Supplemetal Material:
Towards High-fidelity Head Blending with Chroma Keying for Industrial Applications

Hah Min Lew[1]*, Sahng-Min Yoo[2]*‡, Hyunwoo Kang[3]*‡, and Gyeong-Moon Park[4]†
[1]Klleon AI Research, [2]Samsung Research, [3]Hyperconnect, [4]Kyung Hee University
hahmin.lew@klleon.io {yoosahngmin, khw7147}@gmail.com gmpark@khu.ac.kr

This supplementary document provides an intensive insight into our work presented in the main paper, consisting of qualitative comparisons with the state-of-the-art inpainting methods, the notation and visualization of our Head shape and long Hair (H^2) augmentation, more detailed analysis and descriptions of the proposed Foreground Predictive Attention Transformer (FPAT), implementation details on experimental settings and the training objectives, experimental details on the user study, and a head blending video results in our project page.

A. Qualitative Comparisons with Recent Inpainting Models

In this section, we investigate the performance of our CHANGER compared to the state-of-the-art inpainting models through qualitative comparisons.

Baselines. We establish the state-of-the-art inpainting models as follows: (1) Stable Diffusion Inpainting (SDI) [3], (2) Paint-by-Example (PBE) [6], (3) ProPainter [8].

Figure 1 shows the results from head blending video compared with the recent diffusion-based or video-based inpainting models. *SDI* and *PBE* mainly suffered from background generation (green boxes) and artifacts of the foreground region. *ProPainter* showed blurry foreground generation (orange boxes). Our results show not only the highest fidelity in the background inpainting region as well as the foreground but also stability in a time-consistency perspective, which ensures the quality of the video output.

B. Notation and Visualization Summary of H^2 Augmentation

We provide a detailed explanation of the various notations used in our method, especially for the proposed Head shape and long Hair (H^2) augmentation, in Table 1. We also visualize the process of H^2 augmentation in Figure 2. Please refer to the descriptions in the table and figure, to ensure clarity in interpreting our work.

C. More Details in FPAT

C.1. Attention Map of FPAT

The Foreground Predictive Attention Transformer (FPAT) stands at the forefront of our model structure, primarily focusing on the enhancement of the fidelity of foreground blending. In the diverse situations created by various head shape and hairstyle differences between the source and target images, FPAT aims to predict the foreground region and then attend to the predicted foreground region. We demonstrate the effectiveness of the FPAT qualitatively.

Figure 3 presents predicted masks (M), attention maps (*Attention*), and head blending results (Y) obtained by FPAT on various source and target pairs. For each input, two distinct attention maps are depicted: one for the neck (upper row) and another for the cloth (lower row). The small red boxes inside the images in the X column represent the patches used to generate queries for our proposed FPAT transformer layer. The images in the *Attention* column depict the calculated attention derived from these queries and keys, where higher values are represented closer to yellow and lower values closer to blue.

The predicted mask results show that FPAT effectively reconstructs obscured foreground areas caused by long hair. Furthermore, meaningful attention is trained within the predicted region, as depicted in the attention maps. Specifically, during the generation of the neck region (upper row), the model focuses explicitly on the neck area of the target image. In contrast, when generating the occluded attire region (lower row), the model focuses on relevant clothing areas, indicating its ability to create images with attention to pertinent regions.

Project page: https://hahminlew.github.io/changer

*Equal contribution

†Corresponding author

‡Work done at Klleon AI Research

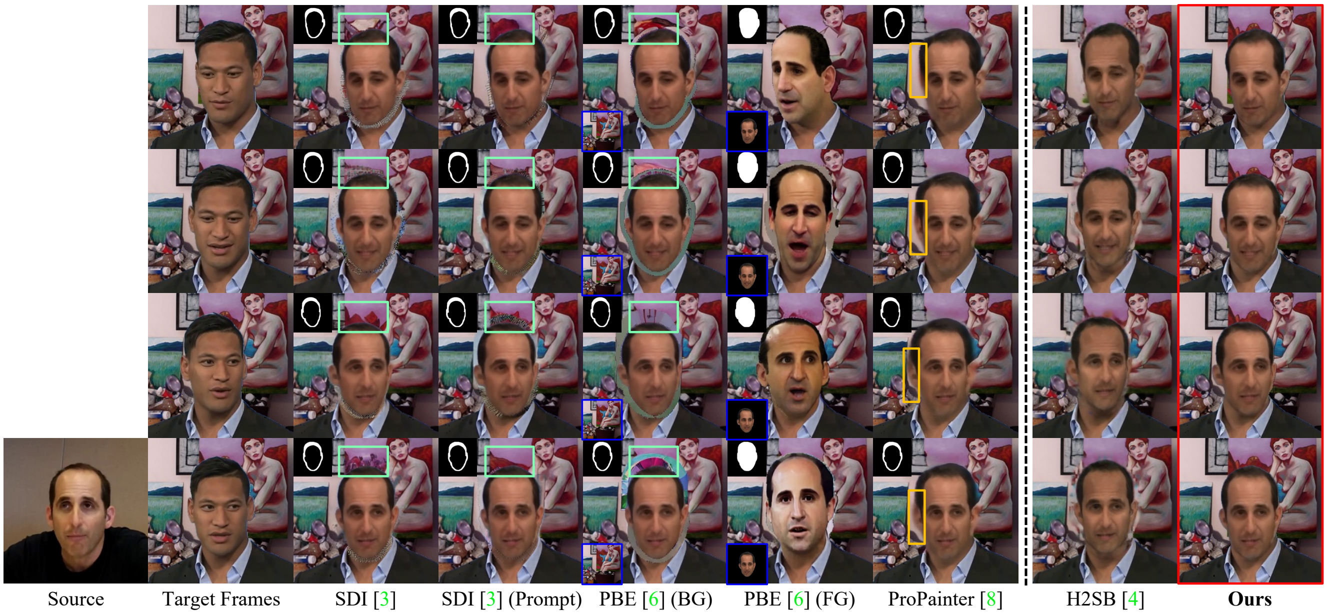

Figure 1. Qualitative comparisons of using recent inpainting baselines [3, 6, 8] and the head blending model [4] on sequential frames of a target video. We tested both scenarios with and without text prompting (Prompt) for *SDI*. For *PBE*, we separated scenarios; the background (BG) and the foreground (FG) references (bottom-left blue boxes of each column).

Notation	Dimension	Description
I_S	$\mathbb{R}^{3\times H\times W}$	Source image.
I_T	$\mathbb{R}^{3\times H\times W}$	Target image.
I_S^{gray}	$\mathbb{R}^{1\times H\times W}$	Gray-scale image from source.
I_T^{green}	$\mathbb{R}^{3\times H\times W}$	Target image with a green screen background.
I_T^{head}	$\mathbb{R}^{3\times H\times W}$	Target head image, used in Head Colorizer, made by only leaving the head region from the target image.
I_T^{body}	$\mathbb{R}^{3\times H\times W}$	Target body image, used in Body Blender, made by excluding head, neck, and background.
$\mathtt{M}_S^{head}$	$\mathbb{R}^{1\times H\times W}$	Head mask from source.
$\mathtt{M}_T^{head}$	$\mathbb{R}^{1\times H\times W}$	Head mask from target.
$\mathtt{M}_{h^1}^{head}$	$\mathbb{R}^{1\times H\times W}$	Augmented head mask made by transformation $\mathcal{T}_{head}$.
$\mathtt{M}_{h^2}^{head}$	$\mathbb{R}^{1\times H\times W}$	Augmented head mask from $\mathtt{M}_{h^1}^{head}$, made by transformation $\mathcal{T}_{hair}$.
$\mathtt{M}_{union}^{head}$	$\mathbb{R}^{1\times H\times W}$	Union mask of $\mathtt{M}_{h^2}^{head}$ and $\mathtt{M}_S^{head}$ during training, union mask of $\mathtt{M}_T^{head}$ and $\mathtt{M}_S^{head}$ during testing.
$\mathtt{M}^{ip}$	$\mathbb{R}^{1\times H\times W}$	Inpainting region subtracting $\mathtt{M}_S^{head}$ from $\mathtt{M}_{union}^{head}$.
M	$\mathbb{R}^{1\times H\times W}$	Predicted foreground mask which is further used as an input of the FPAT blocks.
X	$\mathbb{R}^{3\times H\times W}$	Input for our CHANGER.
Y	$\mathbb{R}^{3\times H\times W}$	Head blended outputs of our CHANGER.

Table 1. **Notations and corresponding descriptions in our CHANGER.**

C.2. Detailed Explanation of FPAT Mechanism

Our FPAT starts with the input z_c, and predicts a foreground region, including the body and the neck, as a binary mask $M \in \mathbb{R}^{h\times w}$ with Foreground-Prediction module. The FPAT block refers to the target body information I_T^{body} and updates z_c using the information of M to generate the neck and body via the Foreground-Aware Transformer block. FPAT patchifies the feature output of the head colorizer $z_c \in \mathbb{R}^{C\times h\times w}$ and get $z_c^p \in \mathbb{R}^{N\times P^2 C}$, where (P, P) is the resolution of the patches and $N = hw/P^2$ is

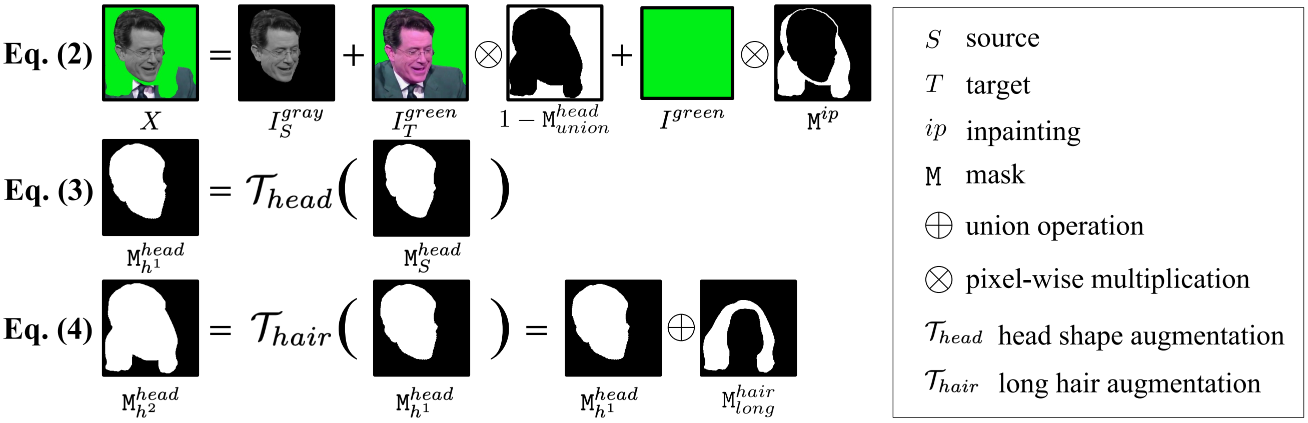
Eq. (2)
X
I_S^{gray}
I_T^{green}
$1 - \mathtt{M}_{union}^{head}$
I^{green}
$\mathtt{M}^{ip}$
Eq. (3)
$\mathtt{M}_{h^1}^{head}$
$= \mathcal{T}_{head}($
$\mathtt{M}_S^{head}$
Eq. (4)
$\mathtt{M}_{h^2}^{head}$
$= \mathcal{T}_{hair}($
$\mathtt{M}_{h^1}^{head}$
$\mathtt{M}_{h^1}^{head}$
$\mathtt{M}_{long}^{hair}$
S source
T target
ip inpainting
$\mathtt{M}$ mask
$\oplus$ union operation
$\otimes$ pixel-wise multiplication
$\mathcal{T}_{head}$ head shape augmentation
$\mathcal{T}_{hair}$ long hair augmentation

Figure 2. **Visualization of H^2 Augmentation.** Eq. (2) is the input X formulation during training. Inspired by [7], we apply the same color jitter to both I_T^{green} and the ground truth during the training phase. Eq. (3) shows the head shape augmentation. Eq. (4) shows the long hair augmentation.

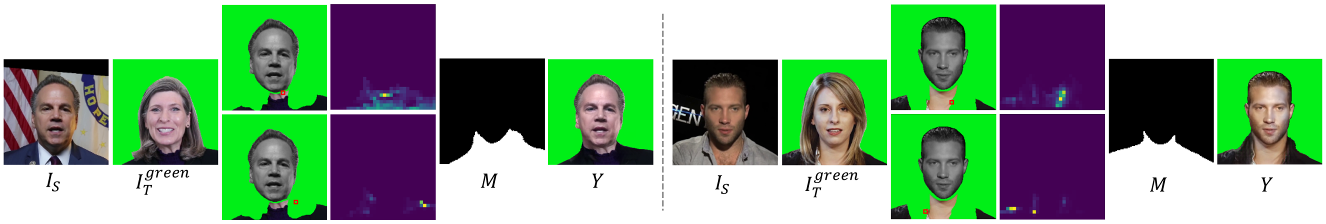

Figure 3. The foreground mask predicted by FPAT (M), the attention map used in the transformer layer (*Attention*), and the head blending result (Y) when input source image I_S and target image I_T are used. We visualize the similarity between the query patch (red box) and each key patch in the depicted image as an attention map. Blue represents low values and yellow represents high values.

the number of patches. FPAT also patchifies the embedded feature of the target body I_T^{body} as $z^p_{body} \in \mathbb{R}^{N \times P^2 C}$, and the predicted body and neck mask M as $M^p \in \mathbb{R}^{N \times P^2}$. Then, FPAT averages M^p along the channel axis to acquire $M^p_{\text{avg}} \in \mathbb{R}^N$ as following:

$$[M^p_{\text{avg}}]_n = \frac{1}{P^2}\Sigma^{P^2}_{m=1} M^p_{nm}, \tag{1}$$

where $[M^p_{\text{avg}}]_n$ is the n-th patch of M^p_{avg} and M^p_{nm} is the (n, m)-th element of M^p. Next, we divide patches into two groups: (1) a set of patches S_b that includes the predicted body and neck parts and (2) a set of patches S_{nb} that does not include them by thresholding M^p_{avg} by following:

$$\begin{aligned} S_b &= \{i \in 1, ..., N \,|\, [M^p_{\text{avg}}]_i \geq \tau\} \\ S_{nb} &= \{i \in 1, ..., N \,|\, [M^p_{\text{avg}}]_i < \tau\}, \end{aligned} \tag{2}$$

where τ is the hyperparameter. Then, FPAT computes the binary mask $M^b \in \mathbb{R}^{N \times N}$ as following:

$$M^b_{ij} = \begin{cases} 0, & \text{if } i, j \in S_b \;,\; i, j \in S_{nb}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{3}$$

where M^b_{ij} is the (i, j)-th element of M^b.

Finally, FPAT masks the attention between a query from the latent representation z^p_c, key and value from the target head feature z^p_{body}.

D. Implementation Details

Experimental Settings. An Adam optimizer [1] with hyperparameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used for every models. We used learning rates 1e-4 and 4e-4 for the generator and discriminator, respectively. We used $\epsilon = 0.5$ in Eq. (4) in our main paper. We used 4 NVIDIA RTX 3090 (24 GB) GPUs to train our CHANGER.

Training Objectives Details. We train the model with $\mathcal{L}_{total}$, which is a summation of **(1)** $\mathcal{L}_{rec}$, the reconstruction loss for the final output head and the ground truth, **(2)** $\mathcal{L}_{hc}$, the reconstruction loss for the output of `ToRGB` block, **(3)** $\mathcal{L}_{mask}$ [2], the loss for the output of the Foreground-Prediction module, **(4)** perceptual loss $\mathcal{L}_{per}$, and **(5)** adversarial loss $\mathcal{L}_{adv}$ for our objective functions.

Corresponding objective functions are as follows:

$$
\begin{aligned}
\mathcal{L}_{total} &= \lambda_{rec}\mathcal{L}_{rec} + \lambda_{hc}\mathcal{L}_{hc} + \lambda_{mask}\mathcal{L}_{mask} \\
&\quad + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv},
\end{aligned} \tag{4}
$$

$$
\mathcal{L}_{rec} = ||Y \otimes \mathtt{M}_S^{head} - I_T \otimes \mathtt{M}_S^{head}||_1, \tag{5}
$$

$$
\mathcal{L}_{hc} = ||Y - I_T^{hc}||_1, \tag{6}
$$

$$
\mathcal{L}_{mask} = ||M_{gt} - M||_1, \tag{7}
$$

$$
\mathcal{L}_{per} = \sum_{i=1}^{L} ||\Phi_i(Y) - \Phi_i(I_T)||_1, \tag{8}
$$

$$
\begin{aligned}
\mathcal{L}_{adv}^{D_I} &= -\mathbb{E}_{I_T \sim p_{data}}[\log(D_I(I_T))] \\
&\quad - \mathbb{E}_{Y \sim p_Y}[\log(1 - D_I(Y))],
\end{aligned} \tag{9}
$$

$$
\mathcal{L}_{adv}^{\mathcal{D}(z)} = -\mathbb{E}_{Y \sim p_Y}[D_I(Y)], \tag{10}
$$

where λ_{rec}, λ_{hc}, λ_{mask}, λ_{per}, and λ_{adv} are weights for the loss $\mathcal{L}_{rec}$, $\mathcal{L}_{hc}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{adv}$, respectively. I_T^{hc} is a target image without neck, and body completion and Φ is a pre-trained VGG19 network [5], and D_I is a discriminator. We used λ_{rec} = 10, λ_{hc} = 10, λ_{mask} = 10, λ_{per} = 1, and λ_{adv} = 1.

E. Project Page

The head blending video results are shown on our project page linked in the footnote of the main paper. The video results demonstrate the effectiveness and robustness of CHANGER in various industrial scenarios and suggest its potential for adoption in the industrial field.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[2] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 3

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[4] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot head swapping in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10789–10798, 2022. 2

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[6] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1, 2

[7] Sahng-Min Yoo, Tae-Min Choi, Jae-Woo Choi, and Jong-Hwan Kim. Fastswap: A lightweight one-stage framework for real-time face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3558–3567, 2023. 3

[8] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 1, 2