Visual Narratives: Large-scale Hierarchical Classification of Art-historical Images

Matthias Springstein[1,2] Stefanie Schneider[3] Javad Rahnama[4] Julian Stalter[3]
Maximilian Kristen[3] Eric Müller-Budack[1,2] Ralph Ewerth[1,2]
[1] TIB – Leibniz Information Centre for Science and Technology, Germany
[2] L3S Research Center, Leibniz University Hannover, Germany
[3] Ludwig Maximilian University of Munich, Germany [4] Reply GmbH, Germany
https://github.com/TIBHannover/iconclass-classification

Abstract

Iconography refers to the methodical study and interpretation of thematic content in the visual arts, distinguishing it, e.g., from purely formal or aesthetic considerations. In iconographic studies, Iconclass is a widely used taxonomy that encapsulates historical, biblical, and literary themes, among others. However, given the hierarchical nature and inherent complexity of such a taxonomy, it is highly desirable to use automated methods for (Iconclass-based) image classification. Previous studies either focused narrowly on certain subsets of narratives or failed to exploit Iconclass's hierarchical structure. In this paper, we propose a novel approach for Hierarchical Multi-label Classification (HMC) of iconographic concepts in images. We present three strategies, including Language Models (LMs), for the generation of textual image descriptions using keywords extracted from Iconclass. These descriptions are utilized to pre-train a Vision-Language Model (VLM) based on a newly introduced data set of 477,569 images with more than 20,000 Iconclass concepts, far more than considered in previous studies. Furthermore, we present five approaches to multi-label classification, including a novel transformer decoder that leverages hierarchical information from the Iconclass taxonomy. Experimental results show the superiority of this approach over reasonable baselines.

1. Introduction

Iconography, as established by Panofsky [38], entails the systematic analysis of content or meaning in visual art, distinguishing these elements from mere formal characteristics. For this purpose, *Iconclass*, short for *Iconographic Classification System*, provides a widely used taxonomy for annotating visual content [51, 53].[1] In particular, the system makes it possible to convey semantically complex narratives, which are found especially in historical, biblical, and literary themes; see Figure 1 for an example. Corpora labeled with *Iconclass* are essential for text-based retrieval of artworks with certain narratives [7]: they provide a foundation for the digital exploitation of the collections. However, despite obvious advantages, galleries, libraries, archives, and museums—known as the GLAM institutions—only sporadically utilize *Iconclass*. This is due, on the one hand, to limited resources for (human) annotation, and on the other hand to the inherent complexity of the system. It is therefore highly desirable to develop an efficient indexing process via automated image classification methods.

Figure 1. Utilizing *Iconclass*, Hans Holbein the Elder's *Last Supper* (1501) could be labeled with the notations `73D24` ("Last Supper [...]") and `41C3` ("laid table [...]").

To date, related work on visual art objects primarily considered classification tasks of image-related metadata features, such as the identification of artists, genres, or creation dates [6, 14, 29, 33, 34, 48, 49, 59]. While these tasks are important, they do not take into account the classification of semantic concepts represented in artworks. Previous studies often focused narrowly on particular subsets of

[1] https://iconclass.org/ (last accessed on 2023-11-08).

narratives [35, 44]. Furthermore, attempts to comprehensively map the entire *Iconclass* system failed to leverage its hierarchical structure: Banar et al. [3] investigated the feasibility of ascribing *Iconclass* notations through cross-modal retrieval, while Cetinic [9] created image-text pairs that were used in an image captioning task. In addition, Vision-Language Models (VLMs) have not yet been exploited for the classification of iconographic concepts, although they achieve impressive performance in many downstream applications, including the classification of metadata in art-historical images, e.g., the country of origin [14].

In this paper, we propose a novel approach to extensively classify hierarchical iconographic concepts in order to mitigate, or at least minimize, the need for manual annotation. Our contributions can be summarized as follows: (i) We propose three strategies that use, for example, LMs [41] and VLMs [15, 26] to automatically create image descriptions using keywords provided by *Iconclass*; (ii) We apply contrastive pre-training with synthetic image-text pairs and show an improved performance for rare concepts; (iii) We present five multi-label classification approaches, including a novel transformer decoder that leverages hierarchical information from the *Iconclass* taxonomy and, to the best of our knowledge, is the first decoder that can handle multi-label classification; (iv) Compared to previous studies [3, 9, 44], we expand the scope of classifiable iconographic concepts by introducing a new data set of 477,569 images with more than 20,000 unique *Iconclass* concepts. The source code, models, and data set will be made publicly available.[2]

The remainder of the paper is structured as follows. In Section 2, we review related work. Section 3 describes our proposed transformer model for Hierarchical Multi-label Classification (HMC) of art-historical concepts, which uses contrastive pre-training with synthesized image-text pairs. Section 4 introduces a novel data set, while Section 5 presents experimental results for several benchmarks. We conclude with Section 6 and outline areas for future work.

2. Related Work

The rapidly advancing field of Computer Vision (CV), fueled by sophisticated deep learning models, is facilitating the in-depth analysis of complex data; a task that, until recently, could only be performed by human experts. The implications of this development are particularly significant when applied to the field of visual art—which frequently encompasses representations and abstract concepts that differ considerably from real-world data. This section reviews related work in CV for the visual arts, as well as HMC, which is crucial for leveraging the hierarchical structure of the *Iconclass* taxonomy.

[2] https://github.com/TIBHannover/iconclass-classification (last accessed on 2023-11-08).

Computer Vision (CV) for the Visual Arts: Research in CV for the visual arts focuses on several key areas including, but not limited to, aesthetic quality assessment [2], human pose estimation [23, 32, 47], sentiment analysis [36, 58], correspondence matching [24, 46], and visual question answering [18]. To date, however, research efforts have largely been devoted to classification tasks of image-extrinsic features, such as the identification of artists, genres, or creation dates [6, 14, 29, 33, 34, 48, 49, 59]. While these tasks are significant, they only address tangible aspects of the domain, leaving, e.g., content-based features relatively unexplored. Indeed, the classification of intrinsic features, particularly those related to iconographic elements, has been inadequately attended to: prevailing studies have often focused narrowly on certain subsets of narratives, such as the prediction of saints [35, 44]. A deviation from this tendency is illustrated by Gupta et al. [21], who applied image captioning models based on an encoder-decoder architecture to art-historical images spanning across nine iconographies. Moreover, there have been attempts to comprehensively map the entire *Iconclass* system. Banar et al. [3] conducted an exploratory investigation into the feasibility of ascribing *Iconclass* notations, with up to five levels of depth, through cross-modal retrieval. Cetinic [9] transformed *Iconclass's* textual correlates into descriptions to create image-text pairs to fine-tune a transformer model, morphing the classification into an image captioning task. Compared to these works, we not only scale our approach to the entire *Iconclass* system of over 20,000 art-historical concepts, but fully exploit its hierarchical structure.

Hierarchical Multi-label Classification (HMC): Approaches to HMC can exploit the hierarchical structure of taxonomies such as *Iconclass*. They have been used for many tasks, e.g., event classification [37] and geolocation estimation [13] involving real-world data like text [22] and image [11]. Hierarchical information provides the opportunity to generate a chain of coarse-to-fine labels describing an object [10, 13] or to unify data sets into a common annotation scheme [42]. Prior work on image classification leverages hierarchical dependencies in several ways: (i) some work maps the hierarchical relationship between individual classes in the *embedding space* [1, 4, 19, 60]; (ii) *hierarchical loss functions* have been presented to take into account the hierarchical information from the ontology during optimization [5, 16, 20, 20, 37, 61]; (iii) *hierarchical architectures* design the architecture of the model so that it can solve a particular hierarchical ontology [50, 57, 62]. Most of the papers investigate problems where each image has only one annotation [11, 37] or use structures that are difficult to transfer from several thousand of classes [20]. Our proposed approach scales to several thousand concepts and is suitable for multiple annotations per sample.

3. Hierarchical Multi-label Classification of Iconographic Concepts

In this section, we introduce our approach to the Hierarchical Multi-label Classification (HMC) of iconographic concepts in images. First, we describe the *Iconclass* notation scheme, which provides a taxonomy of iconographic concepts (Section 3.1). In Section 3.2, we suggest three approaches to synthesize textual descriptions based on keywords for *Iconclass* concepts that are used to pre-train VLMs according to Section 3.3. Finally, we use the image encoder of a pre-trained VLM along with several approaches for HMC, including a novel transformer-based classification decoder that incorporates structured information from the *Iconclass* taxonomy (Section 3.4).

3.1. Iconclass Notation Scheme

While *Iconclass* is explicitly designed for the iconography of Western fine art, it also encompasses universal definitions ranging from natural phenomena to socio-economic aspects [53]. As shown in Figure 2a, each definition within the taxonomy is represented by a unique combination of alphanumeric characters, referred to as the 'notation,' hereafter denoted as *Iconclass* concept C, and an explanatory 'textual correlate' T_C, accompanied by a corresponding set of keywords $\mathbb{K}_C$. A notation comprises at least one digit symbolizing the first level of hierarchy or 'division.' This can be followed by another digit at the secondary level, and one or two (identical) capital letters at the tertiary level. This structure, referred to as the 'basic notation,' can be further supplemented with auxiliary components [51, 53]. Notations may be linked using a colon to establish a relationship between two or more notations, as in `79C52:42E3`.

3.2. Image-Text Pairs for Contrastive Pre-Training

Several recent methods have shown that Vision-Language Models (VLMs) pre-trained with image-text pairs from other domain-relevant [14, 27] or large-scale data sets in general [28, 40] can significantly improve the performance of many downstream applications. As shown in Figure 2b, it is necessary to provide textual descriptions for the corresponding images in order to optimize a VLM with contrastive pre-training (Section 3.3). However, these descriptions can be difficult to obtain, and in our case they are only available for the *Iconclass* concepts (Section 3.1) represented in an image, but not for the image itself.

Similar to Conde and Turgutlu [14], our goal is to create a textual description D for an image I that has been labeled with k *Iconclass* concepts $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$. For this purpose, we leverage the associated keywords $\mathbb{K}_C$ of each *Iconclass* concept $C \in \mathbb{C}$. However, unlike the data set used by Conde and Turgutlu [14], these keywords do not contain a subdivision into categories such as origin, material, or dimension. Thus, we cannot use a generic free-form to create textual descriptions. Instead, we consider the following three approaches to generating pairs of images and textual descriptions for language-supervised pre-training of VLMs; examples are shown in Figure 2a.

Descriptions based on Iconclass Keywords (`KW`) In this baseline strategy, a textual description D is created by comma-separating all unique keywords $\mathbb{K}_C$ from each annotated *Iconclass* concept $C \in \mathbb{C}$ of an image I. However, comma-separating keywords leads (i) to a loss of information when projected into a textual description space, and (ii) to redundancies (e.g., *stained glass*, *glass*) that should be avoided given *CLIP's* limited textual context length. Furthermore, the resulting descriptions are different from those typically used to train CLIP (e.g., LAION-400M [45]), which may increase the optimization time.

Descriptions based on Large Language Models (`GPT`) To address the issues with the aforementioned `KW` approach, we use a Language Model (LM) to generate descriptions based on a set of keywords provided by *Iconclass*. The intention behind this idea is that LMs can generate shorter and more natural image descriptions compared to simply chaining keywords together. To this end, we first fine-tune a GPT-2 model [41] for the task of image captioning, using a set of keywords as input. To train such a model, we extract named entities as keywords from ground-truth image captions provided by *MS COCO* [30] using *Wikifier* [8]. The goal of the LM is to reproduce the ground-truth caption from *MS COCO* using this set of named entities as input. During inference, we use the trained model to generate captions that serve as textual description D for an image I based on a set of keywords $\mathbb{K}_C$ for each Iconclass concept $C \in \mathbb{C}$. In doing so, we often found it helpful to provide the GPT-2 model with the start of the caption, such as "A photo of ..." or "A drawing of ...", along with the keywords.

Descriptions based on Vision-language Models (`BLIP`) The introduction of instruction-based fine-tuning of Large Language Models (LLMs) allows descriptions to be created without the need to optimize the LLM itself for that specific task. Another difference to the previous `KW` and `GPT` approaches is that in addition to the *Iconclass* keywords, the corresponding image is also used as input. We use the *BLIP2* model [26] fine-tuned to instructions [15], which in turn consists of a *CLIP* vision encoder [40] and a *FlanT5* language model [12]. To create a textual description D, we use the corresponding image I, all n associated keywords $K \in \mathbb{K}_C$ for each *Iconclass* concept $C \in \mathbb{C}$, and the following instruction as input: "Create a description of up to three sentences for this image and try to include the terms $\langle K_1 \rangle, \langle K_2 \rangle, \ldots, \langle K_n \rangle$."

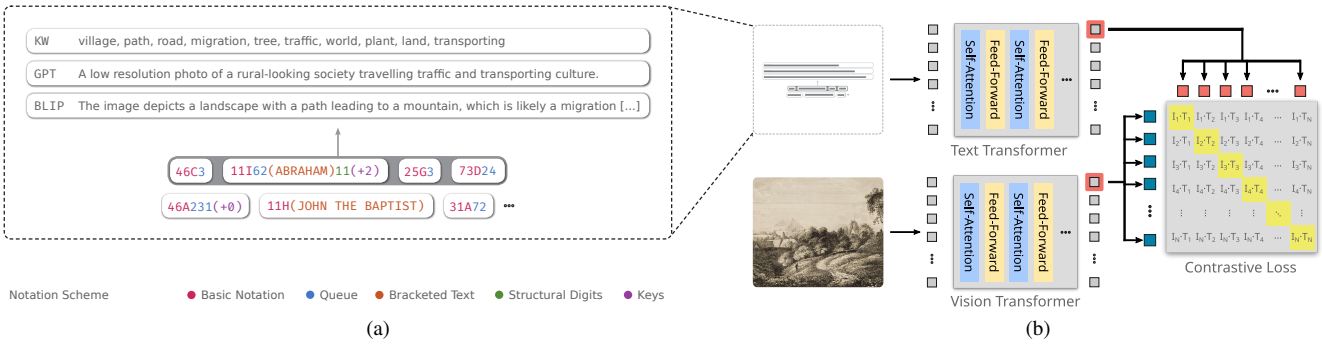

Figure 2. The proposed approach for Vision-Language Model (VLM) pre-training based on images labeled according to the *Iconclass* notation scheme: (a) Based on the keywords $\mathbb{K}_C$ from each annotated *Iconclass* concept $C \in \mathbb{C}$ of an image I, three strategies are used (`KW`, `GPT`, `BLIP`) to create image descriptions (Section 3.2); (b) They are used for contrastive pre-training of *CLIP* [40] (Section 3.3).

3.3. Constrastive Pre-Training

Given a data set containing a set of images and associated annotations according to the *Iconclass* notation scheme, we create textual descriptions based on the associated keywords using *one* of the methods presented in the previous section. As shown in Figure 2b, the Info Noise-Contrastive Estimation (InfoNCE) contrastive loss [52] is used to optimize the VLM model *CLIP* [40] based on these image-text pairs. We refer to Section 5.1 for more details on optimization. After pre-training, the weights of the vision transformer are further optimized during classifier training for the hierarchical classification of iconographic concepts, as explained in Section 3.4.

3.4. Iconographic Concept Classification

Based on the pre-trained image encoder of the VLM, we aim to train an image classifier that predicts the corresponding iconographic concepts. Please note that we do not use the text encoder for classification as there is no textual information available for the images during testing. The *Iconclass* taxonomy consists of L levels of granularity, each with its own set of concepts $\mathbb{C}_l$, $l \in [0, L-1]$. An image is labeled with a set of *Iconclass* concepts $C \in \mathbb{C}$ based on the *Iconclass* notation scheme (Section 3.1). Note that the annotated concept can be at any level of granularity in the taxonomy. The goal of the classifier is to predict the set of ground-truth *Iconclass* concepts for a given image.

For the prediction, we use the embedding from the classification token of the vision transformer [17] as input. This token is also used during the *CLIP* pre-training (Section 3.3). Subsequently, a fully-connected layer with neurons corresponding to the amount of classes (i.e., iconographic concepts) is added as the classification head. In the remainder of this section, we propose one zero-shot classification approach (Section 3.4.1) and four supervised classifiers (Section 3.4.2 to Section 3.4.5).

3.4.1 Zero-shot CLIP-based Classification (`CLIP`)

For zero-shot classification, we measure the similarity between an image and the textual descriptions for all *Iconclass* concepts (as shown in Figure 2b) using a *CLIP* model pre-trained according to one of the strategies in Section 3.3. Unlike the following supervised classifiers, this method does not require a classification head and further optimization. More specifically, for each *Iconclass* concept $C \in \mathbb{C}$, we create a textual description based on the associated keywords $\mathbb{K}_C$. To make this procedure more robust, we follow Radford et al. [40] and combine the keywords $\mathbb{K}_C$ with a set of pre-defined templates to create *hard prompts* (e.g., "This is a photo of ⟨class⟩", "A drawing of ⟨class⟩"). These *hard prompts* serve as input to the text encoder from Section 3.2 to create textual embeddings for the given *Iconclass* concept. Finally, we compute the dot product between the average textual embeddings of all n concepts and the image embeddings produced by the vision transformer, resulting in a similarity vector $\hat{\mathbf{y}} \in \mathbb{R}^n$. Unlike for following classification methods that aim for a binary decision, this approach uses the similarity vector as a ranking to calculate the mean Average Precision (AP) for HMC according to Section 5.2.

3.4.2 Flattened Classification (`Flat`)

To create a baseline classifier, we 'flatten' the concepts $\mathbb{C}_h$ of all hierarchy levels in *Iconclass* (similar to Figure 3). We use a fully-connected layer with as many neurons n as there are concepts at all levels of the hierarchy. A multi-hot encoded target vector $\mathbf{y} \in \{0,1\}^n$ is created that indicates only the annotated *Iconclass* concepts of an image, disregarding the corresponding parent nodes of those concepts according to the *Iconclass* taxonomy. A *sigmoid* function is used as activation in the classification layer to predict a probability vector $\hat{\mathbf{y}} \in \mathbb{R}^n$. The cross-entropy loss between the predicted and target vectors is used for optimization.

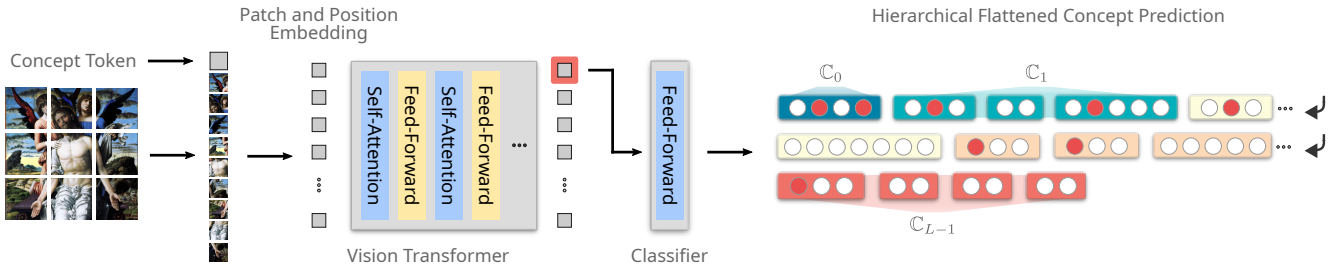

Figure 3. Workflow of the Hierarchical Flattened Classification (`Flat-H`). A vision transformer [17] is used to create a semantic representation of the images. The classification head is used as input for a fully-connected feed-forward layer with *sigmoid* activation that flattens the *Iconclass* taxonomy using as many neurons as iconographic concepts in the whole taxonomy. The colors in the flattened prediction represent all concepts $\mathbb{C}_l$ in a given taxonomy level l; the blocks within the colors have the same parental notation.

3.4.3 Hierarchical Flattened Classification (`Flat-H`)

This classifier extends `Flat` with a 'flattened,' multi-hot encoded target vector $\mathbf{y} \in \{0,1\}^n$ that encodes not only the annotated concepts of an image, but also their corresponding parents according to the *Iconclass* taxonomy. The structure of this approach is shown in Figure 3. For optimization, we follow the approach of *YOLO9000* [42] and compute the cross-entropy loss between the ground-truth vector $\mathbf{y} \in \{0,1\}^n$ and the predicted probabilities $\hat{\mathbf{y}} \in \mathbb{R}^n$ at each level of the taxonomy. For this purpose, we apply the *sigmoid* activation function only to concepts at taxonomy level $l \in [0, L-1]$ that are related (i.e., synsets) to the most likely parent class(es) at level $l-1$. Concepts with other parents are not considered in the loss term. This allows the classifier to learn from structured hierarchical information. It also alleviates the problem of class imbalance, as the number of concepts to be classified as negative classes is significantly reduced. During inference, the probability of a concept can be refined by considering the probabilities of its parents (e.g., by multiplication).

3.4.4 Weighted Flattened Classification (`Flat-W`)

To integrate ontology information, we use a weighting scheme similar to that presented by Müller-Budack et al. [37] for ontology-driven event classification. As for `Flat-H`, we first create a multi-hot encoded target vector $\mathbf{y} \in \{0,1\}^n$ that encodes the annotated concepts of an image and its parents in the *Iconclass* taxonomy. To put more emphasis on annotated iconographic concepts, we assign a weight of $w = 1$ to all concepts that have been labeled for at least one training image, while concepts that have no annotations are weighted with $w = 0.5$. Subsequently, the target vector and the predicted probabilities (using *sigmoid* activation), are multiplied by the corresponding weights. The cosine similarity between the weighted vectors is optimized during training. Unlike Müller-Budack et al. [37] we (i) consider a Hierarchical Multi-label Classification (HMC) task where samples can contain more than one annotation, and (ii) the fraction of nodes assigned with a maximum weight is much higher for our data set (19,829 / 23,113; see Section 4) compared to the *Visual Event Classification Data set* (148 / 409; [37]). As a result, the weighting of concepts may be less significant.

3.4.5 Hierarchical Cross-Attention Transformer (`CAT`)

To better represent the *Iconclass* taxonomy, we introduce a novel approach that does not compute all *Iconclass* concepts in one step, but in an iterative fashion. The general idea behind this approach is similar to the recurrent hierarchical classification approach *HMCN-R* [55] or image captioning [25], where in each iteration, a level of the hierarchy, or a token of the *Iconclass* concept C (Figure 2a), is predicted. In contrast to related work, we use an encoder-decoder structure based on recent transformer architectures. The entire architecture is shown in Figure 4. The vision encoder is the vision transformer, pre-trained according to Section 3.3. We use the original transformer decoder proposed by Vaswani et al. [54] that applies cross-attention to the vision encoder. Cross-attention allows all regions of the image, i.e., the heads of all regions that carry the embeddings, to be considered by the decoder in each iteration of token prediction. This is the main difference to the variants of the `Flat` classifier that use only the classification head.

For each level $l \in [0, L-1]$ in the taxonomy, we learn an individual class embedding representing all *Iconclass* concepts $|\mathbb{C}_l|$ in that level, and an additional *start* and *stop* token, for the initial input and to end the prediction. These L class embeddings are used as input to the decoder.

In the classification layer, we use L separate dense layers with *sigmoid* activation for each level of the hierarchy. Each dense layer consists of $|\mathbb{C}_l| + 1$ neurons outputting the probabilities $\hat{\mathbf{y}}_l \in \mathbb{R}^{|\mathbb{C}_l|+1}$ of the $|\mathbb{C}_l|$ concepts in the taxonomy level l and an additional neuron indicating the probability for the *stop token*. To handle multiple *Iconclass* concepts per image, we repeat this process m times to pre-

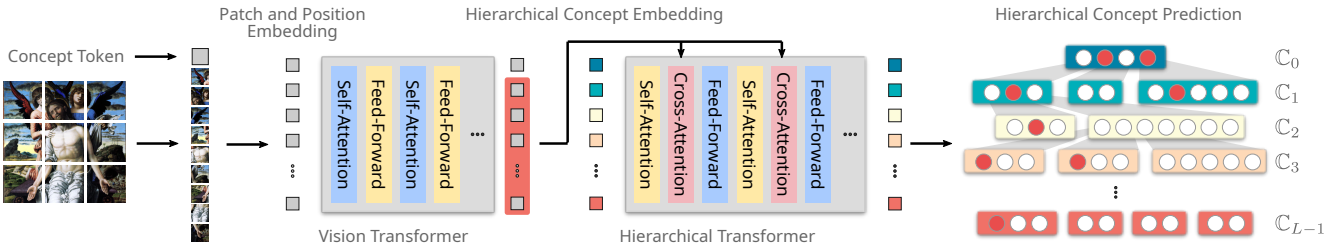

Figure 4. Workflow of the Hierarchical Cross-Attention Transformer (CAT) based on a vision transformer [17] as encoder and an hierarchical decoder extended from Vaswani et al. [54]. The hierarchical decoder applies cross-attention to include features from all image regions and learns individual class embedding for all *Iconclass* concepts $|\mathbb{C}_l|$ in a particular level $l \in [0, L-1]$ of the taxonomy. The CAT model predicts in each iteration the concepts of a level (illustrated with different colors) based on the input embedding from the previous level (parent *Iconclass* concept). Thus, in each step, only one of the blocks in the concepts $\mathbb{C}_l$ is predicted. The details for the optimization of the classifier are visualized in Figure 5.

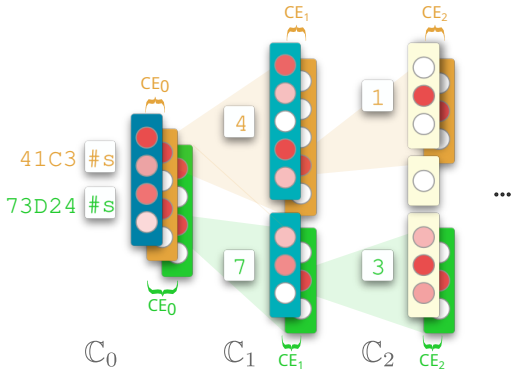

Figure 5. Optimization of the multi-label CAT classification. Using two notations, 41C3 and 73D24 from Figure 1, we apply the CAT model twice. First, the input of the transformer is the sequence #s (*start*), 4, 1, where the ground-truth annotation is highlighted in orange. Second, the cross-entropy CE_l loss between the respective ground-truth and the prediction is calculated for each level $l \in [0, L-1]$ considering only the valid parent.

dict a total of up to m notations. In each case, we use one of the *Iconclass* concepts as the input sequence to the transformer (teacher forcing) and then optimize the respective valid part of $\hat{\mathbf{y}}_l \in \mathbb{R}^{|\mathbb{C}_l|+1}$ with the matching ground-truth vector $\mathbf{y}_l \in \{0,1\}^{|\mathbb{C}_l|+1}$ using the cross-entropy loss CE_l:

$$\mathcal{L}_{\texttt{CAT}} = \sum_{l=0}^{L-1} CE_l\left(\mathbf{y}_l, \hat{\mathbf{y}}_l\right) = -\sum_{l=0}^{L-1}\sum_{c=0}^{\mathbb{C}_l} \mathbf{y}_{l,c} \log\left(\hat{\mathbf{y}}_{l,c}\right) \quad (1)$$

The optimization process is summarized in Figure 5.

To predict more than one path through the taxonomy during inference, we cannot use a greedy decoder or beam search procedure [56], as is common in image captioning, since this would result in only one prediction. Instead, we use a simple solution of repeatedly running the decoder with the current most likely concept as input, which still has child *Iconclass* concepts that have not been predicted yet. To avoid having to repeat the process for each classifier, we can define two termination criteria: (i) we can limit the maximum number of iterations p; (ii) we can stop the process if no concept has a probability above a certain threshold t.

4. Data Sets

Despite increasing efforts to digitize art-historical material, the amount of collections available online utilizing *Iconclass* remains disproportionately low. We are relying on two data sets that unite several institutions and entail a wide range of art-historical objects, such as paintings, emblems, drawings, engravings, and manuscripts: (i) *Iconclass AI Test Set* [39], in the following abbreviated to *ICAI*. The data set contains 87,744 images sampled from the Arkyves database.[3] These images are labeled with a total of 362,561 *Iconclass* concepts, 12,488 of which are unique. (ii) *ICARUS (Iconographic Classification and Representation Understanding)*. In addition, we introduce a novel data set that comprises 477,569 images, providing 1,328,417 annotations for 20,596 unique *Iconclass* concepts. To compile this data set, a total of 19 publicly available collections were harvested from a variety of countries; further details are given in the supplementary material. For machine learning purposes, we divided *ICARUS* into training, validation, and testing sets, with approximate split ratios of 80 %, 10 %, and 10 %, respectively. Details of image pre-processing and duplicate removal are explained in supplementary material, as is the unification of the annotations of *ICAI* and *ICARUS*.

5. Experimental Setup and Results

In this section, we present the network architecture and parameters (Section 5.1), the evaluation metrics (Section 5.2), the experimental results of the pre-training of the VLM models (Section 5.3), as well as the performance of the hierarchical classification approaches (Section 5.4).

[3] https://www.arkyves.org/ (last accessed on 2023-11-08).

5.1. Implementation Details

For *CLIP*, we use the vision transformer variant *ViT-B/16* [17] as vision encoder and a transformer model [54] with twelve layers and eight attention heads as text encoder. For the CAT classification model (Section 3.4.5), we use the transformer presented by Vaswani et al. [54] with three decoder layers and eight attention heads. During training, we use $m = 5$ to achieve a good trade-off between performance and speed. During inference of the CAT classifier, we use $p = 30$ iterations as stopping criterion. Unless otherwise specified, we optimize our models for 40,000 iterations using the *AdamW* optimizer [31] with a batch size of 256 and a learning rate of $1e-4$. More experiments and details on the hyperparameters as well as their selection are included in the supplementary material.

5.2. Evaluation Metrics

Since we consider a Hierarchical Multi-label Classification (HMC) problem for more than 20,000 concepts and significant class imbalance due to the hierarchical structure, typical classification metrics are not suitable. Therefore, we calculate the AP for each *Iconclass* concept and average it over all concepts that have at least a certain number of training images. Setting the number of images to larger thresholds provides insights into the model performance for lower levels (i.e., coarser iconographic concepts) of the taxonomy.

5.3. Contrastive Pre-training with Image-Text Pairs

In this experiment, we evaluate the efficacy of different strategies for creating image descriptions (Section 3.2) for contrastive pre-training of VLMs. For this purpose, we generate a data set using the proposed KW, BLIP, and GPT methods for text-synthesis and train *CLIP* [40] for 40,000 iterations on the *ICARUS* training set. We then fine-tune our CAT approach from Section 3.4 to classify the *Iconclass* taxonomy for another 40,000 iterations and compare the results. To investigate the performance of our pre-training on *ICARUS*, we also fine-tune the CAT classifier on the original *CLIP* model trained on the *LAION-400M* data set [45]. Therefore, no pre-training takes place in this experiment. The results of this experiment are shown in Table 1.

Models that were first pre-trained on one of the synthesized image-text pairs generally outperform the original *CLIP* model trained on *LAION-400M*. In particular, the results improve for concepts that have few training examples in the corpus. This proves that pre-tuning with art-historical images does indeed improve performance for iconographic concept classification. As expected, our novel, more sophisticated strategies for description synthesis, i.e., GPT and BLIP, outperform the KW baseline; the results for BLIP are slightly better than for GPT. Thus, we choose BLIP for all subsequent experiments.

Table 1. Results of contrastive pre-training with image-text pairs on different text generation strategies on the *ICARUS* test set using the CAT classifier. The results show the mean Average Precision (mAP) for all concepts that have at least one image in the test set. The best-performing strategy is denoted in bold.

Strategy	# of Training Images per *Iconclass* Concept			
	> 0	> 10	> 100	> 1000
KW	0.1862	0.2025	0.2545	0.3953
BLIP	**0.1922**	**0.2106**	**0.2596**	**0.3961**
GPT	0.1902	0.2063	0.2583	0.3916
LAION-400M	0.1845	0.2017	0.2540	0.3936

Table 2. Results of the individual classification approaches on the *ICAI* and *ICARUS* test sets using the BLIP text generation. The results show the mAP for all concepts that have at least one example in the test data set. The best-performing classifier per test set is denoted in bold.

Test Set	Classifier	# of Training Images per *Iconclass* Concept			
		> 0	> 10	> 100	> 1000
ICAI	CLIP	0.0033	0.0040	0.0088	0.0323
	Flat	0.0294	0.0378	0.0688	0.1639
	Flat-H	0.0638	0.0777	0.1148	0.2107
	Flat-W	0.0105	0.0134	0.0286	0.0970
	CAT	**0.1715**	**0.1803**	**0.2012**	**0.2737**
ICARUS	CLIP	0.0035	0.0038	0.0080	0.0245
	Flat	0.0394	0.0484	0.0942	0.2265
	Flat-H	0.0789	0.0946	0.1507	0.2935
	Flat-W	0.0134	0.0172	0.0382	0.1294
	CAT	**0.1714**	**0.1889**	**0.2407**	**0.3716**

5.4. Iconographic Concept Classification

In these experiments, we compare our proposed image classification approach CAT presented in Section 3.4.5 with several state-of-the-art baseline methods: FLAT is consistent with a common solution for multi-label classification [43], which uses a *sigmoid* activation together with a cross-entropy loss; FLAT-H is a widely applied technique [42] to exploit hierarchical information; and FLAT-W mimics a state-of-the-art approach [37] that uses ontology information. We have also considered the applicability of other state-of-the-art approaches to HMC. However, they are not applicable, either because they cannot handle multiple classes at the same level of the hierarchy [11], or because they cannot handle the large number of over 20,000 classes considered in our task [20]. As mentioned in Section 5.3, our proposed approaches use the image encoder of the VLM model pre-trained with BLIP descriptions optimized on the *ICARUS* training set. The results are shown in Table 2.

On both test data sets we can see that the CAT approach performs significantly better than all other baseline classification methods. Furthermore, the two best performing approaches, Flat-H and CAT, are those that use masking to optimize only the relevant parts of the *Iconclass* taxon-

(a) 11HH(THERESA) ("the foundress of the reformed (Discalced) Carmelites, T(h)eres(i)a of Avila [...]")

(b) 71B4 (" story of the Tower of Babel (Genesis 11:1–9)")

(c) 92D19217 ("Psyche performing various tasks set to her by Venus")

Figure 6. Results of the CAT model on the *ICARUS* test set. For visualization, we randomly selected three *Iconclass* concepts with $AP > 0.5$ and at least five images in the test set. The images are arranged in descending order of prediction probability. Green borders indicate correctly classified images; red bordered images do not include the selected concept in their ground-truth annotations.

omy. This can probably be explained by the fact that some of the images in the training set are not thoroughly labeled and thus also show unlabeled concepts, leading to a worse optimization for the other methods. We achieve promising results given the complexity of the task and the limited amount of training data for some concepts. A qualitative evaluation conducted with domain experts also showed that our approach can be usefully applied in practice due to its hierarchical architecture, since not only the prediction of the finest level of hierarchical is relevant, but also the prediction of concepts that are superordinate to this level. We see particular value in semi-automated use cases, where potentially relevant concepts can be automatically recommended for an image and then manually confirmed or refined by an expert. Figure 6 illustrates some of the art-historical concepts that were qualitatively analyzed: in addition to the narratives of primarily Christian religion illustrated in Figure 6a and Figure 6b, there are also those of classical mythology (Figure 6c). Visually striking iconographies, such as the story of the *Tower of Babel* (Figure 6b), are reliably classified by the CAT model with the corresponding *Iconclass* concepts, regardless of the painting or printing technique used; this is true for copper engravings as well as for illuminated manuscripts. As shown in Figure 6a, the model occasionally detects similarly predisposed compositions, even if they are false positives. Further information about the qualitative evaluation is given in the supplementary material.

6. Conclusions

In this paper, we have presented a novel approach for Hierarchical Multi-label Classification (HMC) of iconographic concepts. We have introduced three strategies for automatically creating image descriptions to pre-train a state-of-the-art Vision-Language Model (VLM) based on a novel data set comprising 477,569 images for more than 20,000 unique iconographic concepts. Furthermore, we proposed five classification approaches, including a novel transformer decoder that leverages hierarchical knowledge from the *Iconclass* taxonomy, which is the first decoder adopted to the problem of multi-label classification. We have demonstrated that our proposed solution benefits significantly from the adoption of *Iconclass's* structure: if a concept situated at a lower hierarchy level is not detected, the taxonomy allows for an upward traversal, facilitating the identification of a related concept. This decisively increases the potential usefulness of digital collections for research and education in the visual arts.

In the future, we aim to extend our approach to other, particularly non-western, taxonomies such as the Chinese Iconography Thesaurus (CIT), as well as other hierarchical multi-label classification tasks in Computer Vision (CV). It would also be interesting to explore how simultaneous optimization of a VLM and a classifier, which are currently trained separately in two stages, affects the performance. Pre-training LLMs on captions for art-historical documents for description generation is also worth investigating.

Acknowledgements

This work was partly funded by the German Research Foundation (DFG) under project numbers 415796915 and 510048106.

References

[1] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Computer Vision and Pattern Recognition, CVPR 2015*, pages 2927–2936, New York, 2015. IEEE. 2

[2] Seyed Ali Amirshahi, Gregor Uwe Hayn-Leichsenring, Joachim Denzler, and Christoph Redies. JenAesthetics subjective dataset. Analyzing paintings by subjective scores. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Workshop co-located with the European Conference on Computer Vision, ECCV 2014*, volume 8925 of *Lecture Notes in Computer Science*, pages 3–19, Cham, 2014. Springer. 2

[3] Nikolay Banar, Walter Daelemans, and Mike Kestemont. Multi-modal label retrieval for the visual arts. The case of iconclass. In Ana Paula Rocha, Luc Steels, and H. Jaap van den Herik, editors, *International Conference on Agents and Artificial Intelligence, ICAART 2021*, pages 622–629. SCITEPRESS, 2021. 2

[4] Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Neural Information Processing Systems, NeurIPS 2010*, pages 163–171. Curran Associates, Inc., 2010. 2

[5] Luca Bertinetto, Romain Müller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes. Leveraging class hierarchies with deep networks. In *IEEE/CVF Computer Vision and Pattern Recognition, CVPR 2020*, pages 12503–12512, New York, 2020. IEEE. 2

[6] Simone Bianco, Davide Mazzini, Paolo Napoletano, and Raimondo Schettini. Multitask painting categorization by deep multibranch neural network. *Expert Systems with Applications*, 135:90–101, 2019. 1, 2

[7] Hans Brandhorst and Etienne Posthumus. Iconclass. A key to collaboration in the digital humanities. In Colum Hourihane, editor, *The Routledge Companion to Medieval Iconography*, Routledge Art History and Visual Studies Companions, pages 201–218, Milton Park, 2017. Routledge. 1

[8] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. volume 472, 2017. 3

[9] Eva Cetinic. Towards generating and evaluating iconographic image captions of artworks. *Journal of Imaging*, 7(8):123, 2021. 2

[10] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "Flamingo" is my "Bird". Fine-grained, or not. In *IEEE Computer Vision and Pattern Recognition, CVPR 2021*, pages 11476–11485, New York, 2021. IEEE. 2

[11] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *IEEE Computer Vision and Pattern Recognition, CVPR 2022*, pages 4848–4857, New York, 2022. IEEE. 2, 7

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. arXiv:2210.11416, 2022. 3

[13] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at. Query based worldwide image geo-localization using hierarchies and scenes. arXiv:2303.04249, 2023. 2

[14] Marcos V. Conde and Kerem Turgutlu. CLIP-Art. Contrastive pre-training for fine-grained art classification. In *Workshop co-located with the Computer Vision and Pattern Recognition, CVPR 2021*, pages 3956–3960, New York, 2021. IEEE. 1, 2, 3

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP. Towards general-purpose vision-language models with instruction tuning. arXiv:2305.06500, 2023. 2, 3

[16] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *European Conference on Computer Vision, ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 71–84, Cham, 2010. Springer. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words. Transformers for image recognition at scale. In *International Conference on Learning Representations, ICLR 2021*, 2021. 4, 5, 6, 7

[18] 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 Adrien Bartoli and Andrea Fusiello, editors, *Workshop co-located with the European Conference on Computer Vision, ECCV 2020*, volume 12536 of *Lecture Notes in Computer Science*, pages 92–108, Cham, 2020. Springer. 2

[19] Vivien Sainte Fare Garnot and Loïc Landrieu. Leveraging class hierarchies with metric-guided prototype learning. In *British Machine Vision Conference 2021, BMVC 2021*, page 123. BMVA Press, 2021. 2

[20] Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Neural Information Processing Systems, NeurIPS 2020*, 2020. 2, 7

[21] Jahnvi Gupta, Prathmesh Madhu, Ronak Kosti, Peter Bell, Andreas K. Maier, and Vincent Christlein. Towards image caption generation for art historical data. In *AI Methods for*

Digital Heritage co-located with the German Conference on Artificial Intelligence, KI2020 2020, 2020. 2

[22] Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. Hierarchical multi-label text classification. An attention-based recurrent network approach. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *International Conference on Information and Knowledge Management, CIKM 2019*, pages 1051–1060. ACM, 2019. 2

[23] Leonardo Impett and Sabine Süsstrunk. Pose and pathosformel in Aby Warburg's bilderatlas. In Gang Hua and Hervé Jégou, editors, *Workshop co-located with the European Conference on Computer Vision, ECCV 2016*, volume 9913 of *Lecture Notes in Computer Science*, pages 888–902, Cham, 2016. Springer. 2

[24] Tomás Jenícek and Ondrej Chum. Linking art through human poses. In *International Conference on Document Analysis and Recognition, ICDAR 2019*, pages 1338–1345, New York, 2019. IEEE. 2

[25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition, CVPR 2015*, pages 3128–3137, New York, 2015. IEEE. 5

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2. Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597, 2023. 2, 3

[27] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. CLIP-event. Connecting text and images with event structures. In *Computer Vision and Pattern Recognition, CVPR 2022*, pages 16399–16408, New York, 2022. IEEE. 3

[28] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. arXiv:2212.00794, 2022. 3

[29] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The ArtBench dataset. Benchmarking generative models with artworks. arXiv:2206.11404, 2022. 1, 2

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO. Common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision, ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, Cham, 2014. Springer. 3

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR 2019*, 2019. 7

[32] Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas K. Maier, and Vincent Christlein. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. *Journal on Computing and Cultural Heritage*, 16(1):1–17, 2023. 2

[33] Hui Mao, Ming Cheung, and James She. DeepArt. Learning joint representations of visual arts. In Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan, editors, *International Conference on Multimedia, MM 2017*, pages 1183–1191, New York, 2017. ACM. 1, 2

[34] Thomas Mensink and Jan C. van Gemert. The Rijksmuseum challenge. Museum-centered visual recognition. In Mohan S. Kankanhalli, Stefan M. Rüger, R. Manmatha, Joemon M. Jose, and Keith van Rijsbergen, editors, *International Conference on Multimedia Retrieval, ICMR 2014*, pages 451–454, New York, 2014. ACM. 1, 2

[35] Federico Milani and Piero Fraternali. A dataset and a convolutional model for iconography classification in paintings. *Journal on Computing and Cultural Heritage*, 14(4), 2021. 2

[36] Saif M. Mohammad and Svetlana Kiritchenko. WikiArt emotions. An annotated dataset of emotions evoked by art. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *International Conference on Language Resources and Evaluation, LREC 2018*, pages 1225–1238, Miyazaki, 2018. ELRA. 2

[37] Eric Müller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, and Ralph Ewerth. Ontology-driven event type classification in images. In *Winter Conference on Applications of Computer Vision, WACV 2021*, pages 2927–2937, New York, 2021. IEEE. 2, 5, 7

[38] Erwin Panofsky. *Studies in Iconology. Humanistic Themes in the Art of the Renaissance*. Oxford University Press, New York, 1939. 1

[39] Etienne Posthumus. Iconclass AI test set. https://iconclass.org/testset/. Last accessed on 2023-11-08. 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 3, 4, 7

[41] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 3

[42] Joseph Redmon and Ali Farhadi. YOLO9000. Better, faster, stronger. In *IEEE Computer Vision and Pattern Recognition, CVPR 2017*, pages 6517–6525, New York, 2017. IEEE. 2, 5, 7

[43] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *International Conference on Computer Vision, ICCV 2021*, pages 82–91, 2021. 7

[44] Stefanie Schneider, Matthias Springstein, Javad Rahnama, Eyke Hüllermeier, Ralph Ewerth, and Hubertus Kohle. The

dissimilar in the similar. An attribute-guided approach to the subject-specific classification of art-historical objects. In Ralf H. Reussner, Anne Koziolek, and Robert Heinrich, editors, *Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020*, volume P-307 of *LNI*, pages 1355–1364, Bonn, 2020. GI. 2

[45] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M. Open dataset of CLIP-filtered 400 million image-text pairs. arXiv:2111.02114, 2021. 3, 7

[46] Xi Shen, Alexei A. Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In *Computer Vision and Pattern Recognition, CVPR 2019*, pages 9278–9287, New York, 2019. IEEE. 2

[47] Matthias Springstein, Stefanie Schneider, Christian Althaus, and Ralph Ewerth. Semi-supervised human pose estimation in art-historical images. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *International Conference on Multimedia, MM 2022*, pages 1107–1116. ACM, 2022. 2

[48] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. Artpedia. A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing, ICIAP 2019*, volume 11752 of *Lecture Notes in Computer Science*, pages 729–740, Cham, 2019. Springer. 1, 2

[49] Gjorgji Strezoski and Marcel Worring. OmniArt. A large-scale artistic benchmark. *Transactions on Multimedia Computing, Communications, and Applications*, 14(4):88:1–88:21, 2018. 1, 2

[50] Salma Taoufiq, Balázs Nagy, and Csaba Benedek. HierarchyNet. Hierarchical CNN-based urban building classification. *Remote Sensing*, 12(22):3794, 2020. 2

[51] Henri van de Waal. *Iconclass. An Iconographic Classification System. Completed and Edited by L. D. Couprie with R. H. Fuchs*. North-Holland Publishing Company, Amsterdam, 1973–1985. 1, 3

[52] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018. 4

[53] Roelof van Straten. *Iconography, Indexing, Iconclass. A Handbook*. Foleor, Leiden, 1994. 1, 3

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Neural Information Processing Systems, NeurIPS 2017*, pages 5998–6008, 2017. 5, 6, 7

[55] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo C. Barros. Hierarchical multi-label classification networks. In Jennifer G. Dy and Andreas Krause, editors, *International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5225–5234. PMLR, 2018. 5

[56] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system. Bridging the gap between human and machine translation. arXiv:1609.08144, 2016. 6

[57] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. HD-CNN. Hierarchical deep convolutional neural networks for large scale visual recognition. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 2740–2748, New York, 2015. IEEE. 2

[58] Victoria Yanulevskaya, Jasper R. R. Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe. In the eye of the beholder. Employing statistical analysis and eye tracking for analyzing abstract paintings. In Noboru Babaguchi, Kiyoharu Aizawa, John R. Smith, Shin'ichi Satoh, Thomas Plagemann, Xian-Sheng Hua, and Rong Yan, editors, *International Conference on Multimedia, MM 2012*, pages 349–358, New York, 2012. ACM. 2

[59] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne van Noord, and Giorgos Tolias. The Met dataset. Instance-level recognition for artworks. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Neural Information Processing Systems, NeurIPS 2021*, 2021. 1, 2

[60] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels. A hierarchical multi-label contrastive learning framework. In *IEEE Computer Vision and Pattern Recognition, CVPR 2022*, pages 16639–16648, New York, 2022. IEEE. 2

[61] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Large-scale category structure aware image categorization. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Neural Information Processing Systems, NeurIPS 2011*, pages 1251–1259, 2011. 2

[62] Xinqi Zhu and Michael Bain. B-CNN. Branch convolutional neural network for hierarchical classification. arXiv:1709.09890, 2017. 2