Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 32(12): 6698–6716.
DOI: 10.3934/era.2024313
Received: 30 July 2024
Revised: 28 November 2024
Accepted: 05 December 2024
Published: 13 December 2024

Research article

ViT-DualAtt: An efficient pornographic image classification method based on Vision Transformer with dual attention

Zengyu Cai[1], **Liusen Xu**[2], **Jianwei Zhang**[2,3,*], **Yuan Feng**[4], **Liang Zhu**[1] **and Fangmei Liu**[1]

[1] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450003, China

[2] School of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou 450003, China

[3] Research Institute of Industrial Technology, Zhengzhou University of Light Industry, Zhengzhou 450003, China

[4] School of Elechonic Information, Zhengzhou University of Light Industry, Zhengzhou 450003, China

* **Correspondence:** Email: ing@zzuli.edu.cn.

Abstract: Pornographic images not only pollute the internet environment, but also potentially harm societal values and the mental health of young people. Therefore, accurately classifying and filtering pornographic images is crucial to maintaining the safety of the online community. In this paper, we propose a novel pornographic image classification model named ViT-DualAtt. The model adopts a CNN-Transformer hierarchical structure, combining the strengths of Convolutional Neural Networks (CNNs) and Transformers to effectively capture and integrate both local and global features, thereby enhancing feature representation accuracy and diversity. Moreover, the model integrates multi-head attention and convolutional block attention mechanisms to further improve classification accuracy. Experiments were conducted using the nsfw_data_scrapper dataset publicly available on GitHub by data scientist Alexander Kim. Our results demonstrated that ViT-DualAtt achieved a classification accuracy of 97.2% ± 0.1% in pornographic image classification tasks, outperforming the current state-of-the-art model (RepVGG-SimAM) by 2.7%. Furthermore, the model achieves a pornographic image miss rate of only 1.6%, significantly reducing the risk of pornographic image dissemination on internet platforms.

Keywords: pornographic image classification; Vision Transformer; Convolutional Block Attention Module; Multi-Head Attention; Convolutional Neural Network

1. Introduction

In the digital era, a vast amount of visual content inundates the internet at an unpredictable pace, most of which includes pornography. Pornographic images typically feature explicit or suggestive material deemed inappropriate for public viewing. These images not only degrade user experience but also give rise to societal moral concerns and pose security risks. Prolonged exposure to pornography may negatively impact individuals' mental health, relationships, and daily lives. At a societal level, pornography undermines public moral standards and could potentially affect social stability. Therefore, accurately identifying and filtering pornographic images is crucial for maintaining the safety of online communities and protecting user security.

Image recognition is one of the key technologies in digital image processing during the information age [1], particularly crucial in the field of pornographic image classification. Early research predominantly focused on machine learning methods, often using metrics such as the proportion of exposed skin [2] or specific shapes and textures of intimate body parts [3] as classification criteria. However, these approaches have limitations when dealing with complex and diverse pornographic images. The rapid advancement of deep learning technology has brought new hope to pornographic image classification. Deep learning has made significant strides in tasks such as image classification [4], object detection [5], and image retrieval [6], providing robust technical support for pornographic image classification. Transformer [7], as a powerful deep learning architecture, has achieved notable success in natural language processing and is beginning to explore applications in image classification [8, 9]. However, there is a lack of research on its application in pornographic image classification.

To overcome the limitations of traditional methods and explore a more efficient solution for pornographic image classification, we propose a novel hybrid model called ViT-DualAtt. This model combines the strengths of Convolutional Neural Networks (CNNs), attention mechanisms (Multi-Head Attention and Convolutional Block Attention), and Vision Transformers (ViT) to improve the accuracy and robustness of pornographic image classification. CNNs excel at extracting local features. Transformers capture global features through multi-head attention modules and enhance the model's focus on relevant parts of the image through convolutional block attention modules. Their integration enables the model to achieve comprehensive and precise feature representation, significantly boosting classification performance.

This paper, titled "ViT-DualAtt", makes the following three key contributions:

1) We propose a novel hybrid model, ViT-DualAtt, which integrates the strengths of CNN, multi-head attention, convolutional block attention mechanisms, and Vision Transformer. This model aims to enhance the accuracy and robustness of pornographic image classification, overcoming limitations of traditional approaches.

2) ViT-DualAtt comprehensively integrates multi-head attention and convolutional block attention mechanisms. Multi-head attention helps capture global dependencies, while convolutional block attention optimizes the representation of local features, effectively improving the model's understanding of fine-grained features and global contextual information.

3) By utilizing CNN for extracting local features and Transformer's self-attention mechanism for capturing global features, ViT-DualAtt achieves more comprehensive and precise feature

representation. This fusion not only enhances the model's ability to abstract complex image features but also improves accuracy in classification tasks.

4) Extensive experiments on public datasets validate the effectiveness of ViT-DualAtt in pornographic image classification tasks. Results demonstrate a significant performance with an accuracy of 97.2% ± 0.1% and a pornographic image miss rate of only 1.6%. Compared to other state-of-the-art models, our approach shows notable improvements in performance, confirming its practical effectiveness and superiority.

2. Related works

Great progress has been made in the field of pornographic image classification, spanning a range of techniques from traditional machine learning methods to deep learning approaches. Traditional methods rely on manual feature extraction and classical classification algorithms, while modern methods leverage convolutional neural networks and other deep learning models to enhance classification accuracy and robustness. In this section, we first review pornographic image classification based on traditional machine learning methods, then discuss deep learning-based approaches and their latest advancements in this field, and finally introduce the applications of multi-head attention mechanisms and CBAM mechanisms in image classification.

2.1. Methods based on traditional machine learning

Early methods for pornographic image classification mainly relied on traditional image processing techniques. These approaches typically involved preprocessing the images first, followed by using classifiers such as Support Vector Machines (SVM) [10], K-Nearest Neighbors (KNN) [11], and Naive Bayes [12] for classification. For example, Jones and Rehg [13] developed a color model in the RGB color space and determined whether regions in the image were skin areas based on the brightness of each channel. Lin et al. [14] extracted skin regions from images, identified the correlation between skin and non-skin regions, and fed this correlation into an SVM classifier. Lv et al. [15] proposed a pornographic image filtering model based on high-level semantic features, optimizing the Bag of Visual Words (BoVW) model by integrating contextual visual words and spatial-related high-level semantic features to construct a high-level semantic dictionary. Dong et al. [16] introduced a pornographic image detection method that combined image perceptual features with image-text features. This method integrated text information extracted from image filenames, headers, or webpages with visual features such as texture and local shapes, and used an SVM classifier for classification.

While these methods have shown some effectiveness in certain scenarios, they often fall short when dealing with complex and rapidly changing pornographic images. Particularly when confronted with large-scale, high-dimensional datasets of such images, their performance and efficiency can be compromised. With the rise of deep learning, researchers have increasingly turned their attention to neural network-based approaches to better capture advanced features in pornographic images.

2.2. Methods based on deep learning

Deep learning has achieved significant success in the field of image classification, making it a crucial tool for addressing the problem of pornographic image classification. Deep learning methods utilize multi-layer neural network models to automatically learn high-level features from pornographic images, enabling effective classification of complex content. For instance, Ying et al. [17] proposed a deep convolutional neural network based on feature visualization analysis, employing deconvolutional networks to visualize the features extracted by CNNs. Cheng et al. [18] introduced an image classification approach using deep CNNs, incorporating a multitask learning strategy to optimize the network with global and local information. Cai et al. [19] incorporated a CBAM module into the ResNet101 network to classify gory and pornographic images, effectively avoiding the problem of gradient vanishing during deep network training. Additionally, Cai et al. [20] presented a classification method, RepVGG-SimAM, based on RepVGG and simple non-parametric attention mechanisms (SimAM), enhancing neural networks' capability to acquire more effective information and improve inference speed. The abstracts of referenced studies are summarized in Table 1.

Table 1. Reference overview.

Characteristics	References	Methods	Limitations
Classification based on skin area	[13]	RGB color space	It has achieved certain effectiveness in certain scenarios, but appears to be inadequate when dealing with complex and diverse pornographic images.
	[14]	RGB + SVM	
Classification based on shape and texture	[15]	BoVW	
	[16]	BoVW + SVM	
Deep learning	[17]	CNN	The ability of feature extraction and generalization is relatively insufficient.
	[18]	DCNN	
	[19]	RenNet101 + CBAM	
	[20]	RepVGG-SimAM	

2.3. *Multi-Head Attention mechanism*

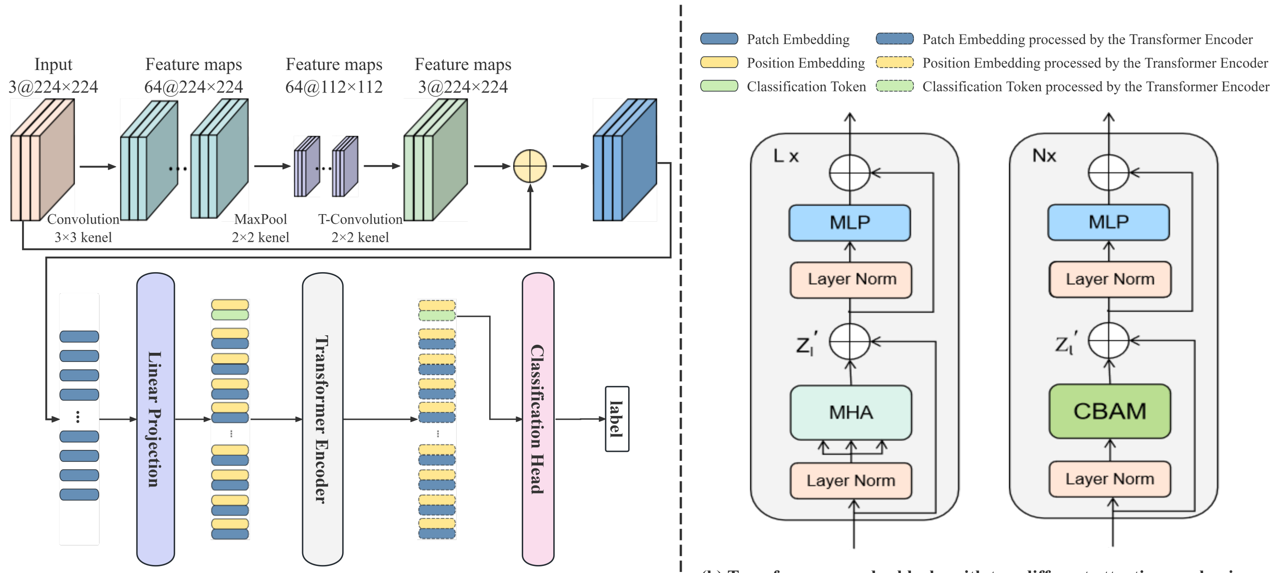
Multi-Head Attention (MHA) has become a powerful tool in the field of deep learning, particularly suited for tasks that require a comprehensive understanding and integration of long-range dependencies within data sequences. Initially demonstrating exceptional performance in natural language processing tasks like machine translation, MHA has since been applied in computer vision, including image classification. MHA works by computing multiple attention heads in parallel, with each head responsible for learning different aspects of relationships and dependencies among the input data. This parallel computation enables the model to capture diverse representations of the data, enhancing its ability to distinguish complex patterns and relationships in intricate datasets. In image classification, MHA enables the model to effectively integrate information across spatial dimensions and capture both local and global dependencies within images. By simultaneously attending to multiple positions at different levels of granularity, MHA helps extract robust features crucial for accurate classification.

2.4. *CBAM mechanism*

The Convolutional Block Attention Module (CBAM) [21] integrates channel attention and spatial attention to enhance neural networks' performance in visual tasks. Channel attention adjusts the importance of each channel, while spatial attention adjusts the importance of each spatial position in the feature maps. By embedding attention mechanisms within each module of convolutional neural networks, CBAM enables dynamic learning and adjustment of both channel and spatial information in feature maps. Channel attention computes importance weights for each channel using global average pooling, which are then learned and used to adjust channel feature maps. Spatial attention employs different convolutional kernels to capture spatial features at various scales and adjusts feature map positions based on learned weights. Widely applied in image processing tasks, especially in image classification [22] and object detection [23], CBAM provides a solid theoretical foundation and practical guidance for its implementation.

3. Proposed method (ViT-DualAtt)

In this section,we introduce an improved pornographic image classification model based on a CNN-Transformer hybrid structure, called ViT-DualAtt. This model combines the strengths of Convolutional Neural Networks and Transformers, incorporating Multi-Head Attention and Convolutional Block Attention Module (CBAM) mechanisms to enhance classification accuracy and robustness. The architecture of the model is shown in Figure 1(b).



(a) Architecture

(b) Transformer encoder blocks with two different attention mechanisms

Figure 1. (a) The architecture of the ViT-DualAtt; and (b) Two transformer encoder blocks with different attention mechanisms are employed. MHA refers to the Multi-Head Attention module, while CBAM denotes the Convolutional Block Attention Module.

3.1. The CNN-Transformer backbone network

3.1.1. Convolutional Neural Network

In the recognition of pornographic images, convolutional neural networks play a crucial role. Their primary function is to extract low-level features such as edges and textures from pornographic images through a series of convolution and pooling operations, thereby capturing fundamental local features. As shown in the upper part of Figure 1(a), the CNN in this model consists of a 3 × 3 convolutional layer, a ReLU activation layer, a max pooling layer, and a 2 × 2 transposed convolutional layer. The feature extraction process is as follows: for an input pornographic image with dimensions of 3 × 224 × 224, the convolution module processes it to produce a feature map of the same size, 3 × 224 × 224. This feature map is then fused with the original image, resulting in an enriched feature map of 3 × 224 × 224, which serves as the input for subsequent steps.

3.1.2. Vision Transformer

The Transformer model was initially designed for processing one-dimensional text data. It has garnered attention for its impressive performance in machine translation and other natural language processing (NLP) tasks [24]. The Vision Transformer [25] represents a novel architecture for visual tasks that builds upon the original Transformer structure. To adapt the Transformer for image processing, researchers have made several modifications and extensions. First, considering that feature maps extracted by convolutional neural networks are two-dimensional, the model needs to transform these feature maps to meet the Transformer's requirements. This is done by dividing the feature map into a series of patches and then applying a trainable linear transformation (also known as an embedding layer) to each patch. Specifically, this transformation is achieved by learning an embedding matrix E, which linearly projects each image patch into a vector space of dimension D. This process not only converts image patches into vector form but also captures the potential relationships between image patches through the learned embedding matrix. For classification tasks, the model requires a special classification token, referred to as v_{class}. This token is combined with the embedded image patches, enabling the model to perform image classification during processing. Since the Transformer model is a sequential model and lacks information about the positions of elements in the input sequence, it is necessary to add positional encodings E_{pos} to each embedded vector. This enables the model to understand the relative positions of the image patches in the original image. The sequence of labeled embedded patches is shown in Eq (3.1):

$$z_0 = \left[v_{\text{class}}\,; x_p^1 E; x_p^2 E; \ldots ; x_p^N E\right] + E_{pos}, E \in R^{\left(P^2 \cdot C\right) \times D}, E_{pos} \in R^{(N+1) \times D}, \tag{3.1}$$

where $x_p^i E$ is the embedding vector of the image block after linear transformation. N is the number of patches generated, P is the size of the patch, C is the number of channels in the original image.

The embedded block sequence z_0 is passed into the Transformer encoder. The encoder consists of two key components: An attention module and a fully connected feed-forward dense block. In the first L layers of the Transformer encoder (as shown on the left side of Figure 1(b)), the model uses the Multi-Head Attention mechanism. The multi-head self-attention mechanism can capture long-range dependencies between different parts of the image. Specifically, it computes multiple attention heads in parallel, with each head learning different aspects of relationships and dependencies within the input data, thus capturing image features on a global scale. This mechanism is particularly suited for capturing global information and long-range dependencies in images. In the subsequent N layers of the Transformer encoder (as shown on the right side of Figure 1(b)), the model introduces the Convolutional Block Attention Module. CBAM enhances feature representation through two modules: channel attention and spatial attention. The channel attention module focuses on which feature maps are more important, while the spatial attention module focuses on which positions in the feature maps are more significant. CBAM effectively captures local features in images, improving the model's sensitivity to details. This design aims to fully utilize both local and global features of the image, thereby enhancing classification accuracy. Each attention mechanism module is followed by a residual connection and layer normalization. Residual connections enable gradients to flow directly back to earlier layers, mitigating the vanishing gradient problem in deep networks and speeding up model convergence. Layer normalization helps stabilize the output of each layer, improving the stability and reliability of model training. After processing through multiple encoder layers, the classification token is extracted and passed to an external classifier to predict the image's class label y. The process of the encoder can be expressed by Eqs (3.2)–(3.6):

$$z_l^{'} = MHSA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1 \ldots L, \tag{3.2}$$

$$z_l = MLP(LN(\mathbf{z}_l^{'})) + \mathbf{z}_l^{'}, \quad l = 1 \ldots L, \tag{3.3}$$

$$z_n^{'} = CBAM(LN(\mathbf{z}_{n-1})) + \mathbf{z}_{n-1}, \quad n = L + 1 \ldots N, \tag{3.4}$$

$$z_n = MLP(LN(\mathbf{z}_n^{'})) + \mathbf{z}_n^{'}, \quad n = L + 1 \ldots N, \tag{3.5}$$

$$y = Softmax(MLP(\mathbf{z}_N^0)). \tag{3.6}$$

3.2. Attention module

Attention is a crucial mechanism that can be utilized in various deep learning models across different domains and tasks [26]. In deep learning, attention mechanisms have become a crucial technology for handling complex data and tasks. By dynamically adjusting the focus of the model on different parts of the input data, they significantly enhance model performance. In this section, we provide a detailed introduction to multi-head attention and convolutional block attention mechanisms.

3.2.1. Multi-Head Attention module

Multi-Head Attention is an important variant of the attention mechanism, building on self-attention by introducing the concept of "multiple heads". This enables the model to focus on different aspects of the input information in parallel. The core idea is to split the input information into multiple subspaces with each head performing its own attention calculation independently. The benefit of this approach is that it enables parallel processing of different pieces of information, thus improving computational efficiency. Since each head has its own set of weight matrices and attention computation process, they can learn different feature representations and dependencies. This diversity of information helps the model capture richer contextual information. The calculation process for multi-head attention is as follows:

Input transformation: For an input sequence X, three learnable weight matrices W_Q , W_K , and W_V are used to generate the Query, Key, and Value matrices, respectively, as shown in Eq (3.7):

$$Q = X \cdot W_Q, K = X \cdot W_K, V = X \cdot W_V. \tag{3.7}$$

Calculating attention weights: For each attention head, compute the dot product of the queries and keys, then apply scaling and the Softmax function to generate the attention weight matrix, as shown in Eq (3.8):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.8}$$

where d_k is the dimension of the keys, used to scale the dot product results.

Concatenate attention heads: The outputs of all attention heads are concatenated and passed through a linear transformation to produce the final attention output, as shown in Eq (3.9):

$$MultiHead(Q, K, V) = Concat\,[head_1, head_2, \ldots, head_n]\, W_O, \tag{3.9}$$

where $head_i = Attention(Q, K, V)$, h is the number of attention heads, and W_O is the linear transformation matrix.

3.2.2. CBAM module

The CBAM module enables the model to dynamically adjust the feature map weights, thereby directing greater focus towards the most crucial feature channels and spatial locations. This results in a notable enhancement in the model's ability to accurately recognize pornographic images, thereby significantly improving the overall recognition accuracy of the pornographic image recognition task. The CBAM module comprises four parts: input features, channel attention module, spatial attention module, and output, as shown in Figure 2. Assuming the dimensions of the input feature map F_{in} are C × H × W, where C represents the number of channels, and H and W represent the height and width of the feature map respectively, in accordance with the established processing sequence of the CBAM module, two principal attention maps can be generated in sequence: A one-dimensional channel attention map M_C and a two-dimensional spatial attention map M_S. The attention process can be represented by Eqs (3.10) and (3.11):

$$F^{'} = M_C(F_{in}) \otimes F_{in}, \tag{3.10}$$

$$F^{''} = M_S(F^{'}) \otimes F^{'}. \tag{3.11}$$

The symbol ⊗ represents element-wise multiplication and $F^{''}$ refers to the final Output Feature.

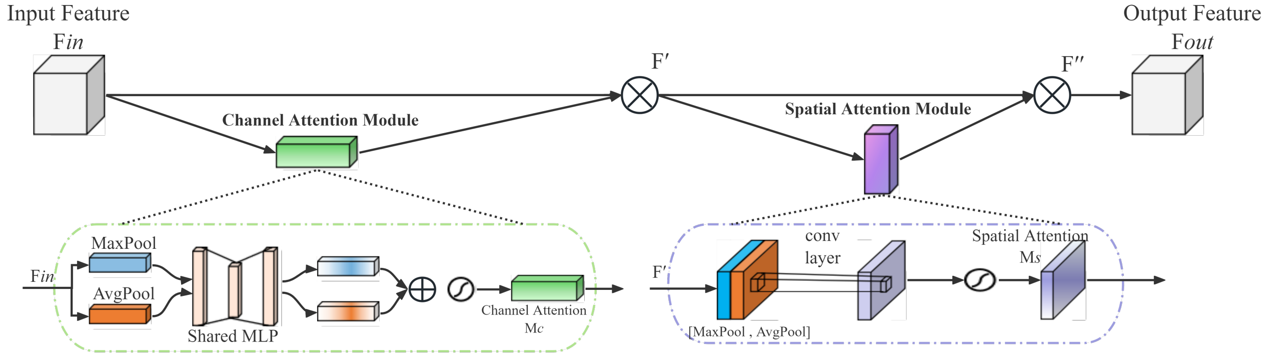


Figure 2. CBAM structure.

After being processed by the channel attention module, the channel dimension of the input feature F_{in} remains unchanged, while the spatial dimension is compressed. This process is achieved through two parallel operations: the maximum pooling layer and the average pooling layer. The input feature

F_{in} is acted upon separately by two pooling layers, resulting in two feature maps of size C × 1 × 1. These feature maps are then inputted into the Shared Multilayer Perceptron (Shared MLP) module. In the Shared MLP module, the number of channels of the feature maps is compressed to $1/R$ times of the original, where R is the reduction rate. This parameter is used to reduce computational complexity and prevent overfitting. The number of channels is then expanded to its original size to match the channel dimension of the original feature map. Prior to the expansion, it typically undergoes a ReLU activation function to increase the model's nonlinearity. Following the Shared MLP process, two feature maps of size C × 1 × 1 are obtained. The final channel attention map M_C is obtained by summing the two feature maps element by element and normalizing them using a Sigmoid activation function. Each element of M_C is a value between 0 and 1, indicating the weight of importance for the corresponding channel. The channel attention mechanism is calculated using the Eq (3.12):

$$M_C(F_{in}) = \sigma\left(MLP(AvgPool(F_{in})) + MLP(MaxPool(F_{in}))\right). \tag{3.12}$$

After being processed by the spatial attention module, the spatial dimension of the input feature $F^{'}$ remains unchanged, while the channel dimension is compressed. The output $F^{'}$ of the channel attention module is passed through both the maximum pooling layer and the average pooling layer to obtain two feature maps of size 1 × H × W. These two feature maps capture the maximum and average response at spatial locations, respectively. This helps the model understand the distribution of spatial information. Subsequently, the 'Concat' operation is employed to fuse these two 1 × H × W feature maps along the channel dimension, resulting in a combined feature map of size 2 × H × W. The splicing operation enables the model to consider both the maximum and average information on spatial locations, resulting in a more comprehensive understanding of spatial attention. Next, a 7 × 7 convolutional layer is applied to the concatenated feature map, producing a single-channel feature map. The convolutional layer merges the spatial information and produces a single-channel feature map with the same dimensions as the original. The resulting convolved feature map is then normalized using a Sigmoid activation function to generate the spatial attention map M_S. Each element of M_S is a value between 0 and 1, indicating the importance weight of the corresponding spatial location. The spatial attention mechanism is calculated using the Eq (3.13):

$$M_S(F^{'}) = \sigma\left(f^{7\times7}([AvgPool(F^{'}); MaxPool(F^{'})])\right). \tag{3.13}$$

4. Experiment and result analysis

4.1. Experimental setup

4.1.1. Experimental data sets

The dataset used in this experiment is the public NSFW_Data_Scrapper project published on GitHub by Alexander Kim, a data scientist. It comprises 67,776 images divided into two categories: Neutral images with 21,917 images (32%) and pornographic images with 45,859 images (68%). Such a ratio ensures that the model is exposed to enough pornographic images during the training process, thereby enabling it to better learn the features for recognizing such images. This, in turn, reduces the probability of misdetection pornographic images in the application of real-world scenarios, and reduces the harm to society. The dataset is divided into a training set, a validation set, and a test set in a 6:2:2 ratio. The training set and validation set data are batched and passed into the model for training after being randomly shuffled. The test data consists of partial datasets from NSFW_Data_Scrapper and NudeNet, comprising a total of 9604 pornographic images and 8209 neutral images. The additional NudeNet data helps evaluate the model's adaptability and generalization on new data, providing a more comprehensive performance assessment.

4.1.2. Training environment and parameter configuration

The experimental platform is provided by a server with the following hardware configuration: CPU: Intel(R) Xeon(R) CPU E5-2680 v4@2.40 GHz, RAM: 256 GB, GPU: NVIDIA TITAN RTX 24 GB, and software configurations of Python 3.9, CUDA 12.0, and Pytorch 1.13.

The experiment utilized the ViT-DualAtt model for training. In the data preprocessing stage, the images were resized to 224 × 224 pixels and standardized before being converted into feature tensors. The model parameters were updated using the SGD optimizer, with a batch size of 32, momentum set to 0.9, and a weight decay coefficient of 5e-5. Regarding the momentum and weight decay parameters, we opted to maintain the original settings of 0.9 for momentum and 5e-5 for weight decay. These parameters are well-established within the context of Vision Transformers and have been shown to effectively support model convergence and generalization. The training process was set for 100 epochs, during which the learning rate was dynamically adjusted using a cosine annealing scheduler to optimize the learning process. The cosine annealing scheduler adjusts the learning rate based on changes in the cosine function, which gradually decays from its initial value to a lower learning rate. This strategy helps achieve fast learning in the early stages of training and fine-tuning in the later stages, improving the final convergence of the model. Detailed training configurations can be found in Table 2.

Table 2. Training parameter settings.

Parameters	Setting
Image size	224 × 224
Epoch	100
Optimizer	SGD
Scheduler	Cosine annealing
Initial learning rate	0.001
Batch size	32
Momentum	0.9
Weight decay	5e-5

4.2. Evaluation criteria

To evaluate the performance of the model, a confusion matrix is used to analyze and compare the differences between the model's predictions and the actual labels. The confusion matrix provides four key metrics: True positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). TP represents the number of pornographic images correctly classified as pornographic by the model, FN represents the number of pornographic images incorrectly classified as neutral by the model, FP represents the number of neutral images incorrectly classified as pornographic by the model, and TN represents the number of neutral images correctly classified as neutral by the model. By combining these four metrics, commonly used evaluation measures can be derived, such as:

Equation (4.1) shows the accuracy of the model, which is the proportion of correctly classified samples to the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.1}$$

Equation (4.2) shows the precision rate, which is the proportion of instances where the model predicts a positive sample that are truly positive.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Equation (4.3) shows the recall, which is the proportion of correctly predicted positive samples of all actual positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Equation (4.4) shows the reconciled average of precision and recall, which is used to comprehensively evaluate the performance of the model, i.e., $F1 - Score$.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

We use the False Negative Rate (*FNR*), which we refer to as the Pornographic Image Miss Rate (*PIMR*) in the context of our study, to evaluate the model's performance in identifying pornographic images. Equation (4.5) shows the calculation of *PIMR*.

$$PIMR = \frac{FN}{TP + FN} \tag{4.5}$$

The *PIMR* represents the percentage of pornographic images that the model fails to correctly identify, out of all actual pornographic images. This is a crucial indicator to assess, given the potential harm that pornographic content can have on society, particularly on minors. A low *PIMR* model accurately captures most pornographic images, reducing the risk of underreporting.

4.3. Experimental result analysis

4.3.1. Training process analysis

During training, model performance was evaluated by tracking accuracy and loss values at each epoch. Figure 3 shows the trend of accuracy and loss of the ViT-DualAtt model over 10 training sessions. In particular, Figure 3(a) shows the changes in accuracy over training epochs, while Figure 3(b) displays the loss trends. In terms of accuracy, all training runs show significant improvement during the first 25 epochs, after which accuracy stabilizes, reaching high levels around 0.97 in the later stages of training. This consistency across runs highlights the model's stability. For loss, the plots reveal a steady decline as training progresses. Initially, the loss drops quickly from around 0.4 to below 0.15, then gradually stabilizes. In the later stages, the loss further decreases and levels off near 0.07, indicating that the model's error has been minimized. Overall, as the number of epochs increases, the model consistently shows a convergent trend across all training runs, with accuracy improving and loss decreasing. These results confirm the model's effectiveness and stability in the classification task.

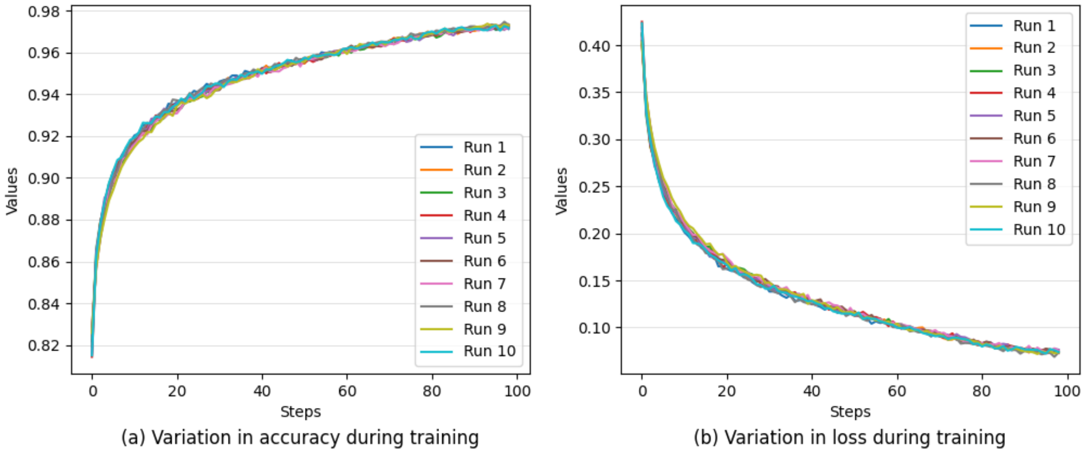


Figure 3. (a) shows the variation in accuracy during training; (b) shows the variation in loss during training.

4.3.2. Comparison of existing algorithms

We conducted comparative experiments between the method proposed in this paper and several existing methods in the literature, with the results shown in Table 3. Analyzing the classification accuracy in the table, we observe that our method performs excellently on the experimental dataset, achieving a 2.7% improvement over current deep learning methods [17–20]. Moreover, this enhancement is more pronounced when compared to traditional machine learning methods. Our approach outperforms the skin area-based classification method [13, 14] by 22.2%, and the shape and texture-based classification method [15, 16] by 4.5%. These results demonstrate that our model exhibits stronger generalization ability and effectiveness in pornographic image classification tasks, providing new insights and methods for future image classification research.

Table 3. Comparison of existing algorithms.

Papers	Methods	Accuracy (%)
[13]	Skin	61.0
[14]	Skin + SVM	75.0
[15]	BoVW	87.6
[16]	BoVW + SVM	84.9
[17]	CNN	86.9
[18]	DCNN	92.7
[19]	ResNet101 + CBAM	93.2
[20]	RepVGG-SimAM	94.5
Ours	ViT-DualAtt	97.2

4.3.3. Comparison of different classification algorithms

To ascertain the efficacy of the models proposed in this paper for the classification of pornographic images, a series of comparative experiments were conducted. The major comparison objects were ResNet101 [27], VGGNet16 [28], EfficientNet_b0 [29], DenseNet201 [30] and Swin Transformer [31]. These algorithms have a wide range of applications and excellent performance in image classification, making them highly valuable for comparison. To ensure the fairness and accuracy of the results, the experiments employed the same dataset, preprocessing steps, and assessment metrics. The comparison results are presented in Table 4. The comparison experiment shows that the model proposed in this paper outperforms other models in terms of accuracy and *PIMR* in the field of pornographic image classification.

Table 4. Classification results of each model.

Methods	Precision (%)		Recall (%)		F1 (%)		Accuracy (%)	PIMR (%)
	Porn	Neutral	Porn	Neutral	Porn	Neutral		
ResNet101 [27]	92.4	89.8	95.4	83.6	93.9	86.6	91.6	4.5
VGGNet16 [28]	92.8	92.1	96.5	84.3	94.6	88.0	92.6	3.4
EfficientNet_b0 [29]	92.3	92.5	96.7	83.2	94.5	87.6	92.3	3.2
DenseNet201 [30]	95.3	93.9	97.1	90.2	96.2	92.0	94.9	2.8
Swin Transformer [31]	98.4	92.5	96.5	96.5	97.4	94.4	96.5	3.5
ViT-DualAtt	97.4	96.5	98.3	94.4	97.8	95.4	97.2	1.6

To further evaluate the classification performance of each model, we computed their receiver operating characteristic (ROC) curves and the corresponding area under curve (AUC) values, as shown in the Figure 4. An increase in AUC value indicates that the model can maintain a high true positive rate and a low false positive rate across thresholds, making it more effective in identifying and classifying pornographic images in practical applications, thereby reducing the risk of misclassification. The AUC value for the ViT-DualAtt model is 0.951, significantly higher than that of the other comparison models. Specifically, ResNet101 has an AUC of 0.892, VGGNet16 has an AUC of 0.882, EfficientNet_b0 has an AUC of 0.891, DenseNet201 has an AUC of 0.913 and Swin Transformer has an AUC of 0.944. These results not only further validate the outstanding performance of the proposed model in the task of pornographic image classification but also demonstrate its strong ability to differentiate between positive and negative samples.

To mitigate the harm caused by the spread of inappropriate images, inference speed is also a crucial evaluation metric. To assess the inference speed of our method, we conducted a comparison with other aforementioned models. Figure 5 illustrates the time required by each model to process varying numbers of images. As the number of input images increases, the processing time for all models exhibits a linear growth trend. ViT-DualAtt demonstrates higher efficiency than ResNet101 and Swin Transformer under all conditions, with consistently lower processing times. However, compared to EfficientNet_b0, the processing time of ViT-DualAtt is slightly higher, while DenseNet201 and VGGNet16 fall between these two models.

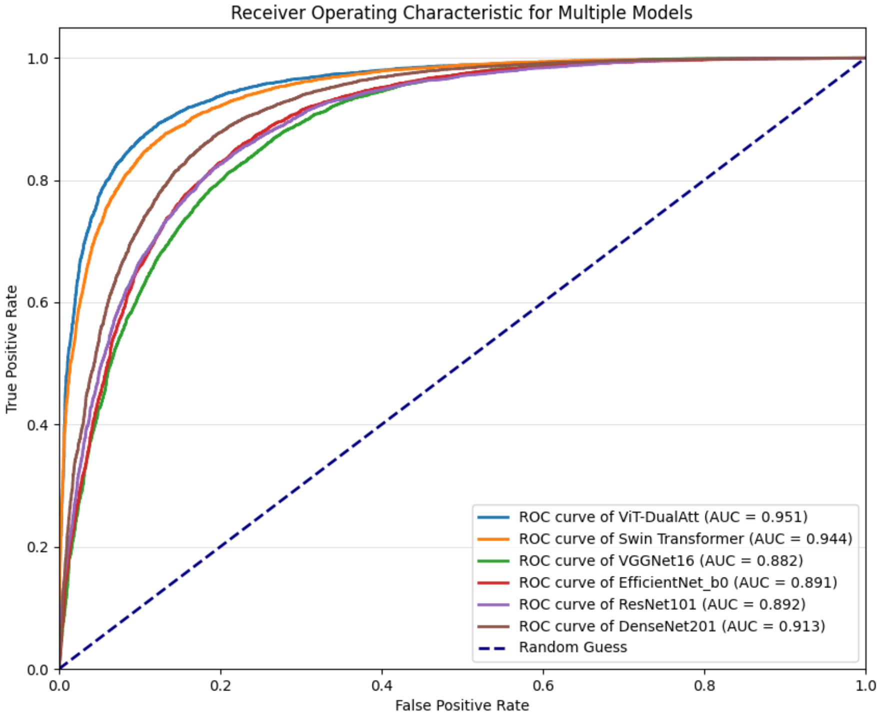


Figure 4. Receiver operating characteristic for multiple models.

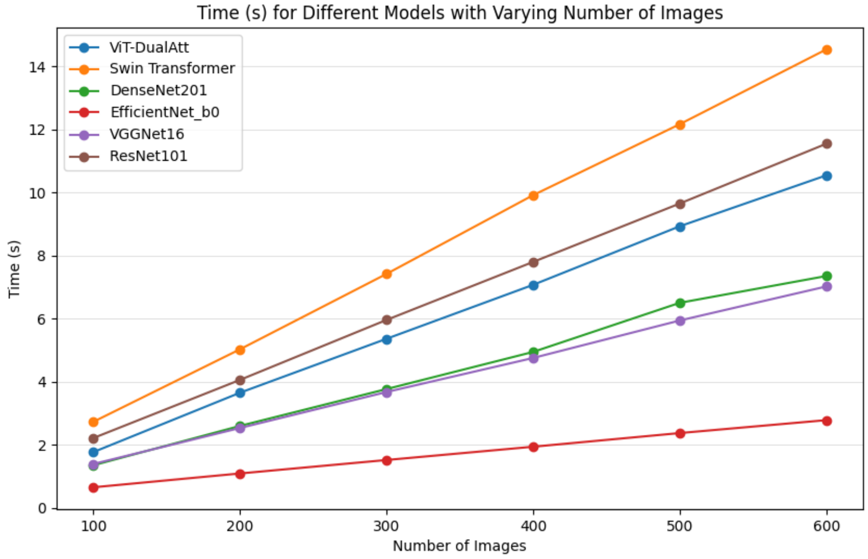


Figure 5. Time(s) for different models with varying number of images.

In summary, although ViT-DualAtt's efficiency is slightly lower than that of models like EfficientNet_b0, its processing time remains within an acceptable range and is accompanied by notable accuracy advantages. Thus, ViT-DualAtt remains a valuable model choice, balancing accuracy and efficiency for practical application.

4.3.4. Ablation comparison experiments

To assess the effectiveness of the ViT-DualAtt model, we conducted ablation experiments by removing different components and analyzing their impact on classification performance. The experimental results are summarized in Table 5. Initially, the original ViT model included only the Multi-Head Attention mechanism, achieving an accuracy of 94.7% with a PIMR of 2.5%. Introducing the ViT-DualAtt model with CNN and MHA (without CBAM) increased accuracy to 96.0% and reduced PIMR to 2.4%, indicating CNN's enhancement in local feature extraction. The ViT-DualAtt model with CBAM and MHA (without CNN) achieved an accuracy of 95.8% and PIMR of 2.0%, highlighting CBAM's significant contribution to local feature attention. When CBAM and CNN were combined but MHA was removed, accuracy improved to 97.0% with PIMR reduced to 1.8%, demonstrating the advantage of CBAM and CNN in capturing both global and local features. The complete ViT-DualAtt model, integrating CNN-Transformer structure, Multi-Head Attention mechanism, and Convolutional Block Attention Module, performed the best with an accuracy of 97.1% and PIMR of 1.6%. These results demonstrate the superiority and robustness of the ViT-DualAtt model in classification tasks, especially in terms of precision and recall, validating its effectiveness and stability in the task of classifying pornographic images.

Table 5. Comparison of ablation experiments.

Methods	CNN	MHA	CBAM	Precision (%)		Recall (%)		F1 (%)		Accuracy (%)	PIMR (%)
				Porn	Neutral	Porn	Neutral	Porn	Neutral		
ViT		✓		92.4	89.8	95.4	83.6	93.9	86.6	91.6	4.5
ViT-DualAtt without CBAM	✓	✓		92.8	92.1	96.5	84.3	94.6	88.0	92.6	3.4
ViT-DualAtt without CNN		✓	✓	92.3	92.5	96.7	83.2	94.5	87.6	92.3	3.2
ViT-DualAtt without MHA	✓		✓	95.3	93.9	97.1	90.2	96.2	92.0	94.9	2.8
ViT-DualAtt	✓	✓	✓	97.4	96.5	98.3	94.4	97.8	95.4	97.2	1.6

In order to verify the specific contributions of MHA and CBAM, we performed visualization operations on the features obtained after model processing. The results are shown in Figure 6. In terms of visual representation, MHA shows a unique role. It enables the model to focus on multiple key areas in the image, presenting a relatively obvious feature aggregation effect. However, while MHA pays attention to key areas, it also gives more attention to the background area. To some extent, this disperses the model's attention resources and has a certain interference on the classification task. On the other hand, CBAM not only strengthens the attention to the relevant areas of the main target, but also filters out more discriminative feature channels through its channel attention mechanism. Moreover, it accurately locates the key local spatial positions of the target with the help of spatial attention, making the model's extraction of target features more refined. The two complement each other and jointly contribute to improving the classification performance of the model.

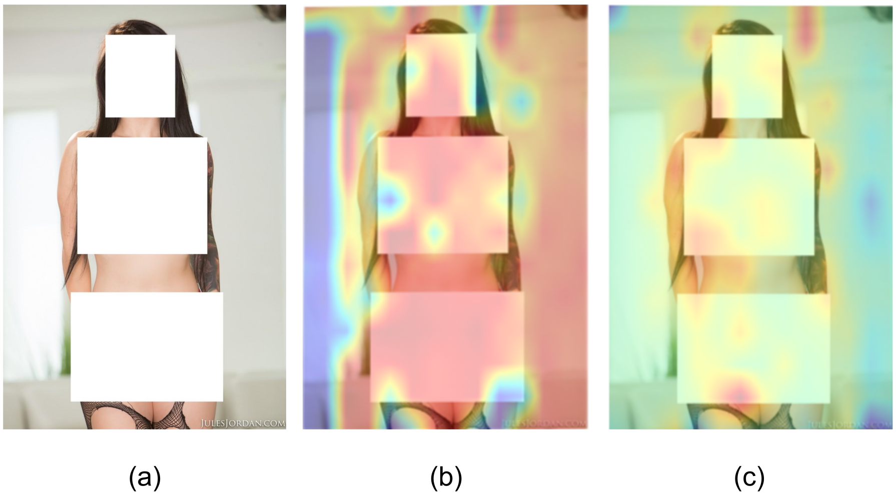


Figure 6. (a) shows the original image; (b) shows the feature visualization diagram after processing by ViT (12-layer MHA); and (c) shows the feature visualization diagram after processing by ViT-DualAtt (6-layer MHA + 6-layer CBAM).

5. Conclusions

We present a novel approach to pornographic image classification: ViT-DualAtt. This model integrates CNN and Transformer architectures within a unified framework and incorporates both multi-head attention and convolutional block attention mechanisms. This integration significantly enhances classification accuracy, achieving a 2.7% improvement over other methods and reaching an accuracy of 97.2% ± 0.1%, underscoring its superior performance and effectiveness in the field of pornographic image classification.

Although ViT-DualAtt demonstrates strong classification accuracy, there remains potential for further optimization of its processing efficiency. Future work will focus on refining the model's architecture by exploring more lightweight attention mechanisms and optimizing the computational workflow to reduce complexity and enhance processing speed. Specifically, techniques such as sparse attention, low-rank decomposition, and knowledge distillation could be utilized to reduce the model's parameter load and computational demands. Additionally, research will investigate the model's adaptability across hardware platforms, including mobile and embedded systems, to support large-scale data processing and real-time applications, ensuring that the model performs efficiently and accurately across diverse deployment scenarios.

While pornographic image classification systems are instrumental in safeguarding user safety and maintaining a healthy online environment, they also bring ethical considerations. One prominent concern is potential bias, as imbalances in training data distribution may lead to suboptimal model performance across certain demographic or cultural groups, resulting in unfair classification outcomes. Furthermore, given the sensitive nature of the content involved, it is crucial that such systems operate with stringent privacy protections to prevent the misuse or leakage of personal data. To address these concerns, future research should emphasize fairness and transparency, exploring bias mitigation techniques and strengthening privacy safeguards in data storage and processing to ensure the system's ethical compliance.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant Number 62072416), the Key Research and Development Special Project of Henan Province (Grant Number 221111210500), and the Key Technologies R&D Program of Henan Province (Grant Numbers 232102211053, 242102211071 and 242102210142).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. Z. Wang, R. Guo, H. Wang, X. Zhang, A new model for small target adult image recognition, *Procedia Comput. Sci.*, **183** (2021), 557–562. https://doi.org/10.1016/j.procs.2021.02.097
2. B. Wang, X. Lv, X. Ma, H. Wang, Application of skin detection based on irregular polygon area boundary constraint on YCbCr and reverse gamma correction, *Adv. Mater. Res.*, **327** (2011), 31–36. https://doi.org/10.4028/www.scientific.net/AMR.327.31
3. Z. Zhao, A. Cai, Combining multiple SVM classifiers for adult image recognition, in *2010 2nd IEEE International Conference on Network Infrastructure and Digital Content*, IEEE, (2010), 149–153. https://doi.org/10.1109/ICNIDC.2010.5657916
4. S. Paheding, A. Saleem, M. F. H. Siddiqui, N. Rawashdeh, A. Essa, A. A. Reyes, Advancing horizons in remote sensing: A comprehensive survey of deep learning models and applications in image classification and beyond, *Neural Comput. Appl.*, **36** (2024), 16727–16767. https://doi.org/10.1007/s00521-024-10165-7
5. C. Zhao, R. W. Liu, J. Qu, R. Gao, Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons, *Eng. Appl. Artif. Intell.*, **128** (2024), 107513. https://doi.org/10.1016/j.engappai.2023.107513
6. R. Shetty, V. S. Bhat, J. Pujari, Content-based medical image retrieval using deep learning-based features and hybrid meta-heuristic optimization, *Biomed. Signal Process. Control*, **92** (2024), 106069. https://doi.org/10.1016/j.bspc.2024.106069
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., Attention is all you need, preprint, arXiv:1706.03762.
8. W. Zhang, G. Chen, P. Zhuang, W. Zhao, L. Zhou, CATNet: Cascaded attention transformer network for marine species image classification, *Expert Syst. Appl.*, **256** (2024), 124932. https://doi.org/10.1016/j.eswa.2024.124932

9. M. Ahmad, U. Ghous, M. Usama, M. Mazzara, WaveFormer: Spectral–spatial wavelet transformer for hyperspectral image classification, *IEEE Geosci. Remote Sens. Lett.*, **21** (2024), 1–5. https://doi.org/10.1109/LGRS.2024.3353909

10. G. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, **42** (2011), 513–529. https://doi.org/10.1109/TSMCB.2011.2168604

11. G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Springer, (2003), 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

12. H. Zhao, I. Liu, Research on test data generation method of complex event big data processing system based on Bayesian network, *Comput. Appl. Res.*, **35** (2018), 155–158.

13. M. Jones, J. Rehg, Statistical color models with application to skin detection, *Int. J. Comput. Vision*, **46** (2002), 81–96. https://doi.org/10.1023/A:1013200319198

14. Y. Lin, H. Tseng, C. Fuh, Pornography detection using support vector machine, in *16th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP 2003)*, (2003), 123–130.

15. L. Lv, C. Zhao, H. Lv, J. Shang, Y. Yang, J. Wang, Pornographic images detection using high-level semantic features, in *2011 Seventh International Conference on Natural Computation*, IEEE, (2011), 1015–1018. https://doi.org/10.1109/ICNC.2011.6022151

16. K. Dong, L. Guo, Q. Fu, An adult image detection algorithm based on Bag-of-Visual-Words and text information, in *2014 10th International Conference on Natural Computation*, IEEE, (2014), 556–560. https://doi.org/10.1109/ICNC.2014.6975895

17. L. Lv, C. Zhao, H. Lv, J. Shang, Y. Yang, J. Wang, Pornographic images detection using high-level semantic features, in *2011 Seventh International Conference on Natural Computation*, IEEE, (2011), 1015–1018. https://doi.org/10.1109/ICNC.2011.6022151

18. F. Cheng, S. Wang, X. Wang, A. Liew, G. Liu, A global and local context integration DCNN for adult image classification, *Pattern Recognit.*, **96** (2019), 106983. https://doi.org/10.1016/j.patcog.2019.106983

19. Z. Cai, X. Hu, Z. Geng, J. Zhang, Y. Feng, An illegal image classification system based on deep residual network and convolutional block attention module, *Int. J. Network Secur.*, **25** (2023), 351–359. https://doi.org/10.6633/IJNS.202303_25(2).18

20. Z. Cai, X. Qiao, J. Zhang, Y. Feng, X. Hu, N. Jiang, Repvgg-simam: An efficient bad image classification method based on RepVGG with simple parameter-free attention module, *Appl. Sci.*, **13** (2023), 11925. https://doi.org/10.3390/app132111925

21. S. Woo, J. Park, J. Lee, I. Kweon, CBAM: Convolutional Block Attention Module, in *Computer Vision – ECCV 2018*, Springer, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

22. S. Yu, S. Jin, J. Peng, H. Liu, Y. He, Application of a new deep learning method with CBAM in clothing image classification, in *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, IEEE, (2021), 364–368. https://doi.org/10.1109/ICESIT53460.2021.9696783

23. J. Liu, H. Qiao, L. Yang, J. Guo, Improved lightweight YOLOv4 foreign object detection method for conveyor belts combined with CBAM, *Appl. Sci.*, **13** (2023), 8465. https://doi.org/10.3390/app13148465

24. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, preprint, arXiv:1810.04805.

25. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.

26. G. Brauwers, F. Frasincar, A general survey on attention mechanisms in deep learning, *IEEE Trans. Knowl. Data Eng.*, **35** (2021), 3279–3298. https://doi.org/10.1109/TKDE.2021.3126456

27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

28. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

29. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, preprint, arXiv:1905.11946.

30. G. Huang, Z. Liu, L. Van Der Maaten, K. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243

31. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

AIMS Press