**LapsCore: Language-guided Person Search via Color Reasoning**  
*(Supplementary Materials)*

Yushuang Wu\(^{123\ast}\) Zizheng Yan\(^{123\ast}\) Xiaoguang Han\(^{123\dagger}\)  
Guanbin Li\(^{13}\) Changqing Zou\(^{5}\) Shuguang Cui\(^{123}\)  
\(^{1}\)SSE, CUHK-Shenzhen \(^{2}\)FNii, CUHK-Shenzhen \(^{3}\)Shenzhen Research Institute of Big Data  
\(^{4}\)Sun Yat-sen University \(^{5}\)HMI Lab, Huawei Technologies  
{yushuangwu@link, zizhengyan@link, hanxiaoguang@, shuguangcui@}.cuhk.edu.cn  
liganbin@mail.sysu.edu.cn aaronzou1125@gmail.com

1 Implementation Details

**Data Augmentation.** In the IC module pre-training on the CUHK-PEDES dataset, we adopt normalization and random horizontal flip with a probability of 0.5. Besides, the input image brightness is randomly adjusted to diminish the indication of grayscale to the color, by using the Pillow ImageEnhance package. This operation forces the model to learn the color information more from text, rather than from grayscale. For the IC\(_f\) and TC module pre-training, the input image is also normalized and randomly horizontally flipped. Considering the class imbalance in color words prediction, we employ under-sampling to those extremely frequent colors, such as black and white. This operation makes the model learn more rich visual representations. Besides, person images are resized into 384×128 when incorporated into NAFS to keep consistent with the setting of NAFS.

**Architecture Details.** (1) IC Encoder. The output of MobileNet’s first 4 conv\_dw\_s2 layers are taken as the input of Multimodal SE-blocks. (2) IC SE-blocks. The structure design refers to similar modules in Tag2Pix [Kim et al, ICCV19] and ManiGAN [Li et al, CVPR20]. The visual feature goes through an average pooling layer and is concatenated with the textual feature into a vector, from which 2 groups of FC-ReLU are adopted to compute the attention weight vector. (3) IC Decoder. The decoder consists of 4 deconv layers, with a series of DeConv-BatchNorm-ReLU operations in each deconv layer. (4) IC\(_f\) Architecture. The input size of IC\(_f\) encoder is 112×112×64, thus the 4 feature maps output by the encoder are 56×56×128, 28×28×256, 14×14×512, and 7×7×1024, with corresponding adjustment in the decoder. (5) Image/Text Backbone. When replacing the backbones of baselines with a ResNet50 and BERT (CMPM/C and NAFS), we also change the feature extractors in IC and TC module accordingly.

2 Extended Qualitative Results

**Colorization Results on the CUHK-PEDES Dataset.** We visualize some text-guided colorization results of the IC module as in Fig. 1. It is observed that (1) image regions are correctly colorized according to the related textual colors; (2) different body parts or clothes are recognized and localized to colorize separately, e.g., [4, 1] and [3, 4] of Fig. 1; (3) even for some striped or plaid shirts, the model can properly handle, e.g., [1, 5], [4, 2], and [4, 5] in Fig. 1. A more interesting phenomenon is that, as shown in Fig. 2, after the joint training, the incorporated IC module generates better results than before. It indicates that not only CMPM/C achieve gain from the incorporation of IC, but also IC gets improved by the image-text matching task.

**Colorization Results on the CUB and Flowers Datasets.** We also visualize some colorization results on these two datasets to validate the generic effectiveness of LapsCore. Although bird and flower images have more divergent appearance compared with person images, colorization is still completed well enough, as shown in Fig. 3. Different components of birds or flowers are distinguished and colorized accordingly, e.g., the yellow stigma of [2, 1], the red center of [4, 1], the yellow center of [6, 2], the pink leg of [4, 3], the orange beak of [6, 3], the brown feather of [1, 4], the yellow belly of [5, 4], and the red crown of [6, 4] in Fig. 3. Such fine-grained recognition is expected to facilitate representation learning and thus promote retrieval accuracy.

**Extended Retrieval Results.** Due to the space limitation in the paper, here we present more visualized retrieval results on the CUHK-PEDES test set in Fig. 4. Compared with the baseline method, our method has the superiority in the following three aspects: (1) higher accuracy to retrieve the correct persons; (2) correct persons are higher up the rankings; (3) high-affinity persons are more reasonable, even for wrong retrievals.

\(^{\ast}\)Equal contribution  
\(^{\dagger}\)Corresponding author

\(^{\ast}\)The image tuple at the 4th row of the 1st column, similarly hereinafter.
Figure 1: Text-guided image colorization results on the CUHK-PEDES test set. Three columns for each image (from left to right) are gray images, original images, and the colorized ones, respectively.

A man wearing a white shirt, a pair of blue jeans, and a pair of black shoes.

This boy feathery behind a larger man. The boy is stocky in build, he wears a light orange shirt dark blue pants, and athletic shoes.

A man wearing a yellow shirt, a pair of blue shorts, and a pair of dark blue shoes.

The man with tan skin is walking through tall grass facing toward the viewer, he has on a blue jacket white collared shirt underneath and dark pants, he is carrying something in his left hand and his right hand is either up his sleeve or feathery.

A woman with long blond hair wearing a denim jacket with a purple shirt underneath with a black hat and gray pants.

The woman has dark hair pulled back into a ponytail, a bright green collared short sleeved shirt, and dark pants.

Figure 2: Comparisons between colorization results of IC module before (the last column) and after incorporation (the third column). The first two columns indicate the grayscale and original images, respectively.
Figure 3: Text-guided image colorization results on the Flowers and CUB test set. Three columns for each image (from left to right) are gray images, original images, and the colorized ones, respectively.

Man is wearing dark orange shorts sneakers and a white short sleeve shirt with orange and yellow print.

The dark haired woman is wearing white pants, light brown shoes, a white shirt, and a peach sweater.

The man is wearing gray pants and a striped blue and white shirt. His head is looking down toward the floor. A black belt is around his waist.

This woman has long black hair, and she’s wearing an olive green shirt over a white t-shirt and blue jeans.

A man wearing a gray shirt, a pair of gray and white shorts and a pair of white shoes.

Figure 4: Language-guided person image retrieval (top 7) comparison on the CUHK-PEDES test set.