

Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation

Zechen Bai* Yuta Nakashima¹ Noa Garcia¹

¹Osaka University

zechenbai@outlook.com n-yuta@ids.osaka-u.ac.jp noagarcia@ids.osaka-u.ac.jp

Abstract

Have you ever looked at a painting and wondered what is the story behind it? This work presents a framework to bring art closer to people by generating comprehensive descriptions of fine-art paintings. Generating informative descriptions for artworks, however, is extremely challenging, as it requires to 1) describe multiple aspects of the image such as its style, content, or composition, and 2) provide background and contextual knowledge about the artist, their influences, or the historical period. To address these challenges, we introduce a multi-topic and knowledgeable art description framework, which modules the generated sentences according to three artistic topics and, additionally, enhances each description with external knowledge. The framework is validated through an exhaustive analysis, both quantitative and qualitative, as well as a comparative human evaluation, demonstrating outstanding results in terms of both topic diversity and information veracity.

1. Introduction

For the general public, art tends to be considered as a mysterious and remote discipline that requires a lot of study to be fully appreciated. In the last few years, many efforts have been made to apply artificial intelligence technologies to the domain of art to make it more accessible [16]. Thanks to the large-scale digitisation of artworks from collections all over the world [22, 66, 49, 61], computer vision techniques have been widely adopted to address different art-related problems [30, 45, 11, 20, 44, 34, 8, 59, 18, 69].

Most of the existing work in the field is focused on the automatic analysis of paintings, addressing problems such as attribute prediction [45, 59], content analysis [11, 21] or style identification [29, 54, 30]. However, there is still an absence of research that conveys in-depth and comprehensive information of artworks for the general public. In other words, most previous work only allows people to under-

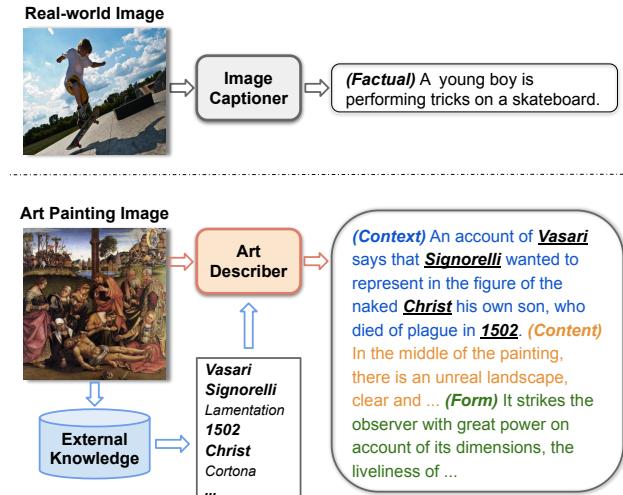


Figure 1. **Art description generation.** Comparison between standard image captioning (top) and art image description generation (bottom). Standard image captioning usually generates a single factual natural sentence to describe the content of a real-world image, while our art image description contains multiple sentences to explain an artwork from different artistic perspectives.

stand a unique aspect of an artwork by providing a single tag, usually associated with either its style or its content.

However, a real understanding of art is much more complex than being able to successfully categorize each piece into a set of pre-defined tags. The intricate relationships between the **content**, **form**, and **context** of each artwork are not at hand in a simple categorization process. In this work, we take a step forward on the field of art understanding and accessibility by proposing to automatically generate rich descriptions of paintings from multiple artistic perspectives. We propose a multi-topic and knowledgeable art description generation framework that produces detailed explanations about different aspects of paintings. Such a service would facilitate a deeper interaction between the general public and artworks, as well as potentially ease the work of art curators by automatically generating comments of paintings.

*Work done during a remote internship at Osaka University.

Art description generation may be seen as a problem similar to image captioning [64, 12, 1], which aims to naturally and factually describe the content of the scene in the image. Compared to conventional image captioning, however, generating descriptions for artworks faces two extra challenges. First, a comprehensive explanation of an artwork requires not only the factual description of its content, but also background knowledge, such as details about its author, the context of the creation process, and so on. This information is rarely contained in the artwork image itself. Second, according to art historians [2], an informative description of an artwork should address three main topics: **content**, **form**, and **context**. While in standard image captioning only the **content** is considered, the distinction among different artistic topics requires to handle a more complex language modeling scheme.

To address these challenges, we introduce a multi-topic and knowledgeable art description generation framework that: 1) introduces external knowledge into the description generation process, and 2) proposes a multi-topic language model to describe the different aspects of the painting. The main idea is illustrated in Fig. 1, where an artwork image is used to both generate a description and retrieve related knowledge from an external source. Our framework follows three steps. Firstly, by training a language model, we generate a masked sentence with fillable slots for the concepts that require external information to be known, such as artist, date of creation, location, *etc.* Moreover, the language model incorporates information about different artistic topics so that the masked sentences are generated according to each topic. Secondly, a knowledge retrieval module is employed to retrieve external information related to the painting from open access databases (*e.g.* Wikipedia¹). Finally, we design a knowledge-filling module that extracts candidate words from the retrieved knowledge and selects the appropriate concepts for each slot.

In our exhaustive experimental section, including quantitative comparisons, qualitative analysis, and human evaluation, we show that our framework generates satisfactory art descriptions more accurately and informatively than others. Overall, our main contributions are:

- We propose the first framework for art description generation that creates multi-topic long descriptions of fine-art paintings. So far, art description generation has been tackled as an image captioning task by only generating short factual sentences about artworks.
- We design a multi-topic language modeling module to generate multi-topic descriptions. Additionally, we annotate an art description dataset with sentence-topic labels based on art historians protocols [2], which we share publicly² to inspire future work not only in art

description but also in general art understanding.

- We leverage a knowledge retriever and train a knowledge filling module as a fill-in-the-blank task to incorporate art information relevant to each painting. This method can be easily applied to other domains.

2. Related Work

Artwork Analysis Computer vision techniques have been widely adopted to address art-related problems [27, 45, 67, 57]. A fundamental task in the field is to extract representative features that can capture the insights of the style [30, 20, 69, 26] or the content of a painting [11, 10, 53, 21] and use them for the automatic analysis of artworks, in tasks such as classification [42, 5, 60, 17], style identification [29, 54, 30, 69], object recognition [11, 9, 21], or image retrieval [5, 11, 9]. Although art categorization is challenging due to the inherent diversity and abstraction in art, it only studies a single aspect of the artwork. However, paintings are complex images full of symbolism. A single label cannot totally represent the elaborated relationship between the depicted elements, the painter’s motivations, and the historical context of the production. For a full comprehension of paintings, we propose to generate coherent language representations in the form of artistic descriptions. Until now, only a few studies [56, 18, 19] have applied multimodal vision and language techniques to the domain of art. While in [18] a system to find paintings given textual descriptions is proposed, other methods [56, 19] predicted answers to questions about artworks. However, generating comprehensive descriptions for fine-art paintings is still rarely studied.

Image Captioning Encoder-decoder image captioning models for natural images are data-driven methods using deep neural networks [64, 70, 1, 48, 73, 36, 74, 41, 65]. The classic scheme [64] combines a convolutional neural network (CNN) as image encoder and a recurrent neural network (RNN) as caption decoder. Several variations have emerged, such as adding attention [70] or using detected objects instead of plain pixels [1]. Although these models obtain good results for natural images, they do not transfer well to cultural images [55]. To generate descriptions for artworks, previous work introduced an ontology [72] and a hierarchical model [71], leveraging low-level features, *e.g.*, image texture and meta-data of cultural images, which heavily rely on feature engineering. Moreover, they could only generate a single and factual sentence about the image **content**. In contrast, we generate multi-topic descriptions, relying on external sources to improve the information quality. External knowledge has been used in image captioning [68, 73, 36, 74, 40, 3], mostly by relying on available tags [40] or texts [3] associated to the target image. Differently, our external knowledge is retrieved by only using the image.

¹<https://www.wikipedia.org>

²<https://github.com/noagarcia/explain-paintings>

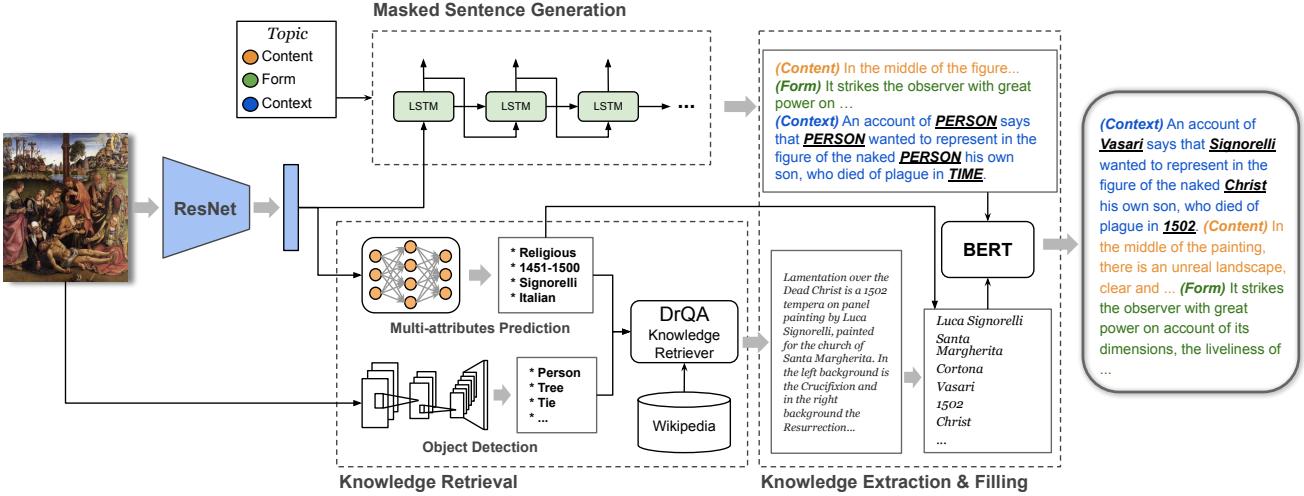


Figure 2. **Proposed framework.** It consists of three parts: masked sentence generation, knowledge retrieval, and knowledge extraction and filling.

3. Method

Our framework contains three main parts: 1) masked sentence generation, 2) knowledge retrieval, and 3) knowledge extraction and filling. As shown in Fig. 2, we first extract D -dimensional visual features from L spatial locations from the image [70] using a pre-trained ResNet [24],³ $V = \{v_1, v_2, \dots, v_L\}, v_i \in \mathbb{R}^D$. Then, in the masked sentence generation part, we input V into the topic decoder to generate multi-topic masked sentences that describe the painting from multiple aspects. These masked sentences have blanked concepts to be filled with knowledge at a later stage. In the knowledge retrieval part, we take the average-pooling vector $\bar{v} = \sum_i v_i / L$ as the global visual feature for multi-attribute prediction. We also employ an object detector to detect visual concepts. The predicted attributes and the detected objects are used to retrieve relevant knowledge from an external source with DrQA [7]. Finally, in the knowledge extraction and filling part, given the generated multi-topic masked sentences and the retrieved knowledge text, we extract candidate knowledge concepts, and use a BERT-based model [14] to get the final description.

3.1. Masked Sentence Generation

Traditional image captioning datasets, such as MSCOCO [38], provide high-quality general captions. Image captioning decoders trained on these corpora predict a probability distribution over a closed vocabulary to generate text. However, these decoders have difficulties in generating specific entities that occur sparsely in the vocabulary. For example, in a vocabulary for art description, the artist name, location, or timeframe may all be in a low frequency. Moreover, the desired art descriptions

³Despite the domain gap, ResNet pre-trained on ImageNet dataset has been shown to work well for art images [51, 59].

should contain external knowledge not directly present in the image. We address these issues by only relying on the decoder to generate masked sentences, which are later completed by the knowledge extraction and filling module.

3.1.1 Data Preprocessing

Given set \mathcal{D} of descriptions about paintings, we obtain a training corpus for the masked sentence generation part by performing Named-Entity-Recognition (NER). Specifically, we apply Stanford CoreNLP name tagger [43] to the descriptions to extract entities of the following types: person, location, organization, ordinal, number, date, and misc. Then, we replace the found entities in the description with their corresponding entity type, e.g.:

An account of **Vasari** says that **Signorelli** wanted to represent in the figure of the naked **Christ** his own son, who died of plague in **1502**.

is transformed into

An account of [person] says that [person] wanted to represent in the figure of the naked [person] his own son, who died of plague in [date].

3.1.2 Topic Decoder

We envisage a decoder that can handle multi-topic description generation: given a visual feature V and a desired topic d , the decoder should generate corresponding topic-related masked sentences. In this section, we first introduce a baseline decoder that generates topic-agnostic masked sentences. Then, we explore two variants to address the multi-topic challenge. Finally, we explain how the multi-topic masked description is generated. Figures illustrating each decoder can be found in Appendix A.

Baseline Decoder Following Xu *et al.* [70], we employ a long short-term memory (LSTM) [25]-based decoder to decode image visual features into masked sentences. The decoder generates one word \mathbf{y}_t (in the one-hot vector representation) at each time step t based on the attention-based visual context vector \mathbf{z}_t , the previous hidden state \mathbf{h}_{t-1} , and the previously generated word \mathbf{y}_{t-1} . Formally:

$$\mathbf{h}_t = \text{LSTM}([\mathbf{z}_t, \mathbf{h}_{t-1}, \mathbf{E}\mathbf{y}_{t-1}]) \quad (1)$$

$$g_{ti} = f_{att}(\mathbf{v}_i, \mathbf{h}_{t-1}) \quad (2)$$

$$\alpha_t = \text{softmax}(\mathbf{g}_t) \quad (3)$$

$$\mathbf{z}_t = \sum_{i=1}^L \alpha_{ti} \mathbf{v}_i, \quad (4)$$

where $[\cdot]$ denotes concatenation, \mathbf{E} is an embedding matrix, f_{att} is a trainable function for predicting the attention weights, for which we use a multilayer perceptron, and $\alpha_t = \{\alpha_{t1}, \dots, \alpha_{tL}\}$ are attention scores that sum to one. The visual context vector \mathbf{z}_t is a dynamic representation of the relevant part of the input image at time t . Based on the LSTM state \mathbf{h}_t and visual context vector \mathbf{z}_t , we calculate the output word probability with a fully-connected layer as:

$$p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, V) = \text{softmax}(\mathbf{W}_y [\mathbf{h}_t, \mathbf{z}_t] + \mathbf{b}_y), \quad (5)$$

where \mathbf{W}_y and \mathbf{b}_y are the parameters in the fully-connected layer. Given the ground truth masked sequence $\mathbf{y}_{1:T}^*$, the weights of the decoder are optimized by minimizing the negative log-likelihood in training as:

$$\mathcal{L}_{\text{mle}} = - \sum_{t=1}^T \log p(\mathbf{y}_t^* | \mathbf{y}_{1:t-1}^*, V). \quad (6)$$

Topic Parallel Decoder Due to the linguistic distinctions that different artistic topics may present, we propose to use different decoders to generate masked sentences for different topics independently, *i.e.*, using different baseline decoders as *sub-decoders* for each topic, respectively. This parallel setting is intuitive as it divides topic-related sentences into different decoding branches, enabling the decoders to not disturb each other. Formally, the parallel decoder can be formulated as:

$$\mathbf{y}_{1:T}^{(d)} = \text{Parallel}(V, d) \quad (7)$$

where d is the topic label, which also serves as a selector for the sub-decoders. Within each sub-decoder, the computation is the same as in the baseline decoder:

$$\mathbf{h}_t^{(d)} = \text{LSTM}^{(d)}([\mathbf{z}_t^{(d)}, \mathbf{h}_{t-1}^{(d)}, \mathbf{E}^{(d)}\mathbf{y}_{t-1}^{(d)}]) \quad (8)$$

Equations (2)–(4) can be written equivalently. The different sub-decoders are optimized separately during training.

Topic Conditional Decoder To improve the computational efficiency as well as to leverage common knowledge among the different topics, we also propose a single-model solution for multi-topic description generation. Inspired by stylized captioning [23], we explore a topic conditional decoder. The conditional decoder injects a topic conditional vector into the decoding process, formulated as:

$$\mathbf{y}_{1:T}^{(d)} = \text{Conditional}(V, d) \quad (9)$$

Specifically, the topic label d is transformed into a N_{topic} -dimensional one-hot vector \mathbf{d}' to represent N_{topic} topics, where each element represents the corresponding topic. Then, we feed \mathbf{d}' into a topic embedding layer and concatenate the resulting vector with the standard inputs of the baseline decoder as:

$$\mathbf{h}_t = \text{LSTM}([\mathbf{z}_t, \mathbf{h}_{t-1}, \mathbf{E}\mathbf{y}_{t-1}, \mathbf{E}_{\text{topic}}\mathbf{d}']) \quad (10)$$

where $\mathbf{E}_{\text{topic}}$ is the topic embedding matrix. To ensure that the generated masked sentences correctly contain the target topic, we employ a topic classifier TC for constrain, *i.e.*, $TC(\text{Conditional}(V, d)) \rightarrow d$. The topic classifier is implemented as in TextCNN [31] and is jointly optimized with the topic conditional decoder with a classical *cross-entropy loss* \mathcal{L}_{ce} . Overall, the objective $\mathcal{L}_{\text{cond}}$ of topic conditional decoder is:

$$\mathcal{L}_{\text{cond}} = \mathcal{L}_{\text{mle}} + \mathcal{L}_{\text{ce}} \quad (11)$$

Multi-Topic Masked Description Given N_{topic} topics, we generate N_{topic} masked sentences, one per topic, with either the parallel or the conditional topic decoder. The full multi-topic masked description is then the concatenation of the N_{topic} independently generated masked sentences.

3.2. Knowledge Retrieval

To address the challenge of generating informative descriptions, we rely on external knowledge bases, such as Wikipedia. We use DrQA [7], an efficient document retriever, to find relevant information. Formally, a given query text q and all articles $c_j \in C$ in the knowledge base are tokenized and encoded as TF-IDF vectors, denoted as $\hat{\mathbf{q}}$ and $\hat{\mathbf{c}}_j$, respectively. Then, a similarity score is computed as:

$$s_j = \frac{\hat{\mathbf{q}}^\top \hat{\mathbf{c}}_j}{\|\hat{\mathbf{q}}\| \|\hat{\mathbf{c}}_j\|} \quad (12)$$

Unlike previous work that uses off-the-shelf tags [40] or text [3] as query to the knowledge base, we automatically constitute a query from the image by extracting: 1) *Artistic attributes* using a multi-task attribute prediction model [17]. Specifically, for each painting we predict its *artist*, *type*, *timeframe*, and *school*; and 2) *Visual concepts* using an object detection model [47] pre-trained in Visual Genome

[33]. Concepts such as person, apple, etc. are extracted to describe the general content of the image, removing visual concepts that are unlikely to appear in paintings, such as cell phone. The words from the two sources are appended together to constitute our q . As an output of the knowledge retrieval module, we return the top-5 article c_j 's with the highest score s_j 's. To improve the ranking accuracy, we pre-process each c_j as 1) stop word removal and word stemming, and 2) bigram TF-IDF.

3.3. Knowledge Extraction and Filling

In this module, we fill the masked concepts in the generated multi-topic masked descriptions with one or several knowledge words. Given the top-5 articles from the knowledge retrieval part, we further narrow down the knowledge space by extracting named entities, again using Stanford CoreNLP. We use the extracted named entities and the *artistic attributes* from Sec. 3.2 to compose a set of candidate words G . Then, we train a BERT-based model [14] as a sequence to sequence task to find appropriate words from G to fill the blanks in the generated multi-topic masked descriptions. Specifically, we generate an input sequence as:

$$SEQ_{in} = [[CLS], y, [SEP], k] \quad (13)$$

where k is a sequence with the concatenation of all the words in G and y denotes the sequence of words in the multi-topic masked description. The output description is:

$$SEQ_{out} = \text{BERT}(SEQ_{in}) \quad (14)$$

where $\text{BERT}(\cdot)$ is trained by minimizing the *cross-entropy loss* to produce the original image descriptions in \mathcal{D} .

4. Experiments

Here we describe the experiments and their results. Implementation details can be found in Appendix B.

Art Dataset We use the SemArt dataset [18], which consists on 21,384 painting images. Each image is associated with an artistic comment and seven attributes, such as artist, title, or date. The dataset is split into 19,244 images for training, 1,069 for validation, and 1,069 for test.

Artistic Topic Annotation To investigate multi-topic description generation, we annotate the original comments in SemArt with their correspondent artistic topic. Following art historian protocols [2] we use three topics: 1) **content**, which describes what the artwork is about, i.e. the *message*; 2) **form**, which describes how the work looks, i.e., the constituent elements of the work independent of their meaning; and 3) **context**, which describes in what circumstances the work is or was. We rely on Amazon Mechanical Turk⁴

⁴<https://www.mturk.com/>

(AMT). We split the original comments on the train and test set into individual sentences and ask workers to annotate each sentence with one of the three topics. Workers are exposed to the image, the original full comment, the title, the artist name, and the year of creation. In total, 17,249 images along with 33,543 sentences are annotated.⁵

Knowledge Base As external source of information, we use the 2016-12-21 dump⁶ of English Wikipedia. For each page, only the plain text is extracted. All structured and non-text data sections such as lists and figures are stripped. After discarding internal disambiguation, list, index, and outline pages, we retain 5,075,182 articles.

4.1. Human Evaluation

The evaluation of generated text is a challenging task [6], due to the complexity of automatically measuring not only grammatical correctness but also veracity, informativeness, and diversity. Automatic metrics designed to evaluate factual tasks such as machine translation (e.g., BLEU [46]) or image captioning (e.g., CIDEr [63]) do not work well on more creative tasks such as ours. Following previous work [15, 52, 62], we based our evaluation on how well humans perceive the text generate by our models.

We conduct a human evaluation on AMT on 100 randomly selected validation paintings. For each painting, we show a generated description to 3 annotators, together with the image, the original SemArt comment, the title, the artist, and the creation year. We ask annotators to rate each description according to the metrics below (higher is better):

- *Understandable*: integer from 1 to 4 measuring if the description can be understood by a human.
- *Relevance*: integer from 1 to 4 measuring if the description is relevant to the given painting.
- *Veracity*: integer from 1 to 4 measuring if the description is correct according to the given information.
- *Content existence*: 1 if the description contains information about the topic **content**, and 0 otherwise.
- *Form existence*: 1 if the description contains information about the topic **form**, and 0 otherwise.
- *Context existence*: 1 if the description contains information about the topic **context**, and 0 otherwise.

Results for different variants of our proposed framework are summarized in Table 1. The baseline model (SAT [70]) do not use topic modeling or external knowledge. The other three models (denoted as Ours) use the parallel decoder, the knowledge retrieval, and knowledge extraction and filling modules. The main difference between them lies in the source of external knowledge used at inference. Specifically, in Ours (Wikipedia), we use Wikipedia as knowledge

⁵We exclude some images with non-meaningful comments from the annotation, e.g. “Catalogue numbers: F 526”.

⁶<https://dumps.wikimedia.org/enwiki/latest/>

Table 1. **Human evaluation.** Human ratings (mean and standard deviation) on generated descriptions according to six metrics.

Model	Knowledge	Understand	Relevance	Veracity	Content	Form	Context
SAT [70] (Baseline)	-	3.62 ± 0.63	1.94 ± 0.94	1.30 ± 0.70	0.35 ± 0.48	0.05 ± 0.21	0.65 ± 0.48
Ours (Wikipedia)	Retrieved Wikipedia	2.71 ± 0.64	2.29 ± 1.04	1.56 ± 0.64	0.73 ± 0.45	0.33 ± 0.47	0.83 ± 0.38
Ours (SemArt)	Retrieved SemArt	2.77 ± 0.62	2.02 ± 1.08	1.39 ± 0.59	0.75 ± 0.44	0.37 ± 0.48	0.90 ± 0.30
Ours (Oracle)	Original SemArt	2.71 ± 0.64	2.49 ± 1.00	1.72 ± 0.70	0.76 ± 0.43	0.38 ± 0.49	0.91 ± 0.29

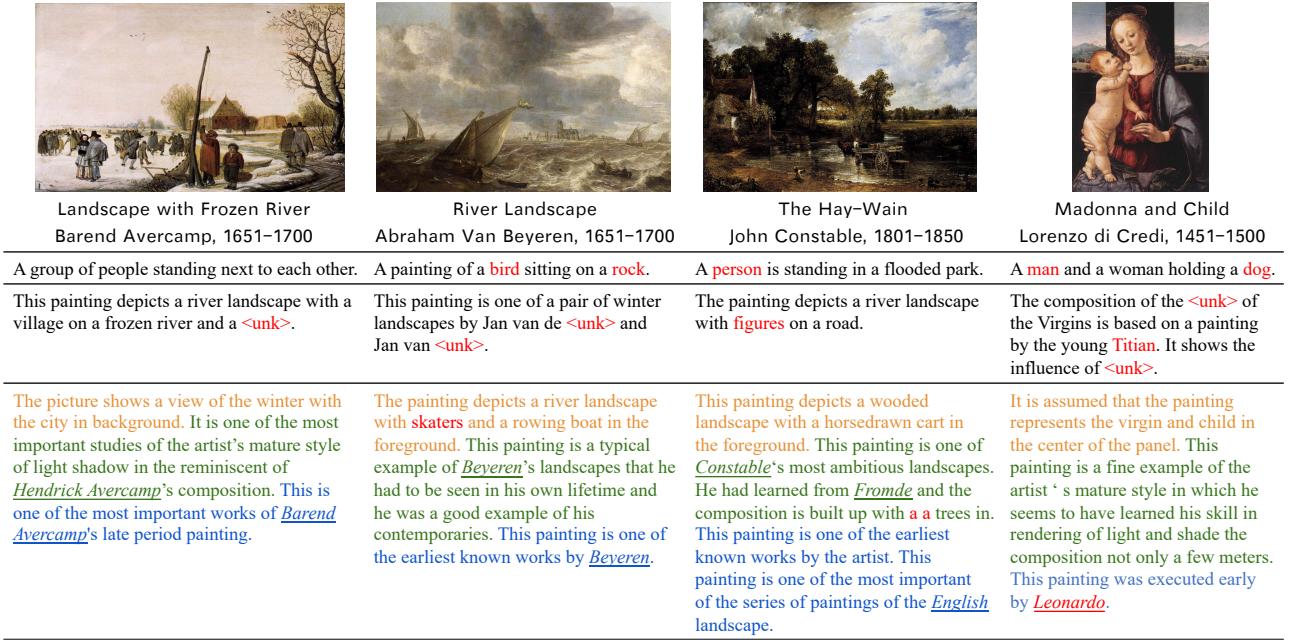


Figure 3. **Qualitative evaluation.** Top row shows four test paintings together with their title, artist, and creation timeframe. The next rows contain the descriptions generated by SAT-Transfer, SAT-Baseline, and our method, respectively. Incorrect words are highlighted in red.

source. In Ours (SemArt), we build a knowledge base with the comments in SemArt and find the most relevant one with knowledge retrieval module. Finally, in Ours (Oracle), we use the original associated comments, assuming a perfect accuracy in the knowledge retrieval module.

Results show that the baseline achieves the highest score in the *Understandable* metric. This is because the text generated is simpler and shorter than the text generated with our models: *e.g.*, the average number of words of SAT output is 30.9, while in Ours (Oracle) it is 71.8. The shorter the sentence, the less prone to contain grammatical errors. Nevertheless, from all the other metrics, which are more related to the informative aspect of the description, we observe that our framework outperforms the baseline by a large margin.

When comparing our three different settings, Ours (Oracle) achieves the best performance in all the metrics, with a large margin in *Relevance* and *Veracity*. This is natural, as the ground truth comments are used as the source of knowledge. In the existence metrics (*i.e.* *Content*, *Form*, and *Context*), the performance of the three models is very close, all of them outperforming the baseline by a large margin. This shows the effectiveness of the topic decoder, which is able

to produce different types of sentences. Ours (Wikipedia) performs the worst among the three settings, as it uses the most challenging source of knowledge and also suffers from a domain-gap between the source of training (SemArt comments) and the source of testing (Wikipedia articles).

4.2. Qualitative Analysis and Examples

We further explore the results of our framework with a qualitative analysis in Fig. 3. We compare three methods. In the first row, we use SAT [70] trained on MSCOCO, namely SAT-Transfer. The generated sentences are very similar to the standard captions generated for natural images. This model suffers from: 1) not containing art-specific information, and 2) a visual domain-gap when transferring from natural images to paintings, with miss-detected concepts, such as ‘dog’ and ‘person’, in *Madonna and Child* and *The Hay-Wain*, respectively. In the second row, SAT is trained on the SemArt dataset, *i.e.* SAT-Baseline. The results show that it can generate some short and understandable sentences as in *The Hay-Wain*. The main problem, however, is that it does not contain background knowledge, leaving some specific knowledge words as *unk* in the output. Finally, the last row



Altarpiece of Calvary
Bernaert van Orley, 1501–1550

Unsupervised

The picture shows NUMBER_ of the scenes from the life of PERSON_. This is the central panel of the central panel in the LOCATION_. The altarpiece was executed for the church of LOCATION_ in LOCATION_.

Semi-supervised

In the centre of the triptych the central <unk> represents the Adoration of the Magi and the <unk>.

The composition is based on a drawing by PERSON_ in LOCATION_. This is NUMBER_ of the most important works of MISC_ art.

Supervised (Annotated)

The picture shows NUMBER_ of the scenes from the life of PERSON_ in the centre of the triptych in the LOCATION_.

This panel is divided into NUMBER_ parts. The composition is influenced by PERSON_.

This painting is NUMBER_ of the most important examples of the artist's career in LOCATION_. The painting is signed and dated lower right.

Figure 4. **Topic model comparison.** The topic labels are predicted with different settings to train the topic decoder.

shows the proposed framework with the parallel decoder and SemArt as source of knowledge. The topic decoder effectively generates different sentences for each topic, shown in different colors. Moreover, each sentence includes relevant knowledge, such as ‘Barend Avercamp’ in *Landscape with Frozen River*, and ‘Beyeren’ in *River Landscape*. More examples can be found in Appendix E.

These results also reveal some of our method limitations. Missdetection of the visual content (a common mistake in traditional image captioning) is shown in *River Landscape*, where the boats are confused by ‘skaters’. In *The Hay-Wain*, there is a syntax error (‘a a’) in the **form** sentence, which is produced in the knowledge filling part, based on BERT, *i.e.* it does not appear in the masked-sentence decoder based on LSTM. We hypothesized that it is because the size of training set is relatively small for BERT to learn the language structure of our task. This can be solved by applying language augmentation techniques. Finally, in *Madonna and Child*, there is an imperfect-knowledge mistake, in which the painting is assigned to the wrong artist.

4.3. Modules Analysis

Exploring Artistic Topic Our topic decoder relies on the availability of a corpus with topic labels for training. For a more general approach, we explore un/semi-supervised topic models to automatically predict the topic label for each sentence. For the unsupervised setting, we use Latent Dirichlet Allocation (LDA) [4], which assumes each topic is a mixture over an underlying set of words. While for the semi-supervised setting, we use Guided-LDA [28], which

incorporates lexical priors, manually set as a list of *seed* words, to LDA. In our case, the *seed* words are obtained from a subset of sentences with topic labels, from which we select the top-10 words with the highest frequency. The larger the subset, the more accurate the *seed* words obtained. Here we use 3245 samples. The artistic comments are split into individual sentences and pre-processed to optimize each topic model method. After training, the topic model predicts a pseudo topic label for each sentence and used to train our parallel topic decoder.

LDA achieves 43.3% accuracy when predicting topic labels on the test set, while Guided-LDA achieves 51.6%. To reduce the influence of the knowledge filling module, in Fig. 4 we compare the generated masked sentences for the different topic models. In the unsupervised setting, the **content** and **context** sentences are generated correctly, while the **form** sentence is confusing. In the semi-supervised setting, the output looks better, even close to the supervised approach, revealing that only a small amount of topic annotations may be necessary for art description generation.

Knowledge Retrieval The impact of the knowledge retrieval module and how it affects the overall framework has been already discussed in Sec. 4.1. According to Table 1, the easiest to retrieve the knowledge, the highest the scores. In Appendix D, we provide accuracy rates for the knowledge retrieval module on its own with different settings. It can be seen that the performance of the knowledge retrieval module is a crucial bottleneck on our system.

4.4. Comparative Evaluation

Although automatic metrics may not correlate well with the evaluation of the knowledge and creativity required to describe art, we include a standard automatic evaluation for completeness. We compare our proposed model against:

Classic image captioning (1) NIC [64], LSTM-based encoder-decoder model without attention; (2) SAT [70], which incorporates soft-attention (note that this corresponds to the baseline decoder in Sec. 3.1.2); and (3) Att2in [48], similar to SAT but in where the attention-derived context visual feature is only input to the cell node of the LSTM.

Stylized image captioning Our conditional topic decoder can be seen as a use case of stylized image captioning. We compare our framework with (4) MScap [23]. In our reimplementation, we regard *topic* as *style*.

Transformer-based image captioning We evaluate the state-of-the-art Transformer (5) OSCAR [35], which is one of the latest multi-modal approaches in vision-language task. We use its original pre-trained weights and finetune it on the SemArt dataset.

Text-summarization As a different perspective, we generate painting descriptions with text-based summarization methods. That is, given the knowledge articles retrieved in

Table 2. **Comparative evaluation.** We compare our models against multiple alternatives. All the methods are trained on the SemArt dataset under equivalent conditions. Methods using external knowledge (LSA and Ours) use Wikipedia as knowledge source. GM:GreedyMatching, S-T: Skip-Thought, EA: EmbeddingAverage.

Model	BLEU-4	CIDEr	METEOR	ROUGE-L	GM	S-T	EA
1 NIC [64]	7.3	39.4	10.9	28.6	71.5	24.8	64.5
2 SAT [70] (Baseline)	6.5	38.6	11.1	27.5	72.9	26.6	73.3
3 Att2in [48]	4.2	26.4	9.7	25.4	69.1	22.4	59.1
4 MScap [23]	0.4	0.1	6.3	14.2	64.0	19.3	50.7
5 OSCAR [35]	0.1	2.0	2.8	11.3	63.0	33.3	84.8
6 LSA [58]	0.2	0.1	8.5	10.9	75.1	37.4	90.6
7 Ours (Parallel decoder)	8.8	9.1	11.4	23.1	77.6	30.9	92.6
8 Ours (Conditional decoder)	0.9	0.4	5.8	14.8	70.1	27.7	89.3

Sec. 3.2, we summarize their content with (6) LSA [58], a language-independent algebraic method.

Ours We compare our method when using the (7) Topic Parallel Decoder and the (8) Topic Conditional Decoder.

Table 2 shows the comparison results. We adopt a wide range of natural language evaluation metrics, including BLEU-4 [46], CIDEr [63], METEOR [13], ROUGE-L [37], GreedyMatching [50], Skip-Thought [32], and EmbeddingAverage [39]. We notice that (1) the scores are much lower than in traditional image captioning: *e.g.*, the BLEU-4 score in MSCOCO is around 30 [64, 70], while in SemArt it is less than 10; (2) the scores among different metrics shows very large variation. On one hand, this phenomenon shows that the proposed task is very challenging. On the other hand, these automatic metrics may not be appropriate to evaluate the richness and diversity of art description generation. Thus we also include the qualitative comparison results in Appendix E.

Firstly, our parallel decoder performs better than the conditional decoder. It is natural as the parallel decoder has larger capacity in network and can fit different topics in a specific manner. When comparing against the classic image captioning (NIC, SAT, and Att2in), our parallel decoder method outperforms them in 5 out of 7 metrics. The performance gain is mainly caused by the masked-sentence-generation-filling schema that reduces the burden of the decoder to handle low frequency words and named-entities. In MScap, we directly let the model to conditionally learn the topic, without the help of masked-sentence-generation-filling, the performance even getting worse. With respect to the Transformer-based approach OSCAR, it performs very poorly in most of the metrics. We deduce this is caused by: 1) the scale of the training set, and 2) the use of Bottom-Up-Top-Down [1] regional features as image input, which are extracted by an object detector pre-trained in natural images and do not consider the visual domain-gap with paintings.

Finally, the text-summarization method LSA performs very poor in most of the metrics except in Skip-Thought, demonstrating that relying only on external knowledge

without “looking” into the image is not the best approach to generate a description for a painting. This is because with summarization methods: 1) it is not possible to control the generation process for the different topics, 2) they heavily rely on the accuracy of the knowledge retrieval module, and 3) descriptions cannot be generated when the external knowledge of a painting does not exist. According to a random subset of 150 annotated images,⁷ 80.7% of the samples do not have a specific article on Wikipedia. However, 71.3% of the them do have artist article, which is useful to find knowledge but not enough for producing a relevant painting description on its own. Moreover, even when retrieving a non-relevant article, our method can generate correct sentences for the **content** and the **form** topics. The ratio of slots per topic on the SemArt dataset, *i.e.*, average number of slots in **content/form/context** sentences, are 0.98/0.91/2.12, respectively, which shows that **content** and **form** require less external data than **context**. Finally, the *artistic attributes* predicted from the image are also used to fill the slots (Sec. 3.3), which may yield to meaningful information even when the retrieved knowledge fails.

5. Conclusion

We proposed the first multi-topic knowledgeable framework for art description generation. To generate multi-topic descriptions, we annotated an art description dataset with sentence-level topic labels. We explored this problem from multiple views, including proposing two types of topic decoder and experimenting with un/semi-supervised as well as supervised settings. Besides, we introduced the use of external knowledge to enhance the background information in the description of the painting. Comprehensive evaluation and comparison showed the effectiveness of our method, which we hope contributes to guide future research.

Acknowledgement This work is partly supported by JSPS KAKENHI Grant Numbers JP20K19822 and JP18H03264, as well as ROIS NII Open Collaborative Research 2021-21S1002.

⁷Details can be found in Appendix D.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. CVPR*, pages 6077–6086, 2018. 2, 8
- [2] Robert Belton. Art history: A preliminary handbook. *British Columbia: University of British Columbia*, 1996. 2, 5
- [3] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proc. CVPR*, pages 12466–12475, 2019. 2, 4
- [4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. 7
- [5] Gustavo Carneiro, Nuno Pinho Da Silva, Alessio Del Bue, and João Paulo Costeira. Artistic image classification: An analysis on the printart database. In *Proc. ECCV*, pages 143–157. Springer, 2012. 2
- [6] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020. 5
- [7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proc. ACL*, 2017. 3, 4
- [8] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 20(9):2491–2502, 2018. 1
- [9] Elliot J Crowley, Omkar M Parkhi, and Andrew Zisserman. Face painting: querying art with photos. 2015. 2
- [10] Elliot J Crowley and Andrew Zisserman. In search of art. In *Proc. ECCV*, pages 54–70. Springer, 2014. 2
- [11] Elliot J Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. 2014. 1, 2
- [12] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional GAN. In *Proc. ICCV*, pages 2970–2979, 2017. 2
- [13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on statistical machine translation*, pages 376–380, 2014. 8
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186, 2019. 3, 5
- [15] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 5
- [16] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108, 2020. 1
- [17] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. In *Proc. ICMR*, 2019. 2, 4
- [18] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proc. ECCV Workshops*, 2018. 1, 2, 5
- [19] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Proc. ECCV Workshops*, 2020. 2
- [20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, pages 2414–2423, 2016. 1, 2
- [21] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *Proc. ECCV*, pages 0–0, 2018. 1, 2
- [22] Google Arts & Culture. <https://artsandculture.google.com/>. Accessed November 2020. 1
- [23] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *Proc. CVPR*, pages 4204–4213, 2019. 4, 7, 8
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 3
- [25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4
- [26] Nikolai Huckle, Noa Garcia, and Yuta Nakashima. Demographic influences on contemporary art with unsupervised style embeddings. *Proc. ECCV Workshops*, 2020. 2
- [27] Katsushi Ikeuchi and Daisuke Miyazaki. *Digitally archiving cultural objects*. Springer Science & Business Media, 2008. 2
- [28] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proc. ACL*, pages 204–213, 2012. 7
- [29] C Richard Johnson, Ella Hendriks, Igor J Berezhnoy, Eugene Brevdo, Shannon M Hughes, Ingrid Daubechies, Jia Li, Eric Postma, and James Z Wang. Image processing for artist identification. *IEEE Signal Processing Magazine*, 25(4):37–48, 2008. 1, 2
- [30] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *BMVC*, 2014. 1, 2
- [31] Yoon Kim. Convolutional neural networks for sentence classification. In *Proc. EMNLP*, pages 1746–1751, 2014. 4
- [32] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 8
- [33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 5
- [34] Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger. Recognizing art style automatically in painting with deep learning. In *Proc. ACML*, pages 327–342, 2017. 1

- [35] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 7, 8
- [36] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *Proc. CVPR*, pages 12497–12506, 2019. 2
- [37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 8
- [38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014. 3
- [39] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016. 8
- [40] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih Fu Chang. Entity-aware image caption generation. In *Proc. EMNLP*. Association for Computational Linguistics, 2020. 2, 4
- [41] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proc. CVPR*, pages 7219–7228, 2018. 2
- [42] Daiqian Ma, Feng Gao, Yan Bai, Yihang Lou, Shiqi Wang, Tiejun Huang, and Ling-Yu Duan. From part to whole: who is behind the painting? In *Proc. ACM MM*, pages 1174–1182, 2017. 2
- [43] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proc. ACL*, pages 55–60, 2014. 3
- [44] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proc. ACM MM*, pages 1183–1191, 2017. 1
- [45] Thomas Mensink and Jan Van Gemert. The Rijksmuseum challenge: Museum-centered visual recognition. In *Proc. ICMR*, pages 451–454, 2014. 1, 2
- [46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, 2002. 5, 8
- [47] Shaqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 91–99, 2015. 4
- [48] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 2, 7, 8
- [49] Rijksmuseum. <https://www.rijksmuseum.nl/en>. Accessed November 2020. 1
- [50] Vasile Rus and Mihai Lintean. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, pages 675–676. Springer, 2012. 8
- [51] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. Deep transfer learning for art classification problems. In *Proc. ECCV Workshops*, pages 0–0, 2018. 3
- [52] Amaia Salvador, Michal Drozdzał, Xavier Giro-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10453–10462, 2019. 5
- [53] Benoit Seguin, Carlotta Striolo, Frederic Kaplan, et al. Visual link retrieval in a database of paintings. In *Proc. ECCV Workshops*, pages 753–767. Springer, 2016. 2
- [54] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):1–17, 2010. 1, 2
- [55] Shurong Sheng and Marie-Francine Moens. Generating captions for images of ancient artworks. In *Proc. ACM MM*, pages 2478–2486, 2019. 2
- [56] Shurong Sheng, Luc Van Gool, and Marie Francine Moens. A dataset for multimodal question answering in the cultural heritage domain. In *Workshop on Language Technology Resources and Tools for Digital Humanities*, pages 10–17, 2016. 2
- [57] Maria Shugrina, Ziheng Liang, Amlan Kar, Jiaman Li, Anagad Singh, Karan Singh, and Sanja Fidler. Creative flow+ dataset. In *Proc. CVPR*, pages 5384–5393, 2019. 2
- [58] Josef Steinberger, Karel Jezek, et al. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100, 2004. 8
- [59] Gjorgji Strezoski and Marcel Worring. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–21, 2018. 1, 3
- [60] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *Proc. ICIP*, pages 3703–3707. IEEE, 2016. 2
- [61] The metropolitan museum of art. <https://www.metmuseum.org/>. Accessed November 2020. 1
- [62] Jukka Toivanen, Hannu Toivonen, Alessandro Valitutti, Oskar Gross, et al. Corpus-based generation of content and form in poetry. In *Proceedings of the third international conference on computational creativity*. University College Dublin, 2012. 5
- [63] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, pages 4566–4575, 2015. 5, 8
- [64] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, pages 3156–3164, 2015. 2, 7, 8
- [65] Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. Show, recall, and tell: Image captioning with recall mechanism. In *Proc. AAAI*, pages 12176–12183, 2020. 2
- [66] Web Gallery of Art. <https://www.wga.hu/>. Accessed November 2020. 1

- [67] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proc. ICCV*, pages 1202–1211, 2017. [2](#)
- [68] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *Trans. PAMI*, 40(6):1367–1381, 2017. [2](#)
- [69] Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. In *Proc. NeurIPS*, pages 6584–6593, 2018. [1, 2](#)
- [70] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057, 2015. [2, 3, 4, 5, 6, 7, 8](#)
- [71] Lei Xu, Albert Merono-Penuela, Zhisheng Huang, and Frank Van Harmelen. An ontology model for narrative image annotation in the field of cultural heritage. In *Workshop on Humanities in the Semantic web (WHiSe)*, pages 15–26, 2017. [2](#)
- [72] Lei Xu and Xiaoguang Wang. Semantic description of cultural digital images: using a hierarchical model and controlled vocabulary. *D-Lib magazine*, 21(5/6), 2015. [2](#)
- [73] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proc. CVPR*, pages 6580–6588, 2017. [2](#)
- [74] Yimin Zhou, Yiwei Sun, and Vasant Honavar. Improving image captioning by leveraging knowledge graphs. In *Proc. WACV*, pages 283–293. IEEE, 2019. [2](#)