2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Neural-Based Cross-modal Search and Retrieval of Artwork

Yan Gong
Department of Computer Science
Loughborough University
Loughborough, UK
y.gong2@lboro.ac.uk

Georgina Cosma
Department of Computer Science
Loughborough University
Loughborough, UK
g.cosma@lboro.ac.uk

Axel Finke
Department of Mathematical Sciences
Loughborough University
Loughborough, UK
a.finke@lboro.ac.uk

***Abstract*—Creating an intelligent search and retrieval system for artwork images, particularly paintings, is crucial for documenting cultural heritage, fostering wider public engagement, and advancing artistic analysis and interpretation. Visual-Semantic Embedding (VSE) networks are deep learning models used for information retrieval, which learn joint representations of textual and visual data, enabling 1) cross-modal search and retrieval tasks, such as image-to-text and text-to-image retrieval; and 2) relation-focused retrieval to capture entity relationships and provide more contextually relevant search results. Although VSE networks have played a significant role in cross-modal information retrieval, their application to painting datasets, such as ArtUK, remains unexplored. This paper introduces BoonArt, a VSE-based cross-modal search engine that allows users to search for images using textual queries, and to obtain textual descriptions along with the corresponding images when using image queries. The performance of BoonArt was evaluated using the ArtUK dataset. Experimental evaluations revealed that BoonArt achieved 97 % Recall@10 for image-to-text retrieval, and 97.4 % Recall@10 for text-to-image Retrieval. By bridging the gap between textual and visual modalities, BoonArt provides a much-improved search performance compared to traditional search engines, such as the one provided by the ArtUK website. BoonArt can be utilised to work with other artwork datasets.**

***Index Terms*—cross-modal, information retrieval, retrieval of artwork, visual-semantic embedding, search engine**

I. Introduction

Creating a retrieval system for artwork images, particularly paintings, is of paramount importance in documenting cultural heritage, facilitating wider public engagement, and fostering advancements in art analysis and interpretation [1], [2]. Existing research for retrieving painting images primarily focuses on using neural networks to classify objects within painting images and match their categories with user queries for retrieval purposes [3]–[5]. However, these studies have limitations in fully comprehending the complex semantic information expressed in user queries. While research specific to information retrieval can uncover the underlying meaning of user queries, text-based information retrieval [6]–[8] lacks the ability to understand the visual content of images, relying solely on textual information such as image tags. On the other hand, cross-modal information retrieval is capable of extracting and comprehending high-level visual semantics in conjunction with textual information, enabling users to obtain more relevant and accurate results [9], which is appropriate for applying to retrieval of painting images. Visual-Semantic Embedding (VSE) networks represent state-of-the-art techniques in cross-modal information retrieval. These networks aim to embed image-description pairs into a shared latent space, enabling the computation of similarity scores for image-to-text and text-to-image retrieval tasks [10], [11]. VSE networks have demonstrated their effectiveness with real-world images in widely used benchmark datasets such as Flickr30K [12], MS-COCO [13], and RefCOCOg [14]. Despite their successful application in these datasets, their potential for implementing retrieval of painting images, especially in the context of being integrated into a cross-modal search engine, remains unexplored, representing an untapped area of research.

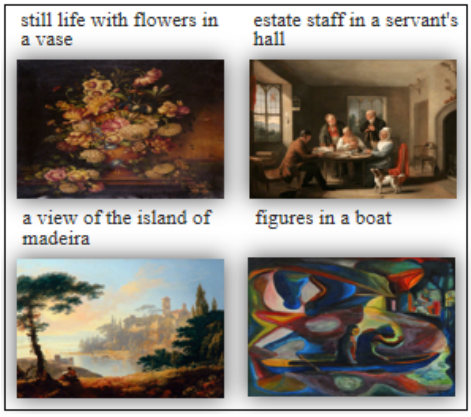

Fig. 1. Examples of painting images with corresponding textual description.

In recent years, there have been significant advancements in the field of VSE. Faghri et al. [15] introduced an architecture that embeds image region features extracted by the faster R-CNN [16] and their descriptions into a shared latent space using a fully connected neural network and a Gated Recurrent Units (GRU) network [17], respectively. Lee et al. [18] augmented VSE networks by employing the attention mechanism to enhance the alignment of image regions with their corresponding words. Li et al. [19] proposed the visual semantic reasoning network, which leverages the Graph Convolution Network (GCN) [20] to extract high-level visual semantics. Chen et al. [21] unveiled a network employing a

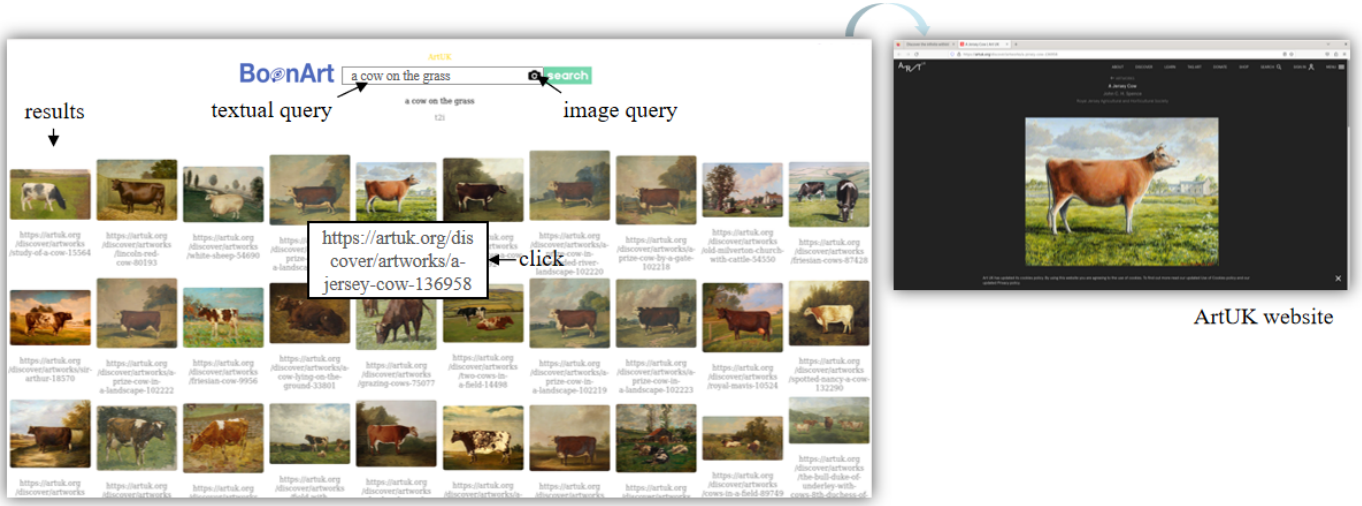

Fig. 2. The front-end of BoonArt enables users to input textual queries for text-to-image retrieval or upload image queries for image-to-text retrieval. The search results are displayed, allowing users to explore painting images, with additional details accessible through clicking links to the ArtUK website.

generalised pooling operator to formulate an optimal strategy for integrating image and description representations. Chen et al. [22] introduced a pre-trained network built upon the transformer model [23] that has been pre-trained on four large datasets [13], [24]–[26]. Recently, Radford et al. [27] proposed a Contrastive Language-Image Pre-training network (CLIP), which leverages 400 million image-description pairs to enable the efficient learning of visual concepts via natural language supervision. To overcome the limitation of the pre-trained Vision Transformers (ViTs) for relation-focused cross-modal information retrieval, Gong et al. [28] proposed a ViT-Relation-focus network (VITR), which employs a local encoder to reason about relations within image regions. This paper primarily focuses on implementing a VSE-based cross-modal search engine designed for retrieving painting images. Figure 1 showcases a variety of painting images, encompassing different styles including realistic, impressionistic, abstract, and still life paintings. The proposed search engine, named BoonArt, incorporates a VSE network, VITR, enhancing its capability in image-to-text and text-to-image retrieval tasks within the painting images, with a particular emphasis on relation-focused cross-modal information retrieval. Specifically, the contributions are as follows.

- BoonArt search engine has the capability to perform image-to-text and text-to-image retrieval for painting images; that BoonArt allows users to search for painting images using textual queries and retrieve textual descriptions along with their corresponding images using image queries; and that BoonArt benefits from a state-of-the-art VSE network, VITR, which can extract and understand high-level visual semantics to improve retrieval performance and enhance the user experience.
- BoonArt's performance is evaluated through experiments using the ArtUK dataset. The results demonstrate that BoonArt outperformes the ArtUK search system (from the ArtUK website https://artuk.org) for text-to-image tasks. In particular, BoonArt can use image queries for retrieval. In contrast, the ArtUK search system lacks this capability.

II. Methodology

BoonArt excels at retrieving painting images from the ArtUK dataset [29]. Users can enable text-to-image requests to search for relevant images based on their textual queries. Additionally, users can perform image-to-text requests to search for descriptions and their corresponding images using image queries. The BoonArt engine is composed of three main components: the front-end, back-end, and a database (images, textual descriptions, and representations, i.e., embeddings). Each component functions as follows.

A. Front-end

As depicted in Figure 2, in the front-end, users are presented with two options for querying: they can either input a textual query in the provided text box to enable the text-to-image request, or they can upload an image query using the designated upload button to enable the image-to-text request. After entering their query, users can initiate the search by clicking the search button. The retrieval results are then displayed on the interface, allowing users to view and explore the painting images. To provide additional information about the retrieved paintings, users can click on the links associated with the images, which will open a new page on the source website of each painting. Users can input their queries with high-level semantics, such as a focus on relations, and BoonArt will provide accurate search results based on the semantics of their queries.

B. Back-end

The core of BoonArt's back-end is a VSE network, and the process of back-end is shown as Figure 3. For the text-to-

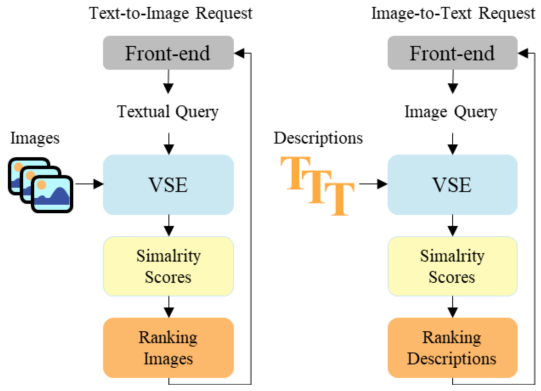

Fig. 3. The back-end of BoonArt utilises a VSE network for text-to-image and image-to-text requests, comparing queries with the dataset's images or descriptions and generating ranked results displayed by the front-end.

image request, the VSE network compares the textual query with all the images in the dataset. It generates similarity scores for each image and ranks them accordingly. The front-end then displays the ranked images. Similarly, for the image-to-text request, the VSE network compares the image query with all the descriptions in the dataset. It calculates similarity scores for each description and ranks them accordingly. The front-end displays the images corresponding to the ranked descriptions.

BoonArt employs the state-of-the-art VSE network VITR [28]. VITR consists of a text encoder for encoding descriptions as global and local representations, a ViT encoder for encoding images as global representations, a local encoder for encoding images as local representations for relational reasoning, and a fusion module that fuses the representations from the encoders to output the similarity score between the image and the description. VITR takes the text and ViT encoders from CLIP to obtain pre-trained knowledge from an extensive dataset of 400 million image-description pairs. Additionally, VITR has been fine-tuned on the RefCOCOg dataset, which enhances the network's ability to learn the reasoning relations of image regions to improve performance in relation-focused cross-modal information retrieval tasks.

C. Database

The database is built upon the ArtUK dataset [29], which consists of 6 783 pairs of painting images and their descriptions, sourced from the ArtUK website (https://artuk.org). To optimise retrieval time, the representations for images and descriptions needed by the VSE network have been pre-encoded and stored. The following files were created and are available in the database: 1) ArtUKimGloRp.npy (13.0 MB) stores the global representations of images; 2) ArtUKimLocRp.npy (1.7 GB) stores the local representations of images; 3) ArtUKdeGloRp.npy (13.0 MB) stores the global representations of descriptions; and 4) ArtUKdeLocRp.npy (999.8 MB) file stores the local representations of descriptions. By directly accessing the saved representation values from these files, the back-end eliminates the need for encoding images and descriptions during the retrieval process, resulting in faster retrieval.

III. Experiments

A. Implementation Details

BoonArt can function with a minimum requirement of a single NVIDIA RTX 3080 graphics card. The integrated VSE network of BoonArt is VITR, which leverages the encoder of 'ViT-L/14' from CLIP. To maintain generalizability in real-world image scenarios, BoonArt's VITR employs zero-shot learning on the ArtUK dataset.

B. A Comparison of BoonArt and the ArtUK Search System for Text-to-Image Retrieval

TABLE I
Performance of BoonArt with the ArtUKA search system on a set of textual queries.

Query	BoonArt	ArtUK
1. a cow in the room	✓	✗
2. a red cow in the room	✓	✗
3. airplane flying with another airplane	✓	✗
4. a man sitting with his dog	✓	✗
5. train in the town	✓	✗
6. a man riding a horse with another horse	✓	✗
7. a brown dog with a white dog	✓	✗
8. white car on the street	✓	✗
9. impression painting of river in a town	✓	✗
10. impression painting of sheep in the rain	✓	✗
11. impression painting of trees and horse	✓	✗
12. impression painting of ruin in the landscape	✓	✗
13. abstract painting of bird on the table	✓	✗
14. abstract painting of a boat behinds another	✓	✗
15. abstract painting of a person with red hair	✓	✗
16. abstract painting of a town in the dark night	✓	✗
17. still life of flower	✓	✗
18. still life of yellow flower	✓	✗
19. still life of cat on the chair	✓	✗
20. still life of fruits on the blue tablecloth	✓	✗

To compare the performance between BoonArt and the ArtUK search system (from the ArtUK website) for text-to-image retrieval, a set of 20 textual queries was generated, as shown in Table I. These queries encompass high-level semantics, including relations. Specifically, there are four queries focused on impression paintings, four queries on abstract paintings, four queries on still life paintings, and an additional eight queries without any specific limitations. The top-ranked retrieved results by BoonArt and the ArtUK search system for these queries are presented in Figure 4. For instance, when queried with 'a red cow in the room', BoonArt retrieves the relevant painting image. However, the ArtUK search system only identifies an image with the color red, which is irrelevant to the query. Similarly, when queried with 'a man sitting with his dog', BoonArt accurately retrieves a relevant painting image. In contrast, the ArtUK search system only finds an image with a man and his dog, overlooking the critical relation of 'sitting with' in the image. Figure 4 highlights the performance of BoonArt in capturing high-level semantics from painting images.

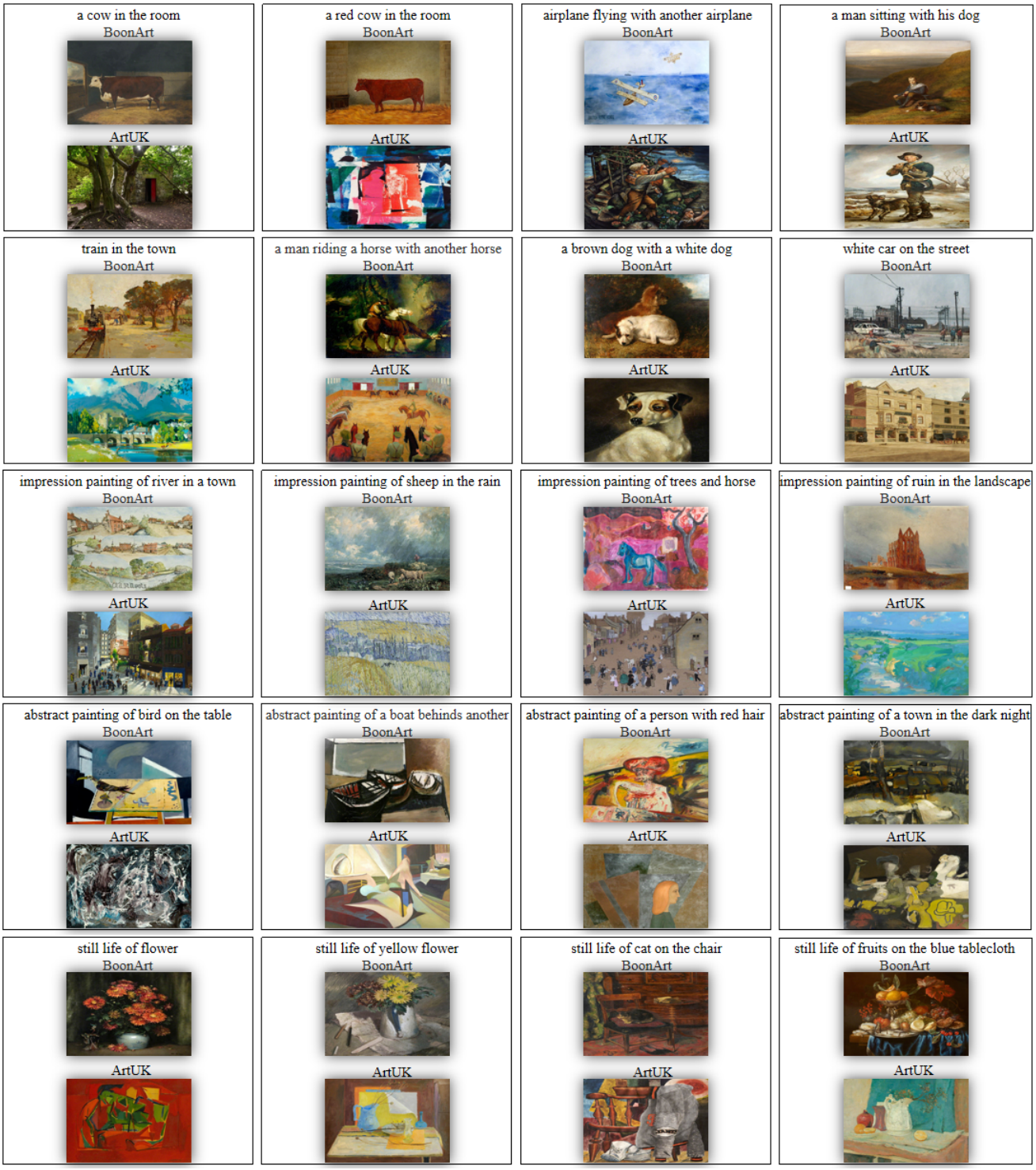

Fig. 4. Comparison between BoonArt and the ArtUK search system for text-to-image retrieval. The figure shows the top-ranked retrieved images for the queries.

Fig. 5. Retrieval results of BoonArt for image-to-text retrieval. The figure shows the top-ranked retrieved descriptions and the images corresponding to the queries.

C. BoonArt's Capability for Image-to-Text Retrieval

The ArtUK search system does not support the use of image queries, therefore, only the results of BoonArt for image-to-text retrieval are presented. BoonArt utilised eight real-world image queries to retrieve the top-ranked descriptions and their corresponding images from the ArtUK dataset, as shown in Figure 5. For example, when a query image depicting two trains in a station was provided, BoonArt successfully retrieved the relevant description 'two trains in a station' and displayed the corresponding image. Similarly, when a query image of a cat on a chair was given, BoonArt retrieved the relevant description 'cat on a chair' along with its corresponding image. These results demonstrate the effectiveness of BoonArt in performing image-to-text retrieval and in facilitating the exploration using visual queries.

D. Quantitative Results of Cross-modal Information Retrieval with BoonArt

To conduct experiments with quantitative results, the ArtUK dataset was randomly partitioned into three subsets: 5 783 for training, 500 for validation, and an additional 500 for testing. The evaluation metric used is Recall@K for cross-modal information retrieval experiments, which measures the percentage of relevant items in the top K retrieved results [30], [31]. The objective is to retrieve at least one relevant item from a given list, and the average Recall is calculated across all evaluated queries to assess retrieval performance.

Table II presents the performance of BoonArt for image-to-text and text-to-image retrieval tasks on the ArtUK test set. Additionally, to highlight the best performance, the results of BoonArt fine-tuned on the training set are also included in Table II.

According to Table II, BoonArt (zero-shot) achieved an average Recall@10 value of 93.0 % for image-to-text retrieval and 94.2 % for text-to-image retrieval in the ArtUK test set. On the other hand, BoonArt (fine-tuned) achieved an average Recall@10 value of 97.0 % for image-to-text retrieval and 97.4 % for text-to-image retrieval. These findings highlight BoonArt's capability in achieving successful cross-modal information retrieval in the ArtUK dataset.

TABLE II
RESULTS OF CROSS-MODAL INFORMATION RETRIEVAL NETWORKS ON THE ARTUK TEST SET. TABLE SHOWS AVERAGE RECALL@K (%) VALUES.

Engine	Method	Image-to-Text [%]			Text-to-Image [%]		
		R@1	R@5	R@10	R@1	R@5	R@10
BoonArt	zero-shot	68.8	88.0	93.0	71.2	90.2	94.2
BoonArt	fine-tuned	77.4	93.4	97.0	80.8	94.6	97.4

E. Evaluation of Retrieval Time

The average retrieval time for each query by BoonArt was experimentally measured. For the image-to-text retrieval task, which involved processing 6 783 textual descriptions, the average retrieval time for one query was 0.18 seconds. Similarly, for the text-to-image retrieval task, which included 6 783 images, the average retrieval time for one query was 0.40 seconds.

IV. CONCLUSION

This paper presents BoonArt, a cross-modal search engine specifically designed for retrieving painting images. To enhance the user experience, BoonArt excels at image-to-text and text-to-image retrieval by extracting high-level visual semantics. It integrates the state-of-the-art VSE network VITR, which focuses on relation-focused cross-modal information

retrieval. Extensive experiments were conducted to evaluate BoonArt's performance, demonstrating its ability to outperform the ArtUK search system in text-to-image retrieval tasks. Furthermore, BoonArt surpasses the limitations of the ArtUK search system by enabling image queries, allowing users to retrieve textual descriptions and their corresponding images. BoonArt enhances the retrieval performance of painting images by bridging the gap between textual and visual modalities, resulting in an improved user experience. Currently, the database includes the ArtUK dataset, but it can be extended to work with new datasets. In future work, the database will be expanded to incorporate various datasets of artworks, and the engine will be evaluated using these datasets.

REFERENCES

[1] G. Castellano, E. Lella, and G. Vessio, "Visual link retrieval and knowledge discovery in painting datasets," *Multimedia Tools and Applications*, vol. 80, pp. 6599–6616, 2021.

[2] K. Li, J. Wang, B. Batjargal, and A. Maeda, "Intuitively searching for the rare colors from digital artwork collections by text description: a case demonstration of japanese ukiyo-e print retrieval," *Future Internet*, vol. 14, no. 7, p. 212, 2022.

[3] E. Crowley and A. Zisserman, "The state of the art: object retrieval in paintings using discriminative regions," in *Proceedings of the British Machine Vision Conference*, 2014, p. 8.

[4] E. J. Crowley and A. Zisserman, "The art of detection," in *Proceedings of the European Conference on Computer Vision Workshops*. Springer, 2016, pp. 721–737.

[5] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, and A. D. Bimbo, "Deep artwork detection and retrieval for automatic context-aware audio guides," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3s, pp. 1–21, 2017.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of International Conference on Computer Vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[7] L. Zhang and Y. Rui, "Image search—from thousands to billions in 20 years," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1s, pp. 1–20, 2013.

[8] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *International Conference on Computer Vision*. IEEE, 2011, pp. 2049–2055.

[9] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: a review," *Computer Science Review*, vol. 39, p. 100336, 2021.

[10] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3536–3545.

[11] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 336–11 344.

[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[14] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 11–20.

[15] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *Proceedings of the British Machine Vision Conference*, 2018, p. 12.

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[18] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 201–216.

[19] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of International Conference on Computer Vision*, 2019, pp. 4654–4662.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of International Conference on Learning Representations*, 2017.

[21] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 789–15 798.

[22] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: universal image-text representation learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 104–120.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[25] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2018, pp. 2556–2565.

[26] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: describing images using 1 million captioned photographs," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1143–1151, 2011.

[27] 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 *Proceedings of International Conference on Machine Learning*, 2021, pp. 8748–8763.

[28] Y. Gong and G. Cosma, "VITR: augmenting vision transformers with relation-focused learning for cross-modal information retrieval," *arXiv preprint arXiv:2302.06350*, 2023.

[29] E. C. Elliot J. Crowley and A. Zisserman, "The ArtUK Paintings Dataset," https://www.robots.ox.ac.uk/ vgg/data/paintings/, 2021.

[30] Y. Gong, G. Cosma, and H. Fang, "On the limitations of visual-semantic embedding networks for image-to-text information retrieval," *Journal of Imaging*, vol. 7, no. 8, p. 125, 2021.

[31] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 138–146.