Supplementary Materials:
Hunting Blemishes: Language-guided High-fidelity Face Retouching Transformer with Limited Paired Data

Anonymous Authors

1 ADDITIONAL RESULTS

1.1 Blemish Detection

We propose a Language-guided Blemish Removal Transformer for automatically retouching face images, which is referred to as LangBRT. One of our design elements is the Text-prompted Blemish Detection module (TBD), which utilizes the prior knowledge encapsulated in a pre-trained vision-language model. Based on the text descriptions of facial blemishes, TBD is able to detect prompt-specific blemishes, which are used to guide the retouching process. In Figure 1, we visualize the blemish detection results for the cases where the textual prompts are "forehead pimples / reflection / messy hair", "dark circle under the eyes", "blemish on the skin" and so on. One can observe that the combinations of blemish prompts are useful for detecting more blemishes. We further perform a visual comparison between LangBRT and the state-of-the-art face retouching method, BPFRe. As shown in Figure 2, LangBRT outperforms BPFRe in blemish detection, since the heat maps of LangBRT indicate the blemish regions, which are more consistent with the manual retouching regions.

1.2 Comparison to State-of-the-arts

To highlight the superiority of our proposed LangBRT with both limited(1%) and full(1%) paired training data, we perform visual comparison to the main competing methods: ABPN and BPFRe on in-the-wild data. A number of representative retouching images of the methods are shown in Figure 3 and Figure 4. One can find that LangBRT achieves better retouching results than the main competing methods both with 1% and 100% FFHQR training data, while at the same time preserving the non-blemish content of the input face images.

Further, to demonstrate the effectiveness of LangBRT, we present high-resolution examples for reflection and dark circle(Figure 5), messy hair and pimples(Figure 6). This confirms the significant advantages of our LangBRT in handling various types of blemishes.

2 SUMMARY OF LANGBRT

LangBRT consists of the following learnable components: the image encoder E, latent transformer T, decoder D and discriminator S, and their network architectures are presented in Tables 1, 2, 3 and 4, respectively. The constituent networks are jointly optimized from scratch, and the training procedure is summarized in Algorithm 1.

3 ETHICAL STATEMENT

This study on facial blemish removal adheres to strict ethical guidelines. Participants were fully informed about the study's purpose and provided voluntary consent. We prioritize participant privacy, protecting their data through secure storage and access restrictions. The proposed retouching technique aims to enhance natural attributes while ensuring transparency and avoiding deceptive alterations. We acknowledge potential biases and limitations, emphasizing ongoing research to promote fair and inclusive application. Overall, we are committed to conducting this study ethically and contributing to responsible face retouching practices.The use of the proposed model for harmful purposes is strongly discouraged.

Algorithm 1 Pseudo-code of training the constituent networks in LangBRT.

```
Input: FFHQR training dataset $X$, pre-defined prompts $\boldsymbol{pt}$ and user-defined prompts $\boldsymbol{ut}$.
Output: Clean face image $\hat{\boldsymbol{y}}$
Initialize Encoder $E$ parameterized by $\theta_E$, latent Transformer $T$ parameterized by $\theta_T$, decoder $D$ parameterized by $\theta_D$, discriminator $S$ parameterized by $\theta_S$, blemish feature mapping layers $\psi$, weighting factor $\gamma$ and $\eta$, learning rates $\epsilon$, and # epochs $\Gamma$.
for $t = 1 \rightarrow \Gamma$ do
    for each mini-batch do
        Sample paired training images: $(x, y) \in X$.
        Feed the raw image $x$ into $E$ to obtain the feature $\boldsymbol{F_x}$.
        Feed $\{\boldsymbol{pt}, \boldsymbol{ut}\}$ into pre-trained Alpha-CLIP text encoder $E_{Text}$ to obtain the features $\{\boldsymbol{F_{pt}}, \boldsymbol{F_{ut}}\}$, mapping $\boldsymbol{F_{pt}}$ feature with $\psi$, and expand the resulting embedding $(\psi(\boldsymbol{F_{pt}}) + \boldsymbol{F_{ut}})$ to estimate the blemish mask $\boldsymbol{m}$, conditioned on $\boldsymbol{F_x}$.
        Feed $\boldsymbol{m}$ and $\boldsymbol{F_x}$ into $T$ and $D$ to synthesize image $\hat{\boldsymbol{y}}$.
        Optimize $E$, $T$ and $D$ by using Adam: $\{\theta_E, \theta_T, \theta_D\} \leftarrow Adam(\nabla(\mathcal{L}^G_{\text{adv}} + \mathcal{L}_{\text{cons}} - \gamma\mathcal{L}_{\text{local}} + \eta(\mathcal{L}_{\text{detc}} + \mathcal{L}_{\text{comp}})), \{\theta_E, \theta_T, \theta_D\}, \epsilon)$
        Optimize $S$ by using Adam: $\theta_S \leftarrow Adam(-\nabla(L^S_{adv}), \theta_S, \epsilon)$
    end for
end for
return $\{\theta_E, \theta_T, \theta_D, \theta_S\}$
```

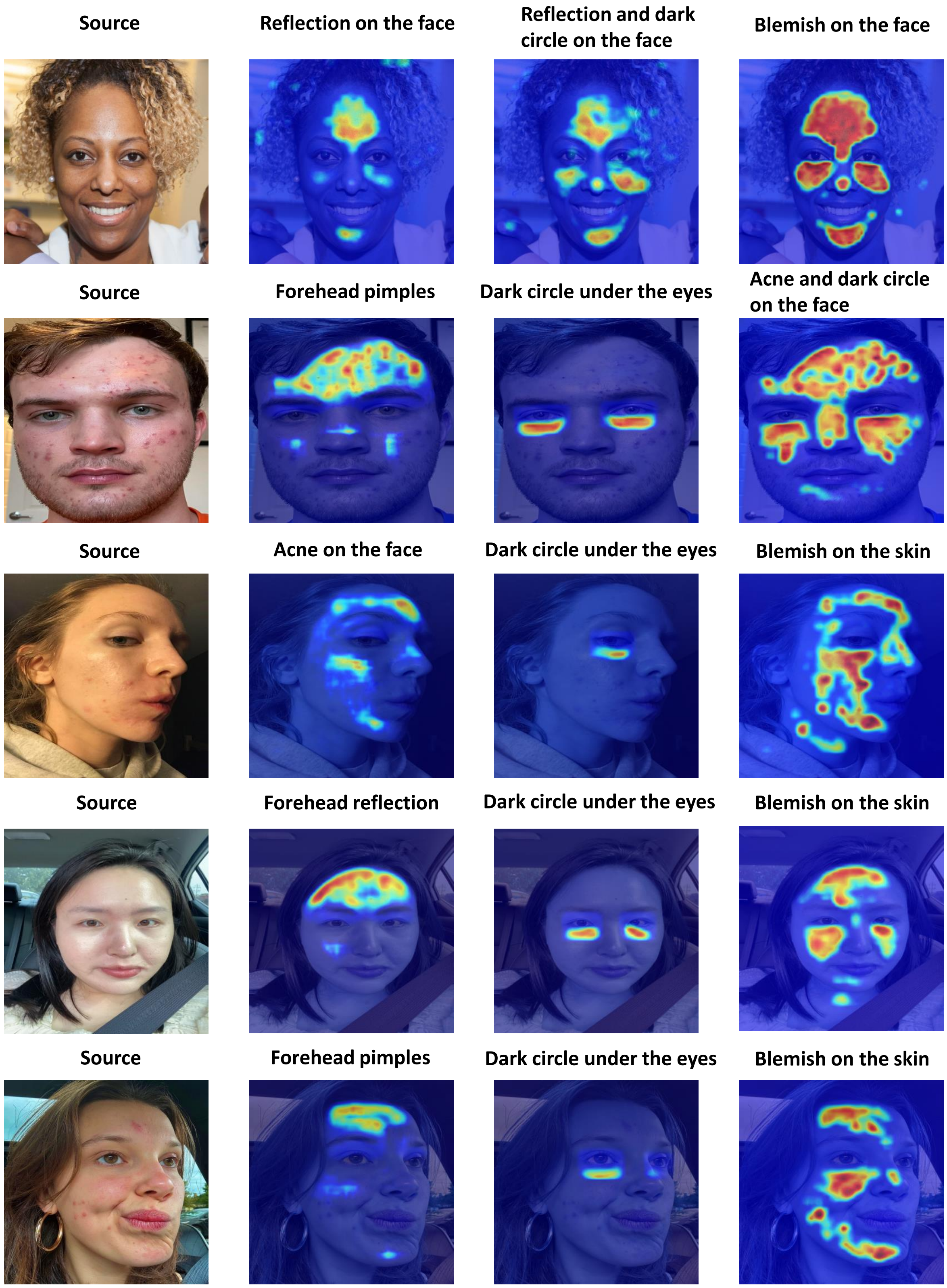

Figure 1: Blemish detection results of LangBRT on in-the-wild images, conditioned on different textual prompts.

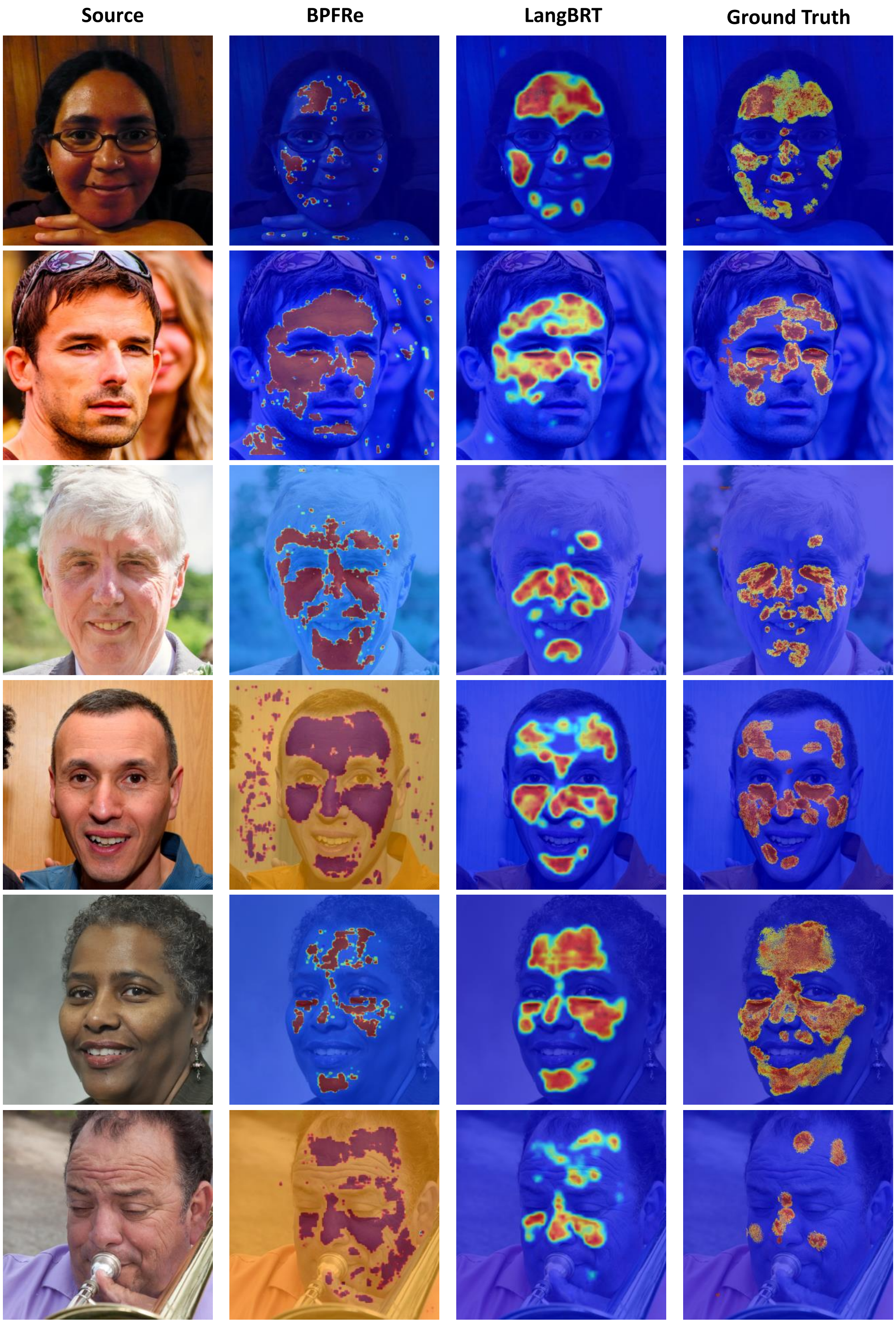

Figure 2: Visual comparison between LangBRT and BPFRe in blemish detection on FFHQR images.

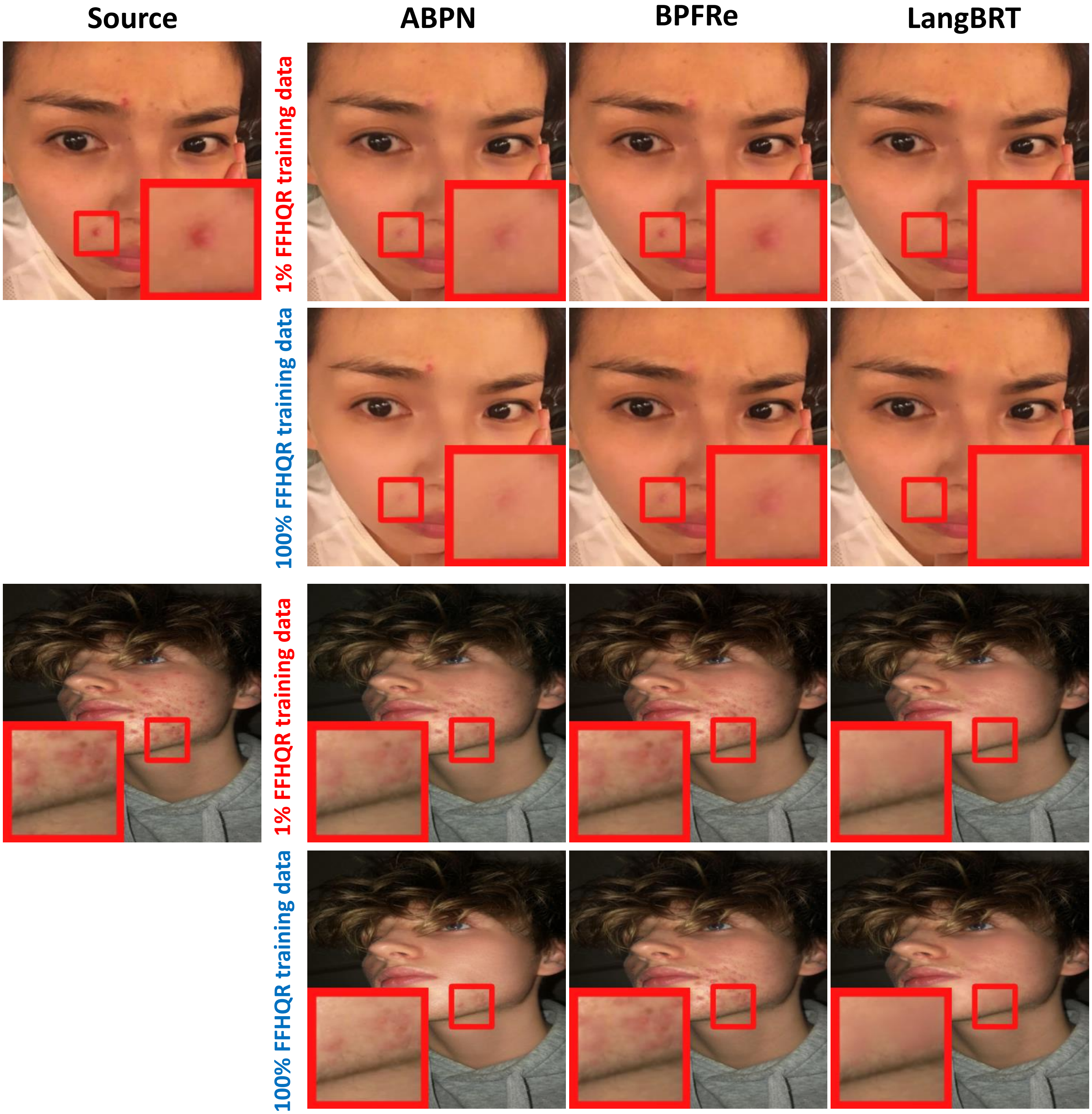

Figure 3: Representative face retouching results of LangBRT and the main competing methods on in-the-wild images.

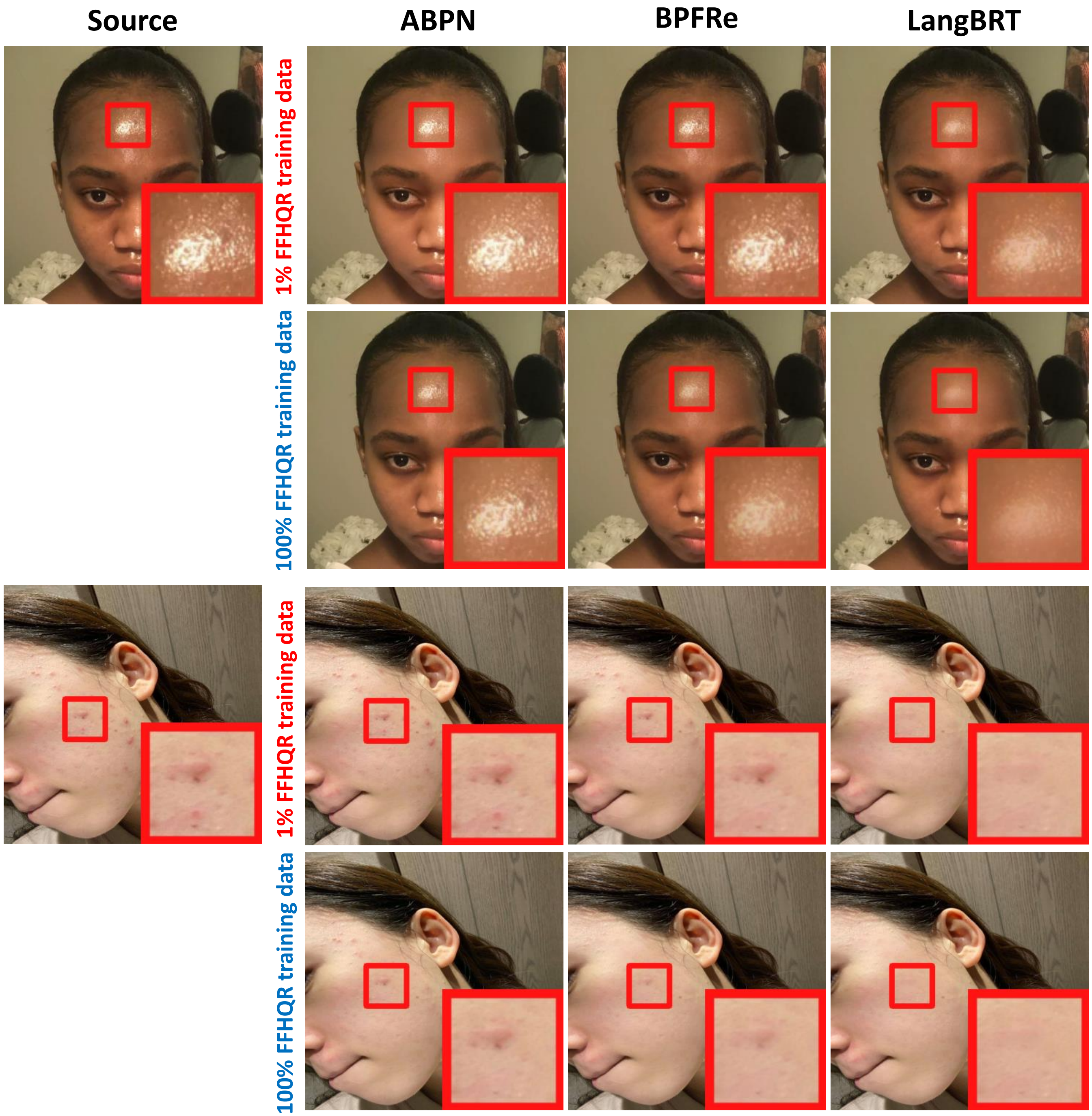

Figure 4: Representative face retouching results of LangBRT and the main competing methods on in-the-wild images.

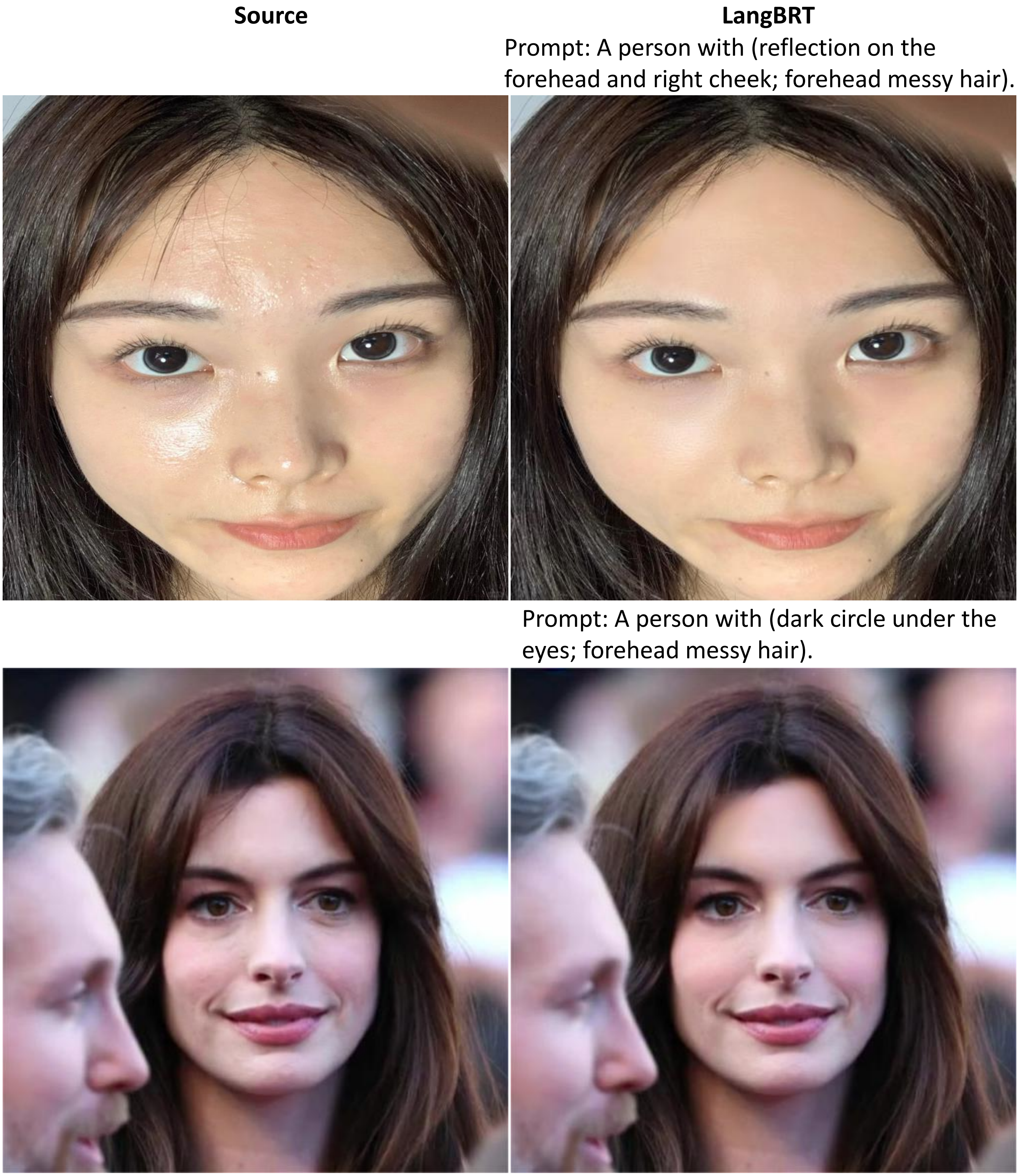
Source
LangBRT
Prompt: A person with (reflection on the forehead and right cheek; forehead messy hair).
Prompt: A person with (dark circle under the eyes; forehead messy hair).

Figure 5: Reflection removal and dark circle removal results on an in-the-wild image.

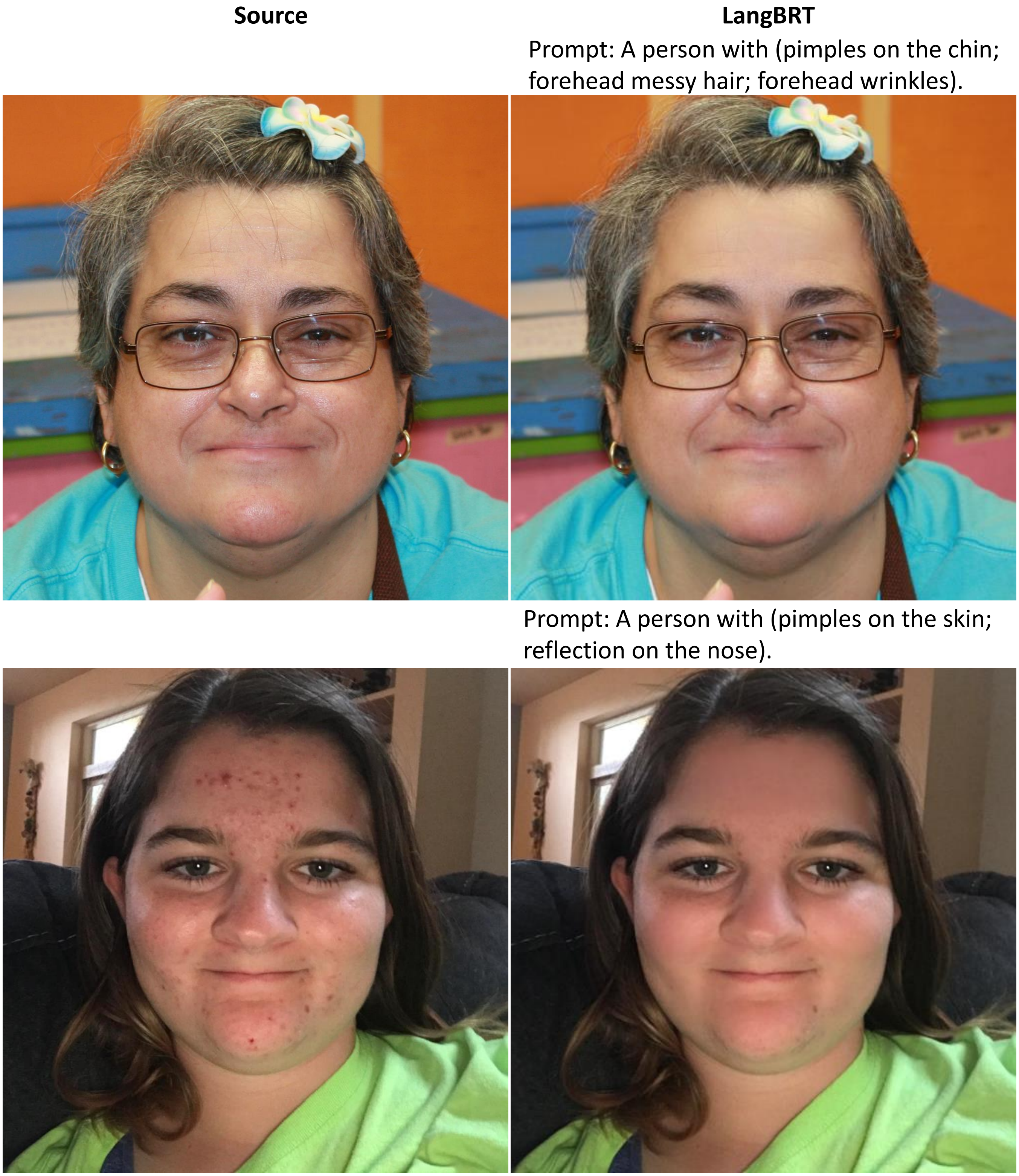
Source
LangBRT
Prompt: A person with (pimples on the chin; forehead messy hair; forehead wrinkles).
Prompt: A person with (pimples on the skin; reflection on the nose).

Figure 6: Messy hair removal and Pimples removal results on an in-the-wild image.

Table 1: The network architecture of the encoder E.

Layer	Activation	Input size	Output size
Conv	Leaky ReLU	$512 \times 512 \times 3$	$512 \times 512 \times 64$
Conv	Leaky ReLU	$512 \times 512 \times 64$	$256 \times 256 \times 128$
Conv	Leaky ReLU	$256 \times 256 \times 128$	$128 \times 128 \times 256$
Conv	Leaky ReLU	$128 \times 128 \times 256$	$64 \times 64 \times 512$
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$

Table 2: The network architecture of the latent transformer T (TCA denotes Target-specific Cross-Attention).

Layer	Window size	Heads	Norm	MLP ratio
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0
TCA	$6 \times 6 \times 16$	8	Layer	2.0

Table 3: The network architecture of the decoder D.

Layer	Activation	Input size	Output size
Conv	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$64 \times 64 \times 512$
ResBlk	Leaky ReLU	$64 \times 64 \times 512$	$128 \times 128 \times 256$
ResBlk	Leaky ReLU	$128 \times 128 \times 256$	$256 \times 256 \times 128$
ResBlk	Leaky ReLU	$256 \times 256 \times 128$	$512 \times 512 \times 64$
ToRGB	-	$512 \times 512 \times 64$	$512 \times 512 \times 3$

Table 4: The network architecture of the discriminator S.

Layer	Activation	Norm Layer	Input size	Output size
Conv	Leaky ReLU	Spectral Norm	$512 \times 512 \times 3$	$512 \times 512 \times 32$
Conv	Leaky ReLU	Spectral Norm	$512 \times 512 \times 32$	$256 \times 256 \times 64$
Conv	Leaky ReLU	Spectral Norm	$256 \times 256 \times 64$	$128 \times 128 \times 128$
Conv	Leaky ReLU	Spectral Norm	$128 \times 128 \times 128$	$64 \times 64 \times 256$
Conv	Leaky ReLU	Spectral Norm	$64 \times 64 \times 256$	$64 \times 64 \times 512$
Conv	Leaky ReLU	Spectral Norm	$64 \times 64 \times 512$	$64 \times 64 \times 512$