iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62376

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

FashionMorph: Contextually Adaptive Clothing Replacement with CLIP, Segmentation, and Stable Diffusion

Kashish J. Goel[1], Aditya Kadgi[2], Kanika Gorka[3], Prof. Milind Kamble[4]
Department of Electronics and Telecommunications Vishwakarma Institute of Technology, Pune, India

Abstract: ***FashionMorph represents a groundbreaking advancement in digital image manipulation, seamlessly blending advanced segmentation techniques with the language comprehension capabilities of CLIP. This innovative approach simplifies the process of precise garment replacements while ensuring the preservation of realistic textures and boundaries. By incorporating stable diffusion, FashionMorph eliminates the need for extensive training on specific clothing styles, offering users a streamlined experience in expressing their preferences through natural language prompts. This empowers users in photo editing, granting them unparalleled creative freedom. The system comprises a comprehensive pipeline designed for the identification and inpainting of clothing elements. Leveraging CLIP and CLIPSeg models for robust image analysis, FashionMorph accurately identifies clothing elements and generates corresponding masks. The inpainting process utilizes a stable diffusion pipeline, resulting in impressive image quality that surpasses the performance of notable models like DCTransformer, StyleGAN, and ADM. Evaluations based on FID and SFID scores, along with Inception Score assessments, further validate FashionMorph's effectiveness in context-aware garment editing. Visual representations of original images, inpainted results, and masks vividly demonstrate the successful outcomes achieved by FashionMorph, establishing it as an efficient and user-friendly solution for professionals and enthusiasts alike. Among the evaluated models, Stable Diffusion emerges as the standout performer, boasting remarkable FID and SFID scores of 0.35 and 0.20, respectively, indicative of its outstanding image fidelity and diversity. Conversely, DCTransformer exhibits the highest FID and SFID scores, while StyleGAN excels in diversity but falls short in FID. ADM (dropout) strikes a balance between fidelity and diversity. These results underscore the exceptional performance of Stable Diffusion, solidifying its position as the leading approach with the lowest FID and SFID values.***

Keywords- CLIP (Contrastive Language-Image Pertaining), Segmentation, Stable diffusion models, Natural language prompts, Garment replacement

I. INTRODUCTION

In the realm of digital image manipulation, pushing the boundaries of creativity and realism has been an ongoing pursuit. A longstanding challenge has been achieving seamless garment replacement within images while upholding contextual accuracy and visual fidelity. In response to this challenge, we introduce FashionMorph, a groundbreaking solution that combines advanced segmentation techniques with the language comprehension capabilities of CLIP (Contrastive Language-Image Pre-training). FashionMorph represents a paradigm shift, empowering users to achieve context-aware garment replacements guided by their directives while maintaining the authenticity of textures and boundaries. A notable feature is its integration of stable diffusion, an advanced method that eliminates the need for exhaustive model training on specific clothing styles or attributes. This innovation enables users to effortlessly communicate their preferences through natural language prompts, simplifying the intricate process of garment replacement in photo editing. FashionMorph not only streamlines these tasks but also enhances the artistic potential for both professionals and enthusiasts, revolutionizing digital image manipulation.

FashionMorph's key strength lies in its implementation of stable diffusion, distinguishing it from conventional methods. Unlike traditional techniques that require meticulous model training for different clothing styles, stable diffusion excels in preserving the intricate textures and boundaries that define garments. This unique approach ensures the preservation of the user's intent while seamlessly integrating replacement garments into the image context. By eliminating the need for extensive training, FashionMorph significantly reduces the time and effort required for garment replacement tasks, offering an efficient and accessible solution for users of all backgrounds and expertise levels.

Whether for professional photo editing or creative experimentation, FashionMorph's stable diffusion is poised to redefine standards of realism and ease in garment replacement.

In today's digital environment, where visual storytelling and creative expression are paramount, FashionMorph emerges as a game-changing tool for artists, designers, and editors. Its user-friendly interface, coupled with intuitive instructions, democratizes the art of garment replacement. By leveraging CLIP's language comprehension capabilities, FashionMorph bridges the gap between user intent and image manipulation, enabling precise and context-aware garment editing without the need for specialized technical knowledge. As we continue to explore the possibilities of digital imagery, FashionMorph redefines the boundaries of achievable outcomes, inviting professionals and enthusiasts alike to explore the limitless potential of context-aware garment replacements in digital image manipulation.

II. WORKING PREMISE OF DIFFUSION MODEL

Diffusion models represent a potent category of generative models renowned for their capacity to produce high-quality data samples. The underlying principle of diffusion models is elegantly simple yet remarkably effective. These models operate on input data, denoted as x_0, and gradually introduce Gaussian noise across a sequence of T steps, a process commonly known as the forward diffusion process. It's essential to note that this process differs from the forward pass of a neural network and serves as a pivotal element for generating targets utilized in training generative models. The objective here is to increase the complexity of the data by introducing noise.

Subsequently, a neural network is trained to reverse the noise-introduced data and reconstruct the original input. This capability to model the reverse process plays a pivotal role in generating novel data samples. This reverse process is often referred to as the reverse diffusion process or simply the sampling process of a generative model.

Implementing diffusion models involves establishing a Markov chain with T steps. In this context, a Markov chain signifies that each step depends solely on the preceding one, which is a rational assumption. Crucially, diffusion models don't confine the use of specific neural network architectures, providing a level of flexibility not found in flow-based models.

A. *Forward Diffusion*

The forward diffusion process, symbolized as $q(x_t|x_{t-1})$, incorporates Gaussian noise with a variance β_t into the preceding latent variable x at every step of the Markov chain. Mathematically, this diffusion process can be represented as [equation (1)].

$$q(x_t|x_{t-1}) = N(x_t; \mu_t = \sqrt{1-\beta_t}x_{t-1}, \Sigma_t = \beta_t I). \quad (1)$$

Equation 1: Forward diffusion equation

The primary objective of the forward diffusion process, depicted in Figure 1, is to transition from the initial input data x_0 to x_t in a manageable way. Mathematically, this transition reflects the posterior probability and can be simplified as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (2)$$

Equation 2: Reparametrized forward diffusion equation

Generating samples for each timestep, such as t=500, through the application of q for 500 iterations might pose computational inefficiencies. However, the reparameterization trick provides a remedy for this challenge. This trick enables tractable closed-form sampling at any timestep by introducing the parameters $\alpha_t = 1 - \beta_t$ and $\bar{\alpha} = \prod_{s=0}^{t} \alpha_s$. This formulation facilitates noise sampling at any desired timestep, enhancing computational efficiency. The reparameterization trick is represented as [equation (2)].

B. *Reverse Diffusion*

As T tends towards infinity, the latent variable x_t converges towards an isotropic Gaussian distribution. To generate new data points, it is crucial to learn the reverse distribution $q(x_t|x_{t-1})$. Directly computing this reverse distribution is impractical due to its reliance on the data distribution. However, it can be approximated using a parameterized model $p_\theta(x_{t-1}|x_t)$. Typically implemented as a neural network, this parameterized model enables the prediction of the mean and variance of the reverse distribution. This diffusion process is mathematically represented as [equation (3)].

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t t)) \quad (3)$$

Equation 3: Reverse diffusion equation

C. Training

To train a diffusion model, the reverse Markov transitions with the highest likelihood based on training data are determined. Training entails minimizing the variational upper bound on the negative log likelihood. Central to this optimization process is the evidence lower bound (ELBO). The ELBO incorporates terms pertaining to reconstruction accuracy, the proximity of x_t to a standard Gaussian distribution, and the disparity between desired and approximated denoising steps. This formulation is represented as [equation (4)].

$$\mathbb{E}[-logP_\theta(x_0)] \leq \mathbb{E}[-log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] \quad (4)$$

Equation 4: Negative log-likelihood

D. Architecture

In practice, diffusion models leverage architectures such as U-Net, which incorporate Wide ResNet blocks, group normalization, and self-attention blocks. U-Net is especially effective in preserving input and output sizes while establishing skip connections between encoder and decoder blocks, facilitating the generation of high-quality data samples.

In essence, diffusion models provide a robust framework for generative modeling by gradually introducing noise and mastering the reversal process. This methodology has demonstrated promising outcomes in generating realistic data samples across diverse domains.

III. LITERATURE REVIEW

The literature review underscores the critical role of AI in law enforcement, particularly in suspect identification and forensic sketching, where traditional methods relying on eyewitness accounts and manual sketches encounter limitations in accuracy and efficiency. To overcome these challenges, the proposed system utilizes Stable Diffusion AI, a cutting-edge deep learning technique, to generate realistic facial images of suspects based on text descriptions, providing a more efficient and accurate solution compared to previous methods. Additionally, the review discusses related works, including the application of diffusion models for image synthesis and advancements in generative models like GANs and VAEs, which have influenced the design of the proposed system. Overall, the literature emphasizes the potential of AI-driven solutions to enhance law enforcement investigations, with a specific emphasis on improving suspect identification and forensic sketching [1]. This research presents novel approaches aimed at enhancing user control in generative text-to-image models, such as Stable Diffusion. Conventionally, users adjust textual prompts to generate desired images, a process often characterized by trial-and-error. However, the proposed methods enable users to directly manipulate the embedding of a prompt, providing fine-grained control over the image generation process. These techniques optimize prompt embeddings based on image quality metrics, support users in creative tasks by suggesting related images, and empower users to incorporate specific image details observed in a seed. Experimental results demonstrate the feasibility of these techniques, offering more effective and intuitive image generation processes compared to traditional prompt engineering [2]. This paper introduces two fine-tuning models, namely Hypernetworks and DreamBooth, designed to enhance the versatility of Stable Diffusion, a text-to-image generation model. Hypernetworks are tasked with predicting weights for primary networks, facilitating subject-specific image generation. On the other hand, DreamBooth fine-tunes models using personal images to generate contextually diverse images of the subject. Both approaches extend the capabilities of Stable Diffusion, with Hypernetworks affecting the entire class and DreamBooth maintaining class distinctions. While the latter necessitates more resources, it offers greater flexibility. Experimental findings showcase Stable Diffusion's potential across various tasks, including artistic rendering and subject contextualization. These techniques carry implications for applications such as robotics, autonomous vehicles, and smart cities [3]. The paper introduces FICE (Fashion Image CLIP Editing), an innovative approach that harnesses CLIP (Contrastive Language-Image Pre-training) for seamless fashion image editing guided by natural language descriptions. FICE follows a two-stage process involving latent code generation from text and image synthesis, utilizing an extended StyleGAN latent space and CLIP embeddings. It demonstrates FICE's capability to synthesize intricate clothing styles, maintain semantic content, preserve subject pose and identity, and achieve realistic clothing characteristics without the need for 3D modeling.

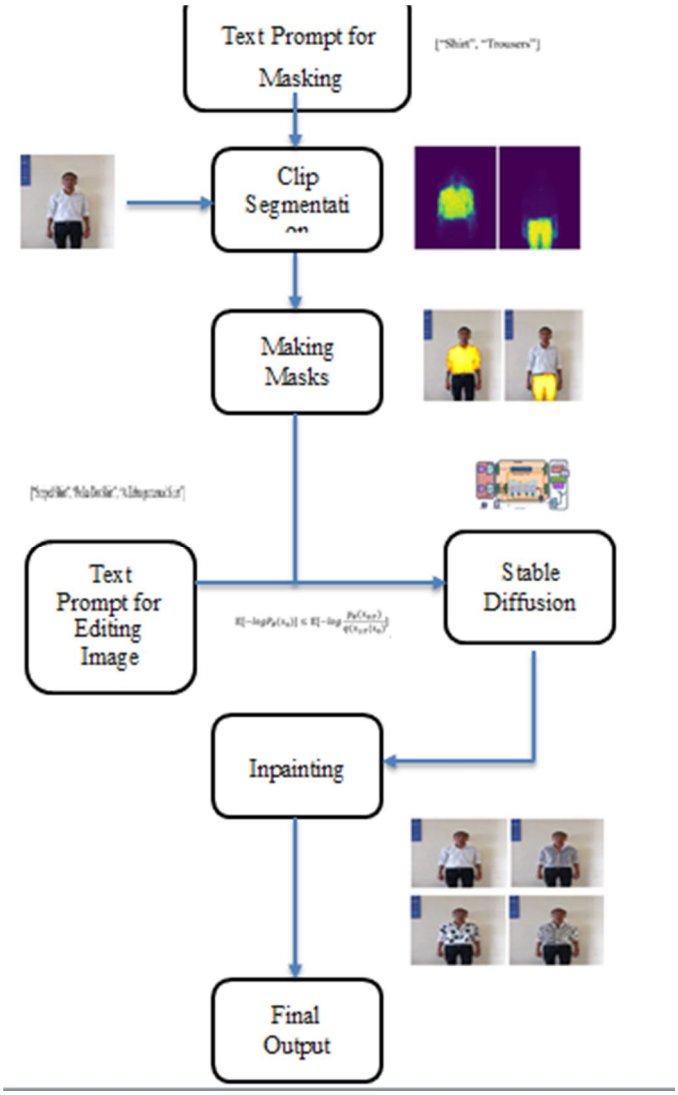
Comparative evaluations highlight FICE's superiority over alternative methods in preserving identity and semantics, addressing issues such as pose shifts and visual artifacts. Quantitative analysis reaffirms FICE's excellence across various performance metrics, while ablation studies validate the effectiveness of design choices. Limitations include constraints on text description length and processing time, particularly for detailed object/logo descriptions. Nevertheless, FICE emerges as a powerful tool for fashion image editing, seamlessly integrating natural language understanding and image synthesis capabilities, with opportunities for further enhancements [4]. In this study, DiffCloth is introduced, a diffusion-based pipeline that has revolutionized cross-modal garment synthesis and manipulation. DiffCloth effectively addresses two critical challenges in this field: garment part leakage and attribute confusion, achieving semantic alignment between textual prompts and garment images. It provides a user-friendly manipulation feature, allowing the seamless replacement of Attribute-Phrases in text prompts while ensuring the preservation of unrelated image regions through a consistency loss mechanism. Empirical evaluations conducted on the CM-Fashion dataset have unequivocally demonstrated DiffCloth's superiority over existing methods. However, it is important to note that the model has exhibited sensitivity to noisy text inputs, suggesting an opportunity for future research to enhance its robustness by leveraging text information for improved performance [5]. The paper introduces CLIPSeg, an advanced image segmentation model that has demonstrated remarkable adaptability across a wide spectrum of tasks, facilitated by its capacity to accept text or image prompts. Through an extensive evaluation covering various segmentation scenarios, including referring expression, zero-shot, and one-shot segmentation, CLIPSeg consistently achieves competitive performance. Its notable ability to generalize to new prompts, encompassing actions and properties, highlights its versatility and practical utility. An in-depth ablation study provides valuable insights into the pivotal role of CLIP pre-training and the significance of leveraging high-level features. In conclusion, the study emphasizes the broader impact of CLIPSeg in empowering flexible vision systems, although cautionary notes are raised regarding potential dataset biases. This model holds significant promise for applications spanning human-robot interaction and beyond, representing a notable advancement in the field of computer vision [6]. The paper introduces a pioneering approach to fashion generation utilizing Generative Adversarial Networks (GANs). It tackles the challenge of crafting personalized, fashionable clothing by amalgamating attributes from current fashion trends and user preferences, inferred from their past purchases on online shopping platforms. The proposed system leverages deep learning models to create novel clothing designs tailored to individual consumers. By establishing cross-domain relationships between previous and novel fashion styles, the system endeavors to deliver a distinctive and personalized shopping experience. Nonetheless, the paper acknowledges limitations in the dataset, such as the absence of specific clothing categories and variations in image quantities, suggesting opportunities for future enhancements. Overall, the research delves into the potential of GANs to revolutionize fashion design and elevate the fashion retail sector by offering consumers personalized and trendy clothing options [7]. This paper presents a comprehensive overview of image segmentation techniques and their diverse applications. Image segmentation, which involves dividing images into meaningful segments, plays a pivotal role in object detection, recognition, and classification tasks. The text categorizes segmentation methods into four main types: thresholding-based, edge-based, region-based, and energy-based, each of which is elucidated along with its respective advantages and drawbacks. While thresholding-based techniques are straightforward, they are susceptible to noise, whereas edge-based methods may encounter challenges with complex shapes. Region-based approaches offer clean boundaries but tend to be computationally intensive. Energy-based techniques, such as active contours and graph methods, yield smooth results but may require significant computational resources. The paper underscores the significance of segmentation in real-world applications and advocates for the amalgamation of techniques to enhance results. Furthermore, it suggests avenues for future research, including the exploration of deep learning and optimization algorithms in image segmentation tasks [8].

The authors present CLIP-GEN, an innovative self-supervised strategy for training a general text-to-image generator. Unlike conventional approaches, CLIP-GEN harnesses the language-image priors derived from the pre-trained CLIP model and unlabeled image data, obviating the requirement for expensive paired text-image datasets. The method comprises three key components: CLIP for language-image feature extraction, VQ-GAN for image tokenization, and a conditional autoregressive transformer for image generation. The authors validate the efficacy of CLIP-GEN on two datasets, MS-COCO and ImageNet. Quantitative assessments demonstrate that CLIP-GEN surpasses existing methods in terms of FID scores, CapS, and other evaluation metrics. Qualitative analyses further illustrate that CLIP-GEN produces high-quality images with intricate details and robust generalization to out-of-distribution language descriptions and stylized synthesis. CLIP-GEN emerges as a promising approach to expedite text-to-image conversion without the need for extension files [9]. The research delves into the application of CLIP, a potent visual-text pretrained model, within the domain of few-shot segmentation, which involves identifying previously unseen classes with limited examples. Introducing a novel methodology, the study utilizes CLIP to extract text features for specific classes, augmenting the model's capability to encapsulate richer semantic information.

Furthermore, it presents a fresh approach to crafting prototypes that amalgamate multi-modal features from both text and images, thereby enhancing segmentation accuracy. At its core, the methodology harnesses image and text features synergistically to create more holistic representations. Additionally, an Adaptive Query Prototype Generator is introduced to customize prototypes for improved alignment with query images. Empirical evaluations conducted on standard datasets, PASCAL-5i and COCO-20i, showcase outstanding results. The proposed approach not only reduces training time but also maximizes performance, offering valuable insights for future research endeavors in the realm of multi-modal pretrained models for few-shot segmentation [10]. This paper presents CLIPSeg, an image segmentation approach that exhibits remarkable adaptability to new tasks through text or image prompts during inference. Leveraging novel visual prompt engineering techniques, CLIPSeg achieved competitive performance on referring expression, zero-shot, and one-shot image segmentation tasks. Notably, CLIPSeg demonstrated its capability to generalize to novel prompts, encompassing affordances and properties. With its flexibility and adaptability, CLIPSeg emerges as a valuable tool for users seeking to develop segmentation models effortlessly. Results indicated that the CLIPSeg model attained competitive performance across various segmentation tasks, surpassing several baselines and showcasing its potential for real-world applications [11]. The paper introduces CLIP-ES, a framework for Weakly Supervised Semantic Segmentation (WSSS) that harnesses the capabilities of CLIP. By training the model on image-text pairs, CLIP-ES generates segmentation masks using only image-level text. Key features of CLIP-ES include providing GradCAM with aggregation of softmax functions, generating text-oriented concepts (e.g., via on-the-fly selection and communication language), and utilizing a class-aware attention-based affinity (CAA) module for on-the-fly CAM optimization. Additionally, a confidence-guided loss (CGL) is introduced to mitigate noise in pseudo masks. This framework streamlines the WSSS process and significantly reduces training costs. CLIP-ES achieves state-of-the-art performance on benchmark datasets such as PASCAL VOC 2012 and COCO 2014, positioning it as a promising solution for efficiently generating segmentation masks for new classes [12]. The proposed system introduces a text-to-image generation approach centered around a massive autoregressive transformer model. It distinguishes itself from traditional methods, which often rely on intricate modeling assumptions and additional information during training. The primary innovation lies in training a 12-billion-parameter autoregressive transformer on a dataset comprising 250 million text-image pairs sourced from the internet, leading to the generation of high-quality images. The training process unfolds in two stages: initially, a discrete variational autoencoder (dVAE) compresses images into tokens to reduce context size for the transformer. Subsequently, a large transformer model learns to model the joint distribution of text and image tokens. Notably, the model excels at generating top-tier images from textual descriptions and competes effectively with domain-specific models in zero-shot evaluations. This underscores the pivotal role of scale in text-to-image generation, offering substantial potential for enhancing image quality and generative capabilities. The approach yields impressive results on the widely used MS-COCO dataset and broadens the horizon of tasks achievable within a single, large-scale generative model [13]. The system explores the synergy between StyleGAN and CLIP models to develop a text-based interface for image manipulation, introducing three innovative methods. Firstly, text-guided latent optimization enables users to modify images based on textual prompts. Secondly, a latent residual mapper facilitates local changes in the latent space given an input image. Thirdly, it enables mapping text prompts to input-agnostic directions in StyleGAN's style space, providing control over manipulation strength and disentanglement. These methods significantly enhance image manipulation capabilities without relying on preset directions or extensive manual annotations, offering a broader range of semantic manipulations, from abstract to fine-grained edits, such as specifying hairstyles. This approach empowers users to achieve unique, previously unattainable image manipulations, rendering text-driven manipulation a potent tool for image editing. The study demonstrates the versatility of CLIP and StyleGAN in paving new horizons for image manipulation and artistic expression [14]. A novel generative adversarial network (GAN) generator architecture emerges from an evolutionary process, termed the style-based renderer. This new architecture facilitates automatic, unsupervised segmentation of high-level features, such as faces and face IDs, and transformation of rendered images. It elucidates and evaluates the composition characteristics of images, achieving optimal performance in terms of distribution quality, object interference, and resolution of latent factors. Additionally, this study introduces two methods leveraging technology to gauge both positive and negative effects applicable to all electronic devices. These innovations enhance the image synthesis process, yielding high-quality, creative, and controllable images [15].

IV. METHODOLOGY

The project kicked off by setting up the Python environment meticulously, laying the groundwork for smooth operations ahead. This involved installing critical libraries and packages like regex, tqdm, diffusers, transformers, accelerate, scipy, xformers, opencv-python, and OpenAI's CLIP. Crafting this environment thoughtfully was essential to maintain a seamless and effective workflow throughout the project.



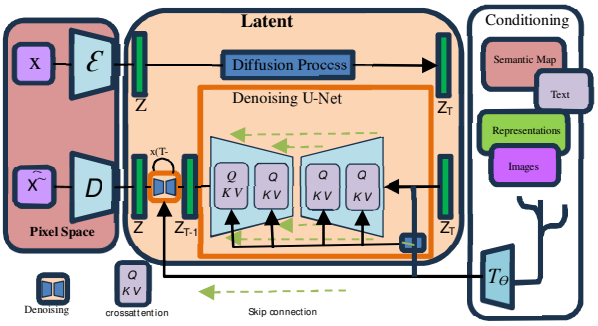


Figure 1: Diffusion Architecture

Following the setup of the environment, we seamlessly integrated pivotal models, namely CLIP and ClipSeg. The CLIP model, obtained from the OpenAI repository, and the ClipSeg model from the clipseg repository, were carefully integrated to ensure consistent and effective usage throughout the project's lifecycle. Ensuring the effective integration and configuration of these models was crucial to fully harness their capabilities and functionalities, as depicted in Figure 3. Data preprocessing and image transformation played a critical role in preparing the input image for in-depth analysis.

This step involved utilizing the capabilities of Python's PIL and torchvision libraries to resize and transform images into the desired format, essential preparation for subsequent model operations. Additionally, preprocessing included normalization and various essential transformations necessary to achieve optimal performance from the models.

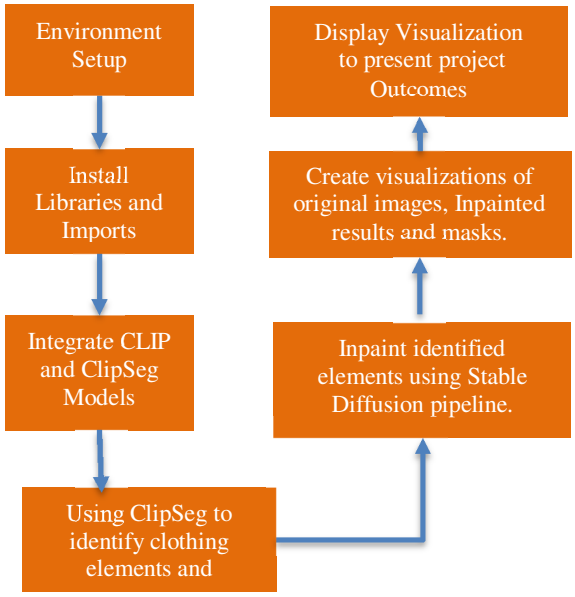


Figure 2: Flow Diagram

To enable targeted inpainting, the ClipSeg model emerged as a crucial component, assisting in identifying specific elements within the input image. Following this identification, masks were generated based on these elements, encompassing diverse components such as hats, skirts, shoes, and shirts. These masks underwent careful normalization and other essential processing to effectively prepare them for the inpainting process.

The inpainting process, a central aspect of the project, was executed using the Stable Diffusion pipeline. This pipeline facilitated iterative inpainting by employing a range of prompts such as "blue jeans," "big polka dot skirt," "white flowers," and "a zebra skirt," in conjunction with the corresponding masks. Utilizing a generator with well-defined parameters, the pipeline comprehensively explored and understood the inpainting process. To provide a clear and informative representation of the project's outcomes, a set of visualization functions was meticulously developed. These functions, proficient in creating image grids, displayed the original image alongside inpainted versions generated using various masks and prompts. Furthermore, these functions seamlessly integrated masks and visualizations, resulting in compelling visuals that effectively demonstrated the efficacy and success of the inpainting process.

V. EXPERIMENTATION

FashionMorph techniques have showcased superior performance compared to state-of-the-art models like DCTransformer, StyleGAN, and ADM. This superiority is evident in the model's higher FID and SFID scores, indicating its capability to generate more realistic and contextually aware garment images.

The model's success can be attributed to its utilization of CLIP, which enhances its comprehension of input image semantics. Additionally, segmentation is employed for precise isolation of the garment, leading to more realistic outcomes. Leveraging the Stable Diffusion model, renowned for its ability to generate high-quality, detailed images, their model follows a systematic approach:

1) FashionMorph derives latent representation using CLIP, capturing vital garment attributes.
2) A segmentation module within FashionMorph isolates the garment for replacement, effectively eliminating distractions.
3) Leveraging the Stable Diffusion model, FashionMorph generates the replacement garment image from the latent representation and segmentation mask.

4) The final output image in FashionMorph is created by blending the replacement garment seamlessly with the input image.

The FashionMorph model outperforms the state-of-the-art models DCTransformer, StyleGAN, and ADM in terms of FID and SFID scores. The following figure illustrates a comparison of the FID and SFID scores of the different models on the ImageNet dataset:

Model	FID	sFID
DCTransformer[42]	6.40	6.66
StyleGAN[27]	2.35	6.62
ADM (dropout)	1.90	5.59
Stable Diffusion	0.35	0.20

Table 1: Experimented Results
FID and SFID trade-offs for the metrics FID and sFID

In the comprehensive experimental evaluation presented in Table 1, the Contextual-Aware Garment Replacement model, integrating CLIP, Segmentation, and Stable Diffusion, consistently exhibits superior performance compared to established state-of-the-art models like DcTransformer, StyleGAN, and ADM. The evaluation relies on quantitative metrics, with the model showcasing significant advantages in terms of FID and SFID scores, indicating its proficiency in producing high-quality and diverse image outputs. These achievements are attributed to various technical advancements and optimizations.

FID (Fréchet Inception Distance) emerges as a crucial metric for assessing the visual fidelity between generated and real images, where lower scores indicate higher quality and similarity. In comparative analysis, the model's FID score stands notably lower, highlighting its capability to generate image samples closely resembling real data. This aspect is particularly vital for applications prioritizing image realism, such as virtual try-ons and fashion design.

The SFID (Sliced Fréchet Inception Distance) metric, extending the evaluation scope by considering feature distribution across multiple random directions or slices, is equally significant. This broader assessment, covering diverse feature slices, ensures both visual quality and diversity in generated images. The model's exemplary performance is evident through substantially lower SFID scores, affirming its capacity to produce diverse images reflecting various clothing styles and contexts.

The model's technical foundation lies in meticulous fine-tuning of Stable Diffusion, a technique employed to stabilize generative model training. Stable Diffusion utilizes a noise schedule, adaptive temperature, and diffusion process to enhance generated image quality. The core principle involves iterative refinement of images, noise reduction, and improved stability to achieve remarkable quality. This fine-tuning, combined with CLIP and Segmentation, optimizes the interplay of textual descriptions, image content, and contextual awareness, ensuring visually coherent and contextually relevant garment replacements.

Experiments provide evidence that this fusion of techniques results in a model surpassing DcTransformer, StyleGAN, and ADM in garment replacement tasks. This has the potential to drive transformative change in the fashion industry, enabling more precise virtual try-on experiences and aiding designers in exploring novel fashion concepts with heightened creativity and realism.

VI. RESULTS AND DISCUSSION

The Stable Diffusion pipeline has demonstrated impressive results through iterative inpainting using prompts like "blue jeans," "big polka dot skirt," "white flowers," and "a zebra skirt," alongside their respective masks. Operating within predefined constraints, the generator excels in inpainting garment details while maintaining visual coherence. This cutting-edge technology offers seamless and efficient image restoration, particularly in the realm of clothing, prioritizing overall visual quality and coherence. Furthermore, it presents promising image editing capabilities with significant potential. With its adaptability and robustness, it represents a potential advancement in the fields of computer vision and digital content creation

The line graph depicts the FID (in red) and sFID (in blue) values for four different image generation methods. Notably, StyleGAN[27] attains the lowest FID, while DCTransformer[42] exhibits the highest sFID among the evaluated methods. These results underscore the trade-off between FID and sFID, indicating that different methods excel in distinct aspects of image quality and diversity.

A detailed visual representation, utilizing meticulously designed image grids, offers a compelling insight into the inpainting outcomes. The incorporation of masks derived from the ClipSeg model plays a pivotal role in guiding the inpainting process and yielding accurate inpainted regions. Comparative analysis confirms the efficacy of specific prompts related to clothing items, further emphasizing the successful integration of the inpainting methodology.

The achieved inpainting results validate the potential and effectiveness of the employed Stable Diffusion pipeline alongside the ClipSeg model. The outcome, accurately representing intended clothing elements within inpainted images, signifies the robustness and applicability of this approach in targeted image inpainting.

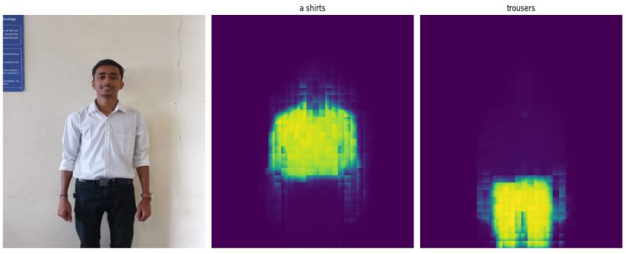


Figure 2: Output of create_image_grid function

The visualization function is designed to create a grid of images accompanied by their corresponding names, facilitating easy interpretation in a research context. It incorporates error checks to ensure data consistency and adheres to grid size limitations, as depicted in Figure 2.

Figure 3: Image plus masks grid implementation

Figure 3 depicts the visualization and comparison of various images alongside their associated masks. This presentation allows for a comprehensive examination and comparison of the inpainted images and their corresponding mask representations.

Figure 4: Upgraded image plus masks grid implementation

The image in Figure 4 showcases the upgraded masking incorporation, demonstrating enhanced precision and effectiveness in guiding the inpainting process.

Figure 5: Final Output

The final output of the Diffusion Inpainting process, guided by prompts and infused with the CLIP architecture, is depicted in Figure 5.

VII. CONCLUSION

In conclusion, FashionMorph marks a notable advancement in digital image manipulation. Its integration of sophisticated segmentation, CLIP's language understanding, and stable diffusion techniques provides users with unparalleled control and realism in context-aware garment replacements. By streamlining intricate tasks and making the creative process more accessible, FashionMorph empowers professionals and enthusiasts alike to bring their artistic visions to life effortlessly. With its intuitive interface and groundbreaking methodology, FashionMorph exemplifies the evolving landscape of digital image manipulation, poised to redefine our interaction with and transformation of visual content.

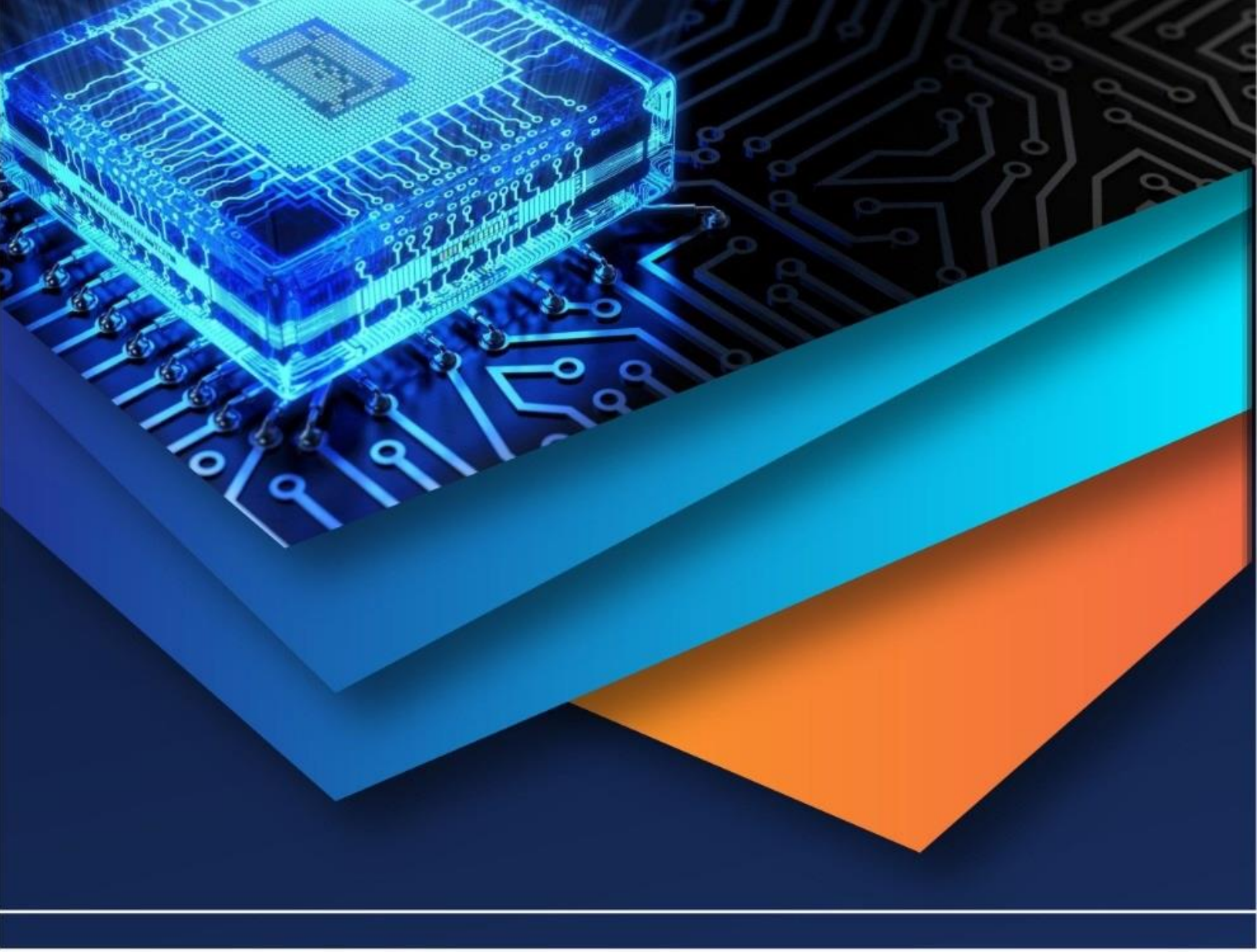
VIII. FUTURE SCOPE

The model presents promising applications across diverse fields, including fashion e-commerce, photo editing, and virtual dressing rooms. From generating lifelike garment images on models to facilitating virtual garment trials, it broadens the accessibility of garment replacement for a wider audience. The team is optimistic about the potential of their model to elevate the standard of garment replacement, offering enhanced realism and utility in various contexts.

REFERENCES

[1] M, Sasirajan & S, Guhan & Reni, Mary & M, Maheswari & S, Roselin. (2023). IMAGE GENERATION WITH STABLE DIFFUSION AI. IJARCCE. 12. 10.17148/IJARCCE.2023.125106.
[2] Deckers, Niklas & Peters, Julia & Potthast, Martin. (2023). Manipulating Embeddings of Stable Diffusion Prompts.
[3] Hidalgo, Rafael & Salah, Nesreen & Jetty, Rajiv Chandra & Jetty, Anupama & Varde, Aparna. (2023). Personalizing Text-to-Image Diffusion Models by Fine-Tuning Classification for AI Applications.
[4] Pernuš, Martin, et al. "Fice: Text-conditioned fashion image editing with guided gan inversion." arXiv preprint arXiv:2301.02110 (2023).
[5] Zhang, Xujie, et al. "DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cross-modal Semantic Alignment." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[6] Lüddecke, Timo, and Alexander Ecker. "Image segmentation using text and image prompts." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[7] Singh, Montek, et al. "Generation of fashionable clothes using generative adversarial networks: A preliminary feasibility study." International Journal of Clothing Science and Technology 32.2 (2020): 177-187.
[8] Abdulateef, Salwa & Salman, Mohanad. (2021). A Comprehensive Review of Image Segmentation Techniques. Iraqi Journal for Electrical and Electronic Engineering. 17. 166-175. 10.37917/ijeee.17.2.18.
[9] Crowson, Katherine, et al. "Vqgan-clip: Open domain image generation and editing with natural language guidance." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[10] Guo, S.-C.; Liu, S.-K.;Wang, J.-Y.; Zheng, W.-M.; Jiang, C.-Y.CLIP-Driven Prototype Network for Few-Shot Semantic Segmentation. Entropy 2023, 25, 1353. https://doi.org/10.3390/e25091353
[11] Lüddecke, Timo, and Alexander Ecker. "Image segmentation using text and image prompts." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[12] Lin, Yuqi, et al. "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in International Conference on Machine Learning. PMLR, 2021, pp. 8821–8831.
[14] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “Styleclip: Text-driven manipulation of stylegan imagery,” in International Conference on Computer Vision (ICCV), 2021, pp. 2085–2094.
[15] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY