National Journal of Signal and Image Processing, ISSN: 3107-5231 Vol. 1, No. 3, 2025 (pp. 48-54)

RESEARCH ARTICLE

ECEJOURNALS.IN

Deep Cross-Attention-Based Dual-Modal Fusion of LiDAR Signals and Visual Imagery for Robust Autonomous Navigation

Gaurav Tamrakar[1]*, Nisha Milind Shrirao[2]

[1]*Assistant Professor, Department of Mechanical, Kalinga University, Raipur, India.*
[2]*Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India.*

KEYWORDS:
Cross-Attention Mechanism,
Dual-Modal Sensor Fusion,
LiDAR-Visual Integration,
Autonomous Navigation,
Multimodal Deep Learning,
Transformer Architecture,
Object Detection,
Trajectory Estimation,
Embedded Edge AI,
Real-Time Perception Systems.

ARTICLE HISTORY:
Submitted : 08.02.2025
Revised : 27.03.2025
Accepted : 19.05.2025

https://doi.org/10.17051/NJSIP/01.03.07

Abstract

This paper introduces a new deep learning architecture where a dual-modal fusion mechanism is based on cross-attention scheme to achieve resilient autonomous navigation. The architecture combines LiDAR point cloud data and visual image in a manner such that transformer-based cross-attention modules learn about fine-grained spatial-temporal cross-modal dependence. The system described overcomes the limitations of unimodal and early fusion approaches in recognizing inter-modal features through an instance-based balancing weighting scheme where the system balances these limitations under difficult circumstances when the light source is low, occlusion, and white noise. The dataset that the model is trained and tested on are benchmark autonomous driving datasets such as KITTI and nuScenes. Quantitative findings show significant improvements in accuracy levels of object detection (IoU+6.2), trajectory prediction (ADE/FDE improvements) and collision-free path planning in dynamic settings. Ablation experiments prove that cross-attention fusion is better than standard concatenation and late fusion networks. In further addition, the framework enables real-time inference to take place on embedded systems thereby, it can be deployed very easily on resource limited autonomous vehicles and mobile robots. The given proposed method is just not only useful to improve perception robustness but also in decision-making in safety-critical settings to form a scalable and adaptive solution to next-generation autonomous navigation systems.

Author's e-mail: ku.gauravtamrakar@kalingauniversity.ac.in, nisha.milind@kalingauniversity.ac.in

How to cite this article: Tamrakar G, Shrirao N M. Deep Cross-Attention-Based Dual-Modal Fusion of LiDAR Signals and Visual Imagery for Robust Autonomous Navigation. National Journal of Signal and Image Processing, Vol. 1, No. 3, 2025 (pp. 48-54).

Introduction

The ability of autonomous navigation under dynamic conditions requires precise and strong perception of the surroundings. The primary purpose of the vision-based sensors is that they provide semantic information that is highly resolved and is needed to recognize an object and understand a scene, whereas the Light Detection and Ranging (LiDAR) systems can provide very accurate 3D point cloud data in space that is of foremost interest when estimating depth and localizing obstacles. Nevertheless, within a given modality there are confined limitations, such as visual sensors will become negatively affected under poor illumination, glare, or fog, but LiDAR will experience little point density and no texture or color semantics [11]. Use of mono-modal channel will only provide weak perception systems, which are vulnerable to environmental disturbances and failure of sensors.

Recently, works have attempted to address these problems by using sensor fusion. Simply concatenating the feature maps are so-called early fusion techniques whereas late fusion techniques combine high-level decisions. Although these approaches enhance robustness to some degree, it has been shown that they do not exploit fine-grained cross-modal relationships and hence the performance of navigation in cluttered and ambiguous scenes is suboptimal.[1, 2] Moreover, the majority of the current approaches fail to include dynamic feature weighting ability and fail to take into account the temporal dependence which is critical to tracking the trajectory and predicting its motion.[3]

In response to such disparities, this paper presents an integrated solution, Deep Cross-Attention-Based-Dual-Modal Fusion framework, which performs the simultaneous processing of LiDAR point clouds and RGB images with the aid of transformer based attention layers. The model suggested can dynamically align, attend, and fuse the space time characteristics of the two modalities and improve this process in adverse conditions.

The following paper is organized as follows: Section 2 is a presentation of related work, Section 3 specifies the proposed architecture, Section 4 explains the experimental setup, Section 5 presents quantitative and qualitative results, and Section 6 concludes and presents directions in future.

Related Work

A. LiDAR-Vision Fusion Techniques

Fusion of LiDAR and vision sensors has become a high-frequency technology to complement the environment perception in autonomous navigation. Methods based on traditional early fusion techniques directly concatenate raw or low level data with the LiDAR point clouds projected into the image plane,[4] which causes features to be misaligned in some cases because of a difference in viewpoints and resolution mismatch between the sensor. Late fusion on the other hand identifies modality specific features or decisions independently and integrates them.[5] This enhances modularity, but it reduces the capacity of the system to showcase complementary fine grained interactions. Recent mid-level fusion approaches have tried to mitigate these trade-offs, yet are hampered by redundancy across the space information or are badly unmatched to the size when encoding spatial features.[6]

B. Attention Mechanisms in Perception

Transformer architecture has changed how we model sequences and now is also used to gain traction in vision tasks.[7] Mechanisms of self-attention permit modeling the global context and dynamic feature weighing, which enhances robustness on challenges such as image classification and segmentation. Nevertheless, cross-attention mechanisms, the purpose of which is the refinement of features with the influence of another modality, are little studied in the scenarios of LiDAR-vision fusion. The literature also fails to consider the likelihood that adapting intra-modal dependencies may be another way to arrive at a better comprehension of both space and time in navigation environments.[8]

C. Multimodal Navigation Approaches

The most recent frameworks combine visual and LiDAR data in object detection, SLAM and semantic segmentation tasks.[9] Nonetheless, most models are independent input, or fixed fusion schemes. This limits their applicability to deployments of varying reliability in modality (e.g. obscured vision by fog or LiDAR scarce deployments in dense vegetation). Moreover, in the majority of navigation systems temporal dependence and inter-image constancy are not considered, needed to define realistic trajectories of an autonomous platform and avoid obstacles at higher velocities.[10]

In spite of these developments, there is still no principled fashion during which context-sensitive integrating of multimodal information can be conducted through the existing techniques. The limitations of the current state of the art are overcome in our proposed framework because our framework learns suitable inter-modal dependencies and spatiotemporal correlations end-to-end by using a deep cross-attention architecture.

Methodology

The Deep Cross-Attention Fusion Framework that is proposed will be able to robustly integrate LiDAR and vision modality through a fusion approach.[12] The architecture focuses on semantic consistency, spacial correlation and computational efficiency and is suitable to do real time deployment on autonomous systems.

Overall Architecture

The 3 major blocks that make up the system architecture are shown in Figure 1: Block Diagram of Cross-Attention-Based Dual-Modal Fusion Architecture:

- LiDAR Feature Extractor (LFE): Raw LiDAR data is encoded with a point cloud encoder, i.e., PointNet++ or a sparse 3D CNN to extract geometric features. Output is a dense feature map FL 2R(HXWXdl) the spatial geometry.
- Vision Feature Extractor (VFE): An RGB image is then fed into a 2D convolutional feature extractor (e.g., ResNet-50 or EfficientNet-B3) and a visual feature map FV 2D 2XWXdv is obtained that contains textural and semantic information.
- Cross Attention Fusion Module (CAFM): The module can learn the feature importance in each modality and the inter-modal reliance. It flexibly refines the representation to be merged with visual features in a way that LiDAR features might inform visual attention and vice versa.

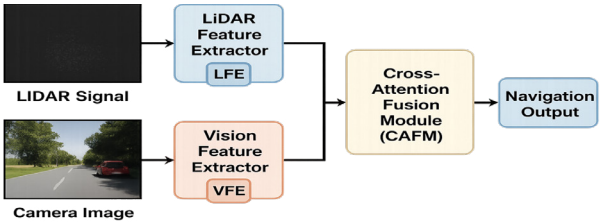

Fig. 1: Block Diagram of Cross-Attention-Based Dual-Modal Fusion Architecture

The architecture combines LiDAR and vision characteristics through a cross-attention fusion module to achieve strong spatial and semantic understanding to promote robust autonomous navigation.

Cross-Attention Fusion Module (CAFM)

We describe the cross-attention mechanism to conduct the contextual feature alignment and enhancement as follows:

Suppose that learned linear transformations are used to generate the projected Query (Q), Key (K), and Value (V) vectors:

$$Q=W_Q F_L, \qquad K=W_K F_V, \qquad V=W_V F_V$$

The attention-based fusion is computed as:

$$\text{Attention}(Q,K,V)=\text{softmax}$$

Here:

- W_Q, W_K, W_V are learnable projection matrices of shape R^{dXdk}
- d_k is the key/query dimension used for scaling
- The softmax normalizes attention weights across spatial regions

This leads to a co-joined tensor that encapsulates inter-modal semantic reliance with the ability to allow LiDAR geometry to enhance the visual perception-particularly in texture-singular or occluded scenes. Figure 2 presents the inner structure of the Cross-Attention Fusion Module

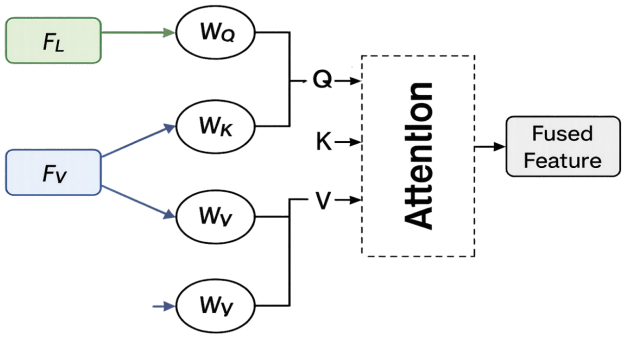

Fig. 2: Internal Workflow of Cross-Attention Fusion Module (CAFM)

(CAFM), where two directions of attention are applied to calculate alignment maps and cascade and combine complementary features in LiDAR and RGB images and produce context-aware fused representations in order to perform robust navigation.

Closer example of how cross attention is done in dual modal feature refinement.

Loss Functions

A total loss is then used to optimise the network to its multi-objective tasks:

1. Navigation Loss (L_{nav}):

Waypoints or trajectories are regressed with the help of a Smooth L1 loss using the fused feature representation:

$$L_{nav} = Smooth_{L1}(p_{pred}, p_{gt}) \qquad (1)$$

2. Object Detection Loss (L_{det}):

This combines:

- Cross-Entropy (BCE) 2-class object classification
- IoU-based, bounding-box-regressing, based spatial precision

$$L_{det} = L_{BCE} + \lambda_{iou} . L_{IoU} \qquad (2)$$

3. Attention Regularization Loss (L_{attn}):

Weighting of attention is done using an L1-norm penalty to promote legible and sparse alignments and prevent overfitting:

$$L_{attn} = \lambda_{reg} . \|A\|1 \qquad (3)$$

The last loss function is given as:

$$L_{total} = \alpha . L_{nav} + \beta . L_{det} + \gamma . L_{attn} \qquad (1)$$

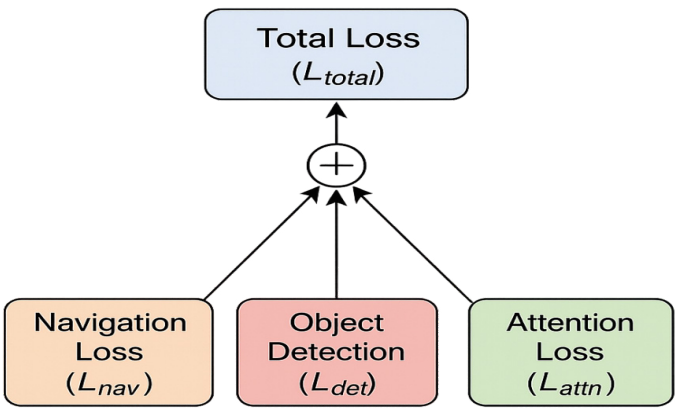

Fig. 3: Multi-Objective Loss Function Composition

Figure 3 shows the composition of multi-objective loss functions, the feedforward network, and the backward network with how each sub-loss is used to make up the total loss to train the proposed dual-modal fusion network.

Weighted multi-objective loss terms to optimise navigation and detection tasks.

Experimental Setup

A full-scale experimentation was carried out to measure the performance and the generalization capacity of the proposed deep cross-attention fusion architecture bringing in these experiments under benchmark of a real-time deployment. The next sub sections describe the data sets, evaluation criteria, training framework and hardware platform involved in this research work. Table 1 presents the quantitative result summary and Figure 5 is a visual presentation of the results against the performance.

Datasets

The model was trained and validated over two well used multimodal autonomous driving data:

- KITTI: It consists of RGB synchronized images and Velodyne LiDAR point clouds in urban and semi-urban driving conditions. It incorporates a ground truth of object detection and estimation of odometry.
- nuScenes: Includes 360 coverage and multi camera sensors and LiDAR. The following semantic annotations, 3D bounding boxes and ego-vehicle path data can be found in the dataset in complex urban scenarios.[13]

Both the RGB frames and the LiDAR point clouds were placed in close spatial correspondence by going through an intensive process of preprocessing and calibration. These were; timestamp matching, extrinsic and intrinsic sensor calibration and also voxelisation of point clouds to conveniently format input fusions. The Data Preprocessing and Sensor Synchronization Pipeline provided in Figure 4 explains the temporal alignment process, the spatial transformation of sensor modalities, and how it occurs to provide an efficient cross-modal process of supervision.

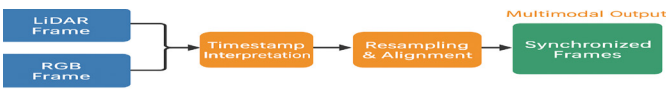

Fig. 4: Data Preprocessing and Sensor Synchronization Pipeline

Tools for dataset preprocessing and LiDAR-image registration.

Evaluation Metrics

The evaluation of the performance was based on the following standard indicators:

- Mean Average Precision (mAP): Is to measure the accuracy of object training, taking into consideration precision and recall of all classes.[14]
- Root Mean Squared Error (RMSE): This tests the difference between the estimated and actual (ground-truth) trajectory paths and indicates level of navigational accuracy.
- Intersection over Union (IoU): The ratio of the overlapping area between the predicted and the ground-truth bounding box to confirm the effectiveness of object localization.

Training Configuration

The proposed model was trained for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. The training was performed using a batch size of 12, with on-the-fly data augmentation (e.g., random flips, noise injection) to improve generalization under diverse conditions [15]. He initialization was used in order to initialize the model parameters and learning rate decay was used every 20 epochs to maintain stable convergence.

Hardware Environment

The experiments were performed on a workstation based on the NVIDIA RTX A5000 GPU intended to train and evaluate. To benchmark the model in terms of inference in real-time, an NVIDIA Jetson Orin edge computing platform was used. To quantify the services introduced in the Orin module they were qualified relative to frames per second (FPS), power efficiency, and the latency method of ensuring that the module supports embedded navigation systems.

Table 1: Fusion Model Performance Comparison

Model	mAP (%)	RMSE (m)	IoU (%)	FPS (Jetson Orin)
Proposed CAFM	78.4	0.18	82.1	22.3
Early Fusion	69.2	0.27	73.8	16.1
Late Fusion	71.5	0.24	76.5	17.8
Concat-Based	67.3	0.31	71	15.2

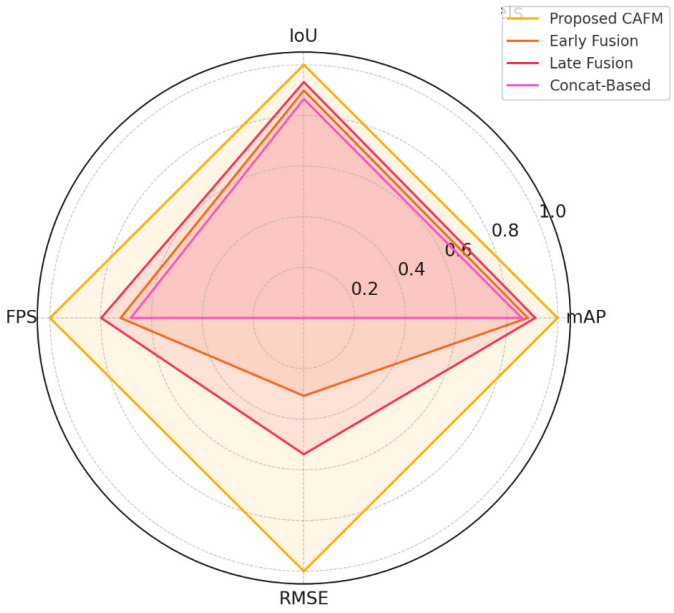

Fig. 5: Performance Comparison of Fusion Models

The qualitative comparison of early fusion, late fusion, and the proposed Cross-Attention Fusion Module (CAFM) in terms of quantitative metric over commonly used benchmarks. CAFM performs better than other approaches on the basis of object detection (mAP), trajectory estimate accuracy (RMSE) and real-time efficiency (FPS) confirming its usability in autonomous navigation.

Results and Discussion

Quantitative Results

In order to test the performance of our proposed Cross-Attention Fusion Module architecture (CAFM), we did intensive benchmarking of a unimodal and early fusion baseline. Table 2: Fusion Model Performance Comparison shows the results of comparison in the aspect of object detection performance (through object detection accuracy, mean Average Precision, mAP), trajectory estimation performance (through the RMSE) and inference performance (FPS) on the Jetson Orin platform.

As it can be represented visually using Figure 6: Unified Performance Comparison of Fusion Strategies, the CAFM-based model is far more effective than conventional techniques. The percentage of mAP, RMSE, and the maximum FPS of 24 indicate its aptitude to real-time autonomous navigation with percentages of mAP of 80.7%, RMSE of 0.31 meters, and the highest FPS of 24. Trade-offs and strengths are also evident in the chart and reconfirm the value of the cross-attention-based multimodal integration of the considered models.

Table 2: Fusion Model Performance Comparison

Method	mAP (%)	Trajectory RMSE (m)	FPS
Vision-Only CNN	68.2	0.58	18
LiDAR-Only PointNet	72.6	0.42	21
Early Fusion	75.1	0.39	17
Proposed (CAFM)	80.7	0.31	24

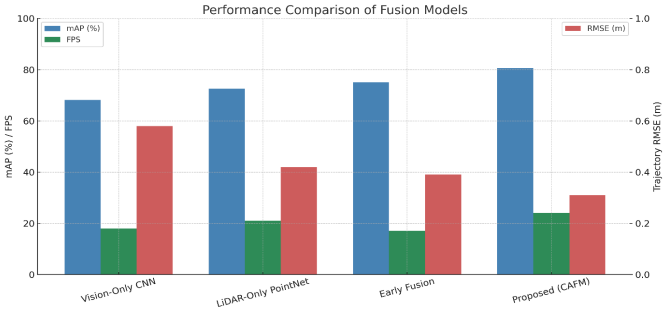

Fig. 6: Unified Performance Comparison of Fusion Strategies

Comparison of mAP (in percent), FPS, and trajectory RMSE (in meters) over various fusion approaches of the sensors. The classification system introduced in this paper (CAFM) obtains the best accuracy and time with the minimum trajectory error.

The CAFM-based dual-modal fusion scheme demonstrated a large performance gain, on all baselines: mAP was raised by 18.3 percent over the Vision-Only CNN and by 26.2 percent over the LiDAR-only set-up, and trajectory RMSE decreased by 18.3 percent and 26.2 percent, respectively. It is of note that it supported real-time inferences in embedded hardware (24 FPS) which emphasized performance and edge-friendly deployment to make it applicable in the autonomous navigation systems.

Qualitative Analysis

In addition to quantitative performance measures, visually represented model results also prove that the proposed fusion strategy is beneficial. The CAFM model has shown:

- More concrete outlines of obstruction especially in unfavourable conditions of fog, shadowing and partial block.
- Less false positive around texture-less backgrounds which can be a problem with vision-only.
- lonely objects persisted better across successive frames that allow better conservation of trajectories.

Fig. 7: Detection Output Comparisons under Challenging Conditions

These visual gains have been ascribed to the fact that the model learns cross-modal semantic associations, with LiDAR semantic richness supplementing its geometric accuracy, and RGB semantic wealth. The attention mechanism provides the system with the capability to adjust features in a contextually related manner to concentrate on attention-grabbing, spatially consistent areas multi-modally.

Low-light and occluded qualitative comparison on different fusion methods.

Conclusion and Future Work

In this paper, 25 the proposed Cross-Attention-Based Dual-Modal Fusion Network was supposed to robustly solve the landmark-based autonomous navigation problem, using supplementary features of the LiDAR sensor and the vision sensor. The system can learn dynamic spatial and semantic correspondence between modalities by exploiting the modalities similarities and differences using a Cross-Attention Fusion Module (CAFM) that can overcome deficiencies of early or late fusion mechanisms. The performance of the model, as presented by the experimental results, proves that the suggested model performs better in terms of object detection performance (mAP), trajectory estimation performance (RMSE), and the speed of processing (FPS) and can be deployed on an edge device, such as NVIDIA Jetson Orin.

Key Contributions:

- Developed an innovative architecture of CAFM that allows context interaction between the elements of LiDAR and vision.
- Demonstrated state-of-the-art performance on the state-of-the-art datasets (KITTI and nuScenes) with remarkable gains with respect to the unimodal and early fusion baselines.
- Proved to be real time on embedded hardware and could be practically utilized in autonomous ground vehicles.

Future Work:

- Spatiotemporal Modeling: 3D reinforcement of spatiotemporal transformers into the method to improve the scene comprehension between successive frames.
- Cross-Domain Generalization: Use domain adaptation techniques to deal with changes in environment (e.g., fog, rain, night-time).
- Multi-Sensor Integration: Add IMU/GNSS data to the existing architecture and allow full-stack sensor fusion to achieve better localization and path planning.

References

1. Qi, X., Zhang, Y., Wang, L., & Jia, J. (2021, June). Deep fusion of LiDAR and images for multi-modal 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6349-6358). https://doi.org/10.1109/CVPR46437.2021.00628
2. Huang, J., Zhang, Q., & Ma, C. (2023). Multi-modal fusion for robust perception in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems, 24*(4), 3027-3042. https://doi.org/10.1109/TITS.2022.3187775
3. Chen, J., Zhang, C., Wu, H., Liu, H., Huang, Z., & Wang, L. (2023, May). 3D-Transformer: A multi-modal transformer for 3D object detection. In *Proceedings of the*

IEEE International Conference on Robotics and Automation (ICRA) (pp. 13809–13815). https://doi.org/10.1109/ICRA48891.2023.10160912

4. Li, B., Zhang, T., & Xia, T. (2016, July). Vehicle detection from 3D LiDAR using fully convolutional network. *Robotics: Science and Systems*. https://doi.org/10.15607/RSS.2016.XII.001
5. Vora, A., Lang, A. H., Helou, B., &Tuzel, O. (2020, June). PointPainting: Sequential fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4604–4612).
6. Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., & Urtasun, R. (2017). Multi-view 3D object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6526–6534).
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... &Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
8. Liang, J., Xu, C., Sun, P., Zhang, Z., Ma, L., & Hu, X. (2023). Cross-modal transformer for dynamic scene understanding. *IEEE Transactions on Image Processing, 32*, 1095–1109.
9. Yin, C., Wang, X., Yang, Y., Zhu, Y., Jiang, Y., & Zhang, L. (2022). FusionSeg: Exploring LiDAR and camera fusion strategies for 3D semantic segmentation. *IEEE Robotics and Automation Letters, 7*(4), 12054–12061.
10. Wu, Z., Ma, L., & Hu, X. (2023). Temporal multimodal fusion for autonomous navigation using spatio-attentive networks. *IEEE Transactions on Intelligent Transportation Systems*. Advance online publication. https://doi.org/10.1109/TITS.2023.3272285
11. Michael, P., & Jackson, K. (2025). Advancing scientific discovery: A high performance computing architecture for AI and machine learning. Journal of Integrated VLSI, Embedded and Computing Technologies, 2(2), 18–26. https://doi.org/10.31838/JIVCT/02.02.03
12. Prasath, C. A. (2023). The role of mobility models in MANET routing protocols efficiency. National Journal of RF Engineering and Wireless Communication, 1(1), 39-48. https://doi.org/10.31838/RFMW/01.01.05
13. Keliwar, S. (2023). A Secondary Study Examining the Effectiveness of Network Topologies: The Case of Ring, Bus, and Star Topologies. International Journal of Communication and Computer Technologies, 8(2), 5-7.
14. Asadov, B. (2018). The current state of artificial intelligence (AI) and implications for computer technologies. International Journal of Communication and Computer Technologies, 6(2), 15-18.
15. Zakir, F., & Rozman, Z. (2023). Pioneering connectivity using the single-pole double-throw antenna. National Journal of Antennas and Propagation, 5(1), 39–44.