IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 January 2023, accepted 25 January 2023, date of publication 10 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243911

RESEARCH ARTICLE

EBAT: Enhanced Bidirectional and Autoregressive Transformers for Removing Hairs in Hairy Dermoscopic Images

YOUNGCHAN LEE AND WONSANG YOU, (Member, IEEE)
Artificial Intelligence and Image Processing Laboratory, Department of Information and Communication Engineering, Sun Moon University, Asan-si 31460, South Korea

Corresponding author: Wonsang You (wyou@kaist.ac.kr)

This work was supported in part by the Support Program for Local Specialized Industry Promotion Business under Grant S3091196; in part by the Start-Up Growth Technology Development Project funded by the Ministry of SMEs and Startups (SME: Small and Medium-sized Enterprise) under Grant S3228660; in part by the Basic Science Research Program under Grant NRF-2022R1F1A1075204; in part by the BK21 FOUR (Fostering Outstanding Universities for Research); and in part by the Regional Innovation Strategy Project under Grant 2021RIS-004 through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, South Korea.

ABSTRACT A great progress in deep learning technologies for skin cancer detection from dermoscopic images has been made for a decade. While its performance is vulnerable to a large amount of hairs densely covering the skin surface, the existing image processing methods frequently fail to remove hairs in hairy skin images. In this paper, we propose, as a deep learning approach to removing hairs, a generative image inpainting network where bidirectional autoregressive transformers (BATs) are employed to learn image features and are systematically integrated with convolutional neural networks (CNNs) in multiple spatial scales in order to reconstruct missing regions. Each patch split from a masked image is unfolded and processed through BATs, and re-folded to constitute diverse shapes of feature maps through kernel-based unfolding-folding operations. By introducing the multi-scale features extracted by collaborative learning of transformers and CNNs to the texture generator network, our method can effectively reconstruct minute details of local regions as well as global structure which might not be easily inferred from neighbor pixels in hairy skin images. Quantitative and qualitative evaluations show not only that our multi-scale dual-modality strategy is much robust to reconstruct hair-shaped missing regions compared to the existing transformer-based image inpainting method called BAT-Fill, but also that our framework outperforms the state-of-the-art image inpainting models in removing hairs from hairy dermoscopic images.

INDEX TERMS Hair removal, skin image, image inpainting, transformer, deep learning, generative adversarial networks.

I. INTRODUCTION

Artificial intelligence (AI) has fast leaped forward as assistive healthcare technologies for medical doctors and patients, and it has been also applied for detecting a skin cancer from dermoscopic images with remarkable performance [1]. One of artifacts in skin image analyses is hairs which cover either widely or locally over the surface of skins [2]. Hairs may seriously perturb analyzing lesion textures and result in reduced accuracy of a deep learning model predicting the cancer type corresponding to the skin lesion. Accordingly, the preprocessing stage of removing hairs from a dermoscopic image is essential in the deep learning framework for skin cancer detection [3].

The hair removal in dermoscopic images can be understood as the problem of image inpainting that aims to reconstruct an incomplete image by filling missing parts naturally [4]. Two major image inpainting approaches to hair removal are

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

present; one is based on image processing, and the other is based on machine learning.

In the image processing approach to hair removal, the Dullrazor software was a pioneering software, introduced by Lee et al., by which hair regions are detected from a grayscale skin image using morphological filtering and are restored through bilinear interpolation with neighbor pixels [5]. E-shaver is its extension toward color skin images where edge filtering and color averaging are employed for hair detection and inpainting respectively [6]. Those early algorithms for removing hairs had been advanced using a diversity of image inpainting techniques including partial differential equation and coherence transport [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. In particular, the fast marching method (FMM), where a skin region hidden by hairs is filled progressively from its border, is one of the robust methods for removing hairs [7], [17]. Since most skin image diagnostic technologies have targeted human skin images with few hairs, the image inpainting methods based on image processing have exhibited quite acceptable performance on such tractable human skin images despite the simplicity of their algorithms. However, their performance had not been verified for hairy skin images whose skin surface is covered with lots of hairs in either quantitative or qualitative manner.

On the other hand, the machine learning approach to hair removal had little advance compared to the image processing approach. It is obviously contrasted with the fast-growing trend of deep learning models for image inpainting including Shift-Net, DeepFill, GMCNN, PartialConv, LBAM [18], [19], [20], [21], [22], [23]. One practical reason for the underdevelopment of machine learning-based hair removal methods would be that an advanced hair removal algorithm with large computational complexity was not necessary in a moderate skin image with few hairs. The other reason is the difficulty in finding such a training dataset that consists of pairs of original skin image with hairs and the corresponding hairless image. To tackle the problem, Talavera-Martinez et al. first employed a convolutional neural network model with simple encoder-decoder architecture to remove hairs in a skin image [3]. They exploited, as a training dataset, pairs of hairless skin images extracted from public datasets as ground truth and their corresponding images with simulated hairs as input data. Bardou et al. examined a variational autoencoder model, where hairs are eliminated as noise, that consequently can be trained without such a dataset as consisting of pairs of hairy and hairless skin images [24]. Li et al. used DeepFill along with gated convolutions to allow free-form image inpainting optimized to hair shapes [25].

Despite the recent noteworthy advance in methodologies for hair removal, the existing methods tend to be vulnerable to such a dermoscopic image in which hairs are densely jammed and tangled, as illustrated in Section IV. The existing methods based on either image processing or convolutional neural networks remove hairs from a skin image by filling a hair region progressively from its border using the properties of local skin texture [7], [25]. However, in such an extreme situation as hairy skin images the surrounding neighbor pixels of a hair region are prone to severe contamination by other hairs covering the skin, which might consequently lead to the reduced performance of reconstructing the skin texture hidden behind hairs.

The transformer can be taken into account as a solution to cope with such technical limitations that may be faced in hairy skin images. It was first introduced by Vaswani et al. in 2017 as a machine translation model that can learn the meaning of a sentence through the attention mechanism which quantifies the contextual intra-relation of words in a sentence [26]. The transformer has been successfully employed mainly for machine translation, and has increasingly applied for a wide range of computer vision problems including image inpainting [27].

An image inpainting model based on bidirectional and autoregressive transformers (BATs), so called BAT-Fill, was recently introduced especially to reinforce the capability of generating diverse contents of missing region [28]. It has a coarse-to-fine network architecture that is composed of the coarse-structure generator based on transformers and the fine texture generator based on generative adversarial network (GAN) [29]. We assessed the applicability of BAT-Fill, the transformer-based image inpainting model, to removing hairs from dermoscopic images. As shown in Section IV, we found that, although its qualitative performance in removing hairs was superior to representative existing methods, fragments of hairs often remain incompletely eliminated in visual inspection and as a consequence the reconstructed skin images used not to be so much as acceptable for clinical use.

In this study, we propose a multi-scale GAN framework, called EBAT (named in the sense of an enhanced network of BAT-Fill), where the image features encoded in multiple spatial scales by both transformers and convolutional neural network (CNN) are jointly learnt to reconstruct fine skin textures as well as global structure. While in BAT-Fill the transformers produce a single low resolution image of coarse structure reconstructed from the corresponding down-sampled input image, the transformers in EBAT generate a set of multi-scale feature maps through kernel-based unfolding-folding operations instead of down-sampling.

The main contributions and novelties of this work can be summarized as follows. First, the transformers are used to extract not coarse and diverse structures but multi-scale image features with long-range dependency, which can enhance the capability of reconstructing fine details and global structure simultaneously. It is through patch-wise unfolding and downsized folding operations that a feature map as an output of BATs can be generated to have an arbitrary shape. Second, the multi-scale feature extractors based on transformers and CNNs are unified and attached to the fine texture generator through multi-scale pathways, which allows collaborative learning with efficient information flows between two backbones. Third, both multi-scale feature extractor and fine texture generator are jointly trained in an end-to-end fashion while in BAT-Fill the coarse-structure generator is completely separated from the fine texture generator in the process of training. Lastly, the inference time was greatly reduced, compared to BAT-Fill, through bypassing the long-winded procedure of diverse contents generation involving millions of pixel-wise operations on CPUs.

using image patches instead of a down-sampled image as an input to transformers.

II. RELATED WORKS

A. CNN/GAN-BASED IMAGE INPAINTING

In this section, we review two representative deep learning models for image inpainting; one is Shift-Net as a CNN-based framework, and the other is DeepFill as a generative adversarial network (GAN) based framework [18], [21].

Shift-Net is an extension of the U-net architecture where the encoder features of the known region are shifted to the decoder through a shift-connection layer and are employed to estimate the missing regions accurately [18]. The network is trained to minimize the guidance loss which is defined as the discrepancy between the predicted feature and the ground-truth features of the missing regions.

On the other hand, DeepFill is a generative image inpainting framework where gated convolution and SN-PatchGAN are applied to enable using free-form masks with arbitrary shapes [21]. Unlike either vanila convolution or partial convolution which uses hard-mask gating, the gated convolution lets free-form masks to be softly updated over layers [20], [21]. SN-PatchGAN is a variant of vanilla GAN where the hinge loss of the discriminator is computed not on a single output value (real or fake) but on all points of the output map, that enables the GAN framework to use free-form masks. It makes a decisive difference with vanilla GAN that is designed based on a single rectangular mask. It also includes contextual attention modules to detect long-range dependencies between distant local regions.

B. BAT-FILL: TRANSFORMER-BASED IMAGE INPAINTING

As illustrated in Figure 1, BAT-Fill, a transformer-based image inpainting method, is composed of a diverse structure generator followed by a texture generator, which aims to achieve both diversity and accuracy [28]. A sequence of bidirectional autoregressive transformers (BATs) are used to reconstruct coarse but diverse structures from a down-sampled masked image.

The transformers fill the missing regions pixel by pixel in an autoregressive manner, which in turn faciliates the diversity of image generation. In addition to the autoregressive modeling, masked language modeling was also adopted, similar to BERT, so that it refers to bidirectional contextual dependency to better predict missing regions especially which have arbitrary shape and surrounding background of rough texture [30].

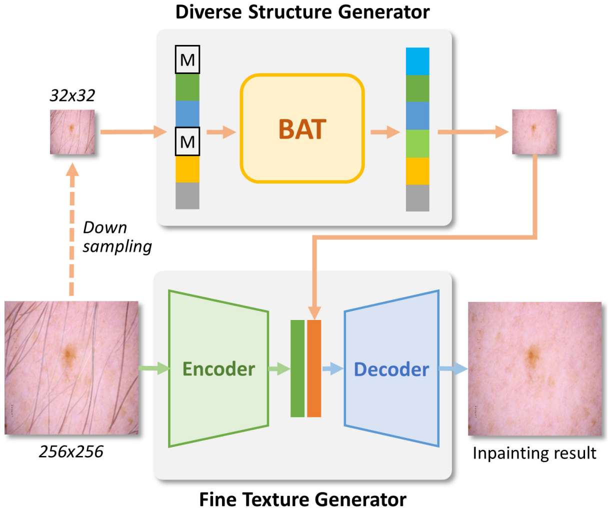

FIGURE 1. The architecture of BAT-Fill as a transformer-based image inpainting method. A masked image is down-sampled and flattened as a word sequence including mask tokens. Missing regions corresponding to mask tokens are filled by the bidirectional autoregressive transformers (BATs), and reshaped into the down-sampled image size. The masked image also passes through a convolutional network; its output feature map is concatenated to the low resolution image reconstructed by BATs, and fed to the fine texture generator as a refinement network.

On the other hand, the texture generator, as the refinement network following the diverse structure generator, is based on generative adversarial learning to regenerate fine-grained details of image texture up-sampled from the low-resolution image which is reconstructed by the diverse structure generator. It also takes advantage of unstained pixels of the input image whose features are encoded by a separate encoder whose architecture resembles the contracting path of U-net [31]. The encoded feature maps are concatenated with the reconstructed low resolution image from BATs, and they are jointly fed to the decoding path of the texture generator to be up-sampled in stages through spatially-adaptive normalization and gated convolutions [21], [32]. The texture generator is completely separated from the diverse structure generator in its training process.

III. PROPOSED METHODS

In the proposed GAN framework, the generator has an encoder-decoder network architecture that is composed of two major parts: a multi-scale feature extractor and a fine texture generator. As illustrated in Figure 2, it could be seen as a similar architecture as U-net but including dual encoding backbones [31]. The multi-scale feature extractor as the encoding part consists of both transformer and CNN backbones that generate the multi-scale feature maps in which missing regions are filled throughout low and high resolutions. The fine texture generator as the decoding part integrates both multi-scale feature maps from the transformer and CNN backbones to reconstruct the fine-grained textures of missing regions through the up-sampling pathways where the context information of lower resolution feature maps are

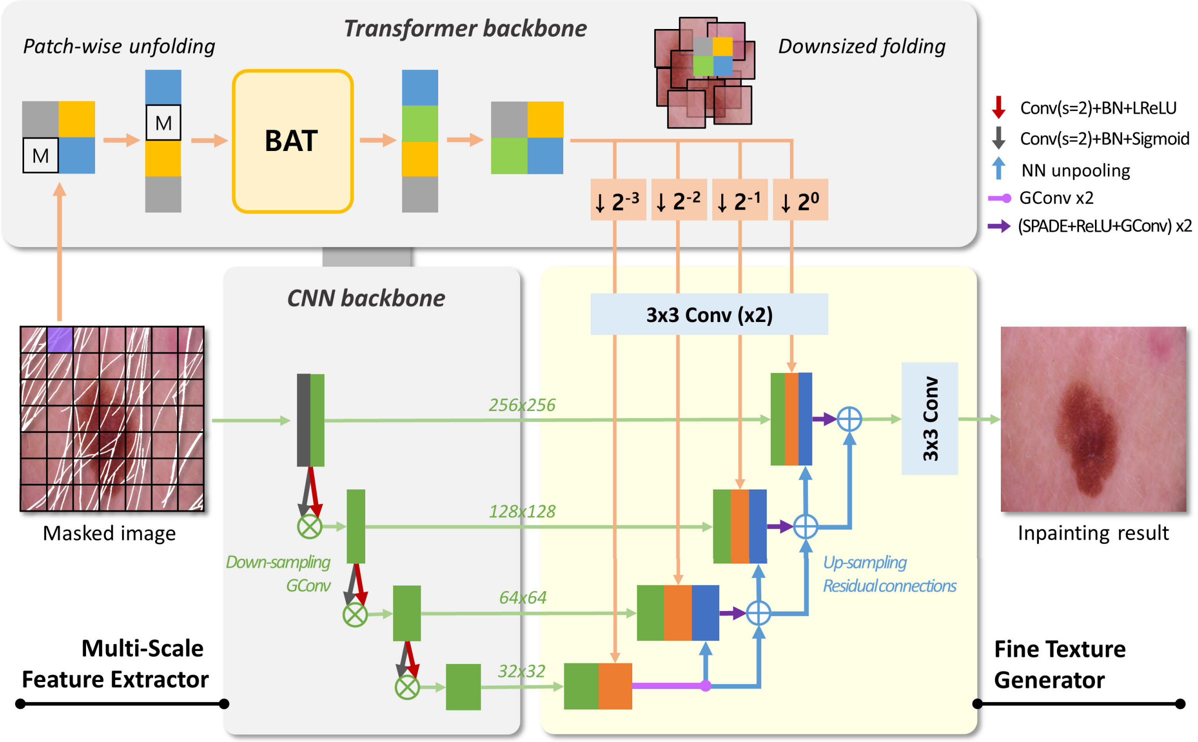

FIGURE 2. **The proposed image inpainting network for removing hairs from hairy dermoscopic images consisting of a multi-scale feature extractor and a fine texture generator. In the transformer backbone of the multi-scale feature extractor, a masked image is split into patches, and each patch is unfolded into a masked sequence including mask tokens denoted as [*M*], fed to BATs, and unfolded to have the equivalent or down-sized shape by 2^{-1}, 2^{-2}, and 2^{-3} times. The masked image also passes through the CNN backbone and its features are encoded in multiple spatial scales. The multi-scale feature maps from both backbones are concatenated to the fine texture generator to reconstruct fine details of missing regions. The ⊕ and ⊗ symbols represent element-wise addition and multiplication respectively.**

propagated to higher resolution layers. The discriminator, the other core element of the GAN framework, has the identical structure to the one in BAT-Fill.

A. MULTI-SCALE FEATURE EXTRACTOR

The multi-scale feature extractor is composed of a transformer backbone and a CNN backbone. The transformer backbone consists of a sequence of BATs in the common manner as the diverse structure generator in BAT-Fill. However, it aims to produce the multi-scale feature maps of the input image whose missing regions are filled taking the long-range dependency between distant regions into account and are readjusted to multiple spatial scales through the operations of pair-wise unfolding and down-sized folding, while the diverse structure generator in BAT-Fill aims to reconstruct the coarse and diverse structure of the down-sampled input image.

1) TRANSFORMER BACKBONE

To permit the generation of multi-scale feature maps, an input masked image of 256 × 256 is not down-sampled but split into 1024 non-overlapping patches of 8 × 8 size. Each patch is flattened into a sequence of length $192 = 8 \times 8 \times 3$ with mask tokens corresponding to a missing region. The patch-wise reshaped input image of 192 × 1024 size is processed through BATs without position embedding [28]. The output feature vector corresponding to each patch is reshaped into the same size of 8 × 8 as the input patch. However, as illustrated in Figure 3, the reconstructed output patches are not located in their original spatial positions but folded through a sliding kernel, and are consequently merged by summing all spatially overlapping values among blocks. The output shape resulted from kernel-based folding is determined by three parameters including kernel size in a spatial dimension (k), striding (s), padding (p), and dilation (d) as follows

$$Y = s(L-1) + d(k-1) - 2p + 1 \tag{1}$$

where Y and L denote the output size and the total number of patches respectively. Supposing that the 8 × 8 patches split from an input image of 256 × 256 size are folded using an 8 × 8 kernel parameterized with stride of 4, padding of 2, and dilation of 1, the folded feature map has the down-sized shape of 128 × 128 based on the equation $Y = 4 \times (32-1) + 1 \times (8-1) - 2 \times 2 + 1$.

While the conventional folding operation for vision transformers has been designed to reconstruct the output feature map with the same shape as the input image has, the folding operation in the proposed method is more flexible so as to produce the different size of output feature map. The shape of the merged feature map is determined depending on the settings of the sliding kernel; in other words, denser striding between blocks leads to larger overlapping areas and

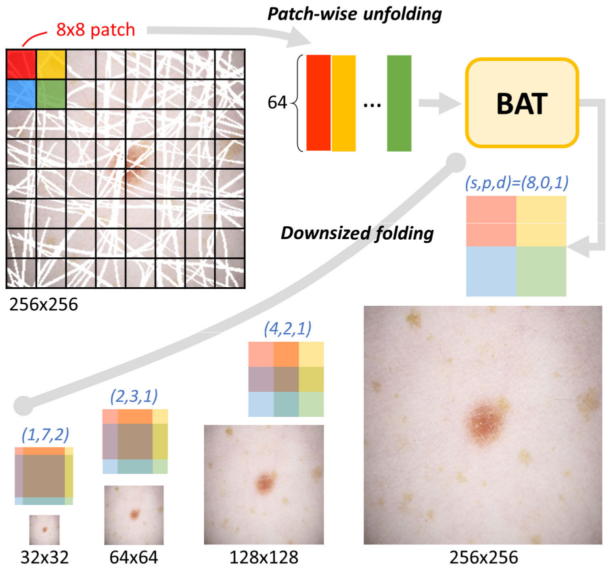

FIGURE 3. Pair-wise unfolding and down-sized folding operations for the transformer backbone. A masked image of 256 × 256 size is split into 1,024 patches of 8 × 8 size. Each patch is unfolded into a sequence of length 192 = 8 × 8 × 3 including mask tokens. After being processed through BATs, the output feature vectors are folded to build an equivalent or down-sized shape by controlling the parameters of stride (*s*), padding (*p*), and dilation (*d*).

in turn a reduced size of the resulting output feature map. The feature map with reduced size is expected to represent the encoded features of coarse structure. On the other hand, the resulting feature map will have exactly the same size as the input image if the maximum striding (as the original distance between subsequent blocks) is applied not to allow overlapping areas between blocks.

2) CNN BACKBONE

The CNN backbone, as the other encoding channel, is almost similar to the the contracting path of the U-net where the input image is gradually down-sampled by the pooling operation with stride 2, however all vanilla convolutions are replaced to the gated convolutions [21], [31]. As shown in Figure 2, an input feature map goes through two pathways of gated convolution; one is down-sampling convolution with stride 2 followed by batch normalization and leaky rectified linear unit (Leaky ReLU) with negative slope of 0.2 as an activation function, and the other consists of convolution with stride 2, batch normalization, and sigmoid function, and two pathways are multiplied together [33]. The gated convolution is designed to train an image inpainting model through soft gating where the mask is allowed to have gating values ranging from zero to one. The mechanism of the gated convolution makes it feasible to use a mask with arbitrary shape rather than a rectangular mask when training the image inpainting network.

B. FINE TEXTURE GENERATOR

The fine texture generator reconstructs fine-grained textures of the missing regions through phased operations that consist of a concatenation of the up-sampled low-level feature map with the feature maps conveyed from both the transformer and CNN backbones of the multi-scale feature extractor, followed by two gated convolutions and spatially adaptive denormalization (SPADE), a residual connection of the up-sampled low-level feature map, and up-sampling by nearest neighbor unpooling, as illustrated in Figure 2 [32]. In the gated convolutions in the texture generator, Leaky ReLU were replaced with ReLU as an activation function [33].

C. LOSS FUNCTION

The loss function for the GAN generator is given, similar to BAT-Fill, to be the combination of a L_1 loss, a perceptual loss L_{per}, and an adversarial loss L_{adv}, which can be mathematically formulated as follows

$$L = \lambda_1 L_1 + \lambda_{per} L_{per} + \lambda_{adv} L_{adv}. \tag{2}$$

where λ_1, λ_{per}, and λ_{adv} are the weighting coefficients corresponding to each loss which were set to be 1.0, 1.0, and 0.2 as done in BAT-Fill [28].

The L_1 loss is defined as the absolute difference between the predicted output image and the corresponding ground truth, while the perceptual loss L_{per} is defined to be the sum of absolute differences in feature maps of special layers in a pretrained VGG-19 network between the predicted output image and the ground truth [34]. On the other hand, the adversarial loss L_{adv} for the generator is defined based on Wasserstein GAN to be the negative expectation of the discriminator output for the predicted output image as the generator aims to delude the discriminator into recognizing the predicted image as the real sample [35].

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) TRAINING CONFIGURATIONS

The proposed method was implemented in the Pytorch framework (version 1.10.1) based on the source codes of BAT-Fill, and was trained on 4 NVIDIA RTX A6000 GPUs with CUDA (version 11.4) [28]. It was optimized using Adam solver with a mini-batch size of 8 and a learning rate of 2×10^{-4} [36].

In particular, hair-shaped masks were alternately employed along with rectangular masks to train the proposed model as well as BAT-Fill, in order to entice the model into increasing the adaptability to hair-like shapes of missing regions in skin images. To increase the complexity and heterogeneity of hair patterns, the hair-shaped masks were generated by stacking in randomly chosen angles (among 0°, 90°, 180°, 270°) a few hair template masks where hair regions are segmented from dermoscopic images using LadderNet [37].

2) TRAIN DATASETS

Two publicly available datasets of CelebA-HQ and ISIC-2020 were separately used to train the proposed model

TABLE 1. Quantitative evaluations of the proposed framework with the state-of-the art image inpainting methods over ISIC 2020 hairless dermoscopic images. Shift-Net, DeepFill v2, and BAT-Fill were trained using the official source codes released in public. All the metrics were assessed in different conditions of hair density: low (the mask ratio of 10%), high (30%), and random densities (whose mask ratios are randomly sampled ranging from 10% to 30%). A standard deviation was parenthesized under the corresponding mean, and the best performance was denoted in bold.

Methods	PSNR↑			SSIM↑			FID↓			LPIPS↓		
	10%	30%	Random	10%	30%	Random	10%	30%	Random	10%	30%	Random
FMM	40.72 (2.59)	36.20 (1.80)	38.20 (3.01)	0.945 (0.025)	0.866 (0.052)	0.900 (0.051)	49.93	164.06	103.87	0.065 (0.027)	0.184 (0.058)	0.127 (0.072)
ShiftNet	35.01 (2.01)	31.56 (1.36)	33.11 (2.19)	0.865 (0.043)	0.741 (0.060)	0.803 (0.078)	89.39	100.12	96.10	0.163 (0.056)	0.263 (0.076)	0.213 (0.068)
DeepFillv2	38.51 (2.26)	33.03 (1.65)	35.60 (2.87)	0.929 (0.033)	0.799 (0.095)	0.864 (0.090)	81.70	183.26	132.82	0.095 (0.045)	0.267 (0.092)	0.185 (0.113)
BAT-Fill	39.26 (2.85)	32.91 (1.93)	35.90 (3.60)	0.919 (0.042)	0.728 (0.120)	0.826 (0.123)	67.64	162.62	119.85	0.146 (0.079)	0.370 (0.127)	0.267 (0.148)
Ours	**43.07** (2.50)	**38.40** (1.78)	**40.80** (2.81)	**0.969** (0.014)	**0.916** (0.031)	**0.942** (0.033)	**16.42**	**32.64**	**35.50**	**0.020** (0.009)	**0.051** (0.017)	**0.037** (0.019)

and evaluate its performance compared to the other methods as described in Section IV-A4 [38], [39]. With the identical experimental settings to BAT-Fill, CelebA-HQ which is a large human face dataset with 30,000 high quality images was used in this study to compare the performance of the proposed method with BAT-Fill. CelebA-HQ was split into 28,000 and 2,000 for training and validation where 1,000 images chosen randomly from the validation set were used for evaluation as well.

On the other hand, the international skin imaging collaboration (ISIC-2020) dataset consists of more than 33,000 dermoscopic images acquired from over 2,000 patients including various types of skin lesion including melanoma (mel), seborrheic keratosis (sk) and nevus (nev) [39]. Adopting the general process of training an image inpainting model using a ground truth image and the synthetically generated mask jointly, we built a training set of hairless dermoscopic images which were manually chosen from ISIC-2020 to train the proposed model along with pre-extracted hair-shaped masks explained in Section IV-A1. The set of hairless images were divided up into three subsets: 3,000 for training, 100 for validation, and 150 for evaluation. Dermoscopic images of diverse resolutions and sizes in ISIC-2020 were reshaped and cropped into 256 × 256.

3) TEST DATASETS

For quantitative evaluation, we built a test dataset including 150 hairless skin images and the corresponding simulated images with fake but realistic hairs, as an alternative solution to the absence of such a dataset as the paired skin images with and without hairs. The simulated images were generated by blending a hairless skin image with one or more stacked hair textures using the Poisson editing algorithm [40]. Note that the hair textures were extracted from hairy skin images by the same method as described in Section IV-A1.

For qualitative evaluation, we built the other test dataset of 30 hairy skin images chosen from ISIC-2020. For a test skin image, the corresponding hair region mask was generated using LadderNet to be used as an input to the common image inpainting pipeline [37].

4) COMPARED METHODS

The proposed method was compared with a few state-of-the-art image inpainting models which were introduced in Sections I and II, including FMM as an image processing based approach, Shift-Net as a CNN-based approach, DeepFillv2 as a GAN-based approach, and BAT-Fill as a transformer-based approach [7], [18], [21], [28].

The default configuration for training the compared methods were set to use the Adam optimizer and the mini-batch size of 8 [36]. As a few exceptions, Shift-Net was trained with the mini-batch size of 1, and the diverse structure generator of BAT-Fill was trained using Adam with decoupled weight decay (AdamW) [41].

5) EVALUATION METRICS

Although there is no unanimous metric for quantitative evaluation in image inpainting as discussed in [42], four well-known metrics were used to evaluate the proposed method on the test dataset of hairless skin images and simulated hairy images described in Section IV-A3. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) with the window size of 11 were adopted which are widely used in image inpainting [43]. The Fréchet inception score (FID) and the learned perceptual image patch similarity (LPIPS) are less widely used compared to PSNR and SSIM but more pertinent to the perceptual quality and diversity of inpainting results respectively [44], [45].

B. QUANTITATIVE EVALUATION

We first trained the proposed method and evaluated its performance using the publicly available human face dataset CelebA-HQ that was previously used to train and test BAT-Fill [28]. Table 2 summarizes our quantitative evaluation results on 1,000 human face images from CelebA-HQ in respect of PSNR and SSIM. It should be noticed that the values for DeepFill v2 and BAT-Fill were quoted from the experimental results reported by Yu et al. in 2021. Interestingly, the proposed method exhibited an increase of 7.98 dB in PSNR but a decrease of 0.105 in SSIM compared to BAT-Fill when the mask ratios were randomly

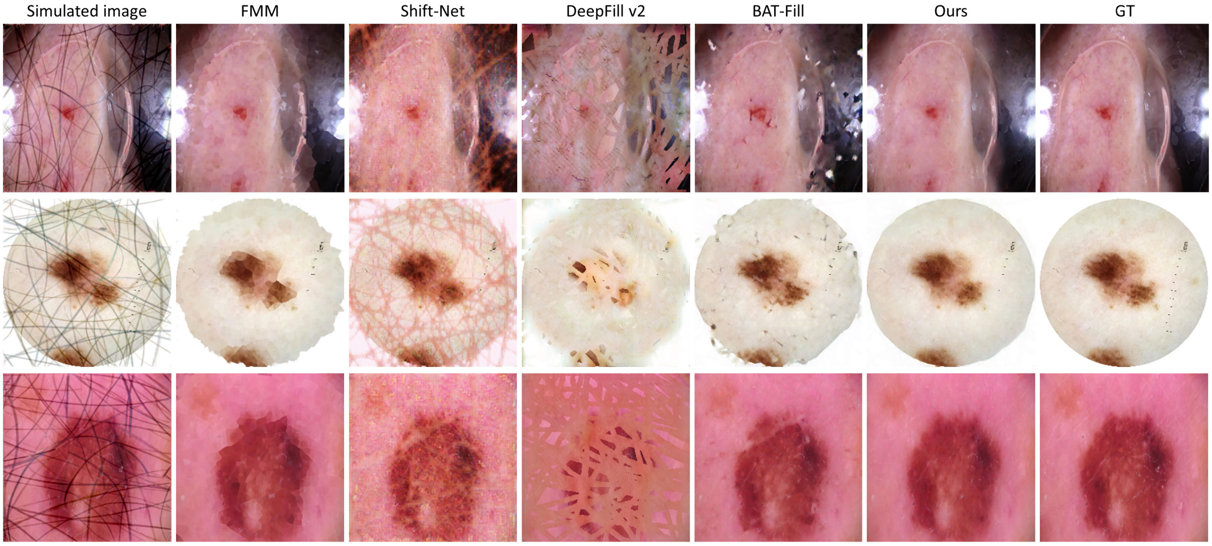

FIGURE 4. Qualitative comparison of the proposed framework with the state-of-the-art image inpainting methods for removing hairs over the simulated hairy skin images from ISIC 2020. The images generated by FMM are seriously blurred impairing fine texture details, and Shift-Net and DeepFill v2 are not successful to get rid of hair marks. The images generated by BAT-Fill are much better but still includes hair fragments which are visually distinguishable. Our framework removes all hairs and accurately reconstructs skin and lesion textures.

TABLE 2. Quantitative evaluations of the proposed framework with the state-of-the art image inpainting methods over CelebA-HQ human face images. The values for DeepFill v2 and BAT-Fill were copied from [28]. Both PSNR and SSIM were assessed for small mask ratios (ranging from 20 to 40%) and large make ratios (40-60%). The best performance was denoted in bold.

Methods	PSNR↑		SSIM↑	
	20-40%	40-60%	20-40%	40-60%
DeepFillv2	25.17	21.21	0.907	0.805
BAT-Fill	27.82	22.40	**0.944**	**0.834**
Ours	**34.33**	**32.06**	0.837	0.733

given ranging from 20% to 60%. It implies that the image generated by our framework is more congruous to its ground truth even in large missing regions, compared to DeepFill and BAT-Fill despite accompanying a nonnegligible loss in structural information.

To assess the proposed method in regard to removing hairs from dermoscopic images, our framework was trained and tested using the ISIC-2020 dataset described in Section IV-A2 [39]. To analyze the effects of hair density on the performance, the hairy images were simulated to have either low or high hair pixel densities (of 10% and 30% on average) by stacking just one or three hair template masks as delineated in Section IV-A1. We carried out a paired t-test to verify the statistical significance of the difference in evaluation results between the proposed method and the other method.

Table 1 shows the statistical evaluation results on the paired set of 150 hairless skin images and simulated hairy images. Compared to the compared methods, PSNR and SSIM of our framework highly increased (with 2.6 dB and 0.042 over the second best method respectively) while FID and LPIPS heavily decreased (with 60.6 and 0.09), which were statistically significant (p-value < 0.05 for Ours vs. all the compared models). It indicates that our EBAT framework outperforms BAT-Fill as well as the other compared models, on all the metrics relevant to either pixel-level accuracy or perceptual quality.

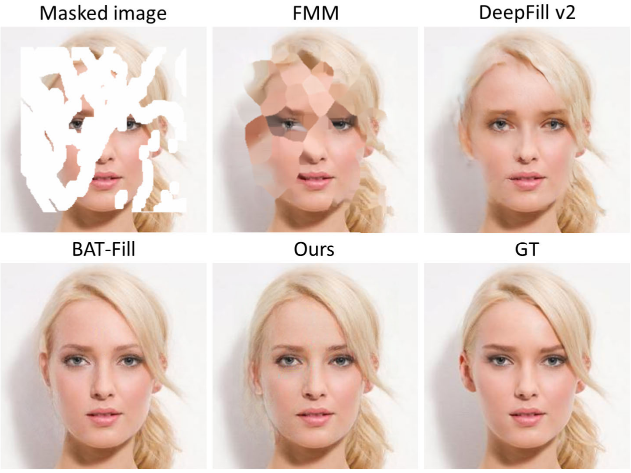

FIGURE 5. Qualitative comparison of the proposed framework with the state-of-the-art image inpainting methods over CelebA-HQ with large masks. The images generated by both our framework and BAT-Fill are more photo-realistic compared to FMM and DeepFill v2, but no significant differences are found between two transformer-based methods.

We measured both the number of trainable parameters and the inference time for the proposed method on a single NVIDIA RTX A6000 GPU and compared them with FMM and BAT-Fill. The average inference time on our proposed framework was 205 ms for one image sample, which is 1.5 times slightly slower than FMM (of 136 ms), while the inference time on BAT-Fill was 187.1 times

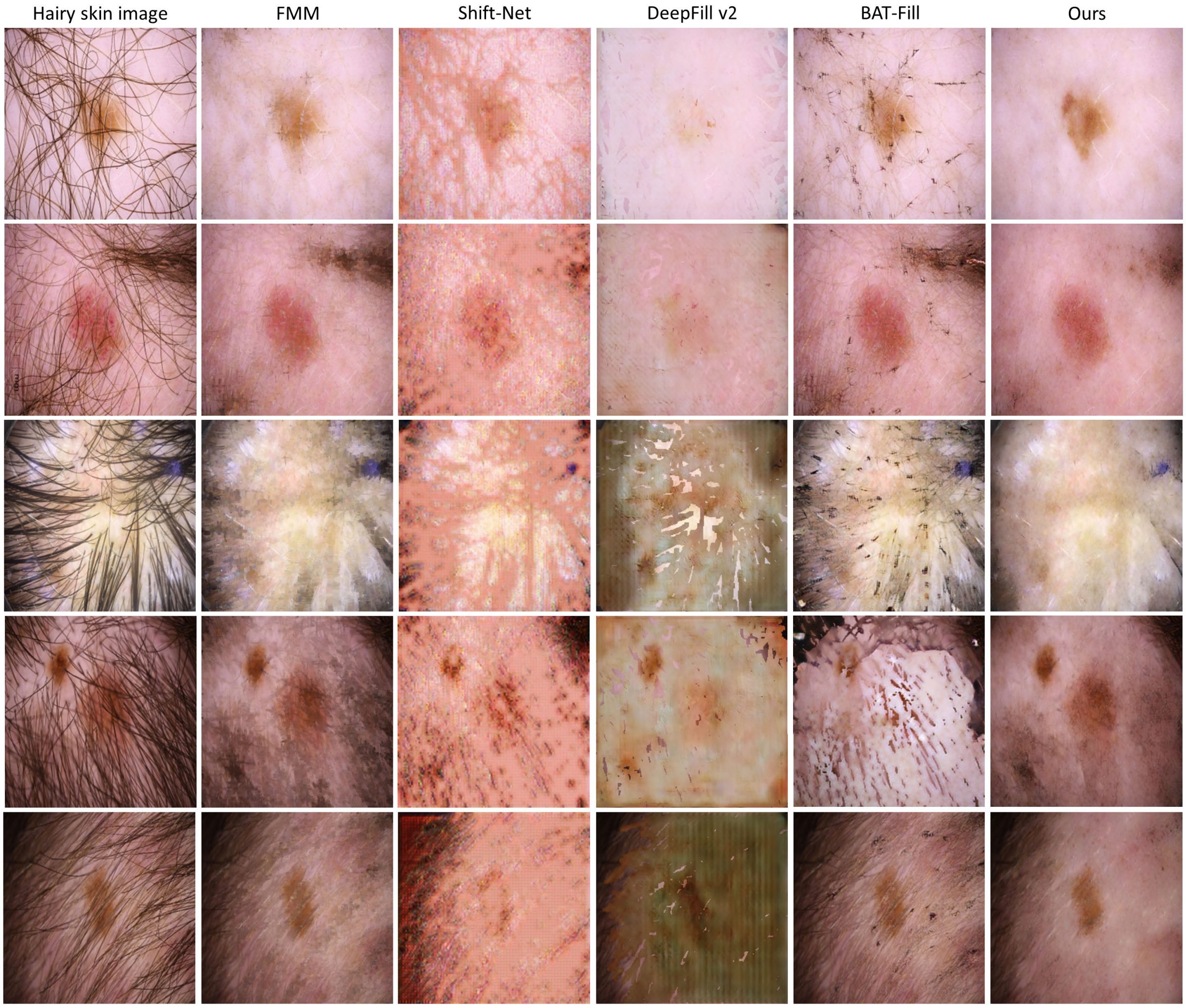

FIGURE 6. Qualitative comparison of the proposed framework with the state-of-the-art image inpainting methods for removing hairs over ISIC 2020 hairy skin images. FMM, Shift-Net, and DeepFill v2 suffer from the blurring artifact and a number of noticeable hair marks, similar with Figure 4 tested over simulated images. On the other hand, our proposed method removed most hairs in hairy skin images and accurately reconstructed minute texture details and global structure of skin lesions.

extremely slower with 38,364 ms on average. Given the fact that the number of trainable parameters is 3.55 times larger in the proposed framework (361.1M) compared to BAT-Fill (101.7M as the sum of 77.2M for the diverse structure generator and 24.5M for the texture generator), it is extraordinary that our method is much faster than BAT-Fill. The sluggish processing time might originate mainly from the pixel-unit repetitive loops, in the top-K sampling process for diverse creation of missing regions, which sample the most likely pixel values from the outputs predicted from BAT.

C. QUALITATIVE EVALUATION

Figure 5 shows the visual comparison between our proposed framework and the compared image inpainting methods over the test dataset of CelebA-HQ. At first glance it is formidable to discriminate the visual differences in generated images between our method and BAT-Fill while both transformer-based methods are superior to DeepFill v2. On the other hand, it can be observed, through such delicate details as hair color and eyebrow, that the image generated by our method is a bit more consistent with the ground truth compared to the other methods.

Figure 4 shows the qualitative comparison of removing hairs over the simulated hairy skin images with high hair density (30%) which were generated on skin cancer lesions including nevus as described in Section IV-A3. In Figure 6, we also showed the qualitative results of removing hairs over authentic dermoscopic images with either sparse hairs or dense hairs. Our method exhibited not only more enhanced accuracy in reconstructing fine texture details and global structure of skin and lesions but also better perceptual quality in synthesizing photo-realistic images, compared to the other methods whose image quality was noticeably deteriorated by blurry and deformed textures as well as incompletely erased hair stains.

TABLE 3. Results of user study. Each entry is a mean opinion score (MOS) for 26 image samples generated using the given methods. The score is ranging from 1 (for the worst) to 5 (for the best), and the best score was denoted in bold. A standard deviation was parenthesized under the corresponding mean.

Method	FFM	ShiftNet	DeepFill	BAT-Fill	Ours
MOS	3.42 (1.21)	2.11 (1.07)	3.78 (1.19)	3.74 (1.12)	**4.18** (0.91)

To evaluate the human-level perceived image quality without references, we obtained the mean opinion score (MOS) by asking 30 participant observers (24 males and 6 females in their 20s) to assess the quality of 26 given images with a score ranging from 1 (worst) to 5 (excellent). The sample images used for the MOS acquisition were randomly chosen from the output images generated over authentic dermoscopic images with a large number of hairs. As summarized in Table 3, our method had the highest average MOS ratings (4.18) which are proven statistically significant compared to all the other methods (p-value < 0.05 for Ours vs. all the other methods).

D. ABLATION STUDY

To figure out the effects of CNN and transformer backbones in the multi-scale feature extractor on the performance of removing hairs, we conducted ablation studies by removing one of two backbones. We compared our method with two ablations: (i) with CNN backbone only and (ii) with transformer backbone only. The first ablation model (i) would be seen as a network architecture similar to the U-net where the CNN backbone and the fine texture generator behave as the encoding and decoding parts respectively [31]. By contrast, the convolutional image features from the CNN backbone are no longer exploited in the fine texture generator in the second ablation model (ii), instead they are replaced with the multi-scale feature maps produced through the transformers.

As shown in Figure 7, the output images were distinctly degraded when using both ablation models. Hairs remain incompletely eliminated which resulted in deteriorating the skin and lesion textures due to hair steins, although the CNN backbone and the transformer backbone seemed to be relatively better in capturing the global structure information and the fine texture details with long-range dependency respectively. On the other hand, the proposed method where both CNN and transformer backbones are integrated was superior to both ablation models in reconstructing both global structure as well as fine-grained textures.

The qualitative analysis for the ablation models is consistent with the quantitative evaluation results over simulated skin images as summarized in Table 4. The proposed method obviously improved all the metrics even in the case of high hair density, which demonstrates the advantage of integrating both CNN and transformer backbones to extract a comprehensive set of image features relevant to global structure and fine-grained textures with long range dependency that cannot be readily captured by convolutional layers.

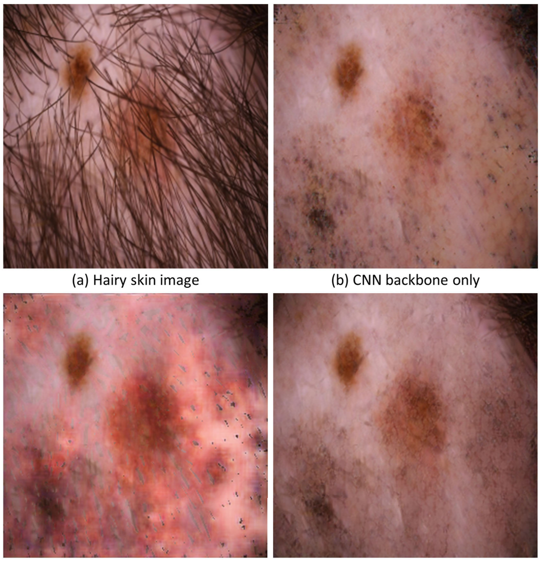
(a) Hairy skin image
(b) CNN backbone only
(c) Transformer backbone only
(d) CNN + Transformer backbones

FIGURE 7. Visual comparison of the proposed method with its ablations for verifying the effect of integrated transformer and CNN backbones. The use of integrated transformer and CNN backbones in the multi-scale feature extractor results in the enhanced performance in removing hairs and reconstructing the fine textures of skin and lesions.

TABLE 4. Quantitative comparison of our proposed framework with its ablations over a hair-simulated image dataset from ISIC 2020 hairless dermoscopic images. The ablation study was conducted for coarse hairs (of mask ratio 10%) or dense hairs (30%). A standard deviation was parenthesized under the corresponding mean, and the best performance was denoted in bold.

Methods	PSNR↑	SSIM↑	FID↓	LPIPS↓
Coarse hairs (10%)				
CNN only	41.41 (2.21)	0.957 (0.017)	47.67	0.039 (0.024)
Trans only	40.64 (2.25)	0.951 (0.019)	53.35	0.057 (0.032)
CNN + Trans	**43.07** (2.49)	**0.969** (0.014)	**17.33**	**0.020** (0.009)
Dense hairs (30%)				
CNN only	36.86 (1.40)	0.889 (0.034)	63.60	0.084 (0.026)
Trans only	35.40 (1.47)	0.873 (0.037)	77.89	0.131 (0.054)
CNN + Trans	**38.40** (1.77)	**0.916** (0.031)	**42.65**	**0.051** (0.017)

V. CONCLUSION

In this study, we present a novel transformer-based generative image inpainting framework, called EBAT, that achieves accurate and realistic reconstruction of skin texture by removing hairs from dermoscopic images. Our model is able to learn both coarse structure and fine skin textures by extracting image features with long-range dependency in multiple spatial scales using dual modalities of bidirectional autoregressive transformers and convolutional neural networks. The multi-scale feature extraction from transformers is facilitated by patch-wise unfolding and down-sized folding operations. Our framework is efficiently trained in an end-to-end manner through manifold pathways laid between multi-scale feature extractor and fine texture generator. The experimental results on removing hairs in both simulated and authentic hairy dermoscopic images show that, in qualitative and quantitative performance of removing hairs from dermoscopic images, our extensive transformer-based image inpainting framework outperforms not only the state-of-the-art image inpainting models but also the latest transformer-based method like BAT-Fill.

Despite the novelties and improved applicability of our model to hairy skin images, it deserves to mention its technical limitations. It is still challenging to remove hairs laid on skin lesions. We find that our method sometimes fails in removing hairs whose color is similar to either skin lesions or skin. The method should be improved to learn diverse features of skin lesions.

As a future work, it is imperative to figure out the potential effects of hair removal on skin cancer classification. The hair removal algorithm needs to be advanced toward reducing the vulnerability of skin cancer classification to hair removal, taking into account the risk such that the texture of skin lesion might be tainted by the process of removing hairs.

ACKNOWLEDGMENT

The authors thank Dr. Soo Ye Kim from Adobe Research, Dr. Jaeyoung Choi from Hankuk University of Foreign Studies, Dr. Bumshik Lee from Chosun University, and Hyunjae Zhang from F&D Partners Corporation, for insightful discussions on image inpainting and its industrial demands, and Bumjin Park and Gyubin Lee from AIIP Lab, for assisting data processing.

REFERENCES

[1] O. T. Jones, R. N. Matin, M. van der Schaar, K. P. Bhayankaram, C. K. I. Ranmuthu, M. S. Islam, D. Behiyat, R. Boscott, N. Calanzani, J. Emery, H. C. Williams, and F. M. Walter, "Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: A systematic review," *Lancet Digit. Health*, vol. 4, no. 6, pp. e466–e476, Jun. 2022.

[2] J. A. Salido, "Hair artifact removal and skin lesion segmentation of dermoscopy images," *Asian J. Pharmaceutical Clin. Res.*, vol. 11, no. 15, p. 36, Oct. 2018.

[3] L. Talavera-Martinez, P. Bibiloni, and M. Gonzalez-Hidalgo, "Hair segmentation and removal in dermoscopic images using deep learning," *IEEE Access*, vol. 9, pp. 2694–2704, 2021.

[4] J. A. A. Salido and C. Ruiz, "Using morphological operators and inpainting for hair removal in dermoscopic images," in *Proc. Comput. Graph. Int. Conf.*, Jun. 2017, pp. 1–6.

[5] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "Dullrazor®: A software approach to hair removal from images," *Comput. Biol. Med.*, vol. 27, no. 6, pp. 533–543, Nov. 1997.

[6] K. Kiani and A. R. Sharafat, "E-shaver: An improved Dullrazor® for digitally removing dark and light-colored hairs in dermoscopic images," *Comput. Biol. Med.*, vol. 41, no. 3, pp. 139–145, Mar. 2011.

[7] F. Bornemann and T. März, "Fast image inpainting based on coherence transport," *J. Math. Imag. Vis.*, vol. 28, no. 3, pp. 259–278, 2007.

[8] F.-Y. Xie, S.-Y. Qin, Z.-G. Jiang, and R.-S. Meng, "PDE-based unsupervised repair of hair-occluded information in dermoscopy images of melanoma," *Comput. Med. Imag. Graph.*, vol. 33, no. 4, pp. 275–282, 2009.

[9] A. Huang, S.-Y. Kwan, W.-Y. Chang, M.-Y. Liu, M.-H. Chi, and G.-S. Chen, "A robust hair segmentation and removal approach for clinical images of skin lesions," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 3315–3318.

[10] Q. Abbas, I. F. Garcia, M. Emre Celebi, and W. Ahmad, "A feature-preserving hair removal algorithm for dermoscopy images," *Skin Res. Technol.*, vol. 19, no. 1, pp. e27–e36, Feb. 2013.

[11] A. Nasonova, A. Nasonov, A. Krylov, I. Pechenko, A. Umnov, and N. Makhneva, "Image warping in dermatological image hair removal," in *Proc. Int. Conf. Image Anal. Recognit.*, 2014, pp. 159–166.

[12] D. Borys, P. Kowalska, M. Frackiewicz, and Z. Ostrowski, "A simple hair removal algorithm from dermoscopic images," in *Proc. Int. Conf. Bioinf. Biomed. Eng.*, 2015, pp. 262–273.

[13] J. Koehoorn, A. Sobiecki, P. Rauber, A. Jalba, and A. Telea, "Effcient and effective automated digital hair removal from dermoscopy images," *Math. Morphol.-Theory Appl.*, vol. 1, no. 1, pp. 1–17, Mar. 2016.

[14] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Hair detection and lesion segmentation in dermoscopic images using domain knowledge," *Med. Biol. Eng. Comput.*, vol. 56, no. 11, pp. 2051–2065, Nov. 2018.

[15] I. Zaqout, "An efficient block-based algorithm for hair removal in dermoscopic images," *Comput. Opt.*, vol. 41, no. 4, pp. 521–527, 2017.

[16] K. Zafar, S. O. Gilani, A. Waris, A. Ahmed, M. Jamil, M. N. Khan, and A. S. Kashif, "Skin lesion segmentation from dermoscopic images using convolutional neural network," *Sensors*, vol. 20, no. 6, p. 1601, Mar. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/6/1601

[17] J. Koehoorn, A. C. Sobiecki, D. Boda, A. Diaconeasa, S. Doshi, S. Paisey, A. Jalba, and A. Telea, "Automated digital hair removal by threshold decomposition and morphological analysis," in *Proc. Int. Symp. Math. Morphol. Appl. Signal Image Process.*, 2015, pp. 15–26.

[18] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-Net: Image inpainting via deep feature rearrangement," in *Proc. Int. Symp. Math. Morphology. Appl. Signal Image Process.*, 2018, pp. 3–19.

[19] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 331–340.

[20] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 89–105.

[21] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.

[22] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8857–8866.

[23] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Process. Lett.*, vol. 51, no. 2, pp. 2007–2028, 2019.

[24] D. Bardou, H. Bouaziz, L. Lv, and T. Zhang, "Hair removal in dermoscopy images using variational autoencoders," *Skin Res. Technol.*, vol. 28, no. 3, pp. 445–454, May 2022.

[25] W. Li, A. N. Joseph Raj, T. Tjahjadi, and Z. Zhuang, "Digital hair removal by deep learning for skin lesion segmentation," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107994.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[28] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 69–78.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, pp. 139–144, Oct. 2020.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, W. M. W. Joachim and F. A. Nassir, and J. Hornegger, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[32] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

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

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. Adv. Neural Inf. Process. Syst.* 1 2017, pp. 1–15.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[37] J. Zhuang, "LadderNet: Multi-path networks based on U-Net for medical image segmentation," 2018, *arXiv:1810.07810*.

[38] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[39] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, p. 34, Jan. 2021.

[40] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH*, 2003, p. 313.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[42] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[44] 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. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[45] 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.

YOUNGCHAN LEE received the B.S. degree in information and communications engineering from Sun Moon University, in 2022. He is currently pursuing the M.Sc. degree in information and communications engineering with the Artificial Intelligence and Image Processing Laboratory (AIIP Lab). His research interests include image inpainting, computer vision, deep learning, and medical imaging.

WONSANG YOU (Member, IEEE) was born in Seoul, South Korea, in 1977. He received the M.Sc. degree in engineering (specialized in computer vision and image processing) from the Korea Advanced Institute of Science and Technology (KAIST), in 2008, and the Ph.D. degree in electrical engineering and information technologies (specialized in brain imaging data analysis) from Otto-von-Guericke University Magdeburg, in 2013. From 2009 to 2012, he worked as a Research Assistant with the Leibniz Institute for Neurobiology, Magdeburg, Germany. From 2013 to 2019, he worked as a Staff Scientist with the Center for the Developing Brain in Children's National Hospital, Washington, DC, USA. He is currently an Assistant Professor with the Department of Information and Communication Engineering, Sun Moon University, South Korea. He is also the Director of the Artificial Intelligence and Image Processing Laboratory (AIIP Lab). His research interests include computer vision, image analysis, deep learning, and medical imaging. He is also working on image inpainting, super resolution, pose estimation, 2D-to-3D reconstruction, and image-to-text generation.