RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis from Prior Knowledge

Jun Cheng1,2, Fuxiang Wu1,2, Yanling Tian3, Lei Wang1,2, Dapeng Tao4
1 CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, CAS, China
2 The Chinese University of Hong Kong, Hong Kong, China
3 Graduate School of Information, Production and Systems, Waseda University, Japan
4 School of Information Science and Engineering, Yunnan University, China
{jun.cheng, fx.wu1, lei.wang1}@siat.ac.cn, tianyanling@fuji.waseda.jp, dptao@ynu.edu.cn

A. Training Details in Caption Matching

Figure 1. Examples of positive and negative samples for training: captions in the middle are positive samples, and the right ones are negative samples, where the querying captions {t} are masked as bold.

Figure 2. Accuracy and the training loss: we dynamically construct the samples for each training epoch and evaluation, as described in Section 3.1 of the main paper.

As described in Section 3.1 of the main paper, we treat captions of an image as a positive sample by selecting a caption as the query t. For its negative sample, another query is randomly selected from captions of an image belonging to a different class. For example, in Fig. 1, the captions in a middle black box are a positive sample, where the query t is marked as bold,
and the rest are its positive context. The captions in the right box are its negative sample by selecting a different querying
caption, masked as a different color, as the negative caption. In training, we dynamically construct the positive and negative
samples and evaluate the models every 2.5K samples, denoted as a step, as shown in Fig. 2. For training loss, we exploit
the log sigmoid for numerical stability. The results show the models can distinguish the compatible context of a query from
others effectively.

B. Additional Experiment Results

We present more semantically consistent experiments in Fig. 3 and more results about combining two captions in Fig. 4.
In Fig. 5, we demonstrate the influence of $N_{sel}$ in Our$_F$ and Our$^{SA}_F$. Finally, we show the more enriching examples with the
given captions in Fig. 6.

Figure 3. The images synthesized with the captions: Our$_F$ and Our$^{SA}_F$ synthesize more semantically consistent images with the real images
than DM-GAN by using the original captions, while Our$^{SA}_F$ is better than Our$_F$. 
Figure 4. Examples synthesized by two captions: the bold black words indicate prominent visual details, while the red words indicate conflicting visual details in captions. Words in black boxes are prominent features in generating steps, specifically, $F^1_{att}$ and $F^2_{att}$. 
Figure 5. The synthesized images with increasing captions: the bold words in a caption indicate the prominent features, and \( I-b \) indicates the image is generated with the captions from the first one to the \( b \)-th one.
Figure 6. Synthesized examples by exploiting the recalling captions: given a caption, caption matching will retrieve the compacted items and select their captions, masked as bold, to synthesize images.