IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 December 2023, accepted 11 January 2024, date of publication 18 January 2024, date of current version 30 January 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3355542*

RESEARCH ARTICLE

# Fine-Grained Human Hair Segmentation Using a Text-to-Image Diffusion Model

**DOHYUN KIM[1], EUNA LEE[2], DAEHYUN YOO[1], AND HONGCHUL LEE[1]**
[1]School of Industrial Management Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Center for Defense Resource Management, Korea Institute for Defense Analyses, Seoul 02455, Republic of Korea

Corresponding author: Hongchul Lee (hclee@korea.ac.kr)

This work was supported in part by the BK21 FOUR funded by the Ministry of Education of Korea, and in part by the National Research Foundation of Korea.

**ABSTRACT** Human hair segmentation is essential for face recognition and for achieving natural transformation of style transfer. However, it remains a challenging task due to the diverse appearances and complex patterns of hair in image. In this study, we propose a novel method utilizing diffusion-based generative models, which have been extensively researched in recent times, to effectively capture and to finely segment human hair. In diffusion-based models, an internal visual representation during the denoising process contains pixel-level rich information. Inspired by this aspect, we introduce diffusion-based models for segmenting fine-grained human hair. Specifically, we extract the representation from the diffusion-based models, which contains pixel-level semantic information, and then train a segmentation network using it. Particularly, to more finely segment human hair, our approach employs the representation from a text-to-image diffusion model, conditioned on text information, to extract more relevant information for human hair, thereby predicting detailed hair masks. To validate our method, we conducted experiments on three distinct hair-related datasets with unique characteristics: Figaro-1k, CelebAMask-HQ, and Face Synthetics. The experimental results show the improved performance of our proposed method across all three datasets, outperforming existing methods in terms of mIoU (mean intersection over union), accuracy, precision, and F1-score. This is particularly evident in its ability to accurately capture and finely segment human hair from background and non-hair elements. This demonstrates the effectiveness of our method in accurately and finely segmenting human hair with complex characteristics. Our research contributes not only to the fine-grained segmentation of human hair but also to the application of generative models in semantic segmentation tasks. We hope that the proposed method will be applied for detailed semantic segmentation in various fields in the future.

**INDEX TERMS** Hair segmentation, fine-grained segmentation, generative model, diffusion model, text-to-image diffusion model, Figaro-1k, CelebAMask-HQ, face synthetics.

## I. INTRODUCTION

Image segmentation is a fundamental task in computer vision that classifies pixels within an image based on target objects. Among various segmentation tasks, human hair segmentation in images has been utilized not only in facial and gender recognition, but also in recent applications such as face editing and style simulation [1], [2], [3]. Although segmentation of general objects has seen significant progress, human hair segmentation remains one of the more challenging image segmentation tasks due to its diverse shapes and appearances [4]. In particular, as shown in Fig. 1, human hair exhibits fine and complex characteristics that make it difficult to distinguish hair accurately from background and non-hair elements. Recent studies [3], [5], [6] have aimed to perform human hair segmentation using multiple levels of feature maps based on the convolutional neural network (CNN). Additionally, CNN-based methods often require additional post-processing modules to generate more precise hair masks [6]. Nevertheless, most existing

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


**FIGURE 1.** Human hair segmentation.

methods still suffer from inaccurate hair boundaries and unsatisfactory segmentation results. In this study, our goal is to more finely segment human hair. We do this by using a novel approach that utilizes representations computed by diffusion-based generative models. These representations contain valuable pixel-level semantic information, which is obtained during the image generation process. We train a segmentation network with these representations, enabling it to achieve more detailed segmentation of human hair.

Significant advancements have been achieved with regard to the semantic segmentation, leading to more precise segmentation results. Most methods [7], [8], [9] based on encoder–decoder network architecture utilized both coarse and fine representations to capture more detailed image features. Furthermore, both [10] and [11] exploit multi-scale features to enhance the contextual understanding in image processing. These methods have been applied in areas requiring fine-grained segmentation, such as anomaly detection [12] and biomedical applications [13], [14], [15]. Subsequently, Vision Transformer [16] (ViT) successfully applied the transformer model [17], which was initially proposed for machine translation, focusing on the attention to the entire context of text in natural language processing (NLP) research, to the field of computer vision (CV), thereby facilitating the consideration of the entire context of images for image segmentation [18]. Alternatively, segmentation research has also leveraged the advantages of the image generation process using generative models [19], [20], [21], [22]. In particular, generative models based on the diffusion process [23], which gradually adds Gaussian noise to an original image and then iteratively removes it to generate the image, have recently shown revolutionary advances in various image generation tasks [24], [25], [26], [27]. According to [21], denoising diffusion probabilistic models (DDPM) [24] capture excellent pixel-level semantic information of the inputted image, thereby providing a valuable semantic representation for segmentation tasks. Additionally, the representation in diffusion-based generative models containing pixel-level dense and rich semantic information have been utilized to perform panoptic segmentation [22]. Inspired by these studies, we aim to utilize a diffusion model that would be superior at capturing pixel-level semantic information to perform fine-grained hair segmentation.

In this study, we propose a novel human hair segmentation method using a diffusion-based generative model, aimed at achieving fine-grained results for intricate and complex hair patterns. Specifically, we extract the internal visual representation from the diffusion-based generative model. Particularly, we utilize a text-to-image diffusion model, conditioned by text, to obtain hair-related representations. By conditioning on the word `'hair'` during the denoising process, we obtain pixel-level representations with enhanced semantic information for hair-related pixels in image. Subsequently, a segmentation network is trained using these representations to perform fine-grained human hair segmentation. To evaluate the effectiveness of the proposed method in segmenting human hair under various conditions, we conducted experiments on three publicly available datasets: Figaro-1k [28], CelebAMask-HQ [29], and Face Synthetics [30]. Figaro-1k requires capturing and separating hair in various environments, while the CelebAMask-HQ involves separating hair from non-hair elements like various accessories. By contrast, Face Synthetics is characterized by its synthetic facial data, which provide more detailed annotations. The proposed method both captured hair effectively and performed fine-grained segmentation of hair from various non-hair elements and background. Specifically, our method achieved significant improvements over existing methods. On Figaro-1k, it achieves mean intersection over union (mIoU) improvement of 4.2% points and 3.1% point rise in F1-score. For CelebAMask-HQ, the model improves mIoU by 1.8% points and F1-score by 0.2% points. On Face Synthetics, it reports mIoU improvement of 1.5% points, 1.2% point accuracy gain, along with 0.9% point increase in precision. Furthermore, in qualitative experiments conducted on the three datasets, the proposed method exhibited more detailed boundary segmentation of human hair compared to existing methods. Particularly, it was effective in capturing complex hair patterns. These results have established that the proposed method, which utilizes pixel-level representations in diffusion-based generative models, predicts more detailed human hair masks. The contributions of this study are summarized as follows:

- We introduce a novel approach for fine-grained human hair segmentation using diffusion-based generative models, which utilize pixel-level semantic information for precise hair segmentation.
- We develop a method that utilizes a text-to-image diffusion model, conditioned with the text `'hair'`, for extracting hair-focused features, which are then used to train a segmentation network.
- We achieve segmentation of human hair by using only pixel-level semantic representation without additional refine modules.
- We conduct extensive experiments on fine-grained hair segmentation by using three distinct publicly available hair-related datasets, which include a wide range of unique characteristics. Our method achieved state-of-the-art performance in fine-grained hair segmentation, demonstrating the effectiveness of our method.

The remainder of this paper is organized as follows. Section II discusses previous research on semantic

segmentation, segmentation using diffusion-based generative models, and human hair segmentation. Section III introduces the proposed method, and Section IV presents the quantitative and qualitative experimental results. Finally, Section V concludes the work and suggests possible directions for further research in other areas.

## II. RELATED WORK

### A. SEMANTIC SEGMENTATION

Semantic segmentation is a task of image segmentation that involves the semantic classification of pixels in an image. There have been significant advances in semantic segmentation in recent years. With the success of CNN on several CV tasks, it has been applied to various image segmentation tasks. Fully convolutional neural networks [7] effectively combined coarse and fine image features using a convolutional-only architecture, facilitating accurate semantic segmentation. UNet [8], another method for medical image segmentation, leveraged a symmetric encoder–decoder architecture to capture contextual information and enable precise localization.

As regards fine-grained segmentation, several studies based on CNN have been developed to capture more detailed image features. Encoder–decoder-based methods, such as SegNet [9], utilize low- and high-level information at multiple encoding stages during decoding to achieve more finely detailed segmentation results. DeepLabV3+ [11] employs Atrous Spatial Pyramid Pooling along with atrous convolution. This facilitates the extraction of dense features from the image and allows for a broader understanding of the image context. It was used to achieve more precise object boundaries. These methods using multi-scale features and encoder–decoder architecture have been mainly applied to the biomedical field because of their ability in capturing detailed representations of the image. Thus, research continues to focus on realizing precise segmentation [31], [32]. Furthermore, an adaptive feature fusion module has been proposed to enhance segmentation performance [33], and a boundary-aware context neural network for capturing contextual information has been proposed to achieve improved segmentation results with detailed object boundaries [34].

Meanwhile, recently, ViT [16] successfully adapted the Transformer architecture [17], originally proposed for machine translation task in NLP, to the field of CV. This approach allows the extraction of image features without spatial information compression, capturing global context in a way that contrasts with the local focus of CNN-based methods. Inspired by success of ViT, it has been widely utilized for feature extraction to various downstream tasks of CV, such as object detection and segmentation [35]. [36], [37], [38] employed the transformer decoding process, which is based on an attention mechanism among multi-scale features, to perform precise segmentation with the global context of the image. In this study, we propose a method for fine-grained semantic segmentation. We leverage the advantages of existing generative models, instead of using the commonly employed discriminative models for segmentation tasks. Particularly, we utilize diffusion-based generative models and suggest a method where semantic segmentation is conducted by extracting and utilizing features computed during the image re-generation process.

### B. DIFFUSION-BASED GENERATIVE MODELS FOR SEGMENTATION

In recent years, diffusion-based models have been applied to a wide variety of image generative tasks, such as image generation [24], [26] and image editing [39], because of their improved ability to generate high-quality images compared with previous methods. Latent Diffusion Models (LDM) [40] employed the latent space of autoencoders for pixel representation, which allows for a reduction in complexity and the preservation of details. Moreover, diffusion-based models can also be useful in image segmentation tasks. SegDiff [41] extended method of diffusion probabilistic models to image segmentation. Reference [21] demonstrated that diffusion-based models can be used in semantic segmentation using only a few labels. In particular, they showed that visual features, extracted in the denoising stage, contain excellent pixel-level semantic information. Additionally, ODISE [22] successfully performed panoptic segmentation using diffusion-based models; they also used the visual features of pixel-level rich semantic information extracted through these models. Motivated by these advantages of diffusion-based generative models, we focus on fine-grained human hair segmentation to leverage the visual features obtained from these models, which contain rich pixel-level semantic information.

### C. HUMAN HAIR SEGMENTATION

Many studies for human hair segmentation have impacted various face-related tasks, such as facial recognition and style translation. Particularly, hair segmentation is required for successful style transformation in the field of face-centered style transfer, such as face translation [42] and face editing [43], [44]. Early research focused primarily on the prediction of hair masks by only using color, spatial, and frequency information [45], [46], [47], [48]. However, this approach was limited by the spatial location of human hair within the images, which required additional image preprocessing. To address the limitations, [49] employed a neural network to classify image pixels that represent human hair. In particular, remarkable success has been achieved using deep CNN; [50], [51], [52] utilized deep CNN to learn various characteristics of human hair for automatic hair segmentation. Recently, some approaches [3], [5], [53] have been based on encoder–decoder network architectures to utilize multi-scale features for predicting finely detailed hair masks. In a subsequent study [6], a border refinement module was added to enhance hair segmentation and refine the details of hair borders. However, most of these existing methods

still suffer from inaccuracies in hair boundary detection and detail. They also require additional pre- or post-processing stages. In this study, we aim to perform fine-grained human hair segmentation using only the visual features that contain pixel-level semantic information extracted through the diffusion-based generative model without the need for additional processing.

## III. PROPOSED METHOD

The framework of the proposed method is shown in Fig. 2. It consists of the following stages. The first stage is feature extraction, during which we extract internal visual features in the reverse process from a pre-trained text-to-image diffusion model. The second stage involves segmentation, where a segmentation network is trained using the extracted visual features to predict human hair masks. Our goal is to predict a fine-grained human hair mask using the internal visual features of the diffusion-based generative models.

In the next subsection, we provide a brief overview of diffusion probabilistic models. We then describe feature extraction with pixel-level semantic information and the segmentation network that utilizes this information to predict the hair masks.

### A. BACKGROUND

We consider the diffusion probabilistic models proposed by [23] and [24]. Diffusion-based generative models generate new images by progressively corrupting an original image with Gaussian noise over a series of steps, and then creating a new image from the corrupted image by iteratively predicting and removing the added Gaussian noise at each step, as shown in Fig. 3.

The forward process gradually corrupts $x_0$ sampled from the original data distribution $q(x)$ by adding Gaussian noise. Specifically, Gaussian noise with variance schedule $\beta_t$ is added to $x_{t-1}$ for each $t \in T$ steps, which can be defined as in

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

The resulting variable $x_t$ can be expressed as in

$$x_t = \sqrt{1-\alpha_t}\epsilon + \sqrt{\alpha_t}x_0. \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\alpha_t = \prod_{i=1}^{T}(1-\beta_i)$. The corrupted image $x_T$ approximately follows a standard normal distribution with a mean of 0 and a variance of 1, with each pixel being independently and identically distributed. Subsequently, the reverse process is the procedure of denoising from $x_T$ back to $x_0$. The model, under parameter $\theta$, predicts the mean and standard deviation of the added noise at each step, which can be defined as in

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

The model learns its parameters through the diffusion process while reconstructing the original image, which involves learning the distribution of the data. During inference, a new image is generated by predicting and removing noise from the learned distribution at each step.

Based on the diffusion process, recent text-to-image generative models [27], [40], [54] generate images that correspond to text descriptions of target images. Specifically, given an initial Gaussian noise and a text embedding extracted from a text encoder like CLIP [55], the model generates a new image that aligns with the text description through the reverse process.

### B. FEATURE EXTRACTION

As aforementioned, diffusion-based models generate new images from Gaussian noise through reverse processing. In text-to-image diffusion models, such as LDM, a UNet architecture with a symmetric encoder–decoder structure is utilized for denoising. Additionally, the denoising process is guided by employing cross-attention between visual features and text embeddings to generate images corresponding to the desired text descriptions. Notably, the visual features computed during denoising correlate strongly with rich, semantically meaningful text descriptions.

For fine-grained human hair segmentation, we focus on extracting the visual features that contain hair-focused semantic information. Therefore, we use the text-to-image diffusion model, which performs the diffusion process conditionally based on textual information about hair. Specifically, we extract the internal visual features during the reverse process in the text-to-image diffusion models, which takes image-text pairs $(x, w)$ as input. The model first performs a forward process on the original image data point $x_0$ by adding Gaussian noise over $T$ steps. The corrupted data $x_T$ are then denoised via the reverse process. We focus specifically on visual features correlated with pixels of hair parts. To achieve this, we condition the text input using the word `'hair'` in $w$ during the reverse process. For text conditioned denoising, the model $f_\theta$ employs a UNet architecture with cross-attention (Diffusion UNet). Using the text embedding extracted from the text encoder as $K$ and $V$, and the visual features as $Q$, the attention operation is performed as expressed in (4).

$$CrossAttention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

This process is computed over T steps, and at each step, the visual features that capture the relevance between visual and textual information are computed. At this point, we extract the visual features from the Diffusion UNet at the last step. As shown in Fig. 4, we take the visual features at different scales from the Diffusion UNet; this enables the gathering of features encapsulating various levels of semantic information. The extracted features are expressed in

$$w_{emb} = TextEncoder(w), \tag{5}$$

$$r = f_\theta(x_t, w_{emb}). \tag{6}$$

The extracted features $r$ contain pixel-level semantic information correlated to the word `'hair.'` As a result,

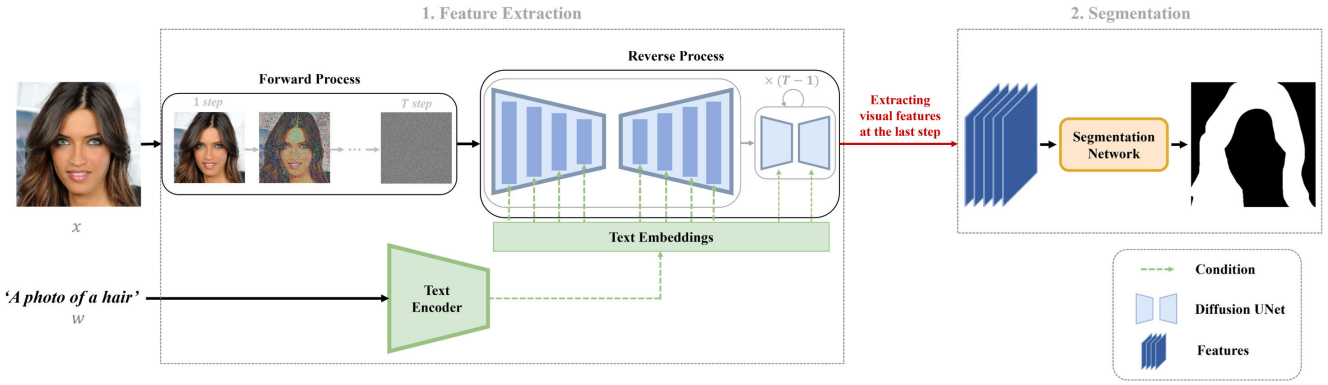


**FIGURE 2.** Framework of proposed method.

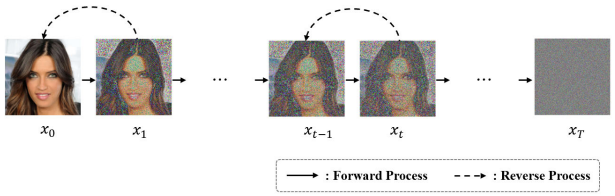


**FIGURE 3.** Diffusion process.

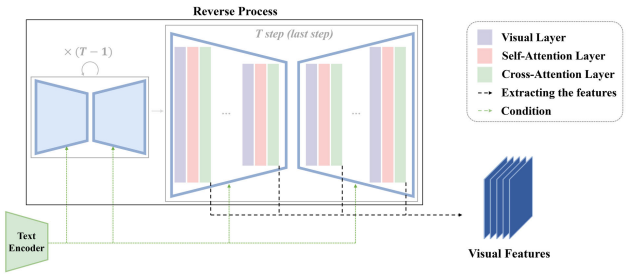


**FIGURE 4.** Feature extraction from Diffusion UNet during reverse process.

the features $r$ that contain more semantic details regarding the pixels representing hair can be obtained. Utilizing the features $r$, we can train the segmentation network to produce a more finely detailed human hair mask.

### C. SEGMENTATION NETWORK

We train the segmentation network by leveraging the features $r$ extracted through the text-to-image diffusion model. Since the features inherently contain pixel-level semantic information related to the hair, their use enables our segmentation network to achieve more finely detailed predictions. These features can also be effectively utilized with direct segmentation mask-based (query-based) models [37], [38]. These models predict segmentation masks by employing an initial query, which is decoded with visual features. This method aligns well with our approach, where the initial query interacts with the extracted features for accurate image segmentation. However, focusing exclusively on hair among various objects can lead to pixel imbalances. Such imbalances could potentially hinder both the training process and the accuracy of predictions. Therefore, it is crucial to select a model that takes these considerations into account. To address this problem, in this study, we utilize Mask2Former [38], which employs binary cross-entropy and dice loss functions.

## IV. EXPERIMENTAL RESULTS

### A. ARCHITECTURE

In these experiments, we utilized the latent diffusion model v1-3 [40] as a text-to-image diffusion generative model and Mask2Former [38] as the segmentation network model. In the feature extraction stage, we paired the image with the word `'hair'` and performed denoising during the reverse process. Text embeddings were extracted using CLIP text encoder [55], by inputting the word `'hair'` into the prompt `'A photo of a hair.'` We maintained all the stages of the diffusion process in a frozen state and trained only the segmentation network. Additionally, because the optimal performance was observed at timestep $T = 0$, we report the results observed at this timestep.

Following [6], we compared our results with those of UNet [8] and DeeplabV3+ [11], which use a pre-trained ResNet [56] as the backbone. Additionally, we included Mask2Former, which uses both a pre-trained ResNet and Swin transformer [57] as the backbone, as an additional comparison method.

### B. DATASETS

To verify the effectiveness of the proposed method, we used three public datasets: Figaro-1k [28], CelebAMask-HQ [29], and Face Synthetics [30] for our experiments. Fig. 5 shows an example of each dataset. Figaro-1k is a hair analysis dataset composed of 1,050 images of various forms of hair in everyday scenarios, as shown in (a). CelebAMask-HQ is a large-scale facial image dataset used for various face-related tasks; we used only the hair annotations. This dataset consists of 30,000 face images, many of which have celebrities wearing various accessories that partially cover their hair,

**TABLE 1.** Comparison of quantitative results of models applied to the Figaro-1k dataset.

| Method | Backbone | mIoU (%) | Accuracy (%) | Precision (%) | F1-score (%) |
|---|---|---|---|---|---|
| UNet [8] | ResNet18 | 89.69 | 96.36 | 95.91 | 95.63 |
| | ResNet50 | 89.40 | 93.69 | 91.91 | 90.94 |
| DeepLabV3+ [11] | ResNet50 | 95.08 | 97.35 | 97.61 | 97.02 |
| Mask2Former [38] | ResNet50 | 91.34 | 96.20 | 96.86 | 97.09 |
| | SwinTransformer | 91.03 | 96.08 | 96.90 | 96.74 |
| Yan et al. [6] | - | 91.25 | 97.23 | 97.33 | 95.15 |
| Ours | - | **95.40** | **98.03** | **98.24** | **98.28** |


**FIGURE 5.** Examples from the Figaro-1k, CelebAMask-HQ, and Face Synthetics datasets.

as shown in (b). Face Synthetics is a diverse face synthesis image dataset that can be used for various face-related tasks. It is composed of 100,000 face images and is characterized by the detailed depiction of faces and hair, as shown in (c). Therefore, this dataset has finer annotations at the pixel-level, and we used only the hair annotations.

### C. EVALUATION METRICS

For quantitative evaluation, we adopted the mean values of the intersection over union (IoU), accuracy, precision, and F1-score. The IoU represents the degree of overlap between the ground truth mask and the predicted mask, whereas the accuracy denotes the number of pixels correctly classified. The precision indicates the accuracy of positive predictions. Finally, the F1-score was calculated by combining precision and recall, providing a balanced measure of both. These quantities are described by Equations (7), (8), (9), and (10), respectively, and were used to conduct the evaluations. In these equations, $p_{cls}$ denotes the number of segmentation classes, $p_{ij}$ represents the number of pixels predicted as class $j$ but in fact belonging to class $i$, and $t_i$ denotes the number of pixels belonging to class $i$.

$$mIoU = \frac{1}{p_{cls}} \sum_i \frac{p_{ii}}{t_i + \sum_j p_{ij} - p_{ii}} \tag{7}$$

$$\text{Accuracy} = \frac{1}{p_{cls}} \frac{\sum_i p_{ii}}{\sum_i t_i} \tag{8}$$

$$\text{Precision} = \frac{1}{p_{cls}} \sum_i \frac{p_{ii}}{\sum_j p_{ij}} \tag{9}$$

$$\text{F1-score} = \frac{1}{p_{cls}} \sum_i \frac{2 \cdot p_{ii}}{2 \cdot p_{ii} + \sum_j p_{ij} + \sum_j p_{ji}} \tag{10}$$

### D. QUANTITATIVE AND QUALITATIVE RESULTS

In this section, we present both the quantitative results and visual outcomes of the proposed method in comparison with those of existing methods, as applied to three public datasets.

#### 1) FIGARO-1K

The experimental results obtained using the Figaro-1k dataset are presented in Table 1 and Fig. 6. This dataset contains images of natural human hair in various forms; hence, the model needed to segment hair with diverse patterns from the background.

As shown in Table 1, the proposed method outperformed all the compared existing methods. In particular, it outperformed the previous state-of-the-art methods by approximately 4% points in mIoU, 1% point in accuracy, and 1% point in precision. This suggests that the proposed method both segmented parts of the hair more accurately than the existing methods and performed well in segmenting the complex boundaries of hair. Furthermore, our method outperformed other methods by about 3% points in F1-score. This also means that it is effective in capturing and distinguishing the complex pattern of hair. Upon reviewing the results of existing methods, it is observed that extracting the visual features from larger Backbone models does not necessarily guarantee higher performance. This suggests that for hair segmentation, the quality of the features is more important than the size of the model. In contrast, [6]

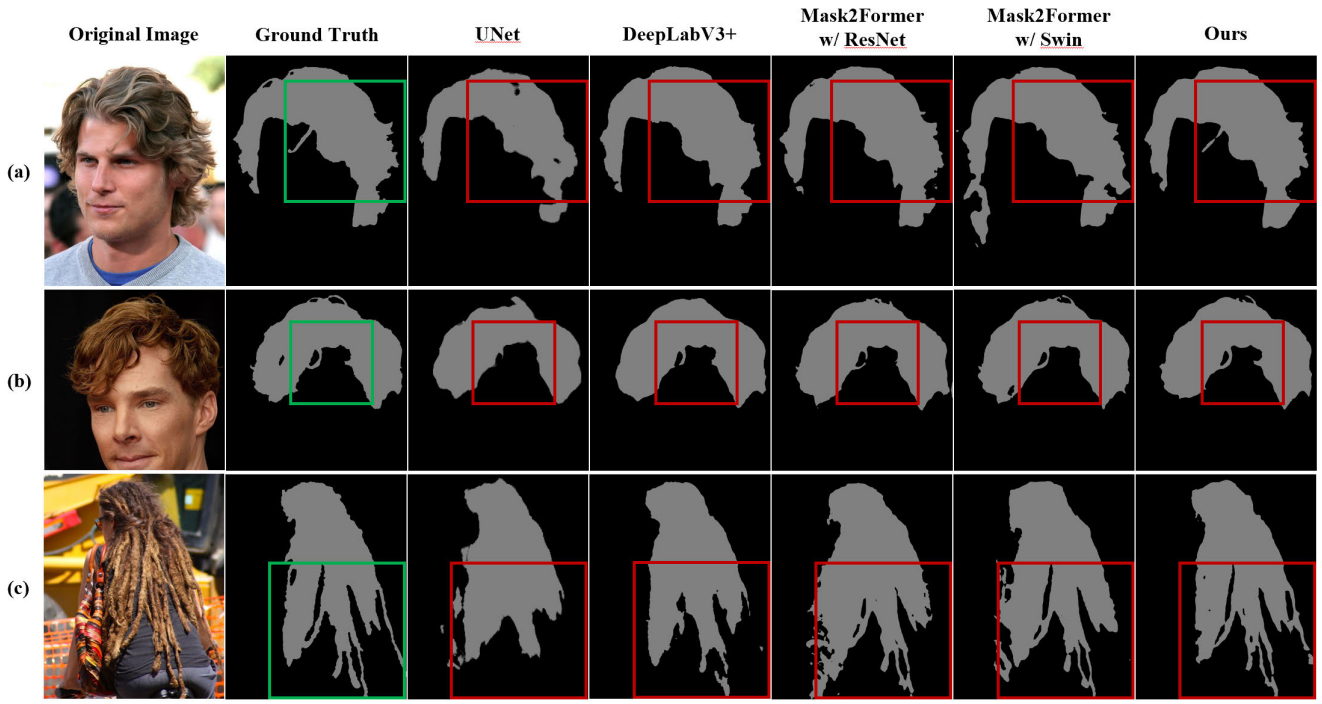


**FIGURE 6.** Comparison of qualitative results of models applied to the Figaro-1k dataset.

**TABLE 2.** Comparison of quantitative results of models applied to the CelebAMask-HQ dataset.

| Method | Backbone | mIoU (%) | Accuracy (%) | Precision (%) | F1-score (%) |
|---|---|---|---|---|---|
| UNet [8] | ResNet18 | 85.16 | 92.34 | 91.49 | 91.85 |
| | ResNet50 | 85.91 | 92.00 | 91.91 | 92.28 |
| DeepLabV3+ [11] | ResNet50 | 90.61 | 94.46 | 94.78 | 94.45 |
| Mask2Former [38] | ResNet50 | 90.89 | 93.86 | 97.68 | 97.66 |
| | SwinTransformer | 90.82 | 93.83 | 97.51 | 97.61 |
| Ours | - | **92.58** | **95.34** | **98.00** | **97.79** |

utilized the initially obtained human hair shape prior and the original image together for human hair segmentation. They also performed this segmentation with an additional refine module. Nevertheless, our proposed method demonstrated superior performance of the segmentation network with only pixel-level semantic visual features without any additional modules.

Fig. 6 shows the original images, ground truth, prediction results of the existing methods, and the prediction results of the proposed method. Overall, recent methods show improved performance in segmenting hair, but they still struggle with hair segmentation when dealing with complex hair shapes. As depicted in (a), (b), the proposed method could segment areas such as boundaries more finely than the existing methods. In particular, (c) depicts that, despite the complex pattern of the hair, the proposed method more accurately captured and segmented hair parts. These results imply that the features of the proposed method that comprise pixel-level semantic information not only succeed in capturing the various forms of hair, but are also effective in segmenting the hair more finely.

### 2) CELEBAMASK-HQ

The results of the experiments on CelebAMask-HQ are presented in Table 2 and Fig. 7. This dataset mainly consists of images focused on the faces of various celebrities, who often wear accessories such as hats and earrings in public settings. The model must distinguish hair from the background as well as from the various non-hair objects such as diverse accessories.

According to the experimental results in Table 2, the proposed method exhibited superior performance with respect to mIoU, accuracy, precision, and F1-score compared to the existing methods. Notably, our method effectively segmented human hair not only in hair-centered images but also in face-centered images. These results indicate that our method can accurately capture only the hair parts and more precisely

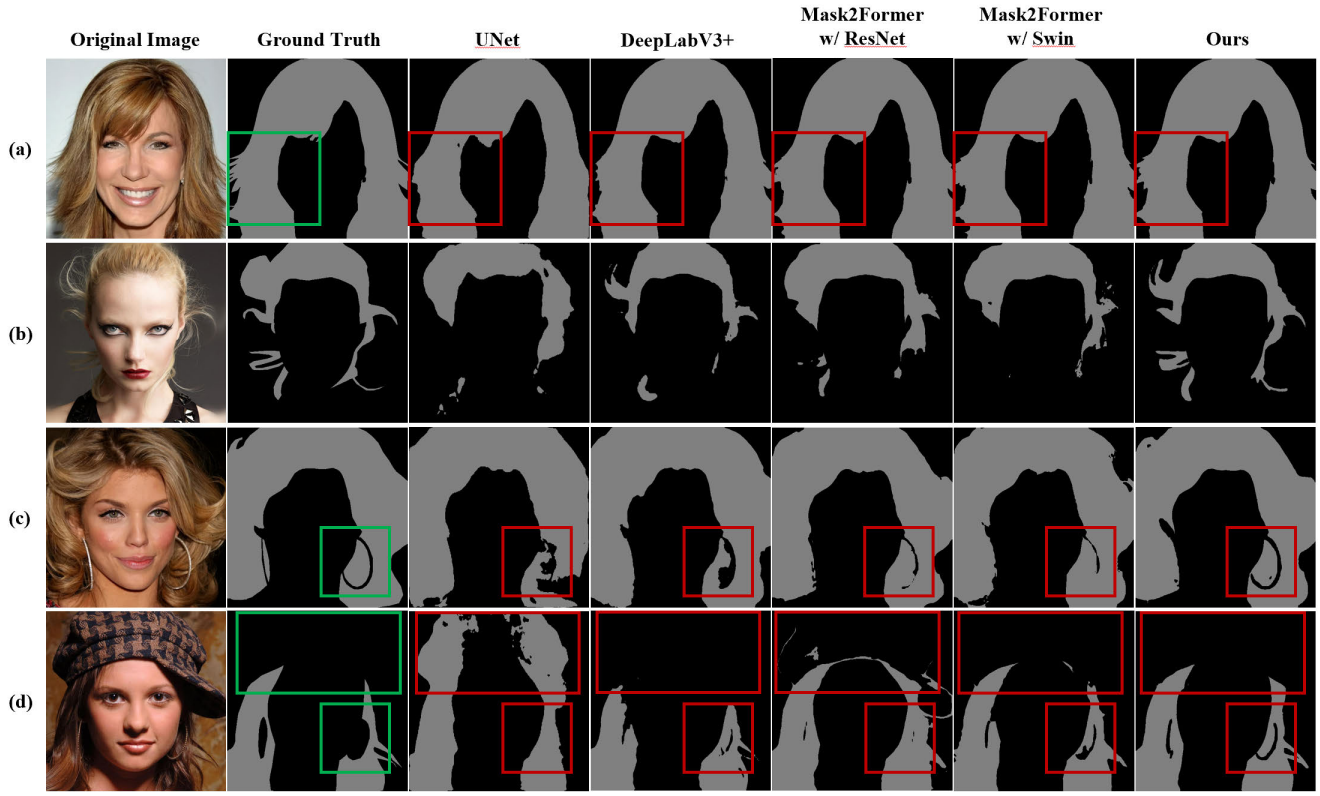


**FIGURE 7.** Comparison of results of models applied to the CelebAMask-HQ dataset.

segmented them while distinguishing hair from the various non-hair objects such as diverse accessories.

Fig. 7 depicts the visual results of the proposed method compared with those of other methods. Overall, as reviewing in the results of existing methods, recent methods have shown improved performance in segmenting hair, but they still face difficulties in achieving detailed segmentation. As shown in (b), in spite of the complex patterns exhibited by the hair around the face, the proposed method captured and segmented hair more effectively than the other methods. Moreover, as shown in (c), (d), even in instances where earrings and hats caused the hair to appear discontinuous, the proposed method successfully focused on and segmented only the hair, thereby performing finer segmentation. In particular, in (d), the predicted hair mask visually aligns more closely with the observed hair patterns than the ground truth in certain aspects. This implies that the visual features extracted from the text-to-image diffusion model, which is focused on the hair parts, are effective in segmenting hair more finely from the various non-hair elements and background.

### 3) FACE SYNTHETICS

Finally, the experimental results using Face Synthetics are reported in Table 3 and Fig. 8. This is a facial synthesis image dataset, containing more fine annotations than the other datasets. Therefore, the model must be capable of capturing the various forms and textures of hair, while also predicting a more fine-grained hair mask.

Based on the quantitative experimental results in Table 3, the proposed method achieved the best performance, surpassing the existing methods across all evaluated metrics. Notably, the highest mIoU, accuracy, and precision scores reflect our method's ability to segment the hair boundaries precisely. Additionally, the highest mIoU and F1-score indicate that our method can more effectively capture the hair parts than the existing methods, distinguishing them from the face and background.

In Table 3, more recent methods have reported increasingly improved performance. However, as shown in Fig 8, there were still limitations in the detailed segmentation of the hair boundaries. The visual prediction results illustrate that, overall, the proposed method predicts the hair mask more finely than the existing methods. In particular, the red boxes in each row indicate that the proposed method segments the boundaries of the hair more finely from the face or the background. These results imply that the pixel-level semantic visual features extracted from the diffusion-based generative model in the proposed method are effective in predicting more detailed hair masks.

In summary of all experimental results, through comprehensive evaluation across three distinct human hair-related datasets, each with its own unique characteristics, our proposed method consistently demonstrated superior

**TABLE 3.** Comparison of quantitative results of models applied to the Face Synthetics dataset.

| Method | Backbone | mIoU (%) | Accuracy (%) | Precision (%) | F1-score (%) |
|---|---|---|---|---|---|
| UNet [8] | ResNet18 | 87.20 | 94.55 | 91.30 | 92.77 |
| | ResNet50 | 87.65 | 92.17 | 91.79 | 91.48 |
| DeepLabV3+ [11] | ResNet50 | 95.42 | 97.83 | 97.40 | 97.59 |
| Mask2Former [38] | ResNet50 | 95.96 | 98.26 | 98.58 | 98.59 |
| | SwinTransformer | 95.67 | 97.61 | 98.44 | 98.45 |
| Ours | - | **97.14** | **98.79** | **99.44** | **99.18** |

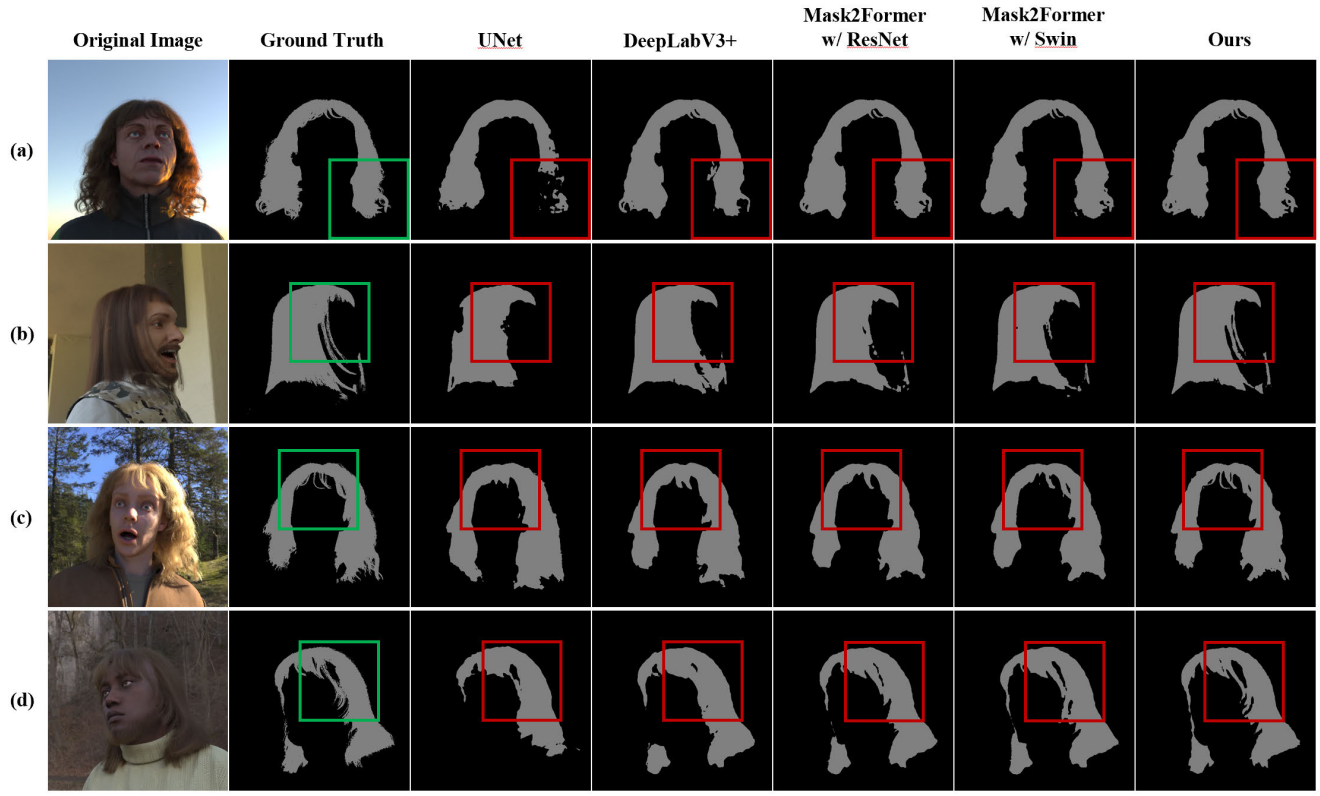


**FIGURE 8.** Comparison of results of models applied to the Face Synthetics dataset.

performances. It not only achieved the highest scores in key metrics (mIoU, accuracy, precision, and F1-score) but also qualitatively surpassed the existing methods. These results conclusively show that our method is highly effective in capturing various forms of hair in both everyday and face-centered images. Moreover, it outperforms other methods in finely segmenting hair, clearly distinguishing it from the background and various non-hair elements. This precision in segmentation is particularly notable in challenging scenarios wherein hair interacts with complex patterns and appearances.

## V. CONCLUSION

In this study, we have proposed a method for the fine-grained segmentation of human hair in various characteristics by utilizing diffusion-based generative models. Human hair, with its diverse shapes and outward appearances, presents a challenge for fine-grained segmentation from the background or from various non-hair elements. Therefore, the model must effectively capture its diverse characteristics to segment hair. To address this issue, we leveraged the advantage of pixel-level semantic information by utilizing the internal visual features of the diffusion-based generative model. Specifically, to obtain the features focused on the parts of hair, we extracted the visual features computed in the text-to-image diffusion model by pairing an image with the word `'hair.'` These visual features were then employed to train a segmentation network, enabling the prediction of more fine-grained hair masks. We conducted experiments with three distinct human hair-related datasets to validate our proposed method. Experimental results indicate that the proposed method effectively captures hair in various patterns and

segments hair more finely from the background and non-hair elements than the existing methods do. Consequently, the proposed method was empirically demonstrated to effectively perform accurate and fine-grained hair segmentation. However, the model's training and inference were conducted independently for each dataset in this study. Therefore, a limitation is that the inference results of a model trained on a specific dataset remain unclear when applied to data with different characteristics. In future research, we aim to develop an integrated model trained on a broader range of data and test it on various datasets to explore a robust method. Meanwhile, we hope that the fine-grained segmentation tasks required in other areas may be addressed through the proposed method.

## REFERENCES


[1] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," *Comput. Graph. Forum*, vol. 23, no. 3, pp. 669–676, 2004.

[2] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN: Face swapping and editing using face and hair representation in latent spaces," 2018, *arXiv:1804.03447*.

[3] H. Zhu, Y. L. Nanjing, and Y. Liu, "Automatic hair segmentation in complex background," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2019, pp. 2496–2501.

[4] D. Borza, E. Yaghoubi, J. Neves, and H. Proença, "All-in-one 'HairNet': A deep neural model for joint hair segmentation and characterization," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[5] T. A. Ileni, D. L. Borza, and A. S. Darabant, "Fast in-the-wild hair segmentation and color classification," in *Proc. VISIGRAPP*, 2019, pp. 59–66.

[6] Y. Yan, S. Duffner, X. Naturel, A. Berthelier, C. Garcia, C. Blanc, and T. Chateau, "Two-stage human hair segmentation in the wild using deep shape prior," *Pattern Recognit. Lett.*, vol. 136, pp. 293–300, Aug. 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Cham, Switzerland, 2015, pp. 234–241.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[12] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8330–8339.

[13] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-UNet: An improved neural network based on unet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21420–21428, 2019.

[14] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[15] K. Radha and Y. Karuna, "Modified depthwise parallel attention UNet for retinal vessel segmentation," *IEEE Access*, vol. 11, pp. 102572–102588, 2023.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 12077–12090.

[19] Y. Xu, F. He, B. Du, D. Tao, and L. Zhang, "Self-ensembling GAN for cross-domain semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 7837–7850, 2023.

[20] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Proc. 5th Int. Conf. Med. Imag. Deep Learn.* (Proceedings of Machine Learning Research), vol. 172, Jul. 2022, pp. 1336–1348.

[21] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," 2021, *arXiv:2112.03126*.

[22] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2955–2966.

[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 37, Jul. 2015, pp. 2256–2265.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 6840–6851.

[25] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456*.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 8780–8794.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, vol. 35. Red Hook, NY, USA: Curran Associates, 2022, pp. 36479–36494.

[28] M. Svanera, U. R. Muhammad, R. Leonardi, and S. Benini, "Figaro, hair detection and segmentation in the wild," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 933–937.

[29] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.

[30] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3661–3671.

[31] M. Yu, K. Pei, X. Li, X. Wei, C. Wang, and J. Gao, "FBCU-Net: A fine-grained context modeling network using boundary semantic features for medical image segmentation," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106161.

[32] Z. Yu, L. Yu, W. Zheng, and S. Wang, "EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation," *Comput. Biol. Med.*, vol. 162, Aug. 2023, Art. no. 107081.

[33] X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet," *Sci. Rep.*, vol. 13, no. 1, p. 7600, May 2023.

[34] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102395.

[35] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2020, pp. 213–229.

[37] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Advances in Neural Information Processing Systems*, vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 17864–17875.
[38] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.
[39] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18208–18218.
[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.
[41] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," 2021, *arXiv:2112.00390*.
[42] M. Lu, F. Xu, H. Zhao, A. Yao, Y. Chen, and L. Zhang, "Exemplar-based portrait style transfer," *IEEE Access*, vol. 6, pp. 58532–58542, 2018.
[43] P. Zhu, R. Abdal, J. Femiani, and P. Wonka, "Barbershop: GAN-based image compositing using segmentation masks," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–13, Dec. 2021.
[44] X. Xia, F. Yu, N. Li, Y. Qu, J. Zhang, and C. Zhu, "Self-attention-masking semantic decomposition and segmentation for facial attribute manipulation," *IEEE Access*, vol. 8, pp. 36154–36165, 2020.
[45] Z.-Q. Liu, J. Guo, and L. Bruton, "A knowledge-based system for hair region segmentation," in *Proc. 4th Int. Symp. Signal Process. Appl.*, vol. 2, 1996, pp. 575–576.
[46] Y. Yacoob and L. S. Davis, "Detection and analysis of hair," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1164–1169, Jul. 2006.
[47] K. Lee, D. Anguelov, B. Sumengen, and S. B. Gokturk, "Markov random field models for hair and face segmentation," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.
[48] C. Rousset and P. Coulon, "Frequential and color analysis for hair mask segmentation," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 2276–2279.
[49] W. Guo and P. Aarabi, "Hair segmentation using heuristically-trained neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 25–36, Jan. 2018.
[50] M. Chai, T. Shao, H. Wu, Y. Weng, and K. Zhou, "Autohair: Fully automatic hair modeling from a single image," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–13, Jul. 2016.
[51] S. Qin, S. Kim, and R. Manduchi, "Automatic skin and hair masking using fully convolutional networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 103–108.
[52] U. R. Muhammad, M. Svanera, R. Leonardi, and S. Benini, "Hair detection, segmentation, and hairstyle classification in the wild," *Image Vis. Comput.*, vol. 71, pp. 25–37, Mar. 2018.
[53] H.-S. Yoon, S.-W. Park, and J.-H. Yoo, "Real-time hair segmentation using mobile-unet," *Electronics*, vol. 10, no. 2, p. 99, Jan. 2021.
[54] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.
[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 8748–8763.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

**DOHYUN KIM** received the B.S. degree in statistics from Cheongju University, Republic of Korea. He is currently pursuing the M.S. degree in industrial and management engineering with Korea University, Republic of Korea. His research interests include developing machine learning algorithms for both structured and unstructured data, such as text, signal, and image, and applying them to solve problems, such as computer vision.

**EUNA LEE** received the B.S. degree in mathematics from the University of Ulsan, in 2009, and the M.S. and Ph.D. degrees in industrial engineering from Korea University, in 2013 and 2023, respectively. She is currently an Associate Fellow with the Korea Institute for Defense Analyses. Her research interests include requirements analysis of military supplies and artificial intelligence.

**DAEHYUN YOO** received the B.S. degree in statistics from Yonesi University, Republic of Korea. He is currently pursuing the M.S. degree in industrial engineering from Korea University, Republic of Korea. His research interests include image generation, machine learning, and deep learning.

**HONGCHUL LEE** received the B.S. degree in industrial engineering from Korea University, in 1983, the M.S. degree in industrial engineering from The University of Texas at Arlington, in 1988, and the Ph.D. degree in industrial engineering from Texas A&M University, in 1993. He is currently a Professor with the Department of Industrial Systems and Information Engineering, Korea University. His research interests include system engineering, system simulations, and artificial intelligence.