82	99.27	98.56	99.27	99.26	99.20	99.42
	ResNet	93.71	96.07	84.46	84.22	96.07	84.28	73.04	84.28	96.94	96.24	98.48
	MobileNet	97.63	98.53	94.18	94.01	98.52	94.06	88.83	94.06	98.76	98.50	98.72
	SwinT	95.69	97.33	89.89	89.22	97.31	89.30	80.82	89.30	93.24	95.42	96.14
	XAViT (proposed)	99.16	99.48	97.95	97.86	99.47	97.89	95.88	97.89	98.64	99.08	99.66

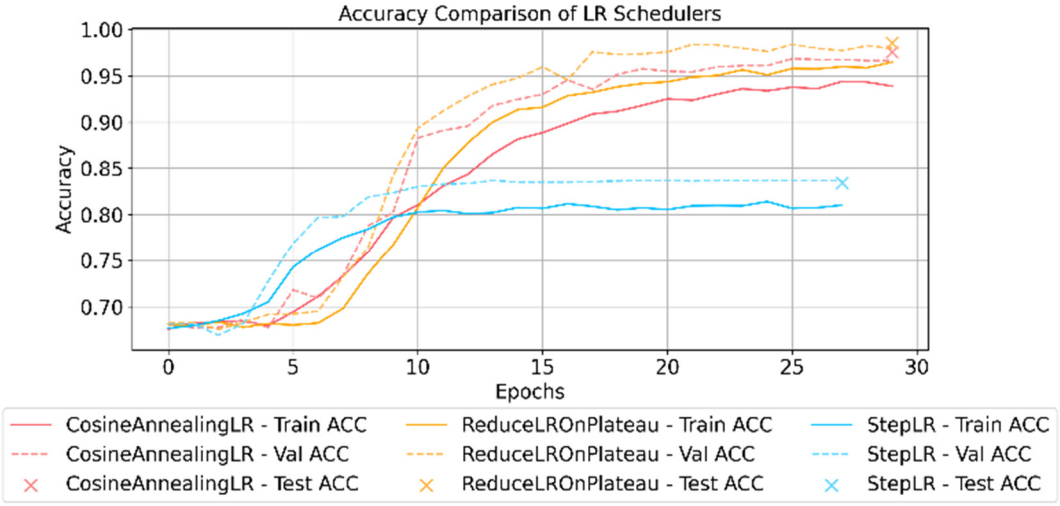

Fig. 5. Comparison of training, validation, and testing accuracy on the Realistic-ArUco dataset.

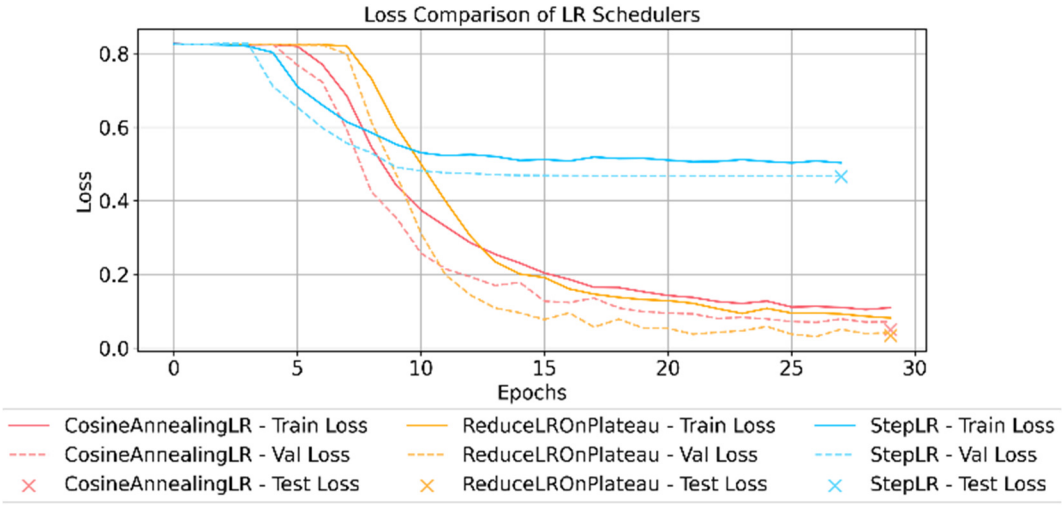

Fig. 6. Comparison of training, validation, and testing loss on the Realistic-ArUco Dataset.

TABLE III. PERFORMANCE OF XAVIT WITH DIFFERENT LEARNING RATE SCHEDULERS ON THE SYNTHETIC-ARUCO DATASET

Type	Split	ACC	PRE	REC	IOU	FOS
RLRP	Train	0.965	0.912	0.912	0.839	0.912
	Val	0.980	0.952	0.950	0.905	0.949
	Test	0.986	0.966	0.965	0.932	0.965
CALR	Train	0.939	0.848	0.847	0.737	0.847
	Val	0.966	0.916	0.916	0.845	0.916
	Test	0.976	0.939	0.939	0.885	0.939
SLR	Train	0.810	0.550	0.523	0.357	0.506
	Val	0.837	0.632	0.604	0.420	0.567
	Test	0.834	0.643	0.591	0.425	0.571

The proposed XAViT model outperformed previous methods, achieving 99.16% accuracy, 97.95% precision, 97.86% recall, 97.89% F1-score, and 95.88% IoU on the Realistic-ArUco dataset. These results highlight the hybrid CNN-Transformer model's superior accuracy and robustness in challenging low-light conditions, where other models often struggle with visual inconsistencies and occlusions. Additionally, XAViT demonstrates versatility by effectively handling various environmental challenges, such as high visual variability and marker size differences. The model's hybrid architecture, leveraging the strengths of both CNNs and Transformers, allows it to dynamically adapt to diverse low-light scenarios, making it highly effective for real-world fiducial marker detection tasks.

TABLE IV. COMPARATIVE PERFORMANCE OF XAVIT AND RELATED STATE-OF-THE-ART MODELS IN MARKER DETECTION

Study, year	Model used	ACC	PRE	REC	IOU	FOS	Notes
[13], 2021	Improved Tiny-YOLOv3	97.20	95.40	94.90	-	95.10	Struggled with small marker detection
[14], 2022	CNN-based detector	-	-	-	-	-	Focused on energy efficiency, not tested in low-light
[15], 2022	CNN	~96.00	-	-	-	-	Marine object detection on embedded systems
[23], 2020	Optical–Inertial	-	-	-	-	-	No marker detection performance metrics
[16], 2024	Wet-ConViT (hybrid)	98.10	96.20	96.70	-	96.40	Wetland classification (satellite data)
[17], 2024	HTViT (hybrid)	98.65	96.50	96.90	-	96.60	General object detection (non-marker specific)
Proposed (XAViT)	CNN + Transformer	99.16	97.95	97.86	95.88	97.89	Best overall results under low-light environments

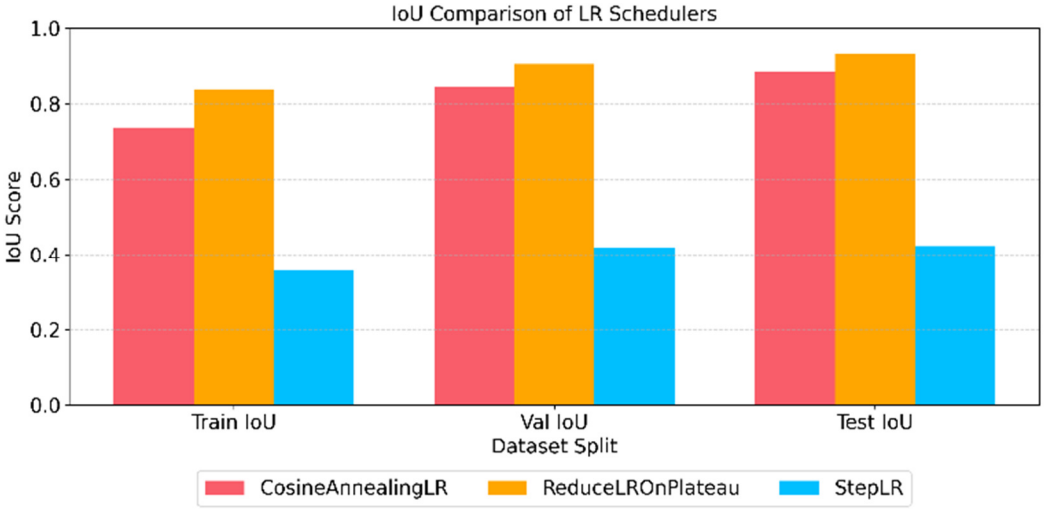

Fig. 7. Comparison of IoU accross different learning rate schedulers (RLRP, CALR, SLR) on the Realistic-ArUco Dataset.The proposed XAViT model was further evaluated through a comparative analysis with recent state-of-the-art studies, as presented in Table IV. Authors in [13] used an improved Tiny-YOLOv3 model for marker detection and achieved 97.2% accuracy, but struggled with small markers and occlusion. Authors in [14] focused on energy-efficient detection, but did not test under low-light conditions. Authors in [15] achieved around 96% accuracy for marine object detection on edge devices, yet lacked robustness testing. Authors in [23] introduced an optical-inertial tracker but without marker classification performance. Hybrid models like Wet-ConViT [16] and HTViT [17] showed good results in other domains, but were not designed for fiducial marker detection.

IV. CONCLUSIONS

This study presents the Extreme ArUco Vision Transformer (XAViT), a novel hybrid Transformer–Convolutional Neural Network (CNN) architecture designed to enhance ArUco marker detection under challenging conditions, including low illumination and extreme rotations. By integrating hierarchical convolutional feature extraction with a multi-head self-attention mechanism, XAViT effectively captures both local and global contextual information, resulting in robust performance across diverse environments. Experimental results demonstrate that XAViT consistently outperforms existing CNN- and Transformer-based models in accuracy, segmentation quality, and generalization to realistic scenarios. The proposed adaptive attention mechanisms contribute significantly to reducing noise and visual distortions while preserving essential marker details. Future work could focus on optimizing the model for real-time deployment through lightweight architectures and exploring advanced loss functions, such as color consistency loss, to further enhance detection fidelity in complex visual settings.

V. LIMITATIONS AND FUTURE WORK

The proposed method's ability to accurately restore color and context is limited under extremely low-light conditions, where objects are nearly invisible. As a result, slight color deviations may appear in enhanced images. Future research should focus on advanced low-light compensation techniques to recover finer details, minimize color distortion, and enhance model performance in these challenging scenarios.

REFERENCES

[1] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 10556–10565, https://doi.org/10.1109/CVPR46437.2021.01042.

[2] J. Liang, Y. Xu, Y. Quan, J. Wang, H. Ling, and H. Ji, "Deep Bilateral Retinex for Low-Light Image Enhancement." arXiv, Jul. 04, 2020, https://doi.org/10.48550/arXiv.2007.02018.

[3] J. Subash and J. Majumdar, "Comparison of Image Enhancement Algorithms for Improving the Visual Quality in Computer Vision Application," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 638–654, Jul. 2022, https://doi.org/10.14569/IJACSA.2022.0130775.

[4] Z. Tian *et al.*, "A Survey of Deep Learning-Based Low-Light Image Enhancement," *Sensors*, vol. 23, no. 18, Sep. 2023, Art. no. 7763, https://doi.org/10.3390/s23187763.

[5] W. Huang, Y. Zhu, and R. Huang, "Low Light Image Enhancement Network With Attention Mechanism and Retinex Model," *IEEE Access*, vol. 8, pp. 74306–74314, 2020, https://doi.org/10.1109/ACCESS.2020.2988767.

[6] R. Khan, Q. Liu, and Y. Yang, "A Deep Hybrid Few Shot Divide and Glow Method for Ill-Light Image Enhancement," *IEEE Access*, vol. 9, pp. 17767–17778, 2021, https://doi.org/10.1109/ACCESS.2021.3054505.

[7] X. Li, Q. Yu, X. Pan, and Z. Yu, "Research on the Contrast Enhancement Algorithm for X-ray Images of BiFeO3 Material Experiment," *Applied Sciences*, vol. 14, no. 9, May 2024, Art. no. 3546, https://doi.org/10.3390/app14093546.

[8] J. Park, A. G. Vien, J.-H. Kim, and C. Lee, "Histogram-Based Transformation Function Estimation for Low-Light Image Enhancement," in *2022 IEEE International Conference on Image Processing*, Bordeaux, France, 2022, pp. 1–5, https://doi.org/10.1109/ICIP46576.2022.9897778.

[9] Y. Zhou, Y. Wang, and W. Cai, "Cycle-enhance: low-light image enhancement based on CycleGan," in *Second International Conference on Electronic Information Engineering, Big Data, and Computer Technology*, Xishuangbanna, China, 2023, pp. 178–183, https://doi.org/10.1117/12.2674697.

[10] Y. Fan *et al.*, "Laser Image Enhancement Algorithm Based on Improved EnlightenGAN," *Electronics*, vol. 12, no. 9, May 2023, Art. no. 2081, https://doi.org/10.3390/electronics12092081.

[11] X. Zhao and L. Li, "A low-light-level image enhancement algorithm combining Retinex and Transformer," in *International Conference on Remote Sensing, Mapping, and Image Processing*, Xiamen, China, 2024, pp. 704–712, https://doi.org/10.1117/12.3029685.

[12] M. He, R. Wang, Y. Wang, F. Zhou, and N. Guo, "DMPH-Net: a deep multi-scale pyramid hybrid network for low-light image enhancement with attention mechanism and noise reduction," *Signal, Image and Video Processing*, vol. 17, no. 8, pp. 4533–4542, Nov. 2023, https://doi.org/10.1007/s11760-023-02687-9.

[13] M. Elgendy, C. Sik-Lanyi, and A. Kelemen, "A Novel Marker Detection System for People with Visual Impairment Using the Improved Tiny-YOLOv3 Model," *Computer Methods and Programs in Biomedicine*, vol. 205, Jun. 2021, Art. no. 106112, https://doi.org/10.1016/j.cmpb.2021.106112.

[14] A. Fisne, A. Kalay, F. Yavuz, C. Cetintepe, and A. Ozsoy, "Energy-efficient computing for machine learning based target detection," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 24, Nov. 2023, Art. no. e7582, https://doi.org/10.1002/cpe.7582.

[15] D. Heller, M. Rizk, R. Douguet, A. Baghdadi, and J.-Ph. Diguet, "Marine Objects Detection Using Deep Learning on Embedded Edge Devices," in *2022 IEEE International Workshop on Rapid System Prototyping*, Shanghai, China, 2022, pp. 1–7, https://doi.org/10.1109/RSP57251.2022.10039025.

[16] A. Radman, F. Mohammadimanesh, and M. Mahdianpari, "Wet-ConViT: A Hybrid Convolutional–Transformer Model for Efficient Wetland Classification Using Satellite Data," *Remote Sensing*, vol. 16, no. 14, Jul. 2024, Art. no. 2673, https://doi.org/10.3390/rs16142673.

[17] K. Ren, T. Zhang, X. Li, Y. Du, and H. Han, "HTViT: an efficient CNN-Transformer hybrid model with high throughput," in *Optoelectronic Imaging and Multimedia Technology XI*, Nantong, Jiangsu, China, 2024, pp. 332–341, https://doi.org/10.1117/12.3036323.

[18] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)." arXiv, Jun. 06, 2023, https://doi.org/10.48550/arXiv.1606.08415.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[20] S. M. S. Zobly, "Comparison of Different Image Enhancement Methods for Effective Whole-Body Bone Scan Image," *Advances in Bioscience and Bioengineering*, vol. 7, no. 3, pp. 55–59, Aug. 2019, https://doi.org/10.11648/j.abb.20190703.16.

[21] A. Arora *et al.*, "Low Light Image Enhancement via Global and Local Context Modeling." arXiv, Jan. 04, 2021, https://doi.org/10.48550/arXiv.2101.00850.

[22] L. Triyono, R. Gernowo, and Prayitno, "MoNetViT: an efficient fusion of CNN and transformer technologies for visual navigation assistance with multi query attention," *Frontiers in Computer Science*, vol. 7, Feb. 2025, Art. no. 1510252, https://doi.org/10.3389/fcomp.2025.1510252.

[23] F. Cutolo, V. Mamone, N. Carbonaro, V. Ferrari, and A. Tognetti, "Ambiguity-Free Optical–Inertial Tracking for Augmented Reality Headsets," *Sensors*, vol. 20, no. 5, Mar. 2020, Art. no. 1444, https://doi.org/10.3390/s20051444.