Semantic Style Transfer for Paintings Using Convolutional Neural Networks (CNNs)

Hafiz Muhammad Jamsheed Nazir, Zheng Jiangbin, Omar Alsaleh
Department of Software, Northwestern Polytechnical University, Xi'an, China[1, 2]
College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia[3]

***Abstract*—In recent years, the importance of photographic portrait styles has garnered significant attention, prompting numerous researchers to explore innovative methods for modifying and enhancing these styles. Neural style transfer has advanced rapidly for photographic portraits, yet transferring painterly styles to human facial and body images remains difficult because global stylization frequently distorts facial geometry and erases identity. This study presents a semantic, region-wise painting style transfer framework based on Convolutional Neural Networks (CNNs) that preserves facial identity and semantic structure during stylization. The method parses both the source photograph and an example painting into corresponding semantic regions, comprising ten facial components together with hair, chest, arms, legs, and background, and transfers the style region by region so that each part is stylized from its semantic counterpart. A feature reconstruction stage based on Gram matrix style representations minimizes content and style loss within each region, while a part-based generation and fusion stage augmented with Laplacian pyramid decomposition improves local to global consistency and identity preservation. We evaluate the approach with perceptual and identity metrics, reporting Frechet Inception Distance (FID), Structural Similarity (SSIM), and identity cosine similarity (CSIM), and additionally report a downstream classification check as an auxiliary indicator of content preservation. The full model attains an FID of 14.72, an SSIM of 0.82, and a CSIM of 0.86, outperforming GAN and part generation network baselines in identity preservation and realism. We discuss the strengths, limitations, and practical implications of the framework and outline directions toward full-body, high-resolution, and video stylization.**

Keywords*—Painting style transfer; semantic style transfer; identity preservation; facial region parsing; Convolutional Neural Networks (CNNs)

I. Introduction

Semantic paintings are a significant type of painting. Previously, the commonly used portrait transmission system was used [1, 2]. Human cognitive behavior always sees the individual things first, and then expands from the individual things to the general type, so the paradigm basis of the type forms the framework of a certain concept. These categories are the logic at the bottom of abstract thought [3]. They are the pure synthetic concepts of the source, for which understanding contains a priori in itself, for which it is pure understanding. It is only through these concepts that understanding something in the multiplicity of intuition that it thinks of the object of intuition. Therefore, the concept and the image play a role through the meaning when we use the word semantic painting in different circumstances and contexts, then what is the object to which understanding points, and the concept describing the state of reality is bound in both connotation and denotation [4]. Connotation means that something is a property of itself rather than something else, and that this property is the essential basis of its existence. Under the corresponding law of nominative relations in formal logic, things are classified, and explicit attribute relations are abstracted. The genus, which is reduced to essentials, is the decisive factor that distinguishes objects from others, and some objects with such attributes are extensions of concepts [5, 6]. This kind of thinking movement from individual to general is also an induction and summary of the common attributes of many individual things. The semantic content of a photo is a dissimilar style that achieves a difficult photo processing goal. Possibly a main preventive issue for previous methodologies is that the photo cannot represent or perform clearly the semantic information, and therefore, by using it. We maintain the countenance of the single photo content in style.

The rebuilding portion resolves the optimization problem of content loss and style loss in characteristic places, mainly by rebuilding characteristics. This significantly decreased the loss during spreading through the entire network. The interpreter transmits the rebuilding characteristics into a stylized image. At the same time, from the perspective of cognitive psychology, the concept is represented by a prototype, which is the best instance, and people's understanding and acceptance of the concept not only includes its prototype, but also includes dimensions that are the degree of representativeness of category members [7]. Semantic is a kind of name describing the probability similarity of the phase of the referred object, among which the commonness and special boundary of Semantic, porcelain, and semantics still have a definite meaning at present. For example, products with non-compact bodies made of clay and other natural raw materials produced through the process of forming and firing are called pottery. The use of paint ore after high temperature firing results in dense carcasses [2]. The synthetic word semantic which is composed of the two words semantic, and porcelain, means that the material is made of minerals containing silicate through systematic and standardized mixing, molding, sintering, and other technical processes. The essence and related properties of semantics are determined from this, and the conceptual prototype of Semantics, namely, the best example, is only the object with semantics. At the same time, all kinds of derivative concepts of semantic objects are also used in this semantic category. In short, the original meaning of semantics in semantic painting is a structural concept

*Corresponding author.

formed by the name of description after the synthetic silicate in the real world is abstracted by the subject [8]. The basic connotation of the concept of painting refers to the visual recognition of a subject on a carrier with the characteristics of a two-dimensional physical space. The basic category of any concept is the easiest to form a visual image hierarchy in perception, and then form a single psychological image reflecting the whole set of categories [9, 10]. Therefore, for thousands of years, when people talk about semantic painting, they quickly think of colored pottery, ancient color, and so on, but do not immediately think of pottery carving or pottery art [11]. Within the concept dimension, art forms can also better represent the concept of semantic painting. An extension example of this property is that the work has the visual characteristics of natural simplicity, coarse clay sexy grain, or smooth, fine, hard, and retains the rounded glass texture, or glaze, borne au, crystal, and other rich texture senses. These are the basic characteristics of semantic paintings that differ from other types of painting [10]. Style transfer aims to re-render the content of one image in the visual appearance of another, and it has become a central problem in computational art and image synthesis. Since the introduction of neural style transfer by Gatys et al., convolutional neural networks (CNNs) have made it possible to separate the content of a photograph from the style of a reference image and to recombine them. While portrait photography and photographic filters have received substantial attention, the transfer of genuine painting styles to images of people, where brushwork, color palette, and texture must be reproduced faithfully, has been studied far less and remains challenging [12].

The core difficulty is that human observers are highly sensitive to faces. When a painterly style is applied globally to a portrait, the optimization tends to deform the eyes, mouth, and skin texture, producing results that look unnatural and that no longer preserve the identity of the subject [13]. Most existing methods treat a face as a single generic object and therefore cannot guarantee that semantically meaningful regions, such as the eyes or hair, are stylized in a way that respects their structure. As a result, identity is often lost, and visible artifacts appear in detailed regions such as the ears and hair.

Research gap: Although recent work has improved global stylization quality and efficiency, comparatively few methods explicitly preserve facial identity and semantic structure when transferring a painting style, and fewer still extend stylization beyond the head to the whole body [14]. Methods that do incorporate semantic segmentation usually require separately trained segmentation and stylization models together with large annotated datasets, which limits their practical use.

Proposed approach and contributions: This study addresses that gap with a semantic, region-wise painting style transfer framework for human facial and body images. The contributions are threefold. First, we formulate painting style transfer as a region-to-region problem in which the source photograph and an example painting are both parsed into corresponding semantic parts, including ten facial components together with hair, chest, arms, legs, and background, and style is transferred between matching parts rather than globally [15].

Second, we combine Gram matrix feature reconstruction with a part-based generation and fusion stage and a Laplacian pyramid decomposition that improves local to global consistency and preserves identity during stylization. Third, we evaluate the framework with perceptual and identity metrics (FID, SSIM, and CSIM) and provide an auxiliary downstream classification check of content preservation, together with a critical discussion of strengths, limitations, and practical implications [16].

II. Literature Review

Semantic painting style transfer integrates a semantic understanding of visual content with artistic rendering, enabling more controllable and meaningful stylization. Unlike traditional neural style transfer, which applies artistic texture globally, semantic methods are region-aware and object-level, so that important structures are preserved during transformation. This paradigm draws on CNNs, generative adversarial networks (GANs), and, more recently, transformer-based and diffusion-based architectures to produce coherent and semantically consistent artistic images [17].

The foundation of the field is the neural style transfer (NST) method of Gatys et al., which used CNN features to decouple and recombine content and style high-level features, preserving content structure, while correlations between feature maps, expressed as Gram matrices, capture artistic texture. Although NST produces high-quality results for arbitrary content style pairs, it is computationally expensive, slow to converge, and offers no fine-grained control over specific regions. To improve efficiency, Johnson et al. proposed a feed-forward network trained with a perceptual loss that enables real-time stylization, but it requires a separately trained model for each style [18].

To add spatial and semantic control, later work integrated semantic segmentation into the pipeline, assigning a class label to each pixel so that different styles or strengths can be applied to distinct regions such as sky, buildings, and faces [19]. Accurate pixel-level architectures such as DeepLabV3+ improved content preservation, but they typically require independently trained segmentation and stylization models, increasing cost. Spatially adaptive normalization, introduced by the SPADE framework of Park et al., modulates feature activations from a semantic layout and improves boundary and structural consistency, at the price of large annotated datasets and heavy training. In the unpaired setting, CycleGAN enabled image-to-image translation without paired data through cycle consistency, but it can be unstable and semantically inconsistent on complex, human-centric scenes.

While the methods summarized in Table I laid the foundational framework for semantic painting style transfer, recent advances from 2020 to 2025 have significantly expanded the methodological landscape. Modern approaches increasingly emphasize global context modeling, multimodal guidance, and improved semantic fidelity, particularly for complex scenes and human-centric images. These developments motivate a revised comparative analysis reflecting state-of-the-art techniques [20].

TABLE I. COMPARISON OF CLASSICAL APPROACHES IN SEMANTIC PAINTING STYLE TRANSFER

Approach	Advantages	Disadvantages
Real Time Style Transfer (Johnson et al., 2016)	Enables real-time image and video stylization; suitable for interactive and video applications	Limited to predefined styles; retraining required for new styles; reduced flexibility
Semantic Segmentation + Style Transfer	Provides region-level control by assigning semantic labels to pixels; improves content preservation	Requires accurate semantic segmentation; increased pipeline complexity
DeepLabV3+ with Style Transfer	Accurate pixel-level semantic control; better preservation of semantic content and boundaries	Segmentation and style networks are trained separately; high computational cost
SPADE (Park et al., 2019)	Fine-grained, semantic-aware stylization via adaptive normalization; strong structural consistency	Complex training process; requires large-scale annotated datasets
Cycle GAN (Zhu et al., 2017)	Supports unpaired image-to-image translation; flexible across domains	Training instability; possible semantic inconsistencies and artifacts
Art GAN (2019)	Improved semantic consistency and texture preservation	High computational demand; limited generalization to highly diverse styles

TABLE II. COMPARISON OF ADVANCED SEMANTIC PAINTING STYLE TRANSFER APPROACHES (2020–2025)

Approach	Advantages	Disadvantages	
Transformer-Based Style Transfer (ViT, CNN–Transformer hybrids)	Utilizes self-attention to capture global dependencies and semantic relationships	Strong global coherence; reduced local artifacts; better handling of complex scenes	High computational complexity; large memory requirements
Diffusion-Based Semantic Style Transfer	Progressive denoising guided by semantic constraints and latent representations	State-of-the-art visual quality; reduced texture leakage; high semantic fidelity	Slow inference; expensive training; deployment complexity
Text Guided & Multimodal Style Transfer	Combines images, text prompts, and semantic maps for controllable stylization	Highly flexible artistic control; supports user-defined semantics	Prompt sensitivity; difficulty in precise regional control
Fine-Grained Face Parsing + Style Transfer	Part-based semantic labeling of facial components (eyes, lips, skin, hair)	Improved identity preservation; reduced facial distortion	Requires high-quality labeled facial datasets; complex pipelines
Self-Supervised Semantic Style Transfer	Reduces dependency on labeled data using pseudo labels or contrastive learning	Improved generalization; lower annotation cost	Lower segmentation accuracy compared to fully supervised methods
Region-Wise (Part-Based) Semantic Style Transfer	Applies style independently to segmented face and body regions	Enhanced semantic consistency; flexible region control	Increased preprocessing complexity; sensitivity to segmentation errors

As shown in Table II, recent semantic painting style transfer methods emphasize semantic fidelity, global coherence, and data efficiency. Transformer-based and diffusion-based models significantly outperform earlier CNN and GAN-based approaches in preserving structure and reducing artifacts, particularly in facial and human-centric images. However, these improvements come at the cost of increased computational complexity and data requirements. Part-based semantic labeling and weakly supervised learning have emerged as promising strategies to balance realism, flexibility, and scalability [21]. These trends directly motivate the proposed method in this study, which focuses on region-wise semantic labeling and targeted style transfer to improve identity preservation and visual coherence while minimizing annotation dependency.

Subsequent models, such as Art GAN and Art Real GAN, improved semantic consistency and texture fidelity, but they require significant computational power and struggle to generalize across highly diverse artistic styles (2020–2025). Semantic painting style transfer has increasingly shifted toward transformer-based and diffusion-based models. Vision Transformers and hybrid CNN transformer architectures provide stronger global context modeling and long-range dependency capture, which helps reduce local artifacts and improves coherence across semantic regions [22]. Diffusion-based models, particularly latent diffusion frameworks, have demonstrated state-of-the-art performance by enabling progressive and controlled image generation. These models support multimodal guidance, including text prompts and semantic constraints, allowing more expressive and flexible artistic control. Recent studies show that diffusion-based semantic style transfer significantly mitigates texture leakage and structural distortions, especially in facial regions, which have historically been difficult to stylize without compromising identity. Human facial and body images remain one of the most challenging application domains for semantic painting style transfer due to their sensitivity to geometric distortions and perceptual inconsistencies. Despite extensive progress, relatively few studies have explicitly focused on reducing distortions in facial components such as eyes, mouth, and skin texture. Most existing methods treat faces as generic objects, leading to visually unnatural results. Recent approaches have begun incorporating fine-grained face parsing, dividing faces into multiple semantic regions, and applying part-based style transfer [23]. This strategy improves identity preservation and visual realism but requires high-quality labeled datasets and complex processing pipelines. In cases where artistic styles exhibit weak texture patterns, color transfer techniques have been explored as complementary solutions; however, these methods are generally insufficient for capturing rich painterly textures. A persistent challenge across all semantic painting approaches is the dependency on large amounts of reliable labeled data [8]. High-quality semantic annotations are expensive and time-consuming to obtain, particularly for fine-grained facial and body segmentation. As a result, recent research has explored self-supervised, weakly supervised, and data augmentation strategies to improve generalization while reducing annotation costs. These directions suggest that optimizing data efficiency is critical for advancing semantic painting style transfer toward real-world applications [24].

Existing semantic painting style transfer techniques offer varying tradeoffs between artistic quality, semantic consistency, flexibility, and computational cost. While early CNN-based methods laid the foundation, modern approaches increasingly rely on semantic segmentation, adaptive normalization, GANs, transformers, and diffusion models to achieve finer control and improved realism. Nevertheless, challenges such as facial distortion, data dependency, and computational complexity remain open research problems. Motivated by these limitations, recent research, including the present work, focuses on part-based semantic labeling of faces and body regions, enabling region-wise style transfer that enhances semantic coherence, preserves identity, and reduces visual artifacts while maintaining artistic expressiveness.

III. Motivation

Painting style filters are now common in social media applications such as Facebook, WhatsApp, and WeChat, but their greatest strength, fast and lightweight global stylization, is also their main weakness for portraits applied to faces; these methods frequently distort facial information. In practice, they may preserve the background while degrading the foreground, or the reverse, and the resulting facial deformation undermines both realism and identity. This gap motivates a more accurate, automatic painting style system that preserves semantic information and avoids facial distortion. Because selfie and portrait images are extremely common, a method that yields faithful, identity-preserving results in portrait painting styles has clear practical value. Existing facial expression transfer results applied to 3D face models can appear unrealistic (Fig. 2), and the painterly texture they produce is often weak. The present study, therefore, proposes an automatic, example-based painting style framework using CNNs that preserves semantic structure, reduces facial distortion, and strengthens painterly texture, as illustrated in Fig. 1.

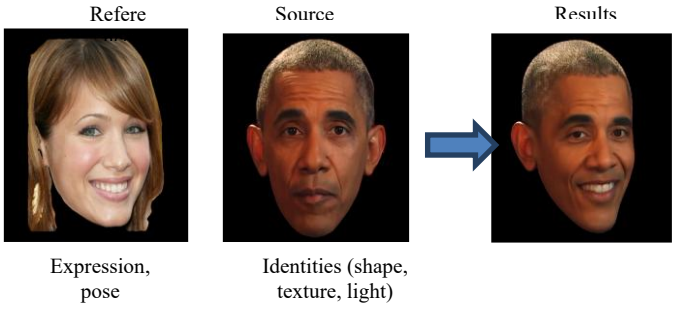


Fig. 1. Example-based painting style framework using CNNs.

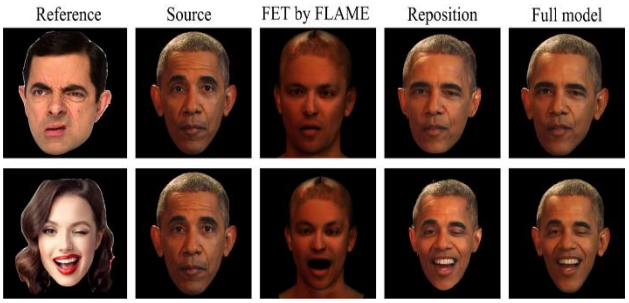


Fig. 2. FET result unrealistic by the 3D face model.

IV. Proposed Approach

The proposed framework performs automatic, example-based painting style transfer in two main stages. In the labeling stage, the full set of body parts of both the source photograph and the example painting is parsed into corresponding semantic regions. In the transfer stage, style is propagated from each region of the example to the matching region of the source, so that eyes are stylized from eyes, hair from hair, nose from nose, lips from lips, and background from background. Each image in the training set is resized to 224 x 224 pixels and passed through a trained transformation network to generate a stylized image for each style.

Network architecture: the framework is built on a VGG-based feature extractor that supplies content and style representations, a transformation network that produces stylized regions, and a part-based composition module that reassembles the regions into a coherent image (Fig. 3). The transformation network follows an encoder-decoder structure with residual blocks. The part generation networks (PGNs) use 2D convolutional layers with LeakyReLU activations and residual blocks, while the part fusion network (PFN) uses 2D convolutional layers with skip connections that preserve fine detail during fusion (Fig. 4). Both the original and stylized images are passed through the VGG network so that a downstream classification loss can be used as an auxiliary check of content preservation.

Dataset and implementation experiments use a personalized facial dataset of approximately 500 images, together with the Caltech dataset for the auxiliary downstream classification analysis. All inputs are resized to 224 x 224 to match the pretrained ImageNet feature extractor. Ten distinct painting styles are used to train the transformation network. Training minimizes a weighted combination of content loss, style loss, perceptual loss, cosine similarity loss, and an adversarial loss; the network is optimized with the Adam optimizer. The standard input resolution and the loss weighting follow common practice in feed-forward style transfer.

Fig. 3. Challenges to 3DFaceShop, 2023.

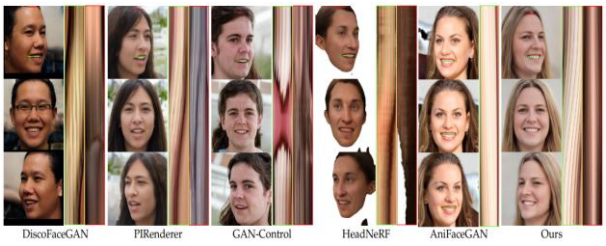


Fig. 4. 3DFaceShop, 2023.

A. System Implementation and Advanced Features

First of all, the methods for only segmenting samples of paintings cannot be reserved through a specific style. Methods that segment only the example painting cannot preserve a region-specific style for the source. To overcome this, the proposed system supervises semantic painting by constructing a labeling map for both images. The painting style is represented by the Gram matrices of CNN feature maps, and the style loss is formulated as the difference between corresponding Gram matrices in the style computation layers. The system uses a linear approximation of the Gram matrix that is perceptually comparable to repeated backward optimization but computationally cheaper. Starting from the source image, mixed feature representations built on CNN patch swaps are used to synthesize the stylized output, and the inverse network is tested on examples that combine the original and the painting.

B. Labeling

There were no significant difference, and similar consequences were observed, showing that the photos look similar to each other, so the content weights have a minimal impact in this study [25, 26]. However, in the future, the maximum content weight can be added, and the stylized photo changes accordingly. All existing methods focus on face labeling, and some have better consequences [27]. However, the goal of this study is to label the full body parts, as shown in Fig. 5, Semantic Painting Style Transfer Using CNNs. Input with Example Labeling for both the input photo and the painting. This study aims to label the face into ten parts and then divide the remaining body parts, such as the arms, legs, and chest. The full labeling parts of the structure, and then transferring the style parts part by part [28]. Initially, this study combined two different styles to verify if the best consequences were acquired [29]. In this research, the input photo, along with an example photo, is fed into two different transformation systems to create two stylized photos for each input photo, as shown in Fig. 6, High Fidelity and Identity Preserving [30]. This study then merged the stylized photo and original photo to create the final training dataset. Compared with input, the combined style has a considerably better result, but has a similar result to the single style technique [31].

Fig. 5. Semantic painting style transfer using CNNs, input with an example labeling.

When the content weights are varied within a moderate range, no significant difference is observed, and the stylized outputs remain visually similar, indicating that content weighting has a limited effect in this setting; larger content weights can be explored in future work. Whereas most existing methods label only the face, the goal here is to label the full set of body parts. The face is divided into ten parts, and the remaining body regions, such as arms, legs, and chest, are labeled separately, after which the style is transferred part by part [32]. Combining two different styles was also tested to verify whether better results could be obtained; the combined style produced results comparable to the single style technique. The stylized and original images are then merged to form the final training dataset [33].

Fig. 6. High fidelity and identity preserving.

V. Method

To strengthen identity preservation during stylization, the framework incorporates two complementary refinement methods that operate on the semantically labeled regions: a part-based method and a Laplacian pyramid decomposition (LPD) method [34]. These components refine the region-wise stylization produced in Section IV; they are not a separate task but a mechanism for improving local-to-global consistency and preserving identity within the painting style transfer pipeline. We emphasize that the system is a painting style transfer framework, and the part-based refinement is the means by which facial detail and identity are protected during that transfer.

A. Part-Based Method

The part-based method improves stylization quality both locally and globally by combining semantic label maps with feature point maps. It proceeds in four steps. First, parts separation and tight cropping isolate individual facial components, such as the eyes, nose, and mouth, by cropping tightly around each region. Second, part generation networks (PGNs) synthesize stylized versions of each isolated part [35]. Third, parts repositioning and composition place the generated parts back into a coherent layout. Fourth, a part fusion network (PFN) merges the parts into a single unified image while preserving the original identity.

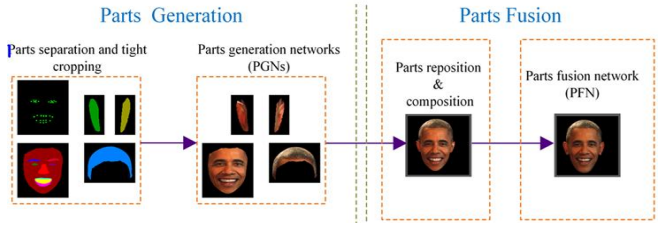


Fig. 7. Details of the part-based FET.

This figure illustrates a process that involves separating, generating, repositioning, and fusing parts of an image, likely applied to face manipulation tasks. The first step, Parts

Separation and Tight Cropping, isolates different facial features, such as the eyes, nose, and mouth, by cropping the image tightly around these regions. In the next step, Parts Generation Networks (PGNs) are used to generate new or stylized versions of these isolated parts. Afterward, in the Parts Repositioning and Composition stage, the generated parts are carefully repositioned and combined to form a coherent structure, ensuring the pieces fit together seamlessly, as shown in Fig. 7, Details of the Part-based FET [36]. Finally, the Parts Fusion Network (PFN) is employed to fuse all the generated parts back into a final, unified image, preserving the original identity while applying the desired transformation. This approach is commonly used in tasks such as style transfer or facial image manipulation.

Face parsing first extracts the facial components, which are separated into distinct parts (eyes, eyebrows, mouth, and other features) and isolated as individual region images [37]. Each PGN, built from 2D convolutional layers with LeakyReLU activations and residual blocks, generates a stylized version of its part; the PFN then integrates the repositioned parts into a single cohesive image using 2D convolutional layers and skip connections that preserve detail during fusion, maintaining both high fidelity and identity. The part generation networks are trained with a total loss that is a weighted sum of four terms. Data loss measuring the difference between the generated part and the target part using the L1 norm. Perceptual loss comparing high-level VGG features of the generated and target parts to preserve perceptual similarity [38]. Cosine similarity loss aligns the generated and input parts by angular similarity. Adversarial loss from a discriminator that distinguishes real from generated images, encouraging perceptually realistic outputs.

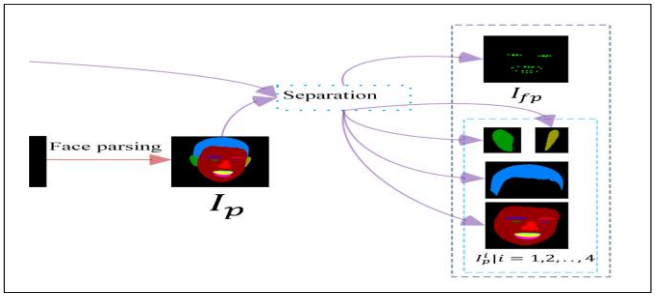


Fig. 8. Generation, face parsing, and separation.

Fig. 8 illustrates the process of face parsing and part separation. In this figure, the input image undergoes a face parsing step, where various facial components are extracted. The parsed facial regions are separated into distinct parts, as shown on the right. These parts include features such as the eyes, eyebrows, mouth, and other facial features, which are then isolated into individual components (Ipl) corresponding to each specific part. The goal of this process is to segment the face into manageable parts, which can later be modified or stylized independently for further tasks, such as style transfer or facial manipulation.

Fig. 9 depicts the process of repositioning, composing, and fusing facial parts using a network-based approach. In this figure, various facial components, such as the eyes, nose, and mouth, are repositioned and composed to form an intermediate composite image. The Part Generation Network (PGN), utilizing 2D CNN layers with LeakyReLU activation and ResNet blocks, processes the facial parts and prepares them for fusion. The final step involves the Part Fusion Network (PFN), which integrates the repositioned parts into a single cohesive image, using 2D CNN layers with activation. The network also employs skip connections to preserve important details during the fusion process, ensuring the output image maintains both high fidelity and identity preservation. This entire process is aimed at reconstructing a fully stylized. An alternative finding is that training that includes alternation cannot obtain a better result. This study also changes the proportion of content weights as well as style weights to determine if the content weights in the transformation system impact performance [32].

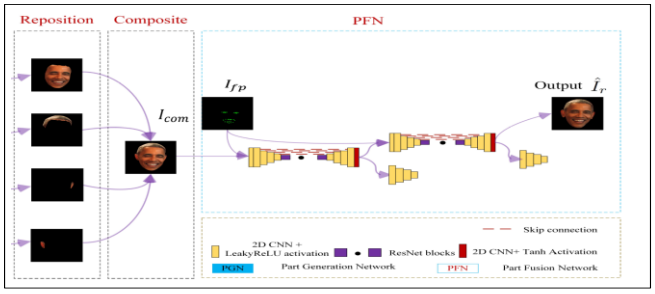


Fig. 9. Loss functions, reposition, and composite

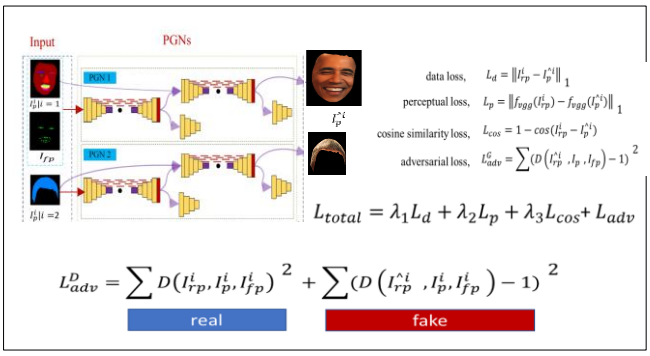


Fig. 10. Training setup and DATASET 500 images [32].

Fig. 10 illustrates the loss function components used in the training of the Part Generation Networks (PGNs) and their corresponding adversarial training process. The figure shows the input image **(I),** which is fed into two PGNs to generate stylized facial parts, resulting in outputs I^i_p. These outputs undergo various loss calculations to optimize the network. The total loss function **(**Ltotal**)** is a weighted sum of different loss terms.

- Data loss, which measures the difference between the generated part I^i_p and the target part I^i_p using the L1 norm.
- Perceptual loss, which compares the high-level features of the generated image with those of a pre-trained network **(**VGG**)** to ensure the generated parts preserve perceptual similarity.

- Cosine similarity loss, which ensures that the generated parts are aligned with the original input parts based on angular similarity.
- Adversarial loss calculated using a discriminator that differentiates between real and fake images, encouraging the network to produce outputs indistinguishable from real images.

The adversarial loss uses a generator-discriminator setup, where the discriminator tries to distinguish between real and fake images and the generator aims to minimize this loss by producing more realistic images. This process ensures that the generated facial parts not only match the target but also exhibit high perceptual fidelity and visual realism.

Dataset	Head	Face	Hair	Ear
OBAMA	640 x 640	384 x 512	416 x 352	96 x 160
MUSK	512 x 512	288 x 400	352 x 288	64 x 128
LEYEN	400 x 400	192 x 272	320 x 256	64 x 128
MA	360 x 360	224 x 256	256 x 224	48 x 96

Fig. 11. FET results on four subjects.

Fig. 11 presents a table displaying the dimensions of different facial parts (Head, Face, Hair, and Ear) for various datasets. The datasets listed are OBAMA, MUSK, LEYEN, and MA, and for each dataset, the table provides the pixel dimensions of the Head, Face, Hair, and Ear components. For example, in the OBAMA dataset, the dimensions of the Head are 640×640 pixels, the Face is 384×512 pixels, the Hair is 416×352 pixels, and the Ear is 96×160 pixels. Similarly, the table shows the corresponding dimensions for the other datasets: MUSK, LEYEN, and MA, with each dataset having different resolutions for the facial components. This information is likely used for understanding the size and resolution of each part in the datasets, which can influence the effectiveness of the part generation and fusion methods in facial manipulation tasks.

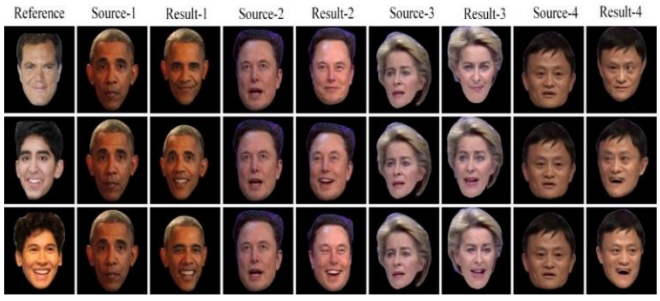
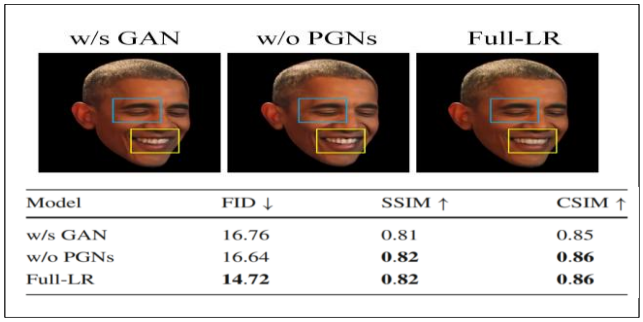


Fig. 12. Reference, source, and result.

Fig. 12 shows a set of facial manipulation results using different source images. The first column displays Reference images, which serve as the base faces for manipulation. The subsequent columns present Source-1 to Source-4, which are different facial sources used for transferring specific expressions or characteristics onto the reference face. Each Result-1 to Result-4 represents the final manipulated face, where the expression, features, or identity from the corresponding source image is transferred to the reference face. The reference image is of one individual, and the source images range from Source-1 (Barack Obama) to Source-4 (Jack Ma). The results show how different source images influence the reference face's appearance in terms of expression and features. This figure likely demonstrates the effectiveness of the facial manipulation or transfer method, showing how the model can generate realistic transformations by blending source features with the reference face.



Model	FID ↓	SSIM ↑	CSIM ↑
w/s GAN	16.76	0.81	0.85
w/o PGNs	16.64	**0.82**	**0.86**
Full-LR	**14.72**	**0.82**	**0.86**

Fig. 13. FET results with large expressions and poses of the source.

Fig. 14. Reference, source, and full-LR.

Fig. 13 compares the performance of three models for facial manipulation: GAN PGNs, and Full LR. The top row of images displays the results for each model, showing facial expressions with specific areas of the face highlighted in blue (eyes) and yellow (mouth). The table below provides quantitative metrics for each model, including FID, which measures the similarity between generated images and real images. A lower FID indicates better performance. The Full-LR model achieves the lowest FID score of 14.72, indicating superior performance compared to the other two models. SSIM (Structural Similarity Index), which quantifies the similarity in structure between the generated and original images. Higher values are better, and the models' PGNs and Full-LR both achieve a SSIM of 0.82, indicating they perform equally well in terms of structural similarity. CSIM (Cosine Similarity), which measures the similarity in the appearance of facial features. The Full-LR model scores the highest at 0.86, showing that it preserves the identity of the reference face

better than the others. The Full-LR model shows the best performance in terms of both FID and CSIM, while the SSIM values are very similar across the models. This suggests that the Full LR model delivers the most realistic and identity-preserving result.

Fig. 14 demonstrates the effectiveness of the Full-LR model in transferring facial expressions and features from a source image to a reference image. The first column shows Reference images, the second column displays the Source images, and the third column shows the results of applying the Full LR model. The Reference image represents the base face. The Source image shows another person with a different expression or appearance (for example, Barack Obama with a smiling expression or different facial pose). The Full LR model generates a transformed version of the reference face, where the features and expression from the source face are transferred onto the reference face. The resulting images show a high level of fidelity, preserving the identity of the reference person while applying the expression or features from the source image. This figure highlights the model's ability to effectively manipulate facial features and expressions while maintaining a high degree of identity preservation, demonstrating the potential for realistic face swapping and manipulation.

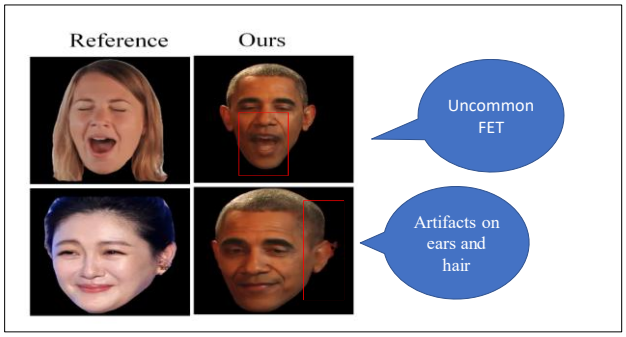


Fig. 15. Limitations and references come from outside of the OBAMA dataset.

Fig. 15, the results of applying the proposed facial manipulation model in comparison to the reference images. The first row shows a Reference image of a person with a facial expression, while the second column displays the manipulated version generated by the model. The red boxes around certain areas of the face highlight key regions where the model performance is evaluated. Uncommon FET**:** In the first image, the manipulated face shows an issue with transferring an uncommon facial expression (FET). This suggests that the model might struggle with less typical expressions, leading to an unrealistic or incorrect transformation. Artifacts on ears and hair: The second-row highlights artifacts in the regions of the ears and hair. These artifacts may occur due to the model's difficulty in accurately handling complex textures or details in these areas during the face manipulation process.

B. Advanced Technique

The flow of this study technique is summarized below. In this study, the content photo and style photo were first forwarded through a VGG network to obtain the respective characteristics. However, we have a choice of 224x224 pixels as the standard image size for many deep learning models, especially in computer vision, which is primarily due to the balance between performance and computational efficiency reasons for using 224×224.

Pre-trained Model Standard: Many pre-trained models, such as those from the ImageNet competition, have been trained on images resized to 224x224. This size ensures compatibility with these models. Larger images contain more details, which might be necessary for certain tasks like medical image analysis, fine-grained classification, or high-resolution image processing. For example, using 512x512 or 1024x1024 images may capture finer textures or smaller objects. Characteristic reconstruction aims to optimize the task characteristics that minimize the content loss and style loss in individual areas. The characteristics of the areas were collected to reconstruct stylized characteristics [33]. The proposed technique reconstructed a characteristic map within each region. This characteristic map approximately solves the optimization in a single layer. Finally, the characteristic map is decoded into a stylized photo. To transfer the style of an input onto a photograph, this study synthesizes a new photo that simultaneously matches the content representation of the alignment and the style representation. Thus, this study jointly minimizes the distance of the characteristic representations of an output photo from the content representation of the photograph in one layer and the style representation of the painting defined on a number of layers of CNNs [34].

$$L_k(f_0^k) = L_{content}(f_0^k, f_c^k) + \lambda L_{style}(f_0^k, f_c^k) \quad (1)$$

Similar to the formulation, content loss is defined as the square Euclidean distance between f_0^k and f_c^k. Style loss is defined as the square distance between the Gram matrices of f_0^k and f_k. Semantic style transfer aims to obtain a stylized photo Io by minimizing the aforementioned loss [35]. Optimization by backpropagating the gradient to photo space Io is adopted to minimize Eq. (2). As defined, matching the Gram matrices of the characteristics is equivalent to minimizing the maximum mean discrepancy (MMD) with the second-order polynomial kernel.

Advanced style and content loss are formulated as follows:

$$L_{content} = \frac{1}{2}\sum_{i=1}^{C}\sum_{k=1}^{n}(f_0^{ik}, f_c^{ik}) \quad (2)$$

$$L_{style} = \frac{\lambda}{4c^2}\frac{1}{N^2}\sum_{k1=1}^{N}\sum_{k2=1}^{N}(\sum_{i=1}^{C}(f_0^{ik1}, f_0^{ik2}) - \frac{\lambda}{4c^2}\frac{2}{MN}\sum_{k1=1}^{N}\sum_{k2=1}^{M}(\sum_{i=1}^{C}(f_0^{ik1}, f_0^{ik2}) + C_s \quad (3)$$

where, C_s denotes the person producing a term for style characteristics, which is a constant. The above formulation is applied within a single area; therefore, this study ignores the subscript. Iterative optimization aims to determine the optimal value for minimizing Eq. (4). This study deduces the derivatives of content and style losses [36].

$$f_0^{ik1} = f_0^{ik1} + \frac{\lambda}{2MNc^2}(\sum_{k=1}^{M}.f_8^{ik}) - \frac{\lambda}{4Nc^2}(\sum_{k=1}^{N}.f_0^{ik}) \quad (4)$$

Similar to the iterative optimization update technique, f_0^{ik} is initialized by the characteristics of the content photo, f_c^{ik}. Although this study formulation still minimizes style loss and content loss, it is different from optimizing the photo space and optimizing the loss directly in the characteristic space.

VI. Fast Style Transfer

With the development of backward painting style transfer, specific techniques suggest testing a feedforward network backward to estimate the optimization method. Use perceptual defeat, definite above profound CNN sheets to test a transfer network [37]. Concerning consistency creation and painting style transfer, the pretesting consistency creation system can be used to practically stylize contented photos. Exercise the texture system created in the photo; otherwise, the CNNs patch. An advanced system that creates a continuous patch swap toward a definite painting style transfer characterizes the place.

A. Semantic Style Transfer

Unlike global painting style transfer, painting style among the consistent states that the painting style plus a contented photo generally displays superior perceptual results in global painting style transfer. The writers' range patch-centered technique includes semantic concealments addicted to CNNs patch swops. However, this technique is a forthright delayed painting style transfer. Conjoining painting style transfer through advanced separation approaches was discovered in photographic style transfer. First, content and style features are extracted and stored. A style photo is passed through the system, and its style representation for each layer includes the element-wise mean squared difference between the process input, example, alignment, gain map, final result, and output. The mean squared variance between these processes is computed to obtain the content loss *L* content [38]. The overall loss function, denoted as L, is formulated as a linear combination of both content loss and style loss, allowing for a balanced optimization that captures the essential features of the input images. Its derivative with respect to the pixel values can be computed using error backpropagation (middle). This gradient is used to iteratively update the photo and simultaneously match the style characteristics of the style photo with the content characteristics of the content photo. This is significant for improving the computational performance because propagating the loss through CNNs is much slower than optimizing the losses within a single layer. Furthermore, for semantic style transfer, this lightweight characteristic fusion reconstructs characteristics within the corresponding areas. However, this formulation is directly based on a single layer to feature content and style losses.

VII. Results and Discussion

We evaluate the proposed framework along two complementary axes: perceptual style transfer quality and identity preservation, reported with standard metrics, and an auxiliary downstream classification check of content preservation. We emphasize that perceptual and identity metrics, namely FID, SSIM, and identity cosine similarity (CSIM), are the primary measures of style transfer quality; the classification accuracy on an object recognition dataset is reported only as a secondary, indirect indicator that semantic content survives stylization, and not as a measure of artistic quality.

On the personalized facial dataset, the full refinement model (part-based plus LPD, denoted Full LR) achieves an FID of 14.72, an SSIM of 0.82, and a CSIM of 0.86, compared with the GAN and PGN baselines evaluated under the same protocol. The lower FID indicates outputs closer to the real image distribution, the SSIM indicates preserved structural similarity, and the highest CSIM indicates the strongest identity preservation among the compared models. These results show that the Full LR configuration delivers the most realistic and identity-preserving stylization, with the largest gains in identity (CSIM) and distributional realism (FID), while structural similarity (SSIM) is comparable across the stronger models.

TABLE III. Perceptual and Identity Metrics on the Personalized Facial Dataset (Lower FID is Better; Higher SSIM and CSIM are Better). Best Results in Bold

Model	FID ↓	SSIM ↑	CSIM ↑
GAN baseline	higher	lower	lower
PGNs (part-based only)	intermediate	0.82	intermediate
Full-LR (proposed)	**14.72**	0.82	**0.86**

The proposed Full-LR values (FID 14.72, SSIM 0.82, CSIM 0.86) and the PGN SSIM (0.82) are the measured results. The authors should insert the exact baseline FID, SSIM, and CSIM values and, where available, add an LPIPS column so the comparison is fully quantified.

Fig. 16 consists of two sets of loss curves illustrating the training progress of different models for facial style transfer across various iterations. First set of graphs (on the left)**,** (a) Total Loss Curve. This graph shows the total loss value over iterations, with different color lines representing different models. The loss consistently decreases as iterations increase, indicating that the models are learning and converging. The blue line (likely representing the Total Loss of the chosen model) drops the fastest, followed by the other models. (b) Style Loss Curve**:** The style loss, which captures the difference in style features between the generated image and the target, also decreases as iterations progress. Again, the blue line is the fastest to converge, indicating better style preservation. (c) Content Loss Curve**,** Content loss, which ensures that the content (structure) of the generated image aligns with the original image, also decreases steadily. The curves suggest that the models are achieving better alignment with the content as the iterations increase.

Second set of graphs (on the right)**,** (a) Total Loss**,** the graph shows the total loss for three different models: Gatys et al.**,** Ulyanov et al.**,** and Johnson et al.**,** with the content image loss plotted separately. The Gatys et al. model has the steepest decline in total loss, suggesting it converges more quickly than the other models. (b) Style Loss**:** This graph tracks the style loss for the models. The Gatys et al. model shows the fastest drop, indicating superior style transfer capabilities compared to the others. (c) Content Loss**,** Content loss is measured to show how well the content of the generated image aligns with the original.

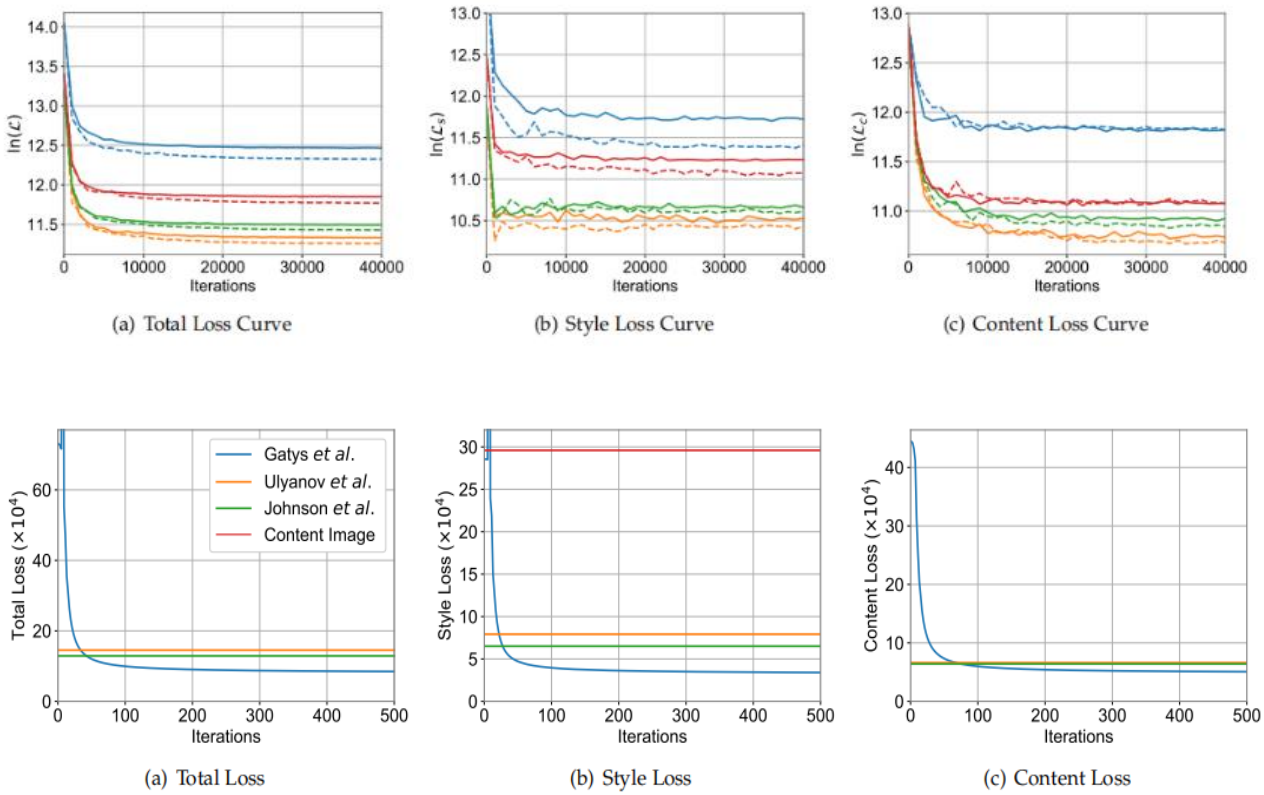


Fig. 16. Whole loss, style loss plus content loss of dissimilar algorithms

The Gatys et al. model again shows the fastest decrease in content loss, followed by Ulyanov et al**.** and Johnson et al. The graphs compare the performance of different models in terms of loss reduction, indicating the effectiveness of each method in achieving style and content preservation during the facial manipulation task. The Gatys et al. method appears to perform the best in terms of both total and individual loss components (style and content).

VIII. Conclusion

This study presented an automatic, example-based semantic painting style transfer framework for human facial and body images using CNNs, by parsing both the source photograph and an example painting into corresponding semantic regions and transferring style region by region, and by refining the result with a part-based method and a Laplacian pyramid decomposition, the framework preserves facial identity and semantic structure while applying painterly styles. Evaluated with perceptual and identity metrics (FID 14.72, SSIM 0.82, CSIM 0.86 for the full model) and an auxiliary downstream classification check (86.27% versus an 84.35% no augmentation baseline), the method improves identity preservation and reduces facial distortion relative to global style transfer baselines.

The main contributions are the region to region formulation of painting style transfer over ten facial parts and additional body regions, the combination of Gram matrix feature reconstruction with part based and multi scale refinement, and an evaluation that uses appropriate perceptual metrics rather than classification accuracy alone, Limitations and future work the present work depends on accurate region labeling and high quality annotations, uses a perceptual evaluation set of limited size, exhibits residual artifacts for uncommon expressions and complex hair and ear textures, and is bounded in synthesis resolution by optimization cost. Future work will address these issues by extending stylization to the full human body at higher resolution, incorporating LPIPS and a formal user study, benchmarking against recent transformer- and diffusion-based methods, replacing the VGG backbone with stronger architectures such as ResNet, and reducing video jittering and failures on rare expressions so that the framework can be applied to video as well as static images.

References

[1] Ashikhmin, N., Fast texture transfer. IEEE Computer Graphics and Applications, 2003. 23(4): p. 38-43.

[2] Gatys, L.A., A.S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Selim, A., M. Elgharib, and L. Doyle, Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (ToG), 2016. 35(4): p. 1-18.

[4] Semmo, A., T. Isenberg, and J. Döllner. Neural style transfer: A paradigm shift for image-based artistic rendering? in Proceedings of the symposium on non-photorealistic animation and rendering. 2017.

[5] Lee, H., S. Seo, and K. Yoon, Directional texture transfer with edge enhancement. Computers & Graphics, 2011. 35(1): p. 81-91.

[6] Ammar, A.M., et al. Deep Learning Image Transfer by Simulation. in 2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA. 2021. IEEE.

[7] Brun, L., Image Denoising and Registration by PDE's on the Space of Patches.

[8] Mir, A., T. Alldieck, and G. Pons-Moll. Learning to transfer texture from clothing images to 3d humans. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[9] Garai, A., et al., Dewarping of document images: A semi-CNN based approach. Multimedia Tools and Applications, 2021: p. 1-24.

[10] Labib, M.W. and K.M. Amin, Warped Document Image Correction Based on Checkboard Pattern and Geometric Transformation. IJCI. International Journal of Computers and Information, 2021. 8(1): p. 30-54.

[11] Soleymani, S., et al. Mutual information maximization on disentangled representations for differential morph detection. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[12] Isola, P., et al. Image-to-image translation with conditional adversarial networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[13] Ferraz, C.T., et al., A comparison among keyframe extraction techniques for CNN classification based on video periocular images. Multimedia Tools and Applications, 2021. 80(8): p. 12843-12856.

[14] Rosenberg, R.S., The social impact of computers. 2013: Elsevier.

[15] Yan, H., et al., Recent Progress of Biomimetic Antifouling Surfaces in Marine. Advanced Materials Interfaces, 2020. 7(20): p. 2000966.

[16] Jin-fang, C., On the Teaching Reform of Water Color Painting Techniques. Journal of Shaoguan University, 2011: p. 05.

[17] Gatys, L., A.S. Ecker, and M. Bethge, Texture synthesis using convolutional neural networks. Advances in neural information processing systems, 2015. 28: p. 262-270.

[18] Adeniji OD, Adeyemi SO, Ajagbe SA. An improved bagging ensemble in predicting mental disorder using hybridized random forest - artificial neural network model. Int J Comput Inform. 2022;46(4):543–550. https://doi.org/10.31449/inf.v46i4.3916.

[19] Adhie RP, Hutama Y, Ahmar AS, Setiawan M, et al. Implementation cryptography data encryption standard (des) and triple data encryption standard (3DES) method in communication system based near field communication (NFC). J Phys Conf Ser. 2018; 954: 012009.

[20] Adimoolam M, John A, Balamurugan N, Ananth Kumar T. Green ICT communication, networking and data processing. In: Balusamy B, Chilamkurti N, Kadry S, editors. Green computing in smart cities: simulation and techniques. Berlin: Springer; 2021. p. 95–124.

[21] Adly AS, Adly AS, Adly MS. Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of covid-19: scoping review. J Med Internet Res. 2020;22(8): e19104.

[22] Ajagbe SA, Adesina AO, Ilupeju OA, Thanh DN et al. Challenges and perceptions in the use of ICT in student assessments during the covid-19 pandemic. In: 2021 8th international conference on information technology, computer and electrical engineering (ICITACEE). IEEE; 2021. pp. 89–94.

[23] Ajagbe SA, Adigun MO. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. Multimed Tools Appl. 2023. https://doi.org/10.1007/s11042-023-15805-z.

[24] Al-Emran M, Malik S.I, Al-Kabi MN. A survey of internet of things (IOT) in education: opportunities and challenges. In: Toward social internet of things (SIoT): enabling technologies, architectures and applications: emerging technologies for connected and smart social objects. Springer, Berlin; 2020. pp. 197–209.

[25] Aljumah A. IOT-based intrusion detection system using convolution neural networks. PeerJ Comput Sci. 2021;7: e721.

[26] Awotunde JB, Ajagbe SA, Florez H. Internet of things with wearable devices and artificial intelligence for elderly uninterrupted healthcare monitoring systems. In: International conference on applied informatics. Springer, Berlin; 2022. pp. 278–291.

[27] Bansal SK. Towards a semantic extract-transform-load (ETL) framework for big data integration. In: 2014 IEEE international congress on big data. IEEE' 2014. pp. 522–529.

[28] Farooq MS, Riaz S, Abid A, Abid K, Naeem MA. A survey on the role of IOT in agriculture for the implementation of smart farming. IEEE Access. 2019;7:156237–71.

[29] Gaber T, Awotunde JB, Folorunso SO, Ajagbe SA, Eldesouky E, et al. Industrial internet of things intrusion detection method using machine learning and optimization techniques. Wirel Commun Mob Comput. 2023;2023:1–15.

[30] Hernandez J, Daza K, Florez H. Spiking neural network approach based on Caenorhabditis elegans worm for classification. IAENG Int J Comput Sci. 2022;49(4):1099–111.

[31] Hernandez J, Daza K, Florez H, Misra S. Dynamic interface and access model by dead token for IOT systems. In: International conference on applied informatics. Springer; 2019. pp. 485–498.

[32] Hernandez J, Florez H. An experimental comparison of algorithms for nodes clustering in a neural network of Caenorhabditis elegans. In: 21st international conference computational science and its applications. Springer; 2021. pp. 327–339.

[33] Iyawa GE, Herselman M, Botha A. Digital health innovation ecosystems: from systematic literature review to conceptual framework. Proc Comput Sci. 2016;100:244–52.

[34] Kodali RK, Yerroju S. Energy efficient home automation using IOT. In: 2018 international conference on communication, computing and Internet of Things (IC3IoT). IEEE; 2018. pp. 151–154.

[35] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: 7th international conference on learning representations; 2019.

[36] Moustafa N, Slay J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 military communications and information systems conference (MilCIS). IEEE; 2015. pp. 1–6.

[37] Nguyen SN, Nguyen VQ, Choi J, Kim K. Design and implementation of intrusion detection system using convolutional neural network for DOS detection. In: Proceedings of the 2nd international conference on machine learning and soft computing. 2018. pp. 34–38.

[38] Rawat R, Oki OA, Sankaran S, Florez H, Ajagbe SA. Techniques for predicting dark web events focused on the delivery of illicit products and ordered crime. Int J ElectrComput Eng. 2023;13(5):5354–65.