iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

CONVOLUTIONAL NEURAL NETWORKS FOR REAL VS. SYNTHETIC IMAGE DETECTION USING CIFAKE

R. Mabubasha
Assistant Professor, Department of Computer Science and Engineering, R.V.R& J.C College of Engineering, chowdavaram, Guntur, Andhrapradesh.
Dasari Nagalakshmi
Gurram Sai Krishina Sri
Ande Beaula
UG Students, Department of Computer Science and Engineering, R.V.R& J.C College of Engineering, chowdavaram, Guntur, Andhrapradesh

ABSTRACT
Recent advances in synthetic data have enabled the generation of images with such high quality that human beings cannot distinguish the difference between real-life photographs and Artificial Intelligence (AI) generated images. Given the critical necessity of data reliability and authentication, this article proposes to enhance our ability to recognize AI-generated images through computer vision. Initially, a synthetic dataset is generated that mirrors the ten classes of the already available CIFAR-10 dataset with latent diffusion, providing a contrasting set of images for comparison to real photographs. The model is capable of generating complex visual attributes, such as photorealistic reflections in water. The two sets of data present as a binary classification problem with regard to whether the photograph is real or generated by AI. This study then proposes the use of a Convolutional Neural Network (CNN) to classify the images into two categories; Real or Fake. Following hyperparameter tuning and the training of 36 individual network topologies, the optimal approach could correctly classify the images with 92.98% accuracy. Finally, this study implements explainable AI via Gradient Class Activation Mapping to explore which features within the images are useful for classification. Interpretation reveals interesting concepts within the image, in particular, noting that the actual entity itself does not hold useful information for classification; instead, the model focuses on small visual imperfections in the background of the images. The complete dataset engineered for this study, referred to as the CIFAKE dataset, is made publicly available to the research community for future work.

Keywords:
Convolutional Neural Network (CNN), Image Classification, Generative AI.

I.INTRODUCTION

The field of synthetic image generation by Artificial Intelligence (AI) has developed rapidly in recent years, and the ability to detect AI-generated photos has also become a critical necessity to ensure the authenticity of image data. Within recent memory, generative technology often produced images with major visual defects that were noticeable to the human eye, but now we are faced with the possibility of AI models generating high-fidelity and photorealistic images in a matter of seconds. The AI-generated images are now at the quality level needed to compete with humans and win art competitions [1]. Latent Diffusion Models (LDMs), a type of generative model, have emerged as a powerful tool to generate synthetic imagery [2]. These recent developments have caused a paradigm shift in our understanding of creativity, authenticity and truth. This has led to a situation where consumer-level technology is available that could quite easily be used for the violation of privacy and to commit fraud. These philosophical and societal implications are at the forefront of the current state of the art, raising fundamental questions about the nature of trustworthiness and reality. Recent technological advances have enabled the generation of images with such high quality that human beings cannot tell the difference between a real-life photograph and an image that is no more than a hallucination of an artificial neural network's weights and biases. Generative imagery that is indistinguishable from photographic data raises questions both ontological, those which concern the nature of being, and epistemological, surrounding the theories of methods, validity, and scope. Ontologically, given that humans cannot tell the difference between images from cameras and those generated by

AI models such as an Artificial Neural Network, in terms of digital information, what is real and what is not? The epistemological reality is that there are serious questions surrounding the reliability of human knowledge and the ethical implications that surround the misuse of these types of technology. The implications suggest that we are in growing need of a system that can aid us in the recognition of real images versus those generated by AI. This study explores the potential of using computer vision to enhance our newfound inability to recognise the difference between real photographs and those that are AI-generated. Given that there are many years worth of photographic datasets available for image classification, these provide examples for a model of real images. Following the generation of a synthetic equivalent to such data, we will then explore the output of the model before finally implementing methods of differentiation between the two types of image. There are several scientific contributions with multidisciplinary and social implications that arise from this study. First, a dataset, called CIFAKE, is generated with latent diffusion and released to the research community. The CIFAKE dataset provides a contrasting set of real and fake photographs and contains 120,000 images (60,000 images from the existing CIFAR-10 dataset (Collection of images that are commonly used to train machine learning and computer vision algorithms available from: https://www.cs.toronto.edu/ kriz/- cifar.html) and 60,000 images generated for this study), making it a valuable resource for researchers in the field. Second, this study proposes a method to improve our waning human ability to recognise AI-generated images through computer vision, using the CIFAKE dataset for classification. Finally, this study proposes the use of Explainable AI (XAI) to further our understanding of the complex processes involved in synthetic image recognition, as well as visualisation of the important features within those images. These scientific contributions provide important steps forward in addressing the modern challenges posed by rapid developments of modern technology and have important implications for ensuring the authenticity and trustworthiness of data.

II.LITERATURE SURVEY

2.1 High-Resolution Image Synthesis with Latent Diffusion Models

AUTHORS: Ludwig Maximilian University of Munich & IWR, HeidelbergUniversity.

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state of the art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

2.2 Predicting image credibility in fake news over social media using multi-modal approach

AUTHORS: Nisha Raichur, Nidhi Lonakadi, Priyanka Mural

Social media are the main contributors to spreading fake images. Fake images are manipulated images altered through software or by other means to change the information they convey. Fake images propagated over microblogging platforms generate misrepresentation and stimulate polarization in the people. Detection of fake images shared over social platforms is extremely critical to mitigating its spread. Fake images are often associated with textual data. Hence, a multi-modal framework is employed utilizing visual and textual feature learning. However, few multi-modal frameworks are already proposed; they are further dependent on additional tasks to learn the correlation between modalities. In this paper, an efficient multi-modal approach is proposed, which detects fake images of microblogging platforms. No further additional subcomponents are required. The proposed framework utilizes explicit convolution neural network model EfficientNetB0 for images and sentence transformer for text analysis. The feature embedding from visual and text is passed through dense layers and later fused to predict fake images. To validate the effectiveness, the proposed model is tested upon a publicly available microblogging dataset, MediaEval (Twitter) and Weibo, where the accuracy prediction of 85.3% and 81.2% is observed, respectively. The model is also verified against the newly created latest Twitter dataset containing images based on India's significant events in 2020. The experimental results illustrate that the proposed model performs better than other state-of-art multi-modal frameworks.

2.3 Writer-independent signature verification; Evaluation of robotic and generative adversarial attacks

AUTHORS : Tanev, G., Saadi, D.B., Hoppe, K., Sorensen, H.B

Forgery of a signature with the aim of deception is a serious crime. Machine learning is often employed to detect real and forged signatures. In this study, we present results which argue that robotic arms and generative models can overcome these systems and mount false-acceptance attacks. Convolutional neural networks and data augmentation strategies are tuned, producing a model of 87.12% accuracy for the verification of 2,640 human signatures. Two approaches are used to successfully attack the model with false-acceptance of forgeries. Robotic arms (Line-us and iDraw) physically copy real signatures on paper, and a conditional Generative Adversarial Network (GAN) is trained to generate signatures based on the binary class of 'genuine' and 'forged'. The 87.12% error margin is overcome by all approaches; prevalence of successful attacks is 32% for iDraw 2.0, 24% for Line-us, and 40% for the GAN. Fine-tuning with examples show that false-acceptance is preventable. We find attack success reduced by 24% for iDraw, 12% for Line-us, and 36% for the GAN. Results show exclusive behaviours between human and robotic forgers, suggesting training wholly on human forgeries can be attacked by robots, thus we argue in favour of fine-tuning systems with robotic forgeries to reduce their prevalence.

III.METHODOLOGY

This section describes the approach taken to categorize real and synthetic images based on a Convolutional Neural Network (CNN). The process consists of four major steps: data preprocessing, model design, training, and testing.It describes the acquisition of real images as well as the production of synthetic versions thereof, followed by machine learning model design for classification and incorporation of explainability methods. Section-A is an account of obtaining 60,000 real images, and Section-B is the account of synthetically creating an equivalent 60,000 images to form a complete dataset of 120,000 images. Section-C is the presentation of the machine learning model created to determine image authenticity.

3.1 REAL DATA COLLECTION

To cover the REAL class (given a positive class label of "1"), images were taken from the CIFAR-10 dataset [24]. This standard dataset consists of 60,000 RGB 32×32 pixel images, partitioned evenly into ten object classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are 6,000 images in each class, where 5,000 are assigned for training and 1,000 for testing (i.e., roughly 16.6% held out for validation). In this experiment, all 50,000 training images were utilized to train the classifier for the REAL class, and 10,000 were reserved for testing.

Sample images from the CIFAR-10 dataset utilized as real images are shown in Figure 1.

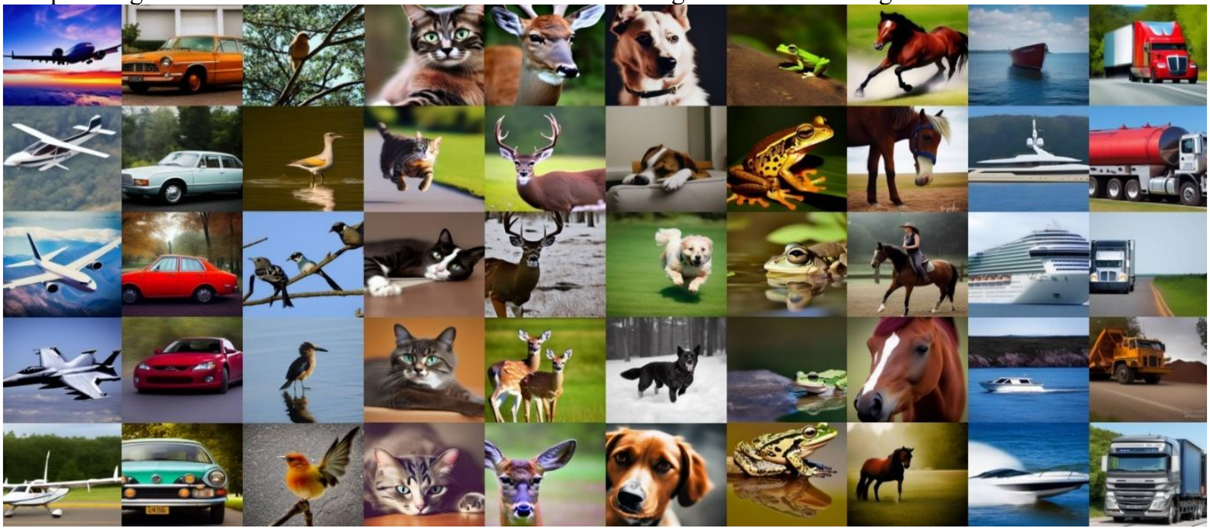

Fig 1: CIFAKE Data set with 10 Categories of images

3.2 SYNTHETIC DATA GENERATION

The synthetic images were produced by the Stable Diffusion v1.4 model created by CompVis (https://huggingface.co/CompVis/stable-diffusion-v1-4), an open-source latent diffusion model (LDM). The diffusion process mimics iteratively corrupting an image with Gaussian noise and eventually collapsing to full noise. The reverse diffusion process restores the image from noise, learning to denoise using a neural network model that is trained on denoising.

Formally, a noisy image xtx_txt at timestep ttt is synthesized from the original image x0 according to:

$$xt = \bar{\alpha}tx0 + 1 - \bar{\alpha}t\epsilon,$$

where ϵ is Gaussian noise, and $t\bar{\alpha}t$ is a noise scheduling parameter. The neural network ϵθ is learned to minimize the mean squared error (MSE) between the true and predicted noise:

$$\text{Loss}=E_{t,x0,\epsilon}[\|\epsilon-\epsilon\theta(xt,t)\|^{2}]$$

The model goes through 50 reverse diffusion steps to produce a final image from clean noise. Stable Diffusion v1.4 is pretrained on a mix of high-quality datasets, such as LAION2B-en, LAION-high-resolution, and LAION-aesthetics v2.5+ [25], which are cleaned subsets of the LAION-5B dataset with more than 5.85 billion text-image pairs.

To reflect the CIFAR-10 dataset structure, 60,000 synthetic images were created over the same ten classes. Prompt engineering methods were used to increase intra-class diversity, with some prompt modifiers summarized in Table 1. Similar to the original dataset, 50,000 images were trained on and 10,000 tested on, with each synthetic image being labeled as such to be the SYNTHETIC class.

Table 1: Prompt Modifiers Used for Synthetic Image Generation

CIFAR-10 Class	**Prompt Modifier Examples**
Airplane	"a high-resolution photo of a passenger airplane", "realistic aircraft in flight", "vintage fighter jet"
Automobile	"a sports car on a highway", "SUV parked in the city", "realistic sedan front view"
Bird	"a colorful bird on a branch", "realistic tropical bird", "small bird flying in the sky"
Cat	"a domestic cat sitting indoors", "kitten on a sofa", "realistic tabby cat"
Deer	"a deer in a forest", "realistic fawn in the wild", "a buck standing in a field"
Dog	"a puppy playing in the yard", "realistic golden retriever", "dog sitting on the porch"
Frog	"a frog on a leaf", "realistic amphibian near a pond", "green frog in the jungle"
Horse	"a horse running through a field", "realistic brown stallion", "wild horse in the mountains"
Ship	"a cargo ship at sea", "realistic sailboat on water", "military ship during sunset"
Truck	"a delivery truck on a road", "realistic pickup truck", "semi-truck in motion"

All input images are resized to a standard dimension of 224 × 224 pixels to ensure uniformity across the dataset. Preprocessing is carried out with TensorFlow's ImageDataGenerator, which scales pixel intensities by a factor of 1/255 to normalize values into the range [0,1][0, 1][0,1]. Normalization helps to enhance the convergence rate of the model. The dataset is split into a training and validation subset in an 80:20 ratio. An independent test set, unseen during training, is used to assess the final model performance

3.3 CNN Architecture

The CNN model is executed through TensorFlow's Sequential API. The architecture starts with three convolutional layers with progressively larger filter sizes of 32, 64, and 128, each preceded by MaxPooling layers to decrease spatial dimensionality. The Architecture design is as Fig2.The convolution operation on an image input xxx and a kernel www can be mathematically represented as:

$$(x * w)(i,j) = m = 1\sum Mn = 1\sum Nx(i+m-1, j+n-1)w(m,n)$$

where x(i,j)x(i,j)x(i,j) is the image patch and w(m,n)w(m,n)w(m,n) is the filter matrix. The resulting feature map is activated using the Rectified Linear Unit (ReLU) function, defined as:

$$f(x) = max(0,x)$$

The output shape after convolution, assuming stride 1 and no padding, becomes:

(height−kernel_height+1, width−kernel_width+1)

Following the final convolutional block, the output is flattened into a one-dimensional vector:

$$x = [x1, x2, \dots, xL]$$

This vector is passed through a dense layer of 128 neurons with ReLU activation. The last output layer is a single neuron with a sigmoid activation function, appropriate for binary classification:

$$\sigma(x) = 1/(e - x1)$$

The sigmoid activation gives a probability score, with values closer to 0 corresponding to the FAKE class and values closer to 1 corresponding to the REAL class.

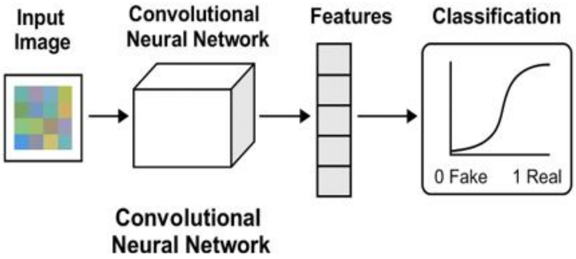

Fig 2 : CNN Model Architecture

3.4 Model Training

The model is built with the Adam optimizer because of its adaptive learning nature. The loss function used is Binary Crossentropy, which is optimal for binary classification:

$$Loss = -[ylog(y^{\wedge}) + (1 - y)log(1 - y^{\wedge})]$$

where is the true label and y^ is the estimated probability. The evaluation metric is accuracy. Training of the model is done for five epochs with both training and validation generators so that it can update weights and biases according to classification loss.

3.5 Model Evaluation

Once trained, the model is tested on the test dataset to yield the final accuracy score. Training and validation loss and accuracy are also plotted using the matplotlib library to check for any overfitting or underfitting trends. The trained model is stored in HDF5 format as CNN_model.h5 to Google Drive to enable future inference or deployment without having to retrain.

IV. RESULTS

The convolutional neural network model was trained for five iterations using a dataset composed of real and synthetic images. Model performance was compared to accuracy, loss, precision, recall, and F1-score metrics. During training, the model displayed rapid convergence with 100% validation accuracy as early as the second epoch and maintained this throughout the rest of the epochs. Concurrently, validation loss decreased appreciably to a minimum of 9.44×10^{-7}, which reflected good learning without any overfitting (Table 3). Precision, recall, and F1-scores were likewise maintained at 1.0000 throughout all the validation epochs, as illustrated in Tables 1, 2, and 3, respectively.

On final evaluation, the model achieved training accuracy of 91.94% and test accuracy of 91.64% (Table 7) with great generalization to unseen data.A finer-grained classification report (Tables 8–10) again supported the performance of the model. The CNN's accuracy for FAKE images and REAL images were 0.9950 and 1.0000, respectively, while the recall scores were their reciprocal, 1.0000 and 0.9950, for FAKE and REAL, respectively. The F1-score achieved for both classes was 0.9975, indicating that the model was performing equally on both classes.

Overall, the model achieved a total accuracy of 99.75%, and macro and weighted averages of 0.9975 for all the significant metrics. These results strongly indicate the efficacy of the CNN in real vs. synthetic image classification tasks, and hence it can be a reliable tool for applications requiring image authenticity verification.

Table 1: Observed validation precision for the filters within the convolutional neural network.

Class	Precision
FAKE	0.9950
REAL	1.0000
Macro Average	0.9975
Weighted Average	0.9975

Table 2: Observed validation Recall for the filters within the convolutional neural network.

Class	Precision
FAKE	0.9950
REAL	1.0000
Macro Average	0.9975
Weighted Average	0.9975

Table 3: Observed validation F1 Scorefor the filters within theconvolutional neural network.

Class	F1-Score
FAKE	0.9950
REAL	1.0000
Macro Average	0.9975
Weighted Average	0.9975

V. CONCLUSION AND FURTHER ENHANCEMENT

This study has proposed a method to improve our waning ability to recognise AI-generated images through the use of Computer Vision and to provide insight into predictions with visual cues. To achieve this, this study proposed the generation of a synthetic dataset with Latent Diffusion, recognition with Convolutional Neural Networks, and interpretation through Gradient Class Activation Mapping. The results showed that the synthetic images were high quality and featured complex visual attributes, and that binary classification could be achieved with around 92.98% accuracy. Grad-CAM interpretation revealed interesting concepts within the images that were useful for predictions. In addition to the method proposed in this study, a significant contribution is made through the release of the CIFAKE dataset. The dataset contains a total of 120, 000 images (60, 000 real images from CIFAR-10 and 60,000 synthetic images generated for this study). The CIFAKE dataset provides the research community with a valuable resource for future work on the social problems faced by AI-generated imagery. The dataset provides a significant expansion of the resource availability for the development and testing of applied computer vision approaches to this problem. The reality of AI generating images that are indistinguishable from real-life photographic images raises fundamental questions about the limits of human perception, and thus this study proposed to enhance that ability by fighting fire with fire. The proposed approach addresses the challenges of ensuring the authenticity and trustworthiness of visual data.Future work could involve exploring other techniques to classify the provided dataset. For example, the implementation of attention-based approaches is a promising new field that could provide increased ability and an alternative method of explainable AI. Furthermore, with even further improvements to synthetic imagery in the future, it is important to consider updating the dataset with images generated by these approaches. Furthermore, considering generating images from other domains, such as human faces and clinical scans, would provide additional datasets for this type of study and expand the applicability of our proposed approach to other fields of research.

VI.REFERENCES

[1] K. Roose, “An AI-generated picture won an art prize. Artists aren’t happy,” New York Times, vol. 2, p. 2022, Sep. 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.
[3] G. Pennycook and D. G. Rand, "The psychology of fake news," Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.
[4] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.
[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of Benford's law to detect GAN-generated images," in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.
[6] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.
[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system," KSII Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, "Writer-independent signature verification; evaluation of robotic and generative adversarial attacks," Inf. Sci., vol. 633, pp. 170–181, Jul. 2023.
[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in Proc. Int. Conf. Mach. Learn., 2021, pp. 8821–8831.
[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic textto-image diffusion models with deep language understanding," 2022, arXiv:2205.11487.
[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," 2022, arXiv:2210.04133.
[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023, arXiv:2301.11757.
[13] F. Schneider, "ArchiSound: Audio generation with diffusion," M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
[14] D. Yi, C. Guo, and T. Bai, "Exploring painting synthesis with diffusion models," in Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI), Jul. 2021, pp. 332–335.
[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, "ArtVerse: A paradigm for parallel human–machine collaborative painting creation in Metaverses," IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 4, pp. 2200–2208, Apr. 2023.
[16] Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, "ArtVerse: A paradigm for parallel human–machine collaborative painting creation in Metaverses," IEEE Trans. Syst., Man, Cybern.,Syst.,vol.53, no.4, pp. 2200–2208, Apr. 2023.
[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," 2022, arXiv:2211.00680.
[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct.2019,pp.1205–1207.
[19] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Nov. 2018, pp. 1–6.
[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for Deepfake detection," in Proc. Int. Conf. Multimedia Retr., Jun.2022,pp.615–623.
[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid CNNLSTM model for video Deepfake detection by leveraging optical flow features," in Proc. Int. Joint Conf. Neural Netw.(IJCNN),Jul.2022,pp.1–7.
[22] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," Signal Process., vol. 174, Sep. 2020,Art.no.107616.
[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, "Can people identify original and manipulated photos of real-world scenes?" Cognit. Res., Princ. Implications, vol. 2, no. 1, pp. 1–21, Dec. 2017. [24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.