SUPPLEMENTARY MATERIAL

A DISCUSSION

Why dose SDC work? In the Semantic Direction Consistency (SDC) loss, our method simply treats the textual semantic feature y_t as semantic direction, which seems unreasonable. In theory, the change direction of visual semantic should align with the change direction of textual semantic. Consider a common example, *"smiling"*. If there exist two description of the target image and the desired image, it can be *"a man is walking down the street"* and *"a smiling man is walking down the street"*. Notably, the main semantic change of the two descriptions is *"smiliing"*. Besides, we can not use a unified text to describe all of the target images. So it makes sense that treating the textual semantic feature as semantic direction.

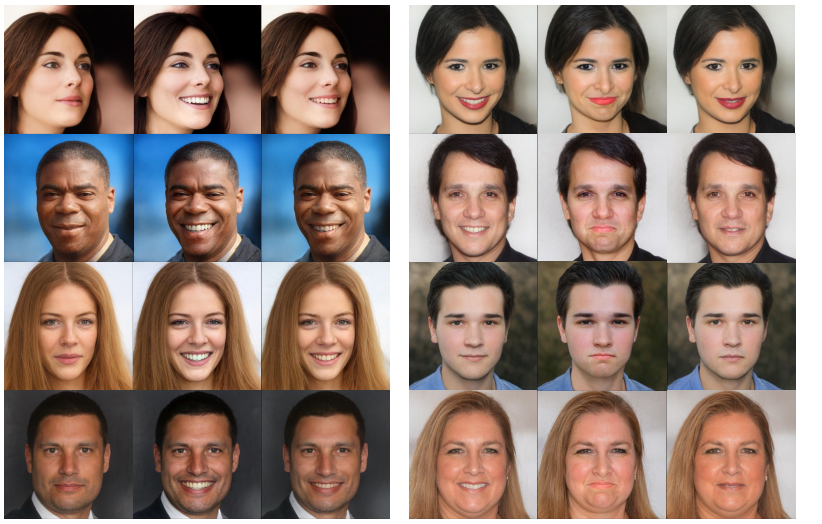

Figure 8: Qualitative results of different choice of semantic direction. For the original textual semantic direction, we mark it as *"smiling"* or *"unsmiling"*. For the semantic direction of two descriptions, we remark it as "two text". We set the description of target image as *"a person"*, and the description of desired image as *"a smiling n unsmiling person"*.

As mentioned above, we want to prove that simply using textual semantic feature y_t as semantic direction could achieve the same effect of using two different descriptions. We experiment on two opposite attributes: *"smiling"* and *"unsmiling"*, and illustrate this results in Figure 8. For the original semantic direction in SDC loss, we mark it as *"smiling"* or *"unsmiling"*. For the semantic direction of two descriptions, we named it as "two tex" manipulation and mark it as "two text". We set the description of target image as *"a person"*, and the description of desired image as *"a smiling \ unsmiling person"*.

After careful comparison, we have the following observations: (1) For the attribute *"smiling"*, our proposed method could correctly manipulate the target images as well as the "two text" manipulation. (2) However, for the attribute *"unsmiling"*, the "two text" manipulation changes irrelevant attributes due to inaccurate descriptions. Obviously it is intractable to describe every target images in the datasets, so we use the textual semantic feature y_t as efficient semantic direction.

B ADDITIONAL RESULTS

In this section, we show more qualitative experimental results.

B.1 RESULTS ON VARIOUS ATTRIBUTES

As shown in Figure 9 and Figure 10, we employ our SDD-Net on other challenging single attributes such as *"black hair"* and *"bangs"*, which needs the method to understand the global semantic information. For *"black hair"* attribute, the SDD-Net projects the $\mathcal{V}_{RT}^{\rightarrow}$ vector to get the final direction of projected vector. Owing to that the hair color is the opposite of given text prompt, the SDD-Net manipulates the target face in a reverse direction. For *"bangs"* attribute, our SDD-Net could correctly transfer the fine-grained semantic feature to the target, which leads to refine attribute transfer.

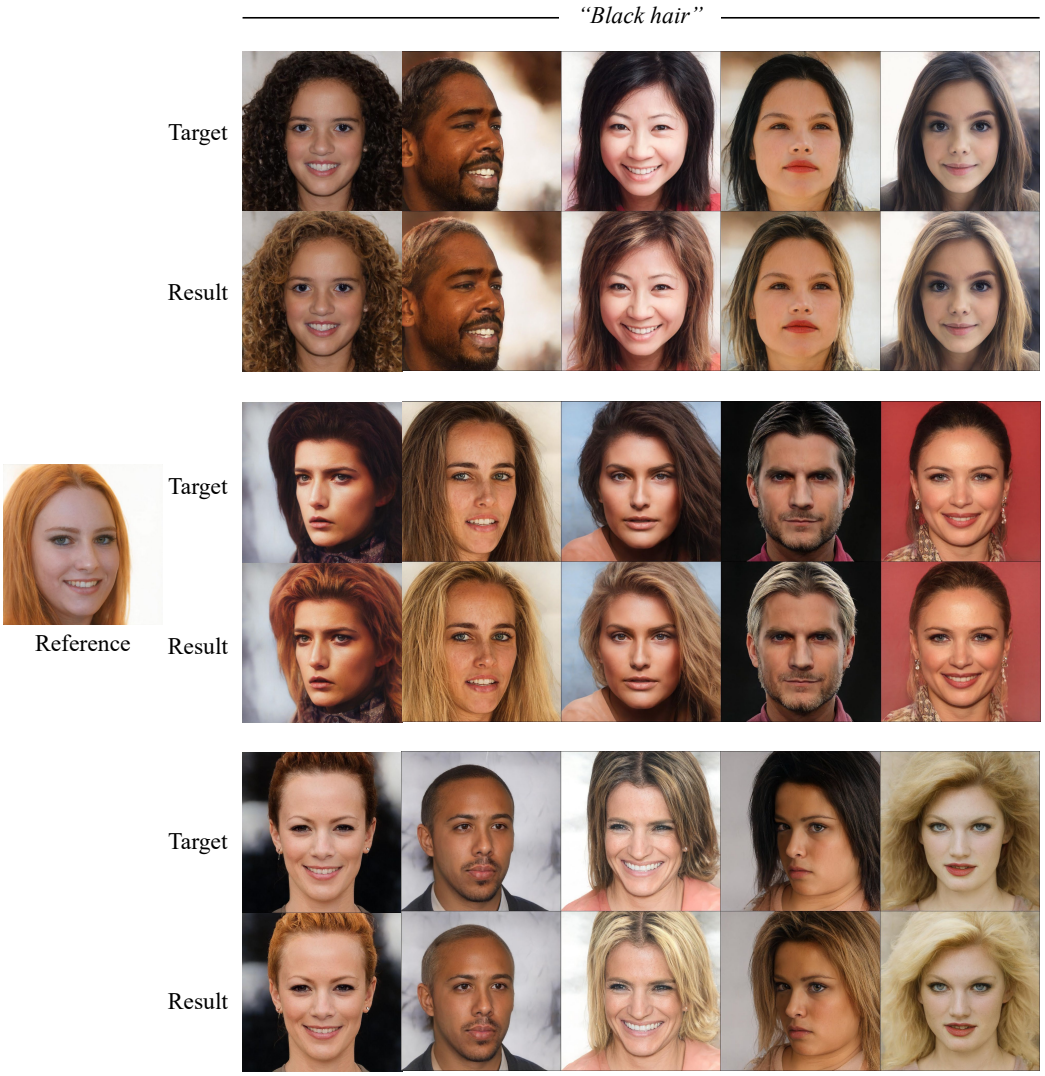

Figure 9: The experiment results on attribute *"black hair"*.

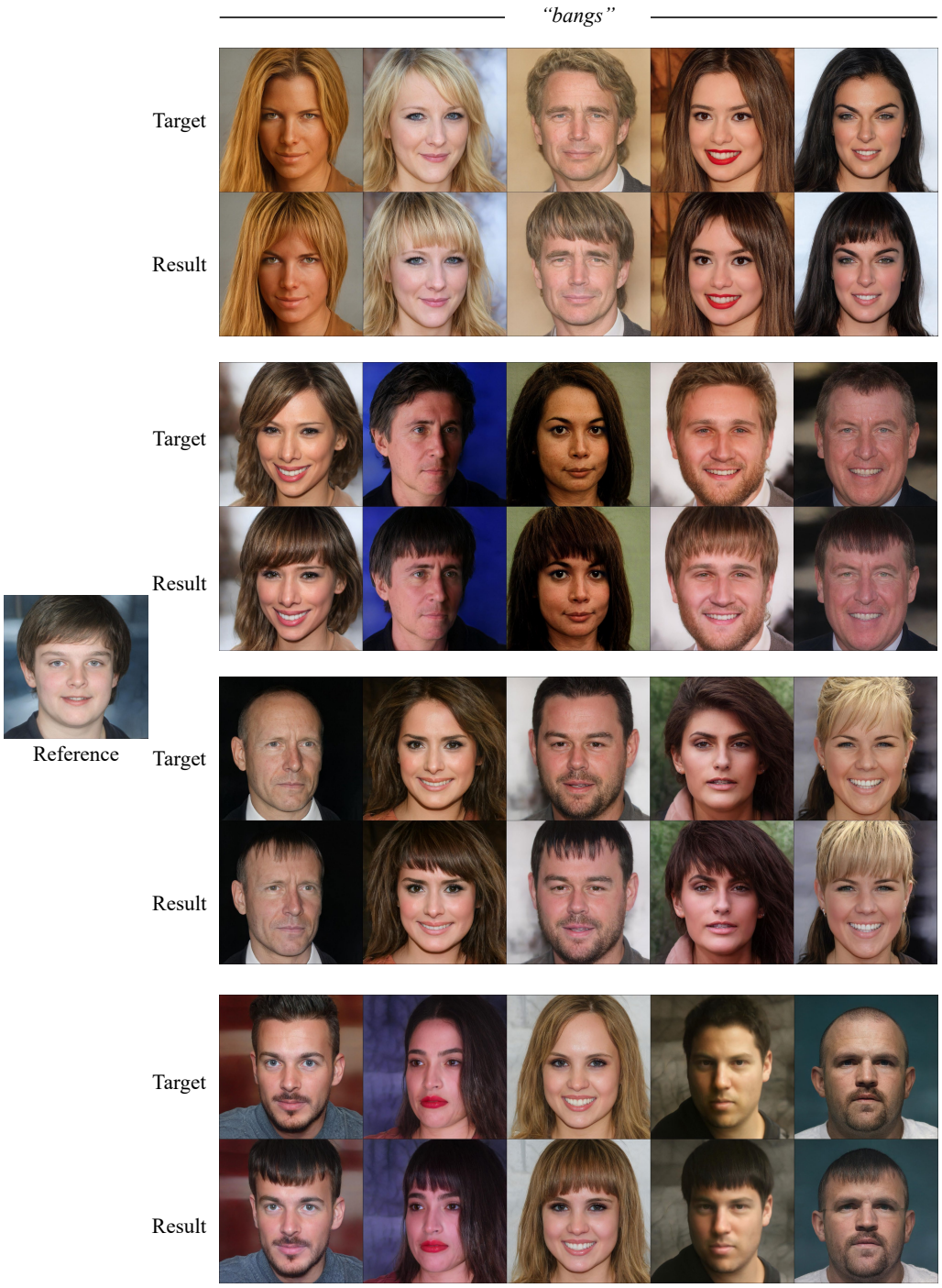

Figure 10: The experiment results on attribute *"bangs"*.

B.2 RESULTS ON IRRELEVANT ATTRIBUTES.

To further verify the effectiveness of our SDD-Net, we design this experiment to transfer irrelevant attributes while preserving the relevant attribute, as illustrated in Figure 11 and Figure 12. We enforce the semantic feature of the manipulated images to change to y_{goal} along the projected vector $\vec{\mathcal{V}}_p$ in CLIP-space. Here we use the vertical vector $\vec{\mathcal{V}}_v = \vec{\mathcal{V}_{RT}} - \vec{\mathcal{V}}_p$, to transfer all of attributes except *"smile"*.

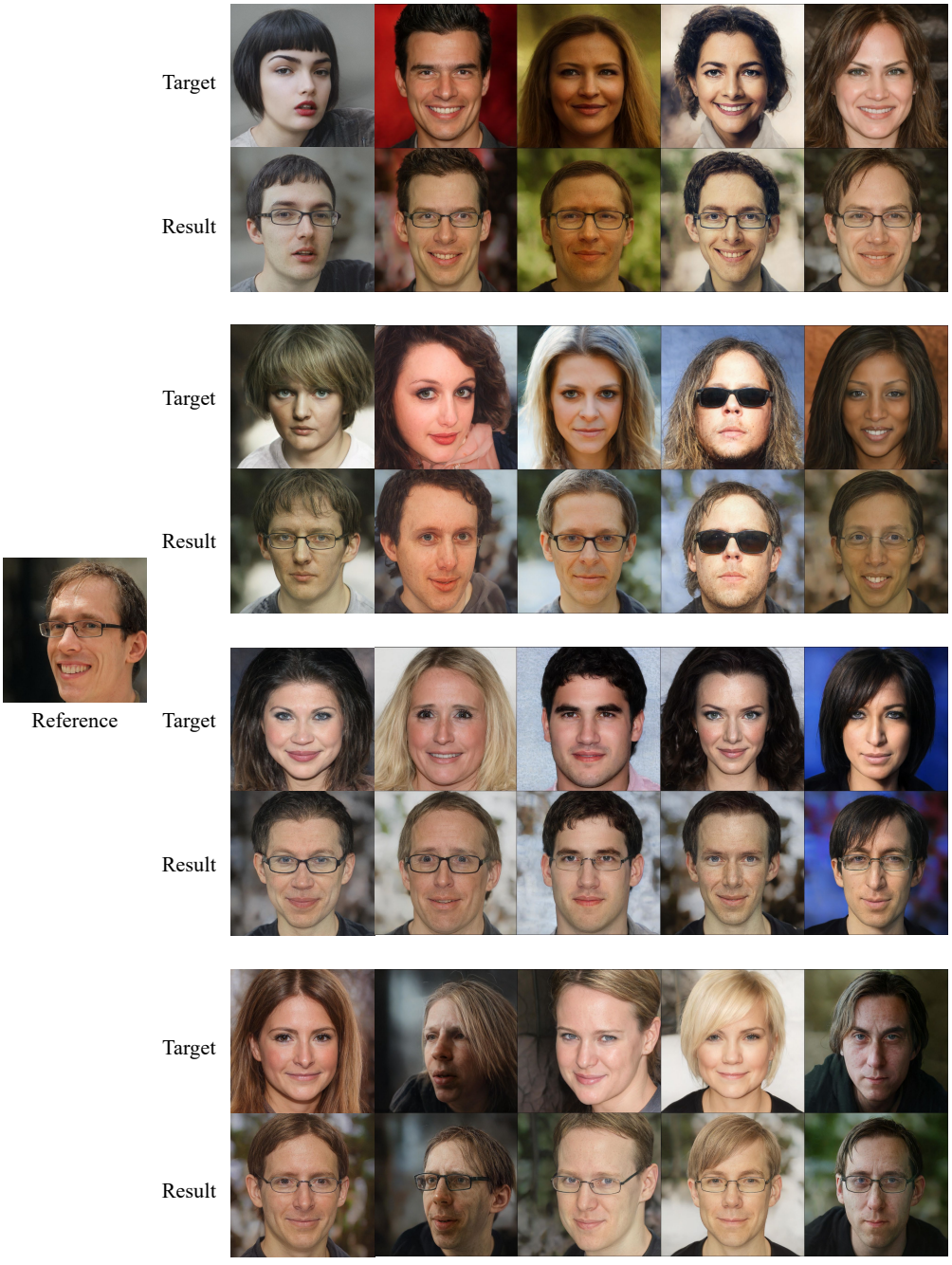

Figure 11: The experiment results on irrelevant attributes transfer. Note that every attributes of the reference, except *"smile"*, are transferred to the target.

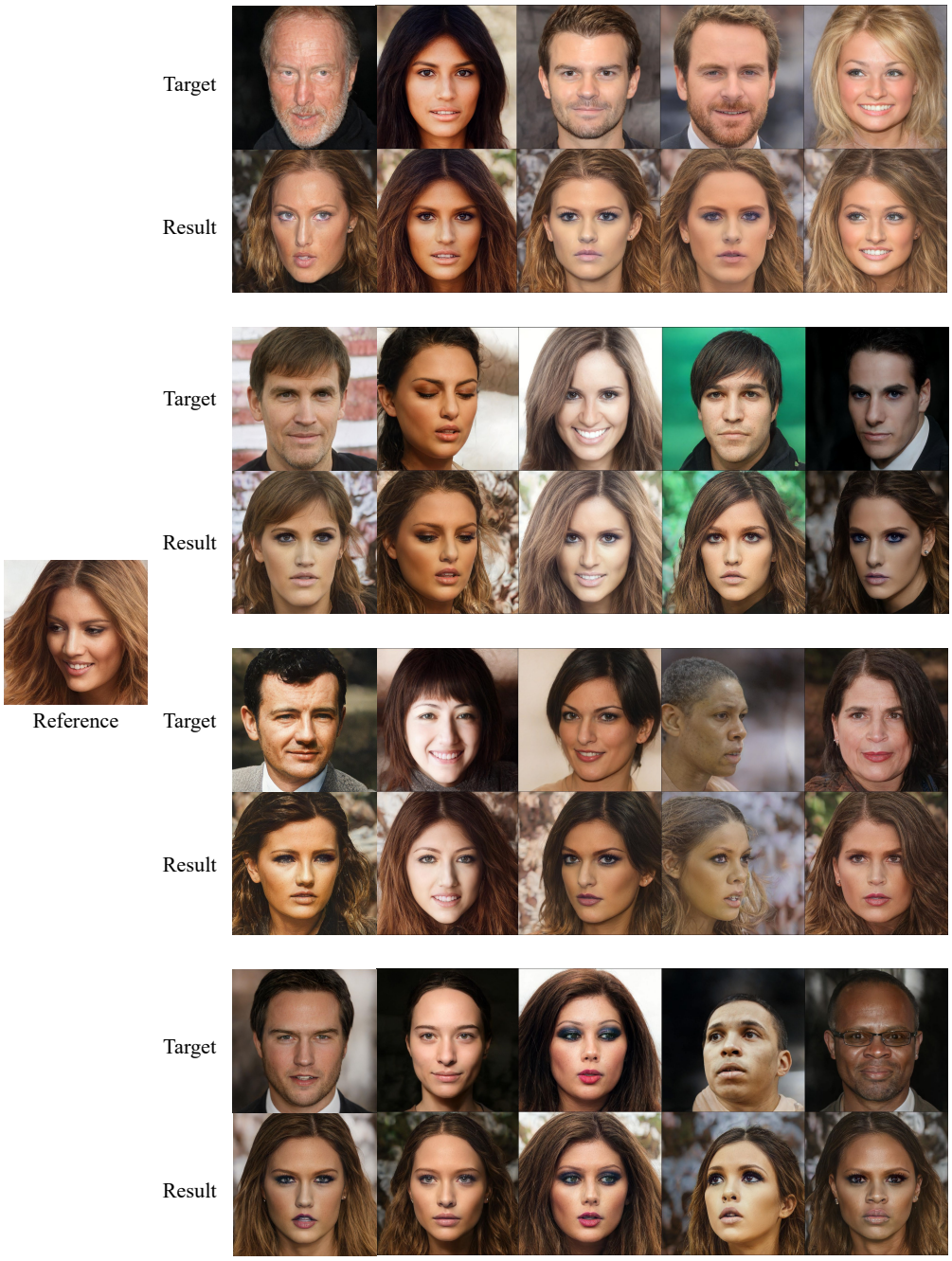

Figure 12: The experiment results on irrelevant attributes transfer. Note that every attributes of the reference, except *"smile"*, are transferred to the target.