Semantic Analysis of Cultural Heritage Data: Aligning Paintings and Descriptions in Art-Historic Collections

Nitisha Jain*[0000−0002−7429−7949], Christian Bartz*†[0000−0002−1800−0442], Tobias Bredow, Emanuel Metzenthin, Jona Otholt, and Ralf Krestel[0000−0002−5036−8589]

Hasso Plattner Institute
University of Potsdam, 14482 Potsdam, Germany
<firstname.lastname@hpi.de>

Abstract. Art-historic documents often contain multimodal data in terms of images of artworks and metadata, descriptions, or interpretations thereof. Most research efforts have focused either on image analysis or text analysis independently since the associations between the two modes are usually lost during digitization. In this work, we focus on the task of alignment of images and textual descriptions in art-historic digital collections. To this end, we reproduce an existing approach that learns alignments in a semi-supervised fashion. We identify several challenges while automatically aligning images and texts, specifically for the cultural heritage domain, which limit the scalability of previous works. To improve the performance of alignment, we introduce various enhancements to extend the existing approach that show promising results.

Keywords: cultural heritage · natural language processing · computer vision.

1 Introduction

Digitized collections of cultural artifacts provide an interesting opportunity for deeper analysis and understanding of our heritage and culture. Several cultural institutions around the world have made continued efforts to digitize their cultural resources and make them widely available for access and analysis by scholars, as well as interested audiences. Massive volumes of art-historic archives have been scanned and digitized by museums and galleries as part of the OpenGlam[1] initiative. These digital collections comprise art catalogues, magazines and art

* both authors contributed equally
† corresponding author
[1] openglam.org

books that contain multimodal data, namely texts and images. For instance, images of paintings are often accompanied by their titles and textual descriptions in these materials.

In order to derive useful insights from these resources, many attempts are being made to leverage machine learning techniques. Several works have focused on deriving useful information and systematic representations from text alone [6, 8, 17, 22, 32, 36]. On the other hand, previous work has paid attention to image analysis techniques for the depicted artworks [9, 20, 35, 37]. However, both research directions typically overlook the rich information embedded in joint analysis of multiple modalities. An important example is the structure and associations of the paintings with their descriptions, that are present in the original resources but usually lost during the digitization process. In this work, we bring attention to the task of aligning the images with their corresponding texts in the context of digitized art collections. As detailed previously [1], one of the most important benefits of image and text alignment is to facilitate multimodal search and retrieval of artworks from online digital archives, by leveraging the textual, as well as image features in conjunction. Additionally, enriching the digital resources with multimodal meta-data can improve results for individual text analysis and image analysis tasks. For example, image classification is difficult for paintings, especially those depicting portraits, and classification models can greatly benefit from textual cues for such cases.

The alignment of images and texts in digitized collections is a non-trivial task due to various challenges that are unique to the cultural heritage domain, including diverse formats, lack of training data, etc. Moreover, different datasets require customizations to the alignment techniques. For instance, varying ratios and types of images and texts are found in different art-historic datasets. On the one hand, art books have longer texts in the context of a few paintings or art styles, artists and so on. On the other hand, auction or exhibition catalogues contain mostly images of paintings or artifacts and shorter texts, mainly title, artist, date, *etc*. Due to this variability, scaling of any existing technique across different datasets becomes difficult.

Within the scope of an ongoing project on multimodal analysis of cultural heritage datasets, we collaborate with the Wildenstein Plattner Institute (WPI)[2] that was founded to promote scholarly research on cultural heritage collections. We have been provided access to a large digitized collection compiled and maintained by WPI. This collection consists of scanned pages of art-historic documents ranging from sales catalogues to art magazines from the 19th century up to today. We refer to our dataset as WPI-Art in the rest of the paper.

Although there are a few existing approaches on image and text alignment (see related work, Section 2), we found that none of the current techniques performed well on the WPI-Art dataset due to several novel challenges (see Section 3). Therefore, we explored several enhancements over existing approaches for overcoming these challenges in the context of the WPI-Art dataset. More specifically, due to the lack of annotated training datasets we follow an approach,

[2] https://wpi.art/

introduced by Cornia *et al.* [5], for identifying the correct alignment. We first extract text and image content from the digitized dataset and derive semantic features from both separately, and then perform semantic alignment to match the artwork images with their correct texts. Building on the work of Cornia *et al.*, we improve the results with several new additions for semantic analysis such as augmenting texts with thesauri tokens, using bag-of-words (BoW) and N-grams representations for text, as well as using the technique of neural style transfer for the images (Section 4). In our experimental results (see Section 5), we show that our proposed enhancements do not only improve results on the WPI-Art dataset, but also on the commonly used SemArt dataset [14].

The specific contributions of this work are as follows: (1) Define distinct challenges for the task of aligning artwork images with their text descriptions in art-historic collections. (2) Reproduce and evaluate an existing semi-supervised approach for image and text alignment on our WPI-Art dataset. (3) Enhance the approach with several customizations for text analysis and image analysis techniques, improving the performance on the WPI-Art and SemArt datasets.

2 Related Work

The multimodal nature of the problem domain we are dealing with is rooted in two modalities. On the one hand, computer vision methods are used for image analysis. On the other hand, methods from the field of natural language processing are leveraged for attaining semantic understanding of the texts. In this section we present related work in the field of image and text alignment.

The overall task of aligning images and texts in a multimodal retrieval setting has been under active research in the past few years [5, 11, 12, 14, 21, 26]. All of the proposed methods attempt to seek a function that embeds the features of images and texts in a common semantic space [26]. Thereafter, retrieval algorithms search for images and texts that are close to each other in this embedding space. While several methods were introduced to perform alignment of photos and their corresponding texts [21, 26, 29], quite a number of methods have been introduced to achieve the same for images of artworks and their corresponding descriptions [5, 11, 12, 13, 14].

Many of these proposed methods only work with supervised learning techniques that require full annotation of a large training dataset [11, 12, 14, 26]. Garcia *et al.* [14] introduced such a supervised retrieval model. They compare multiple methods for embedding images and texts into a common semantic subspace together with different techniques of matching the embedded features of images and texts in the same space. The authors further enhance their proposed model with contextual features, such as the category of depicted scene, the artist's name, or the timeframe of the painting [12, 13]. Furthermore, they propose to use knowledge graphs to enrich the embeddings of their input images.

To train supervised models for the retrieval of artworks and their corresponding descriptions, Garcia *et al.* [14] introduce the SemArt dataset consisting of 21 384 paintings paired with textual descriptions. Furthermore, Stefanini *et*

al. [34] introduced the Artpedia dataset that consists of approximately 3000 images of paintings paired with textual descriptions. Although the Artpedia dataset is smaller than the SemArt dataset, it contains much more interesting content. Each sentence that is paired with an image is further classified as a sentence that describes the visual content of the image, or a sentence that provides art-historical information about the image. The SemArt dataset lacks such a categorization which makes it difficult to apply this dataset for multi-modal retrieval, since several images only contain texts without any description of the visual content. Although the Artpedia dataset seems to be the natural choice for the development of a system, its small size kept us from using it in our experiments.

Apart from fully supervised models, Cornia *et al*. [5] proposed an unsupervised approach for aligning paintings and their corresponding descriptions. They first train an image to text matching model on the standard COCO [28] image captioning dataset in a fully supervised setting. Thereafter, they use the Maximum Mean Discrepancy (MMD), as well as images and texts from the SemArt dataset (without any annotations) to adapt the supervised model in a manner that the embeddings of paintings and texts match, although the model has only been trained on the COCO dataset.

The approach presented in this paper is based on the work of Cornia *et al*. Owing to the lack of availability of large annotated training data, their apporach fits quite well for our use case. We also include the SemArt dataset in our evaluations to enable fair comparison of our reproduction of their model, as well as to demonstrate the improvements from our proposed enhancements, as discussed in Section 4.

3 Challenges for Image and Text Alignment in the Cultural Heritage Domain

Due to the variety and diversity of the available art-historic material, restoring lost associations between artwork images and texts is a non-trivial task. In this section we discuss and illustrate the most prominent challenges that we encountered while working with the WPI-Art dataset.

Loss of Formatting. Due to a variety of formats within our digital collections, we found several examples where there is no one-to-one mapping between the images and texts. Figure 1 shows a common scenario where there are several different text segments, but only one of them is aligned to the image. In this case, it is important to correctly identify which text is associated with the image and take the information about missing associations into account while learning the alignment. To further compound the problem, the irregular text spacing in such old documents makes it difficult to clearly separate the different texts after extraction through OCR via text segmentation techniques. The identification of correct and relevant text segments is also important for art books having longer discussions not only about the artworks, but also about the artists, art styles, *etc*.

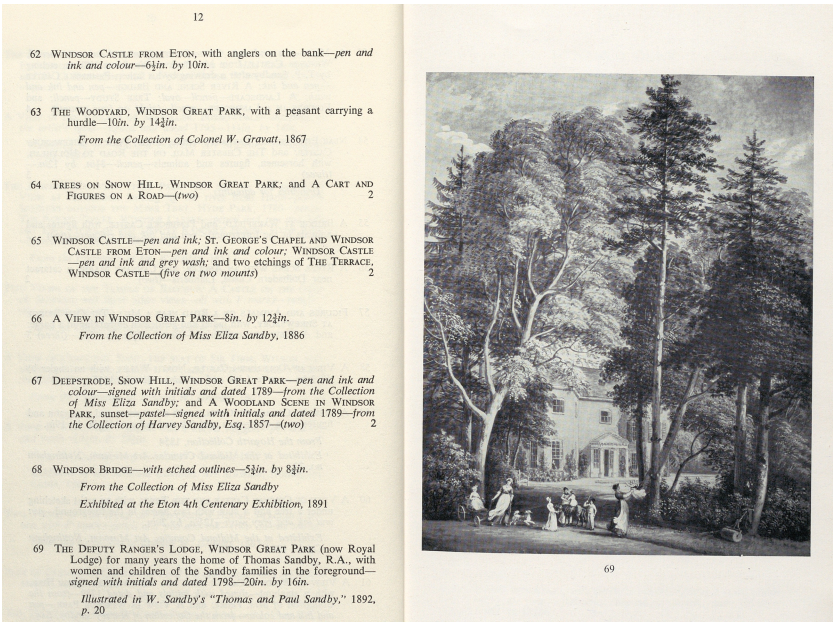

Fig. 1: Mapping difficulties

A calm with a yacht at anchor in the centre with numerous figures firing a salute a row boat with fifteen figures on the right in front and numerous small craft and boats beyond ...

A sloop with numerous figures in the centre lowering sail in collision with another vessel ; a barque beyond on the right ; a man-of-war shortening sail in a gale on the left.

Fig. 2: Similar depictions in images

Moreover, useful formatting cues such as the numbers matching the images with the text segments, are also lost after text extraction from page scans. Without such cues, in many cases it is challenging to perform the alignment, even for human annotators.

Granularity. The complexity of the task also varies depending on the level of granularity - whether the alignment is being performed for a single page of the resource, the whole catalogue/book, or for the whole collection. Different levels face different types of challenges. For instance, a single page of an art catalogue can contain multiple images depicting the same scenery. Generally, in a catalogue, images of paintings showing similar depictions are grouped together. Post-digitization, it becomes difficult to identify the correct associations since the features obtained from image analysis and text analysis are quite similar for such cases. Minute differences in depictions, colors, tones, and style need to be recognized for distinguishing the different mappings and therefore the feature ex-

Fig. 3: Semantic ambiguity in Images

traction needs to be quite elaborate in such a scenario. On the other hand, when performing alignment for images and texts from an entire catalogue or collection, all artworks are placed together in a single corpus. This may bring together artwork images from different pages with very similar depictions and descriptions as shown in Figure 2, making the task of alignment equally challenging.

Lack of Semantic Pointers. For the WPI-Art dataset, we encountered several cases where the textual descriptions, including the painting titles, fail to contain useful information to enable the alignment with the correct images. As discussed in [23], the identification of titles of artworks in textual descriptions is a non-trivial task itself for which existing Named Entity Recognition (NER) tools show sub-optimal performance. Even if these titles can be identified, there are several instances where the titles and even the description text for an artwork do not sufficiently describe the depictions of the painting, as would be identified by an image captioning model. One prominent example is that of portrait paintings where the texts usually elaborate on the person being portrayed, rather than describe the artwork depiction itself. Figure 3 illustrates a few portraits of different women in similar white attire. Even when some indicative features are present, it is challenging to perform alignment due to the overlapping features in the images. In the absence of fine-grained semantic features for guidance, the task of identifying the correct alignment becomes several magnitudes harder.

Training Data Unavailability. Machine learning algorithms require sizeable training data to learn models for performing specific tasks. There are several annotated datasets for the task of aligning image datasets comprising of photographs with texts [28]. However, the same is not true for our use case where the matching has to be performed for images of artworks and paintings. Although we leverage the few available datasets in this work [14], the existing models and techniques

do not scale well to our dataset due to differences in image features, length of text descriptions, as well as formatting styles. Image analysis is harder for artworks due to the ambiguity of depiction and interpretation for several artworks, especially in the modern art genre. Due to this ambiguity, annotations need to be more precise as well as larger in size for the effective training of machine learning models. The unavailability of such annotated training datasets prevents the applicability of existing techniques for art-historic datasets and necessitates the need for further efforts in this direction.

4 Alignment Approach

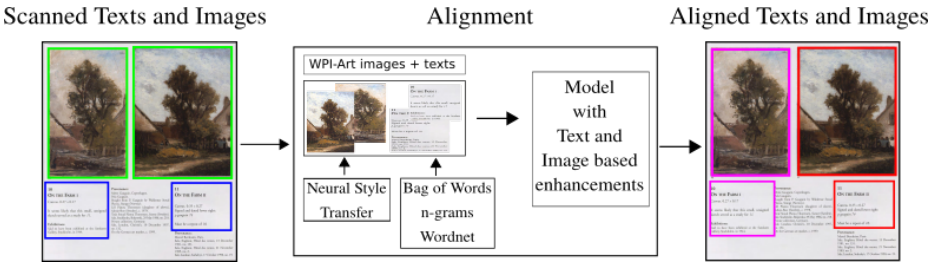

Fig. 4: Overview of our approach

In this section, we discuss our approach for the alignment of artwork images to associated texts and describe the different enhancement techniques in detail. In our previous work [1] we envisioned a framework for aligning images and texts. The overall framework, including the enhancements made in this work is shown in Figure 4.

To apply any existing approaches on the WPI-Art dataset which consists of scanned catalogue pages, as a first step, we perform the automatic extraction of images and texts. Image extraction is performed by using the image localization algorithm from the OpenCV library [3]. We employ the Tesseract OCR engine [33] for the extraction of textual content from the scans and store it as plain text files.

Due to the lack of large annotated datasets in the art domain, we adopted the approach of Cornia *et al.*, which leverages the knowledge from the supervised training of an image-text alignment model trained on non-art documents. After training the supervised alignment model, an unsupervised method based on Maximum Mean Discrepancy is used to align the feature distributions of the unsupervised dataset with the feature distribution of the supervised training set. In this case, the COCO dataset, which consists of photos of common objects, was used as the supervised training dataset and the SemArt dataset was considered as the unsupervised training dataset. In the approach presented by Cornia *et al.* two types of neural networks are used to encode the data. Images are encoded

using a pretrained ResNet-152 [18] architecture, while texts are encoded using a Gated Recurrent Unit (GRU) [4]. Furthermore, words are embedded using pretrained word embeddings (either Glove [31] or FastText [2]). The output of the encoders is passed through one linear layer each, resulting in an embedding vector for each text description or image respectively. The embedding layers need to be trained such that the vectors of a matching text-image pair end up close in the semantic space whilst embeddings of other negative pairs are separated apart. In order to achieve this, two batches of training samples are fed through the networks at each step: a supervised batch (from the COCO dataset) and an unsupervised batch (from SemArt). A triplet hinge loss is used as the loss function. The loss gets reduced over the batch samples either by using a mean or sum operation. The unsupervised batch is used to guide the resulting embeddings towards the target domain of artworks. For this, the Maximum Mean Discrepancy (MMD) loss is used for aligning the distributions of the unsupervised text and image samples embedded with the current state of the model.

In the rest of the section, we discuss the techniques that we have explored for improving the performance of the model by Cornia *et al.*, specifically on the WPI-Art dataset.

4.1 Word Encodings

Since the textual descriptions in our dataset are often composed of short sentences (this is especially true for painting titles), a complex word representation model based on word sequences is not a good way to encode the text. Previous work has also shown that a simple BoW encoding for texts outperformed more complex methods [14]. Therefore, we implemented the BoW encoding for the texts, replacing the GRU encoding in Cornia *et al*. We further extended this technique and used bigrams encoding for the texts, so as to capture the semantics at phrase level as well.

4.2 Vocabulary Augmentation

In order to maximize the information from the textual excerpts and thus increase the chance of aligning the correct images, we enhanced the textual data using a thesaurus. We use WordNet [30] to generate synonyms from the terms in the text descriptions and added these to the training data. The primary motivation for this technique being, two semantically similar terms, that are not trivially matched by the model could be anchored around a shared synonym and hence nudge the model towards matching this pair better.

4.3 Neural Style Transfer

Paintings and photographs follow very different distributions in semantic space [5]. Due to this, conventional deep learning models that are trained on photographic datasets do not work well on art-related images. Convolutional neural networks

(CNNs) for image recognition are typically biased towards texture in their input data [16]. Artworks, however, include varying types of brush strokes that may have no relation to the actual objects depicted in the image. In the absence of large annotated datasets, we leverage the technique of Neural Style Transfer (NST) to bridge the gap between the datasets of photographs and paintings.

First introduced by Gatys *et al.* [15], neural style transfer describes the task of rendering a given input image (content image) in the "style" of another input image (style image), while retaining the content of the content image and applying the style of the style image. In the past years, several other neural style transfer mechanisms have been proposed [24] and successfully applied. In our work, we explored using the same technique as Geirhos *et al.* for image classification, using a stylized version of the ImageNet dataset [7], for improving the retrieval results on the WPI-Art dataset.

5 Results and Discussion

In this paper, our focus is on the application of multimodal retrieval for evaluating the alignment of images and texts extracted from art-historic datasets. We evaluate the approach introduced by Cornia *et al.* [5] and show how our proposed enhancements can help to boost the performance of the baseline approach. In this section, we first introduce our experimental setup. Subsequently, we describe the datasets that we use for training and evaluation. Lastly, we show the results of our experiments on the SemArt dataset and WPI-Art dataset.

5.1 Experimental Setup

In our experiments, we followed the network design introduced by Cornia *et al.* [5]. For the extraction of image features we used a ResNet-152 [18] model pre-trained on the ImageNet dataset [7]. For the encoding of text, we used a GRU [4], while we generated embeddings of our vocabulary using pre-trained word vectors. As long as not stated otherwise, we trained our models for 30 epochs, used the Adam optimizer [25] with a learning rate of $2 \cdot 10^{-4}$, and a batch size of 128. For evaluation, we calculate and report Recall@5 at the sample size 100, of the respective test datasets. We release our code and models for further experimentation[3].

5.2 Datasets

We used different datasets for training and evaluation of our models. For the training of the supervised part of our model, we followed Cornia *et al.* and used the Microsoft COCO [28] dataset, totalling 83 000 samples for training. For training the unsupervised part of the model, we used the train split of the SemArt dataset [14], which accounts for 19 244 samples for training. For

[3] https://github.com/HPI-DeepLearning/semantic_analysis_of_cultural_heritage_data

evaluation, we used the validation split of the SemArt dataset and a subset of the WPI-Art dataset. Thw WPI-Art dataset consists of 93 images of paintings and their corresponding descriptions that we extracted from scans of auction catalogues provided by the Wildenstein Plattner Institute.

5.3 Results

We performed a range of different experiments. First, we attempted to reproduce the results of Cornia *et al.* and created a baseline that was used to ilustrate the effect of our enhancements (see row 1 in Table 1 for our baseline results). Subsequently, we performed experiments using our proposed enhancements introduced in Section 4.

Table 1: Results of our experiments on the SemArt and WPI-Art evaluation datasets with a sample size of 100 or 93, respectively and a batch size of 32 for the BoW bigrams approach. We report Recall@5 as our evaluation metric. **Bold font** indicates the overall best result.

Method	Text Retrieval		Image Retrieval	
	SemArt	WPI-Art	SemArt	WPI-Art
Baseline	0.08	0.20	0.11	0.20
unigram BoW	0.21	0.25	0.26	0.29
unigram BoW + Style Transfer	0.20	0.17	0.27	0.23
unigram BoW + WordNet	0.24	0.20	0.28	0.27
bigram BoW	**0.30**	**0.34**	**0.34**	**0.34**

Enhancing the Text Encoding. Our first enhancement to the baseline model was to replace the GRU text encoder with a simpler BoW encoding, as used in [14]. Using this simple approach of text encoding we were able to improve the performance of our model by a large margin. We further adjusted the text encoding with bag-of-words to consider bigrams instead of single words, which lead to improvements in the results, as can be seen in rows 2 and 5 of Table 1. The results show that a simple text pre-processing technique is better suited for the alignment task since the bag-of-words approach can skip learning the word sequences in the descriptions.

An additional technique that we applied on the text encoding part, was to enrich the text data with a thesaurus. As already described in Section 4.2, we enriched our vocabulary using synonyms extracted from the WordNet [10] hierarchy. Table 1 (see row 4) shows that enriching the vocabulary with WordNet helped to improve our results on the SemArt dataset although not on the WPI-Art dataset. We conclude that enriching descriptions with synonyms is a

promising technique for enabling the alignment of images to texts. Using vocabulary enrichment increases the probability for the text embedding model to find concepts that can be embedded close to the concepts obtained by the image feature extractor.

Enhancing the Image Encoding. Besides improving the text analysis, we also introduced improvements on the image encoding side. Here, we used neural style transfer for the creation of a stylized version of ImageNet, as described in Section 4.3. Unfortunately, using a pre-trained image encoder on stylized images did not improve results, as can be observed in Table 1 (see row 2). One reason for this could be the stylization of only the training data for the feature extractor. The stylization of the images used for training the supervised part of the alignment model could likely improve results, however this is open for future work.

Table 2: Recall@5 on the SemArt (different sample sizes) and WPI-Art evaluation datasets for bigrams BoW approaches using different batch sizes (in parenthenses).

Method	Text Retrieval					Image Retrieval				
	100	300	500	1000	WPI-Art	100	300	500	1000	WPI-Art
bigrams BoW (128)	0.27	0.14	0.10	0.05	0.32	0.28	0.14	0.09	0.05	0.34
bigrams BoW (64)	0.27	**0.19**	**0.15**	**0.10**	**0.34**	0.33	0.19	**0.16**	0.11	**0.40**
bigrams BoW (32)	**0.30**	**0.19**	**0.15**	0.09	**0.34**	**0.34**	**0.24**	0.12	**0.12**	0.34

Batch Size Experiments. Until this point, our improvements already showed substantial improvements over the baseline. Next, we experimented with different batch sizes. Previous work has shown that large batch sizes can harm generalization [19]. A batch size of 128 being quite large, we performed additional experiments with lower batch sizes than 128. From the results in Table 2, it can be seen that decreasing the batch size indeed helped the model to generalize and provided us with our best results on the SemArt and WPI-Art datasets.

Reproducing the Existing Approach Since the code or models from Cornia *et al.* [5] was unavailable, we attempted to reproduce their results for this work. Table 3 shows our replication results (denoted as "baseline"), the results reported by Cornia *et al.* as well as the results of our best performing model. Despite following the design of their system, as outlined in Section 5.1, as closely as possible, our replication of their setup did not show the same results, with our baseline having a considerably lower Recall@5. The reason for this is not clear, it is highly probable that small technical nuances that could not be directly inferred from the paper were missing in our setup. However, with the help of our enhancement

techniques, we were able to improve the results to be on par with the results reported by Cornia *et al*. This shows that our proposed enhancements can prove valuable in improving the performance of existing approaches substantially.

Table 3: Recall@5 on the SemArt evaluation datasets for Cornia *et al*., our re-implementation as baseline, and our best approach. We evaluate on different sample sizes (100, 300, 500, 1000) for retrieval.

Method	Text Retrieval				Image Retrieval			
	100	300	500	1000	100	300	500	1000
Cornia *et al*. [5]	**0.34**	**0.19**	0.12	**0.09**	0.32	0.17	**0.12**	0.07
Baseline	0.08	0.04	0.03	0.02	0.11	0.06	0.04	0.02
bigrams BoW	0.30	**0.19**	**0.15**	**0.09**	**0.34**	**0.24**	**0.12**	**0.12**

6 Conclusion

In this paper, we presented our approach on aligning artwork images to their corresponding descriptions in the digitized art-historic corpus of the Wildenstein Plattner Institute. We showed that analyzing digitized art-historic corpora poses many challenges. One of the greatest challenges is the availability of annotated training data for the training of deep models for multimodal retrieval. To this end, we proposed to leverage a previously introduced, semi-supervised approach that we further extended with various enhancements. The results of our experiments show that though the results of the previous approach could not be fully reproduced, our chosen approach and enhancements are viable and promising on the dataset under consideration.

Future Work. There is ample scope for several improvements based on this work that can lead to performance gains. In this work, we have evaluated the gains of our enhancements individually, in future work we want to further investigate combinations of our proposed enhancements. We also think that further experiments with neural style transfer could prove helpful. In this paper, we have considered stylizing the training data for the feature extractor but have not stylized the COCO dataset that was used as supervised training set, this could lead to potential problems in the distribution alignment phase and could be explored further. On the text analysis side, a possible improvement would be to additionally incorporate a dataset that contains less contextual information, but more descriptions of artworks, such as the Artpedia dataset [34]. Furthermore, a larger dataset extracted from the corpus of the WPI could be of help, since the training dataset does not have to contain matched pairs of artworks and descriptions. Finally, it might be possible to improve our experimental results

by following a different unsupervised domain adaptation approach, as used by Lee *et al.* [27].

Acknowledgement We thank the Wildenstein Plattner Institute for providing access to their art-historic archives.

References

1. Bartz, C., Jain, N., Krestel, R.: Automatic Matching of Paintings and Descriptions in Art-Historic Archives using Multimodal Analysis. In: Proceedings of the International Workshop on Artificial Intelligence for Historical Image Enrichment and Access (AI4HI). pp. 23–28 (2020)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Bradski, G., Kaehler, A.D., Opencv, D.: Dobb's journal of software tools. OpenCV Libr **25**, 120 (2000)
4. Cho, K., Merrienboer, B.v., Gülcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP). pp. 1724–1734 (2014)
5. Cornia, M., Stefanini, M., Baraldi, L., Corsini, M., Cucchiara, R.: Explaining digital humanities by aligning images and textual descriptions. Pattern Recognition Letters **129**, 166–172 (Jan 2020)
6. De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In: Proceedings of the Extended Semantic Web Conference (ESWC). pp. 733–747 (2012)
7. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR. pp. 248–255 (2009)
8. Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., ter Weele, W., Wielemaker, J.: The Rijksmuseum Collection as Linked Data. Semantic Web **9**(2), 221–230 (2018)
9. Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., Mazzone, M.: The shape of art history in the eyes of the machine. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2018)
10. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
11. Garcia, N., Renoust, B., Nakashima, Y.: Context-aware embeddings for automatic art analysis. In: Proceedings of the International Conference on Multimedia Retrieval (ICMR). pp. 25–33. ICMR '19, Ottawa ON, Canada (Jun 2019)
12. Garcia, N., Renoust, B., Nakashima, Y.: Understanding art through multi-modal retrieval in paintings. arXiv:1904.10615 [cs] (Apr 2019)
13. Garcia, N., Renoust, B., Nakashima, Y.: ContextNet: representation and exploration for painting classification and retrieval in context. International Journal of Multimedia Information Retrieval **9**(1), 17–30 (Mar 2020)
14. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the ECCV Workshops (Workshop on Computer Vision for Art Analysis). pp. 676–691 (2018)

15. Gatys, L.A., Ecker, A.S., Bethge, M.: A Neural Algorithm of Artistic Style. arXiv:1508.06576 [cs, q-bio] (2015)
16. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: Proceedings of the International Conference on Learning Representations (Sep 2018)
17. Harris, M., Levene, M., Zhang, D., Levene, D.: Finding Parallel Passages in Cultural Heritage Archives. Journal on Computing and Cultural Heritage **11**(3), 1–24 (2018)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
19. Hoffer, E., Hubara, I., Soudry, D.: Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 1731–1741 (2017)
20. Huang, X., Zhong, S.h., Xiao, Z.: Fine-Art Painting Classification via Two-Channel Deep Residual Network. In: Advances in Multimedia Information Processing (PCM). pp. 79–88 (2018)
21. Huang, Y., Wang, L.: ACMM: Aligned cross-modal memory for few-shot image and sentence matching. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 5774–5783 (2019)
22. Hyvönen, E., Rantala, H.: Knowledge-based Relation Discovery in Cultural Heritage Knowledge Graphs. In: Proceedings of the Digital Humanities in the Nordic Countries Conference (DHN). pp. 230–239 (2019)
23. Jain, N., Krestel, R.: Who is Mona L.? Identifying Mentions of Artworks in Historical Archives. In: Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL). pp. 115–122 (2019)
24. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., Song, M.: Neural style transfer: A review. Transactions on Visualization and Computer Graphics **26**(11), 3365–3385 (2019)
25. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of the International Conference on Learning Represenations (ICLR). San Diego (2015)
26. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. arXiv:1411.2539 [cs] (2014)
27. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced Wasserstein discrepancy for unsupervised domain adaptation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10285–10295 (2019)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
29. Liu, Y., Guo, Y., Liu, L., Bakker, E.M., Lew, M.S.: CycleMatch: A cycle-consistent embedding network for image-text matching. Pattern Recognition **93**, 365–379 (Sep 2019)
30. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
32. Segers, R., Van Erp, M., Van Der Meij, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., Jacobs, G.: Hacking History via Event

Extraction. In: Proceedings of the International Conference on Knowledge Capture (K-CAP). pp. 161–162 (2011)
33. Smith, R.: An overview of the Tesseract OCR engine. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). pp. 629–633 (2007)
34. Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R.: Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Image Analysis and Processing (ICIAP). pp. 729–740 (2019)
35. Thomas, C., Kovashka, A.: Artistic object recognition by unsupervised style adaptation. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 460–476 (2019)
36. Van Hooland, S., Verborgh, R.: Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish your Metadata. Facet Publishing (2014)
37. Yang, S., Oh, B.M., Merchant, D., Howe, B., West, J.: Classifying Digitized Art Type and Time Period. In: Proceedings of the Workshop on Data Science for Digital Art History (DSDAH) (2018)