Interactive Avatar Image Manipulation with Unconstrained Natural Language Instruction using Source Image Masking

Seitaro Shinagawa*1*2 Koichiro Yoshino*1*3
Sakriani Sakti*1 Yu Suzuki*1 Satoshi Nakamura*1*2
*1 Nara Institute of Science and Technology
*2 RIKEN, Center for Advanced Intelligence Project AIP
*3 PRESTO, Japan Science and Technology Agency

Summary

What is an easy way to get a desired image?
- **Image retrieval**: the image should be available in database
- **Hand drawing**: requires much time and drawing skills

A potential way: image generation from natural language caption (Caption2image, cap2image) [Reed et al. 2016]

```
"There is a big red chair and a horse stands on the top of the chair."
```

However, cap2image is not good at modification
- Short text input satisfies many images
- Repetition of detailed long text input frustrates users

Main idea: **Image Manipulation with Instruction (IMI)**

natural language instruction represents the difference between source image and target image

```
I make cap2image interactive toward improving usability
Source image masking (SIM) mitigates the unintentional change in generated images generated by IMI model
```

Baseline (w/o SIM) & proposed (w/ SIM) model

![Diagram of baseline and proposed model]

**Why source image masking?**

A naive model suffers from not mentioned change

```
“make the hair large”
“put a glasses”
```

color of background and clothes are also changed!

We hypothesized covering source image with mask preserves not mentioned part in the instruction

Experiments and Discussion

Objective Evaluation

SSIM histogram between generated and target image with whole test set
- **Score of w/ SIM is higher than that of w/o SIM**

Subjective Evaluation

Crowdworker evaluated the preference of generated images in 5-grade between w/o and w/ SIM
- with test set, 3 evaluation on each sample, considering order effect: 230x3x2 = 1380 in total
- Over 60%, w/ SIM was preferred

Case study

<table>
<thead>
<tr>
<th>Preference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B is much better</td>
<td>A and B are equal</td>
</tr>
</tbody>
</table>

Generated images between w/ and w/o SIM

```
given instruction
make his mouth thicker
make the upper earlobes flare outwards
make her hair longer and smooth
```

w/ SIM model can generate a similar image to the target in early time

Data Collection

<table>
<thead>
<tr>
<th>Source target</th>
<th>Crowdworker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe a difference between these images</td>
<td>make his eyebrows smaller</td>
</tr>
</tbody>
</table>

Experimental settings

```
optimizer: Adam(α = 2.0 × 10^{-4}, β = 0.5)
hidden: φi, φdis: 128, φim: 1024, φimm: 512 × 4 × 4
batch size: 64
vocabulary size: 1892
其他 option: feature matching loss to stabilize training
```

Generated images between w/ and w/o SIM

```
phase 1 500 700 1,000 2,000 4,000 4,350 5,000
```

<table>
<thead>
<tr>
<th>Source target</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>“put a black beard that ends before the ears”</td>
<td></td>
</tr>
</tbody>
</table>