DeltaEdit: Exploring Text-free Training for Text-driven Manipulation
–Supplementary Material–

Yueming Lyu\textsuperscript{1,2}, Tianwei Lin\textsuperscript{3}, Fu Li\textsuperscript{3}, Dongliang He\textsuperscript{3}, Jing Dong\textsuperscript{2,}\textsuperscript{*}, Tieniu Tan\textsuperscript{2,4}
\textsuperscript{1} School of Artificial Intelligence, University of Chinese Academy of Sciences
\textsuperscript{2} CRIPAC, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{3} VIS, Baidu Inc. \textsuperscript{4} Nanjing University
yueming.lv@cripac.ia.ac.cn, jdong@nlpr.ia.ac.cn

1. Network Architecture

The detailed architecture of the proposed Delta Mapper is shown in Table 1. In Table 1, “Linear” indicates one fully-connected layer. The input of the Style Module is 6048 style channels in style space. Here, we split it into three levels and feed them into $M^c_s$, $M^m_s$, and $M^f_s$ respectively. The input of the Condition Module is the concatenation of CLIP image-space embeddings, $\Delta i$ and $i_1$, which have the dimensions of 512 respectively. By taking the concatenation of the output of Style Module and Condition Module, Fusion Module outputs coarse-to-fine features with three sub-modules and then concatenates them to produce the final manipulation direction with 6048 channels.

2. Additional Results

In this section, we provide additional qualitative results to further verify the effectiveness of DeltaEdit.

2.1. Attributes Interpolation.

Given two edited latent codes $s_0'$ and $s_b'$ in $S$ space, we can achieve facial attributes manipulation by interpolation. The new edited latent codes is obtained by $s_I = \omega \cdot s_0' + (1 - \omega) \cdot s_b'$, where $\omega$ is the interpolation weight. As shown in Figure 1, we achieve natural and impressive interpolated results from reference text A (“Blond Hair”) to reference text B (“Bangs”), by setting the weight $\omega$ from 0 to 1 in intervals of 0.2.

2.2. More visual results on different datasets.

We show more visual results on FFHQ dataset in Figure 2, Figure 3 and Figure 4. In addition, we provide more visual results on LSUN cat dataset in Figure 5.

2.3. Multi-attributes Generation.

We introduce multi-attribute generation results in Figure 6. Given the target attributes, we can obtain various generating results.

2.4. How text prompts affect results.

For further exploring the effect of different text prompts on manipulation results, we implement experiments on different text prompts settings and show results in Figure 7, Figure 8 and Figure 9.

\textsuperscript{1}Corresponding author.
Table 1. The detailed architecture of the proposed Delta Mapper.

<table>
<thead>
<tr>
<th>Module</th>
<th>Sub-module</th>
<th>Layers in the module</th>
<th>Input shape</th>
<th>Output shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style Module</td>
<td>$M_s^c$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 3, 512)</td>
<td>(1, 3, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_s^{m_1}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 4, 512)</td>
<td>(1, 4, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_s^{m_2}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 2464)</td>
<td>(1, 2464)</td>
</tr>
<tr>
<td>Condition Module</td>
<td>$M_i^c$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 1024)</td>
<td>(1, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_i^{m_1}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 1024)</td>
<td>(1, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_i^{m_2}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 1024)</td>
<td>(1, 2464)</td>
</tr>
<tr>
<td>Fusion Module</td>
<td>$M_f^c$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 3, 1024)</td>
<td>(1, 3, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_f^{m_1}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 4, 1024)</td>
<td>(1, 4, 512)</td>
</tr>
<tr>
<td></td>
<td>$M_f^{m_2}$</td>
<td>4×(Linear, Leaky ReLU)</td>
<td>(1, 4928)</td>
<td>(1, 2464)</td>
</tr>
</tbody>
</table>

Figure 1. Results of facial attribute interpolation. The styles of the interpolated images continuously transfer from attribute A (“Blond Hair”) to attribute B (“Bangs”) by setting the weight $\omega$ from 0 to 1 in intervals of 0.2.
Figure 2. A variety of edits for facial images on StyleGAN2 FFHQ model. The target attribute included in the text prompt is above each column.
Figure 3. A variety of edits for facial images on StyleGAN2 FFHQ model. The target attribute included in the text prompt is above each column.
Figure 4. A variety of hairstyle and hair color edits using our DeltaEdit. The corresponding hair attributes used in the text prompts are listed above each column.
Figure 5. A variety of edits for cat images, using StyleGAN2 pre-trained on LSUN cats dataset. The target attribute is indicated above each column.
Figure 6. The multi-attribute generation results using the DeltaEdit. The leftmost is the target attributes included in the text prompts, columns form left to right represent various generating results.

Figure 7. Manipulation results under different source texts, where the source text is labeled blue and the target text is labeled green. By taking the category name, such as “face”, as the source text, feasible manipulation can be achieved. However, taking some meaningless qualifiers or nouns unrelated to the target category, e.g. “a”, “church”, as the source text, manipulation cannot be realized. Therefore, we choose “face” as the source text in experiments.
Figure 8. Manipulation results under different target texts, where the source text is labeled blue and the target text is labeled green. [Attr] indicates the target attribute. When directly using “[Attr]” as the target text, the manipulation results are entangled with other attributes and are not that accurate to the target attribute. When choosing to use “face has [Attr]” or “face with [Attr]” as the target text, disentangled and accurate results can be achieved. Thus, we use “face with [Attr]” as the target text in experiments.

Figure 9. Manipulation results under different text prompt settings, where the source text is labeled blue and the target text is labeled green. By revising the source text and the target text, an opposite manipulation result can be reached.