Multi-Modal Modified U-Net for Text-Image Restoration: A Diffusion-Based Multimodal Information Fusion Approach

Ailong Tang, Ling Wei, Zhiping Ni, Qiuyong Huang*
College of Information Science and Engineering, Liuzhou Institute of Technology, Liuzhou 545000, Guangxi, China
E-mail: qiuhqy@163.com
*Corresponding author

Keywords: multimodal information fusion, multi-modal modified U-Net and refine layer, PSNR, SSIM, text image restoration

Recieved: February 10, 2025

Realistic picture restoration is a crucial task in computer vision, with diffusion-based models widely explored for their generative capabilities. However, image quality remains a challenge due to the uncontrolled nature of diffusion theory and severe image degradation. To address this, we propose a Multi-Modal Modified U-Net (M3UNET) model that integrates textual and visual modalities for enhanced restoration. We leverage a pre-trained multimodal large language model to extract semantic information from low-quality images and employ an image encoder with a custom-built Refine Layer to improve feature acquisition. At the visual level, pixel-level spatial structures are managed for fine-grained restoration. By incorporating control information through multi-level attention mechanisms, our model enables precise and controlled restoration. Experimental results on synthetic and real-world datasets demonstrate that our approach surpasses state-of-the-art techniques in both qualitative and quantitative evaluations, proving the efficacy of multimodal insights in improving image restoration quality.

Povzetek: Predlagan je multimodalni M3UNET model, ki z združevanjem besedilnih in slikovnih informacij ter difuzijskim pristopom bistveno izboljša kakovost obnove degradiranih besedilnih slik, generiranih iz besedil.

1 Introduction

The field of multimodal natural language processing has great promise for enhancing the comprehension and production of material that integrates many modalities, such as text and visuals. Important multimodal natural language processing applications include picture captioning, which generates textual descriptions of images automatically [1]. The data growth on the Internet along with our home PCs is a direct result of the recent improvements in technology and gadgets. Online stores rely on this data, which may be found in a variety of formats (text, images, videos, etc.). These websites feature items that are multimodal, meaning they offer both visuals and textual descriptions. Conventional wisdom holds that a single modality is the sweet spot for classification and data retrieval techniques [2]. Deblurring fuzzy text images without understanding the blur kernel is what blind text picture deblurring is all about. Multiple blind text deblurring image algorithms have shown the efficacy of sparsity-based approaches. Nevertheless, the restoration impact of the blur kernel is impacted by the fact that sparse prior based blur kernel estimates techniques do not take the blur kernel's brightness information into account [3].

The process of image captioning enables computers to decipher visual information and provide textual descriptions of images. After its inception, deep learning's application to the task of interpreting picture data and producing descriptive text quickly became a hot topic in the academic community. However, not all examples that represent conceptual concepts can be found using these procedures. The truth is that most of them don't seem to have any bearing on the matching duties. A small number of significant semantic occurrences define the level of similarity [4]. While human painters have been known to effectively restore items that have suffered extensive damage, inpainting algorithms have so far failed to replicate this feat. Artists often make educated guesses about the severely damaged picture and document them in a written description before attempting to restore it [5]. The need for reliable and flexible image translation methods has increased dramatically in recent computer vision research. But conventional approaches have a hard time adapting material to different settings and capturing semantic subtleties [6]. One method that may be used to create a concise summary using both text and pictures is multimodal abstractive summarizing, or MAS [7]. The modern medical industry is seeing phenomenal growth. Accurate information may be gleaned from the merging of several medical picture modalities, allowing for earlier illness detection and better treatment planning. The goal of picture fusion is to create a single, more comprehensive and informative image than each of the individual input photos could have been created by separately processing them [8]. One important strategy for social network sentiment analysis in the last few years has been to combine text and visual data. Still, there are limitations to current methods for efficiently integrating multimodal characteristics and collecting complicated cross-modal information [9].

A lot of people are starting to take notice of multimodal big-language models, which means they might be useful for a wide range of vision-language activities in the future. MLLMs include a lot of outside information into their parameters, but keeping these models up-to-date is difficult, since it requires a lot of computing resources and isn't very interpretable. Both LLMs as well as MLLMs have shown success using retrieval enhancement approaches as plugins [10]. In order to facilitate downstream vision tasks, multimodal image fusion seeks to combine data from many imaging modalities into a single, detailed picture. Although models using transformers are computationally costly, they excel in global modeling [11]. In contrast, existing approaches using local CNNs have trouble effectively capturing global characteristics. One classic method for analyzing sentiment that relies on text is multimodal sentiment analysis. Inconsistent cross-modal feature data, inadequate interaction capabilities, with insufficient feature fusion are still issues that the area of multi-modal sentiment evaluation confronts [12]. The goal of image processing operations like restoration and enhancement is to produce a clean, high-quality output from an imperfect input. In regard to single-task circumstances, approaches based on deep learning have shown better performance for a variety of image processing tasks.

However, their generalizability and practical applicability are constrained since they need to train distinct models for various degradations and levels [13]. The goal of image processing operations like restoration and enhancement is to produce a clean, high-quality output from an imperfect input. In regards to single-task circumstances, approaches based on deep learning have shown better performance for a variety of image processing tasks. However, their generalizability and practical applicability are constrained since they need to train distinct models for various degradations and levels [14]. Throughout a patient's journey with the healthcare system, a multitude of clinical data can be found in various formats, including structured, unstructured, or semi-structured information derived from laboratory results, clinical notes, diagnostic code, imaging, audio, and other observational data.

If we can create a representation system that incorporates data from all these different places, we may potentially combine our models on data that has greater predictive value than noise while integrating the limitations we've learned from more accurate information into them [15]. Modern methods for Multimodal Information Extraction are required due to the meteoric increase in the use of social media and other forms of multimodal communication. The semantic and modality gaps that exist between pictures and text pose considerable problems to the direct Image-Text interactions that are the backbone of current techniques [16]. The explosion of new technology and consumer electronics has led directly to the deluge of data stored on computers and the internet. The majority of these data points are compiled from a variety of sources (picture, video, text, etc.). For online stores, this data is equally crucial. The items sold on these websites are multimodal as they include both visuals and textual descriptions. Classification along with data retrieval systems from the past tended to prioritize only one modality over another [17].

1.1 Research gaps

Because current models aren't very versatile, it's usually better to use specialist restoration models that can only do one task at a time. But there are usually a number of degradations happening at once in real-world photos. In a low-quality photograph, for instance, you may see rain, blur, and noise all at once. Degradation phenomena may interact with one another in complicated ways, and various degradations may call for diverse approaches to treatment. The ultimate success of the repair process depends on the order and mix of these techniques. Utilizing expert information and creating all-in-one models have propelled recent field improvements. Below, we provide a full analysis to help you comprehend this subject and our motivations.

- The majority of current picture restoration techniques rely on neural networks that train robust image-level priors from massive amounts of data in order to approximate the missing data. When pictures contain significant information gaps, nevertheless, these efforts still fail.
- There are additional restrictions in the realm of applications for introducing external priors or using reference photos to provide information. As a counterpoint, text input is both more accessible and offers more flexible information.

The specific research contributions are follows:

- Sorting Degradation into Different Categories. The degradation classes of an input picture are detected automatically by M3-UNET, which then calculates the necessary restoration activities.
- A Restoration Sequence by Adaptation. M3-UNET improves the overall efficiency of the picture restoration process by going beyond the limitations of predefined, human-specified model implementation orders and instead deciding the optimal sequence for applying the

restoration models based on dynamic evaluations of the unique properties of each input image.

- Best Practices for Choosing a Model. For each restoration job, M3-UNET dynamically chooses the best model from the pool based on the exact deterioration features in the input picture, guaranteeing optimum performance.
- Automation of Processes. There is no need for human interaction once M3-UNET determines the restoration process and model selection; the whole restoration process is executed autonomously.

1.2 Objectives

- To propose a dual-modal text-guided image restoring model is suggested as a solution to the problems with current restoration algorithms, which include insufficient context-dependent information, poor results in fixing large broken areas, and uncontrollably reconstruction results.
- To connect the significant a representation gap among visual and textual methods is a major challenge in this task.
- To essentially address the feature illustration between visually and textual modalities are another challenge.

The experimental findings show that M3-UNET outperforms human specialists when it comes to sophisticated degradation. In addition, the system's modular architecture makes it easy to add additional tasks and models, which makes it more versatile and scalable for different uses.

1.3 Contributions

Last but not least, we use attention methods to enhance the diffusion model's denoising M3-UNET with pixel control, text embedding, and picture embedding. In order to recover photos while preserving their structural integrity and level of detail, all modules work together. The following is an overview of our contributions:

- We provide a model for realistic picture restoration based on diffusion that takes into account both visual and textual level information.
- To achieve efficient textual control, we use M3-UNET and refine layers. A pixel-level processor with multi-layer supervision allows us to achieve precise control over individual pixels.
- By using multi-layer attention processes, we include the control information into the diffusion model. Our model outperforms the competition on several datasets using a wide range of picture quality criteria.

2 Related work

Image captions that are both informative and accurate is generated using a new multimodal natural language processing approach that was proposed in [18]. By integrating data from image-related text descriptions, their model outperforms conventional unimodal models in capturing contextual signals and producing captions with more nuances. Their multimodal fusion strategy is proven effective as they validate it with the industry standard dataset, Flickr8K, and get state-of-the-art results. In addition, they emphasize how multimodal NLP has the ability to transform their interactions with computers and the way they understand visual information, and they talk about the advantages and disadvantages of that approach to picture captioning. The research cited in [19] uses neutrosophic fuzzy sets to handle uncertainty in information retrieval tasks and classifies multimodal input. Drawing on previous methods of superimposing text over photographs, that endeavor makes use of both image and text data in an effort to categorize the images with neutrosophic classification methods.

Classification tasks make use of feature representations learned by Neutrosophic Convolutional Neural Networks from the generated pictures. For learning representations of the novel fusion approach, they show how to use an NCNN-based pipeline. Conventional convolutional neural networks suffer when trying to classify noisy data because they are susceptible to test-phase noise that isn't yet recognized. Two large-scale multi-modal classification datasets showed promising results when compared to individual sources using their technique. The two popular multi-modal fusion approaches, early fusion along with late fusion, have also been compared to their technique. Present a new approach to blind text picture deblurring using sparse priors and multi-scale fusion in [20]. To further limit the possible solutions space and get excellent clean pictures, they augment the sparse gradient earlier on the hidden clean text image with the sparse earlier on the high-energy wavelet values of the implicit text image. Optimizing the blur kernel with the latent clean image are done in turn using the semi-quadratic splitting approach. They also take into account the restored blur kernel's brightness feature's potential impact. In order to enhance the quality of the blur kernel, they fuse the generated blur kernels in three channels using a multi-scale fusion approach that is based on the Laplacian weight with saliency weight. Their approach successfully restores blur kernels with text pictures, according to the testing findings. Presented in [21] is a deep learning model that utilizes multimodal feature fusion.

Decoding has taken place in long short-term memory; mask recurrent neural networks are used in the coding layer, while the descriptive text is created. Deep learning makes use of gradient optimization to fine-tune the model's parameters. Dense attention methods may help the decoding layer input the right data preferentially and reduce non-salient data interruption. The goal of using input photographs to train a model is to create captions that, when given the chance, will come close to properly describing the images. The accuracy and proficiency of the model's language acquisition using image description analysis are assessed using several datasets. These tests show that the model correctly describes the input photos every time. This model has learned to describe an input image using words or captions. Classification scores are used to evaluate the model's performance. A 95% improvement in performance is shown by the proposed system while using 100 training epochs and a batch count of 512. Results from experiments conducted on generic picture datasets corroborate the model's ability to understand visual content and produce text. Use of Python frameworks in its implementation and evaluation using performance measures including PSNR, RMSE, SSIM, recall, accuracy, F1-score, and precision are all aspects of that research. By modeling its approach after that of artists' hypothesis, [22] suggests including text description into picture inpainting tasks for the first time; That would give a wealth of information useful for restoring images by fusing multimodal elements.

They present MMFL, a multimodal fusion learning approach to picture inpainting. An image-adaptive word demand component is built to fairly filter the most effective text features, allowing for improved usage of text features they provide a text-image matching penalty and a text-guided attention loss to train the network to focus on items described in text. Their technique outperforms the state-of-the-art in generating fine-grained textures and accurately predicting the semantics of items in the missing areas, according to extensive trials. A groundbreaking method based on multimodal datasets is presented in [23]. They want to improve the picture translation model's ability to understand semantic nuances and improve the accuracy of content adaption by making use of the abundance of data found in multimodal datasets. They want to reinvent image translation technology by fusing information across diverse modalities—images, text, and audio—and bringing unique insights for innovation and growth. By combining deep learning techniques with multimodal data fusion structures, their study aims to fill the gaps in image translation that currently exist. Ensuring robustness along with integrity throughout the analytical process, they painstakingly preprocess and combine data from multiple sources. In a set of carefully planned tests, they compare their method's efficiency to that of more traditional approaches.

Their results demonstrate that their multimodal technique significantly improves translation quality and effectiveness. In addition to establishing a firm groundwork for future research initiatives that study helps push the boundaries of image translating technology forward. They usher in a new age of advancement and innovation in computer vision by illuminating the revolutionary power of multimodal datasets. A successful fusion-based decision-making technique was proposed in [24] to classify social media information into Informative while non-informative categories. The data was analyzed using CrisisMMD and related to seven major natural disasters such as floods, hurricanes, hurricanes, wildfires, etc. The tweets are sorted into many humanitarian categories, including those pertaining to rescue attempts, donations, infrastructure as well as utility damage, impacted persons, and not-humanitarian categories. When compared to baselines based on text tweets, the suggested multi-modal fusion approach achieves 6.98% improvement in the Informative area and 11.2% improvement in the Humanitarian category. When compared to baselines based on picture tweets, the proposed technique achieves a 4.5% improvement in the Informative area and a 6.39 percent improvement in the humanitarian category. When analyzing multimodal data, it is crucial to take into account the interdependencies between the various modalities, as stated in [25]. Their goal is to automate video data analysis by using state-of-the-art deep machine learning along with information fusion techniques that fully consider all interdependencies between and within different modalities. They emphasize the critical significance of human connections in the success of microenterprises with an empirical demonstration that measures the reliability of grassroots merchants in real-time trading on Tik Tok. In order to help with combining information in strategy study that uses multimodal data, they make their data and techniques available.

The importance of nonverbal and vocal communication in reaching strategic goals is emphasized by their research. We show that human contacts are crucial for microenterprises to succeed by analyzing multimodal data (text, photos, and audio) and highlighting the importance of trustworthiness. Their use of explainable AI and data fusion for multimodal information significantly improves the predicted accuracy and theoretical comprehension of trustworthiness assessments. The Dynamic Graph-Text Fusion Network, a multimodal sentiment analysis algorithm, was designed to tackle these issues in [26]. By seeing words as nodes and combining their attributes via their adjacency connections, text features are acquired by using the neighborhood data collection capabilities of Graph Convolutional Networks. The multi-head attention method is also used to concurrently extract extensive semantic data from several subspaces. They use a convolutional attention component to extract features from images. After that, the text and picture characteristics are combined using an attention-based fusion module. The suggested DGFN model is successful, as shown experimentally on the two datasets, where sentiment accuracy for classification and F1 scores increase significantly. One such retrieval-augmented framework for several MLLMs is proposed in [27] as RA-BLIP, which stands for multimodal adaptive retrieval-augmented bootstrapping language-image pre-training.

Taking into account the fact that the visual modality contains redundant information, they started by using the question to guide the gathering of visual data by interacting with a single set of learnable searches, thereby reducing the amount of irrelevant interference that occurs during both retrieval and production. In addition, they provide a pre-trained multimodal adaptive fusion unit that can integrate visual and verbal modalities into a single semantic space, allowing for query text-to-multimodal retrieval. In addition, they provide an ASKG technique for training the generator's brain to autonomously determine the relevance of recovered information, resulting in exceptional denoising performance. The results show that RA-BLIP outperforms the top retrieval-augmented models and achieves considerable performance on open multimodal question-answering datasets. If you're looking for a way to beat CNNs with vision transformers in computer vision tasks, check out FusionMamba, a new dynamic feature improvement framework introduced in [28]. By using dynamic convolution while channel attention methods, the framework extends the visual state-space paradigm Mamba. That model preserves its outstanding global feature modeling capacity while also reducing redundancy and increasing the expressive capacity of local features. Furthermore, a brand-new module known as the flexible feature fusion component has been created by their team.

It integrates the DFEM module, which improves texture and disparity perception, with the CMFM module, which enhances inter-modal correlation and suppresses redundant information, to build a cross-modal fusion model. The experimental results demonstrate that FusionMamba outperforms its competitors and is highly applicable across a range of multimodal picture fusion challenges and downstream tests. The author suggests using CLIP, an interaction between images and text, to build a cross-modal sentiment framework [29]. To extract main image-text characteristics, the model uses pre-trained ResNet50 and RoBERTa. To improve information transmission across multiple modalities, it uses a multi-head attention system for cross-modal features interaction after contrastive learning using the CLIP model. Then, feature networks are fused using a cross-modal gating module, which allows for the regulation of feature weights while integrating features at various levels. For sentiment recognition, the finished product is sent into a fully linked layer. The MSVA-Single and MSVA-Multiple datasets, which are made publically accessible, are used in comparative investigations. On the aforementioned datasets, their model attained 75.38% accuracy and 73.95% F1-score, according to the experimental findings. That shows that the suggested method outperforms current sentiment analysis methods in terms of generalizability and robustness.

For accurate multimodal sentiment while emotion categorization, the authors of [30] suggested a new deep multi-view attentive architecture. There are three stages to the DMVAN model: learning features, learning attentive interactions, and learning cross-modal fusion. For precise categorization, the feature learning step involves extracting visual features from scene and area views as well as textual data from word, phrase, and document levels of analysis. To improve the interaction between visual and textual information, the image-text interaction learning system is used in the attentive interaction learning phase. That mechanism extracts discriminative and sentimental visual features and uses textual information to train image features. To further take use of the complementing qualities of several modalities, a cross-modal fusion instructional component is designed to merge distinct features into a holistic framework. Next, a multi-head attention technique is used to gather and combine sufficient information from the intermediate characteristics to help build a strong joint representation. In other related fields, image processing methods such as photogrammetry, medical imaging, and computer vision have their characteristics and innovations. Some cutting-edge image processing methods will help researchers achieve new breakthroughs. Table 1 summarizes image registration algorithms in other fields.

Table 1: Analysis of image registration algorithms in other fields

Reference	**Proposed method**	Limitations
Authors [18]	Multimodal NLP	Difficult to fuse infrared-visible and multi-focus image fusion.
Authors [19]	Neutrosophic Fuzzy sets Neutrosophic Convolutional Neural Networks	Difficult to propose more rapid and active methods of medical image enhancement and fusion

Authors [20]	High-Energy Wavelet Using The Semi-Quadratic Splitting Approach	Low image quality, performance was not consistent so low efficiency
Authors [21]	Recurrent Neural Networks	Computational complexity was high
Authors [22]	MMFL	Difficult to fuse the images
Authors [23]	combining deep learning techniques	Fused image quality was poor
Authors [24]	fusion-based decision-making technique	Required more computational time to perform the task
Authors [25]	deep machine learning along with information fusion	Implementation was complex
Authors [26]	Dynamic Graph-Text Fusion Network	Fused image quality was poor
Authors [27]	RA-BLIP	Failed to execute in real-time applications
Authors [28]	CNNs with vision transformers	Required more computational time to perform the task
Authors [29]	ResNet50 and RoBERTa	Time consumption was more
Authors [30]	DMVAN model	Difficult to fuse the images

3 Proposed Work

3.1 System Method

Fig 1 shows the proposed work architecture for image restoration.

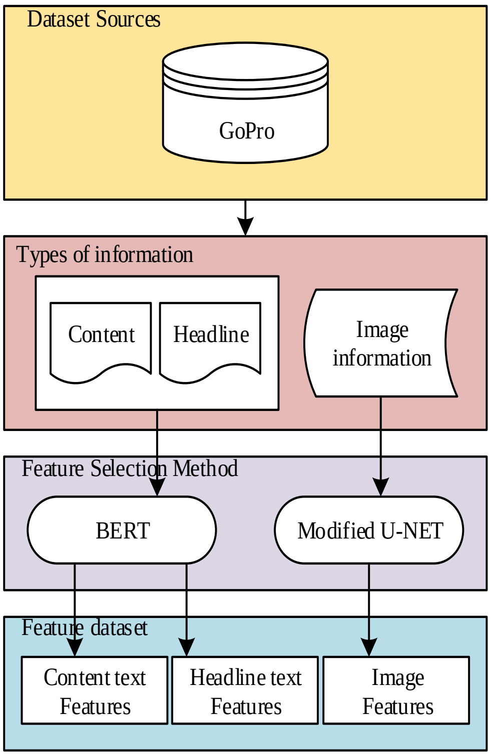


Figure 1: Flow diagram for Modified U-Net Model

Text processing:

- We search for keywords, organizations, and semantic connections in the description language and pull them out.

Image feature extraction:

- The degraded picture is processed using methods such as CNNs to extract features.

Feature fusion

- A selected fusion technique merges the text-derived characteristics with the picture features, making use of attention processes to zero in on pertinent data.

Image reconstruction

- A restored picture is often constructed using the fused characteristics, typically by use of a generative model.

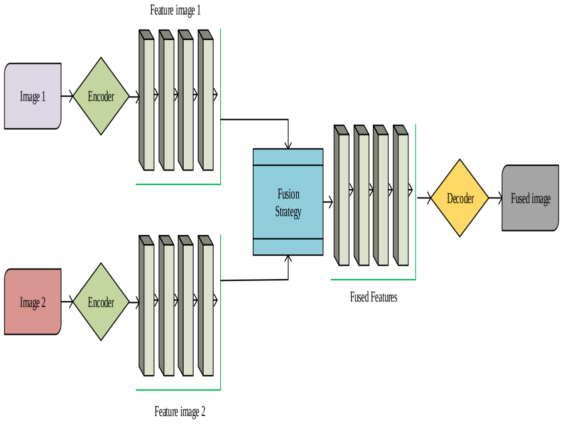
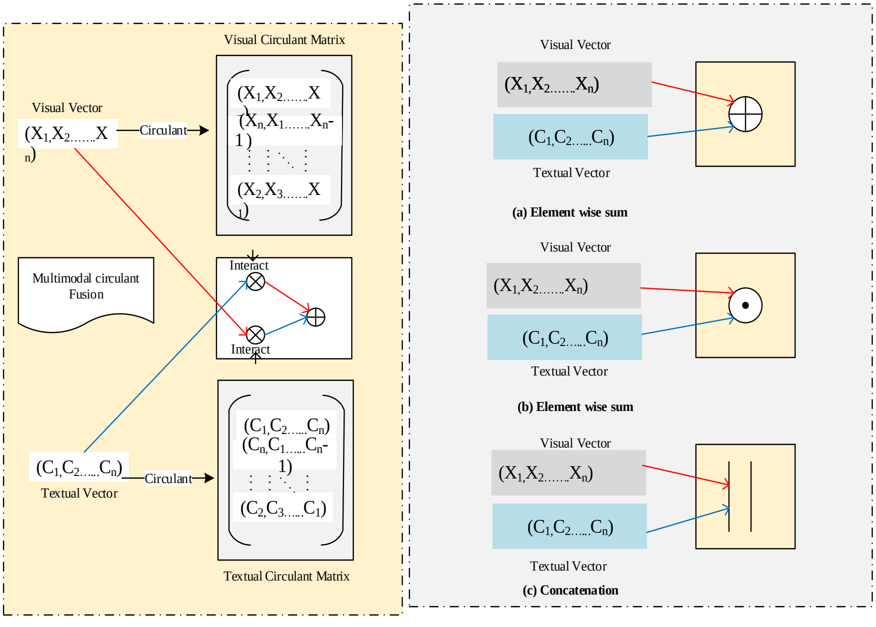


Figure 2: Flow diagram for text image restoration

The BERT and M3-UNET models handle the features of content text, headline text, and images, respectively, during the feature extraction phase. Textual data is processed using the BERT model. It excels in text comprehension because to its bidirectional processing capability, which allows it to pick up on nuanced semantic variations seen in news items' contexts. Using its deep residual network design, the M3-UNET model efficiently finds and interprets visual information pertaining to news when applied to picture data. Very helpful for complicated feature extraction from pictures, its structure enables it to participate in deep learning without training challenges. The flow diagram shows three fusion strategies—early, joint, and late fusion—that attempt to combine text and picture characteristics after feature processing. The last step in picture restoration is to merge these integrated characteristics.



(a) Element wise sum

(b) Element wise sum

(c) Concatenation

(d) Mulitmodal Circulant Fusion

Figure 3: Modified U-Net Model architecture

3.2 Data preprocessing

3.2.1 Data cleaning

This research begins by extracting the information. Then, it uses English tokenization to clean the data by eliminating noise, punctuation marks, graphics, and hyperlinks. Next, stop word removal is applied to the data. Lastly, a spell check is run to fix any misspelled English keywords as well as replace them with the right ones. This ensures that the data feature extraction process is done accurately in terms of interpretation and judgment.

3.3 Feature extraction

Word2vec and Doc2vec

Two text feature extraction approaches are used in this study: Word2vec and Doc2vec. Word2vec is used for word embeddings and word vectors, while Paragraph2vec is used for sentence embeddings and paragraph vectors. Word2Vec captures the semantic associations between words in context by transforming them into vector space. By mapping each phrase to a vector space where similarly-themed sentences are grouped together, Doc2vec depicts sentence characteristics in a manner analogous to how vectors represent individual words. As opposed to Word2Vec's word-level analysis, Doc2Vec allows the research to examine the text at the document level, which incorporates a larger context. By assessing the document's presentation or discussion of subjects, this approach enables the machine to extract the main points of whole articles, which helps in detecting false news. With every piece of text, the popular natural language processing (NLP) package Gensim extracts 50 Word2vec features and 50 Doc2vec features.

BERT feature vectors

This research yields 768-dimensional BERT Base feature vectors. There are two versions of BERT; the one utilized for this study is the Base version, which has 768 hidden states, and the other is the large version, that has 1024 hidden states. The hidden state's size is 768, which means that every token is converted into a vector with 768 dimensions. Token count, hidden state size (768), along with layer count (12) make up the three dimensions of a tensor that is the product of all layer outputs.

Image feature vectors

Computer vision differs from human vision in how it processes visual information. Images are seen by computers as a combination of numerical values, in contrast to humans who can directly perceive form and color. What this means is that images may be thought of as structured collections of numerical data. The picture model used in this research makes use of M3-UNET for feature extraction. M3-UNET has been a popular model for a while now because to its modest size and great performance. Using the 1000-dimensional vector extracted from M3-UNET's last convolutional layer as picture features, the study.

Fusion methods

In order to make the model more accurate and diverse, this research merges text and visual characteristics. The three techniques of fusion are joint, late, and early. Here we will present these three techniques.

3.4 M3-U-Net model

In order to make the approach more flexible for real blurry photos, this research suggests a multi-scale altered U-Net image deblurring system that uses dilated convolution to return a clear image from a dynamically blurred one. By making use of dilated convolution, multi-scale architecture is able to learn the qualities of pictures at various scales, take use of the receptive field's benefits, and extract more information from attributes with less computational cost.

Incorrect camera settings or faulty hardware are common causes of digital photography's flawed, low-quality results. Dynamic blurring, which happens when you take a still shot of an item in motion or when the camera shakes, is the most common kind of visual flaw. on clarify, the image deblurring approach states that a blurred picture is just a convolutional operation applied on a crisp image using a blur kernel.

$$\mathrm{B} = \mathrm{K} \otimes \mathrm{I} + \mathrm{N} \qquad (1)$$

⊎ signifies the convolution operation, N stands for noise, l represents the initial sharp image, K stands for the blur kernel, and B represents the blurred image. The blur kernel, which explains how pixels diffuse along a moving trajectory, is another name for the point-spread function. To restore clarity to hazy photos is the main goal of picture deblurring techniques. Blurred pictures and blurred kernels are used in image deconvolution to recover crisp images. There are two main types of deconvolution algorithms now in use: blind and non-blind. While the latter's blur kernel is known, the former's is unknown. Typically, the results are pictures with a lot of blurs. Nevertheless, we don't sure what the blurred kernels yet crisp pictures are. In order to use the conventional approach to restore a blurred picture to a clear one, it is crucial to precisely estimate the blur kernel. Nevertheless, we still don't know whether the estimated blur kernel deconvolution of a blurry picture yields a proper solution for the approximation crisp image. To fix blurry pictures after blind deconvolution—an ill-posed problem—more data is needed.

One hotspot for study is the topic of picture deblurring methods in dynamic settings. In the realm of picture deblurring techniques, deep learning algorithms has recently attracted a lot of attention. Unfortunately, these approaches often have poor deblurring results and insufficient receptive fields due to a lack of natural connections across different hierarchical levels. A more flexible method is used in U-Net, which successfully integrates characteristics of varying degrees. This solution keeps the total amount of parameters within an acceptable range while also drastically reducing their complexity. The current work presented an enhanced U-Net model to increase the picture deblurring effect, based on these benefits.

One common use of encoder-decoder networks in computer vision is picture deblurring, where they have shown to be effective. The network's architecture is a convolutional neural network (CNN) with symmetrical encoder and decoder components. The input picture is first downsampled by the encoder into a smaller feature map having more channels and rich details, and then the decoder takes that smaller map and upsamples it into a bigger one having fewer channels and deeper details. The feature fusion-primarily skip connection, or residual connection, was introduced by Ronneberger et al. between the decoder and encoder networks. The amount of feature data lost during downsampling increases dramatically as the network depth increases. The feature map that is obtained by encoder downsampling bypasses the multilayer system and links directly to the decoder correspondence to guarantee that the last feature map has enough precise information. In order to get feature maps with better information, the corresponding decoder upsampling uses a combination of shallow and deep features. Figure 1 shows a network topology that is known as a U-Net because of its resemblance to a U-shape. Image convolution computation (blue arrow), skip connection (gray arrow), max pooling (red arrow), and deconvolution (green arrow) are shown, respectively. The max pooling approach is used by the U-Net downsampling technique. This method splits the picture into many 2 × 2 rectangular sections, calculates the highest value for every area, and then uses this information to decrease the quantity of image data while keeping crucial details.

Loss function

We studied multi-scale picture deblurring networks, which included calculating the weight as well as the loss between each scale's output and target images. This article primarily discusses Ye et al.'s lifting-scale iteration architecture, which differs from existing approaches. In this design, each scale iteration is considered an independent deblurring subtask; this means that it continues to have an impact even after the training phase stops. Consequently, the mean absolute error (MAE) is chosen as the loss function, and the sole metric computed is the difference between the target picture and the final deblurred image. L_1 loss is another name for the MAE.

$$\mathrm{Loss} = \frac{\mathrm{L_i} - \mathrm{G_i}}{\mathrm{N_i}} \qquad (2)$$

where $\mathrm{L_i}$ represents the crisp picture, $\mathrm{G_i}$ stands for the deblurred image, and $\mathrm{N_i}$ signifies the quantity of components in $\mathrm{L_i}$ that need to be normalized at the i-th scale.

4 Results & discussion

4.1 Experimental environment

(a). Hardware environment

Powered through six Nvidia GeForce RTX 3090 24-GB GPUs, the experimental system made use of a powerful computing server with an AMD Ryzen Threadripper 3990Y @ 3.70 GHz CPU and 1TB of RAM. This powerful hardware architecture is perfect for deep learning activities like model development and inference because to its large storage capacity and outstanding processing capability. Faster experimentation and convergence are guaranteed by the state-of-the-art hardware, which drastically shortens the training time.

(b). Software environment

The main programming language for this research was Python 3.8, with PyTorch being used for deep learning applications. A flexible and iterative development approach was made possible by Python's flexibility. At the same time, PyTorch supplied the necessary tools for creating and teaching neural networks. We enhanced experimental results by developing, optimizing, and training our models quickly using PyTorch's robust computing capabilities and its automated differentiation function.

The experimental training of the model was accelerated by training it on a GPU, which excels at computationally demanding image processing jobs. Table 2details the experimental setting and setup that were used in this investigation.

Table 2: Experimental settings

Configuration	**Experimental environment**
Windows10	Operating system
Python	Programing language
PyTorch	Deep learning framework
GTX 2080	GPU
16.0 GB	Memory size

(c). Dataset

There are 32,214 distinct scenarios included in the GoPro collection, including both clear and fuzzy images. Two thousand three picture pairings made up the training dataset, whereas eleven thousand eleven picture pairs comprised the test dataset. We enhanced the data from the GoPro along with Real Blur training sets using data improvement methods to make the model more generalizable. Two of the actions were adding Gaussian noise and randomly rotating the data. The data augmentation specifically included random rotations of 90, 180, while 270 degrees, as well as horizontal (left to right) while vertical (upside down) flips. Also included was some Gaussian noise, which had a mean of 0 and a standard deviation of 0.0001. This led to an increase of 1,412 picture pairings in the GoPro training dataset and 1,5032 in the Real Blur training dataset, both achieved by means of these augmentation approaches.

Divide the Data into Three Parts: Training, Validation, and Testing Set of data: In order to train our algorithm, we use the training set to experiment with different hyperparameters. Then, we test it on the validation set and choose the hyperparameters that provide the best results. Selecting hyperparameters in this manner is recommended. The "fully convolutional network" is the ancestor of the U-Net design. Substituting upsampling operators for pooling operations in subsequent levels of a conventional contracting network is the key concept. Therefore, the output's resolution is enhanced by these layers. For the vast majority of customers, this is the reality. Only a small number of cloud providers have internal, on-demand access to massive, homogenous collections of lightning-fast hardware. In the first scenario, optimising hyperparameter searches for hardware efficiency requires tweaks to the search method. It is still possible to decide on a degree of parallelism to use throughout the search even in the second situation, for sets of homogenous systems. Machine learning algorithms of all stripes use hyperparameters, and the amount of these parameters, which may span both numerical and categorical domains, varies from algorithm to algorithm. For this optimization issue, many hyperparameter search techniques have been suggested, from the most basic, like random search, to the most complex, including methods like Probabilistic optimization, gradient-based learning, and bandit-based searches. Both of these methods simplify things by assuming certain things, but they do solve the issue of selecting the next parametrization to test: 1) all hyperparameters are treated similarly throughout the search process, and 2) all search spaces for hyperparameters are equally complicated. Both of these hypotheses are usually not true in real life. Because their suitability to the learning goal is not known with confidence, models with a bigger priority are selected in the search, similar to complexity. The model's performance under alternative parameterizations could be better understood by doing a more comprehensive exploration of the hyperparameter spaces, as this lack of assurance suggests. Since it is impossible to determine how well a model performs throughout the whole hyperparameter domain, it is reasonable to keep searching for low-priority models, but to reduce the total number of searches overall.

Images from the training datasets were arbitrarily cropped to a resolution of 256 × 256 pixels to avoid model overfitting. With an initial learning rate of 1e-4 and a half-life adjustment every 1000 rounds, the training time was determined to 3000 rounds. Additionally, 10 was the batch size. Adam, with parameters A1 = 0.9 as well as A2 = 0.999, was chosen as the network optimization algorithm.

4.2 Experimental datasets

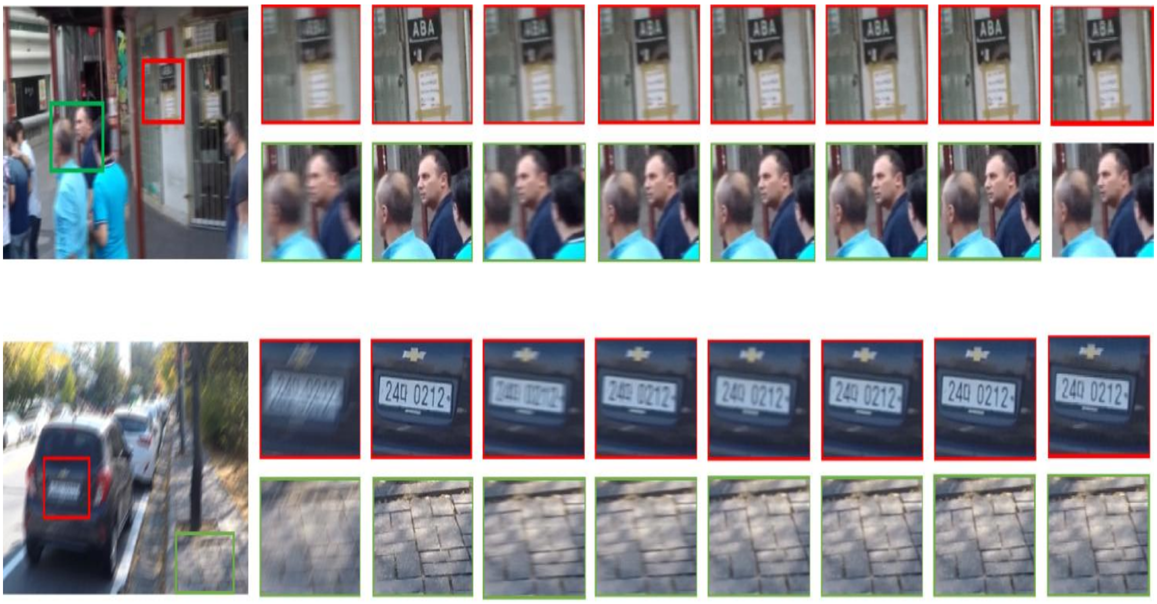
We competed tests on two separate datasets to ensure our model was completely validated. We may train the model in several ways using the data provided by these datasets, and we can assess the performance of the combined approach of visual recognition with language processing for graphic recognition using these datasets.

Table 3: Comparison of experimental results of dilation rate strategy

Strategies	**D = 3**	**D = 4**	**D = 5**	**D = [3,4,5]**
PSNR	**30.33**	29.77	29.88	29.99
SSIM	**0.877**	0.822	0.826	0.723
LPIPS	**0.9**	0.91	0.92	0.914
FID	**30**	32	31.2	31.6

PSNR: greatest signal-to-noise ratio; SSIM: structure-similarity index measurement; LPIPS (Learned Perceptual Image Patch Similarity); FID (Fréchet Inception Distance)

The data is superior to other data sets, as shown by the bold language.



(a), (b), (c), (d), (e), (f), (g), (h)

Figure 5: Comparison of visual effects on the dataset (a), SVM (b), LSTM (c), CNN (d), CLIP (e), VISTANET (f), MMFL (g), U-Net (h). Proposed Model

4.3 Discussion

The synthesizing advice will continually represent the relevant degradation pattern, conditional on the deteriorated text word embedding. Here, we provide a fresh viewpoint: bolstering picture repair with restoration in textual space. To further improve performance, we merge the benefits of text-based prior (which learns degradations at the textual level) with image-based prior (which provides clean guidance).In this article, we provide an innovative way of looking at image restoration. Since content and degradation are associated in images but not in texts, we propose a plug-and-play method that takes degraded images and converts them into text, then eliminates textual deterioration details to get restored text. As opposed to doing reconstruction on the image level, we suggest restoring on the textual level and then using the restored text to aid image restoration. On GoPro test data, our method improves PSNR by 0.35 dB. Figure 5 shows the visual outcomes of our method's restoration efforts, which demonstrate its success by restoring pictures with crisper borders and features. Table 3 shows that when compared to current approaches, our method improves single-image defocus deblurring by +0.35 dB and dual-pixel defocus deblurring by +0.45 dB in terms of PSNR. Figure 5 illustrates that our method's forecast has a more organized structure.

5 Conclusion

In this study, we successfully employ text information to help with picture restoration since text input is more easily accessible and gives information with more flexibility. We apply a M3U-NET-based model and create a simple and effective framework in order to develop a text-based picture restoration approach. This framework enables the user to enter text and get the appropriate image restoration results. The framework uses M3U-NET text-image feature compatibility to improve the combination of picture and text features. Our system is capable of performing a variety of picture restoration tasks, such as image in painting, image super-resolution, and image colorization.

References

[1] Jha, S., Mayer, E., & Barahona, M. (2022, December). Improving information fusion on multimodal clinical data in classification settings. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)* (pp. 154-159). https://doi.org/10.18653/v1/2022.louhi-1.18

[2] Luo, W., Xia, Y., Tianshu, S., & Li, S. (2024, October). Shapley Value-based Contrastive Alignment for Multimodal Information Extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 5270-5279). https://doi.org/10.1145/3664647.3681367

[3] Wajid, M. A., Zafar, A., & Wajid, M. S. (2024). A deep learning approach for image and text classification using neutrosophy. *International Journal of Information Technology*, *16*(2), 853-859. https://doi.org/10.1007/s41870-023-01529-8

[4] Al-Tameemi, I. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2023). Multi-model fusion framework using deep learning for

visual-textual sentiment classification. *Computers, Materials & Continua*, *76*(2), 2145-2177. https://doi.org/10.32604/cmc.2023.040997

[5] Liu, J., Ma, X., Wang, L., & Pei, L. (2024). How Can Generative Artificial Intelligence Techniques Facilitate Intelligent Research into Ancient Books?. *ACM Journal on Computing and Cultural Heritage*, *17*(4), 1-20. https://doi.org/10.1145/3690391

[6] Kumar, P., Malik, S., Raman, B., & Li, X. (2022). VISTANet: VIsual Spoken Textual Additive Net for Interpretable Multimodal Emotion Recognition. *arXiv preprint arXiv:2208.11450*. https://doi.org/10.48550/arXiv.2208.11450

[7] Zong, D., Ding, C., Li, B., Zhou, D., Li, J., Zheng, K., & Zhou, Q. (2023, October). Building robust multimodal sentiment recognition via a simple yet effective multimodal transformer. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9596-9600). https://doi.org/10.1145/3581783.3612872

[8] Ouafa, C., & Tayeb, L. M. (2022). Facial Expression Recognition Using Convolution Neural Network Fusion and Texture Descriptors Representation. *International Journal of Computational Intelligence and Applications, 21*(01), 2250002. https://doi.org/10.1142/s146902682250002x

[9] Chandrasekaran, G., Nguyen, T. N., & Hemanth D, J. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1415. https://doi.org/10.1002/widm.1415

[10] Kaur, P., Malhi, A., & Pannu, H. (2024). Annotate and retrieve in vivo images using hybrid self-organizing map. *The Visual Computer*, *40*(8), 5619-5638. https://doi.org/10.1007/s00371-023-03126-z

[11] Picha, S. G., Chanti, D. A., & Caplier, A. (2024). Semantic textual similarity assessment in chest x-ray reports using a domain-specific cosine-based metric. *arXiv preprint arXiv:2402.11908*. https://doi.org/10.48550/arXiv.2402.11908

[12] Akhmerov, A. K., Vasilev, A. S., & Vasileva, A. V. (2019, June). Research of spatial alignment techniques for multimodal image fusion. In *Multimodal Sensing: Technologies and Applications* (Vol. 11059, pp. 309-317). SPIE. https://doi.org/10.1117/12.2526030

[13] Leonardo, R., Hu, A., Uzair, M., Lu, Q., Fu, I., Nishiyama, K., ... & Ravichandran, D. (2019, December). Fusing visual and textual information to determine content safety. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 2026-2031). IEEE. DOI: 10.1109/ICMLA.2019.00324

[14] Meng, L., Tan, A. H., Wunsch II, D. C., Meng, L., Tan, A. H., & Wunsch II, D. C. (2019). Socially-Enriched Multimedia Data Co-clustering. *Adaptive Resonance Theory in Social Media Data Clustering: Roles, Methodologies, and Applications*, 111-135. https://doi.org/10.1007/978-3-030-02985-2_5

[15] Li, Z., Zhang, D., Du, Y., & Zhang, X. (2024). A Study on the Application of Multimodal Fusion Technology in the Translation of the Historical Literature of Geng Lu Bu. *Applied Mathematics and Nonlinear Sciences.* https://doi.org/10.2478/amns-2024-3626

[16] Song, Y., Lin, N., Li, L., & Jiang, S. (2024, May). A Vision Enhanced Framework for Indonesian Multimodal Abstractive Text-Image Summarization. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 61-66). IEEE. https://doi.org/10.1109/cscwd61410.2024.10580245

[17] Rizmolp, G. (2020). Detection of Abnormalities Using Multimodal Image Fusion and Text Encryption. | ISSN: 2320-2882

[18] Tiwari, M., Khare, P., Saha, I., & Mali, M. (2024). Multimodal NLP for image captioning : Fusing text and image modalities for accurate and informative descriptions. *Journal of Information and Optimization Scienc*es. https://doi.org/10.47974/jios-1626

[19] Wajid, M.A., Zafar, A., Terashima-Marín, H., & Wajid, M.S. (2023). Neutrosophic-CNN-based image and text fusion for multimodal classification. *J. Intell. Fuzzy Syst.*, 45, 1039-1055. https://doi.org/10.3233/JIFS-223752

[20] Li, Z., Yang, M., Cheng, L., & Jia, X. (2023). Blind Text Image Deblurring Algorithm Based on Multi-Scale Fusion and Sparse Priors. *IEEE Access*, 11, 16042-16055. DOI: 10.1109/ACCESS.2023.3245150

[21] Thangavel, K., Palanisamy, N., Muthusamy, S., Mishra, O.P., Sundararajan, S.C., Panchal, H.D., Loganathan, A.K., & Ramamoorthi, P. (2023). A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. *Soft Computing*, 27, 14205-14218. https://doi.org/10.1007/s00500-023-08448-7

[22] Lin, Q., Yan, B., Li, J., & Tan, W. (2020). MMFL: Multimodal Fusion Learning for Text-Guided Image Inpainting. *Proceedings of the 28th ACM International Conference on Multimedia.* https://doi.org/10.1145/3394171.3413982

[23] Zhou, L. (2024). Research on Image Translation Problems Based on Multimodal Data Set Fusion. *International Journal of Computer Science and Information*

Technology. https://doi.org/10.62051/ijcsit.v3n3.03

[24] Kota, S.M., Haridasan, S., Rattani, A., Bowen, A., Rimmington, G.M., & Dutta, A. (2022). Multimodal Combination of Text and Image Tweets for Disaster Response Assessment. *D2R2*. https://soar.wichita.edu/handle/10057/25302

[25] Luo, X., Jia, N., Ouyang, E., & Fang, Z. (2024). Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal*. https://doi.org/10.1002/smj.3597

[26] Li, J., Bai, X., & Han, Z. (2024). DGFN Multimodal Emotion Analysis Model Based on Dynamic Graph Fusion Network. *International Journal of Decision Support System Technology*. https://doi.org/10.4018/ijdsst.352417

[27] Ding, M., Ma, Y., Qin, P., Wu, J., Li, Y., & Nie, L. (2024). *RA-BLIP: Multimodal Adaptive Retrieval-Augmented Bootstrapping Language-Image Pre-training*. ArXiv, abs/2410.14154. https://doi.org/10.48550/arXiv.2410.14154

[28] Xie, X., Cui, Y., Ieong, C., Tan, T., Zhang, X., Zheng, X., & Yu, Z. (2024). *FusionMamba: Dynamic Feature Enhancement for Multimodal Image Fusion with Mamba*. ArXiv, abs/2404.09498. https://doi.org/10.1007/s44267-024-00072-9

[29] Lu, X., Ni, Y., & Ding, Z. (2024). Cross-Modal Sentiment Analysis Based on CLIP Image-Text Attention Interaction. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2024.0150290

[30] Al-Tameemi, I.K., Feizi-Derakhshi, M., Pashazadeh, S., & Asadpour, M. (2023). Interpretable Multimodal Sentiment Classification Using Deep Multi-View Attentive Network of Image and Text Data. *IEEE Access*, 11, 91060-91081. https://doi.org/10.1109/access.2023.3307716