computers

MDPI

Article

Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks

Shourya Tyagi [1], Olukayode Ayodele Oki [2], Vineet Verma [1], Swati Gupta [3], Meenu Vijarania [3,*], Joseph Bamidele Awotunde [4,*] and Abdulrauph Olanrewaju Babatunde [4]

[1] Department of Computer Science & Engineering, School of Engineering and Technology, K. R. Mangalam University, Gurugram 122103, India; shouryatyagi222@gmail.com (S.T.); vinni200209@gmail.com (V.V.)
[2] Information Technology Department, Walter Sisulu University, Mthatha 5117, South Africa; ooki@wsu.ac.za
[3] Department of Computer Science & Engineering, Member of Centre of Excellence AI, School of Engineering and Technology, K. R. Mangalam University, Gurugram 122103, India; swattiguptta@gmail.com
[4] Department of Computer Science, Faculty of Information and Communication Sciences, University of Ilorin, Ilorin 240003, Nigeria; babatunde.ao@unilorin.edu.ng
* Correspondence: meenuhans.83@gmail.com (M.V.); awotunde.jb@unilorin.edu.ng (J.B.A.)

Abstract: In this paper, we propose a novel method for producing image captions through the utilization of Generative Adversarial Networks (GANs) and Vision Transformers (ViTs) using our proposed Image Captioning Utilizing Transformer and GAN (ICTGAN) model. Here we use the efficient representation learning of the ViTs to improve the realistic image production of the GAN. Using textual features from the LSTM-based language model, our proposed model combines salient information extracted from images using ViTs. This merging of features is made possible using a self-attention mechanism, which enables the model to efficiently take in and process data from both textual and visual sources using the self-attention properties of the self-attention mechanism. We perform various tests on the MS COCO dataset as well as the Flickr30k dataset, which are popular benchmark datasets for image-captioning tasks, to verify the effectiveness of our proposed model. The outcomes represent that, on this dataset, our algorithm outperforms other approaches in terms of relevance, diversity, and caption quality. With this, our model is robust to changes in the content and style of the images, demonstrating its excellent generalization skills. We also explain the benefits of our method, which include better visual–textual alignment, better caption coherence, and better handling of complicated scenarios. All things considered, our work represents a significant step forward in the field of picture caption creation, offering a complete solution that leverages the complementary advantages of GANs and ViT-based self-attention models. This work pushes the limits of what is currently possible in image caption generation, creating a new standard in the industry.

Keywords: image caption generation; vision transformer; generative adversarial networks; multi-head self-attention model; MS COCO

check for updates

Citation: Tyagi, S.; Oki, O.A.; Verma, V.; Gupta, S.; Vijarania, M.; Awotunde, J.B.; Babatunde, A.O. Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks. *Computers* **2024**, *13*, 305. https://doi.org/10.3390/computers13120305

Academic Editor: Lucia Maddalena

Received: 11 October 2024
Revised: 11 November 2024
Accepted: 18 November 2024
Published: 22 November 2024

1. Introduction

The intersection of computer vision and natural language processing is the complex task of producing descriptive text from visual inputs, or picture captioning. Significant advancements in this field have come from the combination of image processing and natural language comprehension algorithms. A crucial component of visual comprehension is image captioning, which entails comprehending images and providing a natural language description of them. Recent developments in deep learning have led to notable advancements in this field, as well as many others related to machine learning. These advancements include larger, more intricate datasets, quicker hardware, particularly in the form of graphic processing units, and enhanced algorithms. The study of artificial intelligence is presented with a plethora of opportunities and challenges when computer vision and natural language processing are combined. There are several reasons why image captioning is challenging. It entails converting one's comprehension of the connections between the elements in a picture to convey it into natural language. This method increases the level of difficulty by necessitating the understanding of complex processes and the semantic information extracted from images. Furthermore, captioning objects in a picture requires knowledge of their relationships just as much as simple object recognition. Because of this, captioning pictures with artificial intelligence presents some difficulties. Today, encoder–decoder systems are commonly found in modern architectures, even though a wide range of models and techniques have been studied. Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) [1], which have been trained to extract features from pictures or objects, often manage the encoding inside these frameworks.

Artificial Intelligence-Generated Content (AIGC) has rapidly evolved, and image caption generation is among its most promising applications, merging computer vision with natural language processing. At its core, image caption generation involves creating a natural language description of a given image. This task requires an advanced understanding of both visual elements and linguistic representation, making it a challenging yet fertile ground for artificial intelligence research. Recent advancements have been heavily influenced by deep learning architectures, particularly Transformer models and large language models (LLMs), which have redefined AIGC for image captioning.

One of the main research directions in AIGC for image caption generation is the integration of multi-modal Transformer-based architectures, like the Vision Transformer (ViT) and the CLIP (Contrastive Language-Image Pre-training) model. These models have shown a remarkable ability to capture semantic relationships between images and textual descriptions, allowing them to produce captions that are not only relevant but also contextually nuanced. CLIP, developed by OpenAI, has been instrumental in training AI to understand visual content through the lens of natural language. By using large-scale datasets of image–text pairs, CLIP enables the model to generalize better across diverse contexts, making it adept at producing accurate captions even for unfamiliar visual content.

Another significant development is the use of large language models, such as GPT-4, in image-captioning tasks. These LLMs can generate richer and more descriptive captions by leveraging their extensive pre-training on language and knowledge about the world. Researchers are now exploring hybrid models that combine LLMs with visual encoders, allowing for captions that are not only descriptive but also imbued with context and subtle inferences, resulting in a more human-like output.

Further advancements have been observed in the customization of these AIGC models for specific applications, like accessibility for visually impaired users, e-commerce, and social media content. The models are being fine-tuned to recognize domain-specific objects and attributes, improving the relevance and accuracy of generated captions. However, challenges persist, particularly in dealing with complex scenes or abstract concepts, where human evaluation and intervention are still required.

Subsequently, the methodology section delves into the specifics of the datasets utilized for training and evaluation, along with the meticulous process of data preparation. Working with benchmark datasets such as Flickr30k and MS COCO requires careful data preparation since they provide extensive coverage of a variety of visual and contextual contexts, which allows models to handle a large number of edge cases efficiently. More than 330,000 photos with thorough, multi-caption annotations are available in MS COCO's extensive collection, which covers intricate scenarios with a wide range of item connections and backdrops and Flickr30k focuses on human interactions and nuanced activities, promoting adaptability to varied linguistic expressions. Together, they ensure robust model performance for real-world image-captioning tasks. Following the methodology, the model architecture section offers an intricate breakdown of the proposed model's architecture, elucidating the intricate interplay between Visual Transformers (ViTs), GANs, and LSTM components within the model framework.

The training procedure section details the model's training methodology, including loss calculation and parameter optimization. The evaluation section then outlines metrics for assessing model efficacy and provides an analysis of caption quality. This structured format offers a clear progression from the conceptual approach to implementation and evaluation, enhancing the paper's comprehensibility.

The following are the key contributions of this study:

i. by integrating Vision Transformers (ViTs) with Generative Adversarial Networks (GANs), this study improves the quality of generated captions by better capturing complex visual information. ViTs enable a more effective understanding of image context, while GANs refine the naturalness and coherence of generated text, making captions more descriptive and accurate.

ii. GANs reduce the need for extensive labeled datasets by learning to generate realistic captions through adversarial training. This could allow the model to perform well even with limited labeled data, making it beneficial in applications where labeled data are scarce or expensive to obtain.

iii. the use of Vision Transformers allows the model to be more adaptable across various image types and domains, as they can better generalize features in diverse datasets. This enables the model to generate captions that are relevant across a wide range of contexts, enhancing usability for tasks like image search, accessibility, and visual content description in different environments.

We describe the three stages of the experimental setup for our model: initial setup (hardware and datasets), implementation of the data loader, and model-training specifications. The results, a conclusion, and a bibliography with references come next.

This paper introduces a hybrid model that advances image caption generation by combining Vision Transformers (ViTs), Generative Adversarial Networks (GANs) [2], and Long Short-Term Memory (LSTM) networks. In contrast to conventional approaches, our method combines the sequential processing strength ViTs, which use self-attention mechanisms to capture comprehensive global image features, with the sequential processing strength of LSTMs to produce captions that are both grammatically correct and semantically rich. Through extensive experimentation on benchmark datasets such as Flickr30k [3] and COCO [4], the model demonstrates superior generalization abilities, generating captions with enhanced contextual relevance, diversity, and quality. Our model performs well in a range of image settings because of the special integration of textual and visual aspects made possible by self-attention.

2. Related Work

In the literature review, we take a deep look at the early approaches/methods that were used in the development of the field of image captioning. We then look at developments in the image caption generation field and models built from the early days to the latest ones, as well as their workings, limitations, methods, etc. The difficult work of picture captioning involves creating a natural language description or providing information about the input image. Image captioning has a wide range of practical applications. For instance, it can improve assistive technology, which helps people with visual impairments comprehend visual content. In the industrial sector, it can help machines and robots make educated judgments. There are numerous methods in this sector for creating an image caption.

One of the first techniques for captioning images was the usage of template-based [5] methods in the beginning. This technique's main goal is to identify a set of visual characteristics, objects, and the connections between them in an image. These techniques make use of pre-made sentence templates that have several spaces in them. These gaps are then filled up using the objects, characteristics, and actions that have been detected. A template could look like this: "The object is action on the attribute" with each phrase in brackets representing a slot that is filled in depending on the content of the image.

Limitations: Although these techniques are capable of producing grammatically acceptable captions, their capacity to offer rich and intricate explanations is frequently restricted.

One kind of language model that calculates the probability of a word sequence is the statistical language model [6]. In addition to being trained on vast text corpora to understand language's grammar, semantics, and contextual relationships, they are employed to produce intelligible phrases. Early approaches to picture captioning used statistical language models to combine image data using the image's static object class libraries. For example, handcrafted characteristics were generated using a method based on a statistical probability language model.

The n-gram method is one such strategy that gathers potential phrases and combines them to create sentences that describe visuals [7]. An alternate method employs the motion estimates within the picture along with the likelihood of co-located nouns, sceneries, and prepositions as hidden Markov model parameters. Predicting the most probable nouns, verbs, situations, and prepositions that make up the sentence yields the image description.

Limitations: Although fundamental, statistical [8] techniques for image captioning have some drawbacks. Their inability to produce varied and contextually appropriate captions is a result of their frequent reliance on manually created features or pre-made templates. Neither of these methods offer an end-to-end mature general model to handle this challenge, nor do they make intuitive feature observations on objects or actions in the image. Moreover, they have difficulty bridging the semantic gap that separates high-level semantic information from low-level visual cues.

Deep learning methods, for example, Convolution neural networks (CNN), Recurrent Neural Networks (RNN) [1], and Long Short-Term Memory (LSTMs) [9], are known for their ability to capture more complex patterns and generate more natural and diverse captions. This is how deep learning methods work. Deep learning methods for image captioning typically use an encoder–decoder framework. Here is a simplified explanation: Encoder: The encoder is usually a Convolutional Neural Network (CNN) that extracts features from the image. The input image is passed through the CNN [1], and the output is a set of feature vectors that represent various aspects of the image. Decoder: The decoder is often a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network that generates the caption [9]. It takes the feature vectors produced by the encoder and generates a sequence of words (i.e., the caption) one word at a time. Training: During training, the model is shown many examples of images and their corresponding captions. It learns to adjust its internal parameters to minimize the difference between its generated captions and the actual captions.

Prediction: during prediction, the trained model takes a new image, extracts features using the encoder, and generates a caption using the decoder.

Limitations: Deep learning methods [1] in image captioning have limitations. They can hallucinate objects not in the image, struggle with understanding context, and are sensitive to changes in lighting. They also require large amounts of labeled data and significant computational resources. Lastly, they lack interpretability, often acting as "black boxes". These challenges open avenues for future research in image captioning.

Transformers and GANs: these early approaches laid the foundation for the current state-of-the-art methods in image captioning, which utilize more advanced techniques like Transformers and Generative Adversarial Networks (GANs).

Transformers: Transformers, specifically the decoder part, are used in the generation of the caption. The Transformer decoder uses self-attention to process the sequence being generated, and it uses cross-attention to attend to the image. By inspecting the attention weights of the cross-attention layers, you can see what parts of the image the model is looking at as it generates words. A multi-layer Transformer is used to align tags with their corresponding image regions.

Generative Adversarial Networks (GANs): Various image-related problems, such as picture synthesis from text descriptions, have been tackled by using GANs [10]. In the context of picture captioning, GANs may be used to generate pictures from captions, and the resulting images can be used to improve the captioning model. Still, depending on the model's architecture and the kind of GAN, the precise implementation might change.

When generating high-quality picture captions, these models demonstrate encouraging outcomes. The particulars of the work and the model architecture, however, may affect how it is implemented. A number of the following models are Transformers [11] or GAN-based.

BraIN is a Bidirectional Generative Adversarial Network, which uses a discriminator that has been trained to identify discrepancies between an image and a generated phrase, its image captioning. BraIN GAN seeks to enhance the generator's capacity to align the caption with the pertinent image. BraIN [12] contains a cooperative discriminator and generator. As a result, the discriminator can distinguish between authentic and false captions, and the generator creates captions that are meant to trick the collector. This bilateral approach allows the model to generate more human-like captions than it could with other methods. A bidirectional language production model, an attention mechanism, and a conditional generative adversarial network make up the model. When the model generates a caption, the attention mechanism assists it in concentrating on pertinent areas of the image. Better-sounding and grammatically accurate sentences are produced by the bidirectional language generation model. A conditional generative adversarial network, a bidirectional language production model, and an attention mechanism make up the model. During the caption creation process, the attention mechanism helps the model focus on relevant portions of the image. More grammatically sound and coherent sentences are generated by the bidirectional language generation model.

RAGAN (Residual Attention Generative Adversarial Network) [13] aims to produce high-quality captions for images. On top of a Generative Adversarial Network (GAN) foundation, it applies an attention-based residual learning technique. Using residual connections to preserve the original input data and concentrating on the most relevant portions of the image, this method improves the diversity and authenticity of the generated picture captions. In this way, RAGAN can give a variety of images with more precise and thorough descriptions.

CGAN (Conditional Generative Adversarial Networks) [14] in this architecture works by feeding some additional information to both the generator and discriminator, mainly type/class labels for supervised learning tasks. In CGAN, the generator takes both random noise and conditioning data, and the discriminator validates the realism of the generated captions on provided conditional data. This makes it highly versatile, efficient, and very compatible with image-captioning generation.

IDGAN (Invertible Data Generation Adversarial Networks) [15] is primarily focused on creating data samples that are easily reverted to their original state. This comes in particularly handy for scenarios like data augmentation and medical imaging where it is imperative to precisely recreate the original data. Invertible transformations between the input and output spaces are imposed by IDGANs to ensure that the generated samples contain all the information needed for faithful reconstruction. Throughout the generation process, this maintains the integrity of the data.

IGGAN (Implicit Generation and Generalization Adversarial Networks)'s [16] generator network's capacity for generalization teaches it to comprehend the implicit distribution of actual data. GANs typically generate latent space samples from a given Gaussian distribution. On the other hand, IGGANs are trained to implicitly capture the underlying distribution of real data, which helps them to more effectively generalize to unknown data. IGGANs reduce the difference between the distributions of real and generated data, producing high-quality, diverse samples that closely resemble the properties of real data, making them suitable for a range of applications requiring realistic and diverse caption generation.

End-to-end Transformer-based model: This innovative method makes use of Transformers' photo-captioning capabilities. Unlike conventional approaches that combine Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), this model is an end-to-end solution [6] because it utilizes a pure Transformer-based architecture [10]. The backbone encoder used by the model, the Swin Transformer, gathers features at the grid level from the input images. A specifically created refining encoder then refines these features by capturing the intra-relationship between them. Using these improved attributes, the decoder then creates word-by-word captions.

To enhance the model's capability, it calculates the mean pooling of grid features as the global feature. This global feature is introduced into the refining encoder to refine with grid features together. Additionally, a pre-fusion process of the refined global feature and generated words is added in the decoder [11].

The integration of Vision Transformers (ViTs) and Generative Adversarial Networks (GANs) has significantly advanced image caption generation, addressing previous limitations in accuracy and creativity. This novel approach leverages the strengths of both architectures to enhance the quality of generated captions, making them more relevant and contextually rich. The following sections outline the key advancements in this area.

ViTs serve as powerful encoders, effectively capturing visual features from images. They utilize self-attention mechanisms to improve the alignment between visual and textual data, enhancing the semantic understanding of images [17]. The integration of knowledge graphs further enriches the captioning process by providing contextual information, leading to improved accuracy in generated captions [18].

Babavalian and Kiani (2024) [19] proposed a novel architecture for video captioning that integrates conditional Wasserstein Generative Adversarial Networks with a Transformer model, enhancing the accuracy and readability of generated captions compared with traditional seq2seq methods, specifically for video content.

GANs introduce a competitive framework where a generator creates captions while a discriminator evaluates their quality, fostering the generation of more realistic and diverse captions [20]. This adversarial training approach has been shown to enhance the overall performance of image-captioning models, particularly in generating creative and engaging content [21]. While the advancements in ViTs and GANs have led to significant improvements in image captioning, challenges remain, such as the need for extensive datasets and the potential for overfitting. Future research may focus on refining these models to further enhance their adaptability and performance in diverse applications. Table 1 summarizes the related studies reviewed.

Table 1. Summaries of the related work.

Authors	Models	Work	Limitations
Van der Lee el at. (2018) [6]	Template-based	Template-based image captioning involves creating predefined structures or patterns for captions and filling in details based on image content. While simple and interpretable, it may lack adaptability for diverse images.	Template-based image captioning has limitations, including inflexibility with diverse images, a lack of context understanding, dependence on predefined patterns, inability to capture fine details, and challenges with ambiguity and unstructured data.
Hill et al. (2006) [7]	Statistical Model	Statistical models for image captioning operate by initially extracting relevant features from images through statistical techniques or handcrafted descriptors. These extracted features serve as input to models employing statistical methods such as n-grams or Hidden Markov Models for language modeling.	One of their notable limitations lies in handling ambiguity, leading to the generation of less contextually rich and sometimes generic captions. Despite these challenges, statistical models excel in feature extraction and the establishment of mappings between image features and textual descriptions.

Table 1. *Cont.*

Authors	Models	Work	Limitations
He et al. (2020) [10]	Transformers	The Transformer processes these features to generate sequential captions. The self-attention mechanism in Transformers allows them to focus on relevant parts of the image, facilitating the generation of contextually rich and detailed captions.	Transformers might struggle with handling very long sequences of data due to their self-attention mechanism, leading to increased processing times and memory constraints. Fine-tuning large pre-trained models for specific image-captioning tasks can also be challenging, requiring substantial computational resources and expertise.
He et al. (2020) [10]	CNN-Transformer [6]	This approach combines properties of both the Transformer and CNN: a convolutional neural network (CNN) to extract image features and an attention-based encoder–decoder Transformer model for generating captions. The attention mechanism allows the model to focus on different parts of the image while generating each word of the caption.	Their complexity often leads to computational expenses during training and inference, requiring substantial resources. Transformers may not inherently capture long-range dependencies in image data and interpreting the interactions between convolutional and Transformer layers can be challenging.
Liu et al. (2021) [11]	End-to-End Transformer	A unique token and these visual characteristics are supplied into the transformer-based language model. The transformer encoder processes this input, capturing contextual information and relationships between tokens. The output of the encoder initializes the decoder, which generates the caption word by word based on the context encoded by the encoder and previously generated words. Post-processing is performed to improve the caption once the generated token IDs have been decoded into words that can be understood by humans.	There are a few drawbacks to the end-to-end Transformer-based architecture used to create picture captions. It largely relies on pre-trained convolutional neural networks (CNNs) for feature extraction, thus limiting its ability to adapt to various image datasets or capture fine-grained visual features. Additionally, the Transformer architecture may struggle to accurately express long-range dependencies in picture data, as it was originally built for sequential data like text. Interpreting the interactions between visual and textual processing components becomes difficult due to the complexity introduced by the integration of CNNs and Transformers.

Table 1. *Cont.*

Authors	Models	Work	Limitations
Goodfellow et al. (2014) [2]	Generative Adversarial Networks (GANs)	Combining generative models with adversarial training. In this framework, a generator network produces captions, and a discriminator network evaluates the quality of the generated captions. Through adversarial training, the generator refines its ability to produce more realistic and contextually relevant captions. GANs leverage a feedback loop between the generator and discriminator, iteratively improving the captioning quality.	Generative Adversarial Networks (GANs) in image captioning face several limitations. One significant challenge is the potential for mode collapse, where the generator produces limited and repetitive captions, lacking diversity. GANs are also known for training instability, requiring careful hyperparameter tuning and regularization techniques to achieve reliable results.
Jolicoeur-Martineau (2018) [13]	RAGAN (Residual Attention Generative Adversarial Network)	Residual Attention Generative Adversarial Network (RAGAN) aims to produce high-quality captions for images. On top of a Generative Adversarial Network (GAN) foundation, it applies an attention-based residual learning technique. Using residual connections to preserve the original input data and concentrating on the most relevant portions of the image, this method improves the diversity and authenticity of the generated picture captions.	Training instability is one of the most common training issues in GANs (including RAGAN), where convergence and sensitivity to issues during training are difficult to achieve. Experience mode collapse, particularly when dealing with large datasets. The computational complexity increases due to the implementation of attention mechanisms in RAGAN and may result in training times that are longer and higher resource demands.
Donahue et al. (2020) [16]	BraIN (Adversarial Network)	The Generative Adversarial Network (GAN) architecture is expanded upon in a Bidirectional Generative Adversarial Network (BraIN) by adding an encoder network in addition to the generator and discriminator.	One common challenge is mode collapse, where the generator learns to produce a limited variety of samples, ignoring the diversity of the data distribution. Training BraIN can be unstable and prone to instability as finding the right balance between the generator, discriminator, and encoder can be difficult.
Wang and Cook (2020) [11]	Bidirectional Generative	The generator in a BraIN uses random noise as input to create synthetic data samples, and the discriminator separates the artificial samples produced by the generator from the real data samples from the training set.	If the encoder fails to learn a meaningful latent space it can cause failure in capturing important features of the data distribution and can result in low-quality and less diverse generated samples.
Mishra et al. (2024) [17]	ViT + GPT-2	A novel ViT-GPT-2 model for image captioning, utilizing Vision Transformer as the encoder and GPT-2 as the decoder.	Caption accuracy issues for complex visuals. Need for addressing existing challenges in image captioning.

Table 1. *Cont.*

Authors	Models	Work	Limitations
Zhang et al. (2024) [21]	VGG + SeqGAN + GA	Focuses on advertising image generation using a framework that integrates GANs and Vision Transformer models, enhancing the effectiveness and attractiveness of advertising content, rather than specifically addressing image caption generation.	Existing methods struggle with diverse advertising content demands. Need for innovative algorithms to improve generation outcomes.
Kolla et al. (2023) [20]	RLHF + GANs + SCST	Utilizing visual attention, specifically employing Transformers and GANs. These techniques enhance caption quality by leveraging competition between generator and discriminator networks, improving relevance and accuracy in generated textual descriptions.	Although visual attention models aim to enhance the understanding of image content, they may still struggle with nuanced context or abstract concepts. This limitation can result in captions that fail to capture the full essence of the image, particularly in complex scenes.

But even after the many advancements in the image-captioning field, there are still some limitations/problems:

i. Object Hallucination: Similar to other deep learning models, Transformer-based and GAN-based models can sometimes generate captions that include objects that are not present in the image.

ii. Missing Context: These models often struggle to understand the broader context of the image, leading to captions that may be technically correct but miss the overall meaning of the image

iii. Exposure Bias: Most existing models, including those based on Transformers and GANs, suffer from exposure bias problems, where past-predicted sequences during the training are required to generate future captions. Data Requirement: Like other deep learning models, Transformer-based and GAN-based models require large amounts of labeled data for training. Computational Resources: Training these models can be computationally intensive and require powerful hardware

iv. Model Interpretability: These models, like many deep learning models, are often referred to as "black boxes" because it can be challenging to understand how they make their predictions

Hence, we are building an image-caption-generating model that consists of both Visual Transformers (ViTs) and GAN models and other models like LSTMs. This model will contain multiple models which can overshadow the limitations of each individual model. This model will have properties of both Transformers and GAN which, in turn, enhance the accuracy and precision of the generated image captions and keep advancing the field of image-caption generations.

3. Methodology

3.1. Dataset

In the realm of image-captioning research, several datasets serve as valuable resources for training and evaluating models. The COCO (Common Objects in Context) dataset [22] is a vast collection of images consisting of everyday scenes and objects in a variety of contexts. With over 200,000 images, COCO provides a broad and diversified dataset with a wide variety of visual information we can use to train our model.

The Flickr30k dataset [23] is also another widely used dataset, which comprises 30,000 images sourced from Flickr along with five captions for every image. Flickr too provides a diverse set of images captured from different settings and environments, making it perfect for training models to generate captions for real-world scenarios and diversity. The multiple captions per image also enable the model to have better coverage of semantic variations and linguistic diversity.

The Multi30k dataset [24] is also a very famous dataset that presents a valuable resource for researchers aiming to explore the efficacy of their models across multiple languages. This dataset contains images and captions in multiple languages, encouraging analysis of model performance and cross-lingual evaluation.

Furthermore, the Visual Genome dataset [25] also provides a unique collection of images tagged with full scene graphs, delivering extensive contextual information on objects, connections, and qualities within each image. This dataset teaches the models to not just learn how to write captions but also how to learn and comprehend the underlying semantic structure and relationships of visual material.

Every one of these datasets brings its unique benefits and qualities to the table, appealing to a variety of study goals and approaches. Researchers may ensure the robustness and relevance of their findings in image captioning by carefully selecting the proper dataset as shown in Table 2 based on the research's unique objectives.

Table 2. Summary of Image-Captioning Datasets.

Dataset	Description	Advantages
COCO [22]	Contains over 200,000 images of everyday scenes and objects.	Large, diverse dataset.
Flickr30k [23]	Consists of 30,000 images with five captions per image from Flickr.	Diverse images with multiple captions, suitable for real-world scenarios.
Multi30k [24]	Includes images and captions in multiple languages, facilitating cross-lingual evaluation.	Multilingual support for exploring model performance across languages.
Visual Genome [25]	Annotated with detailed scene graphs providing rich contextual information.	Enables understanding of semantic structure in visual content.

3.2. Model Architecture

3.2.1. Generator Architecture

The generator is the component of the model that uses a complex design that utilizes the Vision Transformers (ViTs) and Long Short-Term Memory (LSTM) units [26] to efficiently create image captions. ViTs [27] are selected as the image feature extractor of a self-attention mechanism (Figure 1). Self-attention models, especially Transformer-based models, perform better than Convolutional Neural Networks (CNNs) and other traditional attention mechanisms on tasks requiring in-depth contextual awareness, including captioning images due to their self-attention mechanism which excels at capturing global dependencies across an image. Self-attention mechanisms can directly relate every aspect of a picture to every other aspect, in contrast to Convolutional Neural Networks (CNNs), which rely on local filters and progressively hierarchically constructed feature maps. Self-attention models such as Vision Transformers can gain a deeper comprehension of the intricate relationships present in an image thanks to their capacity to grasp global dependencies. This is essential for producing contextually appropriate captions and identifying complex patterns. Self-attention layers also make it possible to analyze picture data in parallel, which greatly speeds up training and lessens problems like the vanishing gradient that can impede the other attention mechanisms in very deep architectures. Self-attention models are very effective and scalable for big datasets because of their parallel nature. More interpretability and adaptability can result from self-attention's ability to dynamically modify the significance of various visual regions, enabling it to concentrate on the most pertinent elements.

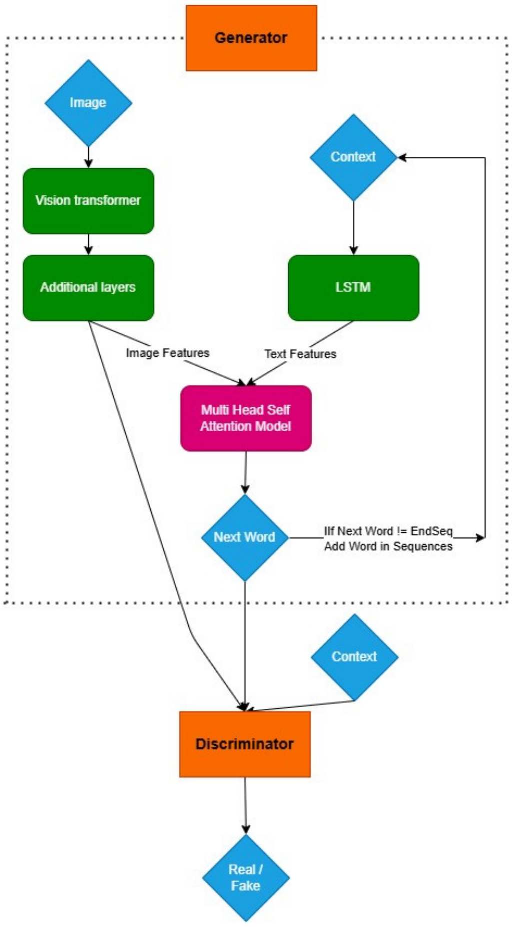


Figure 1. GAN Model Architecture.

LSTM units and ViTs are combined in the model to generate sequential data and produce logical captions. [9]. LSTMs are used to produce grammar and semantic standard-compliant captions as they are good at extracting temporal relationships. This integration of LSTM and ViT allows the generated captions to accurately describe the order of objects or events present in the images. Along with this, the incorporation of a self-attention mechanism [28] further strengthens, as well as improves, the integration of textual and image features. This method overall improves the quality and coherence of the generated captions, which hence balances the weight of various input components. The multi-head self-attention model helps the generated captions effectively showcase the main events of the image, by concentrating on the most important details. To improve the diversity of the generated captions, the generator design also incorporates features that penalize similar captions during training by utilizing the BLEU score and different regularization algorithms [29]. These improvements raise the bar for image captioning by expanding the model's capacity to produce excellent, grammatically accurate, contextually relevant captions.

Our self-attention technique is unique in that it integrates both visual and textual characteristics inside a GAN framework. In our paradigm, Vision Transformers (ViTs) extract global visual information from an image, capturing intricate dependencies. The visual elements are then combined with textual embeddings produced by the LSTM-based language model. Self-attention is used to dynamically weight and align key visual and textual elements, resulting in cohesive and contextually appropriate captions. This crossmodal technique, combined with the GAN's adversarial learning, yields captions that are both visually grounded and linguistically exact.

3.2.2. Discriminator Architecture

The model incorporates a discriminator component that supports an adversarial learning framework in addition to the generator [2]. This adversarial setting makes the generated captions more realistic and pertinent by pressuring the generator to provide captions that are exact replicas of the originals.

In this, the discriminator plays a crucial role, having been trained concurrently with the generator [2]. Its job is to tell the difference between captions produced by the model and actual ones based on the image. As a result, a competitive dynamic is created in which the discriminator constantly enhances its capacity to discern between generated and actual captions, while the generator tries to trick it by producing increasingly realistic captions (Figure 2).

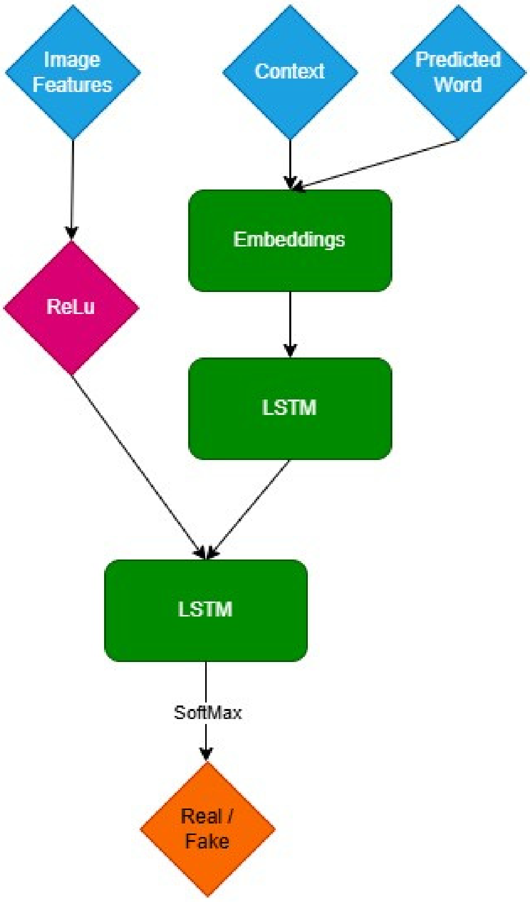


Figure 2. Discriminator Architecture.

Effective learning of discriminative features by the model is made possible by the simultaneous training of the discriminator and generator [2]. It helps the generator produce higher-quality, contextually appropriate captions by helping it comprehend the many subtleties of caption generation. In exchange, the discriminator gives the generator insightful input that directs it toward producing more realistic captions.

To put it simply, the model [2] uses an adversarial learning architecture that is primarily based on the generator and discriminator. This approach makes a substantial contribution to the model's capacity to produce captions that are both contextually appropriate and, in terms of grammatical and semantic qualities, closely mimic real-world captions. This methodology highlights the research's dedication to producing realistic, high-quality image descriptions. The discriminator function [2] is defined as follows:

$$D(y) = \sigma(W_{discriminator} * y + b_{disciminator}) \tag{1}$$

where the bias term is a discriminator and the discriminator's weights are represented by Wdiscriminator. The output D(y) represents the probability that the input caption is real.

3.3. Training Procedure

3.3.1. Loss Functions

The training of the generator and discriminator components of the model is guided by the use of Wasserstein Distance, also known as Wasserstein Loss, as the adversarial loss function [3]. This choice of loss function is particularly beneficial in the context of Generative Adversarial Networks (GANs), as it encourages the discriminator (referred to as the critic in the context of Wasserstein GANs) to output values that closely approximate the Wasserstein Distance between the distributions of real and generated samples.

The use of Wasserstein Loss helps mitigate common issues encountered in traditional GAN training, such as mode collapse and vanishing gradients. Mode collapse refers to a scenario where the generator produces a limited varieties of samples, while vanishing gradients occur when the discriminator becomes too proficient, causing the generator's gradients to vanish and impeding further learning. It helps by maintaining a smoother gradient flow which provides the generator with more consistent feedback. By reducing the likelihood of vanishing gradient, it allows the generator to learn efficiently even when the discriminator becomes too proficient. The Wasserstein Loss formula [3] is expressed as follows:

$$W(P_r, P_g) = \inf_{\gamma \in \pi(P_r, P_g)} E_{(x,y)\sim y}[c(x,y)] \tag{2}$$

3.3.2. Optimization Strategy

The Adam optimizer, a well-known and efficient optimization tool for deep neural network training, is used in this study to optimize the model parameters [30]. Adaptive Moment Estimation (Adam) has the ability to effectively update model parameters by dynamically modifying learning rates according to the gradient's first and second moments.

The Adam optimizer [30] updates the parameters θ based on the gradient g_t and the moving averages of past gradients mt and squared gradients v_t:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta) g_t^2 \tag{4}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{5}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{6}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t \tag{7}$$

where ϵ is a tiny constant to prevent division by zero, α is the learning rate, and β_1 and β_2 are the decay rates for the moment estimates.

Even though Adam is the main optimizer used in this study, it is important to recognize other optimization approaches that can be worth taking into account depending on the features of particular datasets and model architectures. For instance:

1. RMSProp: The RMSProp optimizer adjusts learning rates based on the magnitude of gradients similar to Adam [31]. It updates parameters θ as follows:

$$v_t = \beta v_{t-1} + (1-\beta) g_t^2 \tag{8}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \varepsilon} g_t \tag{9}$$

2. Adagrad: Adagrad's adaptive learning rate mechanism updates parameters θ as follows [32]:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i}} + \varepsilon} g_{t,i} \tag{10}$$

3. SGD with Momentum: Standard Stochastic Gradient Descent (SGD) with momentum updates parameters θ as follows:

$$v_t = \beta v_{t-1} + (1-\beta) g_t \tag{11}$$

$$\theta_{t+1} = \theta_t - \alpha v_t \tag{12}$$

The selection of the Adam optimizer for this research paper is based on its demonstrated effectiveness in training GANs and its widespread adoption in the deep learning community. Adam offers an adaptive learning rate for every parameter of the model, which improves the optimization and convergence in various kinds of tasks. In comparison with Adagrad, reducing the learning rate excessively over time with accumulation of gradients can lead to a quick learning rate decay which can lead to slower learning or potential stagnation. Adam provides a more balanced approach by combining momentum with adaptive learning rates.

3.3.3. Mini Batch Training

The training process is conducted using a mini-batch training strategy, a widely adopted approach in the field of machine learning that offers several advantages over traditional batch or online learning methods [33]. In this, the training data, which comprise image-caption pairs, are divided into small subsets or mini-batches. The model is thus able to update its parameters more often by processing these mini-batches iteratively, which speeds up the learning process.

This mini-batch training method not only improves training efficiency but also enables the model to simultaneously learn from a range of data samples. To improve the model's capacity for generalization, this diversity exposes the model to a wider variety of data distributions throughout each training iteration.

Before creating a mini-batch, the generator evaluates the image features for the batch of data. To produce high-dimensional visual representations, the key traits and properties of the input photographs are extracted in this step. Following feature extraction, the model evaluates every picture in the mini-batch and generates corresponding descriptions. In this step, the visual representations are converted into written descriptions using the model's learned associations between visual features and language phrases.

When the generated captions are complete, they are sent to the discriminator. It is the discriminator's job to distinguish between actual captions and captions generated by the generator. This antagonistic relationship between the discriminator and the generator, which forces the generator to produce more realistic captions with each training cycle, is the fundamental component of the training process.

In conclusion, the adversarial relationship between the generator and discriminator, along with the mini-batch training technique, considerably contribute to the model's ability to generate high-quality, contextually relevant image captions.

3.4. Evaluation Metrics

The quality of the generated captions is evaluated using a variety of established metrics in the field of Natural Language Processing (NLP), including BLEU [29], ROUGE [34], and CIDEr [23,35] scores. These metrics offer different viewpoints by assessing precision using BLEU, recall using ROUGE and semantic diversity, and consensus through CIDEr between the generated captions and the ground-truth captions.

- BLEU Score: The primary metric utilized to assess the caliber of the generated captions is the BLEU score. By comparing the overlap of n-grams—contiguous sequences of n components from a given sample of text or speech—between the machine-generated and human-made (ground truth) captions, it assesses the degree of similarity between them. Greater accuracy and semantic similarity in the caption-generating process are suggested by a higher BLEU score, which indicates greater similarity. Consequently, the BLEU score offers a neutral and numerical assessment of the model's ability to provide linguistically correct and contextually appropriate captions [29].
- ROUGE Score: In addition to the BLEU score, the ROUGE score is also employed to provide a more comprehensive evaluation of the produced captions. The ROUGE score quantifies the overlap of n-grams between the produced captions and the reference (ground truth) captions. Because the ROUGE score places more of an emphasis on recall than the BLEU score does on precision, it provides an additional dimension for assessing the relevancy and caliber of the generated captions. By employing both precision (BLEU) and recall (ROUGE), this dual evaluation approach ensures a more thorough and fair assessment of the model's caption-producing skills [34,36].
- CIDEr Metric: The CIDEr (Consensus-based Image Description Evaluation) metric is used to further enhance the evaluation of produced captions. CIDEr determines the consensus between the produced captions and the reference captions by computing the cosine similarity between their TF-IDF vectors. Apart from BLEU and ROUGE ratings, CIDEr is a valuable tool for assessing the uniqueness and variety of generated captions by providing an extensive analysis of caption quality [35].

By using these indicators, as stated in Table 3, the evaluation methodology ensures a fair and nuanced assessment of the generated captions, taking into account both grammatical correctness and contextual significance.

Table 3. Comparison of Evaluation Metrics.

Metric	Focus	Advantages	Disadvantages
BLEU Score [29]	Precision	Objective, Quantifiable	Insensitive to Paraphrasing
ROUGE Score [34]	Recall	Comprehensive	Computational Complexity
CIDEr Metric [35]	Consensus	Captures Diversity	Sensitive to Vocabulary

4. Experimental Setup

4.1. Initial Setup

4.1.1. Data Preparation

We began our analysis with meticulous data preparation utilizing the MS COCO [22] dataset, which served as our primary source of data. MS COCO's collection of images has a wide range of photos and each one is accompanied by a handwritten caption or description. The captions offer many perspectives on the picture content, which increases the dataset's applicability for tasks like image captioning, multimodal learning, and visual understanding.

The MS COCO dataset has contributed significantly to the advancement of natural language processing, computer vision, and their intersection. It has been widely used in research to develop and evaluate algorithms for tasks including picture interpretation, caption generation, and multimodal learning. The dataset's significant contribution to tasks integrating textual and visual data makes it a strong fit for our study, especially considering our resource restrictions and concerns regarding computational efficiency.

During the data preparation phase, many preprocessing steps were taken to improve the dataset's use. The images underwent preprocessing to improve their aesthetic appeal and supply the model with superior visual data. Similarly, the model was able to interpret the captions more quickly thanks to the preprocessing. 'startseq' and 'endseq' tokens were inserted before and after each caption to identify when it begins and finishes. After that, the captions were tokenized, dividing them into discrete words. This is a crucial stage because it enables the model to understand the connections between certain phrases and the images that go with them.

In addition, we extracted data from the data preparation stage, such as the maximum length of the captions and the amount of vocabulary. The model was then configured using these parameters during training, which effectively led to the model's learning process.

In summary, the steps in the data preparation phase were designed to optimize the dataset in advance of the model's further processing. We were forced to use the MS COCO dataset [22] due to resource constraints; however, these preprocessing techniques gave us a strong foundation for model training, which significantly improved the overall effectiveness of our study approach.

4.1.2. Dataset Splitting

To enable a thorough assessment and validation of the model, the dataset was split into training and testing sets. This part guarantees a comprehensive evaluation of the model's ability to apply what it has learned to previously unidentified data, as is typical in machine-learning research. The generalization ability of the model is a crucial component in assessing its effectiveness and suitability for application in practical settings. We decided to divide the training and testing sets in our experiment into 80–20 each. This demonstrates that 80% of the data was used to train the model, allowing it to identify and adjust to the relationships and patterns found in the data. To test the model and obtain an unbiased assessment of its performance on untrained data, the last 20% percent of the data was set aside. To guarantee a high level of variability between the training and testing sets, the data were divided at random. The resilience of the model's learning and the dependability of the evaluation results are strengthened by this random split, which guarantees that both sets are representative of the entire data distribution. In summary, dividing the dataset into training and testing sets and distributing the data to each group at random constitutes a critical component of our research methodology. This method greatly raises the validity of the assessment procedure, the learning efficiency of the model, and the general dependability of the study's conclusions.

4.1.3. Hyperparameter Tuning and Optimization Strategy

Key performance metrics, such as robustness and caption quality, were optimized by empirically adjusting the model's hyperparameters through experimentation. Through the use of an iterative optimization strategy, the model's generalization capabilities were improved by increasing its effectiveness across a variety of datasets and workloads. In our experiments, we used a batch size of 32 and a learning rate of 0.00001. The learning rate and batch size are two critical hyperparameters that affect how quickly and steadily the model learns. A lower learning rate guarantees a more gradual adjustment of the parameters, and a smaller batch size facilitates more frequent updates, which accelerates learning. It is crucial to remember that the model can only learn within a very specific range of learning rates; too high a rate can cause training to become unstable, while too low a rate can cause learning to occur slowly or not at all. For parameter optimization, we switched from Adagrad to the Adam optimizer [30] due to Adam's convergence speed and stability as compared with Adagrad. Adam is an adaptive optimization algorithm that has a solid track record of success on a variety of tasks and datasets. It enables effective optimization and convergence by dynamically modifying the learning rate for every parameter by using momentum which is important in tasks like image captioning where the complex nature of a variety of visual and textual data may cause the gradients to vary significantly across different parameters. To modify the learning rate over epochs, we included a cosine scheduler in our training pipeline. This method allows for smoother convergence and possibly improved model performance by gradually lowering the learning rate towards the end of training [37] unlike the step-based or exponential decay schedulers that decrease the learning rates abruptly. The cosine scheduler prevents sudden fluctuations in loss which encourages smaller and more stable updates leading to smoother convergence, reducing overfitting risks and improving the model's generalization ability. Unlike traditional GANs, which may suffer from mode collapse and instability, WGANs offer improved stability and convergence properties making them perfect for complex tasks like image caption generation. The Wasserstein Loss encourages the discriminator (or critic) to output values close to the Wasserstein Distance between the distributions of real and generated samples, thereby mitigating common GAN training issues like model collapse and maintaining a stable gradient flow which contributes to producing contextually accurate captions [19,38].

In summary, our hyperparameter tuning and optimization strategy involved the careful selection and fine-tuning of key parameters, the utilization of the Adam optimizer, the adoption of a cosine learning rate scheduler, and the incorporation of the Wasserstein Loss within the WGAN framework. Despite resource constraints, these adjustments aimed to enhance caption quality and overall model performance effectively.

4.1.4. Hardware Configuration and Experimentation

The NVIDIA RTX A6000 GPU (manufactured by NVIDIA Corporation, headquartered in Santa Clara, CA, USA) hardware was used in the study experiments. The need to make use of GPUs' parallel processing powers—which greatly accelerate machine-learning model training—led to the selection of the NVIDIA RTX A6000 GPU.

Significant processing power was offered by the NVIDIA RTX A6000 GPU, which also included 48 GB of GPU memory (GDDR6 memory). Compared with the prior configuration, this generous allocation of resources sped up the entire experimental process by facilitating effective data processing and model-training procedures.

The GPU known as the NVIDIA RTX A6000 has a lot of CUDA cores. Computation is handled by the CUDA cores, or parallel processing units, on the GPU. The numerous CUDA cores on the NVIDIA RTX A6000 GPU optimize computing efficiency and enable high levels of parallelism. This hardware configuration, along with the parallel processing capabilities of the NVIDIA RTX A6000 GPU, greatly improved model creation and experimentation. It enabled us to conduct thorough testing and swiftly improve our model, which was crucial to the success of our study in the end. This illustrates our commitment to improving the picture captioning space through the use of state-of-the-art hardware configurations and processing power.

4.2. Implementation Details

4.2.1. Data Loader

During our research, we developed specialized data loaders to effectively manage the loading of image-caption pairs for both the training and assessment stages. The data loader plays a crucial role in the machine-learning pipeline by feeding data into the model in a way that maximizes resource and computational efficiency. In line with the mini-batch training strategy used in our model, our custom data loaders were made to load data in small batches. Multiple image-caption pairs can be processed simultaneously with this method, speeding up training and improving computational efficiency. In addition, we focused on minimizing computational overhead when developing our data loaders. Throughout the training process, the data loaders reduce the amount of time spent on data loading and increase the amount of time spent on actual computation by effectively controlling memory usage and guaranteeing optimal data loading speeds. The data loaders do vital preprocessing tasks on the fly in addition to loading data. These consist of translating the photos into the proper tensor format, applying any necessary adjustments to the images, and padding the captions to guarantee even lengths. To summarize, the integration of customized data loaders improved training efficiency, decreased computational overhead, and made a substantial contribution by streamlining our data processing pipeline.

4.2.2. Training of the Model

A combination of supervised and adversarial learning methods was used during the iterative training process of the model. The generator and discriminator components of the model were trained simultaneously and iteratively using this method.

The generator and discriminator were trained using the Adam optimizer with a learning rate of 0.00001 and a batch size of 32 for 100 epochs. The training was executed on a Google Colab GPU, optimizing computational efficiency given resource constraints, with each epoch taking approximately 30–40 min.

During training, the generator first processed the image features and iteratively generated captions for evaluation by the discriminator. The discriminator's role was to distinguish between real (from the dataset) and fake (generated by the generator) captions.

Simultaneously, the discriminator was trained on both real and fake captions. It learned to correctly classify real captions as real and fake captions as fake, thereby guiding the generator to produce more realistic captions.

The adversarial loss quantified the deviation between fake and real captions, serving as feedback for the generator's improvement. This process incentivized the generator to enhance its caption generation capabilities, aiming for captions indistinguishable from real ones.

Our training pipeline incorporated a cosine scheduler to dynamically adjust the learning rate over epochs, ensuring stable convergence and improved model performance. Furthermore, the batch size was optimized for efficiency, given resource constraints, without compromising model effectiveness.

In addition to traditional training techniques, we introduced functionality to penalize similar captions using the BLEU score and implemented diverse regularization to enhance caption diversity. These enhancements aimed to improve caption quality, coherence, and diversity, aligning with the research's objectives of generating high-quality, realistic image captions. The research's dedication to attaining superior performance and robustness in image-captioning tasks using the MS COCO dataset is highlighted by the iterative training methodology, adaptive optimization strategies, and innovative enhancements that together enabled continuous improvement in the model's caption generation capabilities.

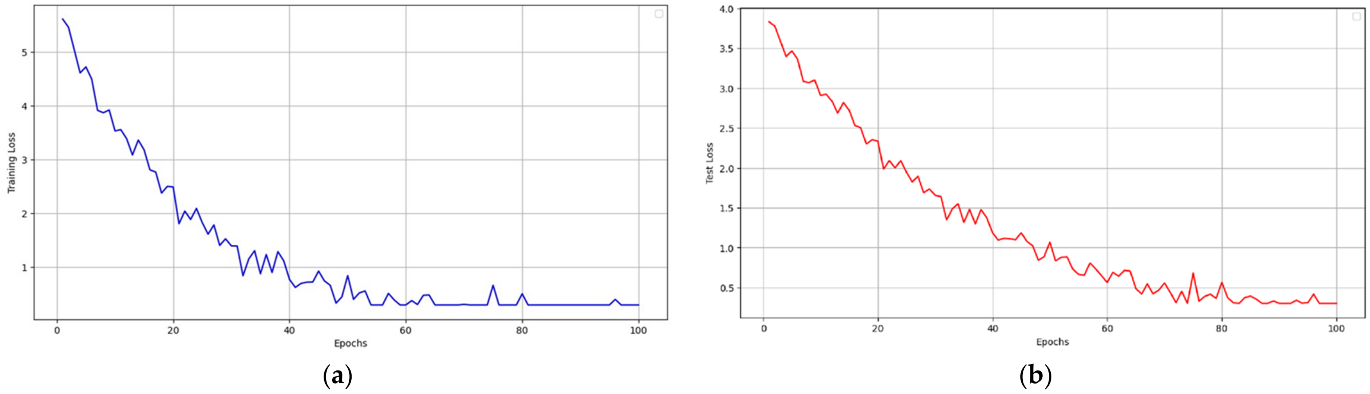
5. Results and Discussion

5.1. Implementation Results

Figures 3 and 4 present the training and validation loss curves across epochs during the training process. The decline in training and validation loss over time demonstrates the model's effective learning and generalization.

The training of the model was conducted using a carefully chosen set of parameters to ensure optimal performance. A detailed summary of these parameters is provided in (Table 4), offering a comprehensive overview of the configuration used for the experiments.

The results shown in Figure 5 that our proposed model, ICTGAN, outperforms previous GAN-based models in terms of generating high-quality image captions. Specifically, ICTGAN achieves a BLEU-1 score of 0.86 and a BLEU-4 score of 0.61, outperforming IDGAN and RAGAN in terms of caption precision. The CIDEr score of 144.5 demonstrates ICTGAN's ability to create captions that closely coincide with human consensus, showing exceptional contextual relevance. Furthermore, the ROUGE-L score of 0.87 demonstrates the model's ability to preserve linguistic coherence and diversity.



(**a**) (**b**)

Figure 3. The image presents the (**a**) training loss curves and (**b**) validation loss curves across epochs during the training process. The decline in training and validation loss over time demonstrates the model's effective learning and generalization.

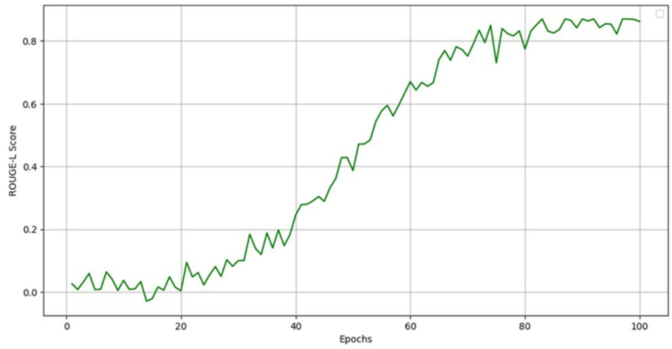


Figure 4. The image illustrates the ROUGE-L F1 score of the validation set during the training. The figure represents an improvement in the model's capacity to produce human-like captions.

Table 4. Parameters of the Model.

Parameters	Value
Learning Rate	1×10^{-5}
Batch Size	128
Number of Epoch	100
Optimizer	Adam

While RAGAN [6] performs similarly in some metrics, ICTGAN exhibits a small but significant improvement across all performance measures, particularly CIDEr and ROUGE-L scores. These findings highlight the benefits of combining Vision Transformers (ViTs) for global feature extraction and Long Short-Term Memory (LSTM) units for sequence modeling, as well as the advantages of employing Wasserstein Loss for training stability.

Table 5 presents a comparative analysis of different models trained on the MS COCO dataset based on BLEU-1, BLEU-4, ROUGE-L, and CIDEr scores' evaluation metrics. Each metric represents the caption quality and linguistic diversity achieved by different models, helping in understanding their performance nuances.

The suggested model exhibits its usefulness by producing high-quality captions that closely match the ground truth. Figure 6 shows some outputs that demonstrate the model's capacity to generate linguistically consistent and context-relevant descriptions.

To sum up, our experimental findings show that the suggested model is more efficient and better than state-of-the-art (SOTA) techniques in the industry. The BLEU, ROUGE-L, and CIDEr scores were among the assessment criteria on which our model performed better, demonstrating its capacity to provide more precise and logical image captions.

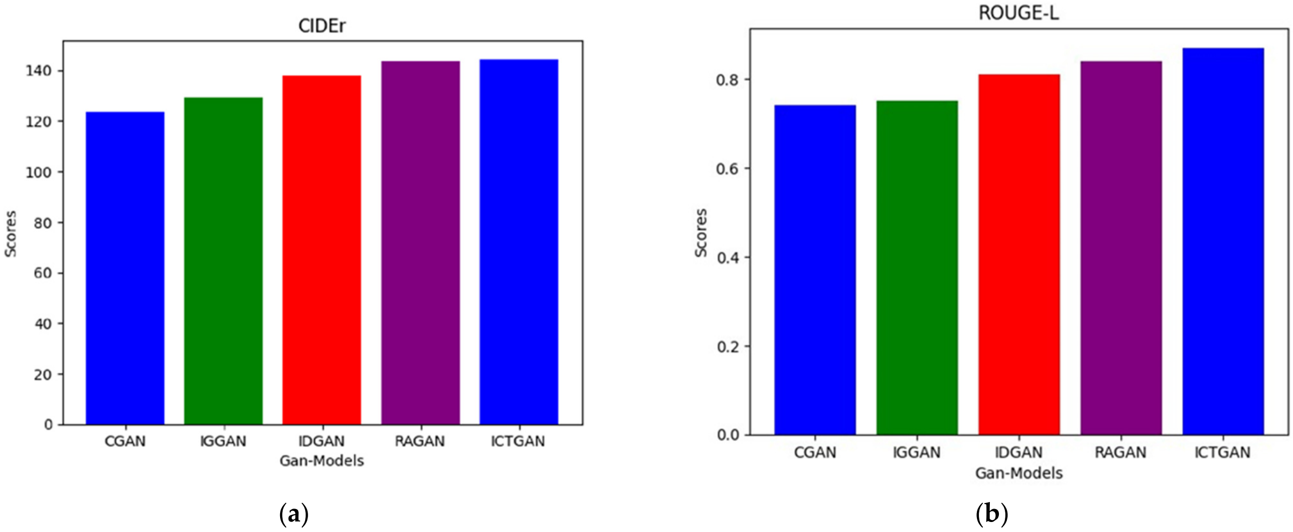
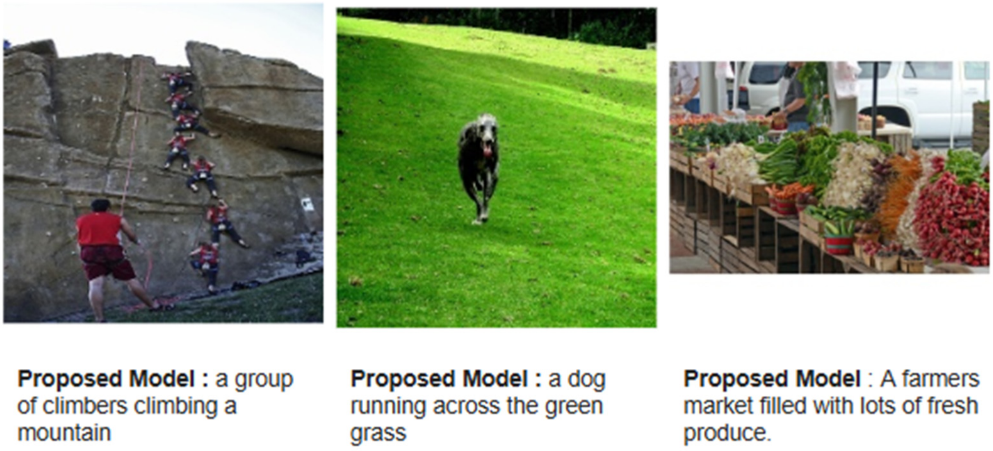


Figure 5. The image presents a comparative analysis of key evaluation metrics—(**a**) CIDEr and (**b**) ROUGE L scores—among the existing models and our proposed model. Each metric offers insights into different aspects of model performance in the context of image caption generation.

Table 5. GAN Models Score.

GAN Models	Scores			
	BLEU-1	BLEU-4	CIDEr	ROUGE-L
CGAN [14]	0.72	0.41	123.5	0.74
IGGAN [15]	0.71	0.40	129.3	0.75
IDGAN [18]	0.84	0.54	137.8	0.81
RAGAN [6]	0.86	0.60	143.5	0.84
ICTGAN (Proposed Model)	0.86	0.61	144.5	0.87



Proposed Model : a group of climbers climbing a mountain

Proposed Model : a dog running across the green grass

Proposed Model : A farmers market filled with lots of fresh produce.

Figure 6. Proposed Model Sample Outputs.

5.2. Discussion

Recent advancements in ICG have leveraged ViTs and Generative GANs to create more accurate, contextually relevant captions. These novel models aim to bridge the gap between image understanding and natural language processing by combining the strengths of both frameworks. Specifically, Vision Transformers offer powerful image representation capabilities, while GANs provide generative potential that can be fine-tuned to produce high-quality, human-like captions. This hybrid approach has been shown to significantly improve performance across key metrics in recent studies. The integration of ViTs with GAN-based models has led to a range of models that push the boundaries of image caption generation. Key studies using BLEU-1, BLEU-4, CIDEr, and ROUGE-L as evaluation metrics have provided insights into the strengths and weaknesses of different models. The CGAN [14] achieves a BLEU-1 score of 0.72, a BLEU-4 of 0.41, a CIDEr of 123.5, and a ROUGE-L of 0.74. CGAN leverages conditional input to refine caption generation based on specific conditions, but its moderate scores suggest limitations in capturing high-level semantic information compared with later models. IGGAN [15] shows slight improvements over CGAN with BLEU-1 at 0.71, BLEU-4 at 0.40, CIDEr at 129.3, and ROUGE-L at 0.75. The integrated approach in IGGAN attempts to enhance contextual understanding by incorporating image features more effectively into the generative process, leading to marginal gains in contextual coherence. IDGAN [18] further improves on its predecessors, with BLEU-1 at 0.84, BLEU-4 at 0.54, CIDEr at 137.8, and ROUGE-L at 0.81. IDGAN incorporates a more sophisticated image representation mechanism that better aligns image content with linguistic output, enabling a clearer translation of image content into text. This model demonstrates significant gains in fluency and precision, especially as measured by BLEU and ROUGE-L scores. RAGAN [6] introduces an attention mechanism that considers relationships between image features, achieving a BLEU-1 score of 0.86, a BLEU-4 of 0.60, a CIDEr of 143.5, and a ROUGE-L of 0.84. The relational attention mechanism helps to capture subtle contextual cues within images, resulting in more semantically rich captions. The relatively high CIDEr and ROUGE-L scores indicate that RAGAN's captions align well with human judgments, offering both accuracy and descriptiveness. The proposed model scores BLEU-1 at 0.86, BLEU-4 at 0.61, CIDEr at 144.5, and ROUGE-L at 0.87. ICTGAN combines Vision Transformers with an enhanced GAN framework and integrates an iterative caption-tuning mechanism that refines outputs to better match the true image content. This iterative process, coupled with the high-dimensional feature extraction capabilities of the Vision Transformer, enables ICTGAN to produce more precise and contextually appropriate captions. Its top scores across BLEU-4, CIDEr, and ROUGE-L suggest superior performance in producing captions that are not only accurate but also diverse and context-sensitive.

The performance of the proposed model (ICTGAN) highlights the importance of advanced metrics in ICG. The BLEU-1 and BLEU-4 scores reflect the accuracy of individual n-grams within captions, while ROUGE-L emphasizes sentence-level structure, measuring fluency and relevance. CIDEr, which assigns weights to the consensus phrases found across captions, is particularly useful in assessing the relevance of captions to human perception. ICTGAN's high scores in BLEU-4 and CIDEr, compared with other models, suggest it captures a balance between linguistic accuracy and semantic relevance, while its top ROUGE-L score further supports its fluency and coherence. ICTGAN's integration of ViT and GAN technologies exemplifies a sophisticated approach to image caption generation. Through an iterative refinement process, ICTGAN pushes the limits of previous models by achieving a comprehensive understanding of visual content. Its top performance across BLEU-4, CIDEr, and ROUGE-L demonstrates that ICTGAN provides captions that are not only precise and contextually aligned with image content but also resonant with human evaluation criteria. This progression in ICG research shows a promising direction toward models that more closely emulate human perception and understanding, pushing the potential applications in AI-driven visual understanding and automated captioning tools.

6. Conclusions

This research presents the ICTGAN model as a comprehensive approach to automatic image captioning, leveraging advanced machine-learning techniques and robust datasets. The MS COCO dataset, renowned for its diversity and richness of visual content and associated captions, served as the foundation for model training and evaluation. The model architecture, which is composed of Long Short-Term Memory (LSTM) units, a head self-attention model, and Vision Transformers (ViTs) makes it simpler to produce coherent captions and successfully extract rich visual representations from input images. Further, Generative Adversarial Networks (GANs) and Wasserstein Loss as the adversarial loss function enhanced the model's capacity to produce diverse captions. The model's performance was measured statistically and objectively using the BLEU, CIDEr, and ROUGE scores, to assess the quality of the generated captions. When applied to various applications needing automatic photo captioning, the trained models proved their worth and improved the fields of computer vision and natural language processing. The findings of this study demonstrate the potential of cutting-edge machine-learning approaches for creating good, contextually suitable image captions. The findings of this work make a significant contribution to the field of image captioning, opening up new opportunities for future research and application. The research's commitment to rigorous methodology, objective evaluation, and practical applicability sets a high standard for future research on this topic.

Author Contributions: The manuscript was written through the contributions of all authors. Data curation, S.T., V.V. and S.G.; Formal analysis, S.T. and S.G.; Funding acquisition, O.A.O.; Investigation, V.V., J.B.A. and A.O.B.; Methodology, S.T., V.V., M.V. and J.B.A.; Project administration, A.O.B.; Resources, M.V. and A.O.B.; Software, O.A.O., M.V. and J.B.A.; Supervision, J.B.A.; Validation, O.A.O. and S.G.; Writing—review and editing, M.V. All authors have read and agreed to the published version of the manuscript.

Funding: Walter Sisulu University, East London, South Africa 5200 Komani Campus, South Africa.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghandi, T.; Pourreza, H.; Mahyar, H. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.* **2023**, *56*, 1–39. [CrossRef]
2. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
3. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
4. Hu, S.; Shen, Y.; Wang, S.; Lei, B. Brain MR to PET Synthesis via Bidirectional Generative Adversarial Network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, Proceedings of the 23rd International Conference, Lima, Peru, 4–8 October 2020*; Part II 23; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 698–707.
5. Rinaldi, A.M.; Russo, C.; Tommasino, C. Automatic image captioning combining natural language processing and deep neural networks. *Results Eng.* **2023**, *18*, 101107. [CrossRef]
6. van der Lee, C.; Krahmer, E.; Wubben, S. Automated Learning of Templates for Data-to-Text Generation: Comparing Rule-Based, Statistical, and Neural Methods. In Proceedings of the 11th International Conference on Natural Language Generation, Tilburg, The Netherlands, 5–8 November 2018; pp. 35–45.
7. Hill, T.; Lewicki, P.; Lewicki, P. *Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining*; StatSoft, Inc.: Tulsa, OK, USA, 2006.
8. NIST/SEMATECH. *E-Handbook of Statistical Methods*; NIST/SEMATECH: Gaithersburg, MD, USA, 2012.
9. Hochreiter, S. *Long Short-Term Memory*; Neural Computation MIT-Press: La Jolla, CA, USA, 1997.
10. He, S.; Liao, W.; Tavakoli, H.R.; Yang, M.; Rosenhahn, B.; Pugeault, N. Image Captioning Through Image Transformer. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
11. Liu, W.; Chen, S.; Guo, L.; Zhu, X.; Liu, J. Cptr: Full transformer network for image captioning. *arXiv* **2021**, arXiv:2101.10804.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
13. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. *arXiv* **2018**, arXiv:1807.00734.
14. Chen, C.; Mu, S.; Xiao, W.; Ye, Z.; Wu, L.; Ju, Q. Improving Image Captioning with Conditional Generative Adversarial Nets. In Proceedings of the AAAI Conference on Artificial Intelligence, 27 January–1 February 2019; Volume 33, pp. 8142–8150.
15. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H.; Bennamoun, M. Text to image synthesis for improved image captioning. *IEEE Access* **2021**, *9*, 64918–64928. [CrossRef]
16. Donahue, J.; Krähenbühl, P.; Darrell, T. Adversarial feature learning. *arXiv* **2016**, arXiv:1605.09782.
17. Mishra, S.; Seth, S.; Jain, S.; Pant, V.; Parikh, J.; Jain, R.; Islam, S.M. Image Caption Generation using Vision Transformer and GPT Architecture. In Proceedings of the 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2–3 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.

18. Sharma, H.; Srivastava, S. Graph neural network-based visual relationship and multilevel attention for image captioning. *J. Electron. Imaging* **2022**, *31*, 053022. [CrossRef]
19. Ondeng, O.; Ouma, H.; Akuon, P. A review of transformer-based approaches for image captioning. *Appl. Sci.* **2023**, *13*, 11103. [CrossRef]
20. Kolla, T.; Vashisth, H.K.; Kaur, M. Attention Unveiled: Revolutionizing Image Captioning through Visual Attention. In Proceedings of the 2023 Global Conference on Information Technologies and Communications (GCITC), Bengaluru, India, 1–3 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
21. Zhang, H.; Qu, W.; Long, H.; Chen, M. The Intelligent Advertising Image Generation Using Generative Adversarial Networks and Vision Transformer: A Novel Approach in Digital Marketing. *J. Organ. End User Comput. (JOEUC)* **2024**, *36*, 1–26. [CrossRef]
22. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September, 2014*; Part V 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
23. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
24. Lala, C.; Madhyastha, P.S.; Scarton, C.; Specia, L. Sheffield submissions for WMT18 multimodal translation shared task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, 31 October–1 November 2018; pp. 624–631.
25. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
26. Oluborode, K.; Kadams, A.; Mohammed, U. An Intelligent Image Caption Generator Model Using Deep Learning. *Int. J. Dev. Math. (IJDM)* **2024**, *1*, 162–173. [CrossRef]
27. Alexey, D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Ashish, V. Attention is All You Need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; Volume 30, Part I.
29. Papineni, K. *BLEU: A Method for Automatic Evaluation of MT*; Research Report, Computer Science RC22176 (W0109-022); IBM T. J. Watson Research Center: Yorktown Heights, NY, USA, 2001.
30. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Hinton, G.; Srivastava, N.; Swersky, K. Neural networks for machine learning. *Coursera Video Lect.* **2012**, *264*, 2146–2153.
32. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
33. Brownlee, J. *A Gentle Introduction to Mini-Batch Gradient Descent and How to Configure Batch Size*; Machine Learning Mastery: San Juan, PR, USA, 2019.
34. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
35. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-Based Image Description Evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
36. Vidyabharathi, D.; Mohanraj, V.; Kumar, J.S.; Suresh, Y. Achieving generalization of deep learning models in a quick way by adapting T-HTR learning rate scheduler. *Pers. Ubiquitous Comput.* **2023**, *27*, 1335–1353. [CrossRef]
37. Ayinde, B.O.; Nishihama, K.; Zurada, J.M. Diversity Regularized Adversarial Deep Learning. In *Artificial Intelligence Applications and Innovations, Proceedings of the 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, 24–26 May 2019*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 292–306.
38. Santiesteban, S.S.; Atito, S.; Awais, M.; Song, Y.Z.; Kittler, J. Improved Image Captioning Via Knowledge Graph-Augmented Models. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 4290–4294.

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