Perceptive Multimodal Generative AI

Rithani M.
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
m_rithani@ch.amrita.edu

Sidesh Sundar S.
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
sidesh260303@gmail.com

Vimal Dharan N.
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
vimalvimal1293@gmail.com

SyamDev R. S.
Department of AI and ML
New Horizon College of Engineering
Bangalore, India
rssyamdev@gmail.com

***Abstract*—Generative AI has moved very fast in all directions-from text generation to image synthesis. However, most AI models remain specialized in one particular task rather than enabling the workflows involving other kinds of media. Perceptive Multimodal Generative AI is a holistic platform that outlines creative workflow enhancement through five core functionalities: comic generation, image style transfer, extraction of text from visual content, template customization, and creation of HE content. The system applies advanced artificial approaches, deep learning, NLP, and OCR in order to make it possible for users to go about their activities with no disturbance.**

The functionality of creating a comic allows users to create fully customizable comics through the integration of narrative generation and image synthesis. Based on VGG19 and GAN-based models, that style could be given to images, maintaining some critical features in its content, such as color and edges.

The text extraction ability means that users can gain access to textual data from images and reuse this data, therefore allowing smooth creative editing. The feature of model customization adjusts AI models for specifying a particular user's preferences, thus ensuring that content agrees with their artistic intentions. Lastly, the platform supports an array of AI-supercharged content-generating activities, including text-to-image generation and narrative creation, and increases opportunities for customized and multimodal outputs. The Perceptive Multimodal Generative AI has capitalized on the need of the wide casual and professional creators for creating friendliness and significantly increased content format variability as an attempt to successfully develop fluid interaction among text, images, and many other media. Scalability and personalization enable the product to create creativity and flex the form, accessibility, and efficiency of AI-driven content creation. It fundamentally redrafts the role of AI in digital creativity, where users can now generate, adapt, and discover new modes of expression seamlessly in the system.

***Index Terms*—Generative AI, Multimodal Content Creation, Text Generation, Image Synthesis, Comic Generation, Image Style Transfer, Optical Character Recognition (OCR), Natural Language Processing (NLP), Model Personalization,AI Content Detection, Deep Learning(DL).**

I. Introduction

Significant advancements in generative AI have made it possible to create various types of material in several modalities, such as text, graphics, and audio. Recent developments in text generation models like GPT-3 and GPT-4 and picture style transfer models like StyleGAN have demonstrated these models' remarkable capabilities in particular fields. The utilization of these models has led to a significant transformation in the approach towards creative work. They enable the production of tailored outcomes, automated content creation, and stylistic modifications [1] [13] . However, let us now discuss the limitations of this technology.Recent rapid developments in artificial intelligence (AI) tools have highlighted the necessity for a robust system that allows users to work in multiple modes simultaneously. Therefore, the objective of this project is to develop a multimodal AI platform that promotes collaboration among different modalities while also improving the distinctive traits of each generative model.The five main which will be the focus of this discussion and the features of the proposed platform are comic generation, image style transfer, image text extraction, template customization, and AI content generation [2].

A. Comic Generation:

Making creative comics with AI-powered storytelling and illustration tools can increase user engagement and creativity. When text production and photos are combined, users may easily create stories that resonate with their audience [3].

B. Image Style Transfer:

This model makes use of advanced neural networks, which allows users to add different artistic styles to their photos, turning ordinary shots into visually striking works of art. This approach serves as a tool to increase the originality of visual storytelling in addition to showcasing the application of AI in artistic expression [4].

Users can concentrate on ideation and refining instead of the technical parts of content generating by streamlining the creative process. To enable people to fully realise the potential of generative AI, this research aims to make the technology more creative and approachable.[5] This platform strives to close the gap between various modes and encourage a more seamless and comprehensive creative experience by providing a range of content creation options and an easy-to-use interface. By combining all of these features into one

platform, people from all backgrounds can take part in and profit from AI-driven creative processes, democratising access to the newest technological advancements [4].

II. Literature Review

AI has taken some recent critical steps toward achieving the elusive goal of molding multifaceted dialogue systems and comic generation, particularly in multimodal approaches, consistency in persona features, and integration with visual styles. Agrawal et al. (2023) proposed an excellent multimodal approach to generating comic dialogue by using the MPDIALOG dataset, which is based on English comics but gives extensive importance to data use ethics. Their work fills a critical lacuna in current state-of-the-art models for dialogue that would be missing systems which can manage visual context as well as persona consistency in multiple character interactions. The methodology they used in persona modelling improved both the naturalness and coherence of dialogue while keeping the authenticity of characters that are essential in comic books.[6]

A closely related study is the historical context of digital comic generation processes discussed by Broden (2024). A comparison was done of the piece published in "The Journal of Graphic Novels and Comics" between digital and traditional styles. Paper acknowledged that further research is required to understand how digital technology transforms artistic expressions in the comics medium but, at the same time, presents an important influence that these tools have had on the aesthetic character of comic art [7].

In a similar line, Yang et al. (2019) introduced emotion-driven text balloons with multi-page layouts to solve problems in the development of self-driving comics. The approach required very little user effort and produced great comic pages; however, they did point out a few limitations to entirely automate systems that have appealing layouts as well as emotion-aware discourse [8]. WEN et al. (2023) meanwhile propose a model of emotion generation which is based on the Valence-Arousal-Dominance space. It includes personality traits in the model, so as to make generated dialogs more believable. Based on their study, modeling mood shifts significantly enhance the correctness of emotion generation, especially for minority emotions that can produce more complex character interaction. [9]

The recent surge in interest in the field of computer vision about parallel advancements in image transfer, specifically artistic style transfer that retains the content of the original image has been largely focused on tackling the essential challenge of balancing computing efficiency, artistic integrity, and reliability [10]. Chen et al (2017) propose the DualAST model, which generalizes previous methods. It captures unique properties of single artists and general creative styles. The above model was more representative of fine-grained artistic details but not scalable because of the computational cost of iterative optimization, especially in real-time applications. [11]

Perhaps, the most significant contribution in this domain has been made by B.M.G et al. in the generation of Generative Adversarial Networks (GANs). GANs can synthesize highly realistic stylized images while adversistically be able to retain the content of the original image. Thus, GANs become more suitable for real-time applications. However, mode collapse and large sets of data to ensure consistent results across the spectrum of artistic fields remain areas of ongoing research. [12]

More recent developments incorporated the attention mechanism into neural network models, as is demonstrated in Cheng et al. (2021). Attention models allow a network to selectively focus on regions critical for both content and style preservation, giving greater control over the transfer process and issues of artifacts and distortions. However, this comes at the price of computational complexity, especially when processing high-resolution images. [12]

Altogether, it revealed that the advancements of comic generation and picture transfer are the continuous pursuit to make the usage of AI in content generation more realistic, coherent, and creative. At the same time, they draw attention toward the necessity of resolving methodological and computational problems so that those systems can be further advanced.

III. Methodology

In our study, we propose an enhanced version of ChatGPT to produce consistent and insightful visual responses—particularly for color enhancement tasks—we present a pure version of the system in this work. as well as side defense. Artificial intelligence and image processing techniques are used in advanced modeling to produce meaningful conversations and accurate results. It has also improved a lot of performance features. and much more. This method generates model outputs based on perception and context by utilizing artificial adversarial networks (GAN), attention strategies, and cognitive loss functions.

A. *Comic Generation*

*1) **Image Generation Using Stable Diffusion**:* The visual creation, which kicks off the comic development process, is driven by the Stable Diffusion model. The figure 1 is an the how this process works:

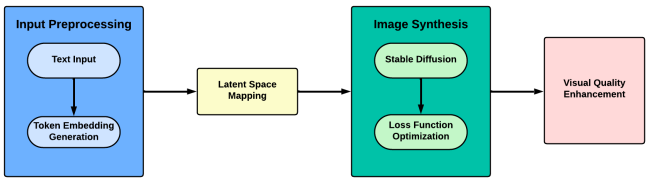

Fig. 1. Comic image generation

1.1 Input Preprocessing

The image generation process starts with a text-based input that describes the comic panel, including specifics about characters, actions, and the setting. For example, a prompt might read, "A superhero flying over a city during sunset." This text is then processed using a pretrained

text encoder, such as CLIP, which converts it into token embeddings—essentially numerical representations capturing the text's meaning. These embeddings serve as the foundation for further processing within the model. [13]

1.2 Latent Space Representation

Next, the tokenized embeddings are transformed [**?**] into a latent space representation. This is a high-dimensional, compact mathematical abstraction of the image to be generated, encoding the image's structure, style, and content. The transformation into this latent vector, z, is performed using a mapping function f, as shown below:

$$z = f(\text{embedding}(prompt)) \tag{1}$$

The latent vector z is crucial, as it forms the underlying framework upon which the final image is built.

1.3 Image Synthesis

With the latent representation in place, the actual image synthesis begins. The Stable Diffusion model [15] iteratively refines a noisy image until it becomes a clear and coherent visual that aligns with the original prompt. This is achieved using a diffusion process driven by the loss function $\mathcal{L}(x_t)$, which measures how closely the current image matches the intended result:

$$x_{t+1} = x_t - \alpha \nabla_x \mathcal{L}(x_t) \tag{2}$$

Here, x_t is the image at step t, and α is a learning rate parameter that guides the model towards a crisp, well-defined image.

1.4 Post-Processing

Post-processing techniques were used to make the visual quality even better and to maintain an aesthetic coherence along with the rest of the panels once the image synthesis has been generated. Color correction, resolution scaling, and fine-tuning of all kinds of visual artifacts could be included in the process. The generated image is well-executed and flows ideally with the overall style of the comic [16].

*2) **Dialogue Generation Using PersonaGPT**:* For the dialogue in each comic panel, **PersonaGPT** is used to ensure that the characters' speech aligns with their established personas. The figure 2 represents the steps involved in the process:

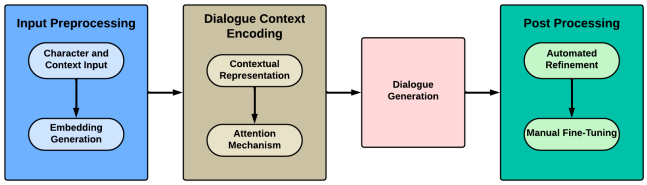

Fig. 2. Comic dialogue generation

2.1 Input Preprocessing

The creation of the dialogue starts with the specification of the character's identity and context of the scene. The context forms the needed background for the conversation, while the persona consists of traits like tone, vocabulary, and attitude. For example, a courageous hero speaks differently than a shy sidekick. All these inputs are then encoded into embeddings that can be used by the model to generate the dialogue in a suitable setting.

2.2 Dialogue Context Encoding

PersonaGPT represents scene information and background from previously created conversations using the transformer-based architecture. During encoding, it ensures the discussion is coherent and flows from previous discussions [17]. With respect to the requirement of making responses logical and authentic to the character's identity, attention mechanism helps the model focus in on relevant parts of dialogue history. It can be represented as the probability of generating a certain sentence:

$$p(d_t|d_{<t}, c) = \prod_{i=1}^{n} p(d_{t,i}|d_{<t,i}, c) \tag{3}$$

In this equation, $d_{<t}$ represents the preceding dialogue, while c is the context vector that encapsulates both the character's persona and the ongoing narrative [18].

2.3 Dialogue Generation

The model then generates the character's dialogue in an autoregressive manner once it has encoded the context. This would ensure that the flow is natural, based each word and sentence on the context, as well as the discourse that has already taken place. During the process, the persona is considered-this therefore customizes the dialogue to the singular nature of the character.

2.4 Post-Processing and Refinement

Once the dialogue is written, it can be refined and read fluently and clearly. It may involve grammatical correction as their sentences may become shorter or longer, and ensuring the tone and style of the dialogue are set right according to the scene's requirement. Now, the speech of the character may be manually edited and improved to suit the story of the comic.

B. STYLE TRANSFER:

The figure 3 shows the complete work flow of the style transfer process. This includes input processing , color space conversion and training with neural networks.

1. Input Preprocessing

Preprocessing attempts to improve the quality of images by preparing for later operations, which may include color enhancement and style transfer. In this step, it minimizes noise, adjusts pixel intensity, and increases contrast and sharpness.

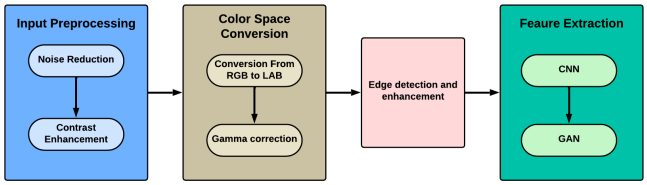

Fig. 3. Methodology for style transfer

One of the most dominant techniques applied is Adaptive Histogram Equalisation [17] (AHE). It even improves the contrast through local alteration of the pixel intensity distribution. It works perfectly on images with various random lighting conditions. The AHE formula is:

$$C(x,y) = \frac{I(x,y) - \min(I)}{\max(I) - \min(I)}\ n$$

where $I(x,y)$ is the contrast pixel value and $C(x,y)$ is the original intensity at pixel (x, y). The Gaussian function,

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

This operator is used to apply additional smoothing with Gaussians in order to reduce noise. The standard deviation σ determines the amount of smoothing; this technique allows noise reduction while the structure in the picture remains intact. Multiscale correction techniques are capable of applying refinement at various degrees of resolution levels. Another important property of the proposed upsampling and subsampling operation at different scales is the change of local contrast and sharpness. This provides for a better general quality of an image as well as more conspicuous features in it. Pre-processing methods used here prepare images for any further operations, like style transfer or color enhancements, making the algorithms work more efficiently and giving better results.

2. Color space conversion

Images are first converted from the RGB color space to a perceptually uniform space, such as CIELAB (LAB), for enhancing image processing because in this space, brightness and hue can be adjusted without affecting each other due to its ability to separate luminance (L) from chrominance components, A and B. A step involved between RGB to LAB conversion usually involves a transformation of RGB values into the XYZ color space. And by equations such as those mentioned above, the XYZ value transforms into the LAB color space. [18]

$$L = 116 \cdot f\left(\frac{Y}{Y_n}\right) - 16$$

$$a = 500 \cdot \left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right]$$

$$b = 200 \cdot \left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right]$$

The white reference values are represented by X_n, Y_n, Z_n. The function $f(t)$ is defined as:

$$f(t) = \begin{cases} t^{1/3} & \text{if } t > \epsilon \\ 7.787 \cdot t + \frac{16}{116} & \text{otherwise} \end{cases}$$

with $\epsilon = 0.008856$.

Gamma correction, which modifies the brightness of the image to enhance visual quality, is expressed as:

$$I_c = I^{\frac{1}{\gamma}}$$

where I is the initial pixel intensity and γ is the correction factor.

3. Edge detection and enhancement

Essentially, an image being processed should be structurally preserved. Sobel, Canny, and Laplacian edge detection algorithms identify areas containing abrupt changes in intensity, which usually represent the edges or boundaries of objects in an image. For example, the Sobel operator uses convolutional masks to compute gradients along both the x and y axes. The magnitude of the gradient is calculated as follows:

$$G = \sqrt{G_x^2 + G_y^2},$$

where G_x and G_y represent the horizontal and vertical gradients. To improve the edges, a blur masking technique can be applied:

$$I_{\text{sharp}} = I + \alpha(I - I_{\text{blurred}}),$$

where I is the original image, I_{blurred} is the blurred version of the image, and α controls the sharpening intensity. Other techniques, such as histogram stretching, can be used to improve the visibility of edges. In addition, non-local means (NLM) filtering is effective in reducing noise while preserving important edge information, thus improving image clarity. [19]

4. Feature Extraction and Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [20] are used to extract hierarchical features that capture global content and local details in images. The convolution operation can be expressed as:

$$(I * K)(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x-i, y-j) K(i,j),$$

where I is the input image, K is the kernel, and k is the size of the kernel.

For advanced image enhancement, Generative Adversarial Networks (GANs) can be used. The loss function for a GAN, which measures the similarity between the generated image $G(z)$ and the target image x, is defined as:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))],$$

where D is the discriminator, G is the generator, and z is a random noise vector.

Adaptive instance normalization (AdaIN) is a technique used in style transfer to approximate the mean and variance of content and style feature maps. It is represented as follows:

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y),$$

where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the content feature map, and $\mu(y)$ and $\sigma(y)$ are those of the style feature map.

IV. Result

In Figure 4, we show the content image and the style image used to produce the stylised image in Figure 5. To evaluate how the main features of the original content image are preserved in in the output, we used an edge map. A edge detection map of the input and output images shows the most important lines and components of the images.

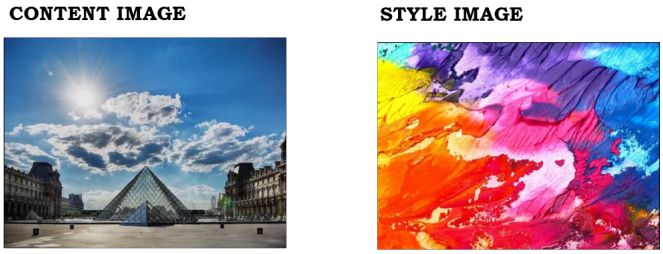

Fig. 4. Content Input and Style input

Comparing the edge maps of the content image and the composited image, we can see that the original structure of the original image is preserved in the final product. This comparison ensures that even though the image has been used from the original image, the important features, such as shapes and dimensions, of the original image are not lost or distorted and figure 6 shows it. This approach helps to verify that the conversion process respects the integrity of the original image, and ensures that key information is preserved in the image output. The figure 7 and figure 8 present the final comic generated by the model we trained. The model was designed to process input images, apply various artistic styles, and generate comic-like visuals based on the training data. These figures demonstrate how the model successfully translated the input data into comic-style imagery, incorporating both visual elements and textual components. The process involved

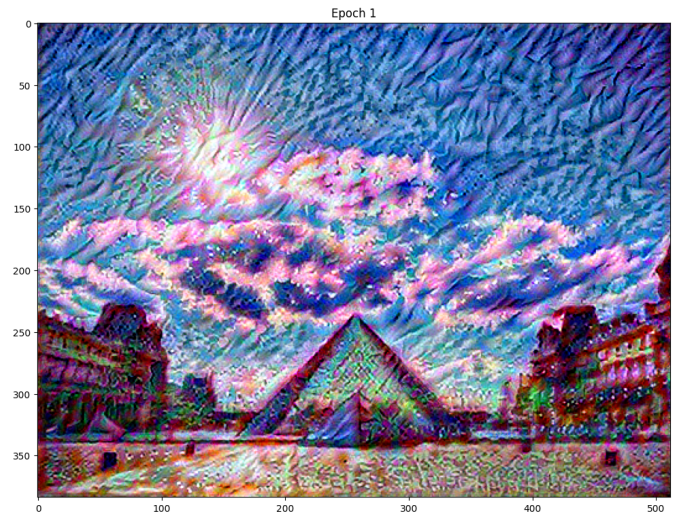

Fig. 5. Stylised Image

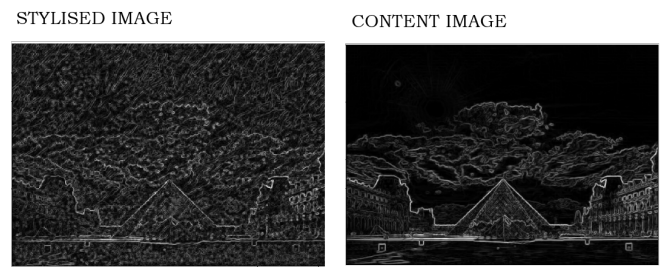

Fig. 6. Edgemap for input and output

training the model to recognize patterns in the input images and blend them with specific comic-style features such as line art, shading, and color adjustments. Additionally, the model was fine-tuned to position speech bubbles and text within the frames, ensuring that the narrative elements fit naturally into the overall composition.

The generated comics in Figures 7 and 8 serve as an example of how the model can be applied to create visually appealing and coherent comic panels. The results illustrate the

Fig. 7. Sequence 1 for comic generation output

effectiveness of the model in balancing artistic style with the preservation of key features from the original images, thus creating a seamless comic format.

Fig. 8. Sequence 2 for comic generation output

V. Conclusion and Future Work

In conclusion, our innovative approach to provide a chatbot which in terms of usability is very rich with all its features, and which has possibilities to enhance the processing capability of real-time data. With the use of open-source data resources like Wikipedia and PubMed, the fetching of further processing of information is more accurate with this chatbot. The vast scope increased by images because of advanced image processing algorithms and caption generation techniques in their application areas expand its range. Additionally, this will leverage the strength of Retrieval-Augmented Generation (RAG), generalize well across various sets, and minimize the risk of plagiarism in AI-generated content. In addition, it integrates techniques for both image translation and 2D-to-3D conversion with better features of image processing.

References

[1] Solanki, A. (2024). Advancements in Artificial Intelligence: A Comprehensive Review and Future Prospects. *International Journal of Computer Engineering Technology*, 2, 53-64.

[2] Jaikumar, V., A. Karunamurthy, Dr. (2023). Recent Advancements in Artificial Intelligence Technology: Trends and Implications. *Quing International Journal of Multidisciplinary Scientific Research and Development*, 2(1), 1-11. https://doi.org/10.54368/qijmsrd.2.1.0003.

[3] Yang, X., Ma, Z., Yu, L., Cao, Y., Yin, B., Wei, X., Zhang, Q., Lau, R. (2021). Automatic Comic Generation with Stylistic Multi-page Layouts and Emotion-driven Text Balloon Generation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2101.11111.

[4] Zhang, F., Zhao, H., Sun, X., Wu, Y., Li, Y. (2023). CBA-GAN: Cartoonization style transformation based on the convolutional attention module. *Computers and Electrical Engineering*.

[5] Taherdoost, H., Madanchian, M. (2024). AI Advancements: Comparison of Innovative Techniques. *AI*, 5, 38-54. https://doi.org/10.3390/ai5010003.

[6] Agrawal, H., Mishra, A. M., Gupta, M., Mausam. (2023). Multimodal Persona Based Generation of Comic Dialogs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

[7] Broden, D. (2024). Refiguring comic book imagery in a post-digital moment: Conspicuousness, hybridity and ontological strangeness. *Journal of Graphic Novels and Comics*.

[8] Wen, Z., Cao, J., Shen, J., Yang, R., Liu, S., Sun, M. (2023). Personality-affected Emotion Generation in Dialogue Systems. *ACM Transactions on Information Systems*.

[9] Yang, X., Ma, Z., Yu, L., Cao, Y. (2019). Automatic Comic Generation with Stylistic Multi-page Layouts and Emotion-driven Text Balloon. *Image and Vision Computing*.

[10] Enhancing Brain Tumor Diagnosis with Generative Adversarial Networks, 2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India, 2024, pp. 846-851. https://doi.org/10.1109/Confluence60223.2024.10463384.

[11] Chen, H., et al. (2021). DualAST: Dual Style-Learning Networks for Artistic Style Transfer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 872-881. https://doi.org/10.1109/CVPR46437.2021.00093.

[12] Wu, H., Sun, Z., Zhang, Y., Li, Q. (2023). Direction-aware neural style transfer with texture enhancement. *Neurocomputing*.

[13] Mohan, G. B., Kumar, R. P., Parathasarathy, S., Aravind, S., Hanish, K. B., Pavithria, G. (2023). Text Summarization for Big Data Analytics: A Comprehensive Review of GPT-2 and BERT Approaches. In R. Sharma, G. Jeon, Y. Zhang (Eds.), *Data Analytics for Internet of Things Infrastructure*. Springer. https://doi.org/10.1007/978-3-031-33808-3_14.

[14] Wang, J., Xu, L., Yu, X. (2023). A non-definitive auto-transfer mechanism for arbitrary style transfers. *Knowledge-Based Systems*.

[15] Cheng, J., Han, Z., Wang, Z. (2023). Stylized image denoising via noise style transfer and Quasi Siamese network. *Signal Processing: Image Communication*.

[16] Ye, W., Li, C., Chen, Y., Liu, Y., Zhou, H. (2023). Multi-style transfer and fusion of image's regions based on attention mechanism and instance segmentation. *Signal Processing: Image Communication*.

[17] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., Haar Romeny, B. T., Zimmerman, J. B., Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*. https://doi.org/10.1016/S0734-189X(87)80186-X.

[18] Nishad, P. M., Chezian, R. (2013). Various colour spaces and colour space conversion algorithms. *Journal of Global Research in Computer Science*, 4, 44-48.

[19] Ganesan, P., Sajiv, G. (2017). A comprehensive study of edge detection for image processing applications. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2017, pp. 1-6. https://doi.org/10.1109/ICIIECS.2017.8275968.

[20] Rithani, M., Kumar, R. P., Doss, S. (2023). A review on big data based on deep neural network approaches. *Artificial Intelligence Review*, 56, 14765–14801. https://doi.org/10.1007/s10462-023-10512-5.