IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 August 2022, accepted 11 September 2022, date of publication 15 September 2022,
date of current version 22 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206771

APPLIED RESEARCH

Text-Guided Sketch-to-Photo Image Synthesis

UCHE OSAHOR, (Student Member, IEEE), AND NASSER M. NASRABADI, (Fellow, IEEE)
Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
Corresponding author: Uche Osahor (uo0002@mix.wvu.edu)

ABSTRACT We propose a text-guided sketch-to-image synthesis model that semantically mixes style and content features from the latent space of an inverted Generative Adversarial Network (GAN). Our goal is to synthesize plausible images from human facial sketches and their respective text descriptions. In our approach, we adapted a generative model termed Contextual GAN (CT-GAN) that efficiently encodes visual-linguistic semantic features pre-trained on over 400 million text-image pairs at different resolutions along the model. Also, we introduced an intermediate mapping network called c-Map that combines textual and visual-based features to a disentangled latent space $\mathcal{W}+$ for better feature matching. Furthermore to maximise the computational performance of our model, we implemented a linear-based attention scheme along the pipeline of our model to eliminate the drawbacks of inefficient attention modules that are quadratic in complexity. Finally, the hierarchical setting of our model ensures that textual, style and content features are synthesised based on their unique fine grained details, which result in visually appealing images.

INDEX TERMS Contextual GAN (CT-GAN), generative adversarial network (GAN), text-guided sketch-to-image synthesis.

I. INTRODUCTION

A text-guided model that translates sketches to images will do great justice in easing the difficulty in generating images from sketches that may find applications in identity recognition, attribute assignment, text-guided synthesis, etc. Prior work [1], [2], [3] has shown exceptional performance in cases when colored images are synthesized from data which contain enough content and style that could control a model towards their respective real image. However, sketches can be seen as images that contain minimal information bounded by pixels that could be translated into photo-realistic images by a suitable generative model. Sketches might contain key structural information that could aid in providing visual meaning, which is crucial in classifying images as valuable or not. However, sketches do not portray any style information regardless of the mode from which they are crafted, and as a result, it becomes pretty difficult to translate sketches to perceptually appealing images. While significant work on image synthesis is still ongoing due to its numerous applications, It is still challenging to synthesize natural-looking images from sketches or labels. Currently, recent methods of image synthesis could be fashioned as a form of text-to-image, image-to-image or sketch-to-image synthesis, or a combination of all three techniques.

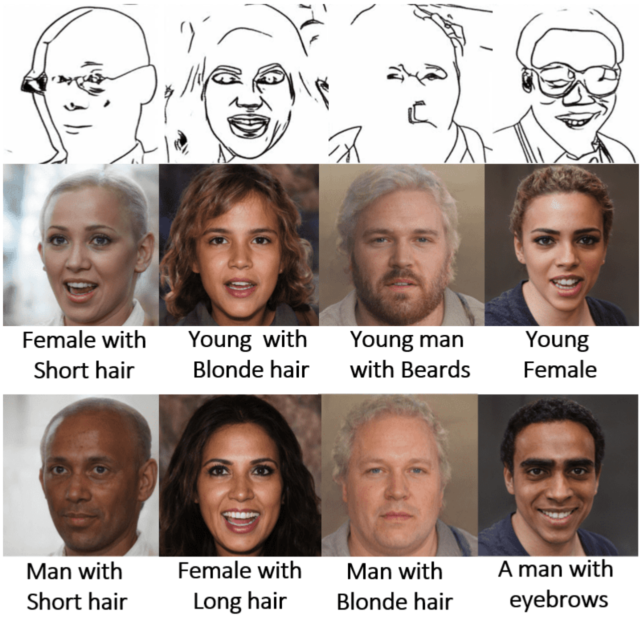

FIGURE 1. Synthesised images from sketches.

The ability to synthesize high quality images is the core goal for most generative adversarial models, and these

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

models [4] are typically designed in a multi-stage fashion [2], [5], where intermediate layers are guided adversarially with discriminator modules or more sophisticated techniques that employ progressive training methods. A multi-stage training process is time consuming and cumbersome, making it infeasible for synthesizing high-quality images with very high resolution. Most importantly, images contain lots of facial details, such as freckles, skin pores, dimples which cannot be obtained by merely upsampling from the lower-resolutions. Latest developments in generative adversarial networks (GANs) have presented an entirely different image generation paradigm that achieves exceptional quality called the StyleGAN [6], [7]. The generator architecture of the StyleGAN is crafted such that it reveals novel ways to control the image synthesis process. The generator starts from a learned constant input which adjusts the style at each convolutional layer based on its dedicated latent code, therefore directly controlling the strength of image features at different scales. Noise is also injected directly into the network which leads to the automatic unsupervised separation of high-level attributes. The StyleGAN approach constructs an intermediate latent space $\mathcal{W}$ that is linear and has a less entangled representation of different factors of variation [7].

In a bid to interact with the feature-rich disentangled latent space of the StyleGAN model, researchers developed a GAN inversion technique [8], [9], [10] which inverts real images into StyleGAN's latent space $\mathcal{W}$, where meaningful manipulation afterward can take place. Such inversion is achieved by training an encoder to map real images into the $\mathcal{W}$ space, which leads to image reconstruction and semantically meaningful image editing. The hierarchical semantic property of the $\mathcal{W}$ latent space inspired the design of numerous cross-modal methods [10], [11], [12], for visual content creation, pretrained on StyleGAN generators.

In this paper, we propose an effective application-based approach for text-guided sketch-to-image synthesis called Contextual-GAN (CT-GAN). Our model aims to imitate a visual-linguistic latent space that efficiently interacts with the latent embeddings of the StyleGAN model, which is aimed at synthesizing highly-appealing images from sketches that are guided by textual descriptions as shown in Figure 1. We capitalize on the power of a Contrastive language-image pre-trained model termed; CLIP [13] trained on over 400 million text-image pairs and a bi-directional encoder classifier model called BERT [14] to learn the unique attributes within each word and their contextual associations. We expand more on the benefits of the CLIP and BERT models in section 3.

Our approach aims to achieve three main goals. The first aim as shown in Figure 2, combines the features of our GAN inversion encoder trained for sketch-to-image synthesis with semantic features extracted from a pre-trained CLIP model for every sketch and text pair. This approach is vital because the extracted frozen weights from CLIP added at different layers, aids faster training convergence, especially for multi-modal training scenarios that combine different input types. We also compute a CLIP loss between each sketch and text description, respectively. Our approach is similar to models that adapt some form of perceptual similarity for standard image-to-image generative models aimed at boosting perceptual quality. Secondly, we introduced the BERT model to properly identify and update the contextual associations between newly derived words defined by the user and the pre-trained text descriptions from the CLIP model. Finally, we deviate from previous techniques that apply a less efficient attention scheme and we rather adapt a novel linear-based attention model that computationally ensures better feature interactions at the semantic level. To elaborate on our method further, we trained an encoder capable of obtaining latent codes that align with the hierarchically semantic arrangement of a pre-trained StyleGAN model, which is inspired originally by [8]. We break down our methods into three core contributions.

- We present a visual-linguistic inversion module structured in a hierarchical fashion, where the inverted code of a given sketch image and learned text descriptions can be found in the $\mathcal{W}$ latent space of a StyleGAN generator.
- Secondly, to aid application-specific implementations, we incorporated a text-based encoder module to update the contextual associations between newly derived text descriptions and the pretrained features of the CLIP model.
- Finally, we eliminated the quadratic nature of previous attention models by adapting a more efficient linear-based attention scheme which effectively permits long sequence interactions on large inputs where contextual information is the key to achieve better feature disentanglement.

II. RELATED WORK

Our work is closely related to the literature of text-guided image-to-image translation, with emphasis on sketch-to-image synthesis which is still a challenging task. Our model is achieved by implementing an improved version of a visual-linguistic GAN inversion model with the additional benefits of a linear-based attention scheme.

A. TEXT-GUIDED IMAGE MANIPULATION

Text descriptions can be used to guide image synthesis, a well structured sentence with significant phrases can be encoded as unique attributes that can in-paint images. For example, Li *et al.* [11] came up with a multi-stage network with a novel text-image combination module to produce high-quality results. Nam *et al.* [15] disentangled different visual attributes by implementing a text-adaptive discriminator, used to provide better fine-grained feature feedback to the generator. Li *et al.* proposed StoryGAN [16], which generates a series of images that are contextually consistent with previously generated images and with the sequence of text descriptions provided by the user. Dong *et al.* [17] proposed an auto-encoder architecture to modify an image according to a given text. Liu *et al.* [18] proposed a multi-modal method that models the visual attributes of an image and learns how to translate them through automatically generated commands.

Qiao *et al.* [19] focused on semantic consistency by enforcing the synthesised images to have the same semantics with the input text description. However, most of the text-based image manipulation methods with significant performance are basically based on the multi-stage framework. Deviating from previous methods, we propose a novel framework that achieves image generation and manipulation without multi-stage processing.

B. GAN INVERSION

It is arguable to infer that a generative model is as good as its latent space. Most importantly, a properly disentangled latent space will be more useful for multi-modal synthesis. GAN inversion was first introduced by Zhu *et al.* [20], and in their approach, the latent space is trained on various image outputs from which a pre-trained GAN most accurately reconstruct a known image. Motivated by their method, recent works have used StyleGAN [7] for this task as well. Generally, inversion methods either directly optimize the latent vector to minimize the error for the given image [21], [22], [23], train an encoder model to map images to the latent space [24], [25], or use a hybrid approach by combining both aforementioned methods [8]. Typically, techniques performing optimization are superior in reconstruction quality to a learned encoder mapping, but require a longer training time.

Our encoder can efficiently embed a given face image into the extended StyleGAN latent space that comprises of both text and sketch images, which we represent as $\mathcal{W}_+$. Some concurrent works improve GAN inversion with better reconstruction like Gu *et al.* [26] who employs multiple latent codes to recover an image, Pan *et al.* [27] optimizes the parameters of the generator together with the latent code, while Karras *et al.* [7] and Abdal *et al.* [28] focused on inverting StyleGAN models by exploiting the layer-wise noises. One important issue omitted by most inversion methods is that they merely consider reconstructing the target image at the pixel level without considering the computational efficiency. Therefore, in this work, we argue that the dependence on the pixel-wise reconstruction loss as the major metric to evaluate a GAN inversion method is not necessarily efficient or the best approach. Instead, we studied the properties of the inverted code obtained at the semantic level and proposed a richer visual-linguistic module coupled with a linear-based attention scheme.

C. ATTENTION TECHNIQUES

The inclusion of attention modules to a network encourages it to focus on specific aspects of an input by weighting the important parts more than irrelevant parts. Hence, attention plays a major impact on improving language and vision applications [29], [30], [31]. A very popular method was introduced by AttnGAN [32] which builds upon StackGAN++ [2] and incorporates attention into a multi-stage feature-refinement pipeline. Their mechanism allows the network to synthesize fine-grained details based on relevant words in addition to the global sentence vector. SEGAN [33] proposed an attention competition module to focus only on the key-words instead of designing an attention weight for each word in the sentence. They achieved this by introducing an attention regularization term [34], [35] that only keeps the attention weights for visually important words. ControlGAN [36] achieved both text-to-image generation and visual attribute manipulation such as category, texture, and color by changing the description without affecting other content. They proposed a word-level spatial and channel-wise attention driven generator which allows the generator to synthesize image regions corresponding to the most relevant words. They also showed how a word-level discriminator can provide the generator with fine grained training signals that disentangles different visual attributes by exploiting the correlation between words and image sub-regions. Indeed attention mechanisms are beneficial to GAN models, but the quadratic computational and memory complexities of most attention mechanisms have limited their scalability for modeling long sequences. In this paper, we propose a preferable linear-based attention mechanism that approximates softmax attention which yields only linear time and space complexity as opposed to quadratic solutions. As compared to traditional attention mechanisms, our method performs the attention operation linearly, while also storing adequate contextual information.

III. MODEL ARCHITECTURE

In our work, we ensure that the derived latent space of the image and text pairs are properly disentangled in order to encourage the augmentation of different facial attributes assigned to each subject. To build such a visually-linguistic model, we combined the efficiency of different semantic rich models that compete with state-of-the-art. Firstly, we adapted a hierarchical structure of visual-linguistic features derived from a sketch-based image encoder that is trained inversely [8] to a common latent space we define as $\mathcal{W}_+$ for both text and image features. We also introduced a computational and semantic efficient attention model that is linear-based which maintains the semantic similarity between text and image pairs. We chose the Multi-Modal CelebA-HQ [37] dataset comprising over 60,000 unique identities and 40 facial attributes as a basis for training our model. The sketches used for our model are curated from a collection of human drawn sketches of unique individuals from different ethnic backgrounds.

A. VISUAL-LINGUISTIC MODEL

In our work, we propose a hierarchy of encoded visual-linguistic features that computes the contextual similarity between images and text pairs extracted at different down-sampled resolutions. The contextual similarity between each pair of text and image is computed for different layers of the model, as shown in Figure 2. Since our design direction is focused on a multi-modal problem where text is used to guide sketch-to-image synthesis, we incorporated a set of linear-based attention models that ensure the local contextual similarity between image regions and specific words are

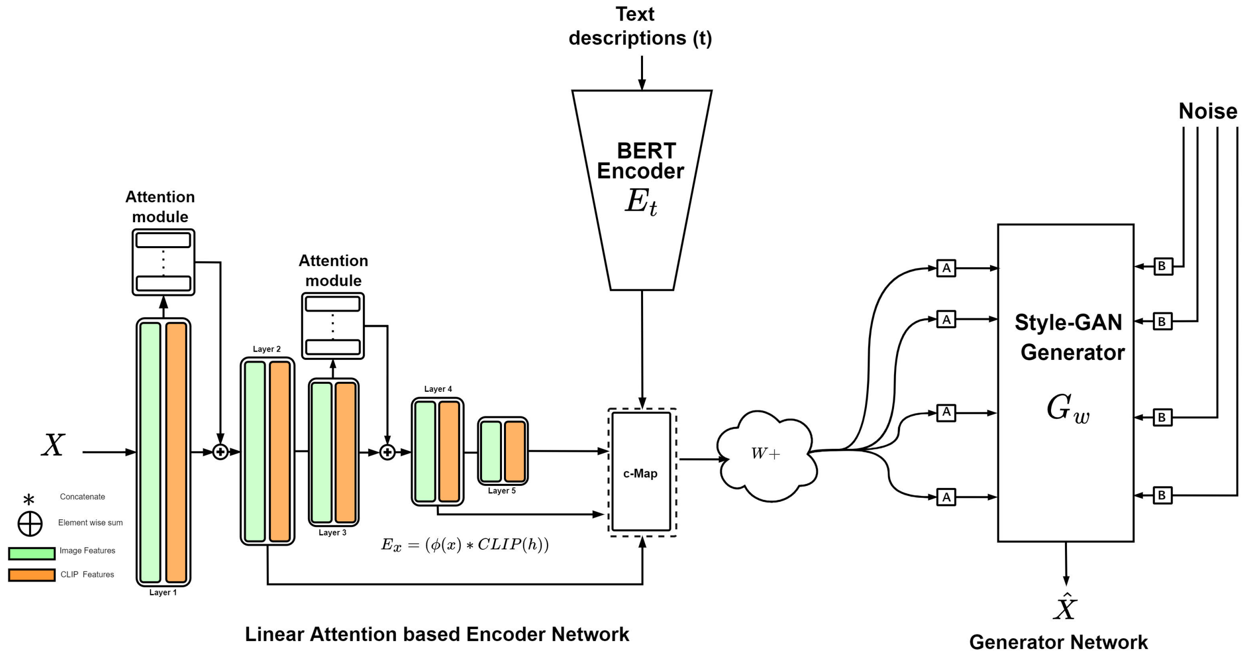

FIGURE 2. The GAN architecture. The structure shows the linear attention-based encoder coupled with a style-GAN generator. Our proposed model encodes a hierarchy of visual features derived from the input sketch image at different resolutions to match the defined style layers of the Style-GAN architecture. Visual features from the visual encoder $\phi(x)$ in green are concatenated with visual features (orange), which are extracted from the pre-trained CLIP vision encoder at different layers represented as $\mathbb{E}_{x_i}(\phi(x) * CLIP(h))$ per layer. A contextual intermediate network (c-Map) maps combined textual and visual features to the disentangled latent space $\mathcal{W}_+$ of the Style-GAN model to synthesise a high resolution image.

learned. In Figure 3a, we break down the linearised attention model to highlight its true benefits. In a structural sense as described in Figure 2, we extracted encoder embeddings into the latent StyleGAN latent space $\mathcal{W}_+$. The embeddings are scaled to a single vector of 512-dimensional vectors. In addition, our technique can easily be plugged into any generative adversarial model without heavily impacting the overall computational complexity.

B. GAN INVERSION MODULE

To build an efficient GAN model that leverages on visual-linguistic features that synthesize plausible images, we must encode a defined set of input pairs to a latent space that semantically maps style and content efficiently. Such a latent space is achieved by inverting the features of an input image to a latent space of a fixed GAN model. To implement such a framework, we adapted the StyleGAN model because it offers the best of image quality at high resolutions and a wider diversity that covers the full spectrum of facial attributes [38]. GAN inversion basically means the reverse mapping of a given image x into a latent space of an already trained GAN model. However, plugging in a new visual or text encoder into an already trained StyleGAN generator comes with a unique set of issues. Firstly, the StyleGAN model is broken down into different sections classified by three main semantic levels of detail, namely: coarse, medium and fine layers [7], [9]. Secondly, the StyleGAN model encodes a style distribution in a hierarchical setting where subsequent semantic affine [9], [39] layers (18 layers in this case) are grouped based on their unique fine grained details [7]. To cater to these two requirements of the StyleGAN model for semantic style interactions, we propose a novel encoder model. Our proposed encoder, similar to the methods applied in [38], [40] encodes a hierarchy of features derived from the input image at different resolutions to match the defined style layers of the StyleGAN architecture. In section 3.5, we describe a contextual intermediate mapping network; **c-Map** that combines textual and visual features from an encoder model $E_x(\cdot)$ to the latent space $\mathcal{W}$ of the StyleGAN model. To further encourage better semantic mixing, we initialize the new latent space $\mathcal{W}_+$ with features from the StyleGAN model. Hence, our encoder is trained to learn visual-linguistic information using an efficient attention scheme. The final semantic output is represented as:

$$E_x(x, h) = (\phi(x) * CLIP(h)) + \overline{w}, \tag{1}$$

where $\phi(x)$, h and $\overline{w}$ denotes the input image embeddings, features from the CLIP model, and the average of StyleGAN latent embeddings, respectively. We use "$*$" to represent the concatenation operation. The training process is given as:

$$\begin{aligned}\mathcal{L}_{E_w} = &\ \|y - G_w(x, h)\|_2^2 + \lambda_1 \|V(y) - V(G_w(x, h))\| \\ &+ \lambda_2 \|\phi(y) - \phi(G_w(x, h))\| - \lambda_3 \mathbb{E}[D_v(G_w(x, h))],\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{D_w} = &\ \lambda_1 \mathbb{E}[D_v(G_w(x, h))] - \lambda_2 \mathbb{E}[D_v(y)] \\ &+ \frac{\lambda_3}{2} \mathbb{E}\|\nabla_x D_v(G_w(x, h))\|_2^2,\end{aligned} \tag{3}$$

We use $D(\cdot)$, $V(\cdot)$ and $\phi(\cdot)$ to represent discriminator, perceptual and style loss, while λ represents adjustable hyper-parameters, more detail on losses is described in section 4. y represents the reference image, we also represent the synthesized output image $\hat{X}$ from the generator as $G_w(x, h)$, to identify the features from the encoder that is made up of the input sketch $\mathbf{x}$, sketch image features $\mathbf{h}$ from CLIP's vision encoder and the combined text and image features from the embedded space $\mathbf{w}$.

C. VISUAL-LINGUISTIC SIMILARITY

A contextual feature space of sentences and image pairs collectively initiates the bases for a potential text-guided image synthesis model. Such visual-linguistic embeddings aim to capture the intrinsic similarity between image and textual features. In our approach, we capitalize on the similarity between texts and images features to which we translate into a latent space $\mathcal{W}_+$ of the StyleGAN model. Our strategy is aimed at improving the visual perceptual appeal of synthesized images for difficult visual inputs like sketches with textual description. To achieve state-of-the-art performance, we utilized a visual-linguistic encoder trained on over 400 million image-text pairs, called Contrastive Language-Image Pre-training (CLIP) [13], [41]. A CLIP model is trained to predict possible (image, text) pairings. To improve performance, CLIP maximizes the cosine similarity of the image and text embeddings of the N real (image, text) pairs in the batch while minimizing the cosine similarity $cos(x, t)$ of the embeddings of the $N^2 - N$ incorrect pairings. In Figure 6 and 7, we provide visually appealing results and compelling quantitative results. For image synthesis, using a pretrained generator and a pre-trained text encoder for the two-fold goal, we define the optimization problem as:

$$w^* = arg \min_{w} \|w - E_x(G_w(x, h))\|_2^2 , \quad (4)$$

we obtain an inverted latent code w as the initialization for the optimization, using the inversion module [38]. $G_w(x, h)$ represents latent embeddings w which is a combination of the image x and CLIP pre-trained embeddings h.

D. TEXT ENCODER

Every facial sketch is associated with a textual descriptions generated from a corpus of texts which are based on unique attributes of the Multi-Modal-CelebA-HQ [38], [42] dataset. To build such a text-guided model, we used BERT to learn the unique attributes within each word and their contextual associations. BERT relies on a transformer-attention mechanism that learns the contextual relationships between words in a text. The input to the encoder for BERT is a sequence of tokens (words), which are first converted into vectors and then processed into the neural network. Semantic vectors indicated as $t \in D^{D \times T}$ from text descriptions are extracted [14] where all of the sub-words in an input sentence are mapped to a set of embeddings, E_t. Each embedding $t \in E_t$ is computed as the sum of a token embedding, specific to the sub-words. The input embeddings E_t are then passed through a multi-layer Transformer network that builds a contextualized representation of the sub-words. Pre-training is done using a combination of two language modeling objectives: Firstly, a masked language modeling where some parts of the input tokens are randomly replaced with a special token (i.e., [MASK]), and the model needs to predict the identity of those tokens. Secondly, a sentence prediction where the model is given a sentence pair and trained to classify whether they are two consecutive sentences from a document. Finally, an output layer and objective are introduced and fine-tuned on the task data from pre-trained parameters [14], [32]. We compute the similarity between these features; text w_t and image w_v embeddings from the latent space. Hence, the multi-modal similarity is learned by the given expression:

$$\mathcal{L}_{E_t} = \sum_{i=1}^{L} \|w_v - w_t\|_2^2. \quad (5)$$

where $w_v, w_t \in \mathcal{R}^{L \times C}$, are the features obtained from the image and text embeddings; $w_v = E_x(x, h)$ and $w_t = E_t(\cdot)$. All features are of the same shape for L layers, each with a c-dimensional latent embedding.

E. CONTEXTUAL MAPPING

A mapping scheme is necessary to transfer the embeddings from both image and text pairs. In our case, we implemented a contextual mapper (**c-Map**) that translates contextual features into a latent embedding; $f : \mathcal{Z} \longrightarrow \mathcal{W}$, of semantic vectors derived from abstracted data. We also rescaled the images and then extracted the local feature matrix from the last layer of the image encoder $E_x(\cdot)$, which we feed to the **c-Map** network. The mapping network mimics a small convolutional network, which gradually reduces the spatial size using a set of 2-strided convolutions followed by LeakyReLU activations [9], [43] and a series of fully connected layers to ensure the model is aware of the inherent information between text and images, which is crucial for feature disentanglement as shown in Figure 3. A fully connected layer learns features from all the combinations of the features of the previous layer, while a convolutional layer relies on local spatial coherence with a small receptive field (3 × 3 kernel, in most cases). We then added the fully connected layers to address the problem of shared weights in conventional architectures, which prevents the convolutional layers from generating subtle variations in different spatial zones which are needed to produce realistic images.

Furthermore, our approach stands out from previous techniques because in the design of our mapping model, we payed attention to the shortcomings of current contextual encoders that operate at a pixel-to-pixel correspondence [9], [10]. An encoder that operates at a pixel-to-pixel correspondence will be subject to locality bias, which is a major limitation when handling non-local transformations [44]. In our approach, we implemented a model that operates at a global level where multi-modal synthesis is easier to achieve. Since StyleGAN provides a layer wise representation, our mapping framework can sample style vectors defined by $\mathcal{W} \in \mathcal{R}^{512}$ which makes hierarchical style mixing efficient.

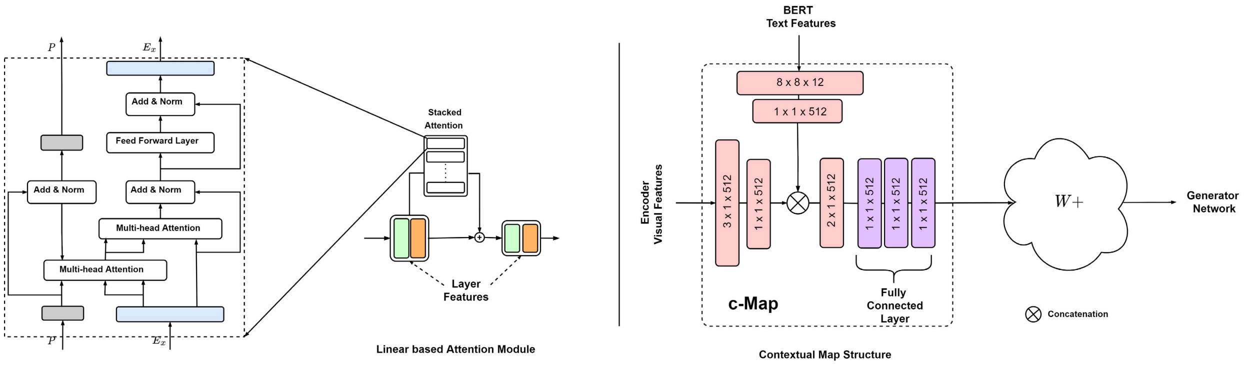

FIGURE 3. Linear-based attention module: A new sequence $P \in R^{l*d}$ with fixed (constant) length is introduced as an input which is referred to as the pack attention. A second scheme is implemented on the input sequence $E_x \in R^{l*d}$ called the unpack attention. The image to the right depicts contextual matching mechanism of our model. The concatenated features from the vision and text-based encoder, result in a visual-linguistic embedding that is fed to the generator model.

F. LINEAR ATTENTION MODULE

Despite the wide adaptation of attention modules for long-range dependency modeling, attention still has a serious drawback; It produces a quadratic solution in both time and memory cases and as a result, the memory and computational complexities of the entire attention module are quadratic. Such computational drawbacks makes the algorithm much slower than a linear-based model and effectively forbids the application of attention on large inputs or visual-linguistic applications where the contextual information is key to achieve better feature disentanglement. Therefore, to address the shortfall, we applied a novel approach by using a linear nested method similar to [45]. At the core of their method, they decoupled the regular attention function into two nested attention operations, both of which have linear efficiency. A new sequence with fixed (constant) length is introduced as an input which is referred to as the pack attention. Formally, if $P \in R^{l \times d}$ denotes the extra input sequence with fixed length l and d as the dimension. The pack attention first packs context sequence C to the output pack attention Y_P with P as the query sequence, given as:

$$Y_P = Attn(P, C). \tag{6}$$

We note that since the length of P is a constant l, the complexity of pack attention is $O(lm)$, which is linear with respect to m. To unpack the sequence back to the length of the original query sequence X, a second scheme is implemented on the input sequence X, given as:

$$Y_X = Attn(X, Y_P). \tag{7}$$

We also incorporated a position-wise feed-forward network and layer normalization, with a final derivation given as:

$$X', P' = LayerNorm((FFN(X, P, C)), P). \tag{8}$$

To match the attention conventions to the rest of our model, we represent $X' := E_x \in R^{l*d}$ and $P' := P \in R^{l*d}$ as the outputs of the feed forward $FFN(\cdot)$. The attention model is illustrated in Figure 2 and 3.

G. TARGET ATTRIBUTE SELECTION

To effectively extract the attributes of a text description, it's important to identify the keywords within the sentence that describe a subject of interest. In our case, we focused on facial features derived from sketches or images. Our goal in general was to use the text descriptions to guide style and content based features from a GAN inverted latent space [8], [9], [38]. In an image synthesis setting, most facial attributes can at least be classified into high and low level features (hair color, freckles, skin color, etc), which should be sufficient to compose a human face without heavily altering the identity of the subject [7]. In our approach, we use the StyleGAN's hierarchical generator arrangement of style and content to mix semantic fragments of both text and images features to synthesize a face [7], [38]. We extracted disentangled features from an inverted latent space $\mathcal{W}_+$ at a specified feature size. To aggregate attribute-specific style and content features from the GAN layers, we implemented a text-guided mix strategy by separating the features into w_c as the visual feature and w_s as the textual embedding. In general, the layers of the StyleGAN model can be segmented into high-level styles such as face shape, earring, eye glasses and head pose, layers in the middle control the hairstyle, hair color and facial expression. The final layers control skin color, age gender and other stochastic fine-grained details.

IV. LOSS FUNCTIONS

In this section, we give a detailed description of the different loss functions used in our model. Overall, our loss functions catered to perceptual appeal, identity preservation, visual-linguistic similarity and proper feature matching with the StyleGAN generator.

A. PERCEPTUAL LOSS

Although the GAN loss and the reconstruction loss are used to guide the generators, they fail to reconstruct perceptually appealing images. Hence, we incorporated the perceptual loss introduced in [1]. The perceptual loss function basically measures high level differences, such as content and style dissimilarity, between images. We added the perceptual loss using a pre-trained VGG-16 [46] network $V(\cdot)$. The perceptual loss calculates the L_1 distance between the features of real and encoded images from the encoder E_w. The perceptual loss $\mathcal{L}_p$ for our proposed network is defined as:

$$\mathcal{L}_p = \left\| V(\mathbf{y})^{c,w,h} - V(\hat{\mathbf{x}})^{c,w,h} \right\|, \tag{9}$$

where $V(\cdot)$ is used to denote a particular layer of the VGG-16 and c, w, and h denote the layer dimensions.

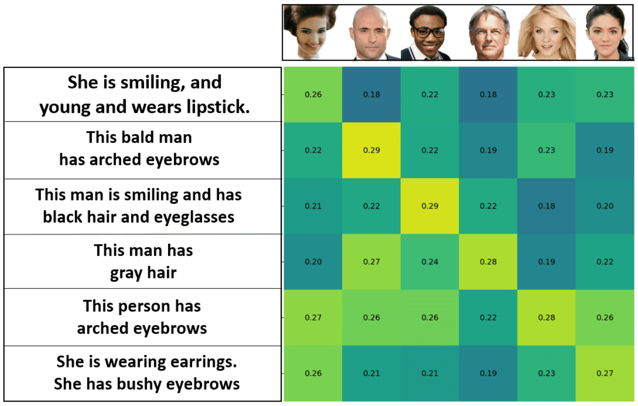

FIGURE 4. The encoder model similarity matrix, showing confident matches along the diagonal that indicate the similarity between the images of interest and their appropriate text descriptions. The prediction outputs along the diagonal is enhanced by the inclusion of the vision section of the pre-trained CLIP model.

B. STYLE LOSS

The style transfer loss comprises of the content and style [1]. The content in our case is derived from the encoder model $E(\cdot)$. Basically, the expectation over the entire spatial space of the feature maps are compared with a loss function, so as to ensure similarity. These are obtained by taking the gram matrix $G^{\phi}(\cdot)$ of the outputs of X and $\hat{X}$ given by:

$$G^{\phi} = (\phi_i(\mathbf{y})_{c,w,h})(\phi_i(\mathbf{x})_{c',w,h}), \tag{10}$$

where $Qi(.)$ represents a filter output function, c' is a transposed channel form of c. The style reconstruction loss for both images is thus an L_1 loss of each computed gram matrix:

$$\mathcal{L}_{sty} = \left\| Q_i^{\phi}(\mathbf{y}) - Q_j^{\phi}(\hat{\mathbf{x}}) \right\|. \tag{11}$$

C. IDENTITY PRESERVING LOSS

To preserve identity during the synthesis, we applied a pre-trained Light CNN2 [47] face recognition network to extract meaningful feature representations that improve the identity preserving ability of the network. We calculated the identity preserving loss L_{id} as the summation of the feature-level difference between the synthesized and the real image, given as:

$$\mathcal{L}_{id} = \left\| P_{id}(\mathbf{y}) - P_{id}(\hat{\mathbf{x}}) \right\|_2^2, \tag{12}$$

where we consider $P_{id}(\cdot)$ as output features from the last fully connected layers of Light CNN network.

TABLE 1. Quantitative comparison of text-guided image manipulation. We compare our method with state-of-the-art techniques; TediGAN [38], ManiGAN [11] in terms of FID, accuracy (Acc.) and realism (Real).

Mode	Method	FID	Acc.(%)	Real.(%)
CelebA	ManiGAN [11]	117.89	27.41± 0.21	10.01± 0.08
	TediGAN [38]	101.25	38.30± 0.01	46.50± 0.03
	CTGAN	**100.17**	**38.85± 0.01**	**45.31± 0.32**
Non-CelebA	ManiGAN [11]	143.39	16.41± 0.11	7.41± 0.08
	TediGAN [38]	**129.27**	40.00± 0.21	45.41± 0.08
	CTGAN	138.10	**42.40± 0.14**	**47.41± 0.08**
Open-Text	ManiGAN [11]	141.51	9.01± 0.11	10.41± 0.08
	TediGAN [38]	113.57	68.01± 0.30	43.24± 0.02
	CTGAN	**106.42**	**69.17± 0.10**	**44.81± 0.08**

D. OVERALL OBJECTIVE FUNCTION

We sum up all the loss functions defined above to compute the overall objective given as:

$$\mathcal{L}_{CT-GAN} = \mathcal{L}_{E_w} + \mathcal{L}_{D_w} + \mathcal{L}_{E_t} + \lambda_1 \mathcal{L}_p + \lambda_2 \mathcal{L}_{sty} + \lambda_3 \mathcal{L}_{id}, \tag{13}$$

where variables λ_1, λ_2, λ_3 are the hyper-parameters used as a weight factor of the different loss terms.

V. EXPERIMENTAL SETUP

We use the Multi-Modal CelebA-HQ [40] dataset for the text-guided multi-modal image synthesis. It's a large-scale dataset which has a high-quality semantic segmentation map, sketch, descriptive texts, and images with transparent background. The text structure comprises of ten unique single sentence descriptions for each image in CelebA-HQ [6]. For training, we divided the dataset into 80% training and 20% test samples, respectively.

A. TRAINING

To train the StyleGAN inversion module [8], we combined features from an image encoder and the features from the CLIP model [13], which was trained on over 400 million image and text pairs, as defined in Equation (1). We adapt this technique in order to achieve better semantic meaning between text and images. In retrospect, we built a visual-linguistic encoder combined with a BERT [14] text encoder to produce embeddings that match the latent space $\mathcal{W}_+$ of StyleGAN. In our design, we adapted the BERT encoder specifically to fine tune the model newly added text descriptions that are of interest for our model. In line with the approach implemented in [10], we trained only the encoder and discriminator while the generator weights are frozen.

B. EVALUATION

We compared our proposed method with similar approaches applied for text and text-guided image synthesis models such as AttnGAN [32], TediGAN [38] ControlGAN [36] and DFGAN [48]. For evaluation, we used techniques similar to [38] to evaluate image quality, diversity, accuracy, and the degree of realism. Following the previous methods [36], [49], we also evaluated the quality of generated or manipulated images using the Frechèt Inception Distance (FID) [50] and the Learned Perceptual Image Patch Similarity (LPIPS) [51].

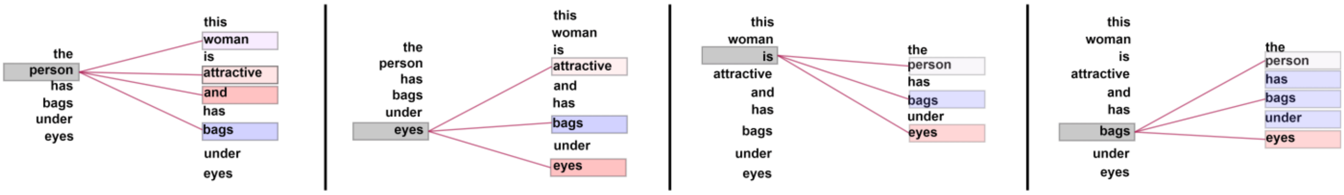

FIGURE 5. Textual semantic interactions between words of sentences, to reflect the relationship between words and texts.

The accuracy of generation is evaluated by checking the level of similarity between text and images.

1) QUANTITATIVE COMPARISON

In our experiments, we evaluate the FID score and also conduct a user study on accuracy and realism by selecting images randomly from both CelebA and Non-CelebA datasets with randomly chosen descriptions, similar to the approach in [38]. The quantitative results are shown in Table 1. Compared with results obtained from ManiGAN [11] and TediGAN [38] our proposed strategy achieves a better FID, accuracy, and realism. The results obtained indicate high-quality synthetic images, with modifications that are aligned with the given text descriptions.

We also compared our method to the optimization technique from Karras [6], the encoder from pSp [38] and IDInvert [8]. The pSp method proposes an auto-encoder training approach, where the encoder is trained alongside the generator to generate latent codes. In IDInvert, images are embedded into $\mathcal{W}_+$ and then optimized over the generated image. In our approach, we applied a linear based visual-linguistic encoder. Table 2 presents a quantitative evaluation measuring the different inversion methods. We computed the structural similarity (Similarity), mean square error (MSE), LPIPS scores as well as the ‘‘Runtime’’ for each model. Compared to other encoders, CT-GAN preserves the original image’s perceptual similarity and subject identity.

2) QUALITATIVE COMPARISON

To analyze overall image quality, we checked different aspects of interest applied for most attribute guided image synthesis models. Firstly, we compared the ability of our model to replicate the visual representation of different user attributes against previous methods that have some form of text or attribute guided modeling. Obtained results shown in Figure 7 and 8 confirm that ControlGAN [36] and DFGAN [48] produce similar results that are quite consistent with the attribute descriptions when compared with our method. However, when compared with AttnGAN [32] and FaceID [39], we see considerable degradation in the image attribute representation, especially for sketch based images. Attributes like lipstick, age, and hair strands were not properly replicated in the synthesized images when compared to our method. We can attribute the weak performance by the nature of the generator adapted [6] and attention model. Our model depends on the latent space of the StyleGAN model [6], [7], [8], which contains better disentangled embeddings of the facial attributes.

TABLE 2. Quantitative comparison of encoders. A comparison of our method with state-of-the-art models; Karras [52], IDInvert [8] and pSp [9] in terms of similarity, MSE, LPIPS, and runtime.

Method	↑ *Similarity*	↓ *LPIPS*	↓ *MSE*	↓ *Runtime*
Karras [52]	0.77	0.11	0.02	182.01
IDInvert [8]	0.35	0.22	0.06	0.032
pSp [9]	0.19	0.19	0.03	0.105
pSp w/o ID [9]	0.49	0.23	0.04	0.064
CT-GAN	0.57	0.16	0.03	0.109

VI. COMPONENT ANALYSIS

In this section, we evaluate the key components that define the performance of the our model. We evaluate the Linear-based attention model for the encoder as well as the visual-linguistic impact of the BERT [14] and CLIP [13] models, respectively. Our findings throw more light on the potential of our model in general.

A. VISUAL-LINGUISTIC ABILITY

Our contextual model comprises of a text based encoder [14] and a visual-linguistic encoder that encodes text [13] and image pairs in a single shot. To get a better understanding on the semantic interaction within words of the same sentence and vice versa, we used the *BertViz* visualizer [12] to reflect the semantic interactions between words and text. We setup a comparison of two short sentences; *A* and *B* as shown in Figure 5. Our goal is to visually show how each word relates to a sentence both locally (within the same sentence) and globally (of a different sentence). Our experimental setup compares two randomly selected sentences:

A: *“The person has bags under eyes”*

B: *“This woman is attractive and has bags under eyes”*

We see that when we choose an identity based word like *“person”* in a comparison between sentences *A* and *B*, represented as $A \xrightarrow{person} B$, we observe that attributes such as: *“woman”*, *“attractive”*, *“bags”*, *“eyes”* show higher color activations. This kind of relationship between different sentences confirms that the association of attributes is consistent regardless of position. In the same fashion, we compared $B \overset{is}{\rightarrow} A$, in this context, “is” is a linking verb used to describe attributes. The neuron activations clearly highlight key words; *“under”*, *“person”*, *“bags”*, *“eyes”* that indicate the quality of semantic association required for an effective visual-linguistic model. We further check the response of the model to the most attended words in unseen text descriptions (test case). We show high resolution examples in Figure 6 and 9 to confirm that the generated images are modified relative to the attribute words in the sentence.

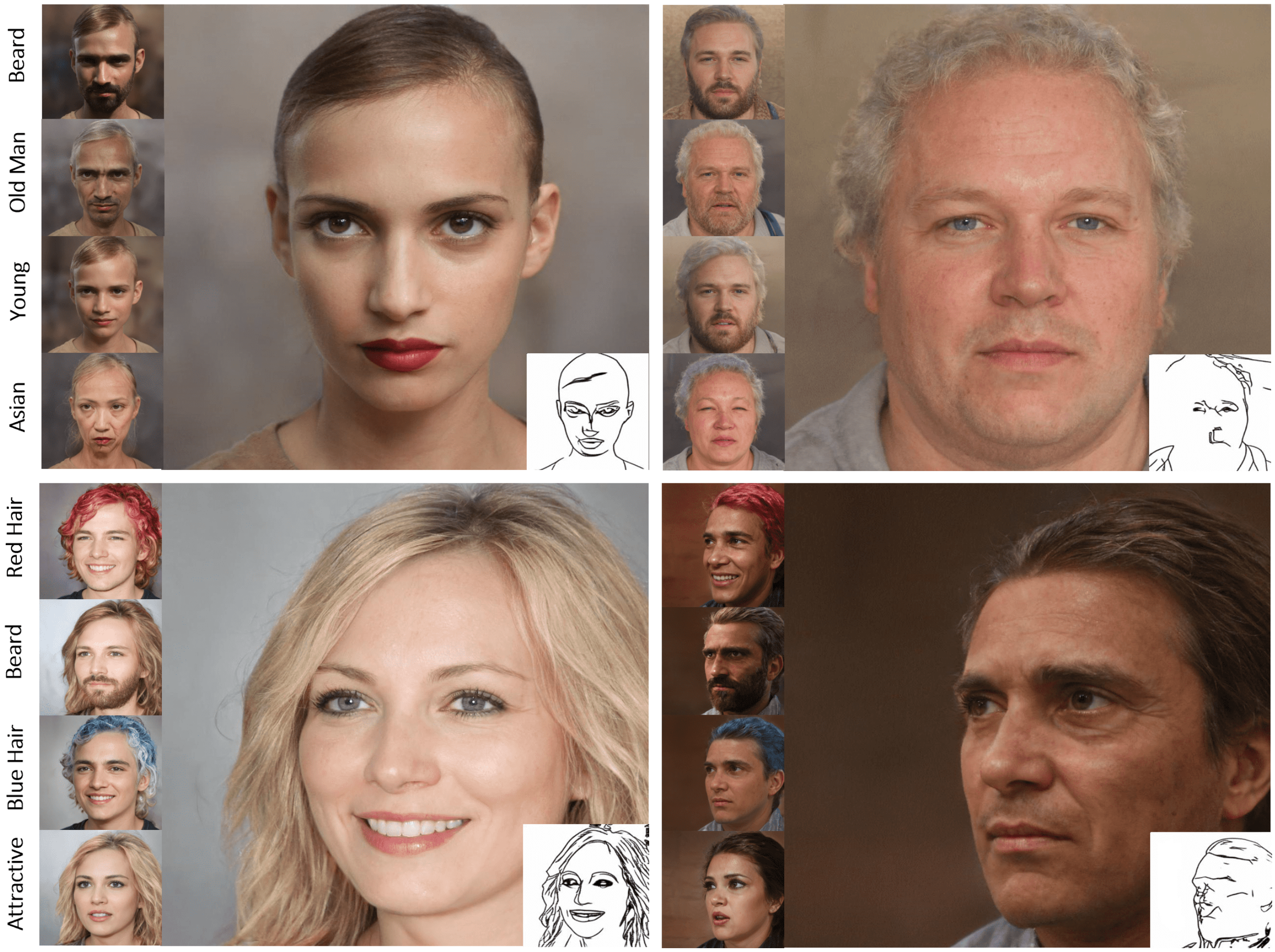

FIGURE 6. Sketches with compound sentences reflecting different attribute combinations.

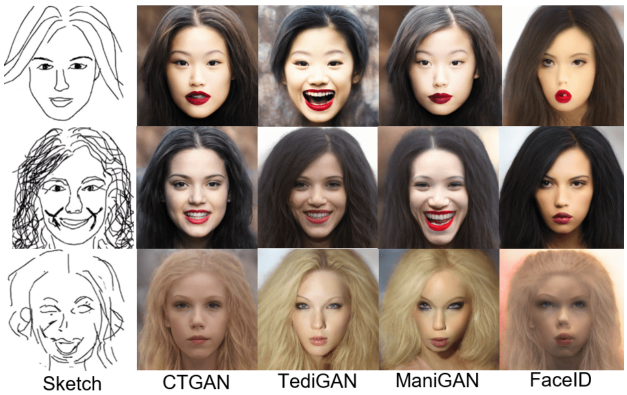

FIGURE 7. Model comparison of images with different attribute combinations.

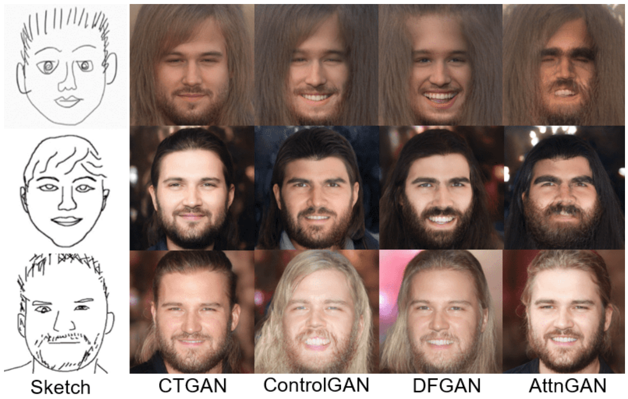

FIGURE 8. Model comparison of images with different attribute combinations.

Our findings indicate that the model is sensitive to the visual-linguistic relationship between image-word pairs. We used the CLIP model to confirm that the semantic features for our encoder model is efficient for text-guided synthesis [13]. In Figure 4, we compared six unique faces with attributes and textual descriptions. The similarity matrix shows confident matches along the diagonal that clearly indicate the similarity between the given image of interest and the appropriate sentence.

B. EFFECTIVENESS OF THE ATTENTION MODEL

In the design of the attention model, we considered the semantic-level relationship of image and word pairs for both channel and spatial representation at the image sub-region and pixel level. To have a better understanding of the benefits

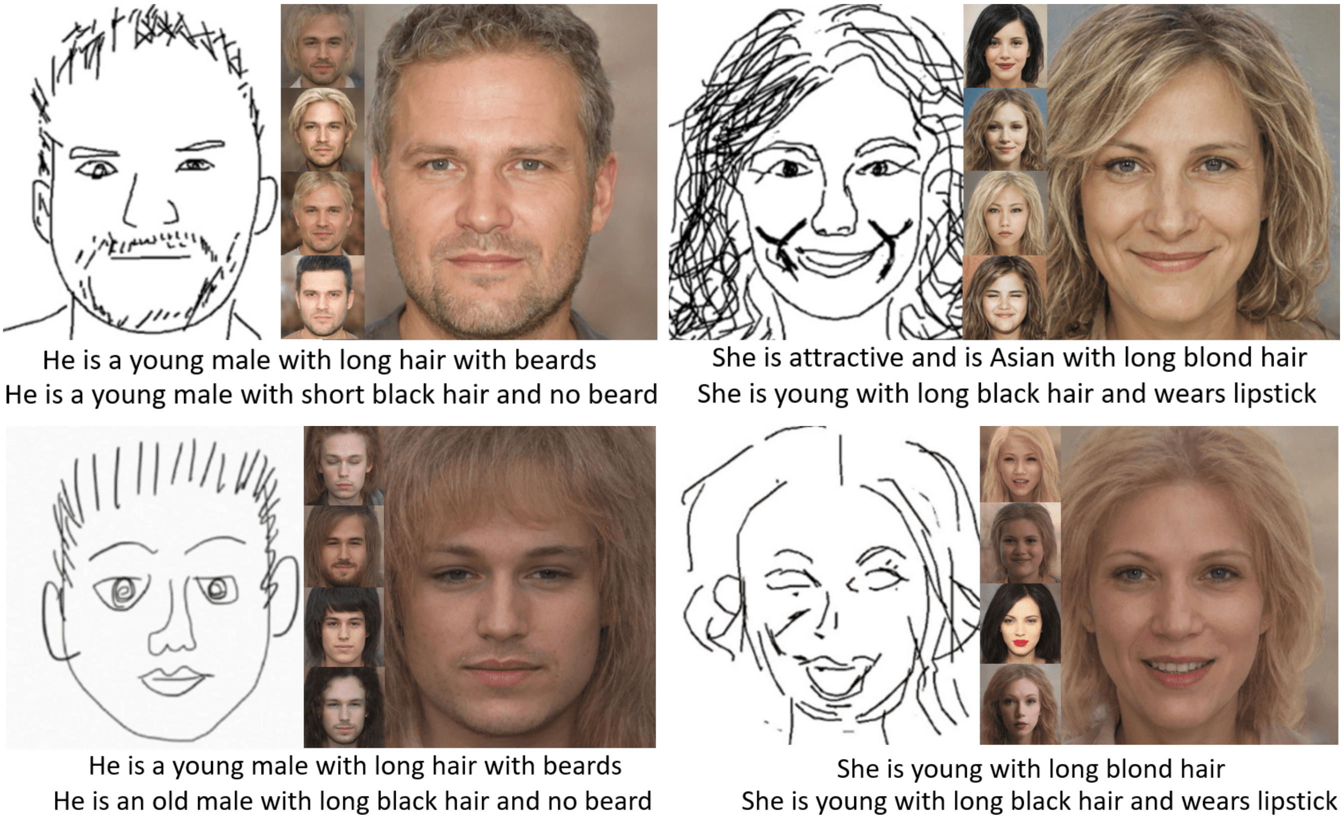
He is a young male with long hair with beards
He is a young male with short black hair and no beard
She is attractive and is Asian with long blond hair
She is young with long black hair and wears lipstick
He is a young male with long hair with beards
He is an old male with long black hair and no beard
She is young with long blond hair
She is young with long black hair and wears lipstick

FIGURE 9. Quality of synthesised images derived from poorly drawn sketches reflecting age, gender, etc.

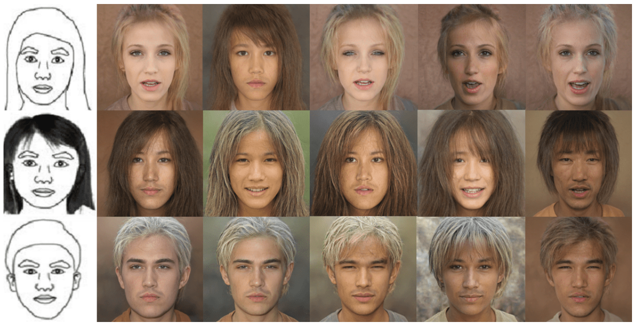

FIGURE 10. Identity preservation. The images represents consistency of the subjects identity for different attributes.

of our approach, we visualized a combination of text based attributes to showcase the power of the attention model adapted. We extend our empirical experiments by implementing different ablation tests on the model to ascertain how the attention scheme benefits the model.

We setup our ablation studies to check the quality of sketch to image synthesis in combination with different degrees of text-based attribute combinations. Most models depend on basic text allocations. However, in our approach, we show different promising text combinations that verify the robustness of our model.

1) GOOD SKETCHES

We chose a collection of good sketches that will set a basis for comparison against *"Bad"* sketches. Good sketches in this case represent facial sketches that contain most of the facial details while the bad sketches have some missing facial details. We guide the image synthesis process of each sketch with different composed sentences, which we arrange in different orders of complexity. Every sentence has key attributes (*"blue hair", "old", "beards" and "pale face"*) that can be visually identified from the results obtained. The aim is to observe the performance of the model at different sketch types and text complexity. As expected, the model easily replicates the sketch into the expected image without compromising the identity portrayed by the sketch regardless of the attribute combinations used as shown in Figure 11.

2) BAD SKETCHES

As the sketches are slightly degraded, the model still maintains its ability to reproduce plausible images at high quality as reflected in Figure 12. We observe that the model still tries to synthesize similar identities, which show the attentiveness of the model to the perceptual identity of the subject. Lastly, we pay key attention to the ability of the model to synthesize images and still maintain the contextual meaning of the sentence for each case study. We can clearly identify the subjects regardless of the poor sketch provided. Our approach confirms that regardless of the sketch quality, visual-linguistic property of the model is still maintained.

C. TEXT-GUIDED ANALYSIS

A combination of visual and text based features in a single generative model creates a set of interesting properties that need to be explored to identify the possibilities of text-guided

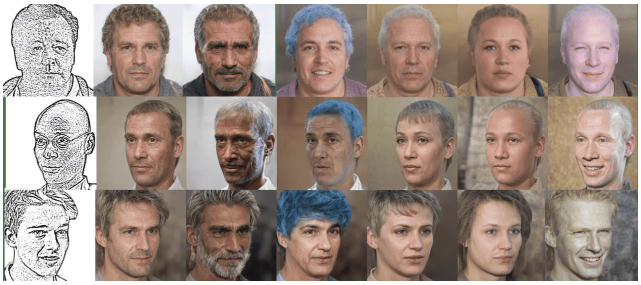

FIGURE 11. Images synthesised from good crafted sketches.

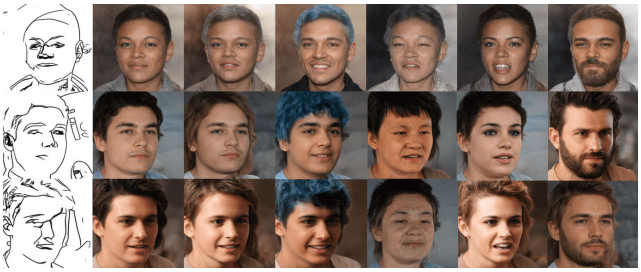

FIGURE 12. Images synthesised from bad crafted sketches.

synthesis. In our approach, we pay more attention to sketches; which is a more difficult problem to solve. To understand the intricacies of our model, we set up a set of experiments to verify the robustness of our model. We looked closely at the ability of the model to maintain the identity of a given subject even when attributes are changed. We also use different sketch templates to analyse the models' ability to synthesise images without compromising identity and perceptual quality.

1) IDENTITY PRESERVATION

The ability to reproduce similar images of the same identity is crucial for sketch base synthesis. In Figure 10, we showcase our model's ability to reproduce identity-consistent images. Our results reflect the synthesised images when we change the subject's attributes such as "gender", "age" and "hair color". We set up the test cases using simple and complex attribute combinations. A sample of the short sentences used are expressed below:

Female → *Male* : *"He is Asian and has short hair."*

Female → *Male* : *"He is a young male."*

Female → *Female* : *"She has long hair and is slim"* Overall, the attributes extracted form the short sentences don't alter the identity of the subjects.

2) COMPOUND SENTENCES

A collection of sentences could be used to synthesise images from sketches. To test the model's response to sentence-based attributes we checked different combination of texts comprising of different attributes. We define compound sentences as a collection of text-based attributes that reflect ***"gender + ethnicity"***,***"hair + age"***, ***"gender + beards"***, ***" beards + gender + age "***, ***"age + ethnicity + gender "***, etc.

Our results confirm consistency in perceptual quality, identity and sketch diversity. Sketch diversity in this case highlights the fact that the model is able to synthesize an image and still represent the attributes within the specified sentence. In Figure 9 we show the perceptual quality of images generated using some compound sentences.

VII. CONCLUSION

In our work, we showed that a text-guided sketch-to-image GAN model can be visually appealing and still portray all the facial attribute within an associated text description. Our model leveraged on the hierarchical structure of the state-of-the-art StyleGAN model to combine visual-linguistic features from a properly disentangled latent space. From our findings, we observe that introducing the CLIP features to our framework encourage better contextual meaning to our results without comprising the identity of the facial results across board. We also confirm that adapting a linear-based attention module aids in generating plausible images.

REFERENCES

[1] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, Oct. 2016, pp. 694–711.

[2] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5908–5916.

[3] W. C. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9416–9425.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–6.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2018, *arXiv:1710.10196*.

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[8] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," 2020, *arXiv:2004.00049*.

[9] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A StyleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[10] Z. Jun-Yan, Z. Richard, P. Deepak, D. Trevor, E. A. Alexei, W. Oliver, and S. Eli, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, 2017, pp. 465–476.

[11] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7877–7886.

[12] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics Syst. Demonstrations*, Jul. 2019, pp. 37–42. [Online]. Available: https://www.aclweb.org/anthology/P19-3007

[13] R. Alec, K. J. Wook, H. Chris, R. Aditya, G. Gabriel, A. Sandhini, S. Girish, A. Amanda, M. Pamela, C. Jack, K. Gretchen, and S. Ilya, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139, Jul. 2021, pp. 8748–8763.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[15] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: Manipulating images with natural language," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 42–51.

[16] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. E. Carlson, and J. Gao, ''StoryGAN: A sequential conditional GAN for story visualization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6322–6331.

[17] H. Dong, S. Yu, C. Wu, and Y. Guo, ''Semantic image synthesis via adversarial learning,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5707–5715.

[18] Y. Liu, M. D. Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, and B. Lepri, ''Describe what to change: A text-guided unsupervised image-to-image translation approach,'' in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1357–1365.

[19] T. Qiao, J. Zhang, D. Xu, and D. Tao, ''MirrorGAN: Learning text-to-image generation by redescription,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.

[20] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, ''Generative visual manipulation on the natural image manifold,'' in *Proc. ECCV*, 2016, pp. 597–613.

[21] Z. C. Lipton and S. Tripathi, ''Precise recovery of latent vectors from generative adversarial networks,'' 2017, *arXiv:1702.04782*.

[22] A. Creswell and A. A. Bharath, ''Inverting the generator of a generative adversarial network,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1967–1974, Jul. 2019.

[23] R. Abdal, Y. Qin, and P. Wonka, ''Image2StyleGAN: How to embed images into the StyleGAN latent space?'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440.

[24] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, ''Invertible conditional GANs for image editing,'' 2016, *arXiv:1611.06355*.

[25] Y. Nitzan, A. H. Bermano, Y. Li, and D. Cohen-Or, ''Disentangling in latent space by harnessing a pretrained generator,'' 2020, *arXiv:2005.07728*.

[26] J. Gu, Y. Shen, and B. Zhou, ''Image processing using multi-code GAN prior,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3009–3018.

[27] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, ''Exploiting deep generative prior for versatile image restoration and manipulation,'' 2020, *arXiv:2003.13659*.

[28] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, ''StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows,'' *ACM Trans. Graph.*, vol. 40, pp. 1–21, May 2021.

[29] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2015, *arXiv:1409.0473*.

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, in JMLR Workshop and Conference Proceedings, vol. 37. Lille, France: JMLR.org, Jul. 2015, pp. 2048–2057.

[31] T. Luong, H. Pham, and C. D. Manning, ''Effective approaches to attention-based neural machine translation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Lisbon, Portugal: The Association for Computational Linguistics, Sep. 2015, pp. 1412–1421, doi: 10.18653/v1/d15-1166.

[32] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, ''AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018 pp. 1316–1324.

[33] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, ''Semantics-enhanced adversarial nets for text-to-image synthesis,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10500–10509.

[34] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, ''A structured self-attentive sentence embedding,'' 2017, *arXiv:1703.03130*.

[35] S. Li, S. Bak, P. Carr, and X. Wang, ''Diversity regularized spatiotemporal attention for video-based person re-identification,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.

[36] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, ''Controllable text-to-image generation,'' *CoRR*, vol. abs/1909.07083, 2019. [Online]. Available: http://arxiv.org/abs/1909.07083

[37] Z. Liu, P. Luo, X. Wang, and X. Tang, ''Deep learning face attributes in the wild,'' in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[38] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, ''TediGAN: Text-guided diverse face image generation and manipulation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2256–2265.

[39] Y. Nitzan, A. H. Bermano, Y. Li, and D. Cohen-Or, ''Face identity disentanglement via latent space mapping,'' *ACM Trans. Graph.*, vol. 39, pp. 1–14, May 2020.

[40] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, ''Towards open-world text-guided face image generation and manipulation,'' 2021, *arXiv:2104.08910*.

[41] S. Shen, L. Harold Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, ''How much can CLIP benefit vision-and-language tasks?'' 2021, *arXiv:2107.06383*.

[42] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, ''MaskGAN: Towards diverse and interactive facial image manipulation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.

[43] S. Barua, S. M. Erfani, and J. Bailey, ''FCC-GAN: A fully connected and convolutional net architecture for GANs,'' 2019, *arXiv:1905.02417*.

[44] A. Jahanian, L. Chai, and P. Isola, ''On the 'steerability' of generative adversarial networks,'' 2020, *arXiv:1907.07171*.

[45] X. Ma, X. Kong, S. Wang, C. Zhou, J. May, H. Ma, and L. Zettlemoyer, ''Luna: Linear unified nested attention,'' 2021, *arXiv:2106.01540*.

[46] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[47] X. Wu, R. He, Z. Sun, and T. Tan, ''A light CNN for deep face representation with noisy labels,'' *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[48] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X. Jing, ''DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis,'' 2020, *arXiv:2008.05865*.

[49] B. Li, X. Qi, P. H. S. Torr, and T. Lukasiewicz, ''Lightweight generative adversarial networks for text-guided image manipulation,'' 2020, *arXiv:2010.12136*.

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, ''GANs trained by a two time-scale update rule converge to a local Nash equilibrium,'' in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6626–6637.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[52] P. Kancharla and S. S. Channappayya, ''Quality aware generative adversarial networks,'' vol. abs/1911.03149, 2019. [Online]. Available: http://arxiv.org/abs/1911.03149

UCHE OSAHOR (Student Member, IEEE) received the B.Eng. degree in electrical engineering from the University of Maiduguri, Borno State, Nigeria, and the M.Sc. degree in electrical and electronic engineering with specialization in control and instrumentation from Obafemi Awolowo University, Nigeria. He is currently pursing the Ph.D. degree in electrical engineering with West Virginia University, Morgantown, WV, USA. Since 2018, he has been involved in deep learning research under the guidance of Prof. Nasser M. Nasrabadi. His research interests include applications of deep learning, machine learning, and image processing.

NASSER M. NASRABADI (Fellow, IEEE) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. In 1984, he was with IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was with the Philips Research Laboratory, New York, NY, USA, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor with the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. From 1991 to 1996, he was an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY, USA. From 1996 to 2015, he was a Senior Research Scientist with the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests include image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, neural networks, and image processing. He is a fellow of the International Society for Optics and Photonics, ARL, and SPIE. He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS, SYSTEMS AND VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS.