IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2397-2407
Journal homepage: http://iieta.org/journals/ts

Application of Multi-Modal Neural Networks in Verifying the Authenticity of News Text and Images

Feng Li[1,2], Meiling Xu[1,2], Marshima Mohd Rosli[1*]

[1] College of Computing, Informatics and Mathematics Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[2] College of Computer and Information Engineering, Hebei Finance University, Baoding 071051, China

Corresponding Author Email: marshima@fskm.uitm.edu.my

https://doi.org/10.18280/ts.400606

Received: 3 August 2023
Revised: 26 October 2023
Accepted: 7 November 2023
Available online: 30 December 2023

Keywords:
multi-modal neural networks, news authenticity verification, anomalous image block search, gated cooperative attention, tampered and forged images, modality fusion, information security

ABSTRACT

In the digital information era, the authenticity of news media is crucial for shaping public opinion and maintaining social stability. The verification of authenticity, especially in news content rich in both text and images, has become a significant task in safeguarding online information security. Traditional methods, often relying on single-modality analysis, are ineffectual against the complex interplay and manipulation techniques possible between news text and accompanying images. Addressing this, research in multi-modal neural networks has emerged, aiming to enhance verification effectiveness by integrating information from both text and images. However, limitations exist in existing research regarding key issue resolution and modality fusion strategies, including inadequate exploration in searching for anomalously similar image blocks, reducing false alarms, and post-processing of tampered images. This study systematically investigates these challenges in news text and image authenticity verification, proposing a novel multi-modal fusion method. This method encompasses three components: a joint gated co-attention mechanism, a filtering gate mechanism, and a joint cooperative representation. These effectively combine text and image information, enhancing the model's ability to discern complex manipulations. The findings of this study not only advance theoretical research in multi-modal fusion but also provide a powerful tool for news media to verify authenticity, significantly contributing to the fight against fake news and maintaining the integrity of information dissemination.

1. INTRODUCTION

With the advent of the new media era, the speed and scale of information dissemination have accelerated and expanded unprecedentedly. Network news, characterized by its immediacy and wide reach, has garnered immense public attention [1-4]. However, the issue of news authenticity, particularly in news texts accompanied by images, has become increasingly prominent [5-7]. The intuitive visual nature of images makes them one of the most influential elements in information dissemination, but this has also led to a rise in the manipulation and forgery of images to mislead readers [8-10]. Therefore, investigating the authenticity verification of news text and images is an urgent priority.

In this context, the significance of this study is particularly notable. Authenticity verification is not only linked to the interests of information consumers, ensuring the healthy development of the public opinion field, but also plays a positive role in preventing and combating the spread of fake news, thus maintaining social stability and harmony [11, 12]. Traditional methods of verification often rely on single-modality analysis, whereas the combination of news text and images conveys richer information and more complex signals of authenticity. Hence, research on the application of multi-modal neural networks in the authenticity verification of news text and images holds substantial theoretical and practical significance [13-16].

However, current research methods still have many flaws and deficiencies [17, 18]. Traditional single-modality authenticity verification methods for text or images often overlook the intrinsic correlation between text and images. Existing multi-modal studies focus primarily on modality fusion strategies and lack adequate research on key issues such as the search for anomalously similar image blocks, the elimination of false alarms, and post-processing operations for tampered images. These limitations restrict further improvements in the efficiency and accuracy of authenticity verification [19-21].

To address these issues, this paper first explores the key problem processing in news text and image authenticity verification, including the search for anomalously similar patch blocks and post-processing techniques for tampered and forged images. Building on this, the paper proposes a multi-modal fusion method tailored for news text and image authenticity verification. This method employs a joint gated co-attention mechanism, a filtering gate mechanism, and a joint cooperative representation. Not only does it innovate multi-modal fusion theory, but it also significantly enhances

the accuracy and robustness of news text and image authenticity verification through the effective integration of various sub-modules. The research presented in this paper offers a more effective tool for authenticity verification in news media and provides solid technical support for information consumers in discerning truth from falsehood, holding significant research value and broad application prospects.

2. KEY ISSUE PROCESSING IN NEWS TEXT AND IMAGE AUTHENTICITY VERIFICATION

During the process of verifying the authenticity of news text and images, the verification of the image part is particularly crucial, as images can be easily manipulated and often have a more direct impact on readers. One of the core challenges in image verification is accurately identifying tampered areas. This process involves searching for anomalously similar patch blocks and post-processing operations of tampered and forged images, among other aspects. Specifically, due to the high-dimensionality of image data, a large amount of pixel information must be processed and analyzed when searching for anomalous blocks. Handling this high-dimensional data requires complex algorithms and substantial computational resources, thereby impacting efficiency. Simultaneously, if an algorithm is well-trained on specific types of images, it may not generalize to new or unknown types of image manipulations, leading to a high false alarm rate. Manipulators might finely post-process the tampered image to eliminate or blur traces of tampering. For instance, re-compressing the image to hide editing traces makes detection of tampering more challenging. These factors collectively pose severe challenges to the efficiency and accuracy of image manipulation detection in news text and image authenticity verification. Figure 1 presents a flowchart of the key factors in news text and image authenticity verification.

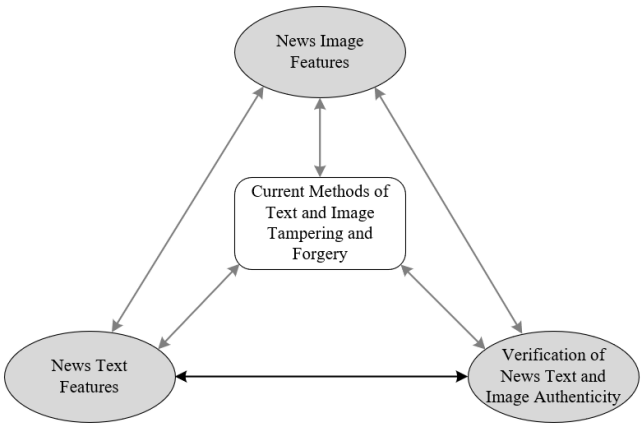


Figure 1. Flowchart of news text and image authenticity verification

In response to these issues, this study first employs a convolutional neural network (*CNN*)-based nearest neighbor image patch block search algorithm for the detection of anomalously similar patch blocks. The rationale and advantage of this algorithm lie in the utilization of *CNN*'s excellent feature extraction capabilities and the efficiency of nearest neighbor matching. *CNN*s can automatically learn deep, robust feature representations from images, capturing the essential attributes of image content rather than merely relying on surface pixel values, thereby exhibiting strength in resisting image noise and handling image deformations. When random reference patch blocks are extracted and their deep features are identified, the nearest neighbor search method can swiftly and effectively find target patch blocks with the most similar features in the entire dataset. Moreover, using nearest neighbor search facilitates the processing and identification of diverse and complex image manipulation methods, as it can recognize tampered areas that may not be visually obvious but exhibit significant differences in feature representation.

The proposed algorithm initially assigns labels to selected reference patch blocks in the input image. Since the reference patch blocks are randomly extracted from the training set, and each block is assigned a label representing its category or attributes, such as whether it has been tampered with, these patch blocks and their labels are used as the dataset for training the *CNN*. During the training process, the *CNN* learns how to classify these patch blocks based on image content through multi-level feature extraction, thereby establishing a rich representation for each patch block in the feature space. Suppose any patch block in the news image is represented by Φ_o, and the selected reference patch blocks in the image are represented by Φ^-_o, the following formula defines the label:

The overall concept of the algorithm: Firstly, a label is assigned to each selected reference patch block in the input image. The label for the patch block plane is defined by Eq. (1):

$$X_{\Phi_o} = \begin{cases} 1, \Phi_o = \overline{\Phi_o} \\ 0, \Phi_o \neq \overline{\Phi_o} \end{cases} \tag{1}$$

The hidden layers of the *CNN* are specifically designed to extract deep features of images, capturing the essential content of image patch blocks, including texture, shape, edges, and colors. Utilizing these deep features, the *CNN* can calculate the similarity between two image patch blocks. In practice, when a target patch block requires verification, the algorithm assesses its similarity to each reference patch block in the training set using the feature representations from the trained *CNN* model. Suppose the pixel intensity of the target patch block is represented by $d(z)$, and that of the reference patch block by $d(z_0)$. Pixel similarity is measured using Euclidean distance, calculated as follows:

$$R_v\left(d(z), d(z_0)\right) = \left\| d(z) - d(z_0) \right\|_2^2 \tag{2}$$

Using the features extracted by the *CNN* and the calculated similarity, the algorithm determines the reference patch block most similar to the target patch block. The label value of this reference block then becomes a key basis for judging the attributes of the target patch block. If the target patch block has the highest similarity with a reference patch block labeled as "tampered", it can be inferred that the target patch block is likely tampered as well. The specific steps for reference patch block propagation update are as follows:

(1) The nearest neighbor patch block function specifies how to measure the similarity between two patch blocks. It defines how to compute the distance in the feature space to determine which reference patch block is closest to the given target patch block. This function typically considers features such as the intensity or color values of the pixels in the image patch block, texture information, and edge directions. Suppose the nearest

neighbor patch block function is represented by d: $S \rightarrow E^2$, where S represents the reference patch block and E represents the possible offsets between nearest neighbor patch blocks.

(2) Once the nearest neighbor patch block function is defined, it can be used to perform nearest neighbor searches. This process typically involves finding the reference patch block most similar to the target patch block in a predefined training set library of patch blocks. The search can be accelerated using approximate nearest neighbor search methods. After determining the nearest neighbor, its correspondence with the target patch block is checked to judge whether the target patch block might have been tampered with.

Selecting the reference patch block centered around the pixel $o(z, t)$, assume the center of the j-th neighboring patch block is represented by (u, k), and let the propagation $w=(z+E,t+E) \leftrightarrow a(u,k,j)$ begin from the neighboring patch block. Further checks are made to see if the current Euclidean distance stored in o and the Euclidean distance between w and $a=(u-u,k-1,j)$ satisfy a greater-than relationship. If so, the reference patch block is updated and the iteration of the above steps is repeated.

(3) In the *CNN*, the activation tensors of the hidden layers can capture complex features of the image content, containing more information than the original pixel values. These activation tensors, as image descriptors, are used to extract representations for each patch block. Such representations include not only the visual content of the image but also its higher-level abstract information, which is crucial for detecting more refined and complex image manipulations. As the target patch block passes through the network, its activation tensor is compared with that of the reference patch blocks. In this way, the network learns which features are key to distinguishing between normal and tampered patch blocks, thus updating the reference patch blocks accordingly.

In news text and image authenticity verification, tamperers often use *JPEG* compression to conceal traces of image manipulation, as *JPEG* compression can reduce high-frequency information in the image through re-encoding, thereby blurring tampering traces and making them harder to detect. This study implements a prediction error filter constraint in the post-processing feature filtering module during network training. The aim is to enable the network to identify and suppress feature distortions introduced by *JPEG* compression while learning tampering trace features. The prediction error filter constraint guides the network to focus on those features that remain stable even after *JPEG* compression. These features are more likely to be the actual characteristics related to tampering, rather than artifacts produced during compression. Such network design enhances the system's ability to detect tampering activities, particularly when tampered images are re-compressed to hide traces, effectively resisting interference from compression noise and improving the reliability and efficiency of news text and image authenticity verification.

In images compressed using *JPEG*, prediction errors change significantly due to block effects, and the prediction error filter can identify compression traces by learning these changes. The post-processing feature filtering module in this study is designed with two layers. The first layer, the *JPEG* compression feature filtering layer, specifically addresses feature distortions caused by compression. By simulating the process of *JPEG* compression, the network learns to ignore features of non-tampered areas caused by compression. The second layer, the Gaussian noise filtering layer, further refines processing by filtering features based on noise patterns. The advantage of this layered design approach lies in its targeted manner of dealing with various common post-processing methods in tampered images. Through this approach, the network not only learns tampering features but also adaptively identifies and suppresses compression and noise features that might conceal or simulate tampering traces. Suppose the filter weight at the location (z,t) in the proposed window is represented by $\mu_j(z,t)$, and the filter weight at the center (0,0) of the proposed window is represented by $\mu_j(0,0)$, the following formula defines the learning rule for all J filters in the first layer:

$$\begin{cases} \mu_j(0,0) = 0 \\ \sum_{z,t \neq 0} \mu_j(z,t) = 1 \end{cases} \tag{3}$$

In the context of news text and image authenticity verification, tampered images are often post-processed by adding Gaussian noise to mask tampering traces. This noise can obscure key tampering features, presenting significant challenges for detection. To enhance the robustness of tampering detection, this study designs a Gaussian noise feature filtering layer, aimed specifically at identifying and weakening changes introduced by Gaussian noise, thereby preserving and strengthening the actual tampering trace features in the image. The advantage of this method is its effectiveness in distinguishing real tampering traces from pseudo-traces caused by Gaussian noise. This is crucial for reducing false positives, especially when dealing with images disturbed by complex noise patterns. The Gaussian noise feature filtering layer significantly improves the performance of the detection algorithm under various noise level conditions, enhancing the credibility and accuracy of the detection results. This ensures that even under adverse conditions with noise presence, the news text and image authenticity verification system can function effectively, providing users with more reliable image authenticity analysis results.

To combat the interference caused by attempts to mask tampering traces or simulate them as shooting noise, the proposed Gaussian noise filtering algorithm is implemented by adding an additional penalty term to the loss function. This penalty term is the L2-norm of the loss gradient of the noiseless image, with the core idea being to reinforce the model's smoothness of local neighborhood loss changes caused by noise during the training process. Designing the loss function in this way provides direct mathematical guidance to stabilize the loss gradient, reducing unnecessary variations caused by noise. Such a smoothing constraint helps the model better distinguish between real tampering traces and pseudo-traces caused by noise in the image, improving the precision of tampered area detection. Suppose the weighting coefficient is represented by η, and the binary weighted cross-entropy is represented by $R(z,t)$, the designed loss function expression is given by:

$$M(z,t) = R(z,t) + \eta \left\| \nabla_z R(z,t) \right\|_2 \tag{4}$$

The binary weighted cross-entropy definition is:

$$R(z,t) = -\sum_{u=1}^{2} \log \left(\frac{r^{\beta_{u,k} x_u}}{\sum_{k=1}^{2} r^{\beta_{u,k} x_k} + \sum_{j=1}^{2} (1-r)^{\alpha_{u,k}(1-x)_k}} \right) \tag{5}$$

Suppose the output of the filtering layer is represented by x, and the weights in the z and t directions are represented by $\beta_{u,k}$ and $\alpha_{u,k}$, then the formula is:

$$\beta_{u,k} + \alpha_{u,k} = 1, \beta_{u,k} \geq 0, \alpha_{u,k} \geq 0 \tag{6}$$

In news text and image authenticity verification, training the Gaussian noise filtering layer is a key step, ensuring the model can effectively distinguish between real tampering traces and pseudo-traces caused by noise. The specific steps are detailed below:

(1) First, construct two training sets: one containing news images artificially added with Gaussian noise, represented by $U_{H\text{-}NO}$, to simulate tampering situations encountered in reality, and another containing noiseless news images, represented by U_{CL}. These two sets are used to train the model so it can identify tampering traces under both conditions.

(2) During the training process, the loss functions for noisy and noiseless images are summed. This balances the model's learning for both types of images, ensuring it does not become biased toward one type and can adapt to different noise environments. Suppose the batch size is represented by B, with B images extracted per batch, the first J images being Gaussian noise images $U_{H\text{-}NO}$, i.e., $\{U^{m}{}_{H\text{-}NO}, U^{2}{}_{H\text{-}NO}, ..., U^{J}{}_{H\text{-}NO}, U^{J+1}{}_{H\text{-}NO}, U^{J+2}{}_{H\text{-}NO}, ..., U^{B}{}_{H\text{-}NO}\}$.

(3) In each training batch, calculate the difference between the model's output and the true labels, i.e., the loss. This loss reflects the current performance of the model and is used in subsequent gradient descent steps. Suppose the weight parameter is represented by β, and the losses for $U_{H\text{-}NO}$ and U_{CL} are summed, further calculating the loss for the batch of B images based on the following formula:

$$\begin{aligned} LOSS(u) &= \sum_{u \in U_{CL}} M(z_u, t_u) \\ &+ \beta \sum_{u \in U_{H-NO}} M(z_u, t_u) \end{aligned} \tag{7}$$

(4) Assign a category label to each image in the training set, indicating whether the image has been tampered with. These labels are vital for supervised learning, as they serve as the "truth" benchmark during training. The real labels for $U_{H\text{-}NO}$ are represented by $M_{Hb\text{-}TR}$, and for U_{CL} by $M_{v\text{-}TR}$; the predicted labels for $U_{H\text{-}NO}$ are represented by $M_{Hb\text{-}TR}$, and for U_{CL} by $M_{v\text{-}PR}$.

(5) Utilize the method of maximum likelihood estimation to determine the direction of model parameter updates by calculating the partial derivatives of the loss function with respect to the model parameters. This process aims to find the point in the parameter space that minimizes the loss function. By gradient descent or other optimization algorithms, update the model parameters in the direction of the calculated gradient, aiming to reduce the total loss for the batch of news images. As training progresses, the model will gradually improve in distinguishing tampered from non-tampered images, maintaining high performance even in the presence of Gaussian noise.

3. MULTI-MODAL FUSION APPROACH FOR NEWS TEXT AND IMAGE AUTHENTICITY VERIFICATION

The core challenge in verifying the authenticity of news text and images lies in effectively identifying and verifying the authenticity of the image and text information in news content. Particularly in modern media, misinformation is often propagated through a clever combination of manipulated images and misleading text. Hence, researching a multi-modal fusion approach for news image authenticity verification is crucial. Multi-modal fusion methods can utilize different information dimensions of images and text, implementing joint analysis of these two types of information through technologies like deep learning, thereby increasing the accuracy of detecting misinformation. Additionally, these methods help reveal potential inconsistencies between text and images, as well as hidden tampering traces in the context, effectively combating fake news and protecting the public from the impact of false information. In summary, developing advanced multi-modal fusion techniques not only meets the urgent need for efficient authenticity verification methods but is also an important research direction for enhancing information audit quality and ensuring media communication integrity. Figure 2 demonstrates the principle of news text and image authenticity verification based on a multi-modal fusion neural network.

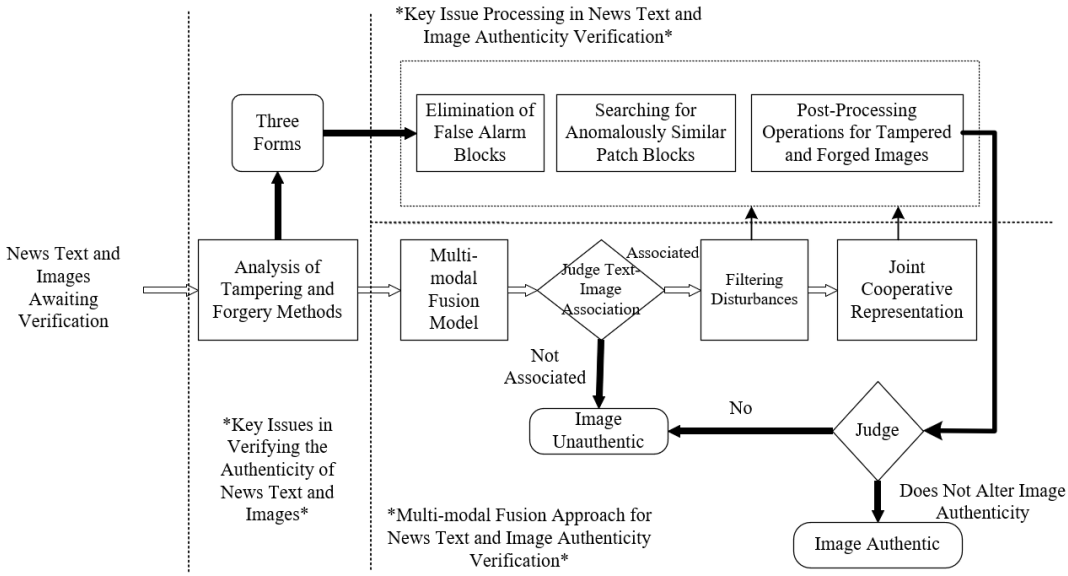


Figure 2. Principle of news text and image authenticity verification based on multi-modal fusion neural network

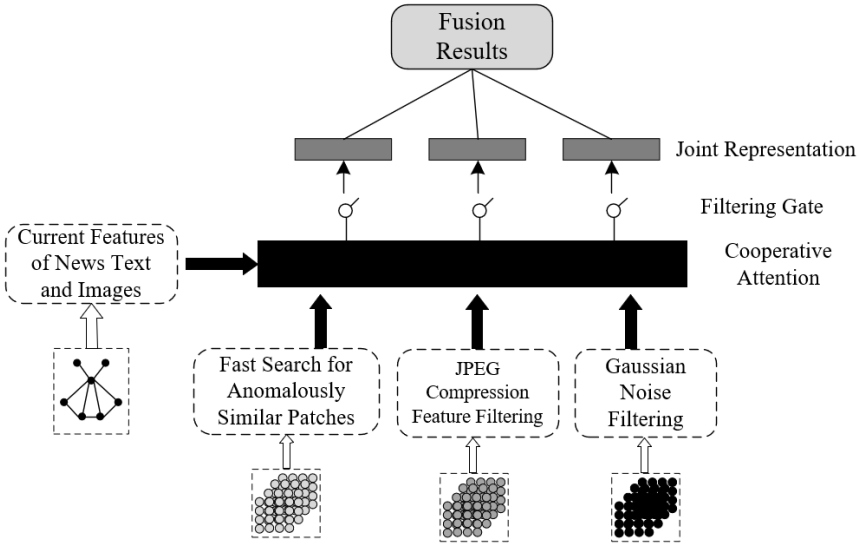


Figure 3. Framework of multi-modal fusion neural network model

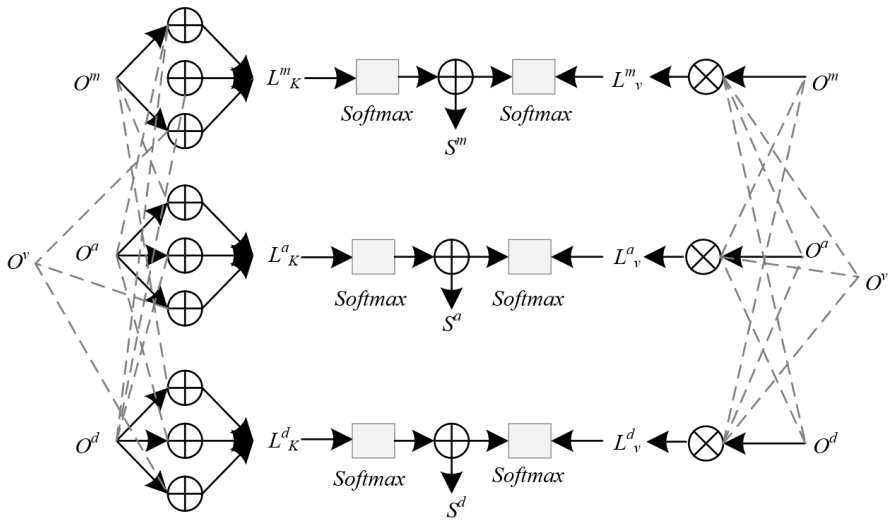


Figure 4. Principle framework of joint gated co-attention mechanism

To effectively distinguish and verify the authenticity of news content, this paper designs three components for the proposed multi-modal fusion neural network model to enhance the model's performance and adaptability. First, the joint gated co-attention mechanism aims to process and analyze the complex associations in image and text data. By focusing on highly correlated information through the attention mechanism, the model's ability to detect subtle manipulations is improved. Second, the filtering gate mechanism effectively reduces interference from noise and irrelevant features, focusing on the identification of tampering traces, which is crucial for precise detection. Finally, the joint cooperative representation and the integration of various sub-modules further optimize the overall performance of the model. This ensures that information from different modalities is comprehensively considered and utilized, enhancing the model's robustness against diverse tampering methods. Figure 3 shows the framework of the multi-modal fusion neural network model.

One of the main challenges in verifying the authenticity of news text and image content is the complex association between images and text, as both modalities may contain different tampering information. To accurately capture and utilize the relationships in these complex multi-modal data, the joint gated co-attention mechanism designed in this paper adopts a more advanced strategy than the traditional conditional attention mechanism. This attention mechanism jointly learns and extracts relevant features between the two modalities, instead of relying solely on the features of a single modality as a condition. This mechanism helps improve the ability to handle diverse tampering scenarios, maintaining high accuracy across a wide range of applications and meeting the stringent requirements of news text and image authenticity verification. Figure 4 presents the principal framework of the joint gated co-attention mechanism.

Typically, the conditional cooperative attention mechanism uses data from one modality to guide the attention distribution of another modality. In practice, this is often achieved through linear combination, i.e., combining the feature vectors of one modality with those of another to generate a new vector that fuses information from both modalities. This step is fundamental in constructing interactions between modalities. Let the three multi-modal sub-modules be represented by O^m, O^a, and O^d. Suppose the conditional association matrices between these multi-modal sub-modules are represented by L^m_v, L^a_v, and L^d_v, the specific implementation process is as follows:

$$\begin{cases} L_v^m = \begin{bmatrix} O^m \oplus O^a \\ O^m \oplus O^v \\ O^m \oplus O^d \end{bmatrix} = \begin{bmatrix} O_2^{ma} \\ O_2^{mv} \\ O_2^{md} \end{bmatrix} \\ L_K^a = \begin{bmatrix} O^a \oplus O^m \\ O^a \oplus O^v \\ O^a \oplus O^d \end{bmatrix} = \begin{bmatrix} O_2^{am} \\ O_2^{av} \\ O_2^{ad} \end{bmatrix} \\ L_K^d = \begin{bmatrix} O^d \oplus O^m \\ O^d \oplus O^a \\ O^d \oplus O^v \end{bmatrix} = \begin{bmatrix} O_2^{dm} \\ O_2^{da} \\ O_2^{dv} \end{bmatrix} \end{cases} \tag{8}$$

Unlike the conditional cooperative attention mechanism, the joint gated co-attention mechanism further calculates a joint association matrix. This matrix reflects not just a unidirectional conditional relationship, but captures the bidirectional or multidimensional relationships between the two modalities. This allows for a more comprehensive consideration of the interaction between images and text, helping to reveal more detailed signs of tampering. Suppose matrix multiplication is represented by $\otimes$, the specific definition formula for calculating the additional joint association matrices $L^m{}_K$, $L^a{}_K$, and $L^d{}_K$ in this paper's method design is as follows:

$$\begin{cases} L_K^m = O^m \otimes \begin{bmatrix} O^m \\ O^v \\ O^d \end{bmatrix} = \begin{bmatrix} O_2^{ma} \\ O_2^{mv} \\ O_2^{md} \end{bmatrix} \\ L_K^a = O^a \otimes \begin{bmatrix} O^m \\ O^v \\ O^d \end{bmatrix} = \begin{bmatrix} O_2^{am} \\ O_2^{av} \\ O_2^{ad} \end{bmatrix} \\ L_K^d = O^d \otimes \begin{bmatrix} O^m \\ O^a \\ O^v \end{bmatrix} = \begin{bmatrix} O_2^{dm} \\ O_2^{da} \\ O_2^{dv} \end{bmatrix} \end{cases} \tag{9}$$

After obtaining the conditional and joint association matrices between O^m, O^a, and O^d, these matrices need to be normalized, commonly using the *softmax* function. This step ensures that the attention weights between different modules are comparable and highlights the most relevant features, providing a basis for subsequent feature weighting. Finally, using the normalized association matrices, the attention weights for each modality are calculated. These weights reflect the relative importance of each part after integrating all modal information. Through weighted averaging, the model can focus on those features that are most important for authenticity verification, while ignoring irrelevant or noisy parts:

$$\begin{cases} S^m = \text{softmax}\left(L_K^m\right) \oplus \text{softmax}\left(L_v^m\right) \\ S^a = \text{softmax}\left(L_K^a\right) \oplus \text{softmax}\left(L_v^a\right) \\ S^d = \text{softmax}\left(L_K^d\right) \oplus \text{softmax}\left(L_v^d\right) \end{cases} \tag{10}$$

In news text and image authenticity verification, the high similarity of multi-modal data may make it difficult for the model to distinguish which features are useful signals for the final prediction and which are noise. In this case, the model might mistake noise for meaningful information, thereby reducing prediction accuracy. To address this issue, this paper introduces a novel filtering gate mechanism designed to dynamically adjust the weights during the fusion of different modal data through the gating mechanism, thereby filtering out noise that could interfere with the model's judgment. The specific implementation is as follows:

$$\begin{cases} \alpha_a^m = \delta\left(Q_a^m \left(\left\| g^m - g^a \right\|_D\right)^{-1}\right) \\ \alpha_v^m = \delta\left(Q_v^m \left(\left\| g^m - g^v \right\|_D\right)^{-1}\right) \\ \alpha_d^m = \delta\left(Q_d^m \left(\left\| g^m - g^d \right\|_D\right)^{-1}\right) \\ \\ \alpha_m^a = \delta\left(Q_m^a \left(\left\| g^a - g^m \right\|_D\right)^{-1}\right) \\ \alpha_v^a = \delta\left(Q_v^a \left(\left\| g^a - g^v \right\|_D\right)^{-1}\right) \\ \alpha_d^a = \delta\left(Q_d^a \left(\left\| g^a - g^d \right\|_D\right)^{-1}\right) \\ \\ \alpha_m^d = \delta\left(Q_m^d \left(\left\| g^d - g^m \right\|_D\right)^{-1}\right) \\ \alpha_a^d = \delta\left(Q_a^d \left(\left\| g^d - g^a \right\|_D\right)^{-1}\right) \\ \alpha_v^d = \delta\left(Q_v^d \left(\left\| g^d - g^v \right\|_D\right)^{-1}\right) \end{cases} \tag{11}$$

Suppose the filtering gate is represented by α, and the learnable parameters in the filtering gate mechanism are represented by Q and n. The F-norm is denoted by $\|_F$. The parameters g^m, g^a, g^v, and g^d satisfy the following definitions:

$$\begin{cases} g^m = \text{Tanh}\left(Q^m \otimes O^m + n^m\right) \\ g^a = \text{Tanh}\left(Q^a \otimes O^a + n^a\right) \\ g^v = \text{Tanh}\left(Q^v \otimes O^v + n^v\right) \\ g^d = \text{Tanh}\left(Q^d \otimes O^d + n^d\right) \end{cases} \tag{12}$$

In the filtering gate mechanism, this paper chooses to use the F-norm as the matrix norm to measure the similarity between multi-modal data modules. The F-norm provides a method to measure the magnitude of matrix elements. It is calculated by squaring the absolute value of all elements in the matrix and then taking the square root, offering an intuitive measure of matrix "size" or "energy." In the context of multi-modal data fusion, the F-norm can be used to assess the similarity or difference between different modules, providing a simplified yet effective way to measure similarity for the gating mechanism.

News text and image authenticity verification requires the precise analysis and judgment of the authenticity of images and text in news content, and the identification of any mismatches or signs of tampering. In this context, the process of joint collaborative representation becomes key to ensuring that the model can effectively fuse multi-modal data.

First, temporary joint representations are calculated based

on the associations of different multi-modal sub-modules. At this stage, the model assesses the associations between each submodule, i.e., the interrelationships between image and text content. These relationships are typically established through learned features. This temporary joint representation provides a preliminary representation that includes multi-modal association information for the fusion process. The specific calculation process for the temporary V^m_y, V^a_y, and V^d_y is as follows:

$$\begin{cases} V_y^m = \begin{bmatrix} \alpha_a^m \otimes O^a \\ \alpha_v^m \otimes O^v \\ \alpha_d^m \otimes O^d \end{bmatrix}^Y \otimes S^m \\ V_y^a = \begin{bmatrix} \alpha_m^a \otimes O^m \\ \alpha_v^a \otimes O^v \\ \alpha_d^a \otimes O^d \end{bmatrix}^Y \otimes S^a \\ V_y^d = \begin{bmatrix} \alpha_a^d \otimes O^m \\ \alpha_a^d \otimes O^a \\ \alpha_v^d \otimes O^v \end{bmatrix}^Y \otimes S^d \end{cases} \tag{13}$$

Next, the model processes this temporary joint representation using the filtering gate parameters. The function of the filtering gate parameters is to dynamically adjust the contribution of different modal data in the final representation, filtering out parts with low relevance or identified as noise, thereby retaining only the information most beneficial for authenticity verification. The joint representation of the three sub-modules is given by the following formula:

$$\begin{cases} V^m = V_y^m \otimes S_s^m \\ V^a = V_y^a \otimes S_s^a \\ V^d = V_y^d \otimes S_s^d \end{cases} \tag{14}$$

For parameters S^m_s, S^a_s, and S^d_s, the formulas are:

$$\begin{cases} S_s^m = \mathrm{softmax}\left(O^m \oplus O^s\right) \\ S_s^a = \mathrm{softmax}\left(O^a \oplus O^s\right), O^a = O^m \oplus O^a \oplus O^d \\ S_s^d = \mathrm{softmax}\left(O^d \oplus O^s\right) \end{cases} \tag{15}$$

After adjustment by the filtering gate parameters, the model calculates a weighted final joint representation. This representation is a composite of all sub-module information, reflecting the optimal combination of multi-modal data after selective filtering. This representation will serve as the basis for the next steps of fusion and classification tasks. The formula for the final joint representation is:

$$V = V^m \oplus V^a \oplus V^d \tag{16}$$

Finally, using the final joint representation, the model calculates the final fusion result through the joint gated cooperative attention mechanism. This fusion result not only considers the interactions between multi-modal data but also ensures, through the gating mechanism, that information from each modality is appropriately considered in the fusion process. Such a fusion result will provide a basis for the final judgment of news text and image authenticity. Suppose the weights for the fusion of various multi-modal sub-module data are represented by Q^m, Q^m, Q^d, and Q^v, then the formula is:

$$\hat{O}^{b+1} = \begin{pmatrix} Q^m \otimes \left(O^m \oplus V_y^m\right) + Q^a \otimes \left(O^a \oplus V_y^a\right) \\ + Q^d \otimes \left(O^d \oplus V_y^d\right) + Q^v \otimes V \end{pmatrix} \tag{17}$$

The entire process of joint collaborative representation emphasizes a thorough understanding of the complex relationships between multi-modal data and builds upon this to effectively filter and fuse information. This method significantly enhances the accuracy and efficiency of news text and image authenticity verification, meeting the high standards of reliability and authenticity required in the news industry.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1 shows the performance of *ComNet*, *GAN*, and the proposed method in detecting tampered and forged news text and images at different *JPEG* image quality factors. From the analysis of the data, it can be seen that our method significantly outperforms *ComNet* and *GAN* methods in accuracy across all *JPEG* quality factor settings. Notably, at a 65% image quality factor, our method achieves an accuracy rate of 85.7%, compared to 44.6% for *ComNet* and 44.8% for *GAN*. This significant gap indicates that the proposed method maintains high detection accuracy even at higher compression rates. In the *MIX* case, where the test set includes images with various compression rates, the proposed method achieves an accuracy of 92.1%, compared to 72.8% and 73.2% for *ComNet* and *GAN*, respectively. This further proves our method's good adaptability to *JPEG* compression, being able to handle images at different compression levels without significantly affecting accuracy. The conclusion is that the proposed method successfully suppresses the model's learning of *JPEG* compression features, meaning the model does not overly adapt to image feature changes caused by *JPEG* compression but focuses more on identifying tampering trace features in news text and images. By reducing sensitivity to *JPEG* compression noise, the proposed method can more effectively detect the authenticity of news content, accurately identifying tampering even in compressed images.

Table 1. Accuracy of tampered and forged news text and image detection under JPEG compression

JPEG Image Quality Factor (%)	*ComNet* (%)	*GAN* (%)	The Proposed Method (%)
65	44.6	44.8	85.7
70	51.8	53.1	92.3
80	65.3	65.9	87.6
90	81.4	84.1	92.4
95	91.5	92.7	95.8
MIX	72.8	73.2	92.1

Table 2 shows the performance of *ComNet*, *GAN*, and the proposed method in detecting target removal news text and images under different signal-to-noise ratio (*SNR*) conditions.

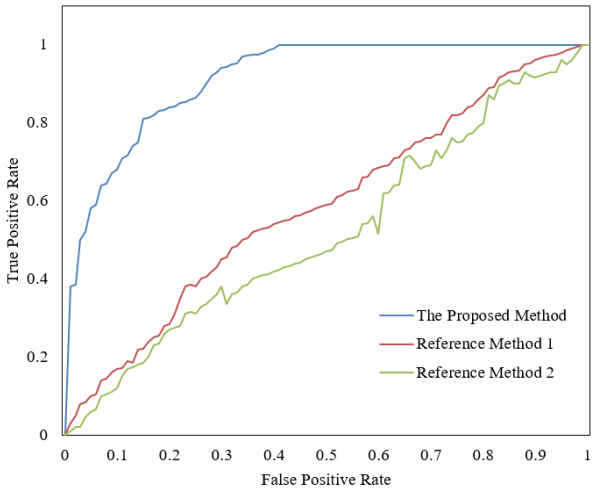
From these data, it is evident that the proposed method consistently achieves higher detection accuracy than *ComNet* and *GAN* across all SNR settings. This demonstrates our method's effectiveness in handling Gaussian noise, thus improving the accuracy of detecting the authenticity of news text and images. At a 25dB SNR, the proposed method reaches an accuracy rate of 83.12%, compared to 52.14% for *ComNet* and 51.24% for *GAN*. This large gap shows that our method still maintains good detection performance at higher noise levels. In the *MIX* situation, where the test set contains images with various SNR levels, the proposed method achieves an accuracy rate of 88.36%, compared to 72.89% and 72.68% for *ComNet* and *GAN*, respectively. This further proves the proposed method's higher stability and robustness when processing image sets with various noise levels. The conclusion is that the proposed method, by adjusting the filtering mechanism, successfully minimizes feature loss caused by Gaussian noise. This capability provides high resistance to interference from Gaussian noise, allowing for more accurate detection of target removal actions in news text and images under noisy conditions.

Table 2. Accuracy of target removal news text and image detection with added gaussian noise

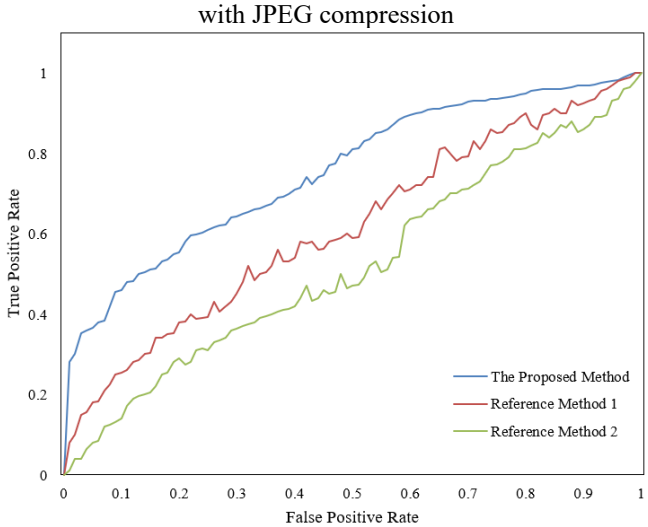
Signal-to-Noise Ratio(*dB*)	*ComNet* (%)	*GAN* (%)	The Proposed Method (%)
25	52.14	51.24	83.12
30	66.39	64.89	85.46
35	73.58	72.13	88.74
40	84.52	83.52	90.12
45	93.68	92.36	93.62
MIX	72.89	72.68	88.36

In Figure 5, Reference Method 1 is based on the bootstrapping multi-view representation detection scheme, and Reference Method 2 is the *MCAN* model. The ROC curves of the three methods in the figure suggest the following conclusions: across the entire range of false positive rates from 0 to 1, the true positive rate of the method proposed in this paper remains at a higher level. Particularly in the lower false positive rate region, the true positive rate of this method rapidly escalates to above 0.81. This indicates that even in cases where a small fraction of tampered text and images are mistakenly marked as authentic, the proposed method can correctly identify most of the tampered content. At higher false positive rates, the true positive rate of this method approaches or reaches 1, signifying that under more lenient classification thresholds, this method can identify almost all tampered text and images. In summary, the proposed method demonstrates superior performance on the ROC curve compared to Reference Methods 1 and 2, especially in maintaining a high true positive rate in low false positive rate regions. This indicates strong robustness and a low false alarm rate for the proposed method in detecting *JPEG* compression tampered news images and recognizing tampered text and images with added Gaussian noise. This is attributed to the method's more effective learning and recognition of features related to *JPEG* compression, along with more accurate detection capabilities of tampering trace features, thus offering better resistance to interference caused by *JPEG* compression. Simultaneously, this method effectively suppresses noise interference, exhibiting strong robustness.

In Figure 6, Model Variant 1 is the model without the joint gated cooperative attention mechanism, Model Variant 2 is without the filtering gate mechanism, Model Variant 3 is without joint cooperative representation, Model Variant 4 is without the search for anomalously similar patch blocks, and Model Variant 5 is without the post-processing feature filtering module. Sample Set 1 is the test set, and Sample Set 2 is the validation set. According to the data in Figure 6, the performance of different model variants in the task of detecting tampered and forged news text and images can be compared and analyzed. It is evident from Figure 6 that the multi-modal fusion method proposed in this paper has higher accuracy in detecting tampered and forged news text and images. In both sample sets, the error rate of the proposed method is lower than that of the other model variants, showcasing its superior robustness and effectiveness. These results emphasize the importance of the combination of technologies such as the joint gated cooperative attention mechanism, filtering gate mechanism, and joint cooperative representation in practical application. These technological integrations provide a robust framework for verifying the authenticity of news text and images, significantly enhancing the ability to detect tampered content. Particularly in processing multi-modal data, the proposed method effectively integrates information from different modalities, reducing errors and enhancing the overall accuracy of detection.



(a) Detection of tampered and forged news text and images with JPEG compression



(b) Detection of tampered and forged news text and images with added gaussian noise

Figure 5. ROC curves for the detection of tampered and forged news text and images with post-processing operations

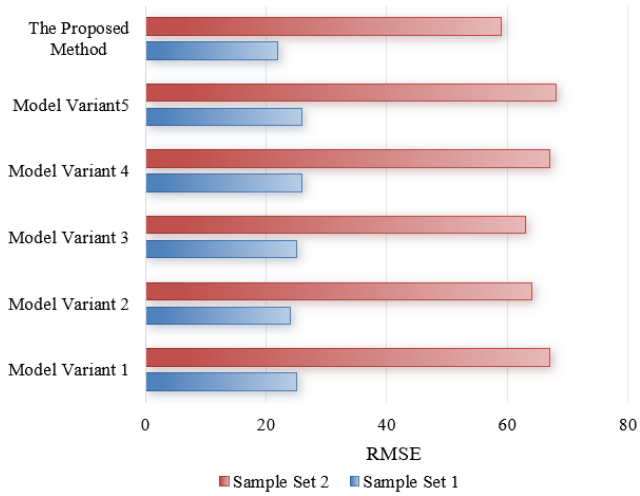
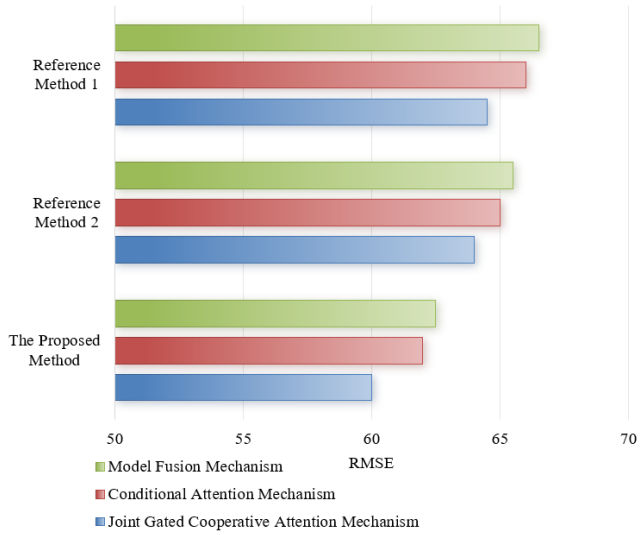
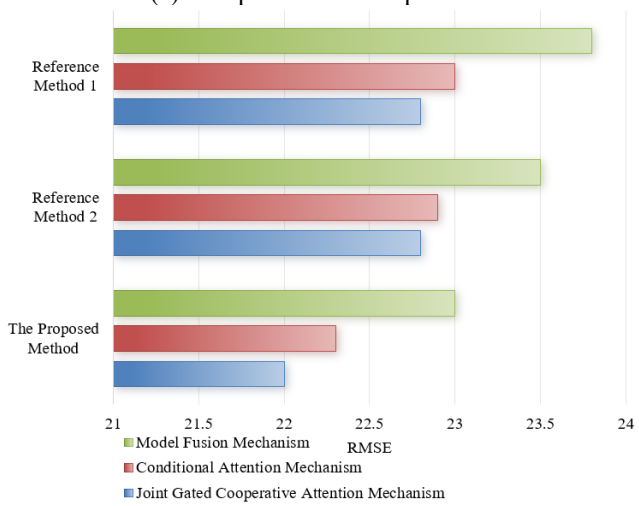


Figure 6. Error comparison of different models in detection of tampered and forged news text and images



(a) Comparison on Sample Set 1



(b) Comparison on Sample Set 2

Figure 7. Error comparison of different multi-modal fusion methods in detection of tampered and forged news text and images

The error comparison data for tampered and forged news text and image detection provided in the Figure 7 allows for an analysis of the performance of different multi-modal fusion methods. In the comparison within Sample Set 1, when applying the joint gated cooperative attention mechanism, the error rate of the proposed method is lower by 4.5% and 4% compared to Reference Method 1 and Reference Method 2, respectively. This indicates that the joint gated cooperative attention mechanism is more effective in fusing multi-modal information in Sample Set 1, thereby reducing the error rate. In the comparison within Sample Set 2, with the joint gated cooperative attention mechanism, the error rate of the proposed method is 22%, slightly lower compared to Reference Method 2 and Reference Method 1, showing that all three methods demonstrate high accuracy in the tests of Sample Set 2. In summary, whether in Sample Set 1 or Sample Set 2, the proposed method consistently exhibits a lower error rate under three different multi-modal fusion mechanisms, highlighting its superior performance in the task of detecting tampered and forged news text and images. Particularly, the introduction of the joint gated cooperative attention mechanism enhances performance across different datasets and fusion mechanisms. This suggests that the multi-modal fusion method and its key technology-the joint gated cooperative attention mechanism-proposed in this paper are not only innovative in theory but also truly enhance the accuracy and robustness of news text and image authenticity verification in practice.

Table 3. Performance evaluation metrics of different methods in detection of tampered and forged news text and images

Methods	***Dice* Coefficient**			***Hausdorff*95(*mm*)**		
	Object Class	**Scene Class**	**Text Class**	**Object Class**	**Scene Class**	**Text Class**
Reference Method 1*t*	0.678	0.911	0.832	47.25	4.785	6.581
Reference Method 2	0.725	0.878	0.841	38.95	5.235	6.742
The proposed method	0.812	0.921	0.856	35.22	4.658	5.268
Methods	***Sensitivity***			***Specificity***		
Reference Method 1*t*	0.662	0.912	0.835	0.989	0.989	0.989
Reference Method 2	0.715	0.925	0.836	0.989	0.987	0.989
The proposed method	0.814	0.935	0.844	0.990	0.991	0.991

Table 3 provides a series of metrics for performance evaluation in tampered and forged news text and image detection, including *Dice* coefficient, *Hausdorff*95 distance, Sensitivity, and Specificity. Each metric evaluates the performance of detection algorithms from different perspectives. The table reveals that the proposed method has higher *Dice* coefficients in the object, scene, and text classes than Reference Methods 1 and 2, indicating a significant advantage in accurately identifying tampered areas. The proposed method also shows better performance in the *Hausdorff*95 distance, particularly in the object class, meaning the detected tampering boundaries are closer to the actual tampered edges. In terms of sensitivity, the proposed method also surpasses both reference methods, especially in the object and scene classes, indicating fewer missed tampered areas. In specificity, the proposed method is slightly better or equivalent to the reference methods, demonstrating its reliability in avoiding false alarms. Overall, the multi-modal fusion method presented in this paper exhibits superior performance in the task of detecting tampered and forged news text and images.

It demonstrates higher accuracy and robustness both in accurately identifying tampered areas and avoiding false alarms and missed genuine tampered areas. These results emphasize the effectiveness of the multi-modal fusion method presented in this paper and prove its potential in practical applications.

Table 4. Performance comparison of different methods in detection of tampered and forged news text and images

Methods	*Dice* Coefficient			*Hausdorff*95(*mm*)		
	Object Class	Scene Class	Text Class	Object Class	Scene Class	Text Class
Model Variant 1	0.734	0.915	0.824	34.562	4.235	7.124
Model Variant 2	0.728	0.897	0.773	33.258	6.485	16.362
Model Variant 3	0.726	0.883	0.831	35.697	8.121	11.234
Model Variant 4	0.724	0.889	0.845	38.264	5.462	6.785
Model Variant 5	0.695	0.915	0.841	47.231	4.785	6.523
The Proposed Method	0.814	0.925	0.905	31.247	4.125	5.247

From the data in Table 4, the performance of different model variants and the multi-modal fusion method presented in this paper in the task of detecting tampered and forged news text and images can be compared. The table shows that the proposed method has higher *Dice* coefficients in the object, scene, and text classes than all model variants, achieving 0.814, 0.925, and 0.905, respectively. This means the proposed method has a significant advantage in accurately identifying tampered areas, better matching the true tampered regions. The proposed method also exhibits lower *Hausdorff*95 distances in the object, scene, and text classes compared to the model variants, at 31.247mm, 4.125mm, and 5.247mm, respectively. Particularly in processing text class tampering, the proposed method shows a significant advantage over Model Variant 2, whose *Hausdorff*95 reaches 16.362mm, indicating a considerable improvement in the precision of determining tampered boundaries. Combining the *Dice* coefficient and *Hausdorff*95 distance evaluation metrics, it can be concluded that the multi-modal fusion method presented in this paper excels in the task of detecting tampered and forged news text and images. It not only more accurately identifies and matches tampered areas but also more precisely delineates tampering boundaries, thereby providing more accurate and robust detection performance.

5. CONCLUSION

This paper delves into the technique of searching for anomalously similar patch blocks within news text and image content, which is crucial for detecting content tampering. By identifying abnormal repetitive patterns in images, it is possible to effectively locate areas that may have been tampered with. The paper analyzes common post-processing techniques in tampered images, such as *JPEG* compression and the addition of Gaussian noise, which are often used to conceal tampering traces, thereby increasing the difficulty of tampering detection. A novel multi-modal fusion method is proposed, achieved through the joint gated cooperative attention mechanism, filtering gate mechanism, and joint collaborative representation. This effectively integrates image and text information to identify the authenticity of news content.

Through metrics such as the *Dice* coefficient, *Hausdorff*95 distance, sensitivity, and specificity, the proposed method demonstrates superior performance at various levels compared to existing methods. These metrics collectively attest to the new method's accuracy and robustness in identifying and locating tampered areas. Across different test sets and validation sets, the method shows a lower error rate compared to other model variants and comparative methods, underscoring its effectiveness in practical application.

This paper makes significant advancements in the field of news text and image authenticity verification, both theoretically and practically. The proposed multi-modal fusion method is not only theoretically innovative but also proven effective in practice through experimental results. The introduction of the joint gated cooperative attention mechanism, along with the application of the filtering gate mechanism and joint cooperative representation, collectively enhances the model's ability to capture the complex associations between multi-modal data, improving the precision and robustness of tampered news text and image detection. Additionally, the method can resist interference from post-processing operations in tampered images, such as *JPEG* compression and Gaussian noise, further demonstrating its robustness. Ultimately, this research provides an effective technical approach for the field of news text and image authenticity verification and lays a solid foundation for future studies.

REFERENCE

[1] Maria, K., Stelios, T., Evangelos, G. (2023). Technical university of Crete February 2023 readers' satisfaction from online news websites. In Novel & Intelligent Digital Systems Conferences, Athens, Greece, pp. 52-61. https://doi.org/10.1007/978-3-031-44097-7_5

[2] Im, J., Park, E. (2022). Effects of political orientation on sentiment features: The case of online news outlets in South Korea. Telematics and Informatics, 74: 101882. https://doi.org/10.1016/j.tele.2022.101882

[3] Lisičar, H., Katulić, T., Jurić, M. (2022). Implementation of GDPR transparency principle in personal data processing by Croatian online news sites. In 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, pp. 1264-1269. https://doi.org/10.23919/MIPRO55190.2022.9803637

[4] Shahu, A., Melem, A., Wintersberger, P., Michahelles, F. (2022). Nudgit-reducing online news consumption by digital nudges. In Adjunct Publication of the 24th International Conference on Human-Computer Interaction with Mobile Devices and Services, Vancouver, BC, Canada, pp. 1-5. https://doi.org/10.1145/3528575.3551447

[5] Hlaing, M.M.M., Kham, N.S.M. (2020). Defining news authenticity on social media using machine learning approach. In 2020 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, pp. 1-6. https://doi.org/10.1109/ICCA49400.2020.9022837

[6] Mahmud, M.A.I., Talukder, A.A.T., Sultana, A.,

Bhuiyan, K.I.A., Rahman, M.S., Hasan, T.H.P., Rahman, R.M. (2023). Toward news authenticity: synthesizing natural language processing and human expert opinion to evaluate news. IEEE Access, 11: 11405-11421. https://doi.org/10.1109/ACCESS.2023.3241483

[7] Francis, E., Monroe, A., Sidnam-Mauch, E., Ivancsics, B., Washington, E., McGregor, S.E., Bonneau, J., Caine, K. (2023). Transparency, trust, and security needs for the design of digital news authentication tools. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1): 1-44. https://doi.org/10.1145/3579534

[8] Pande, S.D., Rathod, S., Joshi, R., Chvan, G.T., Jadhav, D., Phutane, P., Gonge, S., Kadam, K. (2022). Fake news identification using regression analysis and web scraping. International Journal of Safety and Security Engineering, 12(3): 311-318. https://doi.org/10.18280/ijsse.120305

[9] Akpulat, M., Bicakci, K., Cil, U. (2013). Revisiting graphical passwords for augmenting, not replacing, text passwords. In Proceedings of the 29th Annual Computer Security Applications Conference, New Orleans, Louisiana, USA, pp. 119-128. https://doi.org/10.1145/2523649.2523672

[10] Sreenivasulu, V., Wajeed, M.A. (2021). Image based classification of rumor information from the social network platform. Traitement du Signal, 38(5): 413-1421. https://doi.org/10.18280/ts.380516

[11] Chen, Y., Chen, J., Li, M. (2021). Numerical modeling of a new virtual trajectory password architecture. In Journal of Physics: Conference Series, Guangzhou, China, pp. 012013. https://doi.org/10.1088/1742-6596/2068/1/012013

[12] Yang, G.C., Hu, Q., Asghar, M.R. (2022). TIM: Secure and usable authentication for smartphones. Journal of Information Security and Applications, 71: 103374. https://doi.org/10.1016/j.jisa.2022.103374

[13] Chu, X., Sun, H., Chen, Z. (2020). Passpage: Graphical password authentication scheme based on web browsing records. In Financial Cryptography and Data Security: FC 2020 International Workshops, AsiaUSEC, CoDeFi, VOTING, and WTSC, Kota Kinabalu, Malaysia, pp. 166-176. https://doi.org/10.1007/978-3-030-54455-3_12

[14] Abraheem, A., Bozed, K., Eltarhouni, W. (2022). Survey of various graphical password techniques and their schemes. In 2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), Sabratha, Libya, pp. 105-110. https://doi.org/10.1109/MI-STA54861.2022.9837719

[15] Rajendra, A.B., Sheshadri, H.S. (2013). Enhanced visual secret sharing for graphical password authentication. In International Conference on Graphic and Image Processing (ICGIP 2012), Singapore, pp. 682-686. https://doi.org/10.1117/12.2010934

[16] Manjula Shenoy, K., Supriya, A. (2019). Authentication using alignment of the graphical password. In Proceedings of the 3rd International Conference on Advanced Informatics for Computing Research, ICAICR 2019, Shimla, India, pp. 1-5. https://doi.org/10.1145/3339311.3339332

[17] Tang, Z.Y., Tian, C.X., Ye, G.X., Li, J., Wang, W., Gong, X.Q., Fand, D.Y. (2020). A recognition method for text-based captcha based on CGAN. Chinese Journal of Computers, 43(8): 1572-1588.

[18] Das, R., Singh, T.D. (2023). Image-text multimodal sentiment analysis framework of Assamese news articles using late fusion. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6): 1-30. https://doi.org/10.1145/3584861

[19] Wajid, M.A., Zafar, A., Terashima-Marín, H., Wajid, M.S. (2023). Neutrosophic-CNN-based image and text fusion for multimodal classification. Journal of Intelligent and Fuzzy Systems, 45(1): 1039-1055. https://doi.org/10.3233/JIFS-223752

[20] Zhang, J., Wang, Y. (2022). SRCB at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Seattle, United States, pp. 585-596. https://doi.org/10.18653/v1/2022.semeval-1.81

[21] Liu, X., Wang, Z., Wang, L. (2021). Multimodal fusion for image and text classification with feature selection and dimension reduction. In Journal of Physics: Conference Series, Nanjing, China, pp. 012064. https://doi.org/10.1088/1742-6596/1871/1/012064