This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2024.3388958

Towards Generated Image Provenance Analysis via Conceptual-Similar-guided-SLIP Retrieval

Xiaojie Xia, Liuan Wang, Jun Sun, *Member, IEEE*, Akira Nakagawa

***Abstract*—With the prevalence of state-of-the-art generative models, photorealistic synthetic images can now be easily generated. However, the generated images may replicate contents from the original training images, which can lead to potential legal issues. In this paper, we propose a novel method called *Conceptual-Similar-guided Self-supervised Language-Image Pre-training* (CS-SLIP) that leverages both image and text modalities for the generated image provenance. Besides the self-supervised learning branch and contrastive learning branch, a conceptual-similar branch is designed to guide the model to learn a better feature representation of image-text-pairs. We also adopt the re-ranking method to refine the initial matching candidates via the cross-modal bi-directional retrieval. Extensive qualitative and quantitative experiments are conducted, which demonstrate that the replication indeed exists in the generated images, and our proposed method can effectively retrieve the most similar images from the training corpus to achieve the goal of generated image provenance analysis.**

***Index Terms*—Generated image provenance, image retrieval, cross-modal**

I. Introduction

Recently, generative AI has become a prominent research area [1], [2]. Image generation models can synthesize realistic images that are indistinguishable from natural ones. However, since these models are trained on existing image datasets, they may learn to reproduce some content of the images from the training datasets [3], [4]. This may cause copyright issues. Engineering a methodology to ascertain the provenance of generated images is of critical importance.

Image retrieval techniques could potentially offer a viable mechanism towards the solution of generated image provenance. It is a fundamental task that aims to find the most similar images from a collection of images given a query image [5], which could plausibly be the sources that the generated image replicates. There are many applications, such as pedestrian tracking [6], remote sensing landmark recognition [7], and product searching [8], etc.

Current image retrieval methods mainly rely on analyzing the visual features of the query image and comparing them with the image database [9]. Benefiting from the rapid development of deep learning, numerous methods on image feature extraction have been proposed for image retrieval [10], [11]. Cross-modal retrieval studies have emerged in recent years, such as text-to-image retrieval [12], [13] and image-to-text retrieval [14], [15]. The basis idea of them is to establish the connection between image features and text semantics, and implicitly use the information from different modalities. The development of large visual models has provided strong support for multi-modal research, such as CLIP [16], using a large amount of image-text data to build the association between images and texts. However, existing research either retrieves images by image feature only or performs cross-modal retrieval that searches for samples in one modality given a query in another modality. Additionally, several studies have focused on achieving better cross-modal features. For example, [17] proposes a joint embedding in a manifold learning framework to enhance image retrieval performance.

We aim to develop a method that identifies potential replication and analyze the provenance of generated images via the image retrieval method. Our proposed approach deviates from the image-only methods by using image-text-pairs as input to improve the feature learning in image retrieval. Beyond conventional self-supervised learning and contrastive learning, we introduce a conceptual-similar branch to guide a better joint representation by integrating the textual information. The enriched image representations forms the basis for initial retrieval results through feature matching. Additionally, we design a re-ranking method that refines the results of initial coarse matching results by considering the retrieval scores of bi-directions, image-to-text and text-to-image. The final retrieval results may reveal that whether the replication occurs in the generated image and what source it imitates.

II. Related work

A. Image retrieval and cross-modal retrieval

The main steps of image retrieval are feature extraction and similarity measurement. Many researchers have proposed many methods on feature extraction [18], [19], [20], which is the process of extracting features from an image that can represent its content. Recent approaches adopt strong deep learning methods as architectural backbones for retrieval [21], [22], [23]. Similarity measurement is the process of comparing the features between a query image and the database images and ranking them according to their similarity [5].

Cross-modal retrieval involves retrieving relevant information from different modalities [24], and learn new representations for different modalities in a shared subspace [25], [26]. Recent research explores more state-of-the-art techniques for retrieval, such as large-scale pretraining [16], [27] and domain adaptation [28], [29].

B. Re-ranking methods

Re-ranking is a technique that leverages high-confidence retrieved samples to refine the initial retrieval results, which

Xiaojie Xia, Liuan Wang and Jun Sun are with Fujitsu R&D Center CO., LTD, Beijing, China. (e-mail: xiaxiaojie@fujitsu.com, liuan.wang@fujitsu.com, sunjun@fujitsu.com).

Akira Nakagawa is with Fujitsu Research, Kawasaki, Japan. (e-mail: akira@fujitsu.com).

is widely used for various image retrieval tasks, such as person re-identification [30] and vehicle re-identification [31]. Recently, there are more studies on cross-modal retrieval. For example, [32] introduced re-ranking for improved cross-modal retrieval and [33] proposed a method by re-ranking the results based on modality-driven clues.

C. Generated image provenance

Generated image provenance is still an under-explored area of research. A few existing studies [3], [34] have found evidence that generative image models can memorize individual images from the training data. Copy detection is a relevant task of finding unauthorized copies of copyright media [35]. Self-supervised methods [23], [36], [37] learn strong feature representations to enhance the copy detection. Particularly relevant literature on image copy detection task are DINO [38] and SSCD [39], which are built on previous self-supervised methods and optimized on the specific task.

III. Method

In this section, we present our Conceptual-Similar-guided Self-supervised Language-Image Pre-training (CS-SLIP) method as a fundamental technique towards the generated image provenance analysis. We also describe our initial image retrieval method that focuses on image features. Finally, we introduce our cross-modal re-ranking method that refines the retrieval results.

A. CS-SLIP: Conceptual-Similar-guided SLIP

We illustrate the CS-SLIP method for image-text-pair training in Fig. 1. Our method consists of three branches: self-supervised learning (SSL) branch, contrastive learning branch and conceptual-similar-guided branch.

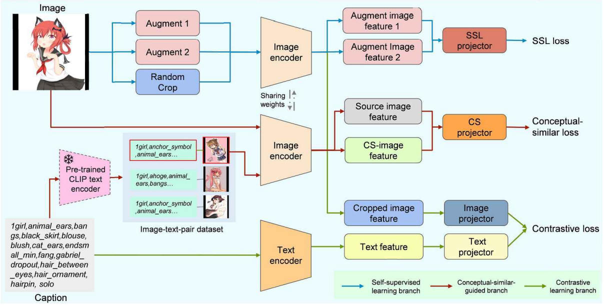
Image
Augment 1
Augment 2
Random Crop
Image encoder
Sharing weights
Image encoder
Augment image feature 1
Augment Image feature 2
SSL projector
SSL loss
Source image feature
CS-image feature
CS projector
Conceptual-similar loss
Pre-trained CLIP text encoder
1girl,anchor_symbol ,animal_ears…
1girl,ahoge,animal_ears,bangs…
1girl,anchor_symbol ,animal_ears…
Image-text-pair dataset
1girl,animal_ears,bangs,black_skirt,blouse, blush,cat_ears,endsmall_min,fang,gabriel_dropout,hair_between_eyes,hair_ornament, hairpin, solo
Caption
Text encoder
Cropped image feature
Image projector
Text feature
Text projector
Contrastive loss
Self-supervised learning branch
Conceptual-similar-guided branch
Contrastive learning branch

Fig. 1. Overview of proposed Conceptual-Similar-guided Self-supervised Language-Image Pre-training (CS-SLIP).

The pretraining method SLIP [27] combines language supervision and self-supervision together, which demonstrates that self-supervised learning with language supervision signal enables more effective visual representation learning. In our proposed method, the self-supervised learning branch and contrastive learning branch follow the same setting with SLIP. We design a conceptual-similar-guided branch containing a weight-sharing image encoder and integrate it with other branches to form a new framework.

The main idea of the conceptual-similar-guided branch is to search the conceptually similar image from the training dataset and enhance the feature learning. The typical self-supervised learning methods learn the image features from different augments of the same image or instance, which depend heavily on the visual similarity. However, many images with similar concepts, such as expressions, actions and backgrounds, may differ in visual appearance but have similar captions. In this branch, we leverage the text modality to learn more diverse and implicit image features by the guidance of the most conceptual similar images with the input images.

We firstly encode the text using the pre-trained CLIP text encoder, which is frozen during model training. The text vector is then matched with all the embeddings of the original texts in the training corpus to find the most conceptually similar one. The corresponding image associated with the most similar text is fed into the image encoder along with the original image as a contrastive image pair. It enables us to learn a better visual representation that accounts for the visual variances as well as the information from textual modality.

We compute three objectives on the different branches and sum them with appropriate weights to balance their contributions. For the self-supervised learning branch and contrastive learning branch, we followed the settings in [27]. For conceptual-similar-guided learning branch, we use the same objective with the self-supervised learning. We assign weighted value as the objective scales to balance the total loss.

It is worth mentioning that our proposed training process receives image-text pairs as input, however, the ground truth of images may not always be available. With the aid of state-of-the-art image-to-text models, such as BLIP [40], we can automatically generate the captions of the images. Then the images and corresponding captions can be fed into the proposed CS-SLIP for training as well to learn the cross-modal representations.

B. Initial image retrieval by concentrating image features

In this subsection, we design an initial image retrieval to obtain a small set of candidates for further searching. The image encoder takes the query image as input and extracts its CS-SLIP image feature. To enhance the image feature and increase the discriminability, we introduce the pretrained CLIP image feature to concatenate with the CS-SLIP image feature for a better representation.

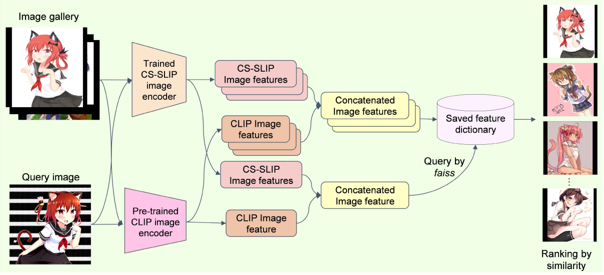

Fig. 2. Illustration of initial image retrieval by concentrating image features.

Fig. 2 illustrates the initial image retrieval process. We firstly extract the concatenated image features of all the images to build an image feature dictionary by faiss[1], an effective tool that enables efficient similarity search and clustering of dense vectors. Then the query image feature is obtained in a same

[1] https://github.com/facebookresearch/faiss

way. Feature matching can be performed by comparing the distances between query image feature and the image feature dictionary to return the top-ranked candidates, which are passed into the next step. For the generated image provenance analysis, we treat the generated image as a query and compare it with a gallery of authorized images.

C. Retrieval refining via cross-modal re-ranking

The initial image retrieval may not produce satisfactory results, because the image features are extracted from the visual modality, even though the model was trained in a multi-modal fashion. We design our re-ranking method to incorporate cross-modal information based on the assumption in [41]. That is, the image and text pairs can be mutually retrieved forwardly and reversely. Moreover, we fully consider the similarity scores from the initial retrieval rather than just ranking positions.

The re-ranking method is illustrated in Fig. 3. Suppose there are N candidates from the initial retrieval and their corresponding texts are extracted from the original dataset. The scores between the query image and candidates have been computed as S. If there are no available descriptions of the images, we can generate the corresponding captions using a pre-trained image-to-text model, such as BLIP [40] and DeepDanbooru[2], forming $N+1$ image-text pairs.

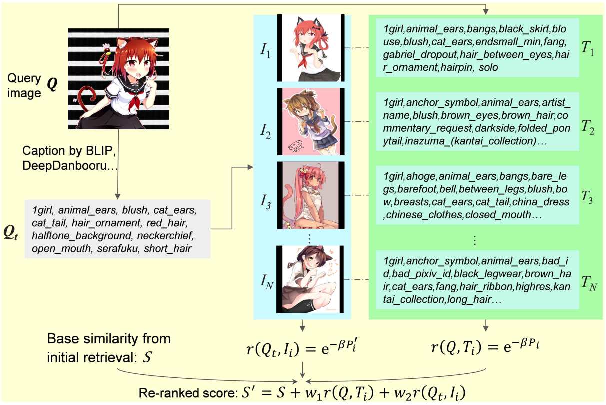
Query image Q
Caption by BLIP, DeepDanbooru…
Q_t
1girl, animal_ears, blush, cat_ears, cat_tail, hair_ornament, red_hair, halftone_background, neckerchief, open_mouth, serafuku, short_hair
I_1
I_2
I_3
I_N
1girl,animal_ears,bangs,black_skirt,blouse,blush,cat_ears,endsmall_min,fang, gabriel_dropout,hair_between_eyes,hair_ornament,hairpin, solo
T_1
1girl,anchor_symbol,animal_ears,artist_name,blush,brown_eyes,brown_hair,commentary_request,darkside,folded_ponytail,inazuma_(kantai_collection)…
T_2
1girl,ahoge,animal_ears,bangs,bare_legs,barefoot,bell,between_legs,blush,bow,breasts,cat_ears,cat_tail,china_dress,chinese_clothes,closed_mouth…
T_3
1girl,anchor_symbol,animal_ears,bad_id,bad_pixiv_id,black_legwear,brown_hair,cat_ears,fang,hair_ribbon,highres,kantai_collection,long_hair…
T_N
Base similarity from initial retrieval: S
$r(Q_t, I_i) = e^{-\beta P_i'}$
$r(Q, T_i) = e^{-\beta P_i}$
Re-ranked score: $S' = S + w_1 r(Q, T_i) + w_2 r(Q_t, I_i)$

Fig. 3. Retrieval refining via cross-modal re-ranking.

To explain our re-ranking method in detail, we use the image query Q as an example. Let Q_t be the caption of Q, and $(I_1 - T_1, I_2 - T_2, \ldots, I_N - T_N)$ be candidate image-text pairs. First, we compute the image-text scores between Q and $T_i \in (T_1, T_2, \ldots, T_N)$ and rank them to obtain the position list $(P_1, P_2, \ldots, P_N) \in (0, 1, \ldots, N)$. To assign higher values to top-ranked texts, we define the image-text refining score as

$$r(Q, T_i) = e^{-\beta P_i} \tag{1}$$

where β is the refining coefficient. Next, we rank Q_t with $I_i \in (I_1, I_2, \ldots, I_N)$ to get the position list $(P_1', P_2', \ldots, P_N') \in (0, 1, \ldots, N)$. Similarly, we define the text-image refining score as

$$r(Q_t, I_i) = e^{-\beta P_i'} \tag{2}$$

Then, we fuse the base similarity and the refining score to obtain the refined similarity as

$$S' = S + w_1 r(Q, T_i) + w_2 r(Q_t, I_i) \tag{3}$$

where S is the base similarity computed from concatenated features, w_1 and w_2 are the refining factors. This score incorporates both visual similarity and cross-modal information comprehensively. Finally, we re-rank the candidates according to this score and return the most relevant retrieval results.

IV. Experiments

This section reports extensive experiments that validate the effectiveness of our approach. We execute quantitative evaluations on image retrieval tasks and compare our method with state-of-the-art baselines. Furthermore, we perform some qualitative analysis on the generated image provenance.

A. Quantitative evaluation

Dataset: There is no ideally suitable datasets with image-text-pairs on image retrieval task. Alternatively, we can generate the captions of the images. To quantitatively evaluate the model performance, we made a hand-crafted-similar image dataset based on the anime dataset Danbooru2017[3], which provides over 300,000 images in normalized 512px×512px form together with the full tags of each image. Following the introduction of the 2021 Image Similarity Dataset and Challenge [35], we created a dataset for our experiment in a similar manner. We refer the readers to the appendix for more details on the dataset.

Evaluation: We used $recall@k$ for the evaluation, which measures the proportion of correct matches among the top k. Higher is better for this metric. We consider two kinds of evaluation conditions: comparison with different methods and different settings or representations in our models.

We evaluated the self-supervised model SimCLR [37] and the state-of-the-art methods on image retrieval and copy detection, i.e., DINO [38], and SSCD [39]. We followed the original setting and trained them on Danbooru2017 dataset. The evaluation was performed on the hand-crafted-similar dataset and the result is shown in Table I. Our method outperformed others significantly on the recall from $k = 1$ to $k = 100$. The results showed that the introduction of text information could be of great help to the image retrieval accuracy.

TABLE I
Evaluation result for image retrieval on the hand-crafted-similar dataset.

Methods	R@1	R@5	R@10	R@50	R@100
SimCLR[37]	66.80%	70.45%	72.50%	78.15%	80.95%
DINO[38]	78.00%	83.20%	84.90%	89.25%	91.20%
SSCD[39]	84.15%	91.10%	94.05%	95.35%	97.00%
Ours	90.10%	96.00%	97.35%	99.25%	100.00%

Fig. 4 presents several examples of retrieval results using transformed images as queries, comparing the performance of different methods. It is evident that our proposed method consistently demonstrates superior performance in most cases when juxtaposed against other approaches.

We also compared the performance of adopting different settings or representations from our model as an ablation study shown in Table II. The results show that using only *CLIP-pretrained* or *SLIP* features leads to poor performance. The

[2] https://github.com/KichangKim/DeepDanbooru

[3] https://www.gwern.net/Danbooru2017

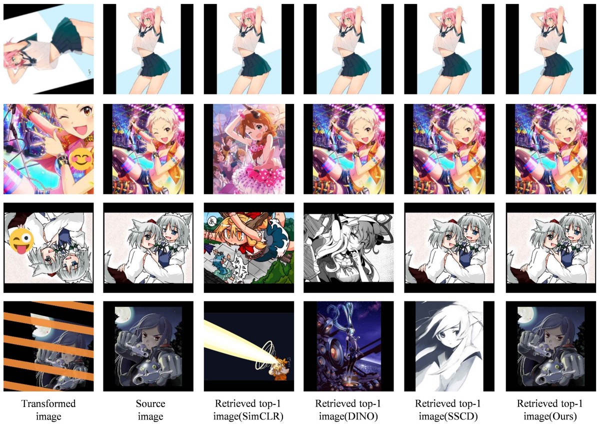
Transformed image
Source image
Retrieved top-1 image(SimCLR)
Retrieved top-1 image(DINO)
Retrieved top-1 image(SSCD)
Retrieved top-1 image(Ours)

Fig. 4. Retrieval results under different methods.

combination of these two features enables the image feature to capture both the image information learned by the training dataset and large-scale dataset. Thus, the final image feature has a better vision representation to achieve a better performance. The conceptual-similar-guided training exploits the text information effectively. The re-ranking method improves the retrieval ranking when a small retrieval k value is required.

TABLE II
ABLATION STUDY ON PROPOSED CONCEPTUAL-SIMILAR-GUIDED-SLIP METHOD VIA RE-RANKING.

Methods	*CS*	*SLIP*	*CLIP-pretrained*	*Re-ranking*	*R*@1	*R*@5	*R*@10	*R*@50	*R*@100
SLIP		✓			55.50%	66.30%	68.50%	76.30%	79.45%
CLIP			✓		63.35%	73.90%	75.90%	83.45%	85.25%
CS-SLIP	✓	✓			61.30%	70.75%	74.15%	81.50%	83.75%
SLIP-CLIP		✓	✓		88.51%	94.26%	96.01%	98.51%	99.41%
CS-SLIP-CLIP	✓	✓	✓		89.75%	95.70%	97.00%	99.05%	100.00%
CS-SLIP-CLIP-ReRank	✓	✓	✓	✓	90.10%	96.00%	97.35%	99.25%	100.00%

* *CS*: conceptual-similarity-guided training; *SLIP*: adopting the image feature extracted from the image encoder in our proposed method. *CLIP-pretrained*: the image feature is pretrained on a large-scale dataset; *SLIP* and *CLIP-pretrained* are eligible stand for that the two image features are concatenated. *Re-ranking*: using the re-ranking processing to refines the retrieval results based on the similarity scores.

B. Qualitative evaluation on generated image provenance

In this part, we present a qualitative experiment of the proposed method on generated image provenance analysis. Current synthesized images by the state-of-the-art generation model are already creative to some degree. In other words, they are not just simple copies of the original images but are more similar in style or object-level. These similar images are difficult to retrieve via traditional image matching methods. Our method combines semantic and vision features, and uses conceptual-similar contrastive pairs to guide training, which can be more suitable for retrieving generated images. As there is no ground truth available for the source of generated images, we provide some illustrative examples to show the utility of our method.

We conducted our experiment on stable diffusion and used diffusers[4] in PyTorch as the tool to realize all the functions. The weights[5] trained on the public LAION-5b [42] dataset was adopted as the initialization parameters during our stable diffusion model training. Since the LAION dataset is very large, the proposed retrieval may be difficult due to computational and storage limitations. For a relatively narrow retrieval, we fine-tuned the generative model on Danbooru2017 dataset with 20 epochs to fully learn the distribution. The model was saved for the text-to-image generation.

For the prompts to generate images by the trained stable diffusion model, we random selected some captions from Danbooru2017 dataset. We called the generated images in-caption generated images. We also synthesized images by random prompts and the analysis can be found in appendix. To our knowledge, the generated image may be more likely similar to the source image with the same caption, but not absolutely. There is no distinct evidence that the replication definitely exists. The ground truths of query and source image pairs are extremely hard to obtain. We just visualize some samples to prove that our proposed method has the ability to retrieve the similar images of the generated images.

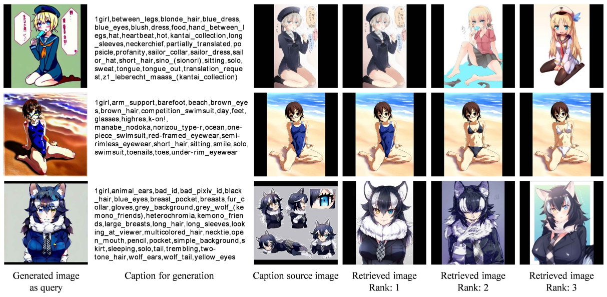
1girl,between_legs,blonde_hair,blue_dress,blue_eyes,blush,dress,food,hand_between_legs,hat,heartbeat,hot,kantai_collection,long_sleeves,neckerchief,partially_translated,popsicle,profanity,sailor_collar,sailor_dress,sailor_hat,short_hair,sino_(sionori),sitting,solo,sweat,tongue,tongue_out,translation_request,z1_leberecht_maass_(kantai_collection)
1girl,arm_support,barefoot,beach,brown_eyes,brown_hair,competition_swimsuit,day,feet,glasses,highres,k-on!,manabe_nodoka,norizou_type-r,ocean,one-piece_swimsuit,red-framed_eyewear,semi-rimless_eyewear,short_hair,sitting,smile,solo,swimsuit,toenails,toes,under-rim_eyewear
1girl,animal_ears,bad_id,bad_pixiv_id,black_hair,blue_eyes,breast_pocket,breasts,fur_collar,gloves,grey_background,grey_wolf_(kemono_friends),heterochromia,kemono_friends,large_breasts,long_hair,long_sleeves,looking_at_viewer,multicolored_hair,necktie,open_mouth,pencil,pocket,simple_background,skirt,sleeping,solo,tail,trembling,two-tone_hair,wolf_ears,wolf_tail,yellow_eyes
Generated image as query
Caption for generation
Caption source image
Retrieved image Rank: 1
Retrieved image Rank: 2
Retrieved image Rank: 3

Fig. 5. Obvious similar retrieval results found in the generated images.

Fig. 5 shows the obvious similar retrieval found of the generated images. The visualizations here were selected from images with a similarity score above 0.8. The generated images are listed in the first column as the query to be analyzed the provenance. The second and third column present the prompts from the Danbooru2017 dataset and the corresponding original images that match them. The rightmost three columns display the images with the highest similarity scores retrieved by our proposed method.

We observed that there were some distinct replications from the source dataset. The displayed generated samples exhibit similitude with the source retrieved images, both in visual appearance and semantic meaning. Not all generations were likely to match the reference image sharing the same generation caption. The reason may be that: 1) the trained image generation model is not perfect enough to describe all the tags from the caption; 2) the caption is not representative for the source image sometimes. Our method can successfully retrieve the most analogous source images by fully utilizing the vision modality and the cross-modal information.

V. CONCLUSION

In this work, we proposed a novel method that leverages both image and text modalities towards the generated image provenance analysis. Our method consists of self-supervised learning branch, contrastive learning branch and a designed conceptual-similar branch, incorporating image and textual features. We also proposed a re-ranking technique to refine the initial coarse retrieval results. We performed extensive qualitative and quantitative experiments to show that our approach can reliably achieve detection of duplicated content across generated images. As future work, we plan to explore more reasonable method on generated image provenance by fully integrating the principle of generative models.

[4] https://github.com/huggingface/diffusers/
[5] https://huggingface.co/runwayml/stable-diffusion-v1-5

References

[1] M. Suzuki and Y. Matsuo, “A survey of multimodal deep generative models,” *Advanced Robotics*, vol. 36, no. 5-6, pp. 261–278, 2022.

[2] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, “Text-to-image diffusion model in generative ai: A survey,” *arXiv preprint arXiv:2303.07909*, 2023.

[3] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, “Diffusion art or digital forgery? investigating data replication in diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6048–6058.

[4] H. Hu and J. Pang, “Membership inference of diffusion models,” *arXiv preprint arXiv:2301.09956*, 2023.

[5] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, “Deep learning for instance retrieval: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, “Person re-identification meets image search,” *arXiv preprint arXiv:1502.02171*, 2015.

[7] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. de With, “Aggregated deep local features for remote sensing image retrieval,” *Remote Sensing*, vol. 11, no. 5, p. 493, 2019.

[8] M. Hendriksen, M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper, and M. de Rijke, “Extending clip for category-to-image retrieval in e-commerce,” in *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Springer, 2022, pp. 289–303.

[9] J. M. Patel and N. C. Gamit, “A review on feature extraction techniques in content based image retrieval,” in *2016 international conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE, 2016, pp. 2259–2263.

[10] P. Chhabra, N. K. Garg, and M. Kumar, “Content-based image retrieval system using orb and sift features,” *Neural Computing and Applications*, vol. 32, pp. 2725–2733, 2020.

[11] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, “Exploiting deep features for remote sensing image retrieval: A systematic investigation,” *IEEE Transactions on Big Data*, vol. 6, no. 3, pp. 507–521, 2019.

[12] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, “Camp: Cross-modal adaptive message passing for text-image retrieval,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5764–5773.

[13] T. Yu, J. Liu, Z. Jin, Y. Yang, H. Fei, and P. Li, “Multi-scale multi-modal dictionary bert for effective text-image retrieval in multimedia advertising,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 4655–4660.

[14] J. Luo, Y. Shen, X. Ao, Z. Zhao, and M. Yang, “Cross-modal image-text retrieval with multitask learning,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2309–2312.

[15] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, “Cross-modal graph matching network for image-text retrieval,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1–23, 2022.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] H. Dong, Z. Wang, Q. Qiu, and G. Sapiro, “Using text to teach image retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1643–1652.

[18] J. Yue, Z. Li, L. Liu, and Z. Fu, “Content-based image retrieval using color and texture fused features,” *Mathematical and Computer Modelling*, vol. 54, no. 3-4, pp. 1121–1127, 2011.

[19] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, “Cnn vs. sift for image retrieval: Alternative or complementary?” in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 407–411.

[20] F. Radenović, G. Tolias, and O. Chum, “Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 3–20.

[21] M. Berman, H. Jégou, A. Vedaldi, I. Kokkinos, and M. Douze, “Multigrain: a unified image embedding for classes and instances,” *arXiv preprint arXiv:1902.05509*, 2019.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[24] L. Zhen, P. Hu, X. Wang, and D. Peng, “Deep supervised cross-modal retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.

[25] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, “Cross-modal retrieval with cnn visual features: A new baseline,” *IEEE transactions on cybernetics*, vol. 47, no. 2, pp. 449–460, 2016.

[26] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, “Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 4, pp. 1–23, 2021.

[27] N. Mu, A. Kirillov, D. Wagner, and S. Xie, “Slip: Self-supervision meets language-image pre-training,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*. Springer, 2022, pp. 529–544.

[28] Y. Liu, Q. Chen, and S. Albanie, “Adaptive cross-modal prototypes for cross-domain visual-language retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 954–14 964.

[29] J. Munro, M. Wray, D. Larlus, G. Csurka, and D. Damen, “Domain adaptation in multi-view embedding for cross-modal video retrieval,” *arXiv preprint arXiv:2110.12812*, 2021.

[30] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[31] N. Jiang, Y. Xu, Z. Zhou, and W. Wu, “Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking,” in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 858–862.

[32] D. Mandal and S. Biswas, “Query specific re-ranking for improved cross-modal retrieval,” *Pattern Recognition Letters*, vol. 98, pp. 110–116, 2017.

[33] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, “Database-adaptive re-ranking for enhancing cross-modal image retrieval,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3816–3825.

[34] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, “Extracting training data from diffusion models,” *arXiv preprint arXiv:2301.13188*, 2023.

[35] M. Douze, G. Tolias, E. Pizzi, Z. Papakipos, L. Chanussot, F. Radenovic, T. Jenicek, M. Maximov, L. Leal-Taixé, I. Elezi *et al.*, “The 2021 image similarity dataset and challenge,” *arXiv preprint arXiv:2106.09672*, 2021.

[36] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[38] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[39] E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, “A self-supervised descriptor for image copy detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 532–14 542.

[40] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, 2022.

[41] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and re-ranking,” in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 12–20.

[42] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *arXiv preprint arXiv:2210.08402*, 2022.