The Dissimilar in the Similar. An Attribute-guided Approach to the Subject-specific Classification of Art-historical Objects

Stefanie Schneider[1], Matthias Springstein[2], Javad Rahnama[3], Eyke Hüllermeier[3], Ralph Ewerth[2,4], Hubertus Kohle[1]

Abstract: Due to the increasingly unmanageable number of art-historical inventories made available in digital form, methods that computationally arrange larger amounts of objects are becoming more important. The category of similarity, which is fundamental in all areas of art-historical description, gains new relevance in this context. In this paper, we propose a novel approach to the subject-specific classification of art-historical objects that utilizes expert-based attributes, i.e., significant figurative motifs. We evaluate our procedure on a concrete use case, representations of saints in the visual arts. A representative data set of saints images is collected and a semi-supervised learning technique applied to enrich the data set with neural style transfer as well as to improve the joint training of saints and their attributes. We show that this technique outperforms other approaches.

Keywords: Semi-supervised Learning; Semi-supervised Image Classification; Art Analysis; Digital Humanities

1 Introduction

The category of similarity is fundamental in all areas of art-historical description: in the history of style, the specification of formal characteristics determines the assignment of artistic phenomena to stylistic attitudes; in iconography, definitions of content are constituted by the observation of comparable—or similar—conventions of representation. Similarity also plays a central role in art-historical practice. When Wölfflin compares a portrait of Albrecht Dürer with one of Frans Hals–inter alia, in the form of the categories of the "linear" and the "painterly"—, he is assuming that the two works were painted in different ways while belonging to the same genre [Wö15]. Decisive for the persuasiveness of this procedure is the determination of the 'dissimilar in the similar': for only (or especially) where a common set of phenomena exists do possible differences become visible and plausible.

Because of the increasingly unmanageable number of art-historical inventories made available in digital form [MG14], two questions arise. Firstly, how can the manifold concepts

[1] Ludwig-Maximilians-Universität München, Institut für Kunstgeschichte, Zentnerstr. 31, 80798 München, {stefanie.schneider@itg.uni-muenchen.de, hubertus.kohle@lmu.de}.

[2] Technische Informationsbibliothek (TIB), Welfengarten 1b, 30167 Hannover, {matthias.springstein@tib.eu, ralph.ewerth@tib.eu}.

[3] Universität Paderborn, Fachgruppe Intelligente Systeme und Maschinelles Lernen, Pohlweg 51, 33098 Paderborn, {javad.rahnama@uni-paderborn.de, eyke@upb.de}.

[4] Forschungszentrum L3S, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover.

doi:10.18420/inf2020_127

of similarity be considered to relate larger amounts of objects computationally? Secondly, are existing methods suitable for such heterogeneous inventories and, if so, to what extent can they be adopted and optimized? Previous studies on the automatic detection, recognition, or identification of objects relevant to image science focus either on small visually distinctive sub-fields, e.g., ballad prints [TBO14] and tinted drawings [Ya11], or larger non-specialised data sets, e.g., WikiArt [HWS16], that predominantly feature well-known Western artists and art periods. They thus do only partially account for the great diversity of historical artefacts and lack the generalizability necessary for this domain.

In this work, we concentrate on the broader category of iconographic similarity and propose a generic approach to the subject-specific classification of art-historical objects that utilizes expert-based attributes of the classification system Iconclass, i.e., figurative motifs significant from an art-historical point-of-view. This is the first attempt to actively exploit Iconclass in automatic classification tasks, to the best of our knowledge. We evaluate our procedure on a concrete use case, representations of saints in the visual arts. This example is advantageous because it is usually possible to clearly assign the saint and the attributes identifying him or her: the attributes are placed in a spatially comprehensible relationship to the person, i.e., they are positioned close to it, even if sometimes hidden. The latter is especially true for phases of art history in which, as in 16th-century Mannerism, the clear legibility of a picture's content was not the main focus. Since many art-historical narratives, especially those of Christian religion and classical mythology, feature sufficiently informative attributes (or attribute-like concepts), this approach is widely applicable.

The contributions are as follows: *(i)* collection of a representative data set of saints, *(ii)* a novel approach to attribute-guided classification that utilizes Iconclass, and *(iii)* application of a semi-supervised learning technique to enrich the data set with neural style transfer as well as to improve the joint training of saints and their attributes.

2 Related Work

Due to the recent growth in computerized analysis of cultural heritage, we primarily discuss studies that address the categorization of art-historical objects.

To classify art periods such as Baroque and Symbolism, Hentschel et al. [HWS16] contrast Fisher Vectors and a Support Vector Machine with a Convolutional Neural Network (CNN) pre-trained on ImageNet and fine-tuned on WikiArt. Anwer et al. [An16] extend on this methodology by also utilizing information about local regions of interest with a Deformable Part Model. In earlier and less relevant works, Gatys et al. [GEB15] train a CNN to capture, separate, and reconstruct the content of an object, and its style, whereas Saleh and Elgammal [SE15] combine low-level and high-level features to categorize style, genre, and artist. A CNN is trained on top of the last layer of an ImageNet-trained network to capture additional semantic features. More recently, Bianco et al. [Bi19] propose a multitask-multibranch CNN to simultaneously classify style, genre, and artist. In contrast, Yang et al. [Ya18] encode

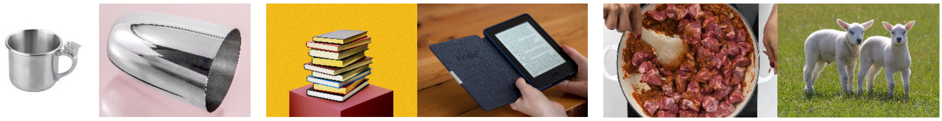

Fig. 1: Images of the attributes “baptismal cup”, “book”, and “lamb”, retrieved from *Google Image Search* respectively.

complementary material to assist visual feature learning in CNNs for style classification. Sabatelli et al. [Sa18] investigate the general effect of fine-tuned CNNs in artist, material, and type classification tasks.

However, studies rarely incorporate concepts significant to iconography. In one of the few exceptions, Gonthier et al. [Go18] propose a multiple instance learning (MIL) technique for the weakly-supervised detection of art-historically specific objects. However, as image-level annotations are only gathered for 7 classes, the generalizability of the approach remains unclear, especially for concepts with high in-class variability. In this work, we focus entirely on a unified set of art-historically relevant classes that are of a comparably high visual and narrative complexity: representations of saints in the visual arts. Like Yang et al. [Ya18], we utilize historical context information—here expert-based attributes that are linked to the respective class, i.e., the respective saint—to improve the subject-specific classification of concepts with high in-class variability.

3 Data

Data set collection Our data set consists of two kinds of images: art-historical and non-art-historical, i.e., real-world imagery.

A total of 19 publicly available inventories, collections, institutions, and web portals are first harvested to gather depictions of saints in the visual arts.[5] The obtained reproductions are extremely varied and, e.g., include stained glass paintings of the Middle Ages, 16th-century emblems as well as Polish folk woodcuts. Each source is at least partially indexed by experts with the decimal classification system Iconclass that was specially conceived for the Western motifs of the visual arts [Wa85]. It thus also contains definitions of male and female saints, where each saint is provided with an explanatory textual correlate including a list of possible attributes.[6] This information is used to retrieve real images of the attributes from *Google*

[5] artemis.uni-muenchen.de, https://www.bildindex.de/, http://ballads.bodleian.ox.ac.uk/, https://corpusvitrearum.de/, emblematica.library.illinois.edu, heartfield.adk.de, https://inkunabeln.digitale-sammlungen.de/, http://manuscripts.kb.nl/, http://www.museen.thueringen.de/, https://www.nga.gov/, https://datenbank.museum-kassel.de/, https://sammlung.belvedere.at/, http://pauart.pl/app, https://realonline.imareal.sbg.ac.at/, https://www.rijksmuseum.nl/en, https://rkd.nl/en/, sammlung.staedelmuseum.de, http://www.virtuelles-kupferstichkabinett.de/de/, and https://vitrosearch.ch/de, respectively (all accessed April 28, 2020).

[6] All other notations are accompanied by a list of keywords, some of which can be defined as attributes, or at least have attribute-like properties.

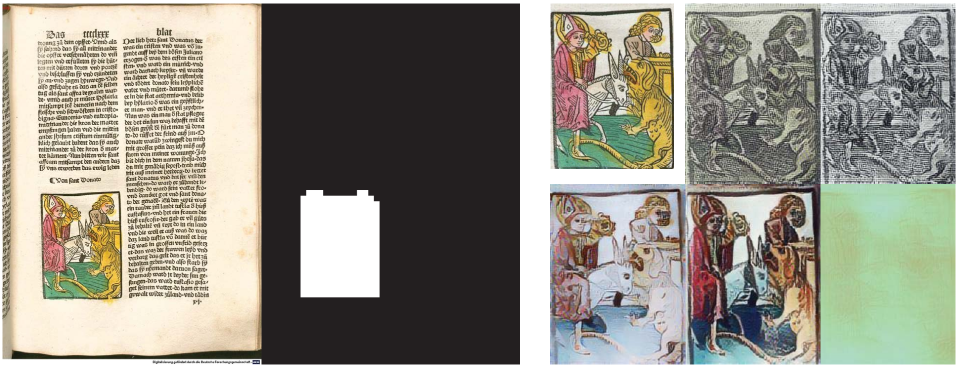

Fig. 2: Detection of bounding boxes (left) and application of style transfer to enrich the data set (right).

Image Search, i.e., photographs taken in recent years. As shown in Figure 1, not all images include the desired attribute in the narrow sense; e.g., a modern e-reader was found as well as lamb meat. In so doing, we collect 21,479 images of 239 saints and 124,133 images of 343 attributes for training and testing our procedures.

Data set preprocessing Many of the previously harvested representations are scans and contain background noise or further information, e.g., signatures of the artist or linear color control charts of the institution responsible for the reproduction. Two preprocessing steps are necessary to use these representations for training a neural network. Relevant image content is first detected using a DeepLabv3 image segmentation model trained on 100 examples from the afore-introduced saints data set [Ch17]. The overlapping image regions thus identified are then integrated into one rectangular region. If a region has a width or height of less than 100 pixels, it is discarded. An example prediction of the trained DeepLabv3 model is shown on the left in Figure 2. As the images of the saints and the images of the attributes originate from highly different domains, we deploy neural style transfer to enrich the data set and bridge the gap between domains [Gh17].[7] Up to 5 variations of the original image are created, where we choose a random image of a saint as a style image.

As depicted in Figure 2, not all images that are generated in this way are recognizable. On the one hand, this is due to the fact that style images are randomly selected and applied from all available saints images. On the other hand, the segmentation introduces errors; therefore, images are selected for style transfer that do not show any saint. On the basis of these steps, the number of images containing (representations of) saints increases to 25,667; the subsequent style transfer further increases the number to 120,626. The number of images depicting attributes increases to 403,788.

[7] Geirhos et al. [Ge19] also show that such techniques increase the robustness of neural versus textural change.

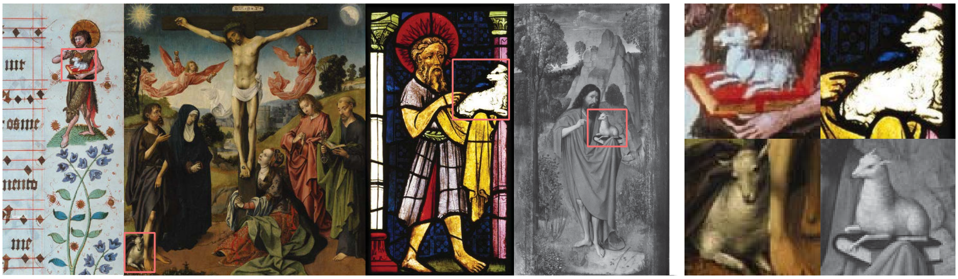

Fig. 3: Four representations of Saint John the Baptist with the exemplary selected attribute "lamb".

4 Attribute-guided Classification

The idea behind our approach is as follows: generally, a saint cannot be identified exclusively by his or her physiognomy, but by a set of pictorial signs, *attributes*, that exemplify a special event in his life or take up characteristics of her status or profession. A distinction must be made between attributes characterizing a (larger) group of saints and attributes that are narratively significant for a particular saint. While, e.g., the staff serves as a general sign of holy abbots, John the Baptist is often accompanied by a lamb to recall the acclamation in which he refers to Christ as the "Lamb of God" (Figure 3). Since most attributes act as binding signifiers, they are often featured prominently in the fore- or background of an image and can thus support the computer-aided classification of saints. We assume that the joint appearance of even relatively trivial appearing or art-historically unspecific attributes, whose artistic depiction has hardly changed over time, is sufficient for this purpose.[8]

Two problems arise. On the one hand, a saint can be identified by more than one attribute; however, not *all* attributes need to be present in the image of a saint. On the other hand, the images found via *Google Image Search* do not always show the desired attribute, or solely modernized versions of it, as already illustrated in Section 3. We thus propose a semi-supervised learning technique based on FixMatch [So20]. The original objective of FixMatch is to use unlabeled data for training an image classifier. In doing so, unlabeled images for which the model predicts a high probability are automatically assigned to a concept and used for the training process. In our case, we use this technique to automatically annotate attributes in images of saints that were *not* originally annotated.

The training process for a batch is shown in Figure 4. During each iteration, the model forwards two batches of labeled images, $B_{l,s}$ for saints and $B_{l,a}$ for attributes, as well as two batches of unlabeled images, $B_{u,s}$ for saints and $B_{u,a}$ for attributes. It then determines the probability for the concept saints, p_s, and for the concept attributes, p_a, independently for each input image of the batch. The supervised loss L_l—applied to $B_{l,s}$ and $B_{l,a}$,

[8] This is in stark contrast to Gonthier et al., who state that "more specific objects or attributes such as ruins or nudity" are needed to detect [Go18, p. 2].

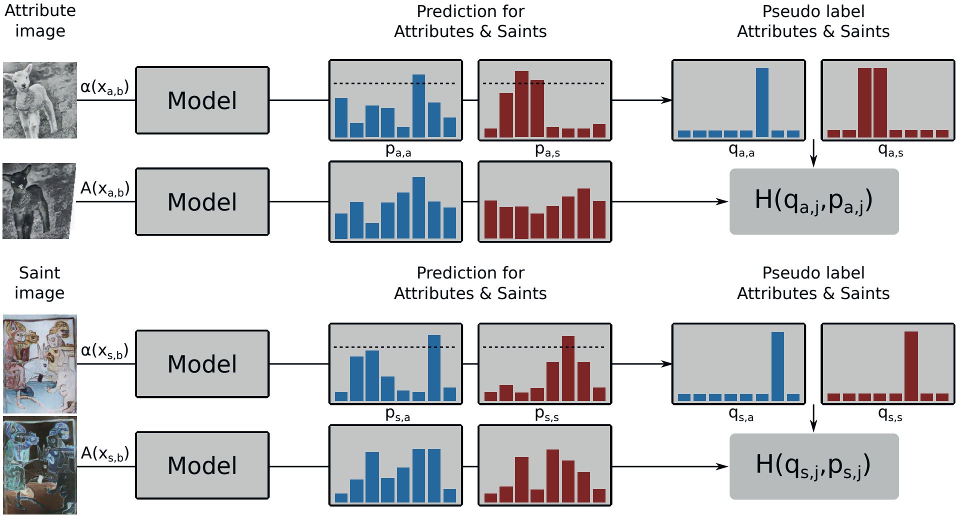

Fig. 4: Visualization of the semi-supervised learning technique. During each iteration, the system predicts a probability distribution for attributes (*blue*) and saints (*red*) that is used to generate pseudo-labels. These labels are then used as optimization targets for the same image with a different augmentation strategy.

respectively—results from the cross-entropy $H(\cdot)$ between the encoded label $\hat{y}_i$ and the prediction $p_{i,i}$ for an input $x_{i,b}$:

$$L_l = \sum_{i \in \{s,a\}} \frac{1}{B_{l,i}} \sum_{b=1}^{B_{l,i}} H\left(\hat{y}_{i,b}, p_{i,i}\left(y \mid \alpha\left(x_{i,b}\right)\right)\right) \tag{1}$$

For unlabeled data, we first compute the model's predicted class distribution for a *weakly*-augmented α-version of the sample $x_{i,b}$ in each subset $B_{u,s}$ and $B_{u,a}$. To create an artificial label, we assign a value of one for each prediction of a concept that is greater than a threshold τ; all other concepts are set to zero:

$$\hat{q}_{i,j,b} = \begin{cases} 1 & \text{if } p_{i,j}\left(y \mid \alpha\left(x_{i,b}\right)\right) \geq \tau \\ 0 & \text{if } p_{i,j}\left(y \mid \alpha\left(x_{i,b}\right)\right) < \tau \end{cases} \tag{2}$$

The unsupervised loss L_u results from the *strongly*-augmented version A of the image $x_{i,b}$ and the pseudo-label $\hat{q}_{i,j,b}$, as long as there is at least one prediction above the threshold τ:

$$L_u = \sum_{j \in \{s,a\}} \sum_{i \in \{s,a\}} \frac{1}{B_{u,i}} \sum_{b=1}^{B_{u,i}} \mathbb{1}\left(\max\left(p_{i,j}\left(y \mid \alpha\left(x_{i,b}\right)\right)\right) \geq \tau\right) H\left(\hat{q}_{i,j,b}, p_{i,j}\left(y \mid A\left(x_{i,b}\right)\right)\right) \tag{3}$$

The final loss L is simply the sum $L = L_l + L_u$. Since all images that contain neither a saint nor an attribute have a low probability of showing any relevant concept, they are

Attribute	AP	Attribute	AP	Attribute	AP	Attribute	AP
peacock feather	0.871	tablet	0.859	ducal hat	0.030	cope	0.016
scissors	0.870	hackle	0.845	net	0.026	stake	0.015
monstrance	0.867	tiara	0.844	Spes	0.026	Turk	0.014
staircase	0.866	broom	0.840	head	0.019	three	0.011
clog	0.865	wreath	0.837	mitre	0.017	two	0.007

Tab. 1: Best and worst classification results based on the data set with 343 attributes retrieved from *Google Image Search*. Average Precision (AP) is used to measure the retrieval performance.

automatically excluded during training. This procedure offers two advantages. When an attribute is recognized in the image of a saint, it is automatically annotated; in this way, there is feedback from attributes in images that were *not* originally annotated. Second, images that are not recognizable by the model after style transfer are excluded from training.

5 Experiments

We employ a ResNet-50 architecture pre-trained on ImageNet [He16]. The optimization is carried out using Stochastic gradient descent (SGD) with Nesterov momentum of 0.9 [Su13]. The initial learning rate is set to 0.01. The data set is split into training, validation, and test with a splitting ratio of 3:1:1. We evaluate the model with the highest accuracy on the validation set on the test set, respectively. Mean Average Precision (mAP) is used to measure the retrieval performance of our system for the entire test set.

Attribute classification We first evaluate whether the attributes data set is generally suitable for the prediction of saints. The model achieves a performance of 0.354 mAP. As shown in Table 1, attributes that are difficult to define ("three") or cannot be found by *Google Image Search* ("mitre") lead to poor classification performance, whereas objects still common in modern everyday life ("scissors") naturally show more promising results.

Joint training of saints and attributes Our approach to jointly train saints and attributes is compared to two baseline strategies, with and without style transfer, respectively. Thus, both saints classifiers do not use explicitly defined visual attributes during training. Random horizontal flip is used as augmentation step. In addition, we use *RandAugment* for the FixMatch approach, which applies a random transformation with a defined strength from a fixed set [Cu19]. We moreover use style-transferred images from the saints and attributes data set, respectively, as unlabeled input for FixMatch. The performance of the procedure is reported for the 49 saints with the most images, and only for images after the bounding box detection (see Section 3). As shown in Table 2, the proposed system performs best, mAP = 0.136, when a threshold of $\tau = 0.5$ is chosen. If the threshold is set too high, not

Method	$B_{l,s}$	$B_{u,s}$	$B_{l,a}$	$B_{u,a}$	mAP	Accuracy
Random					0.021	0.054
Saints (without Style Transfer)	16	0	0	0	0.131	0.250
Saints (with Style Transfer)	16	0	0	0	0.118	0.246
Saints and Attributes (without Style Transfer)	8	0	8	0	0.120	0.241
Saints and Attributes (with Style Transfer)	8	0	8	0	0.128	0.252
FixMatch (τ = 0.4)	8	8	8	8	0.093	0.210
FixMatch (τ = 0.5)	8	8	8	8	**0.136**	**0.260**
FixMatch (τ = 0.6)	8	8	8	8	0.134	0.245

Tab. 2: Scores of the classification methods based on the data set with 49 saints. $B_{l,s}$ and $B_{l,a}$ denote the batch sizes of labeled images for saints and attributes, respectively, $B_{u,s}$ and $B_{u,a}$ the batch sizes of unlabeled images for saints and attributes, respectively. The best performing approach is bold.

enough images are selected for training or not all concepts in an image are selected. If the threshold is set too low, however, too many concepts are selected. We chose 0.5 as a starting point because it is commonly used to generate binary decisions after a sigmoid activation.

A closer look at the results shows that saints are more accurately classified if their depictions are limited to few narratives, or a certain stage of life is primarily illustrated, e.g., in the case of Jerome (AP = 0.432), even if differing materials or techniques are used. If, on the other hand, a saint can be represented in many strongly varying ways that are not related to any specific constellation of attributes, such as Bernard (AP = 0.036), classification results drop immensely. This is especially true for saints, like Helena (AP = 0.020), for whom there are few examples or many visually distinctive ones, e.g., engravings, stained glass paintings, or early preparatory drawings. These findings illustrate that the enormous complexity of the domain, in which an object can be depicted in various ways, is often only insufficiently manageable—even with common augmentation techniques and fine-tuned networks. The underlying phenomenon, referred to as the "cross-depiction problem" [WCH16, p. 1], might possibly be weakened by more sophisticated domain adaptation techniques [TK18]. Moreover, to mitigate the dependency on non-art-historical imagery and further improve classification, the harvested collections could be exploited more extensively, since many attributes are listed in Iconclass as separate notations.

6 Conclusion

In this work, we introduced a new data set and task for the identification of saints in the visual arts. We suggested a novel deep-learning approach that utilizes expert-based attributes to support the subject-specific classification especially of concepts with high in-class variability. The proposed semi-supervised joint training technique increases the performance compared to multiple baselines. In the future, we will apply this procedure to the classification of other art-historically relevant narratives and motifs that can possibly also be improved by the use of visual attributes. To further improve the discrimination of saints (or other individuals relevant to art history), we plan to explore different loss functions, e.g., contrastive or triplet loss, as they are successfully used in face recognition tasks.

Acknowledgements

This work is financially supported by the German Research Foundation (DFG) under project number 415796915.

References

[An16] Anwer, R. M.; Khan, F. S.; van de Weijer, J.; Laaksonen, J.: Combining Holistic and Part-based Deep Representations for Computational Painting Categorization. In: Proceedings of the 2016 ACM International Conference on Multimedia Retrieval. Pp. 339–342, 2016.

[Bi19] Bianco, S.; Mazzini, D.; Napoletano, P.; Schettini, R.: Multitask Painting Categorization by Deep Multibranch Neural Network. In: Expert Systems with Applications. Vol. 135, pp. 90–101, 2019.

[Ch17] Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation, 2017, arXiv: 1706.05587, URL: http://arxiv.org/abs/1706.05587.

[Cu19] Cubuk, E. D.; Zoph, B.; Shlens, J.; Le, Q. V.: RandAugment. Practical Data Augmentation with No Separate Search, 2019, arXiv: 1909 . 13719, URL: http://arxiv.org/abs/1909.13719.

[Ge19] Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; Brendel, W.: ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. In: 7th International Conference on Learning Representations. 2019.

[GEB15] Gatys, L. A.; Ecker, A. S.; Bethge, M.: A Neural Algorithm of Artistic Style, 2015, arXiv: 1508.06576, URL: https://arxiv.org/abs/1508.06576.

[Gh17] Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; Shlens, J.: Exploring the Structure of a Real-time, Arbitrary Neural Artistic Stylization Network. In: British Machine Vision Conference. 2017.

[Go18] Gonthier, N.; Gousseau, Y.; Ladjal, S.; Bonfait, O.: Weakly Supervised Object Detection in Artworks, 2018, arXiv: 1810.02569, URL: https://arxiv.org/abs/1810.02569.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Identity Mappings in Deep Residual Networks. In: Computer Vision – ECCV 2016. Vol. 9908, Springer, pp. 630–645, 2016.

[HWS16] Hentschel, C.; Wiradarma, T. P.; Sack, H.: An Approach to Large Scale Interactive Retrieval of Cultural Heritage. In: Proceedings of the 23th IEEE International Conference on Image Processing. Pp. 3693–3697, 2016.

[MG14] Mensink, T.; van Gemert, J.: The Rijksmuseum Challenge. Museum-centered Visual Recognition. In: Proceedings of the International Conference on Multimedia Retrieval. Pp. 451–454, 2014.

[Sa18] Sabatelli, M.; Kestemont, M.; Daelemans, W.; Geurts, P.: Deep Transfer Learning for Art Classification Problems. In: Computer Vision – ECCV 2018 Workshops. Vol. 48, Springer, 2018.

[SE15] Saleh, B.; Elgammal, A. M.: Large-scale Classification of Fine-art Paintings. Learning the Right Metric on the Right Feature, 2015, arXiv: `1505.00855`, URL: `https://arxiv.org/abs/1505.00855`.

[So20] Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; Raffel, C.: FixMatch. Simplifying Semi-supervised Learning with Consistency and Confidence, 2020, arXiv: `2001.07685`, URL: `https://arxiv.org/abs/2001.07685`.

[Su13] Sutskever, I.; Martens, J.; Dahl, G. E.; Hinton, G. E.: On the Importance of Initialization and Momentum in Deep Learning. In: Proceedings of the 30th International Conference on Machine Learning. Pp. 1139–1147, 2013.

[TBO14] Takami, M.; Bell, P.; Ommer, B.: An Approach to Large Scale Interactive Retrieval of Cultural Heritage. In: Eurographics Workshop on Graphics and Cultural Heritage. Pp. 87–95, 2014.

[TK18] Thomas, C.; Kovashka, A.: Artistic Object Recognition by Unsupervised Style Adaptation. In: Computer Vision – ACCV 2018. Vol. 29, Springer, 2018.

[Wa85] van de Waal, H.: Iconclass. An Iconographic Classification System. Completed and Edited by L. D. Couprie with R. H. Fuchs. North-Holland Publishing Company, Amsterdam, 1973–1985.

[WCH16] Westlake, N.; Cai, H.; Hall, P.: Detecting People in Artwork with CNNs. In: Computer Vision – ECCV 2016 Workshops. Vol. 9913, Springer, 2016.

[Wö15] Wölfflin, H.: Kunstgeschichtliche Grundbegriffe. Bruckmann, Munich, 1915.

[Ya11] Yarlagadda, P.; Monroy, A.; Carque, B.; Ommer, B.: Recognition and Analysis of Objects in Medieval Images. In: Proceedings of the ACCV Workshop on Computer Vision. Pp. 296–305, 2011.

[Ya18] Yang, J.; Chen, L.; Zhang, L.; Sun, X.; She, D.; Lu, S.; Cheng, M.-M.: Historical Context-based Style Classification of Painting Images via Label Distribution Learning. In: Proceedings of the 26th ACM International Conference on Multimedia. Pp. 1154–1162, 2018.