Supplementary Material for *Separating Content and Style for Unsupervised Image-to-Image Translation*

A Algorithm of Squeeze-Selection-and-Excitation for Content-style Separation

The input of the SSE module is the feature F_k, which is encoded from the input image. Then this module is aim to separate the content code and style code from F_k. The detailed algorithm is shown in Algorithm 1.

Algorithm 1: Squeeze-Selection-and-Excitation (SSE) for content-style separation

Input: The *k-th* residual block's feature map F^k, selection ratio r
Output: The corresponding content code c^k and style code s^k

$a \leftarrow \texttt{channels\_of}(F^k)$ ▷ a is the number of channels
$T_1 \leftarrow \texttt{adaptive\_average\_pooling}(F^k)$ ▷ Squeeze
$T_2 \leftarrow \Phi_1(T_1)$ ▷ Φ_1 is a Multi Layer Perceptron (MLP)
$I \leftarrow \texttt{index\_of\_descend\_sort}(T_2, 1)$ ▷ I is the channel index
$I_c \leftarrow$ first $\lfloor a * r \rfloor$ elements in I ▷ High correspondence are selected as content
$I_s \leftarrow I \setminus I_c$ ▷ The rest are taken as style
$c^k, t \leftarrow F^k[I_c], F^k[I_s]$
$s^k \leftarrow \Phi_2(t)$ ▷ Φ_2 is an encoder for compressing the style code
$F_k = F_k \cdot T_2$ ▷ Excitation
return c^k, s^k

B More Details About the Valid Units for Interpretation.

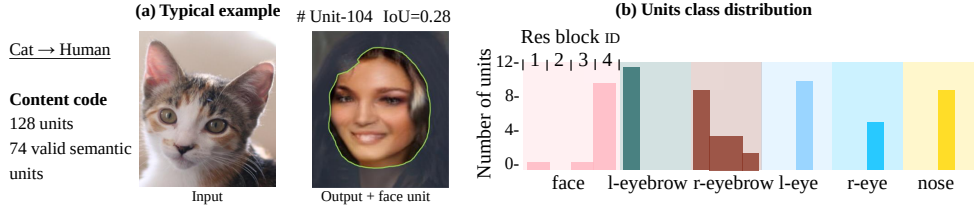

Figure 9: Interpretation of the feature maps in content code, which is learned on Cat⇆Human task. (a) shows the valid face unit focuses on the correct region of the output image. (b) illustrates that all the semantic parts have multiple valid units in content code.

In detail, Fig. 9 (a) shows an example of the activated face features on the generated human image. We find the valid face unit reflects the shared attention face area between the input and output images. There are 6 classes for different semantic parts in Cat⇆Human

dataset: face, left/right eye brow, left/right eye and nose. Fig. 9 (b) shows the units class distribution. Different bins in the same semantic histogram denote different content codes yield from four different residual blocks in encoder. There is one peak value in the face histogram, which indicates that the last residual block yield content code that contains most of the facial features. The residual block ID with larger value extracts more global features [□], it can be interpreted that the facial units gather global information. The other semantic units are illustrated in the other histograms with different colors. The semantic units between different domains exist at similar position of the SCS-UIT. This is because the semantic segmentation is domain-invariant, and the SCS-UIT can learn the similar semantic parts within the same units in the encoder.

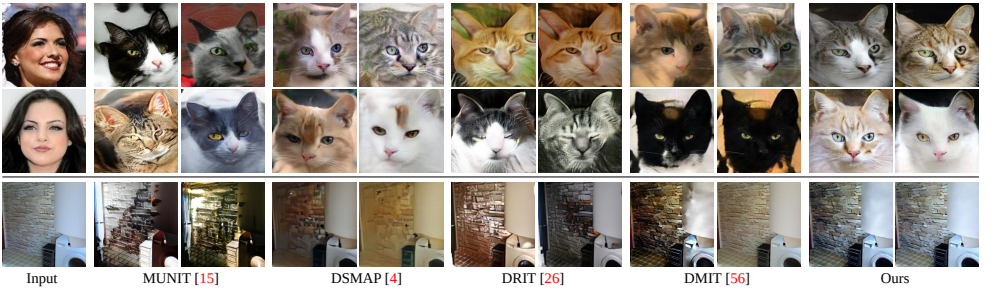

Figure 10: More qualitative comparison on Human to Cat task (Top) and on the CG to Real task (Bottom).

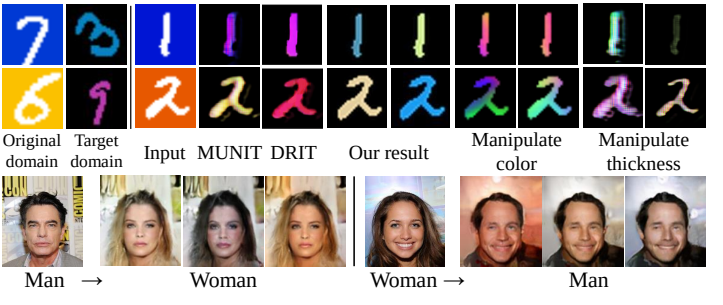

Figure 11: Comparison and manipulation results on Colored-MNIST dataset(top) and multimodal results on Man⇆Woman Task (bottom).

C More Visual Comparison Results

We show more visual results in Fig. 10 on Human to cat task and CG to Real task.

Take the 'Colored MNIST' as an example, which in shown in top row of Fig. 11. To manipulate the styles of the generated images, we add random noise to the intermediate style code by $s_1 = s_1 + e$, where $e \sim U(0,1)$, and s_1 is defined in Fig. 3. we find more than one colors appear on the digits number, which shows more various and possible results than target domain. Furthermore, we can manipulate these units by adding image morphology operation (*e.g.*, dilation and erosion). It can be observed that the digital number become thicker and thinner.