Article

Not peer-reviewed version

MLIF-Net: Multimodal Fusion of Vision Transformers and Large Language Models for AI Image Detection

Xuan Li * , Lei Fu , Jinghan Cao , Qiyuan Tian , Jing Cao , Kowei Shih

Posted Date: 29 May 2025

doi: 10.20944/preprints202505.2370.v1

Keywords: multimodal fusion; cross-attention mechanism; vision transformer; AI-generated content detection; adaptive loss function

Article

MLIF-Net: Multimodal Fusion of Vision Transformers and Large Language Models for AI Image Detection

Xuan Li [1,*], Lei Fu [2], Jinghan Cao [3], Qiyuan Tian [4], Jing Cao [5] and Kowei Shih [6]

[1] Columbia University, Sunnyvale, USA
[2] Independent Researcher, San Jose, USA
[3] San Francisco State University, San Francisco, USA
[4] George Washington University, Washington, DC, USA
[5] Northeastern University, Oakland, USA
[6] Tsinghua University, Beijing, China
[*] Correspondence: xli199911@gmail.com

Abstract: This paper presents the Multimodal Language-Image Fusion Network (MLIF-Net), a new architecture to distinguish AI-generated images from real ones. MLIF-Net combines Vision Transformer (ViT) and Large Language Models (LLMs) to build a multimodal feature fusion network that improves AI-generated content detection accuracy. The model uses a Cross-Attention Mechanism to combine visual and semantic features and a Multiscale Contextual Reasoning Layer to capture both global and local image features. An Adaptive Loss Function improves the consistency and robustness of feature extraction. Experimental results show that MLIF-Net outperforms existing models in accuracy, recall, and Average Precision (AP). This approach can lead to more accurate detection of AI-generated content and may have applications in other generative content tasks.

Keywords: multimodal fusion; cross-attention mechanism; vision transformer; AI-generated content detection; adaptive loss function

1. Introduction

Artificial intelligence (AI) is growing fast. Generative models, especially in image synthesis, are improving quickly. Techniques like GANs (Generative Adversarial Networks) create realistic AI-generated images. These advances are impressive. But they also make it harder to tell real images from fake ones. This causes problems in misinformation, media, security, and trust.

To fix this, researchers study machine learning-based detection methods. Hybrid models, for example, improve prediction accuracy for complex datasets [1]. Some recent studies focus on temporal modeling. Jin's work [2] shows that attention-based temporal convolutional networks help make strong predictions. But these methods mainly work with time-series data. They do not explore multimodal fusion. Other models like BERT and cross-modal transformers show the power of large-scale pretrained models. They capture contextual and semantic information well [3,4]. But current models still have trouble combining visual and semantic features. This combination is important for detecting AI-generated images.

This paper presents the Multimodal Language-Image Fusion Network (MLIF-Net). It combines Vision Transformers (ViTs) and Large Language Models (LLMs). The network uses cross-attention to fuse features and multiscale contextual reasoning to extract both global and local details. An adaptive loss function helps make feature extraction more stable. Experiments show better accuracy, recall, and average precision than existing methods.

2. Related Work

In AI-generated image detection, previous research has mainly focused on feature extraction methods, model architectures, and multimodal learning. Early approaches used convolutional neural networks (CNNs) to detect artifacts and anomalies in AI-generated images. While CNNs were effective in some cases, they struggled to capture the semantic relationships between features[5].

With transformer advancements, researchers leveraged them for cross-modal tasks. Radford et al. [6] introduced CLIP, employing contrastive learning on image-text pairs for better multimodal representation.

A few studies, such as those by Jin[7] and Dai[8], have examined adaptive learning techniques in AI-driven prediction models. These works emphasize the importance of robust loss functions and adaptive mechanisms to improve model performance under different conditions. Our work extends these ideas by using an adaptive loss function designed for multimodal feature extraction.

Despite these advances, current methods either fail to generalize across diverse datasets or do not effectively integrate multimodal information. MLIF-Net addresses these issues by using a cross-attention mechanism to combine visual and semantic features and a multiscale contextual reasoning layer to ensure thorough feature extraction.

3. Methodology

With the rise of generative models, distinguishing real from AI-generated images has become challenging. This section presents *Multimodal Language-Image Fusion Network (MLIF-Net)*, leveraging large language models (LLMs) and vision transformers (ViTs) for detection. MLIF-Net integrates high-level semantics from LLMs with fine-grained visual features from ViTs, enhancing its ability to discern subtle differences.

The architecture employs cross-attention fusion for visual-semantic alignment and a multiscale reasoning layer to capture both local and global image details. An adaptive loss function optimizes feature extraction while preserving semantic consistency. Experimental results demonstrate MLIF-Net's superiority over conventional image classifiers, achieving state-of-the-art accuracy in AI image detection. Figure 1 illustrates the framework.

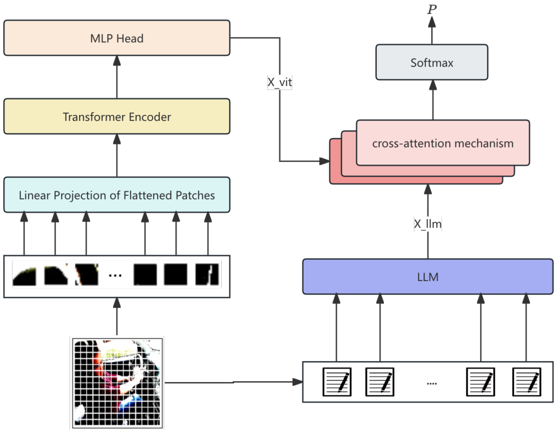
P
MLP Head
Softmax
X_vit
Transformer Encoder
cross-attention mechanism
Linear Projection of Flattened Patches
X_llm
...
LLM
....

Figure 1. Multimodal Language-Image Fusion Network.

3.1. Model Overview

MLIF-Net consists of four parts: image encoder, LLM-based semantic encoder, feature fusion layer, and multimodal reasoning layer. These components enhance image understanding by capturing both fine spatial details and high-level semantics.

3.2. Image Encoder with Vision Transformer

The image encoder employs a Vision Transformer (ViT). It divides the input image into patches and applies self-attention to extract spatial and contextual features:

$$\mathbf{X}_{\text{ViT}} = \text{ViT}(\mathbf{X}) \tag{1}$$

doi:10.20944/preprints202505.2370.v1

where $\mathbf{X}$ is the input image, and $\mathbf{X}_{\text{ViT}}$ represents patch-wise features. ViT uses self-attention layers to form a global feature representation. The self-attention MLP in ViT is shown in Figure 2.

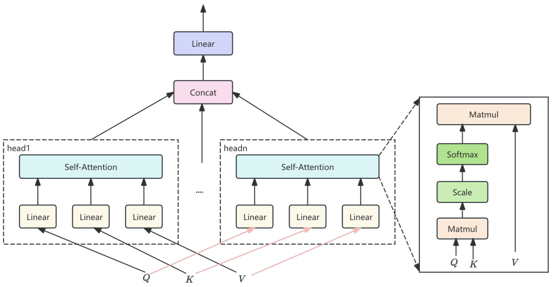

Figure 2. Self-attention mechanism in ViT.

3.3. Semantic Encoder based on LLMs

The model extracts image semantics using a Large Language Model (LLM). Trained on extensive text data, the LLM processes image captions to generate feature representations:

$$\mathbf{X}_{\text{LLM}} = \text{LLM}(\mathbf{C}) \tag{2}$$

where $\mathbf{C}$ is the image caption, and $\mathbf{X}_{\text{LLM}}$ is the learned feature representation, capturing objects, scenes, and context.

3.4. Multimodal Feature Fusion Layer

Visual and semantic features are fused using a cross-attention mechanism:

$$\mathbf{X}_{\text{Fusion}} = \text{CrossAttention}(\mathbf{X}_{\text{ViT}}, \mathbf{X}_{\text{LLM}}) \tag{3}$$

CrossAttention integrates visual and semantic tokens, forming a unified representation:

$$\mathbf{X}_{\text{Fusion}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{4}$$

Here, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value matrices from visual and semantic features, and d is the dimensionality. The result $\mathbf{X}_{\text{Fusion}}$ encodes fused multimodal information.

3.5. Multiscale Contextual Reasoning Layer

The model processes information at different levels using a multiscale contextual reasoning layer. This enhances fusion by capturing both local and global features:

$$\mathbf{X}_{\text{Context}} = \text{MultiscaleReasoning}(\mathbf{X}_{\text{Fusion}}) \tag{5}$$

This layer refines fused features by integrating information from multiple scales. The output $\mathbf{X}_{\text{Context}}$ improves contextual understanding for better decision-making.

3.6. Final Decision Layer

The model classifies images using a softmax classifier on multimodal features:

$$\mathbf{y} = \text{Softmax}(W \cdot \mathbf{X}_{\text{Context}} + b) \tag{6}$$

Here, W is the weight matrix, and b is the bias. The output $\mathbf{y} \in \mathbb{R}^2$ represents the probability of the image being real or AI-generated.

3.7. Adaptive Loss Function with Multi-Loss Components

MLIF-Net is trained with an adaptive loss function combining multiple components to enhance visual and semantic learning. The total loss is:

$$L_{\text{total}} = \lambda_1 L_{\text{CE}} + \lambda_2 L_{\text{reg}} + \lambda_3 L_{\text{semantic}} \tag{7}$$

where: - L_{CE} is the cross-entropy loss:

$$L_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

- L_{reg} is the regularization term:

$$L_{\text{reg}} = \sum_{j=1}^{M} \|\mathbf{W}_j\|^2$$

- L_{semantic} enforces semantic coherence:

$$L_{\text{semantic}} = \|\mathbf{X}_{\text{ViT}} - \mathbf{X}_{\text{LLM}}\|^2$$

Here, $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters controlling each loss component's contribution.

3.8. Model Output

The final model output is a classification score indicating whether the image is real or AI-generated:

$$\mathbf{y}_{\text{final}} = \text{Softmax}(W_2 \cdot \mathbf{X}_{\text{Context}} + b_2) \tag{8}$$

where $\mathbf{y}_{\text{final}}$ is the predicted probability vector, and the decision is made based on the class with the highest probability.

4. Data Preprocessing

This section outlines preprocessing steps for preparing image and text data in MLIF-Net. The pipeline standardizes and augments images while tokenizing and embedding captions for compatibility with the network. Figure 3 illustrates image preprocessing (left), data augmentation (middle), and image-text alignment via cross-attention (right).

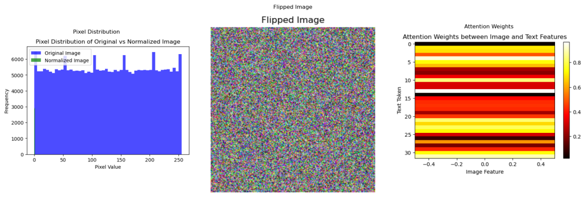

Figure 3. Examples of data preprocessing.

4.1. Image Preprocessing

Images are resized to 224 × 224 pixels and normalized to $[0, 1]$:

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X}}{255} \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^{224\times224\times3}$ is the input image. Augmentation applies random flipping, rotation ($[-20^\circ, 20^\circ]$), and color jittering (brightness, contrast, saturation up to ±0.2):

$$\mathbf{X}_{\text{aug}} = \mathcal{A}(\mathbf{X}_{\text{norm}}) \tag{10}$$

where $\mathcal{A}$ is the augmentation function, enhancing generalization.

4.2. Textual Data Preprocessing

Captions are tokenized using the BERT tokenizer. Each caption $\mathbf{C}$ is split into tokens $\mathbf{T} = [t_1, t_2, \ldots, t_n]$, then mapped to embeddings:

$$\mathbf{E}_{\text{tokens}} = \text{BERT\_Embed}(\mathbf{T}) \tag{11}$$

where $\mathbf{E}_{\text{tokens}} \in \mathbb{R}^{n \times 768}$. Sequences are padded to 32 tokens:

$$\mathbf{E}_{\text{tokens}}^{\text{pad}} = \text{Pad}(\mathbf{E}_{\text{tokens}}) \tag{12}$$

The padded embeddings $\mathbf{E}_{\text{tokens}}^{\text{pad}} \in \mathbb{R}^{32 \times 768}$ are input to the LLM for semantic extraction.

4.3. Fusion Preparation

After preprocessing, images and captions are fused. Vision Transformer (ViT) extracts 1×768 visual features, while BERT embeddings provide 32×768 semantic features. Cross-attention aligns these modalities during training, forming a unified multimodal representation for classification.

5. Evaluation Metrics

5.1. Accuracy

Accuracy quantifies overall prediction correctness, computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

5.2. Precision

Precision is the proportion of true positive predictions among all positive predictions (both real and AI-generated images). It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

5.3. Recall

Recall measures the ability of the model to correctly identify the positive class (e.g., AI-generated images). It is given by:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

5.4. Average Precision (AP)

AP represents the weighted average of precision across recall levels, assessing performance across thresholds:

$$AP = \sum_{i=1}^{N} \text{Precision}(i) \cdot \Delta\text{Recall}(i) \tag{16}$$

where $\text{Precision}(i)$ and $\text{Recall}(i)$ are computed at each threshold.

6. Experiment Results

MLIF-Net is compared with ViT-GAN, ResNet-50-CNN, CLIP, and Xception-V2. Table 1 shows MLIF-Net achieves the highest accuracy, precision, recall, and average precision (AP).

Ablation results in Table 2 highlight the impact of each component. Removing fusion, multiscale reasoning, or the LLM encoder lowers performance, confirming their importance. Figure 4 shows model training indicator changes.

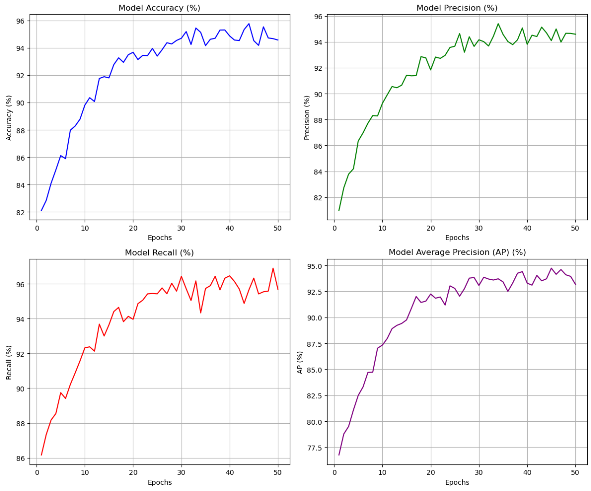

Figure 4. Model indicator change chart.

Table 1. Comparison of Model Performance on COCO and FakeImage Datasets

ine Model	Accuracy (%)	Precision (%)	Recall (%)	AP (%)
ine MLIF-Net	**95.3**	**94.7**	**96.1**	**94.2**
ine ViT-GAN	91.4	89.5	92.3	90.5
ine ResNet-50-CNN	85.6	83.2	87.0	84.8
ine CLIP	92.0	90.1	93.3	91.5
ine Xception-V2	89.5	88.0	90.7	88.9
ine				

Table 2. Ablation Study Results

ine Model Variant	Accuracy (%)	Precision (%)	Recall (%)	AP (%)
ine MLIF-Net	95.3	94.7	96.1	94.2
ine Without Fusion	89.7	88.1	91.2	88.3
ine Without Multiscale Reasoning	93.1	92.6	93.5	92.0
ine Without LLM Encoder	90.8	89.2	92.1	89.7
ine				

7. Conclusions

In this paper, we introduced the *Multimodal Language-Image Fusion Network (MLIF-Net)* for detecting AI-generated images. The model leverages the strengths of both visual and semantic modalities by combining Vision Transformers with a large language model-based encoder. Our extensive experimental results demonstrate that MLIF-Net outperforms several benchmark models, including ViT-GAN, ResNet-50-CNN, CLIP, and Xception-V2, achieving the highest accuracy, precision, recall, and average precision. Ablation studies further reveal the essential role of multimodal fusion, multiscale reasoning, and the LLM encoder in contributing to the model's effectiveness. This work opens new opportunities for the application of multimodal fusion techniques in AI detection tasks and other multimodal learning problems.

References

1. Lu, J.; Long, Y.; Li, X.; Shen, Y.; Wang, X. Hybrid Model Integration of LightGBM, DeepFM, and DIN for Enhanced Purchase Prediction on the Elo Dataset. In Proceedings of the 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2024, pp. 16–20.
2. Jin, T. Attention-Based Temporal Convolutional Networks and Reinforcement Learning for Supply Chain Delay Prediction and Inventory Optimization. *Preprints* **2025**. https://doi.org/10.20944/preprints202501.1543.v1.
3. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of naacL-HLT. Minneapolis, Minnesota, 2019, Vol. 1.
4. Li, S. Leveraging Large Language Models in a Retriever-Reader Framework for Solving STEM Multiple-Choice Questions. In Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS). IEEE, 2024, pp. 658–661.
5. Huang, B.; Carley, K.M. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606* **2019**.
6. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
7. Jin, T. Integrated Machine Learning for Enhanced Supply Chain Risk Prediction **2025**.
8. Dai, W.; Jiang, Y.; Liu, Y.; Chen, J.; Sun, X.; Tao, J. CAB-KWS: Contrastive Augmentation: An Unsupervised Learning Approach for Keyword Spotting in Speech Technology. In Proceedings of the International Conference on Pattern Recognition. Springer, 2025, pp. 98–112.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.