LapsCore: Language-guided Person Search via Color Reasoning

Yushuang Wu[123*] Zizheng Yan[123*] Xiaoguang Han[123†]
Guanbin Li[43] 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
liguanbin@mail.sysu.edu.cn aaronzou1125@gmail.com

Abstract

*The key point of language-guided person search is to construct the cross-modal association between visual and textual input. Existing methods focus on designing multimodal attention mechanisms and novel cross-modal loss functions to learn such association implicitly. We propose a representation learning method for **la**nguage-guided **p**erson **s**earch based on **co**lor **re**asoning (LapsCore). It can explicitly build a fine-grained cross-modal association bidirectionally. Specifically, a pair of dual sub-tasks, image colorization and text completion, is designed. In the former task, rich text information is learned to colorize gray images, and the latter one requests the model to understand the image and complete color word vacancies in the captions. The two sub-tasks enable models to learn correct alignments between text phrases and image regions, so that rich multimodal representations can be learned. Extensive experiments on multiple datasets demonstrate the effectiveness and superiority of the proposed method.*

1. Introduction

Language-guided person search has attracted considerable attention because of its promising application in intelligent surveillance. As shown in Figure 1, it aims to retrieve the person from a large image database that best matches the natural language description query. Compared with image-based and attribute-based person ReID, language queries are easier to obtain than image queries and provide more comprehensive and accurate descriptions than attributes.

There exist two main challenges in the task of language-guided person search. First, it is difficult to compute the visual-textual affinity and construct the image-text alignments, resulting from the cross-modal gap. Secondly, person search is a fine-grained retrieval task: (i) Text provides very detailed descriptions to the target person; (ii) Person images have fine intra-class differences in appearance.

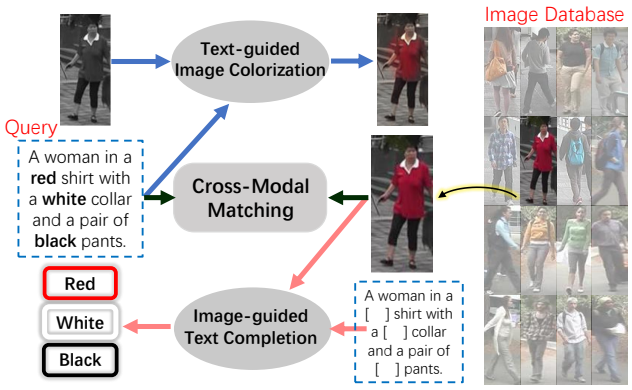

Figure 1: The task of language-guided person search is to retrieve the person that best matches the given textual query from a large image database.

After the pioneering work [20] of language-guided person search, many efforts have been devoted to handling the challenges of this task. [2, 39, 28, 36] design advanced models to learn better representations of image and text. Attention mechanisms are developed in [20, 10, 24, 5] to build the local image-text association. [19, 33, 39, 28] propose novel loss functions to narrow the distance between visual and textual features. However, all of these methods implicitly learn the cross-modal local association, which leaves a rigorous test to the models' learning capability. From numerous experiments of language-guided person search, we observe that colors play a significant role in retrieval. Faced with personal images, human beings tend to accept visual colors to extract the appearance information, and then understand the clothes or ornaments related to these colors. Thus we are inspired to propose a novel representation learning method *LapsCore*, by solving color reasoning sub-tasks, which guide the model to **explicitly** learn **fine-**

*Equal contribution
†Corresponding author

grained cross-modal associations.

As displayed in Figure 1, the first sub-task, text-guided image colorization (IC), is to colorize a gray image according to its text description. In this task, models are facilitated to correctly probe rich color information from the text and align them to the corresponding image regions. For instance in Figure 1, not only the word "red" should be extracted, but also the semantic meaning of "shirt" requires to be paired with "red", and the spatial region in the image indicating "shirt" should be colored in red. Hence the text-to-image local association can be constructed. As for the opposite direction, image to text, the other sub-task image-guided text completion (TC) is designed. Specifically, in each description sentence, all color words are removed, and these vacancies are required to be completed by exploiting the paired colorful images. In this way, valid image regions can be saliently represented and then associated with related text phrases. Although the color reasoning tasks are uncomplicated for human beings, they require models' comprehensive cross-modal understanding to solve them. By using these two sub-tasks, better multimodal representations can be exploited in the main task, image-text matching. Furthermore, we propose another "color" reasoning sub-task, IC_f, aiming to complete image features with missing channels using captions, which generalizes the IC task from image color channel completion into feature semantic channel completion. Given feature representations of the input image, we partially mask some channels, and the caption is utilized to recover them. In this process, general textual information including colors can be probed and exploited. Therefore it endows our method the robustness in cases that colors are not the dominant information in captions.

To tackle the first sub-task IC, we convert it into a pixel-wise regression problem. The original images are processed into gray ones as input, and paired captions are used to recover the original images. The TC task can be treated as a Visual Question Answering (VQA) problem, where the question is a sentence with a color word vacancy and the answer is one of candidate colors. In the image feature channel completion sub-task, we first pre-train a feature extractor on the person ID classification task, then visual feature maps are masked for recovering using captions. Extensive experiments are conducted on the language-guided person search dataset, CUHK-PEDES [20]. The proposed method is proved to produce impressive performance improvements. Verification on general image-text retrieval datasets also confirms its effectiveness, including Caltech-UCSD Birds [26], Oxford-102 Flowers [26], Flickr30k [25], and MSCOCO [21].

In summary, the main contributions of our work include:

- A novel representation learning approach *LapsCore* is proposed to facilitate learning the fine-grained cross-modal association explicitly. It works by solving **color-reasoning** sub-tasks, image colorization, text completion, and image feature channel completion.
- Extensive experiments are conducted on the challenging language-guided person search dataset, CUHK-PEDES. *LapsCore* proves effective to bring considerable performance gain and achieves the **state-of-the-art** results.
- The proposed method is demonstrated as **generic** to be incorporated into different baselines and bring improvements. The effectiveness is also confirmed on other cross-modal retrieval tasks, which is expected to give inspirations to other researchers.

2. Related Work

2.1. Image-text Matching

Early works explore various models and architectures to handle the image-text matching problem. Multimodal convolutional neural networks [23] utilize convolutional architectures to extract image and text features and build cross-modal matching relations. Two-branch neural networks with multiple layers of linear projections are designed in [31] to learn joint embedding, and get applied to address image-text matching in [30]. A recurrent residual fusion block is adopted in [22] to gather visual and textual representations into a more discriminative embedding space. A selective multimodal LSTM is adopted in [9] to measure local similarities between a pair of image and sentence. Some effective loss functions are specifically designed for narrowing the modality gap as well. A cross-modal projection matching loss and classification loss are proposed in [37] to learn the discriminative joint embedding of images and text. In [28], an adversarial loss is introduced to learn the modality-invariant feature representations. More recently, in addition to understanding both modality inputs globally, some methods focus on partial alignments and visual object relationships. In [35], they view the visual inputs as a sequence of objects and attempt to adaptively control the information flow across modalities. Apart from sequences, [32] visual and textual inputs are represented as scene graphs for better cross-modal matching. Another group of works achieves accuracy improvements through introducing external information, for instance, pose information [12], saliency information [11], and consensus knowledge [29]. Our approach works in a novel task-oriented manner to explicitly learn the fine-grained cross-modal association. The color-reasoning tasks can make the model effectively understand colors and thus related objects (clothes, bags, shoes, etc.), in both modalities.

2.2. Language-guided Person Search

Language-guided person search is also known as text-based person ReID or retrieval. The pioneering work [20] first introduces this task. It establishes a large-scale dataset

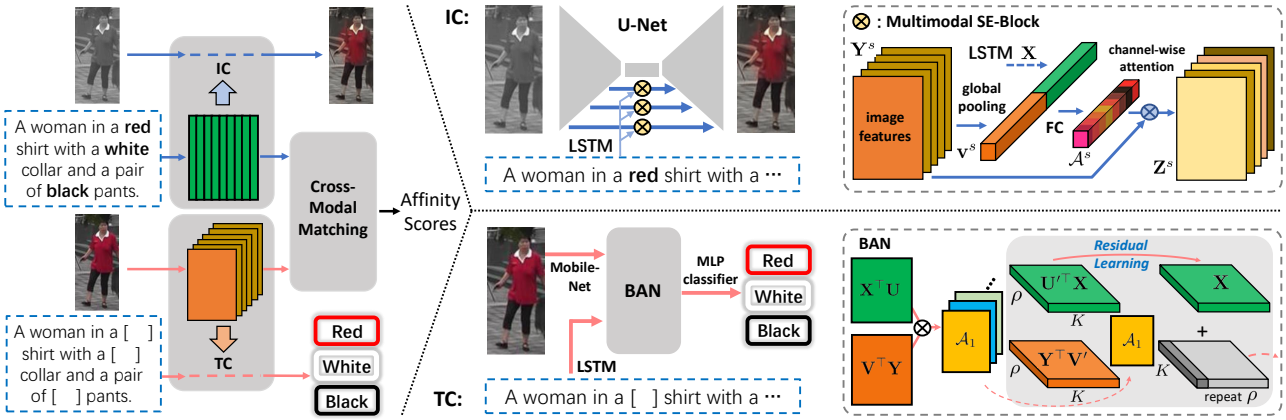

Figure 2: **Left**: the overview of cross-modal matching incorporating a text-guided image colorization task (IC) and an image-guided text completion task (TC); **Upper Right**: the pipeline of IC that adopts an LSTM and a U-Net with multimodal SE-blocks; **Lower Right**: the pipeline of TC that uses a Bilinear Attention Network (BAN).

CUHK-PEDES and proposes a benchmark model GNA-RNN. Based on this work, a two-stage matching framework is proposed in [19], which takes identity information into account and improves performance. Early works contributed to model design include [39], where a dual-path CNN is proposed to project image and text features into the same latent space. Subsequent efforts work on more effective mechanisms to probe the cross-modal association. Authors of [10] design a description-strengthened fusion-attention network to make the discriminative words visually sensitive. A patch-word matching model with an adaptive threshold mechanism is used in [2] to compute text-image affinity. In [24], a multi-granularity image-text alignments model are deployed to explore global-global, global-local, and local-local relationships. More recently, attributes information [1, 34, 38] proves to be beneficial to cross-modal matching. These methods carefully mine numerous attributes from captions as supervision for attribute classification, so that image or text features are learned with discrimination and aligned indirectly. The state-of-the-art method NAFS [4] computes the cross-modal similarity between multi-scale visual regions and textual words/phrases. In comparison with the methods all above, our method tactfully bridges the cross-modal associations in a task-oriented manner. Thus the learning of multimodal representations is guided and strengthened by the color-reasoning tasks. Moreover, *LapsCore* is established on colors, which leads to fine-grained representations.

3. Methodology

In this section, we introduce the proposed method *LapsCore*. As illustrated in Figure 2 (left part), *LapsCore* works on generating representative multimodal features via two color reasoning sub-tasks, text-guided image colorization (IC) and image-guided text completion (TC).

3.1. Text-guided Image Colorization

The IC task aims to exploit text descriptions to colorize gray images, which are processed from original images into grayscale ones. In this task, the model endeavors to comprehend the caption, and probe valid information for colorization. Thus the text-to-image association can be constructed.

The overall task can be converted into a pixel-wise regression problem. The multimodal regression model, denoted as f_{ic}, takes pairs of a gray image, I_{gray} and a description sentence, T_{color} as input, and outputs the recovered image. Original colorful images I_{color} are set as targets, and a pixel-wise Mean Square Error loss $\mathcal{L}_{ic}$ is used:

$$\mathcal{L}_{ic} = \left\| f_{ic}(I_{gray},\ T_{color}) - I_{color} \right\|_2^2$$

To handle this task, we adopt a U-Net framework, which encodes the gray image, and decodes it into colorful ones by fusing text information, as illustrated in Figure 2 (upper right). In the encoding stage, we extract multi-scale visual features from the input. Denote a feature map of scale s as $\mathbf{Y}^s \in \mathbb{R}^{h_s \times w_s \times c_s}$, where h, w, c indicate the height, width and channel, respectively. In the textual branch, the description sentence is tokenized and fed into an embedding layer. Then an LSTM [7] extracts the textual feature $\mathbf{X} \in \mathbb{R}^N$.

In the decoding stage, the visual features should be fused with textual features for colorization. Thus we design multimodal SE-blocks that apply a channel-wise attention mechanism as in [14, 18], so that text information can take effect on the image feature channels. Operations in the multimodal SE-blocks are illustrated in Figure 2 (the upper-right gray dotted frame). At first, the visual feature map $\mathbf{Y}^s$ is compressed into a feature vector $\mathbf{v}^s \in \mathbb{R}^{c_s}$ through global pooling. Concatenated with the textual feature vector $\mathbf{X}$, $\mathbf{v}_s$ is then fed into a two-layer Multi-layer Perceptron and a softmax layer to generate an attention vector $\mathcal{A}^s \in \mathbb{R}^{c_s}$. Finally, $\mathcal{A}^s$ is utilized to update $\mathbf{Y}^s$ into a multimodal rep-

resentation $\mathbf{Z}^s$ with the same dimensions, written as:

$$\mathbf{Z}_i^s = \mathbf{Y}_i^s \cdot \mathcal{A}_i^s,$$

where the subscript $i \in \{1, 2, \ldots, c_s\}$ indicates the index of channel, $\mathbf{Z}_i^s, \mathbf{Y}_i^s \in \mathbb{R}^{h_s \times w_s}$, and $\mathcal{A}_i^s$ is a scalar.

The decoder of U-Net is made up of several deconvolution layers. At first, the last $\mathbf{Y}^s$ in the encoder goes through the first deconvolution layer to generate a feature map $\mathbf{W}^s \in \mathbb{R}^{h_s \times w_s}$. Each $\mathbf{W}^s$ is concatenated with the SE-block output, $\mathbf{Z}^s$, and passes the deconvolution layer to generate a larger $\mathbf{W}^{s'}$. As the last step, given the $\mathbf{W}^s$ from the last deconvolution layer, a simple upsampling and convolution are employed to predict the target.

3.2. Image-guided Text Completion

The dual task TC requires utilizing colorful images to complete text descriptions with color words vacancies. For each sentence, all color words are removed to create a 'colorless' description. And these vacancies should be filled through analyzing foreground colors in different image regions. In this way, image-to-text relations can be bridged.

This task can be treated as a VQA problem. The VQA model, denoted as f_{tc}, takes a colorful image, I_{color} and a text sentence with vacancies, T_q as input, and outputs the missing color words. The target answer is the color words T_a removed from the original descriptions. A typical CrossEntropy loss $\mathcal{L}_{tc}$ is employed, formulated as:

$$\mathcal{L}_{tc} = \text{CrossEntropy}\left(f_{tc}(I_{color},\, T_q),\, T_a\right)$$

We refer to the structure of a popular VQA model, Bilinear Attention Network (BAN), to tackle the TC task, and [15] is recommended for more details. See Figure 2 (lower right), visual and textual features are extracted from input data by a MobileNet and LSTM. Denote textual features as $\mathbf{X} \in \mathbb{R}^{N \times \rho}$ and visual features as $\mathbf{Y} \in \mathbb{R}^{M \times \phi}$, where N is the sequence length, ρ is the LSTM output dimension, ϕ indicates the channel number of the MobileNet output, and $M = h \times w$ is the product of spatial dimensions. Given two modality features $\mathbf{X}$ and $\mathbf{Y}$, several bilinear attention maps $\mathcal{A}_g$ are generated by computing the affinity scores between features patches, formulated as:

$$\mathcal{A}_g = \text{softmax}\Big(((\mathbf{1} \cdot \mathbf{p}_g^\top) \circ \mathbf{X}^\top \mathbf{U})\mathbf{V}^\top \mathbf{Y}\Big),$$

where $\mathbf{U} \in \mathbb{R}^{N \times K}$ and $\mathbf{V} \in \mathbb{R}^{M \times K}$ are projection matrices, $\mathbf{1} \in \mathbb{R}^{\rho}$ is an all-one vector, $\mathbf{p}_g \in \mathbb{R}^K$ where g indicates the attention map index, $\mathcal{A}_g \in \mathbb{R}^{\rho \times \phi}$, and $\circ$ denotes Hadamard product.

With the assistance of attention maps, $\mathbf{X}$ and $\mathbf{Y}$ are fused into joint representations. A residual learning approach is used to increase the representational capacity. In the g_{th} residual block, the output $\mathbf{F}_{g+1} \in \mathbb{R}^{K \times \rho}$ is computed as:

$$\mathbf{F}_{g+1} = \mathbf{P}^\top \text{BAN}_g(\mathbf{F}_g, \mathbf{Y}; \mathcal{A}_g) \cdot \mathbf{1}^\top + \mathbf{F}_g,$$

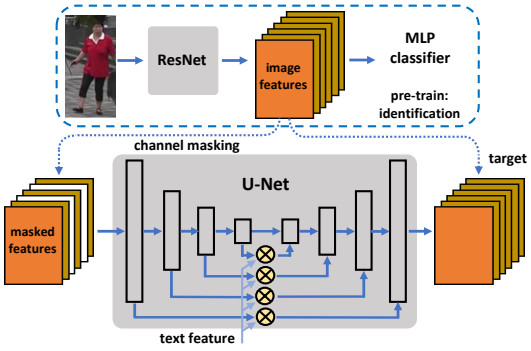

Figure 3: The pipeline of IC$_f$ module for text-guided image feature channel completion. It accepts image features with missing channels as the input and aims to recover them.

where $\mathbf{1} \in \mathbb{R}^{\rho}$ is an all-one vector, and the projection matrix $\mathbf{P} \in \mathbb{R}^{K \times C}$. $\mathbf{X}$ is used as the initial input $\mathbf{F}_0$ by setting N to K [15]. BAN$_g$ is the function for generating intermediate representations, defined as $\mathbf{f}_g = \text{BAN}_g(\mathbf{F}_g, \mathbf{Y}; \mathcal{A}_g)$, where $\mathbf{f}_g \in \mathbb{R}^C$, with its k_{th} element computed as:

$$\mathbf{f}_{g,k} = (\mathbf{F}_g^\top \mathbf{U}')_k^\top \mathcal{A}_g (\mathbf{Y}^\top \mathbf{V}')_k,$$

where $\mathbf{U}' \in \mathbb{R}^{N \times K}$, $\mathbf{V}' \in \mathbb{R}^{M \times K}$, $(\mathbf{X}^\top \mathbf{U}')_k \in \mathbb{R}^{\rho}$, $(\mathbf{Y}^\top \mathbf{V}')_k \in \mathbb{R}^{\phi}$, and the subscript k for the matrices indicates the index of column.

Given joint feature representations output by the last residual block, a Multi-layer Perceptron (MLP) classifier is adopted to predict the color category for each word vacancy.

3.3. Generalized IC: Feature Channel Completion

Colorful images are made up of 3 channels 'YCbCr', and grayscale images are the result of removing two color channels 'Cb' and 'Cr'. Rethinking the IC task, it aims to exploit the textual color information to recover the two missing channels. Although this method can bridge the cross-modal association through colors, text information may not be effectively learned when colors are trivial in descriptions (e.g., in the MSCOCO dataset). Thus we propose a generalized version of IC, denoted as IC$_f$, which turns to complete the missing channels of **image features** using text.

As illustrated in Figure 3, a ResNet18 [6] is pre-trained in an identification task to extract rich representations from images, and "frozen" afterward as a feature generator. We mask some channels of the image features, and feed the masked features into the completion model, with intact features as the target. The completion model and the loss function in IC$_f$ are the same as ones in IC, except that input and output scales are adjusted correspondingly.

3.4. Incorporation

The proposed method can be incorporated into popular image-text matching algorithms [37, 33, 28, 4], as a multimodal representation learning method. The Cross-Modal

Projection Matching and Classification (CMPM/C) model [37] employs a generic framework as in [33, 28], which adopts an LSTM and a MobileNet [8] as the textual and visual feature extractor, respectively. We choose CMPM/C here as the cross-modal matching module to implement *LapsCore*, and it can be easily generalized to other methods of this framework. To incorporate, we remove the feature extraction layers in CMPM/C, replaced by the representation layers of IC and TC modules, as illustrated in the left part of Figure 2. Define the matching loss in CMPM/C as $\mathcal{L}_{cmp}$, then the overall multi-task loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{cmp} + \lambda_1 \mathcal{L}_{ic} + \lambda_2 \mathcal{L}_{tc},$$

where $\lambda_1, \lambda_2 \in \mathbb{R}^+$ are scalar factors that balance the importance of each sub-task. The manner of incorporating IC_f into CMPM/C is similar, with the multi-task loss written as:

$$\mathcal{L} = \mathcal{L}_{cmp} + \lambda_3 \mathcal{L}_{ic_f},$$

where $\lambda_3 \in \mathbb{R}^+$ is a balance factor.

4. Experiments

We evaluate the proposed *LapsCore* on the language-guided person search task in this section. The experimental setup and implementation details are introduced first. Then both quantitative and qualitative results are provided to verify the superiority of *LapsCore*. Finally, ablation studies are conducted for further analysis.

4.1. Experimental Setup

Datasets. The CUHK-PEDES dataset [20] is a challenging dataset of the focused task, which collects 40,206 images of 13,003 person identities from several person identification datasets. Each image is described by two natural language sentences. The training, validation, and test set are made up of 11,003, 3,078, and 3,074 images, and 11,003, 1,000, and 1,000 persons, respectively.

Evaluation Metrics. Recall@k (k = 1, 10) or R@k [13] are used as the evaluation metrics for the focused task. Recall@k indicates the proportion of successful retrievals where at least one ground-truth is included in the top-k scoring images. In extended experiments (Section 5), AP@50 [26] is also used to measure the average precision among all test classes, computed as the ratio of sharing the same class with the query in top-50 scoring results.

Baselines. As mentioned in Section 3.4, we incorporate *Laps-Core* into a generic framework, CMPM/C [37], to verify its generic effectiveness. We also deploy an advanced version CMP_adv, which replaces the MobileNet and LSTM in CMPM/C with a ResNet50 [6] and BERT [3] as the feature extractors. In addition, we also implement a state-of-the-art (SOTA) method NAFS [4] as the baseline to further demonstrate the superiority of *LapsCore* and its capability of contributing to the SOTA algorithm.

Table 1: Recall@k accuracy (%) comparison of different methods on the CUHK-PEDES dataset.

Method	Recall@1	Recall@10
GNA-RNN [20]	19.05	53.64
GLA [2]	43.58	76.26
Dual Path [39]	44.40	75.07
TIMAM [28]	51.30	82.40
TIMAM + BERT [28]	54.51	84.78
ViTAA [34]	55.97	83.52
CMAAM [1]	56.68	84.86
CMPM/C [37]	49.37	79.27
CMP_adv	55.05	85.09
NAFS [4]	61.50	87.51
CMPM/C + TC&IC	53.33	83.20
CMP_adv + TC&IC	57.00	85.62
NAFS + TC&IC	**63.40**	**87.80**

Table 2: Recall@k accuracy (%) of our method incorporated into 3 baselines on the CUHK-PEDES dataset.

	CMPM/C		CMP_adv		NAFS	
	R@1	R@10	R@1	R@10	R@1	R@10
Baseline	49.4	79.3	55.1	85.1	61.5	87.5
+TC	51.8	81.9	56.2	85.3	62.5	87.6
+ IC	52.5	82.8	56.4	85.4	62.7	87.6
+ IC_f	52.7	82.7	56.3	85.4	63.2	87.6
+TC&IC_f	53.0	82.9	56.8	85.4	63.3	87.7
+TC&IC	**53.3**	**83.2**	**57.0**	**85.6**	**63.4**	**87.8**

4.2. Implementation Details

Image Colorization. Gray images are resized into 224× 224 as the input. A MobileNet [8] (pre-trained on ImageNet) encoder is adopted to extract 4 feature maps with varied scales (w, h are equal and set to 56, 28, 14, and 7). The decoder is made up of 4 de-convolution layers. In the text branch, the embedding size and the hidden dimension in the one-layer bi-LSTM are set to 512. A max-pooling operation on the output of all time units is employed to generate the final textual feature vector.

Text Completion. The frequency of all color words are counted first, then those with a frequency over 1,000 (top-14) are selected as the color candidates for completion. In the CUHK-PEDES training set, 95.3% of sentences contain at least one candidate color. For each training sample, we randomly choose one color word and create only one vacancy for prediction. In the BAN model, a MobileNet [8] (pre-trained on ImageNet) is used in the visual branch, and a bi-LSTM with embedding size 512 and one 512-dim hidden layer extracts textual features. 4 glimpses are used as in [15], which leads to adequate accuracy and low complexity.

Image Feature Channel Completion. As mentioned in Section 3.3, a "frozen" ResNet18 [6] acts to generate features, which is pre-trained on the identity classification task.

Figure 4: Comparison of top-7 retrieval results between our method and the baseline, given the same queries. The results are arranged from left to right in descending order of the affinity scores. The red frames indicate the correct retrievals.

On the Flickr30k and MSCOCO datasets, image features are generated by a ResNet50-based Faster R-CNN [27], pre-trained on the COCO2017 object detection task. The conv1 layer output ($112\times112\times64$) is utilized as the completion target. In every two feature channels, one is set to zeros to generate masked features. The encoder in the IC_f module employs a ResNet50 on the Flickr30k and MSCOCO datasets, and a MobileNet for other datasets.

Training and Testing. Before joint training, either IC or IC_f module is pre-trained for 20 epochs using an Adam [16] optimizer, with a learning rate of 0.001. The TC and CMP module are pre-trained for 10 and 20 epochs with a learning rate of 0.0002, respectively, using Adam as well. Then all modules are jointly trained for 40 epochs using Adam. The mini-batch size is set to 64 and the learning rates are set to 0.0002. In the loss function, λ_1 and λ_2 are set as 10 and 1, respectively. In the testing phase, all text and image features are extracted and the cosine similarity between all image-text pairs can be computed.

4.3. Experiment Results

Quantitative Results. We evaluate our method based on all three baselines. Performance comparisons with existing SOTA algorithms are presented in Table 1. Numerical results demonstrate that our method can consistently bring improvements to different baselines. For the generic CMPM/C [37], the proposed method can bring a considerable 0.04 gain in Recall@1 and Recall@10. Comprehensive results are listed in Table 2 to demonstrate the effectiveness of IC, TC, and IC_f module separately. IC is usually competitive with IC_f, and incorporating both IC and TC enable further improvements than using either one only. Furthermore, *LapsCore* proves effective to improve the SOTA method NAFS [4], and achieves a 63.40% Recall@1 rate, which surpasses the SOTA performance by around 0.02.

Qualitative Results. Given the same language queries, retrieval results of the baseline (CMPM/C) and our method (CMP+IC&TC) are visualized in Figure 4. In comparison, our method is more effective to retrieve the matched persons (the first row). It also uncovers that *LapsCore* enables the model's sensitivity to colors, which makes the retrieval results more reasonable. For instance in the second row, most of the top-scoring images retrieved by our method satisfy "light orange shirt" and "dark blue pants", while the baseline can not. A similar case can be observed in the third row of Figure 4. Besides, our model can well colorize the input gray images, even for unseen test images. Related colorization visualization is included in supplementary materials. More impressively, a further experiment is conducted by altering the color words in captions to change the colorization of the same image. As displayed in Figure 5, our model exactly learns the meaning of colors and associates with related regions, rather than from simple memorization. We also output the "segmentation" map as a by-product, by computing the distance of the last layer's outputs when alternating color words, shown in the last column of Fig-

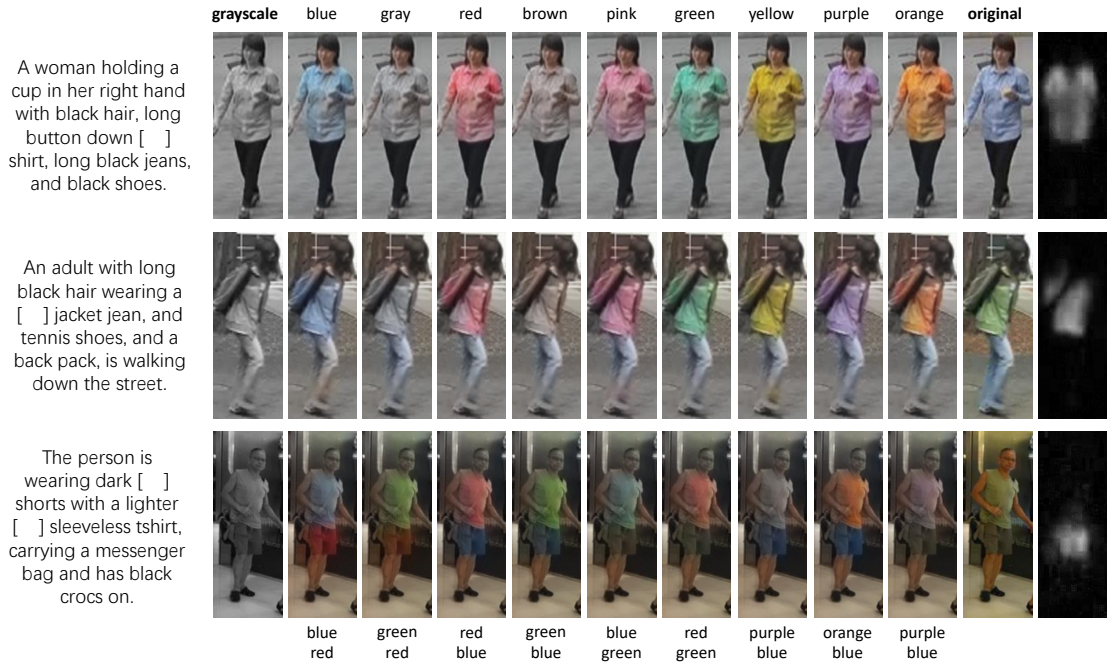

Figure 5: Visualization of colorization results by changing the color words in captions. Top-9 frequent colors (apart from white and black) are chosen for visualization. Images in the last row are colorized with different color combinations.

Table 3: Results of 3 groups of ablative experiments.

	Recall@1	Recall@10
CMPM/C [37]	49.4	79.3
+ IC-Sketch	51.5	82.7
+ IC-Grayscale	**52.5**	**82.8**
+ TC-Object	50.1	81.7
+ TC-Color	**51.8**	**81.9**
+ IC$_f$-Deeper	51.2	82.7
+ IC$_f$-Spatial	51.6	82.4
+ IC$_f$-Mask4	51.7	82.7
+ IC$_f$-Mask2	**52.7**	**82.7**

ure 5. It can be observed that the model learns to segment upper/lower body clothes implicitly through color reasoning tasks, which may inspire future research.

To investigate the effect of some settings in the design of color reasoning tasks, we perform a series of experiments for analysis. A CMPM/C model [37] with a MobileNet backbone is adopted as the baseline, and all the following ablative experiments are conducted on the CUHK-PEDES dataset [20]. We chooses one important variant for each module design of *LapsCore*, the source of colorization in IC, the vacant word type for completion in TC, and the feature selection and masking manners in IC$_f$.

4.4. Ablation Studies

Gray Images in IC. Apart from grayscale images, sketch images, as another kind of gray image, can be an alternative as the source of the colorization sub-task. Sketch images

Figure 6: **Left**: Word cloud of color dictionary and object dictionary. The bigger font size indicates the larger word frequency; **Right**: Visualized examples of grayscale and sketch images generated from original colorful images.

further give up much grayscale information and retain only contours, as shown in Figure 6. Thus it increases the difficulty of colorization, which may be detrimental to the effectiveness of learning. Experiments are performed to compare the difference between using the sketch and grayscale images as the colorization source. As shown in Table 3, using grayscale images leads to higher performance.

Color Words in TC. Some objects are of great importance in the description of pedestrian appearance. Specifically, words like "glasses", "hat", "dress" in the query may filter out numerous unrelated images and act as the retrieval key. Based on this observation, we construct an object dictionary with 26 most frequent object nouns, as visualized in Figure 6, which cover 96.8% of query sentences. However,

Table 4: Recall@1 accuracy (%) and AP@50 comparison of different methods on the CUB and Flowers datasets.

Method	CUB		Flowers	
	Img2Txt R@1	Txt2Img AP@50	Img2Txt R@1	Txt2Img AP@50
GMM+HGLMM [17]	36.5	35.6	54.8	52.8
Word CNN-RNN [26]	56.8	48.7	65.6	59.6
IATV [19]	61.5	57.6	68.4	70.1
CMPM [37]	64.6	62.1	67.7	66.1
CMPM/C [37]	67.9	64.3	69.7	68.9
CMP_adv	70.3	66.4	75.7	72.2
CMPM/C + IC$_f$	67.9	64.6	73.2	69.2
CMPM/C + TC&IC	68.0	66.0	75.2	71.4
CMP_adv + IC$_f$	71.1	67.1	76.5	72.4
CMP_adv + TC&IC	**72.3**	**69.5**	**77.9**	**73.3**

using the object dictionary to replace the original color dictionary results in lower accuracy, as shown in Table 3. One of the reasonable explanations is that most of the objects have a more complex semantic meaning than colors, which makes it harder to learn the cross-modal association.

The Settings in IC$_f$. In our basic setting, the image features for completion are selected from the ResNet18's conv1 layer output. Then one in every two feature channels is set as zeros to generate masked features. Ablation experiments are performed on 3 aspects, the features chosen for completion, the extent of masking, and the manner of masking. First, deeper features, the output of ResNet18 conv2_x is set as the target, which causes a Recall@1 drop by 0.015. Intuitively, deeper features are more abstract and thus difficult to complete. Secondly, we mask one channel in every four channels to reduce the completion difficulty. This modification also results in a decrease of 0.01 in Recall@1, resulted from the greatly increased difficulty of recovering. Thirdly, the image features are masked spatially, rather than channel-wisely. In this way, the performance is also influenced. One potential reason is the high correlation between spatial feature patches makes the completion task too easy. All results above are listed in Table 3 for comparison.

5. Extended Experiments

In this section, we implement extended experiments on other 4 datasets to prove the versatility of *LapsCore* to be applied in other cross-modal retrieval tasks.

CUB and Flowers Datasets. The Caltech-UCSD Birds (CUB) [26] and the Oxford-102 Flowers (Flowers) [26] datasets collect images of various species of birds and flowers, respectively, with 10 descriptions per image, and training set shares no categories with the test set. Experiments on the two datasets further verify the effectiveness of the proposed method. *LapsCore* is implemented on two baselines, CMPM/C and CMP_adv, with quantitative results listed in Table 4. It is observed that on both two datasets and in both retrieval directions, two baselines can obtain

Table 5: Recall@k accuracy (%) of the baseline and our methods on the Flickr30k and MSCOCO datasets.

Method	Flickr30k					
	Img2Txt			Txt2Img		
	R@1	R@5	R@10	R@1	R@5	R@10
CMPM [37]	37.1	65.8	76.3	29.1	56.3	67.7
CMPM/C [37]	40.3	66.9	76.7	30.4	58.2	68.5
CMPM/C + IC$_f$	**44.1**	**70.1**	**80.0**	**31.1**	**59.4**	**71.0**
Method	**MSCOCO**					
	Img2Txt			Txt2Img		
	R@1	R@5	R@10	R@1	R@5	R@10
CMPM [37]	23.9	51.5	65.4	18.9	43.8	56.9
CMPM/C [37]	24.6	52.3	66.4	19.1	44.6	58.4
CMPM/C + IC$_f$	**29.1**	**59.1**	**71.8**	**20.7**	**48.0**	**61.3**

improvements with the incorporation of *LapsCore*. Qualitative results of colorization output visualization are also provided in supplementary materials.

Flickr30k and MSCOCO Datasets. Unlike all above datasets that include one single primary category (person, bird, or flower), images in the Flickr30k [25] and MSCOCO [21] datasets contain a wide range of components, and captions are also more comprehensive. In addition, both datasets contain large amounts of general images and captions, where colors are not dominant (only around 1/3 sentences contain frequent color words). Experiments on the two datasets aim to evaluate the effectiveness of the IC$_f$ module. As shown in Table 5, the proposed method brings an impressive gain of the recall rates in both retrieval directions. Especially in the image-to-text direction, the IC$_f$ module brings more than 4% improvements to CMPM/C in Recall@1 on both two datasets.

6. Conclusion

In this paper, we proposed *LapsCore*, which uses two color reasoning sub-tasks to improve representation learning for language-guided person search. The first one aims to exploit text information to colorize a gray image. And in a dual direction, the colorful image is utilized to complete color word vacancies in the caption. In addition, we propose to complete visual feature channels, which is applicable in the general image-text matching task where colors are not dominant in captions. Both quantitative and qualitative experimental results, with extensive ablation studies as well, demonstrate the superiority of the proposed approach.

Acknowledge The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, and by Guangdong Research Project No. 2017ZT07X152,.

References

[1] Surbhi Aggarwal, Venkatesh Babu RADHAKRISHNAN, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2617–2625, 2020. 3, 5

[2] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1879–1887. IEEE, 2018. 1, 3, 5

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[4] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021. 3, 4, 5, 6

[5] Jing Ge, Guangyu Gao, and Zhen Liu. Visual-textual association with hardest and semi-hard negative pairs mining for person search. *arXiv preprint arXiv:1912.03083*, 2019. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[9] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017. 2

[10] Zhong Ji, Shengjia Li, and Yanwei Pang. Fusion-attention network for person search with free-form natural language. *Pattern Recognition Letters*, 116:205–211, 2018. 1, 3

[11] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5754–5763, 2019. 2

[12] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. *arXiv preprint arXiv:1809.08440*, 2018. 2

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5

[14] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9056–9065, 2019. 3

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 4, 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. 8

[18] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 3

[19] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899, 2017. 1, 3, 8

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, 2017. 1, 2, 5, 7

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 8

[22] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4107–4116, 2017. 2

[23] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015. 2

[24] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *arXiv preprint arXiv:1906.09610*, 2019. 1, 3

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 8

[26] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016. 2, 5, 8

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 6

[28] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5814–5824, 2019. 1, 2, 4, 5

[29] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. *arXiv preprint arXiv:2007.08883*, 2020. 2

[30] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 2

[31] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[32] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020. 2

[33] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language person search with mutually connected classification loss. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2057–2061. IEEE, 2019. 1, 4, 5

[34] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *European Conference on Computer Vision*, pages 402–420. Springer, 2020. 3, 5

[35] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773, 2019. 2

[36] Shizhou Zhang, Yifei Yang, Peng Wang, Xiuwei Zhang, and Yanning Zhang. Attend to the difference: Cross-modality person re-identification via contrastive correlation. *arXiv preprint arXiv:1910.11656*, 2019. 1

[37] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701, 2018. 2, 4, 5, 6, 7, 8

[38] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. Hierarchical gumbel attention network for text-based person search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3441–3449, 2020. 3

[39] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017. 1, 3, 5