IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 July 2022, accepted 26 August 2022, date of publication 5 September 2022, date of current version 7 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3204040

RESEARCH ARTICLE

An Edge Filter Based Approach of Neural Style Transfer to the Image Stylization

SHUBHAM BAGWARI[1], KANIKA CHOUDHARY[1], SURESH RAIKWAR[1], RAHUL NIJHAWAN[2], SUNIL KUMAR[2], AND MOHD ASIF SHAH[3]

[1]Department of Computer Science and Engineering, TIET, Punjab 147004, India
[2]School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248007, India
[3]Department of Economics, Bakhtar University, Kabul 1001, Afghanistan

Corresponding author: Mohd Asif Shah (ohaasif@bakhtar.edu.af)

ABSTRACT Transferring artistic styles onto any image or photograph has become popular in industry and academia in recent years. The use of neural style transfer (NST) for image style transfer is getting more popular. Convolution Neural Networks (CNN) based style transfer provides a new edge and life to the images, videos, and games. The re-rendering procedure of the content of one image with the style of another using various models and approaches is widely used for image style transfer. However, there are many drawbacks, including image quality, enormous loss, unrealistic artefacts, and the style of localized regions being less compared to the desired artistic style. For the reason that transfer technique fails to capture detailed, miniature textures and keep the true artwork's texture scales. We propose a multimodal CNN that stylizes hierarchically with several losses of increasing sizes while considering faithful representations of both colour and luminance channels. We may transfer not only large-scale, evident style cues but also subtle, exquisite ones by effectively handling style and texture cues at different sizes using various modalities. Our approach providing aesthetically pleasing results and is more comparable to multiple desirable creative styles using colour and texture cues at different scales.

INDEX TERMS Convolution neural network, deep learning, image proccesing, neural networks, neural style transfer.

I. INTRODUCTION

The Neural Style Transfer (NST) is a technique to create a stylized image by bringing together a style reference image with a content image [1]. In NST, the content image and a style reference image (artwork by any well-known artist) have been used as input to produce a stylized image (appears to be the same image as the content image, however painted with the style of the style reference image) by blending both of the input images, as shown in Fig. 1.

The NST can be used in a variety of disciplines, such as scene recognition using video style transfer, speech recognition using speech style transfer, and many more [1], [2]. It has some commercial and industrial applications [3] like Prisma, meSTro, clinical practice, and more industrial applications remain to be seen.

There are many key challenges while performing style transfer. First, one has to do with two competing objectives to accomplish. On the one hand, the researcher are trying to produce very robust local impacts (e.g. sky colour in a simple image), and they do not want the transfer to have any geometric effect (for example, window grids should stay as grids and not be distorted). The diversity of real-world scenes adds another layer of challenge. ''The scene's semantics should be preserved during the transfer.'' For example, in a cityscape, the appearance of buildings should mimic the appearance of other buildings, and the sky should match the sky.''

Our work is inspired by the certainty that the number of style transfers has increased exponentially in the last decade. Over the NST, a lot of work has been done. There are many

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

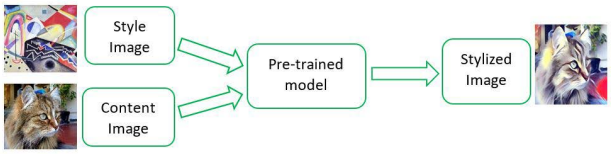

FIGURE 1. Generic model of neural style transfer.

commercial apps are available in the market, which is liked by most of the people.

In the arts, style transfer (ST), or repainting an existing artwork in a different style, is a difficult nevertheless exciting issues. Because of the groundbreaking work of Gatys *et al.* [4], where a pre-trained deep learning network for visual identification is utilised to attain both style and content representations, and yields visually spectacular outputs, this challenge has recently become a popular topic both in industry and academia. Unfortunately, due to the online iterative optimization technique, the transfer run time is unreasonably long. To tackle this situation, a feed-forward network (FFN) could be train offline using the same loss criterion to produce visually similar stylized results (nonetheless still a bit inferior). At application time, only a single FFN inference pass is required. This results in a hundreds of times quicker computational algorithm [5], [6].

Whereas previous work has produced visually satisfactory results for a variety of artworks, there are two major drawbacks:

1) Because current FFN [5], [6] is trained on a certain style of image resolution, a scale mismatch occurs when the resolution is changed (larger or smaller). Implementing a model trained with a size 256 style guide to higher-resolution images, for example, would result in results with a smaller texture scale compared to artistic style.
2) Current networks on high-resolution images are frequently inadequate to grasp tiny, sophisticated textures, such as brushstrokes, of numerous types of artworks.

Since it has been proved that these FFN perform good on paintings including complex, larger-scale textures and strokes that are evident, such as Vincent van Gogh's The Starry Night, creative styles are far more diverse. Because diverse artistic styles are distinguished by delicate, subtle brushes and strokes, we have found that the outcomes of these ST networks are frequently unsatisfactory for a wide range of artistic styles.

This paper proposed edge filter based hierarchical deep convolutional neural network (CNN) architecture for efficient artistic ST in this paper. It consists of four parts:

1) It provides a hierarchical network and an edge filter based training technique that can understand both larger-scale texture distortions, and coarse and delicate, fine brushstrokes of an artistic style using several scales style image.
2) For style transmission, our network handles both luminance channels and colour channels, rather than just RGB colour channels.
3) Our end-to-end CNN network architecture and hierarchical training technique permits to blend various models into single network to accommodate increasingly higher image sizes.
4) We demonstrate that our hierarchical ST network could effectively capture both complex texture patterns and coarse via testing.

We distinguish our hierarchical ST network as a multi-modal transfer (MT) from FF-ST networks of single stylisation loss [5], [6], named as solitary transfer because it is trained including various stylisation losses at various scales utilising fusion of methods. The benefits of MT on learning several levels of textures, such as style, fine brushwork, colour, and huge texture distortion. It's worth noting that our method can more accurately mimic the artwork's brushwork. It could be utilized to train a combination model to stylize a single image using a variety of different aesthetic styles.

The organisation of this paper is as follows. This paper starts with a literature review of recent studies of artistic rendering methods using CNNs in Section II. The proposed method defined in Section III. In Section IV we conduct experiments and analysis.. We have discussed the present challenges and issues in Section V. Finally, conclusion of the paper and delineates several promising directions for future research in Section VI.

II. LITERATURE REVIEW

CNN's usage proved crucial; Gatys *et al.* first attempted recreating famous artwork styles using natural images [2]. Using a pre-trained CNN, the researcher advanced a technique that used the content of an image as feature responses along with the style of artwork as summary feature statistics. In their experiments, the studies showed that a CNN could extract content information from an arbitrary image and style information from a well-known artwork. Gatys *et al.* proposed a method using CNN feature activations towards reassemble the style of famous artworks along with the content of a specified image, based on this discovery [2]. Their algorithm's main idea is to iteratively improve an image to fulfil desired CNN feature distributions that include both the artwork's style and the photo's content information. This paper categorized ST in various ways.

1) NST WITH PATCH

The MRF-based NST algorithm inspired Li and Wand's [7] work. They solve the efficiency problem by developing a Markovian feed-forward network that utilises adversarial training. A Patch-based Non-parametric technique with MRFs is similar to their approach [8]. Because of their patch-based design, their method surpasses Johnson *et al.* and Ulyanov *et al.* methods in preserving intelligible textures in involuted images. However, as both techniques do not take this into consideration semantics, it underperforms with

non-textured styles (such as face pictures). Their algorithm also has other issues, such as a lack of attention to detail and differences in brush strokes, both of which are essential visual elements. The output of Li and Wand [7] is somewhat less remarkable. The training process is not as stable as it could be because it [7] is based on a Generative Adversarial Network (GAN). However, It has been observed that GAN-based style transfer is a promising direction.

Peng *et al.* [9] are able to transfer universal face photo-sketch styles without being limited by drawing styles. As a result, the method can produce texture, lighting information, and shading in the synthesised results, which is useful for face recognition in law enforcement situations. For patch representation, they plan to use more advanced deep NN architectures. For universal-style transmission, it is also conceivable to combine complementary representation skills of various deep networks. This technique may additionally be utilised in-order-to face sketch-to-photo conversion by swapping roles of the sketches and photos.

2) NST WITH TEXT

Zhang *et al.* [10] looked into a more generic manner of representing style/content information, assuming that all styles/contents have the same feature. This technique was shown to be more generalisable. However, this strategy is only applicable to Chinese and English. Hence, having limited applicability.

The architecture presented by Zhu *et al.* [11] enables the possibility of style transfer between several styles and languages. Given the appropriate reference sets, it can generate text in any style-content combination. As a result, it is predicted to generalise new language styles and contents effectively. However, testing shows that this model does not perform well for font and texture generation on a distinct IR without fine-tuning. Future research will focus on finer-grained deep similarity fusion for improved text ST results across languages.

Chen *et al.* [12] style transfer network is the first to link back to the traditional text mapping methods, providing a fresh outlook on NST. However, the auto-encoder could include semantic segmentation as an additional layer of supervision in the region decomposition, resulting in a more spectacular region-specific transfer. Incorporating the proposed approach into video and stereoscopic applications is also intriguing.

3) NST ON REAL-TIME

Adaptive Convolutions (AdaConv) is a general AdaIN extension proposed by Chandran *et al.* [13] that allows for the real-time transfer of both structural and statistical styles. In addition to style transfer, their method may easily be broadened to style-based image production and different jobs in which AdaIN has been utilized so far.

Xu *et al.* [14] proposes VTNet, which is a real-time temporally coherent stylized video generator that is edge-to-edge trained from effectively limitless untagged video data. The temporal prediction branch, along with the stylizing branch of VTNet, transmits the style of a reference image toward the source video frames. Reduced loss, on the other hand, can accomplish novelty in this.

4) NST WITH VGG NETWORK

For N target styles, Li *et al.* [15] develops an N-dimensional one-hot vector as a selection unit for style selection. When it comes to arranging target styles, any one of the selecting unit's components indicates a specific style. Li takes a sample of the appropriate noise map $f(I_s)$ from a uniform distribution for any one of the selecting unit's components, then feeds $f(I_s)$ to get the matching style encoded features from the style sub-network $f(I_s)$. The required stylised outcome can be created by providing the combination of style encoded features $\mathcal{F}(f(I_s))$ and the content encoded features $Enc(I_c)$ toward the decoder component Dec of the style transfer network: $I = Dec(\mathcal{F}(f(I_s)) \oplus Enc(I_c)$. Zhang and Dana work [16] first passes every one style image within a style set along a pre-trained VGG network to get multi-scale feature activations $f(I_s)$ in various VGG layers. Thus, through their proposed inspiration levels, multi-scale $f(I_s)$ are mixed with multi-scale encoded features $Enc(I_c)$ from distinct layers in the encoder. The inspiration layers are intended to modify $f(I_s)$ to correspond the required magnitude, and they also include a learn-able light matrix that may be used to optimise feature mapping and help reduce function with an objective.

Based on whitening and colouring transforms, Yoo *et al.* [17] wavelet adjusted the transmission (WCT2). The method permits features to keep their spatial features and statistical aspects of the VGG feature space during stylisation. However, eliminating the need for semantic labels should be correct for a perfect performance thus far.

5) NST WITH IMAGE FEATURE

Johnson *et al.* [5] and Ulyanov *et al.* [6] proposed methods. Both methods are based on the alike idea: producing a stylised output including a single forward pass and pre-training feed-forward style-specific network during testing. They are primarily a difference in network architecture, with Johnson *et al.* approach broadly followingRadford *et al.* [18] network but with fractionally stridden convolutions along with residual blocks, and Ulyanov *et al.* generator network being a multi-scale architecture. Shortly after [5], [6] Ulyanov *et al.* [19] found that implement normalisation into each individual image instead of a collection of images (specifically Batch Normalisation) results in a large advancement in stylisation attribute. At the time that batch size is set to 1, instance normalisation (IN), which is the same as batch normalisation, is used to normalise a single image. IN's style transfer network converges faster than Batch Normalisation (BN) and produces better visual results. Instance Normalisation is a type of style normalisation in that style of each content image is directly normalised towards the appropriate style [20], according to one interpretation. As a result, the

objective is simpler to understand because the rest of the network simply cares about content loss.

Wang *et al.* [21] propose a sequence-level feature sharing technique for long-term temporal consistency, as well as a dynamic inter-channel filter to enhance the stylization impact ourselves. Temporal consistency further can be used in conjunction with GAN to improve performance.

Although the aforementioned Per-Style-Per-Model (PSPM) approaches may generate stylised images two orders of magnitude quicker compared to earlier image-optimisation-based NST techniques, each style image necessitates the training of distinct generative networks, which is inflexible and long-drawn-out. For example, numerous artworks (for example, impressionist paintings) have identical paint strokes but vary primarily in their colour palettes. Probably redundant to train a different network for each of them. Multiple-Style-Per-Model Neural Methods(MSPM) is offered to increase PSPM's versatility by combining various styles into a single model. While dealing with such problems have commonly two approaches: i) associating each style to a little range of parameters in a network [22]; and ii) employing a single network, such as PSPM, but with inputs for both style and content [15], [16]. In contrast, For multiple styles, Li *et al.* [15] and Zhang and Dana [16] algorithms Use the same trainable network lights on a single network. However, the model size problem that has been addressed appears to be some interplay between distinct styles, which has a minor impact on the stylisation quality.

In both NPR and NST, an aesthetic appraisal is a key concern. Many academics in the subject of NPR stress the importance of aesthetic judgement as in [23], Researchers proposes two phases to investigated such problem. These issues are progressively analytical precisely the area of NST along with NPR matured, explained in [23], Researchers require some decisive criteria to evaluate advantages of their suggested strategy over the prior art, as well as a method to analyse the appropriateness of one methodology to a certain case. Most NPR and NST articles, on the other hand, evaluate their suggested approach using metrics generated or instinctive visual comparisons from various user studies [24].

This technique successfully avoids distortion and achieves appropriate Photorealistic Style Transfers (PST) in a wide range of settings, including mimicking aesthetic edits, the time of day, season, and another, according to Luan and Paris [25]. This study proposes the first solution to the problem. However, some other breakthroughs and enhancements can be made.

The breakthrough image stylization method developed by Cheng *et al.* [26] includes an additional structure representation. Firstly, the depth map represents the global structure. Secondly, the image edges represent the local structure details. It perfectly describes the spatial distribution of all elements in an image and the formation of notable items. The method provides terrific visual effects, especially when processing images hypersensitive to structural deformation, such as images having many items possibly at various depths or notable items with distinct structures, as demonstrated by testing results. Even so, there is scope for improvement.

With the Adversarial distillation learning technique, Qiao *et al.* [3] can produce clear images in a short period. Although the network has been well-trained, it still has unrealistic artefacts and a high computational cost. The difficulty of simultaneously improving efficiency in PST and visual quality remains unsolved.

For style transfer between unpaired datasets, Li *et al.* [27] efficiently handle and preserve the features on essential salient locations. Two new losses are proposed to improve the overall image perception quality by optimising the generator and saliency networks. Furthermore, tasks with modified saliency objects, such as dog2cat translation, are not suited for the proposed SDP-GAN.

Ling *et al.* [28] use the background to expressly formulate the visual style, which they then apply to the foreground. Despite its progress, the proposed method still has two major drawbacks. To begin with, it is unclear why employing RAIN exclusively in the encoder yields such a low gain. Second, the model will reduce the aesthetic contrast and attenuate the sharp foreground object in examples with sharp foreground objects and dark backgrounds.

The mismatch between the human sense of stylisation quality and the classic AST style loss was studied by Cheng *et al.* [29]. During training, the core cause of the problem was identified as the style-agnostic cluster of sample-wise losses. They used a novel style-balanced loss with style-aware normalisation to obtain theoretical limitations for the style loss. In the future, more substantial limitations for the style loss could be derived toward enhancing style-aware normalisation.

Because of superior image-object retention, Lin *et al.* [30] outperforms rival systems in terms of achieving higher nighttime vehicle detection accuracy. However, its uni-modality is a disadvantage. They want to try explicitly encoding a random noise vector to the structure-aware latent vector in the future to achieve model diversity when executing a unitary image-to-image translation.

Virtusio *et al.* [31] propose a single-style input style transfer strategy that focuses on an H-AI encouraged architecture which involves human control across the stylization procedure, allowing for a variety of outputs. This method is inspired by the various perceptual qualities discovered in a sole style image. For example, it could include a variety of different colours and textures. However, neural style transfer is focused on accelerating the process. On the other hand, these works may only discover a limited amount of styles.

Xiao *et al.* [32] approach can efficiently handle the issues of blurring, poor diversity, and distortion of the output images relative to other data augmentation methods. Despite the lack of realistic images, extensive experimental findings show that this strategy can still be effective. They expanded the field of deep learning's application possibilities for simulated images. However, the image quality created is too low, style

FIGURE 2. Style image used to obtain stylized images for each content image (Fig.4(a)).

transfer takes a long time, and the quantity of augmented data produced unavoidably brings out the CNN over-fitting issue.

III. PROPOSED METHOD

Wang *et al.* [33] introduced a multi-modal CNN that executes hierarchical stylization with multiple losses of increasing scales while taking into deliberation precise depiction of both colour and luminance channels.

The advantage of the method in [33] is the use of colour and texture cues at various scales. It can provide an aesthetically pleasing outcome and is further comparable to multiple desirable creative styles. However, other losses could be investigated to convey creative expression at various scales better. In order to generalize the method in [33] for larger images, we investigated alternative loss networks with less memory and produced good images.

The proposed method is composed of two primary building blocks: loss block, and (LB) feed-forward multi-modal block (FFMB), as shown in Fig 3. The FFMB is a deep residual hierarchy CNN. It has three sub-networks: fusion block, resolution enhance block and tuning block are represented by Ψ_1, Ψ_2, and Ψ_3. At the uppermost level, the FFM Block receives an image α as input, and each input is run through a filter grid box and trained to produce various output images β_k of enlarging dimensions.

$$\hat{\beta_k} = f([\cup_k^{i=1}\delta_i]); \alpha \tag{1}$$

The loss network is then used to calculate a stylization loss for each output image separately. The total loss is calculated as a weighted average of all stylization losses. The loss network and the definition of total loss will be shown in Sec III-A.

When applied to larger images, the Multimodal Transfer Network (MT Network) stylizes the image hierarchically at test time to achieve the same stylization effect and accurate textural scale of the artworks. The input β_1 image is first downsized to 256 pixels using a bilinear downsampling layer, then stylized using the fusion block to capture the artwork's large colour and texture characteristics. The styled result is then upsampling into 512 and sent to the output β_2 by the resolution enhance block, which improves the stylization strength. The image is then scaled to 1024 pixels. Finally, the tuning block refines the output by removing local pixelization artefacts. After these three stages of processing, the highest-resolution and most visually pleasing output β_3 is obtained. While it demonstrates the method with a two-level hierarchy, the same principle can be applied recursively to allow for the stylization of increasingly more prominent photos.

A. LOSS FUNCTIONS

The single stylization loss function is first introduced in this section, followed by a hierarchical stylization loss function used to train our MT network.

1) LOSS FUNCTION: SINGLE STYLIZATION

For fast style transfer, similar to the loss defined in previous work [5], [6], Gatys *et al.* provided the stylization loss [4], in which a loss network (a pre-trained VGG- 19 network optimized for object recognition [34]) extracts the image representations.

To determine how well the generated image $\hat{\beta_k}$ integrates the content of the content target $\beta_c$with the texture and style cues of the style target β_s, two perceptual losses are defined.

Content Loss Function: It is used to measure the dissimilitude between $\hat{\beta_k}$ and β_c. In the l-th layer of the i-th feature map of the loss network applied to image α denoted by $F_i(\alpha)$. The content loss is the squared-error loss between the two feature representations at layer ł.

$$L_{content}(\hat{\beta_k}, \beta_c, l) = \sum_{i=1}^{N_l} \left\| F_i^l(\hat{\beta_k}) - F_i^l(\beta_c) \right\|_2^2 \tag{2}$$

– i.e., the content loss compares the feature maps generated through the corresponding layers directly, making it appropriate for determining spatial content similarity.

Style or Texture Loss Function The correlation between feature mappings in each layer of the loss network, according to Gatys *et al.*, can be viewed as texture representations of an image [4], [35]. The Gram matrix, whose elements are pairwise scalar products between the feature maps, provides such correlations:

$$G_{ij}^l(\alpha) = \left\langle F_i^l(\alpha), F_j^l(\alpha) \right\rangle \tag{3}$$

A set of Gram matrices G^l, l *in* L is used as the texture representations, which eliminate the spatial details nevertheless keep the intensity distribution and statistical outline of the colour of an input image. As a result, the texture loss function is defined as follows:

$$L_{text}(\beta_k\hat{}, \beta_s) = \sum_{l \in L} \left\| G^l(\hat{\beta_k}) - G^l(\beta_s) \right\|_2^2 \tag{4}$$

Certainly, the stylization loss for each output $\hat{\beta_k}$ from the MT network is described as a weighted sum of the the texture loss and content loss.

$$L_s(\hat{\beta_k}, \beta_c, \beta_c) = \omega_c L_{content}(\hat{\beta_k}, \beta_c) + \omega_2 L_{text}(\hat{\beta_k}, \beta_s) \tag{5}$$

where ω_c and ω_s are the weights of the content loss and style or texture loss, respectively.

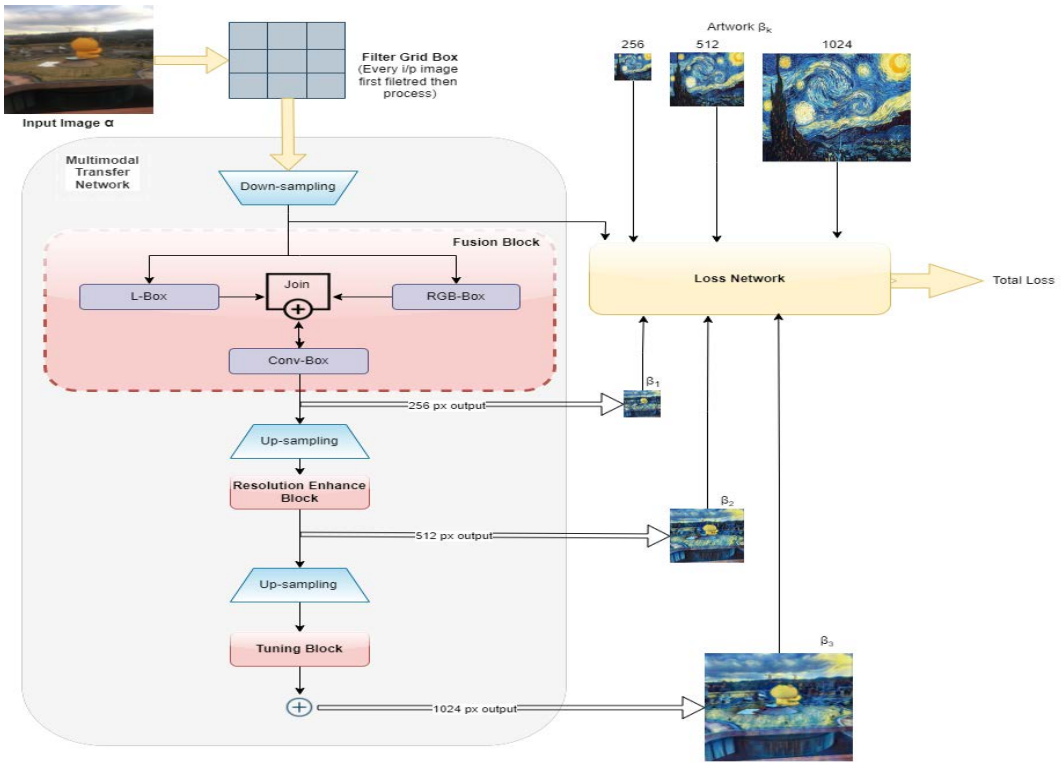

FIGURE 3. Architecture of proposed method.

2) LOSS FUNCTION: HIERARCHICAL STYLIZATION

The MT Network can produce K output results of K increasing sizes (K = 3 in the network shown in Fig. 2). Then, for each output result $\hat{\beta}_k$, stylization loss is computed.

$$L_S^k(\hat{\beta}_k, \beta_c^k, \beta_s^k) = \omega_c L_{content}(\hat{\beta}_k, \beta_c k) + \omega_s L_{text}(\hat{\beta}_k, \beta_s^k) \tag{6}$$

where β_c^k and β_s^k are the corresponding content target and style target, which are the input to the subnet that outputs β_k, and are the scaled versions of the artwork β_s. We can control the types of artistic characteristics that are learned for different subnets by training them with different style scales. And intended to demonstrate that the concept may simply be expanded further layers.

Since such stylization losses are calculated using the outputs of many layers over the whole network. A total loss (e.g., a weighted mixture of all stylization losses) cannot be used to propagate and update the weights backwards directly.

Consequently, a parallel criterion is employed to back-propagate the weights for distinct ranges of layers using varying stylization losses. The hierarchical stylization loss function L_H, which is a weighted sum of such losses, is defined as follows:

$$L_H = \sum_{k=1}^{K} \mu_k L_S^k(\hat{\beta}_k, \beta_c^k, \beta_s^k) \tag{7}$$

where the weight of stylization loss L_S^k is represented by ω_k.

Therefore, during the end-to-end learning on natural images $x \sim X$, each subnet denoted by Ψ_k is trained to minimise the parallel weighted stylisation losses that are computed from the latter outputs $\hat{\beta}_i$ (i $\geq$ k) (latter means it comes later in the feed-forward direction) as in

As a result, throughout the learning on natural images $x \sim X$, Ψ_k represents every subnet is trained to minimise parallel weighted stylisation losses computed from the latter outputs $\hat{\beta}_i$ (i $\geq$ k)(later implies it occurs later in the feed-forward direction), as in

$$\beta_k = arg \min_{\beta_k} E_{x\sim X}\left[\sum_{i\geq k}^{K} \omega_i L_S^k f(\cup_{j=1}^{k}\beta_j, x), \beta_c^i, \beta_s^i\right] \tag{8}$$

In proceeding, suppose f_{-1} is representing the general back-propagation function, and the weight updates (gradients) of the subnet β_k can be written as for each iteration.

$$\Delta\Psi_k = \begin{cases} f^{-1}(\mu_k L_S^k) & k = K \\ f^{-1}(\mu_k L_S^k, \Delta\Psi_{k+1}) & 1 \leq k \leq K \end{cases} \tag{9}$$

As a result, both the stylization loss at the current level L_S^k and the gradients of the latter subnets influence the weights of the current subnet β_k.

Although all of those subnets are built for various reasons, Eq. 8 showed that they are not fully independent. Earlier subnets also help to reduce the latter's losses. Thus, a shallower CNN structure can be used for further blocks(subnets), saving both runtime and compute memory.

FIGURE 4. Comparison between different models. One style image and different content images are taken for better visualization and understanding. Every model's produced stylized images.

IV. EXPERIMENTAL RESULTS

For the experiment, the coco dataset an open-source image dataset has been used. This coco dataset consists of more than 80,000 RGB images with sizes greater than 256 × 256. All the images are real and artistic. While performing the experiment, the content and stylized images should have the same size for comparison. However, in the dataset, every image has a larger size than 256 × 256, but the problem is that every image has a different size. Different sizes for training the model could train. However, in testing, the problem will arise while interpreting the qualitative and quantitative comparison. To handle this problem, we changed the size of every image to 256 × 256, 512 × 512, and 1024 × 1024 with the python script. The size change of every image because the NST-based method produced images in generic sizes like 256×256, 512×512, and 1024×1024. This conversion helps to compare the qualitative analysis of the images. The experiment is performed using the tools like TensorFlow-GPU 1.15, CUDA 10.2, and Python 3.7.6 on NVIDIA GeForce GTX 166Ti GPU. The methods in [2], [5], [6], [7], [15], [16], [20], [24], and [3] have been implemented to conduct the analysis of the existing methods of NST by using COCO dataset.

Further, the PSNR (Peak Signal to noise ratio) has been used as the qualitative evaluation metrics. This metrics compare two images for their perceptual measurements. The PSNR is calculated using Eq.(10)

$$PSNR = 10 * log_{10}(\frac{R^2}{MSE}) \tag{10}$$

where, R is the maximum possible intensity value in the image and MSE is the mean squared error.

The experiment has been conducted using different style images as presented in Table 1. However, the PSNR values have been computed between the content image (shown in Figure 4 (a)) and the stylized image (obtained by different methods in [2], [5], [6], [7], [15], [16], [20], [24], and [3] and our method by using style image, shown in Figure 2) shown in Figure 4(b)-(k). The reason behind using the same style and content images is easily comparable.

Figure 4(a) shows the content images. Each column in Figure 4(b)-(j) present stylized image obtained by methods in [2], [5], [6], [7], [15], [16], [20], [24], and [3] respectively and 4(k) is our model. In method [24], the stylized image has poor clarity and shapes, due to blending of minute details by VGG network. However, method in [7] generated visually pleasing stylized images. Methods [2], [5], [6] have generated

TABLE 1. Style images: Done by famous artists, available in public domain.

SN	Name and Year	Author
1	Divan Japonais, 1893	Henri de Toulouse-Lautrec
2	Edith with Striped Dress (1915)	Egon Schiele
3	Head of a Clown (1907)	Georges Rouault
4	Landscape at Saint-Remy (1889)	Vincent van Gogh
5	Portrait of Pablo Picasso (1912)	Juan Gris
6	Ritmo plastico del 14 luglio (1913)	Severini Gino
7	The Tor of Babel (1563)	Pieter Bruegel the Elder
8	Three Fishing Boats (1886)	Claude Monet
9	Trees in a Lane (1847)	John Ruskin
10	White Zig Zags (1922)	Wassily Kandinsky

TABLE 2. Comparison of PSNR between different models.

SN	Method Name	PSNR
1	Gatys Style [2]	9.2368
2	Johnson style [5]	9.0401
3	Ulyanov style [6]	9.2769
4	Huang style [20]	9.2104
5	Li and Wand style [7]	8.8874
6	Li Diverse style [15]	9.0512
7	Zhang style [16]	9.4303
8	Li Universal style [24]	9.5635
9	Yingxu Style [3]	9.5667
10	**Our method**	**13.0638**

TABLE 3. Timings of different models at different dimensions.

Methods	Time(s)		
	256 x 256	512 x 512	1024 x 1024
Gatys et al. [2]	14.32	51.19	200.3
Li and Wand [7]	0.015	0.055	0.229
Johnson et al. [5]	0.014	0.045	0.166
Ulyanov et al. [6]	0.022	0.047	0.145
Zhang and Dana [16]	0.019	0.059	0.230
Li et al. [15]	0.017	0.064	0.254
Li et al. [24]	0.620	1.139	2.947
Yingxu et al. [3]	0.060(avg.)	0.1(avg)	0.256(avg)
Our Model	0.054	0.107	0.201

much better stylized images due to texture analysis. Methods in [15] and [16] have obtained dark stylized images. Further, method in [7] generated over-saturated stylized images, as shown in Figure 4(i). The method in [3] can generate balanced and visually better-stylized images, compared to other methods due to perceptual-aware distillation and pixel-aware distillation. Still, it is lacking with luminance. Our proposed architecture segregates RGB and luminance blocks, which brings a hike in PSNR value. This comparison shows that the methods based on perceptual-aware distillation, pixel-aware distillation, and edge-filter-based multimodel architecture have produced pleasing results.

Moreover, Table 3 presents different existing methods based on perceptual loss, adversarial loss, multi-scale gram loss, and pixel-aware loss. These methods have been proved to obtain promising results. In these, the perception aware loss and pixel-aware loss based methods are computationally fast compared to other methods, as presented in Table 3.

The computation time (in seconds) of methods in [2], [5], [6], [7], [15], [16], [20], [24], and [3] have been presented in Table 3 using images with sizes $256 * 256$, $512 * 512$ and $1024 * 1024$. The computation time of the method in [2] is very high due to the use of VGG network. Methods in [7] and [16] support statistics of storing encoded style, which speeds up the stylization process. The computation time of [16] is very high and the GPU runs out of memory. The computation time of [5], [7], and [15] is similar due to the utilization of same architecture by both of these methods. This comparison indicates that perceptual loss and edge filter based methods are computationally fast and produce visually promising results.

V. CHALLENGES AND ISSUES

The advancement in NST is remarkable, and many of the algorithms are actively in use in the industry. As evidenced by testing findings, images hold numerous items at the varied intensity or prominent items with distinct structures. Still, performance can be improved. In both NPR and NST, aesthetic evaluation is a critical challenge. This paper mentioned key challenges in this NST industry and explores various ways of dealing with them in future research in this part. Because NST and NPR are so closely related, numerous important problems in NPR, are also future challenges for NST research. Despite the fact that some existing methods still could improve efficiency, resolving image quality, and optimization issues in NST. There are still many issues and problems to be resolved. For a flawless result so far, removing necessary semantic labels should be accurate, enormous loss, and still many to be figured out. The method provides amazing visual effects, particularly when processing images vulnerable to structural distortion.

VI. CONCLUSION AND FUTURE SCOPE

Various applications have used style transfer to give the end-user a new dimension in recent years. DL-based strategies have been found in the literature to help resolve various issues with new NST approaches in Section II. This research explores various NST methodologies for resolving image quality, enormous loss, and optimisation issues in NST. From section III to IV this paper presented an edge filter-based hierarchical training system (multimodal transfer) for learning artistic style cues at several scales, fine texture structure scale disparity issue, comprising colour, and generating significantly more visually appealing stylised results. This paper have presented different NST-based methods [2], [3], [5], [6], [7], [15], [16], [20], [24] and these methods have been evaluated with qualitative and quantitative parameters. The Edge-filter based proposed method capable of generating well stylized image, as can be seen in Figure 4. The qualitative parameter's PSNR value for our proposed method is higher compared to other methods, as discussed Section IV.

Moreover, optimising loss of various models and techniques will be explored. The plan to design a method has been discussed. Despite the fact that this is a continuation of our earlier study [22], there are still some intriguing difficulties to investigate further. For example, the auto-encoder could use semantic segmentation to provide additional supervision during region breakdown, resulting in more spectacular region-specific transfer. Incorporating the proposed approach into video and stereoscopic applications is also exciting. Future work can also be derived to propose more advanced deep neural architecture and a tighter bound for style loss and improvement.

REFERENCES

[1] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 11, pp. 3365–3385, Nov. 2020.

[2] A. L. Gatys, S. A. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[3] Y. Qiao, J. Cui, F. Huang, H. Liu, C. Bao, and X. Li, "Efficient style-corpus constrained learning for photorealistic style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 3154–3166, 2021.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 694–711.

[6] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proc. ICML*, vol. 1, 2016, p. 4.

[7] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.
[8] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.
[9] C. Peng, N. Wang, J. Li, and X. Gao, "Universal face photo-sketch style transfer via multiview domain translation," *IEEE Trans. Image Process.*, vol. 29, pp. 8519–8534, 2020.
[10] Y. Zhang, Y. Zhang, and W. Cai, "A unified framework for generalizable style transfer: Style and content separation," *IEEE Trans. Image Process.*, vol. 29, pp. 4085–4098, 2020.
[11] A. Zhu, X. Lu, X. Bai, S. Uchida, B. K. Iwana, and S. Xiong, "Few-shot text style transfer via deep feature similarity," *IEEE Trans. Image Process.*, vol. 29, pp. 6932–6946, 2020.
[12] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Explicit filterbank learning for neural image style transfer and image processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2373–2387, Jul. 2021.
[13] P. Chandran, G. Zoss, P. Gotardo, M. Gross, and D. Bradley, "Adaptive convolutions for structure-aware style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7972–7981.
[14] K. Xu, L. Wen, G. Li, H. Qi, L. Bo, and Q. Huang, "Learning self-supervised space-time CNN for fast video style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 2501–2512, 2021.
[15] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Diversified texture synthesis with feed-forward networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3920–3928.
[16] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–16.
[17] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9036–9045.
[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[19] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6924–6932.
[20] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1501–1510.
[21] W. Wang, S. Yang, J. Xu, and J. Liu, "Consistent video style transfer via relaxation and regularization," *IEEE Trans. Image Process.*, vol. 29, pp. 9125–9139, 2020.
[22] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2017, pp. 1897–1906.
[23] J. E. Kyprianidis, J. Collomosse, T. Wang, and T. Isenberg, "State of the 'art': A taxonomy of artistic stylization techniques for images and video," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 5, pp. 866–885, Jul. 2013.
[24] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[25] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4990–4998.
[26] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 909–920, 2020.
[27] R. Li, C.-H. Wu, S. Liu, J. Wang, G. Wang, G. Liu, and B. Zeng, "SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 374–385, 2021.
[28] J. Ling, H. Xue, L. Song, R. Xie, and X. Gu, "Region-aware adaptive instance normalization for image harmonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9361–9370.
[29] J. Cheng, A. Jaiswal, Y. Wu, P. Natarajan, and P. Natarajan, "Style-aware normalized loss for improving arbitrary style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 134–143.
[30] C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, "GAN-based day-to-night image style transfer for nighttime vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 951–963, Feb. 2021.
[31] J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, "Neural style palette: A multimodal and interactive style transfer from a single style image," *IEEE Trans. Multimedia*, vol. 23, pp. 2245–2258, 2021.
[32] Q. Xiao, B. Liu, Z. Li, W. Ni, Z. Yang, and L. Li, "Progressive data augmentation method for remote sensing ship image classification based on imaging simulation system and neural style transfer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9176–9186, 2021.
[33] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5239–5247.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, vol. 1, 2015, pp. 262–270.

SHUBHAM BAGWARI received the B.Tech. degree in computer science engineering from Graphic Era Hill University, Dehradun, India, and the M.E. degree in software engineering from the Thapar Institute of Engineering and Technology (Deemed to be University), Patiala, India. He is currently working as a Project Associate with TIH-iHub Drishti Foundation, IIT Jodhpur, India. His current research interests include data analysis, deep learning, and image processing.

KANIKA CHOUDHARY received the B.Tech. degree in computer science engineering from Chitkara University, Punjab, India. She is currently pursuing the M.E. degree in computer science engineering with the Thapar Institute of Engineering and Technology (Deemed to be University), Patiala, India. She is currently working as a Research Scholar with the Thapar Institute of Engineering and Technology (Deemed to be University). Her current research interests include data analysis and visualization, deep learning, and image processing.

SURESH RAIKWAR received the Ph.D. degree in image processing from the ABV-Indian Institute of Information Technology and Management, Gwalior, Madhya Pradesh, India, in 2019. He is currently working as an Assistant Professor with the Thapar Institute of Engineering and Technology, Patiala, Punjab, India. He is the Developer of a device Dradh Drishti Prabardhak (DDP) to assist Loco-pilots during poor weather conditions, recognized by Niti-Aayog, Government of India, under Atal-Innovation Mission. He has published papers in various reputed journal, transactions and international conferences. His main research interests include computer vision, image processing, pattern recognition, machine learning, and generative adversial networks. He is an Active Reviewer of highly reputed journals in the field of image processing such as IEEE TRANSACTIONS ON IMAGE PROCESSING, *IET-Image Processing*, ELECTRONICS LETTERS. He has chaired and co-chaired couple of international conferences.

RAHUL NIJHAWAN received the M.Tech. and Ph.D. degrees from IIT Roorkee with several prestigious achievements. Some of them include: Received young scientist award, received the Best Ph.D. Thesis Award, published more than 25 research articles in very reputed international journals/conferences. He has guided several research projects. He is an expert in the field of machine learning and computer vision. Gate percentile: 99.96% in computer science. He offers of a Postdoctoral Fellowship from USA, Hong Kong, and France.

SUNIL KUMAR received the B.Tech. degree in computer science from Kurukshetra University, Kurukshetra, in 2006, the M.Tech. degree in computer science from MMU, Ambala, in 2011, and the Ph.D. degree from the University of Petroleum and Energy Studies, Dehradun, India, 2021. He is currently an Assistant Professor (Selection Grade) with the Cybernetics Cluster, School of Computer Science, University of Petroleum and Energy Studies (UPES). He is a Founder Member of the OPEN Community under the University of Petroleum and Energy Studies. He has more than 16 years of teaching experience at reputed NAAC accredited universities such as the University of Petroleum and Energy Studies, Dehradun (NAAC accreditation A). He has published more than 20 research articles (including SCOPUS and SCI indexed publications), edited/authored various books/book chapters, and served as a reviewer for various journals and conferences. His area of research is WSN, deep learning, the IoT, metaheuristic optimization, and data mining.

MOHD ASIF SHAH received the B.A., M.A., and Ph.D. degrees with sound teaching and research skills.

He is currently working as an Associate Professor with Bakhtar University (IACBE Accredited), Kabul, Afghanistan. He has been earlier working as an Assistant Professor at FBS Business School, Bengaluru, Karnataka, India, and Lovely Professional University, Punjab, India (AACSB Accredited). He has also served as a Lecturer at the Jamia College of Education and also helped his department with teaching assistance during the Ph.D. He has published more than forty research papers (SCI/WOS/UGC indexed) with 32 citations, the H-index is four Google Scholar. He has attended more than thirty workshops and faculty development programs sponsored by the Government of India, and other agencies. Else than this, he has an excellent grasp of the subject material, with more than five years of using platforms like CANVAS, LMS, and UMS for online teaching.