Self-Enhancement Improves Text-Image Retrieval in Foundation Visual-Language Models

Yuguang Yang[1] Yiming Wang[1,2] Shupeng Geng[1] Runqi Wang[1]
Yimi Wang[1] Sheng Wu[1] Baochang Zhang[1,3,*]
[1]Beihang University [2]Shanghai Jiao Tong University [3]Zhongguancun Laboratory
{guangbuaa, yiming.wang, gengshupeng, bczhang}@buaa.edu.cn

Abstract

*The emergence of cross-modal foundation models has introduced numerous approaches grounded in text-image retrieval. However, on some domain-specific retrieval tasks, these models fail to focus on the key attributes required. To address this issue, we propose a self-enhancement framework, **A³R**, based on the CLIP-ViT/G-14, one of the largest cross-modal models. First, we perform an **Attribute Augmentation** strategy to enrich the textual description for fine-grained representation before model learning. Then, we propose an **Adaption Re-ranking** method to unify the representation space of textual query and candidate images and re-rank candidate images relying on the adapted query after model learning. The proposed framework is validated to achieve a salient improvement over the baseline and other teams' solutions in the cross-modal image retrieval track of the 1st foundation model challenge without introducing any additional samples. The code is available at https://github.com/CapricornGuang/A3R.*

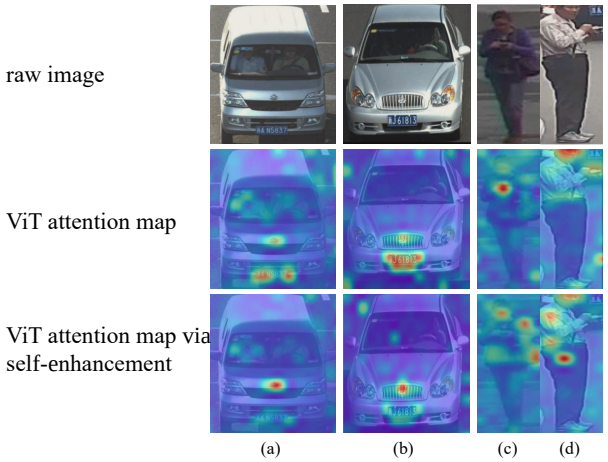

Figure 1. Raw images and their corresponding attention maps of CLIP-ViT/G-14 before and after self-enhancement. Before attribute augmentation, the model primarily attends to sample-specific features, such as vehicle accessories, license plates, and human heads. But after attribute augmentation, the model demonstrates a notable shift towards class-specific features, specifically vehicle logos, human attire, and backpacks.

1. Introduction

Image retrieval has been widely applied in automatic public video surveillance that obtains several relevant images from vast image databases based on user queries [3]. Traditional retrieval methods rely on attribute recognition to identify the desired images [6, 9], which struggle to handle the customized retrieval query and thus lack the flexibility to various user needs. Recently, cross-modal foundation models have gained popularity for their ability to unify text and image representations [1, 7, 10], the large-scale pretraining data equips them with the generalization ability to handle a wide range of real-world scenes, so as to enable them with seamless translation between text and images, enhancing the flexibility and adaptability of the retrieval process.

However, as shown in Figure 1, on some class-specific retrieval tasks, the large-scale pretraining makes these foundation models focus more on the sample-specific attributes rather than the class-specific attributes required. To address this issue, we propose a self-enhancement framework in this work, aiming to deploy the visual-language foundation model in the context of traffic retrieval to further improve the accuracy of text retrieval for images. We perform data and retrieval augmentation before and after model learning, respectively: (i) We recognize that the most valuable information benefiting fine-grained retrieval is the attribute description, so we utilize the rich prior knowledge of foundation models to perform zero-shot ***Attribute Augmentation*** to augment the textual description with various attributes;

*Corresponding author. This paper was partially supported by the "One Thousand Talents Plan" innovation leading talent funding projects in Jiangxi Province, China, and National Innovation and Entrepreneurship Training Municipal Project, China.

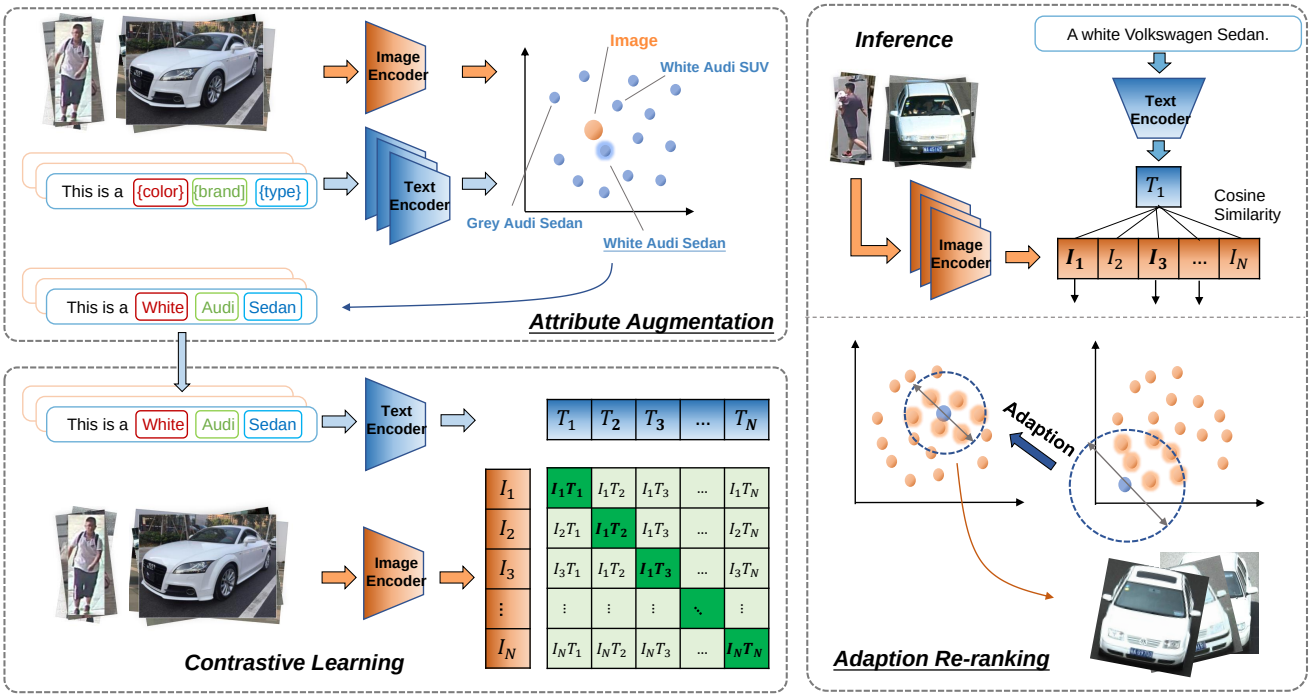

Figure 2. Full pipeline of A^3R framework. *Attribute Augmentation* and *Adapation Re-ranking* are our self-enhancement method.

(ii) We note that similarities between candidate images often overshadow the similarities between images and textual queries in cross-modal retrieval, so we propose the ***Adaption Re-ranking*** method to extract query-specific information from candidate images and instead leverage the original textual query as a similarity constraint. We first align the representation space between the re-ranking query and candidate images, then perform cross-modal re-ranking to retrieve the final expected images. By combining these two strategies, We name our framework $\boldsymbol{A^3R}$. Experiments show that our A^3R framework enables models to achieve salient improvements over the pure fine-tuning paradigm.

2. A^3R: Self-Enhancement Framework

We resort to CLIP-ViT/G-14 [10] for our backbone (§2.1), which achieves dramatic cross-modal performances in many tasks. However, it performs poorly in domain-specific tasks like cross-modal retrieval. Therefore, we propose a zero-shot self-enhancement framework based on attribute augmentation (§2.2) and adaption re-ranking (§2.3) to improve the capability of fine-tuned visual-language models. Figure 2 illustrates our framework.

2.1. Architecture and Formulation

We formulate each sample as a package of image, text, and attribute (I, T, A), where attributions can be viewed as the subset of texts that present the most distinguishing information, such as "*color*", etc. First, I and T are input to image and text encoders, respectively, which are both transformer-based modules generating image and text embeddings w^I and w^T. Then, contrastive learning is employed in the following training. Specifically, assume that N encoded samples $\{(w_i^I, w_i^T)\}_{i=1}^N$ are processed in parallel. Let image embedding matrix be $M_I = \{w_i^I\}_{i=1}^N$ and text embedding matrix be $M_T = \{w_i^T\}_{i=1}^N$, then the text-image similarity matrix F are formulated as

$$F = (M_I)(M_T)' \in \mathbb{R}^{N \times N}. \tag{1}$$

The fine-tuning target is to maximize N matched pairs' similarity and minimize that of $N^2 - N$ mismatched pairs.

2.2. Attribute Augmentation

We note that the most valuable information in the texts is the attributes. However, they are often missing in the labels of vehicle datasets. Therefore, we perform zero-shot data augmentations aimed at filling in the missing attribute elements in the samples. We focus on three attribute elements: *color*, *brand*, and *type*. For the missing attribute element, we exhaust all possible cases of this attribute element and concatenate them after the existing text T, respectively, and then pass all concatenated text and the corresponding only image into the CLIP-ViT/G-14 model to obtain embeddings of the image and all texts. We select the text with the highest image-text cosine similarity as the augmented text T.

2.3. Adaption Re-ranking

Re-ranking is vital for improving retrieval performance [2]. This process typically involves two key steps: re-sorting query embedding and similarity refinement. However, in cross-modal retrieval, the re-sorting query can not be directly taken from the textual query due to the contrastive-based language-image alignment strategy. This strategy aligns different modalities without emphasizing the need for shared representation spaces, leading to the similarities between candidate images often overshadowing the similarities between images and textual queries.

To address this, we propose Adaption Re-ranking, a plug-and-play non-parameter modal adapter utilizing singular vector decomposition (SVD) to extract query-specific information from candidate images considering the text-image similarity. Specifically, given the encoded textual query embedding $M_q = \{w_0^T\}$ and M candidate encoded image embeddings $M_c = \{w_i^I\}_{i=1}^M$ processed in parallel, we first compute the query-candidate similarity as follows:

$$S = (M_c)(M_q)' \in \mathbb{R}^{M\times 1}. \tag{2}$$

To simplify computations, we broadcast the query-candidate similarity metric S into $\widetilde{S} = \{S\}_{i=1}^D \in \mathbb{R}^{M\times D}$. SVD is then performed on M_c under the constraint of similarity S, yielding:

$$M_s = M_c \odot \widetilde{S} = U\Sigma V^T, \tag{3}$$

where $U \in \mathbb{R}^{M\times M}$, $V \in \mathbb{R}^{D\times D}$ are unitary matrixes, and $\Sigma \in \mathbb{R}^{M\times D}$ is a diagonal matrix whose diagonal entries correspond to the singular values of M_s sorted in descending order. By utilizing the column vector $U_1 \in \mathbb{R}^{M\times 1}$ of U *w.r.t.* the largest singular value, the principal vector of the encoded candidate images $M_q^* \in \mathbb{R}^{1\times D}$ can be represented as $M_q^* = U_1' M_c$, which is used as the re-ranking query.

Subsequently, we incorporate the k-reciprocal encoding algorithm [5] to further refine the re-ranking results. The k-reciprocal encoding algorithm is a well-established re-ranking technique. It aims to calculate a new similarity measure between a query image and a candidate image, based on their k-reciprocal nearest neighbors:

$$M_c^* = k\text{-reciprocal}(M_q^*, M_c). \tag{4}$$

3. Experiments

3.1. Setup

Dataset. We use the open-source **PA100k** pedestrian dataset [4] and the **BIT-Vehicle** vehicle dataset [8]. Unlike previous approaches relying on one-hot encoded attribute labels, these datasets incorporate attribute-level natural language text annotations corresponding to the images. The pedestrian dataset is constructed by human images of 21 unique attributes concerning sex, dress, and walking position, while the vehicle dataset contains auto images of 11 colors, 6 vehicle types, and 65 vehicle brands. We note that the vehicle images in the test set were obtained through web scraping, resulting in significant variations in data distribution, while that of pedestrian images is relatively consistent.

Metric. The evaluation metric used is the mean Average Precision ($mAP@K$). Here, K indicates that the top$K$ retrieval results are used in the evaluation. The calculation of $mAP@K$ is as follows:

$$mAP@K = \frac{1}{m} * \sum_{i=1}^{K} p(i) * \Delta r(i), \tag{5}$$

where m is the total query times in the evaluation set, $p(i)$ and $r(i)$ denotes the precision and recall of the top$i$ retrieval results, respectively, $\Delta r(i) = r(i) - r(i-1)$, and $r(0) = 0$.

Implementation. We adopt the 2B-parameter CLIP-ViT/G-14 [10] as our backbone. To address the variations in image proportions between vehicles and pedestrians, we apply zero-padding and resize all images to a size of 224. The training process utilizes the Adam optimizer with a batch size of 25. Given the significant distribution disparities between the training and test sets, the performance improvement on the validation set can not be reflected on the test set directly. Therefore, we employ a small learning rate of 4e-7 and perform only 5 epochs of fine-tuning. The training process utilizes one A100 GPU, while the inference process is executed on one NVIDIA RTX 3090 GPU.

3.2. Main Results

The leaderboard ranking is determined based on the calculation of $mAP@10$. Our competition performance is shown in Table 1. Our proposed method ranks 6th on the leaderboard with a score of 0.75027 without introducing any additional datasets. These results indicate that A^3R can capture both textual and visual information, enabling accurate retrieval in real-world scenarios.

Rank	Team Name	Score	Rank	Team Name	Score
1	MiniModel	0.82382	**6**	**IPCL(ours)**	**0.75027**
2	njust	0.82223	7	VIPS	0.74710
3	DiamondH	0.81990	8	432	0.73654
4	CASHIPS	0.76865	9	BBH~	0.72753
5	HZHv2	0.76268	10	mARapper	0.72697

Table 1. Leaderboard A of the Cross-modal Track in the Foundation Model Challenge.

3.3. Analysis

Re-ranking. Figure 3 illustrates the effect of the re-ranking algorithm. As shown in sub-figure (a), the vanilla

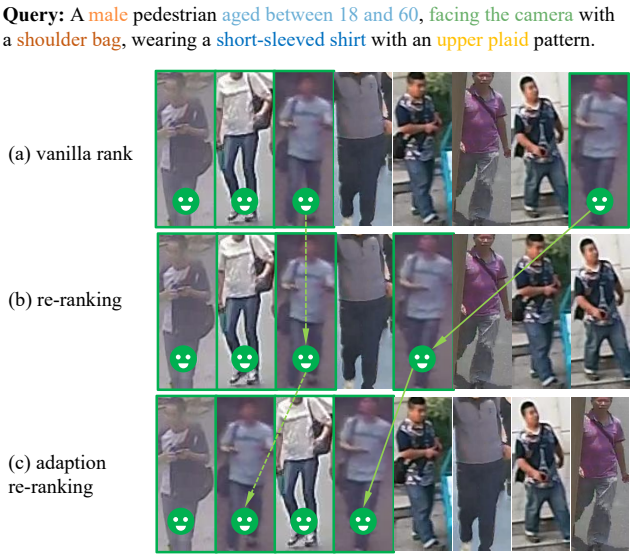

Figure 3. Visualization of the re-ranking algorithm. ☺ indicates the query-matched objects.

searching results have been already accurate to some extent, but some relevant objects are not ranked at the top. Comparing sub-figure (b) and (c), we observe that the proposed adaptation re-ranking strategy effectively promotes lower-ranked target samples to higher positions, as shown by the solid lines in the figure. Furthermore, our method can further improve the retrieval performance by fine-tuning the already well-ranked correct retrievals, as shown by the dashed lines in the figure.

Model Parameter. Figure 4 illustrates the performance of CLIP based on ViT/L-14, ViT/g-14, and ViT/G-14. ViT/g-14 exhibits approximately five times higher computational complexity than ViT/L-14, while ViT/G-14 shows a sevenfold increase. Analyzing the performance increase at each expansion fold, we observe a slight decline in performance improvement from the second expansion compared to the first expansion (2% $\rightarrow$ 1.8%). Considering the challenge associated with ultra-large models, this evidence indicates that current visual foundation models have room for improvement before reaching their performance limits.

4. Conclusion

This paper explores a new avenue to improve the performance of foundation models with their inner knowledge, so-called "self-enhancement". The proposed A^3R framework performs zero-shot attribute augmentation to augment the downstream dataset before model learning while harnessing the internal relationship between the text and image representation space to conduct cross-modal re-ranking to retrieve the final expected images. This self-enhancement framework can be transferred to any cross-modal retrieval scenario with good domain generality.

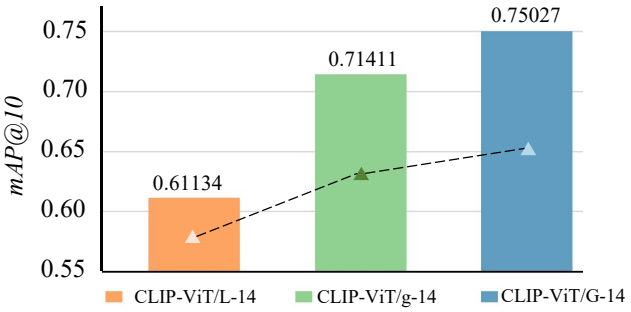

Figure 4. Comparison of the $mAP@10$ metric of backbones of different parameter sizes.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et. al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[2] Mayuri D. Joshi, Revati M. Deshmukh, Kalashree N. Hemke, and et. al. Image retrieval and re-ranking techniques-a survey. *SIPIJ*, 5(2), 2014. 3

[3] Rajiv Kapoor, Deepak Sharma, and Tarun Gulati. State of the art content based image retrieval techniques using deep learning: a survey. *Multimedia Tools and Applications*, 80(19):29561–29583, 2021. 1

[4] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. In *British Machine Vision Conference 2018*, page 142. BMVA Press, 2018. 3

[5] Danfeng Qin, Stephan Gammeter, Lukas Bossard, and et. al. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pages 777–784. IEEE Computer Society, 2011. 3

[6] Rodolfo Quispe, Cuiling Lan, Wenjun Zeng, and et. al. Attributenet: Attribute enhanced vehicle re-identification. *Neurocomputing*, 465:84–92, 2021. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, and et. al. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1

[8] Jun Sang, Zhongyuan Wu, Pei lin Guo, and et. al. An improved yolov2 for vehicle detection. *Sensors*, 18, 2018. 3

[9] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *WACV*, pages 981–990, 2023. 1

[10] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and et. al. Scaling vision transformers. In *CVPR*, pages 12104–12113, 2022. 1, 2, 3