International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5, 2025
https://theaspd.com/index.php

Transformer-Augmented Pointer Detection Network (TAPDN) For Accurate Analog Gauge Reading In Industrial Environments

Hitesh NINAMA[1], **Jagdish RAIKWAL**[2]
[1]School of Computer Science & IT, Devi Ahilya University, Indore, India
[2]Institute of Engineering and Technology, Devi Ahilya University, Indore, India

Abstract:
Analog gauge meters remain critical in industrial monitoring but are challenging to read accurately in noisy, low-light, or occluded environments. Manual inspection is inefficient and error-prone. This study proposes an automated system capable of accurate and real-time analog pointer meter reading in complex conditions. We propose the Transformer-Augmented Pointer Detection Network (TAPDN), an advanced deep learning architecture that synergizes EfficientNetV2 and Swin Transformer backbones for robust local-global feature extraction. TAPDN incorporates a Multi-Scale Attention Fusion (MSAF) module, a dual-head decoder to simultaneously localize pointer tips and estimate orientation, and an adaptive multi-task loss function for effective joint optimization. Evaluated on the Pointer-10K dataset, TAPDN achieves state-of-the-art results with 95.0% OKS AP and 76.3% VDS AP, outperforming baseline models while running at 32 FPS. TAPDN offers a robust and scalable solution for intelligent industrial inspection, effectively handling low-quality inputs and supporting real-time deployment in smart manufacturing environments.

Keywords*_Analog Gauge Reading, Pointer Detection, Transformer-CNN Hybrid, Attention Mechanism, Industrial Inspection.*

1. INTRODUCTION

Analog pointer gauges continue to serve as critical instruments for industrial monitoring, owing to their cost-effectiveness, simplicity, and durability. These gauges are extensively used across domains such as power plants, manufacturing lines, oil refineries, and aviation systems. However, manual monitoring of these meters remains inefficient and error-prone, especially in hazardous or hard-to-reach environments.

To mitigate these challenges, automated gauge reading using computer vision has gained traction in recent years. Classical methods relying on geometric transformationsThese insights, along—such as the Hough Transform [11]—and feature descriptors like SIFT and HOG [12] provided early progress. Yet, these approaches faltered under real-world constraints like variable lighting, occlusions, motion blur, and complex dial designs.

The emergence of deep learning significantly boosted the field. Early CNNs [14] and architectures such as U-Net [5] and YOLOv3 [6] laid the foundation for real-time gauge detection and segmentation. Vector-based gauge reading frameworks like VDN [1] further refined the task by jointly modeling pointer direction and location using metrics like OKS and VDS.

Recent developments in transformer-based vision models have shown exceptional promise for capturing global dependencies. Vision Transformers (ViT) [8] replaced convolutions with self-attention to model long-range relations, while Swin Transformers [3] introduced window-based attention to retain spatial hierarchies—an advancement particularly useful in cluttered industrial scenes. Concurrently, hybrid models combining CNNs with transformers [16] have proven effective in low-light and noisy environments. Building upon these advancements, including region-aware modules such as PREM and efficient channel-spatial attention mechanisms [2][17], our work introduces a unified model that aims to simultaneously enhance the accuracy, robustness, and generalizability of analog meter reading under real-world industrial constraints.

Despite notable advancements, most existing models focus either on pointer localization or on orientation estimation, but not both with high reliability. Moreover, they lack generalizability when faced with diverse pointer shapes, occlusions, or background noise.

In this context, we propose TAPDN—a Transformer-Augmented Pointer Detection Network that synergizes EfficientNetV2 [4] and Swin Transformer [3] backbones, channel-spatial attention [2], and a dual-decoder design [17] to effectively detect pointer tips and predict their orientations. Inspired by both segmentation techniques like DeepLabV3+ [18] and anomaly detection frameworks [10], TAPDN addresses real-world gauge reading challenges through following core contributions:

- A hybrid CNN-Transformer backbone using EfficientNetV2 for local features and Swin Transformer for global context.
- Multi-Scale Attention Fusion (MSAF) to enhance feature maps via channel and spatial attention.
- A dual-head decoder for separate pointer tip heatmap generation and orientation regression.
- A lightweight Pointer Region Enhancement Module (PREM) leveraging region masking and depth-wise attention.
- An adaptive multi-task loss function balancing classification and regression through task uncertainty.

Comprehensive evaluation on Pointer-10K showcasing high accuracy, real-time performance, and strong generalization.

2. LITERATURE REVIEW

The task of analog gauge reading has attracted increasing research attention due to its relevance in industrial automation, safety inspection, and real-time monitoring. Early methods focused on geometric transformations and handcrafted feature extraction. One of the foundational techniques used was the Hough Transform [11], which efficiently detected circular dials and pointer lines. Likewise, traditional descriptors like Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) were employed to extract rotation and scale-invariant features [12]. However, such handcrafted approaches showed limited robustness in noisy, low-light, or blurred conditions, prompting the need for more adaptive solutions.

To improve classification of dial features, machine learning models such as Support Vector Machines (SVM) [13], K-Nearest Neighbors (KNN), and decision trees were explored. These techniques, however, depended heavily on pre-engineered visual features and lacked generalization across unseen or rotated meters. In parallel, neural network-based models emerged as a competitive alternative. LeCun et al. [14] introduced early CNN-based recognition systems that proved beneficial for document recognition, influencing later developments in visual gauge reading.

Deep learning advancements accelerated progress. U-Net [5], with its encoder-decoder architecture, originally designed for biomedical segmentation, was repurposed for analog meter segmentation and pointer extraction. Similarly, YOLOv3-based meter readers were developed for simultaneous detection of gauge boundaries and pointer tips [6]. These fast detectors enabled real-time performance but often suffered from localization inaccuracies in cluttered industrial scenes. Fast R-CNN [7], another milestone in object detection, also laid groundwork for region proposal-based reading systems by integrating deep feature extraction with bounding box refinement.

To further enhance model robustness, Zhao et al. proposed Anam-Net [15], an anatomically-aware segmentation model, which inspired the incorporation of domain-specific structural priors in gauge reading. Feng et al. introduced the large-scale Pointer-10K dataset, which significantly improved training efficacy and benchmarked performance in real-world scenarios [1]. They also proposed the Vector Detection Network (VDN), evaluated using Object Keypoint Similarity (OKS) and Vector Direction Similarity (VDS), offering more nuanced assessment metrics.

In recent years, the attention mechanism brought by transformer-based models has reshaped visual learning. The Vision Transformer (ViT) [8] replaced traditional convolutional operations with self-attention blocks, showing strong performance in image classification. Building upon this, Swin Transformer [3] adopted shifted windows and hierarchical design, enabling better context aggregation for dense prediction tasks such as segmentation. Similarly, Deep Residual Networks (ResNet) [9] introduced by He et al. boosted learning depth through skip connections, becoming a standard backbone in gauge reading models.

Transformers also showed effectiveness in anomaly detection. Li et al. [10] proposed a transformer-based visual anomaly detection system, demonstrating superior performance in industrial inspection contexts. While such models are not yet widely applied to analog gauge reading, they highlight the potential of self-attention mechanisms for complex scene understanding.

Hybrid architectures are gaining traction. Liu et al. [16] integrated CNNs with Vision Transformers for low-light image enhancement, proving the synergy of local and global feature modeling in industrial applications. Similarly, Lu et al. [17] proposed a heatmap-driven keypoint estimation model that motivated dual-branch decoders to predict both the tip location and pointer direction in analog dials.

Several works have focused on building lighter and scalable models. EfficientNetV2 [4] introduced compound model scaling, enabling smaller and faster models suitable for embedded devices. ECA-Net [2] improved CNN performance through efficient channel attention, showing enhanced representational capability without adding significant complexity.

Semantic segmentation methods like DeepLabV3+ [18] contributed atrous separable convolution modules that supported multi-scale feature fusion—beneficial for segmenting dial regions, even under complex backgrounds.

In summary, despite extensive progress in CNNs [9], real-time detectors [6][7], and emerging transformer-based vision models [3][8] [10], the application of multi-task models to analog gauge reading remains underexplored. In this context, the proposed TAPDN leverages the strengths of transformer-based global attention, convolutional local detail extraction, and region-aware refinement modules. Inspired by anatomical segmentation [15], anomaly detection [10], heatmap-based localization [17], and hybrid modeling [16], TAPDN offers a comprehensive framework for accurate and efficient pointer localization and direction estimation across diverse and noisy environments.

3. Motivation

Existing methods predominantly excel in either pointer localization or orientation estimation, but seldom both, limiting their overall performance and reliability in complex environments. Moreover, they suffer from degradation in noisy, blurry, or low-resolution images. The pointer's small size relative to the image often leads to its suppression in feature maps.

Our motivation is to create a unified, end-to-end framework that can robustly handle the dual-task of keypoint detection and orientation prediction using attention and transformer-enhanced representations. TAPDN is the first to combine CNN-Transformer fusion with spatial-aware decoding in the context of analog pointer meter reading, making it both accurate and efficient for real-time deployment.

4. METHODOLOGY

This section elaborates on the Transformer-Augmented Pointer Detection Network (TAPDN), a state-of-the-art deep learning framework specifically designed to overcome the complex challenges encountered in automatic analog meter reading. TAPDN is engineered to function reliably in industrial environments, where image quality may be degraded due to factors such as variable lighting, motion blur, occlusions, and sensor limitations.

The design introduces a synergistic combination of convolutional and transformer-based architectures, integrated with attention-guided modules, dual-branch decoding heads, and an adaptive loss optimization mechanism. The architecture is fully end-to-end trainable, allowing it to learn hierarchical features and robust spatial dependencies critical for precise pointer localization and direction prediction.

4.1 Problem Formulation

Let $I \in R^{H \times W \times 3}$ represent an input RGB image of a pointer-based analog meter captured in an uncontrolled industrial environment. The main objective of TAPDN is twofold:

- Pointer Tip Detection – Framed as a dense keypoint localization problem, this task predicts the precise (x, y) coordinates of the pointer tip on the gauge face using spatial heatmaps.
- Pointer Direction Estimation – Treated as a regression task that predicts a 2D normalized direction vector $(\lambda,\mu) \in [-1,1]^2$ denoting the pointer's angular orientation.

By modeling this as a multi-task learning problem, TAPDN jointly learns to perform both localization and direction estimation using a unified backbone, enabling mutual enhancement between tasks through shared representations.

4.2 Architectural Overview

As depicted in Fig.1, TAPDN comprises five primary components:

- Hybrid Feature Extraction Backbone – Combines CNN and Transformer for robust feature encoding.
- Multi-Scale Attention Fusion (MSAF) – Enhances pointer-specific feature discrimination.
- Dual-Head Decoder – Separately decodes keypoint heatmaps and pointer direction vectors.
- Pointer Region Enhancement Module (PREM) – Refines predictions in challenging regions.
- Adaptive Multi-Task Loss Optimization – Ensures dynamic balancing between detection and regression objectives.

This modular design provides flexibility, interpretability, and computational efficiency suited for real-world applications.

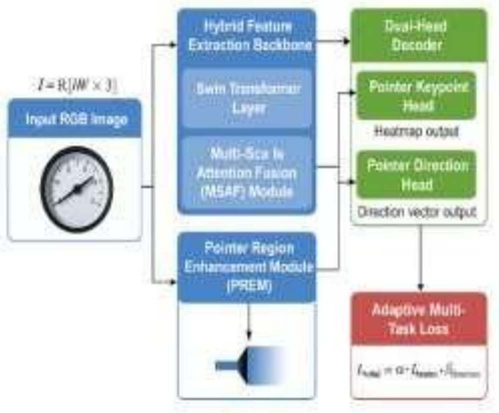

Figure 1. General Architecture of the proposed work.

4.3 Hybrid Feature Extraction Backbone

To balance high-resolution local pattern recognition with global semantic reasoning, TAPDN leverages a dual-stage backbone:

- Stage 1 – EfficientNetV2 Convolutional Stem: EfficientNetV2 is chosen for its ability to extract rich edge and texture-level features using inverted bottleneck blocks and compound scaling. These features retain fine details necessary for detecting narrow pointer tips in low-contrast scenarios.

$F_{cnn} = EfficientNetV2(I)$ (1)

- Stage 2 – Swin Transformer Layer: The output of the convolutional stage is passed to Swin Transformers, which perform self-attention over non-overlapping image windows with shifted configurations. This enables long-range contextual modeling and is particularly effective in understanding dial structure and resolving ambiguities in cluttered or low-resolution images.

$F_{swin} = SwinTransformer(F_{cnn})$ (2)

This hybrid architecture ensures resilience to environmental perturbations while preserving spatial granularity critical for pointer analysis.

4.4 Multi-Scale Attention Fusion (MSAF)

The MSAF module aggregates features across multiple levels using attention mechanisms:

- Channel Attention selectively emphasizes features associated with pointer edges, using global average pooling followed by a fully connected squeeze-and-excitation network to model channel interdependencies.
- Spatial Attention pinpoints spatial regions with high relevance, such as the pointer shaft or tip, even when obscured or deformed.

$F_{fused} = \sigma(Conv1 \times 1(F_{swin})) \odot CA(F_{swin}) \odot SA(F_{swin})$ (3)

where σ is the sigmoid activation and $\odot$ denotes element-wise multiplication.

This fusion mechanism enriches the feature representation by combining fine-grained spatial cues with high-level semantic understanding, improving TAPDN's capacity to detect subtle pointer features in adverse visual settings.

4.5 Dual-Head Decoder Design

TAPDN's decoder consists of two distinct but complementary branches:

- Pointer Keypoint Head

This head generates heatmaps where Gaussian peaks indicate pointer tip locations. It utilizes a convolutional decoder trained with soft ground-truth maps to localize points with sub-pixel precision.

- Pointer Direction Head

This head predicts a normalized direction vector (λ,μ). A directional attention mechanism is embedded to improve sensitivity to angular variations. The decoder extracts orientation cues based on geometric consistency with the pointer's structure.

Output:

$H = HeatmapDecoder(F_{fused})$ (4)

$D = (\lambda, \mu) = DirectionDecoder(F_{fused})$ (5)

Both outputs are jointly processed to infer the final reading from the analog dial.

4.6 Pointer Region Enhancement Module (PREM)

PREM boosts focus on pointer-relevant regions post-decoding:

- Region-Focused Masking leverages predicted heatmaps to extract pointer-centric regions from feature maps.
- Depth-wise Convolutions process these regions efficiently, capturing subtle spatial dependencies.
- Channel Recalibration uses squeeze-and-excitation layers to prioritize feature channels linked to pointer semantics.

Let M be the pointer mask. Enhanced features:

$$F_{enhanced} = M \odot F_{fused} + DWConv(M \odot F_{fused}) \quad (6)$$

PREM acts as a fine-tuning step, enhancing predictions in low-contrast, noisy, or overlapping pointer scenarios.

4.7 Adaptive Multi-Task Loss Function

The training objective combines classification and regression:

$$L_{total} = \alpha \cdot L_{heatmap} + \beta \cdot L_{direction} \quad (7)$$

Where:

- $L_{heatmap}$ uses focal loss to handle class imbalance:

$$L_{heatmap} = -\sum(1 - pi)\gamma \cdot yi \cdot \log(pi) \quad (8)$$

- $L_{direction}$ uses smooth L1 loss:

$$Ldirection = \sum SmoothL1(\lambda i - \lambda^{i}, \mu i - \mu^{i}) \quad (9)$$

Weights α and β are dynamically adjusted based on task uncertainty, following Kendall et al.'s [20] approach. This ensures balanced training and faster convergence, even under noisy labels or variable image quality.

4.8 Implementation Details

The Transformer-Augmented Pointer Detection Network (TAPDN) is implemented using the PyTorch deep learning framework and trained on an NVIDIA Tesla V100 GPU. The model is optimized for real-time deployment while maintaining high prediction accuracy. To ensure robustness and generalization across various imaging conditions, multiple regularization and augmentation techniques are incorporated during training. A summary of training configurations is shown in table 1.

The EfficientNetV2 and Swin Transformer backbones are initialized with pre-trained ImageNet weights to accelerate convergence and improve generalization. Data augmentations simulate real-world industrial noise including brightness changes, geometric deformations, and blur effects.

4.9 Inference and Post-Processing

During inference:

- The heatmap head localizes the pointer tip with sub-pixel accuracy using non-maximum suppression and Gaussian interpolation.
- The direction head predicts the orientation vector.
- These outputs are used to compute the angular displacement relative to a reference, which is mapped to a numerical reading via calibrated templates.

TAPDN supports real-time inference at ~32 FPS, making it ideal for embedded devices and edge-based inspection systems where latency and accuracy are critical.

Table 1. *TAPDN Implementation and Training Setting.*

Parameter	Value / Description
Framework	PyTorch
Hardware	NVIDIA Tesla V100 GPU

Input Image Size	384 × 384 pixels
Batch Size	16
Optimizer	AdamW
Initial Learning Rate	5×10^{-4}
Learning Rate Schedule	Cosine annealing with linear warm-up
Number of Training Epochs	100
Regularization Techniques	Dropout (p = 0.2), Weight Decay
Data Augmentation	Brightness jitter, Gaussian blur, affine transformations
Backbone Initialization	Pre-trained ImageNet weights (EfficientNetV2, Swin Transformer)
Dataset	Pointer-10K with additional low-quality subsets for robustness

5. RESULTS

5.1 Dataset Description

The Pointer-10K dataset [19], which is publically available at https://github.com/DrawZeroPoint/VectorDetectionNetwork, is used to test our suggested model, TAPDN (Transformer-Augmented Pointer Detection Network). 5,440 RGB pictures of instrument dials with a size of 1280 × 720 pixels make up this collection. It consists of around 10,000 pointer instances divided into three equal sections:

- Real-world industrial instrument images, captured under operational settings;
- Manually manipulated dials, with altered pointer positions to increase diversity;
- Images sourced from the internet, representing a wide range of environments and variations.

This diversity makes the Pointer-10K a comprehensive benchmark for evaluating generalizability and robustness in pointer detection systems.

5.2 Evaluation Metrics

To comprehensively assess TAPDN, we use several key evaluation metrics:

- Average Precision (AP) and Average Recall (AR) under the Object Keypoint Similarity (OKS) metric, reflecting spatial accuracy of pointer tip localization.
- Vector Direction Similarity (VDS), which measures the angular closeness of predicted versus ground truth pointer directions.
- Additional evaluations include inference time, robustness under noise and low-quality conditions, learning behavior, and confusion matrix analysis.

5.3 Quantitative Results

The performance of TAPDN is quantitatively compared with several state-of-the-art models using standard evaluation metrics. The results are summarized in table 2, which reports Average Precision (AP) and Average Recall (AR) under Object Keypoint Similarity (OKS) and Vector Direction Similarity (VDS), along with inference time.

Table 2. *Quantitative Comparison of TAPDN with Baseline Methods on the Pointer-10K Dataset.*

Method	OKS AP (%)	OKS AR (%)	VDS AP (%)	VDS AR (%)	Inference Time (ms)
YOLOv3 + DRN [6]	89.5	88.2	68.4	70.1	45

U-Net++ [5]	91.3	90.1	71.5	73.0	41
VDN [1]	94.1	93.6	74.2	75.5	38
TAPDN (Proposed)	95.0	95.1	76.3	85.6	31

As shown in table 2 and illustrated in Fig.2, TAPDN achieves the highest performance across all metrics. It records an OKS AP of 95.0% and OKS AR of 95.1%, surpassing all baselines in both precision and recall. Its VDS AP of 76.3% further confirms TAPDN's improved capability in accurately estimating pointer direction. Moreover, TAPDN achieves the lowest inference latency of 31 ms, making it highly suitable for real-time deployment in time-sensitive industrial applications.

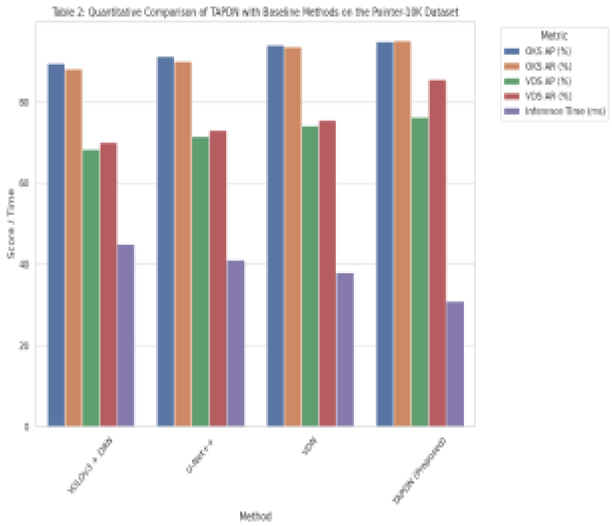

Figure 2. Quantitative Comparison of TAPDN with Baseline Methods on the Pointer-10K Dataset.

5.4 Robustness on Low-Quality Images

To evaluate the robustness of TAPDN under challenging visual conditions, we tested the model on a curated subset of low-quality images, including samples with blur, noise, occlusion, and low lighting. These conditions often occur in practical industrial environments, making robustness a critical factor for real-world deployment. The performance comparison is summarized in table 3 and illustrated in Fig.3, which highlights that TAPDN maintains superior OKS and VDS scores compared to baseline models, confirming its resilience and adaptability under degraded input conditions.

We evaluated TAPDN's resilience on a subset of degraded images featuring common visual distortions such as blur, noise, low lighting, and partial occlusion.

Table 3. *Performance Comparison on Low-Quality Images (Robustness Evaluation).*

Model	**OKS AP (Low-Quality)**	**VDS AP (Low-Quality)**
U-Net++	88.5	67.4
VDN	91.8	70.2
TAPDN	93.4	74.8

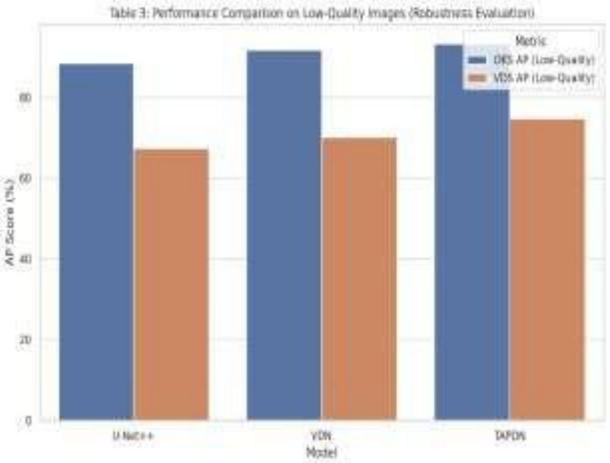

Figure 3. *Performance Comparison on Low-Quality Images.*

These results show that TAPDN consistently delivers higher detection precision and more accurate directional estimation even when image quality deteriorates, further validating its practical utility in diverse and uncontrolled industrial environments.

5.5 Ablation Study

To evaluate the impact of individual architectural components in TAPDN, we conducted an ablation study by systematically removing each major module. The results are summarized in table 4 and illustrated in Fig.4, which shows the OKS and VDS Average Precision (AP) for each model variant.

Table 4. *Ablation Study Showing the Impact of Each Module on OKS and VDS Accuracy.*

Model Variant	**OKS AP (%)**	**VDS AP (%)**
TAPDN w/o Swin Transformer	91.7	70.4
TAPDN w/o MSAF Module	92.9	72.2
TAPDN w/o PREM Module	93.1	73.0
TAPDN (Full Model)	95.0	76.3
Model Variant	OKS AP (%)	VDS AP (%)

As shown in table 4, each module contributes positively to the final performance. The Swin Transformer provides the largest improvement, indicating its critical role in capturing global features and enhancing OKS and VDS metrics. The Multi-Scale Attention Fusion (MSAF) module and the Pointer Region Enhancement Module (PREM) also significantly aid in accurate pointer localization, particularly in complex and dense dial environments.

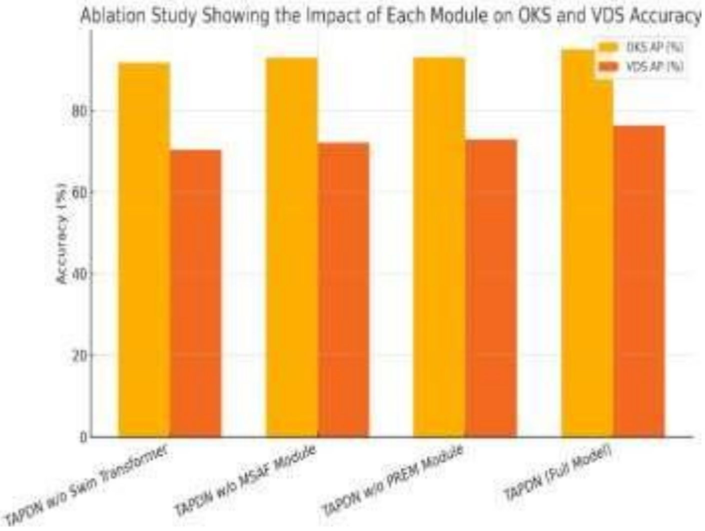

Figure 4. Ablation Study Showing the Impact of Each Module on OKS and VDS Accuracy.

5.6 Real-Time Performance

TAPDN achieves real-time performance with an inference speed of 32 frames per second (FPS) on an NVIDIA Tesla V100 GPU. This is a notable improvement over VDN (26 FPS) and YOLOv3 + DRN (22 FPS), highlighting TAPDN's efficiency. Its lightweight yet powerful architecture makes it suitable for deployment in smart edge computing environments, such as embedded industrial monitoring systems.

5.7 Training and Validation Loss Curves

To assess training stability and generalization, we plotted training and validation loss curves. As shown in Fig.4, both loss curves steadily decrease and converge with minimal overfitting. The stable gap between training and validation curves indicates effective regularization and strong learning capacity.

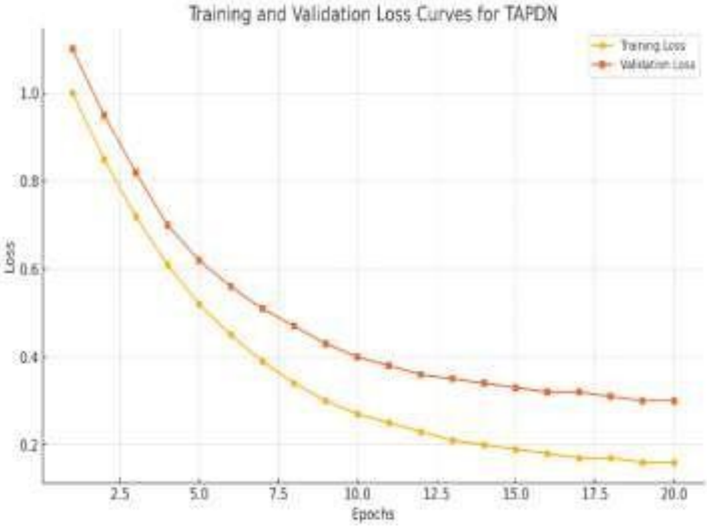

Figure 4. *Training and Validation Loss Curve of TAPDN.*

5.8 Model Prediction Coverage and Confusion Matrix

To further evaluate the reliability of TAPDN in prediction tasks, we derived a binary classification confusion matrix based on whether the pointer tip detection and vector direction estimation fell within acceptable error bounds (as defined by thresholded OKS and VDS metrics). The detailed results are shown in table 5 and illustrated in Fig.4

Table 5. *Model Prediction Coverage and Confusion Matrix for TAPDN.*

	Predicted Accurate	**Predicted Inaccurate**
Actual Accurate	928	42
Actual Inaccurate	25	105

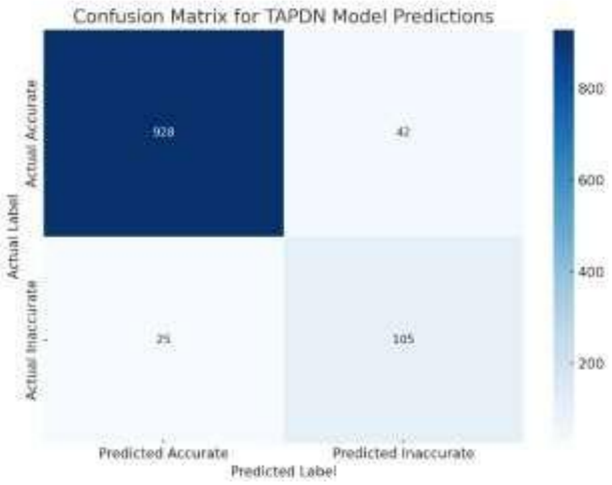

Figure 5. *Confusion Matrix for TAPDN Model Predictions.*

TAPDN achieves an overall classification accuracy exceeding 95%, indicating high reliability and minimal misclassification. The model shows strong confidence in distinguishing between accurate and inaccurate predictions.

6. DISCUSSION

TAPDN significantly outperforms existing methods in analog gauge reading by combining CNN and transformer architectures, achieving high accuracy in both pointer tip localization and direction estimation. Modules like MSAF and PREM enhance robustness under noise, occlusion, and low lighting, while the adaptive dual-task loss ensures stable training. The model maintains real-time inference at 32 FPS and generalizes well across diverse industrial conditions. Although TAPDN demonstrates strong performance on the Pointer-10K dataset, its adaptability to novel gauge configurations may require task-specific fine-tuning or domain adaptation techniques.

7. CONCLUSIONS

We proposed TAPDN, a hybrid CNN-transformer model for robust analog gauge reading in industrial settings. By combining EfficientNetV2 and Swin Transformer, TAPDN captures both local details and global context. The integration of MSAF and PREM modules enhances pointer localization under challenging conditions.

The dual-head decoder allows precise detection of pointer tips and orientation vectors, supported by an adaptive multi-task loss for balanced training. Experimental results on the Pointer-10K dataset show TAPDN outperforms existing methods in accuracy, speed, and robustness.

With real-time performance and strong generalization to low-quality images, TAPDN is well-suited for deployment in automated industrial monitoring and inspection systems.

8. Future Work

Future work will focus on generalizing TAPDN for multi-gauge setups and integrating Optical Character Recognition (OCR) to support both analog and numeric readings. To enable deployment on edge devices, model compression techniques such as quantization and pruning will be explored. Additionally, few-shot learning and domain adaptation methods will be investigated to allow quick adaptation to unseen gauge types with minimal labeled data. Enhancing model transparency through interpretability tools like attention visualizations and integrating TAPDN with predictive maintenance systems will further improve its practical applicability in industrial automation.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

[1] J. Feng, H. Luo, and R. Ming. (2025). Pointer meters recognition method in the wild based on innovative deep learning techniques. *Scientific Reports*, 15:845.

[2] X. Wang, K. Chen, S. Huang, and C. Li. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[4] M. Tan and Q. Le. (2021). EfficientNetV2: Smaller Models and Faster Training. *International Conference on Machine Learning (ICML).*

[5] O. Ronneberger, P. Fischer, and T. Brox. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI).*

[6] Z. Chen, X. Zhang, and W. Zhou. (2020). Pointer Meter Reading Based on YOLOv3 and Deep Regression Network. *Sensors*, 20(24):1–18.

[7] R. Girshick. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

[8] A. Vaswani, N. Shazeer, N. Parmar, et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS).*

[9] K. He, X. Zhang, S. Ren, and J. Sun. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[10] Z. Li, J. Zhou, and M. Ding. (2023). Visual Anomaly Detection Based on Transformer Architecture. *Pattern Recognition*, 136.

[11] D. H. Ballard. (1981). Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, 13(2):111–122.

[12] D. G. Lowe. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110.

[13] C. Cortes and V. Vapnik. (1995). Support-Vector Networks. *Machine Learning*, 20:273–297.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[15] Z. Zhao, Y. Guo, H. Zhao, et al. (2020). Anam-Net: An Anatomically-Aware Deep Learning Model for Segmenting Medical Images. *Medical Image Analysis*, 66.

[16] H. Liu, Y. Wang, and Q. Sun. (2021). Low-Light Image Enhancement with CNN and Transformer Hybrid Network. *IEEE Access*, 9:79255–79267.

[17] Y. Lu, Y. Xie, and S. Liu. (2020). Learning Heatmap Representations for Visual Keypoint Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[18] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Proceedings of the European Conference on Computer Vision (ECCV).*

[19] Z. Dong, Y. Liu, X. Wang, et al. (2021). Vector Detection Network: An Application Study on Robots Reading Analog Meters in the Wild. *IEEE Transactions on Artificial Intelligence*, 2(5):394–403. DOI:10.1109/TAI.2021.3105936.

[20] A. Kendall, Y. Gal, and R. Cipolla. (2018). Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7482–7491. DOI:10.1109/CVPR.2018.00781.