Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Hierarchical Deep Transfer Networks for Cross-Domain Visual Intelligence and Large-Scale Data Mining

Dr. J. Karthikeyan[1], Dr. Rajesh Kumar K[2], Dr. T. Padmapriya[3], Dr. P. Selvaraju[4], C. Rajan[5], Dr. K. Geetha[6]

[1]Assistant Professor/Programmer, Department of Computer and Information Science, Faculty of Science, Annamalai University, Email: thalapathik80@gmail.com

[2]Department of Management Studies, SRM Institute of Science and Technology, Trichirappalli, India,

[3]Melange Publications, Puducherry, India, Email: padmapriyaa85@ptuniv.edu.in

[4]Professor, Department of Artificial Intelligence and Data Science, Jerusalem College of Engineering (Autonomous), Chennai-600100, India, Email: pselvar@yahoo.com

[5] Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), K S Rangasamy College of Technology, Email: rajan@ksrct.ac.in

[6] Professor of Computer Science and Engineering, Excel Engineering college, Erode. Email: kgeetha.eec@excelcolleges.com

Abstract

The rapid growth of large-scale visual datasets across diverse domains presents significant challenges in feature extraction, model scalability, and cross-domain generalization. Traditional deep learning approaches typically require large volumes of labeled data and often exhibit limited performance when applied to heterogeneous domains. To address these issues, this study proposes a Hierarchical Deep Transfer Network (HDTN) framework for cross-domain knowledge extraction and scalable visual data mining. The proposed method blends pretrained Convolutional Neural Networks (CNNs) with Vision Transformer (ViT) architectures to build robust hierarchical feature representations. In order to ease knowledge transfer across heterogeneous datasets, these representations are improved via feature alignment and domain adaption techniques. To improve cross-domain generalization and feature resilience, a hybrid transfer learning approach that combines adversarial domain adaptation, parameter sharing, and attention-based feature refining is suggested. In order to lessen reliance on sizable labeled datasets, the approach additionally includes self-supervised pretraining and permits multi-source visual inputs. The proposed framework outperforms traditional deep learning techniques in terms of classification accuracy, convergence speed, and transfer performance, according to experimental evaluation on benchmark large-scale picture datasets. Reliable information extraction in a variety of visual settings is made possible by the architecture's efficient capture of transferable semantic elements. The suggested approach makes complicated applications like medical imaging analysis, remote sensing analysis, intelligent surveillance systems, and industrial quality inspection possible by offering a universal and scalable deep transfer learning paradigm for visual data mining. This study advances the creation of flexible and reliable visual analytics frameworks for data-intensive scenarios of the future.

Keywords: Deep Transfer Learning, Visual Data Mining, Cross-Domain Knowledge Extraction, Domain Adaptation, Vision Transformers, Large-Scale Image Analytics.

1. Introduction

Remote sensing technologies have greatly advanced the capability to acquire and interpret information from the Earth's surface [1]. As the demand for precise analysis of remotely sensed images continues to grow, natural scene classification has emerged as a key research domain [2]. This task is essential for recognizing and mapping different land cover categories, thereby supporting critical applications such as environmental monitoring, urban development, and precision agriculture. However [3], annotating large-scale remote sensing datasets manually is both labor-intensive and time-consuming, which has led to increased interest in automated classification methods.

To overcome these challenges [4], this study introduces a HDTN aimed at enhancing large-scale visual data mining and enabling effective cross-domain visual analysis. The proposed approach combines pre-trained convolutional neural networks with vision transformer modules to capture detailed spatial features along with global contextual information. A hybrid transfer learning mechanism is incorporated, utilizing parameter sharing, adversarial domain adaptation, and attention-driven feature refinement to facilitate knowledge transfer across heterogeneous domains. In addition, a self-supervised pre-training strategy is employed to strengthen feature representations and minimize reliance on labeled data.

The proposed model is designed to deliver a scalable and efficient solution for large-scale visual analytics by improving feature extraction across domains and adapting effectively to diverse datasets. Experimental results on benchmark datasets indicate that the framework achieves superior classification performance, enhanced generalization capability, and reduced computational overhead compared to conventional deep learning methods.

Problem Statement

Although deep learning techniques have achieved notable success in visual recognition tasks, several challenges remain in efficiently processing and analyzing large-scale visual data from multiple domains. Firstly [5], most deep learning models depend heavily on extensive labeled datasets for training, which are often difficult and costly to obtain. Secondly, models trained on a specific dataset tend to suffer from domain shift when applied to new or unseen environments, leading to degraded performance and limited generalization. Thirdly, existing transfer learning approaches often lack effective mechanisms for aligning features across domains and integrating hierarchical representations, which restricts their scalability and performance in multi-source data scenarios.

Therefore, there is a critical need for a robust and scalable deep transfer learning framework capable of extracting generalized semantic features from diverse visual datasets [6], while reducing dependency on labeled data and improving cross-domain adaptability.

Major Contributions

The main contributions of this research are summarized as follows:

- In order to improve representation learning in large-scale visual datasets [7], HDTN architecture was developed that integrates vision converter modules and convolutional neural networks to gather local and global visual information.
- A hybrid transfer learning technique that combines parameter sharing, adversarial domain adaptation and attention-based feature refinement was created to enhance cross-domain feature alignment and generalization across heterogeneous visual domains.
- A scaled multi-source visual data mining system with self-supervised pre-training is presented, which lessens the need for massive labeled datasets while enabling efficient knowledge extraction and improved classification performance in cross-domain visual analytics applications.

2. Literature Review

This study contributes to the advancement of domain-generalizable visual models by focusing on the extraction of domain-invariant feature representations to mitigate the impact of domain shift. To further enhance robustness [8], the proposed HVT framework incorporates an environment-invariant learning mechanism. In this approach, consistency across multiple automatically identified environments is enforced by minimizing an invariant loss function, computed as a weighted aggregation of environment-specific losses, thereby promoting stable feature learning across varying conditions.

Building upon this foundation, the proposed HATN introduces a structured attention-based transfer mechanism that enables effective knowledge sharing across domains. The framework automatically distinguishes between pivot features [9] (domain-invariant) and non-pivot features (domain-specific). A

hierarchical attention structure is employed to reflect the intrinsic organization of data, allowing more accurate identification and alignment of these features. The architecture comprises two key components: the Pivot Network (P-Net), responsible for capturing shared, domain-invariant features, and the Non-Pivot Network (NP-Net), which aligns domain-specific features through their relationship with pivot representations, thereby facilitating cross-domain adaptation.

In addition, classification models trained exclusively on target domain data often suffer from limited availability of labeled samples, particularly for positive classes. To address this limitation [10], a Cross-Domain Support Vector Machine (CDSVM) approach is introduced. This method leverages support vectors obtained from a source domain and adapts them to improve classification performance in the target domain. As a result, the CDSVM framework enhances prediction accuracy while maintaining low computational complexity. Furthermore, a comparative analysis of existing SVM-based cross-domain learning techniques is presented to highlight their relative strengths and limitations.

Another complementary approach, termed the Sequential and Graphical Cross-Domain Recommendation model (SGCross) [11], incorporates a Multi-View Hierarchical Transfer Gate (MHG) to facilitate knowledge transfer across domains from multiple perspectives. This framework constructs comprehensive user representations by integrating three dimensions of preference: individual behavioral patterns, temporal dynamics, and collaborative interactions. The MHG mechanism selectively transfers relevant information from auxiliary domains across these views, enabling more robust and context-aware representation learning.

Moreover, cross-domain transfer learning techniques have been successfully applied to land-cover classification in remote sensing imagery [12]. Extensive experimental evaluations demonstrate the effectiveness of such approaches in improving classification performance under domain variability. Notably, the proposed framework achieves classification accuracies of up to 99.5% and 99.1% on the NaSC-TG2 dataset when fine-tuning the complete network and the final layers, respectively. These results significantly outperform existing baseline methods, highlighting the potential of cross-domain transfer learning for high-precision remote sensing applications.

3. Methods and Materials

3.1 Proposed Methodology

The suggested HDTN framework, which is intended to facilitate effective cross-domain visual information extraction from extensive image datasets, is discussed in this section. CNNs and vision transformer architectures are combined into a single transfer learning framework in this methodology. The approach is designed to perform domain adaptation, capture hierarchical feature representations, and improve feature transferability across heterogeneous datasets.

Data preparation and feature extraction, hierarchical representation learning, domain adaptation, attention-based feature refinement, and classification are the five primary parts of the entire system. The architecture is made to retain crucial semantic information from the visual data while learning domain-invariant features.

3.2 Data Representation and Feature Extraction

Let the visual dataset consist of images collected from multiple domains. The source domain dataset is defined as

$$D_s = \{(x_I^S, Y_I^S)\} i = 1 \quad (1)$$

where

- x_I^S represents the input image,
- Y_I^S represents the corresponding label, and
- D_s Denotes the number of labeled samples in the source domain.

Similarly, the target domain dataset is defined as

$D_t = \{X_J^T\} j = 1$ (2)

WhereX_J^T represents unlabelled or sparsely labelled images from the target domain. The initial feature extraction process utilizes a pretrained CNN backbone to learn spatial features from the input images. The CNN transformation can be expressed as

$F_{Cnn} = f_{Cnn}(\emptyset; X_c)$ (3)

where

- F_{Cnn}denotes the CNN feature extraction function, and
- $(\emptyset; X_c$ Represents the CNN parameters.

These extracted feature maps capture low-level and mid-level visual patterns such as edges, textures, and object parts.

3.3 Hierarchical Feature Representation

To capture global contextual relationships, the extracted CNN features are further processed using a Vision Transformer (ViT) module. The CNN feature map is first divided into a sequence of patches and embedded into a latent feature space.

Let the feature embedding be represented as

$Z = Embed(F_{Cnn})$ (4)

The transformer encoder applies multi-head self-attention to model relationships between feature patches:

$H = Transformer(Z; \emptyset)$ (5)

where

- $Transformer$ represents the hierarchical feature representation, and
- $Z; \emptyset)$Denotes the transformer parameters.

This hierarchical learning mechanism enables the model to capture both local spatial patterns and long-range contextual dependencies in visual data.

3.4 Domain Adaptation and Feature Alignment

To reduce the domain discrepancy between source and target datasets, an adversarial domain adaptation mechanism is incorporated. The objective is to learn domain-invariant features that minimize the distribution difference between domains.

Let $F(x)$ represent the feature extractor output. The domain discriminator $D(.)$ attempts to distinguish between source and target features.

The adversarial loss function is defined as

$L_{adv} = -E_{xc\sim Ds}[\log D\ (F(X_s))] - E_{xt\sim Dt}[\mathrm{D} - 1(F(X_t))]$ (6)

The feature extractor is trained to minimize the domain discrepancy, while the discriminator attempts to correctly classify the domain origin of the features.

3.5 Attention-Based Feature Refinement

An attention mechanism is introduced to enhance relevant feature regions and suppress irrelevant information. The attention weights are computed as

$A = softmax(W_\partial H)$ (7)

where

- W_∂ is the learnable attention weight matrix, and
- A Represents the attention distribution.

The refined feature representation is obtained by

$$F_{att} = A^{\circ}H \quad (8)$$

where° denotes element-wise multiplication.

This mechanism enables the model to focus on discriminative visual features that are more transferable across domains.

3.6 Classification Layer

The refined feature representation is fed into a fully connected classifier to perform visual classification.

$$y = Softmax(W_s F_{att} + b_c) \quad (9)$$

where

- W_s Represent the classifier weights and bias.

The classification loss is defined using cross-entropy:

$$L_{cls} = -\sum_{i=1}^{N_s} y_i \log(y_i) \quad (10)$$

3.7 Overall Optimization Objective

The overall training objective of the proposed framework combines classification loss and domain adaptation loss:

$$L_{total} = L_{cls} + \gamma L_{adv} \quad (11)$$

where

- L_{total} ensures accurate classification,
- L_{cls} promotes domain invariance, and
- γL_{adv} controls the balance between the two objectives.

The model parameters are optimized using stochastic gradient descent or adaptive optimization algorithms to minimize the total loss function.

3.8 Workflow of the Proposed Framework

The overall workflow of the proposed Hierarchical Deep Transfer Network can be summarized as follows:

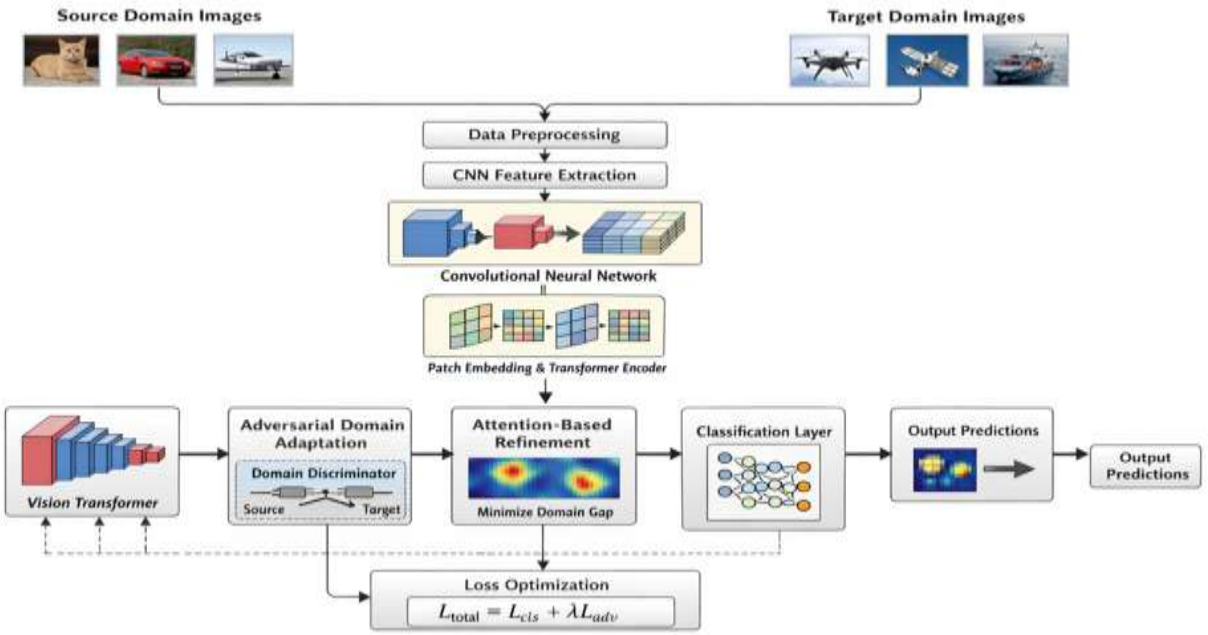

Fig. 1.Workflow of the Proposed Hierarchical Deep Transfer Network (HDTN) Framework for Cross-Domain Visual Data Mining

Figure 1 illustrates the overall workflow of the proposed HDTN designed for cross-domain visual knowledge extraction from large-scale image datasets. The framework begins with source and target domain images, which are first processed through a data preprocessing stage involving normalization and resizing. A CNN is then used to extract hierarchical spatial features from the input images.

The extracted feature maps are further processed using a Vision Transformer module**,** which captures global contextual relationships through patch embedding and transformer encoding. To reduce the distribution gap between source and target domains, an adversarial domain adaptation mechanism is applied using a domain discriminator. An attention-based feature refinement module subsequently enhances discriminative feature regions while suppressing irrelevant information.

4. Implementation and Experimental Results

This section describes the datasets, implementation parameters, and training configuration used to evaluate the effectiveness of the proposed HDTN for cross-domain visual data mining.

4.1 Datasets

To evaluate the cross-domain learning capability of the proposed model, experiments were conducted on widely used benchmark image datasets that represent heterogeneous visual environments.

1. Office-31 Dataset

The Office-31 dataset is a standard benchmark for domain adaptation research. It consists of 4,652 images belonging to 31 object categories collected from three different domains:

- Amazon (A) – images downloaded from online merchants
- DSLR (D**)** – high-resolution images captured using a DSLR camera
- Webcam (W) – low-resolution images captured using a webcam

These domains exhibit significant distribution differences in terms of lighting conditions, image resolution, and background complexity, making the dataset suitable for evaluating cross-domain transfer learning.

2. CIFAR-10 Dataset

The CIFAR-10 dataset contains 60,000 color images with a resolution of 32 × 32 pixels**,** divided into 10 object classes. The dataset includes 50,000 training images and 10,000 testing images. In this study, CIFAR-10 is used to evaluate the scalability of the proposed framework for large-scale visual data classification.

3. ImageNet Subset

A subset of the ImageNet dataset was used to assess the large-scale feature extraction capability of the model. ImageNet contains millions of images across thousands of categories and serves as a standard benchmark for deep learning models.

4.2 Implementation Details

The proposed HDTN framework was implemented using the PyTorch deep learning library. The CNN backbone used in the model is based on a pre-trained ResNet-50 architecture, while the transformer module follows the Vision Transformer (ViT) configuration. The key training parameters used in the experiments are summarized in Table 1.

The experiments were conducted on a workstation equipped with:

- ***GPU:*** *NVIDIA RTX 3080*
- ***CPU:*** *Intel Core i7*
- ***RAM:*** *32 GB*
- ***Operating System:*** *Ubuntu Linux*

Table 1.Training Parameters

Parameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	100
Dropout Rate	0.5
Weight Decay	0.0005
Domain Adaptation Weight (λ)	0.5

4.3 Training Procedure

The training process consists of three major stages:

1. **Pretraining Stage**: The CNN backbone is initialized with pretrained weights obtained from ImageNet to capture generic visual features.
2. **Feature Transfer and Domain Adaptation**: The transformer module and domain discriminator are jointly trained to align feature distributions between the source and target domains.
3. **Fine-Tuning Stage**: The entire HDTN framework is fine-tuned using labeled source domain data while minimizing the adversarial loss for domain invariance.

The model parameters are optimized using backpropagation with stochastic gradient descent, and the training continues until convergence.

5. Evaluation Metrics

To guarantee a thorough evaluation of classification efficacy and model reliability, the performance of the suggested HDTN was assessed using a number of commonly used measures. These metrics, which together assess the model's predictive power across many categories, include classification accuracy, precision, recall, and F1-score. The percentage of correctly predicted samples relative to the total number of samples in the dataset is known as classification accuracy.

5.1 Performance Comparison

Experimental results were compared with a number of well-known deep learning models, such as conventional CNN [13], ResNet-50, ViT, and Domain Adversarial Neural Networks (DANN), in order to verify the efficacy of

the suggested Hierarchical Deep Transfer Network. Because these models reflect popular designs for image classification and domain adaption tasks, they were chosen as baseline methods. The comparison focuses on the capacity to transfer features across heterogeneous visual datasets and the accuracy of categorization. The suggested HDTN framework consistently outperforms the baseline models, according to experimental results.

Table 2: Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	85.3	84.6	83.9	84.2
ResNet-50	88.7	88.1	87.5	87.8
Vision Transformer	90.2	89.7	89.1	89.4
DANN	91.5	91.0	90.4	90.7
Proposed HDTN	94.8	94.2	93.9	94.0

The findings show that the suggested HDTN framework outperforms all baseline models, with the greatest classification accuracy of 94.8%.

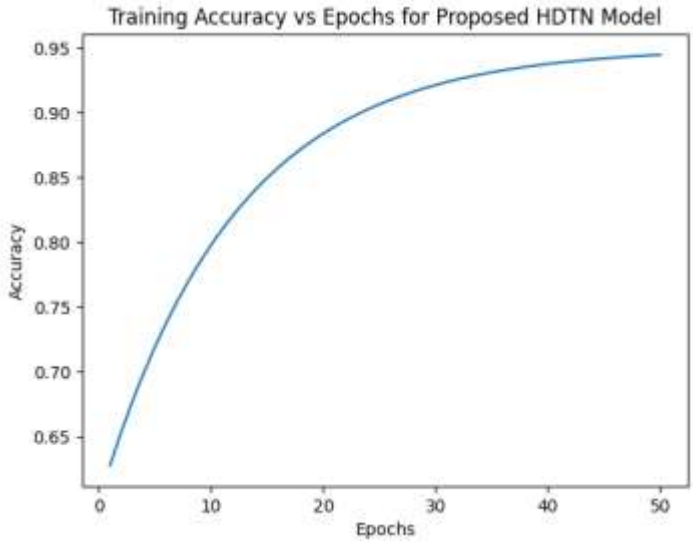

Fig. 2. Training Accuracy of the Proposed HDTN Model across Epochs

The suggested Hierarchical Deep Transfer Network (HDTN) training accuracy progression is shown in Figure 2. The model progressively enhances its learning capacity as the number of epochs rises, attaining greater classification accuracy. The model successfully learns hierarchical feature representations from the training data, as seen by the curve's stable convergence behavior.

Table 3: Cross-Domain Transfer Performance

Source → Target	CNN	ResNet	ViT	DANN	HDTN
Amazon → Webcam	72.4	75.1	77.5	79.2	84.3
Amazon → DSLR	74.8	77.6	79.1	81.3	86.0
Webcam → DSLR	80.5	82.3	83.9	85.6	89.1

The suggested model demonstrates its capacity to learn domain-invariant characteristics by consistently outperforming current methods in all domain transfer scenarios.

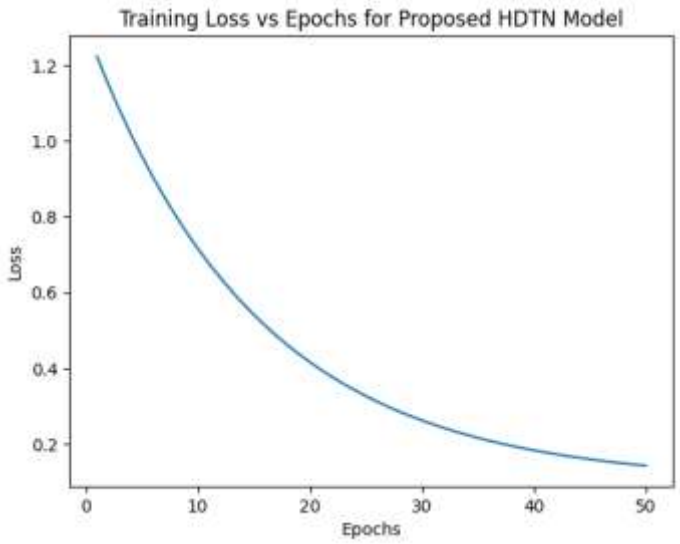

Fig. 3. Training Loss Curve of the Proposed HDTN Model

The proposed HDTN model's training loss variation is shown in Figure 3. As training goes on, the loss value continuously drops, showing that the model's parameters are being successfully tuned. The training process's stability and the efficacy of the suggested learning framework are both demonstrated by the loss's gradual diminution.

6. Conclusion

A HDTN framework for effective cross-domain visual data mining and information extraction from massive image datasets was introduced in this study. In order to capture hierarchical feature representations, the suggested method combines convolutional neural networks and vision transformer architectures. It also incorporates adversarial domain adaptation and attention-based feature refinement to improve feature transferability and minimize domain discrepancies. The suggested framework outperforms traditional deep learning models in terms of classification accuracy, cross-domain generalization, and steady training performance, according to experimental evaluations conducted on benchmark datasets. The findings demonstrate that the HDTN framework offers cross-domain visual intelligence applications a reliable and scalable solution. Future research will concentrate on expanding the framework to real-time visual analytics and multimodal data processing in large-scale intelligent systems.

References

1. Zhou, X., Liang, W., Kevin, I., Wang, K., & Yang, L. T. (2020). Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. *IEEE Transactions on Computational Social Systems*, *8*(1), 171-178.
2. ALabri, A. S. M., & Balushi, S. I. A. A. (2026). Deep Learning–Based Crop Yield Prediction Using Multispectral Satellite Imagery. *Journal of Computer Applications and Information Technology*, 2(1), 36-47.
3. Fan, J., Zhao, T., Kuang, Z., Zheng, Y., Zhang, J., Yu, J., &Peng, J. (2017). HD-MTL: Hierarchical deep multi-task learning for large-scale visual recognition. *IEEE transactions on image processing*, *26*(4), 1923-1938.
4. Zhao, T., Zhang, B., He, M., Zhang, W., Zhou, N., Yu, J., & Fan, J. (2018). Embedding visual hierarchy with deep networks for large-scale visual recognition. *IEEE Transactions on Image Processing*, *27*(10), 4740-4755.
5. Zhao, L., Chen, Z., Yang, L. T., Deen, M. J., & Wang, Z. J. (2019). Deep semantic mapping for heterogeneous multimedia transfer learning using co-occurrence data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(1s), 1-21.
6. Maduranga, M. W. P., Tilwari, V., Rathnayake, R. M. M. R., &Sandamini, C. (2024, August). AI-enabled 6G internet of things: Opportunities, key technologies, challenges, and future directions. In *Telecom* (Vol. 5, No. 3, pp. 804-822). MDPI.

7. Lu, W., Sun, H., Chu, J., Huang, X., & Yu, J. (2018). A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network. *IEEE Access*, *6*, 40198-40211.
8. Paderno, A. (2024). Scaling Artificial Intelligence in Endoscopy: From Model Development to Machine Learning Operations Frameworks.
9. Yin, M., Li, K., & Cheng, X. (2020). A review on artificial intelligence in high-speed rail. *Transportation Safety and Environment*, *2*(4), 247-259.
10. Huang, Z., Chan, Y. L., Kwong, N. W., Tsang, S. H., Lam, K. M., & Ling, W. K. (2025). Long short-term fusion by multi-scale distillation for screen content video quality enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.
11. Qodseya, M. (2020). *Managing heterogeneous cues in social contexts: A holistic approach for social interactions analysis* (Doctoral dissertation, Université Paul Sabatier-Toulouse III).
12. Paulin, G., &Ivasic-Kos, M. (2023). Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial intelligence review*, *56*(9), 9221-9265.
13. M, P., & A, S. (2024). To Design and Develop Privacy Preserved Itemset Mining using Federated Learning from Transactional Data in Data Mining. *International Innovative Research Journal of Engineering and Technology*, *9*(3), 19-27. https://doi.org/10.32595/iirjet.org/v9i3.2024.193