National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Instance-level recognition for artworks

Diploma Thesis

of

NIKOLAOS-ANTONIOS YPSILANTIS

Supervisor: Stefanos Kollias
Professor, NTUA

Athens, February 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Instance-level recognition for artworks

Diploma Thesis

of

NIKOLAOS-ANTONIOS YPSILANTIS

Supervisor: Stefanos Kollias
Professor, NTUA

Approved by the examination committee on 15th of February 2022.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Stefanos Kollias	Georgios Tolias	Andreas-Georgios Stafylopatis
Professor, NTUA	Assistant Professor, CTU in Prague	Professor, NTUA

Athens, February 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

(Signature)

.....
Nikolaos-Antonios
Ypsilantis

Electrical & Computer
Engineer, NTUA

15th of February 2022

Περίληψη

Αυτή η διπλωματική εργασία παρουσιάζει την δημιουργία ενός νέου συνόλου δεδομένων μεγάλης κλίμακας (large-scale dataset) για το πρόβλημα της αναγνώρισης σε επίπεδο οντότητας (Instance-Level Recognition, ILR), στο πεδίο των έργων τέχνης (artwork domain). Το σύνολο δεδομένων αυτό, που ονομάζεται σύνολο δεδομένων Met, εμφανίζει μια σειρά από προκλήσεις όπως μεγάλη ομοιότητα μεταξύ των κλάσεων του (large inter-class similarity), κατανομή μακριάς ουράς (long-tail distribution) και πολύ μεγάλο αριθμό κλάσεων. Βασιστήκαμε στη συλλογή ανοιχτής πρόσβασης του μουσείου The Met για να σχηματίσουμε ένα μεγάλο σύνολο εκπαίδευσης (training set) περίπου 224.000 κλάσεων, όπου κάθε κλάση αντιστοιχεί σε ένα έκθεμα του μουσείου, με φωτογραφίες που τραβήχτηκαν σε συνθήκες στούντιο. Το σύνολο αξιολόγησης (test set) αποτελείται σε ένα ποσοστό από φωτογραφίες που απεικονίζουν εκθέματα και τραβήχτηκαν από επισκέπτες του μουσείου, εισάγοντας μια μετατόπιση κατανομής (distribution shift) μεταξύ αυτού και του συνόλου εκπαίδευσης. Αποτελείται επιπλέον από ένα σύνολο εικόνων που δεν σχετίζονται με εκθέματα από το μουσείο Met, κάτι που κάνει το πρόβλημα να παίρνει χαρακτήρα ανίχνευσης εκτός κατανομής (out-of-distribution detection). Το προτεινόμενο benchmark ακολουθεί το παράδειγμα άλλων πρόσφατων συνόλων δεδομένων για το ILR σε διαφορετικά πεδία, ώστε να ενθαρρυνθούν προσεγγίσεις εφαρμόσιμες ανεξαρτήτως πεδίου. Προκειμένου να προσφερθεί μια βάση για μελλοντικές συγκρίσεις, αξιολογούμε κατάλληλες μεθόδους σε αυτό. Η αυτοεποπτευόμενη (self-supervised) και η εποπτευόμενη συγκριτική (supervised contrastive) μάθηση συνδυάζονται αποτελεσματικά για την εκπαίδευση CNNs που παράγουν την αναπαράσταση εικόνας (image representation) η οποία χρησιμοποιείται σε συνδυασμό με έναν μη παραμετρικό ταξινομητή (non-parametric classifier), υποδεικνύοντας μια υποσχόμενη κατεύθυνση για το ILR.

Λέξεις Κλειδιά

αναγνώριση σε επίπεδο οντότητας, αναγνώριση έργων τέχνης, σύνολο δεδομένων αναγνώρισης μεγάλης κλίμακας, ταξινόμηση κοντινότερου γείτονα

Abstract

In this thesis, the creation of a new dataset and benchmark for large-scale instance-level recognition (ILR) in the domain of artworks is addressed. The proposed dataset, called the Met dataset, exhibits a number of different challenges such as large inter-class similarity, long-tail distribution, and many classes. It relies on the open access collection of The Met museum to form a large training set of about 224k classes, where each class corresponds to a museum exhibit with photos taken under studio conditions. The evaluation set is primarily composed of photos taken by museum guests depicting exhibits, which introduces a distribution shift between training and testing. It is additionally composed of a set of images not related to Met exhibits making the task resemble an out-of-distribution detection problem. The proposed benchmark follows the paradigm of other recent datasets for ILR on different domains to encourage research on domain independent approaches. In order to offer a testbed for future comparisons, a number of suitable approaches are evaluated. Self-supervised and supervised contrastive learning are effectively combined to train CNNs that produce the image representation used in combination with a non-parametric classifier, showing a promising direction for ILR. The dataset webpage (also contains reference code) is: http://cmp.felk.cvut.cz/met/ .

Keywords

instance-level recognition, artwork recognition, large-scale recognition dataset, knn classification

to my Grandma, Ελένη

Contents

Περίληψη 1

Abstract 3

Preface 15

0 Εκτεταμένη Ελληνική Περίληψη 17
- 0.1 Περιγραφή προβλήματος . . . 17
- 0.2 Συνεισφορές εργασίας . . . 19
  - 0.2.1 Το σύνολο δεδομένων Met και το συνοδευτικό benchmark . . . 19
  - 0.2.2 Πειραματική αξιολόγηση σχετικών μεθόδων . . . 21

1 Introduction 23
- 1.1 Contributions . . . 25
- 1.2 Publications . . . 26
- 1.3 Structure of the Thesis . . . 26
- 1.4 Authorship . . . 26

2 Background and related work 27
- 2.1 Image retrieval . . . 27
  - 2.1.1 Global and local descriptors . . . 27
- 2.2 Contrastive learning . . . 28
  - 2.2.1 Self-supervised contrastive learning . . . 28

3 The Met dataset 31
- 3.1 Dataset overview . . . 31
- 3.2 Dataset collection . . . 32
  - 3.2.1 Image sources . . . 32
  - 3.2.2 Annotation . . . 32
- 3.3 Dataset statistics . . . 39
- 3.4 Comparison to other datasets . . . 41
  - 3.4.1 Artwork datasets . . . 41
  - 3.4.2 ILR datasets . . . 41
- 3.5 Benchmark and evaluation protocol . . . 42
  - 3.5.1 Splits . . . 42
  - 3.5.2 Metrics . . . 43

4 Methods **45**
4.1 Image representation . . . 45
4.2 kNN classification . . . 46
4.3 Pretrained models . . . 47
4.4 Training on the Met . . . 48
4.4.1 Learning with a classification objective . . . 48
4.4.2 Representation learning for the kNN classifier . . . 50

5 Experimental evaluation **55**
5.1 Implementation details . . . 55
5.2 Image representation and kNN classifier components . . . 56
5.3 Pretrained backbones and kNN classifier . . . 56
5.4 Training on the Met dataset . . . 57
5.5 Long tail recognition and kNN classifier . . . 60

6 Conclusion **61**

Appendices **63**

A Dataset extras **65**
A.1 Dataset hosting and maintenance . . . 65
A.2 License . . . 65
A.3 Flickr users . . . 65
A.4 Extra image examples . . . 67

B Additional results **71**
B.1 Descriptor dimensionality . . . 71
B.2 Local descriptors . . . 71
B.3 Approaches for long-tail recognition . . . 72
B.4 Mini dataset . . . 72
B.5 OOD ratio . . . 72

Bibliography **79**

List of Figures

1 Το γλυπτό Clytie του William Henry Rinehart. Πηγή: https://www.metmuseum.org/art/collection/search/11922 17

2 Παραδείγματα ILR στο σύνολο δεδομένων που προτείνεται σε αυτήν την εργασία, που καταδεικνύουν τη δυσκολία του προβλήματος. Για να λύσει το πρόβλημα, ένα μοντέλο πρέπει να αντιστοιχίσει τη εικόνα ερωτήματος (query/test image) με αυτήν από τις εικόνες εκπαίδευσης (train images) που ανήκει στην ίδια κατηγορία (σε αυτή την περίπτωση που απεικονίζει το ίδιο έκθεμα). Οι μικρές διαφορές μεταξύ των κατηγοριών απαιτούν μοντέλα τα οποία εστιάζουν σε λεπτομέρειες, ώστε να προβούν σε σωστή κατηγοριοποίηση. 18

3 Σχεδιάγραμμα του προβλήματος ILR. Υπάρχει ένας ταξινόμητης έργων τέχνης (artwork classifier) ο οποίος έχει εκπαιδευτεί σε ένα μεγάλο σύνολο εκπαίδευσης (training set) από έργα τέχνης. Στην φάση της αξιολόγησης, κατά την οποία δίνεται ως είσοδος στον ταξινομητή μια εικόνα ερωτήματος (query image), ο ταξινομητής θα πρέπει να μπορεί να αναγνωρίζει το εικονιζόμενο έργο τέχνης, εάν απεικονίζεται κάποιο, και επίσης να παρέχει ένα μέτρο εμπιστοσύνης (confidence measure) για την πρόβλεψή του. 19

4 Παραδείγματα εικόνων εκθεμάτων (exhibits) και εικόνων ερωτημάτων (queries) από το σύνολο δεδομένων Met. που επιδεικνύουν την ποικιλομορφία στην γωνία θέασης, τον φωτισμό και το θέμα. Οι εικόνες εκθεμάτων και οι εικόνες ερωτημάτων από την ίδια κατηγορία βρίσκονται σε διακεκομμένες γραμμές. Παρουσιάζονται και οι εικόνες περισπασμού (distractor queries), που είναι υποσύνολο του συνόλου αξιολόγησης, και οι οποίες δεν απεικονίζουν εκθέματα του μουσείου. 20

5 Απαιτητικά παραδείγματα αναγνώρισης από το σύνολο αξιολόγησης του Met, με χρήση του ταξινομητή κοντινότερου γείτονα για την προσέγγιση με την κορυφαία απόδοση. Οι εικόνες ερωτήματος (query/test images) παρουσιάζονται δίπλα στον πλησιέστερο γείτονά τους από τις εικόνες εκθεμάτων του Met του οποίου η κλάση αποτελεί και την πρόβλεψη. Επάνω σειρά: σωστές προβλέψεις. Μεσαία σειρά: λανθασμένες προβλέψεις. Παρουσιάζουμε επίσης μια εικόνα από τη σωστή (ground-truth) κλάση. Κάτω σειρά: προβλέψεις για εικόνες περίσπασης (OOD-test) οι οποίες έχουν λάβει υψηλό μέτρο εμπιστοσύνης από τον ταξινομητή. Ο σκοπός για αυτές είναι να αποκτήσουν χαμηλή εμπιστοσύνη. 21

1.1 Clytie by William Henry Rinehart. Image source: https://www.metmuseum.org/art/collection/search/11922 . . . 23
1.2 Examples of instance-level recognition on the dataset proposed in this work, demonstrating the difficulty of the task. A proposed model has to match the test image to the one from the train images that belongs to the same class (in this case, depicts the same exhibit). The low-inter class variability requires models to attend to a high-level of detail in order to correctly classify such cases. . . . 24
1.3 Schematic of the artwork recognition task. There exists a classifier that has been trained on a large training set. At test time, given a query (test) image, the classifier should be able to recognize the depicted artwork, if any, and also provide a confidence measure for its prediction. . . . 25
1.4 Samples from the Met dataset of exhibit and query (Met and distractor) images, demonstrating the diversity in viewpoint, lighting, and subject matter of the images. Exhibit images and queries from the same Met class are indicated by dashed lines. . . . 26

3.1 Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border. . . . 33
3.2 Examples of other-artwork distractor queries. These distractor queries depict artworks that do not belong to the Met collection. . . . 34
3.3 Examples of non-artwork distractor queries. These distractor queries do not depict artworks. . . . 34
3.4 An overview of the Met dataset collection and annotation process. . . . 35
3.5 An example of using the interface of The Met collection website. . . . 35
3.6 Examples of Met queries captured by our team. . . . 35
3.7 Example annotation step during the filtering stage of the annotation process. In this phase, invalid images are discarded, *i.e.* images containing visitor faces, images not depicting exhibits, or images with more than one exhibit. . . . 36
3.8 Example annotation step during the annotation stage of the annotation process. In this phase, queries are labeled with the corresponding Met class. 37
3.9 Example of text-based search on the manual search engine provided by the Met. . . . 38
3.10 Example annotation step during the verification stage of the annotation process. In this phase, annotators verify the correctness of the labeling per query. . . . 38
3.11 Example annotation step during the distractor verification stage of the annotation process. In this phase, annotators verify that other-artwork distractor queries are true distractors and do not belong to The Met collection. . . . 39
3.12 Number of exhibit images per time period. . . . 39

3.13 Number of images and classes by department. Met queries are assigned to the department of their ground-truth class. Some departments that do not contain queries but contain exhibit images are not shown. 40

3.14 Number of distractor images by Wikimedia category. Top categories shown: art-related categories in solid blue and generic categories in dash purple. 40

3.15 Left: number of Met classes versus number of training images per class. Right: number of Met classes versus number of query images per class. Both vertical axes are shown in logarithmic scale, for better visualization. 41

3.16 The number of photographers versus the Met queries that belong to them. 41

3.17 First two rows: example of calculating ACC and GAP for a set of queries and their corresponding predictions and confidences. In the first row, ACC calculation is shown. It is calculated only on the Met queries, which are either correctly (green) or incorrectly classified (red). In the second row, for the GAP calculation, which takes into account distractors as well (grey), the queries are sorted by confidence in descending order. The confidence, the precision and the binary indicator of correctness at each rank of the sorted list is shown. Then, calculation of GAP is straightforward. Last two rows: Ways to incrementally achieve optimal GAP starting from the example of the top rows: 3rd row) to make as many predictions correct for the non distractor queries, 4th row) to place missclassified examples, which include distractors, in the bottom of the list 44

4.1 Example of the calculation of class confidences for a query image using the proposed kNN classifier. The similarity values in this example are fictional and only exist for the purposes of the figure. 47

4.2 Overview of the pipeline for kNN classification. The backbone produces the image descriptor after processing the image at multiple scales. It is subsequently whitened by PCA-whitening and used as the input to the kNN classifier. 47

4.3 The schematic of the DNet classifier. The backbone part of the classifier produces the image descriptor, which is subsequently fed to the cosine similarity classifier that produces the vector of logits. The latter contains the similarity of the image descriptor with every Met class. The entire pipeline is amenable to end-to-end training. 50

4.4 The Siamese architecture used along with the contrastive loss. A pair of images (in this example, a negative pair) is fed into the same backbone, producing their representations. These are then fed to the contrastive loss, along with their corresponding label. 51

4.5 Examples of training pairs used by the representation learning methods that are trained on the Met training set. First row: SimSiam uses only image augmentations to form positive pairs, without the use of any supervision. Second row: Con-Syn also uses image augmentations for the formation of the positive pairs, however it uses supervision from the Met labels to form hard negative pairs. Third row: Con-Syn+Real additionally picks the positive pair from a pool that contains the augmentation plus all the other images from the same class as the anchor. Fourth row: Con-Syn+Real-closest limits the pool of candidate positive pairs to either the augmentation or the closest image (as measured in the representation space) that comes from the same class as the anchor. 53

5.1 Examples of incorrect and correct classification of test images for R18IN (baseline) and R18IN Con-Syn+Real-closest (R18IN*), respectively. The test images are shown next to their nearest neighbor from the Met exhibits that produced the respective prediction per method. 58

5.2 Examples of hard negative pairs formed by the approaches that use the Contrastive loss on the Met training set. These examples additionally demonstrate the large inter-class similarity of the dataset. Images are shown as squares only for the purposes of this figure. 59

5.3 Challenging examples from the Met dataset for the top performing approach. Test images are shown next to their nearest neighbor from the Met exhibits that generated the prediction of the corresponding class. Top row: correct predictions. Middle row: incorrect predictions; an image of the ground truth class is also shown. Bottom row: high confidence predictions for OOD-test images; the goal is to obtain low confidence for these. 59

5.4 Accuracy improvement of the kNN classifier over the parametric one for varying number of training images per class. DNet is trained with AF loss for the parametric classifier, while the embeddings learned with this setup are used for the kNN classifier. Relative improvements are reported in percentage for the different embedding variants. 60

A.1 Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border. 68

A.2 Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border. 69

A.3 Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border. 70

B.1 Performance with a kNN classifier versus dimensionality for different backbones. Two approaches are combined by simple representation concatenation before PCAw and is denoted by "+". *: Contrastive *Syn+Real-Closest* training on the Met dataset. 71

List of Tables

3.1 Comparison to art datasets. † For datasets with multiple kinds of annotations, the task with the largest number of classes is reported. 42

3.2 Comparison to instance-level recognition datasets. 42

3.3 Number of images and classes in the Met dataset per split. Met exhibits images are from the museum's open collection, while Met query images are from museum visitors. Query images contain distractor images too (denoted by the +1 class) while the rest of val/test classes are subset of the train classes. 43

5.1 Recognition performance for kNN classifier on representation obtained from ResNet18 pretrained on ImageNet. MS: multi-scale representation. †: tuning k, τ only with Met queries, and without distractor queries in the validation set. 56

5.2 Comparison of recognition performance for kNN classifier with representation from backbone networks pretrained for different tasks. Relative improvements compared to the corresponding network trained on ImageNet are shown in parentheses. 57

5.3 Performance comparison for different types of training on the Met dataset. Training starts from the result of pretraining on ImageNet (IN) or that of SWSL. Baseline: not trained on the Met. 58

B.1 Performance of R18IN with kNN classification with different amount (percentage of their total number) of distractor queries in the validation (for tuning k, τ) and test set. Ratio lower than 100 is achieved by removing the appopriate amount of distractor queries. 73

Preface

This thesis concludes my 5-year studies at the National Technical University of Athens' School of Electrical and Computer Engineering. It's been a fantastic journey that has provided me with a wealth of knowledge, experiences, wonderful memories, and great friends. I feel obligated to say a few words about the people who have helped me get to where I am today.

To begin, I'd like to express my heartfelt gratitude to Assistant Professor Giorgos Tolias, my advisor and mentor over the past year, without whom none of this would have been possible. His constant guidance and support were not only invaluable in completing this thesis, but also in shaping my first steps into the research world. I will be eternally grateful for the opportunities he provided and the trust he placed in me. I am overjoyed to be continuing for my PhD studies at the lab where I have been working with him for the past year, the Visual Recognition Group in the Department of Cybernetics of CTU in Prague.

Next, I'd like to thank my collaborators Noa, Guangxing, Sarah, and Nanne, who were coauthors of the respective publication. Working with them was an incredible experience. I'd also like to thank my supervisor from NTUA, Professor Stefanos Kollias, for our excellent collaboration.

Many thanks to all of my friends from NTUA for making my student years a little more enjoyable, despite all of the hard work that we all had to put in daily. There are too many to name, but I'd like to mention Nikos Antoniou and Sotiris Karapiperis in particular for being not only amazing friends, but also some of the brightest minds I've ever met. Also, I'd like to thank the friends I made at the Czech Technical University in Prague for making my time as an expat a little easier.

Last but not least, I'd like to thank some people outside of academia. First and foremost, my parents for their endless support over the years, as well as my sister for always being there for me. Finally, Christina, for giving me the motivation to keep going when I needed it the most. I am thankful to have her in my life.

Κεφάλαιο 0

Εκτεταμένη Ελληνική Περίληψη

0.1 Περιγραφή προβλήματος

Η οπτική αναγνώριση ή κατηγοριοποίηση (visual recognition or classification) επιτυγχάνεται με κατηγορίες (κλάσεις) ορισμένες σε διαφορετικά επίπεδα λεπτομέρειας. Για παράδειγμα, το έκθεμα που φαίνεται στην εικόνα 1 κατηγοριοποιείται ως "Clytie" του William Henry Rinehart, ως γλυπτό ή ως έργο τέχνης, από την άποψη της αναγνώρισης σε επίπεδο οντότητας (instance-level recognition, ILR) [1], της λεπτομερούς κατηγοριοποίησης (fine-grained recognition) [2] ή της γενικής κατηγοριοποίησης (category-level recognition, CLR) [3], αντίστοιχα. Η κατηγοριοποίηση σε επίπεδο οντότητας, αποτελεί ειδική περίπτωση της οπτικής κατηγοριοποίησης που στοχεύει στην αναγνώριση συγκεκριμένων αντικειμένων και όχι μόνο της γενικής (σημασιολογικής) κατηγορίας τους. Εφαρμόζεται σε διάφορους τομείς, όπως τα προϊόντα, τα αξιοθέατα, οι αστικές τοποθεσίες και τα έργα τέχνης. Αντιπροσωπευτικά παραδείγματα πρακτικών εφαρμογών της είναι η αναγνώριση τοποθεσιών (place recognition) [4, 5], η αναγνώριση και η ανάκτηση αξιοθεάτων (landmark recognition and retrieval) [6], η αντιστοίχιση προϊόντων (street-to-shop product matching) [7, 8, 9] και η αναγνώριση έργων τέχνης (artwork recognition) [10]. Υπάρχουν διάφοροι παράγοντες που καθιστούν το ILR απαιτητικό πρόβλημα. Ο αριθμός των κατηγοριών (κλάσεων) που πραγματεύεται είναι γενικά πολύ μεγάλος φτάνοντας μέχρι και την τάξη των 10^6 σε μερικά σύνολα δεδομένων, με πολλές κλάσεις να αντιπροσωπεύονται από λίγα ή ένα μόνο παραδείγμα-

Figure 1. *Το γλυπτό Clytie του William Henry Rinehart. Πηγή:* https://www.metmuseum.org/art/collection/search/11922

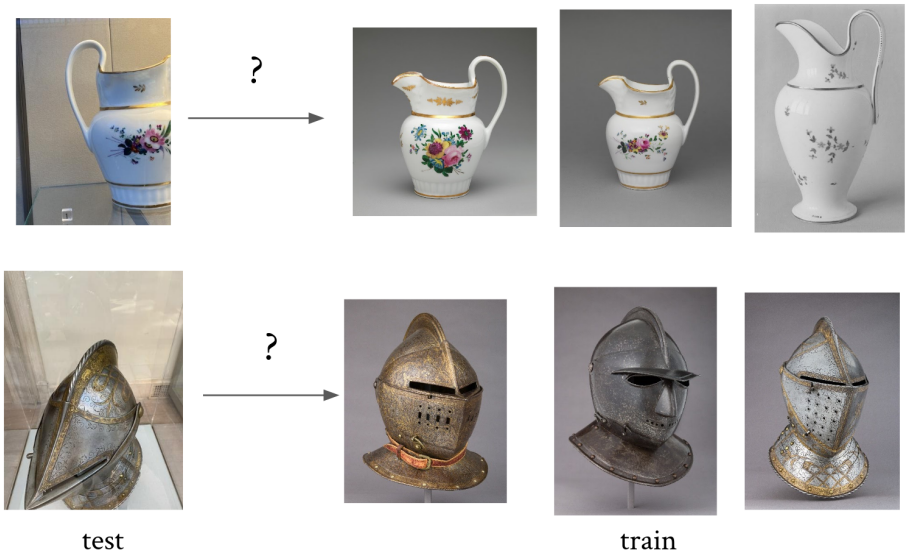

Figure 2. *Παραδείγματα ILR στο σύνολο δεδομένων που προτείνεται σε αυτήν την εργασία, που καταδεικνύουν τη δυσκολία του προβλήματος. Για να λύσει το πρόβλημα, ένα μοντέλο πρέπει να αντιστοιχίσει τη εικόνα ερωτήματος (query/test image) με αυτήν από τις εικόνες εκπαίδευσης (train images) που ανήκει στην ίδια κατηγορία (σε αυτή την περίπτωση που απεικονίζει το ίδιο έκθεμα). Οι μικρές διαφορές μεταξύ των κατηγοριών απαιτούν μοντέλα τα οποία εστιάζουν σε λεπτομέρειες, ώστε να προβούν σε σωστή κατηγοριοποίηση.*

τα, ενώ η μικρή οπτική διαφοροποίηση μεταξύ των κλάσεων κάνει το πρόβλημα ακόμη πιο δύσκολο, όπως φαίνεται και στα παράδειγματα αναγνώρισης στην εικόνα 2. Λόγω αυτών των δυσκολιών, γίνεται συχνά η επιλογή να αντιμετωπίζεται η κατηγοριοποίηση σε επίπεδο οντότητας ως πρόβλημα ανάκτησης σε επίπεδο οντότητας (instance-level retrieval) [11]. Συγκεκριμένες εφαρμογές, *π.χ.* στον τομέα των προϊόντων ή της τέχνης απαιτούν δυναμικές ενημερώσεις του συνόλου κατηγοριών. Εικόνες από νέες κατηγορίες προστίθενται συνεχώς, δίνοντας στο ILR χαρακτήρα αναγνώρισης ανοιχτού συνόλου (open-set recognition) [12].

Παρά τις πολλές πρακτικές εφαρμογές και τις δύσκολες πτυχές του προβλήματος, το ILR έχει προσελκύσει λιγότερη προσοχή από το πρόβλημα της γενικής κατηγοριοποίησης (CLR), το οποίο συνοδεύεται από δημοφιλή σύνολα δεδομένων (datasets) και benchmarks, όπως το ImageNet [13], που εξυπηρετούν ως βάση δοκιμών ακόμη και για άλλα προβλήματα οπτικής αναγνώρισης. Μια κύρια αιτία για αυτό είναι η έλλειψη συνόλων δεδομένων μεγάλης κλίμακας (large-scale datasets), των οποίων η δημιουργία και η επισημείωση (annotation) τους για το πρόβλημα του ILR είναι μια πολύ κοπιαστική διαδικασία. Κατά συνέπεια, πολλά σύνολα δεδομένων περιλαμβάνουν θόρυβο στις επισημειώσεις τους [1, 10, 6]. Σε αυτήν την εργασία, καλύπτουμε αυτό το κενό εισάγοντας ένα σύνολο δεδομένων για το ILR στον τομέα των έργων τέχνης.

Ο τομέας της τέχνης έχει τραβήξει μεγάλη προσοχή στην κοινότητα της όρασης υπολογιστών. Μια δημοφιλής γραμμή έρευνας επικεντρώνεται σε μια συγκεκριμένη πτυχή της ταξινόμησης, την πρόβλεψη χαρακτηριστικών (attribute prediction) [14, 15, 16, 17, 18]. Σε αυτήν την περίπτωση, τα χαρακτηριστικά (attributes) αντιστοιχούν σε διάφορα είδη μεταδεδομένων για ένα έργο τέχνης, όπως το στυλ, το είδος, η περίοδος, ο καλλιτέχνης και άλλα.

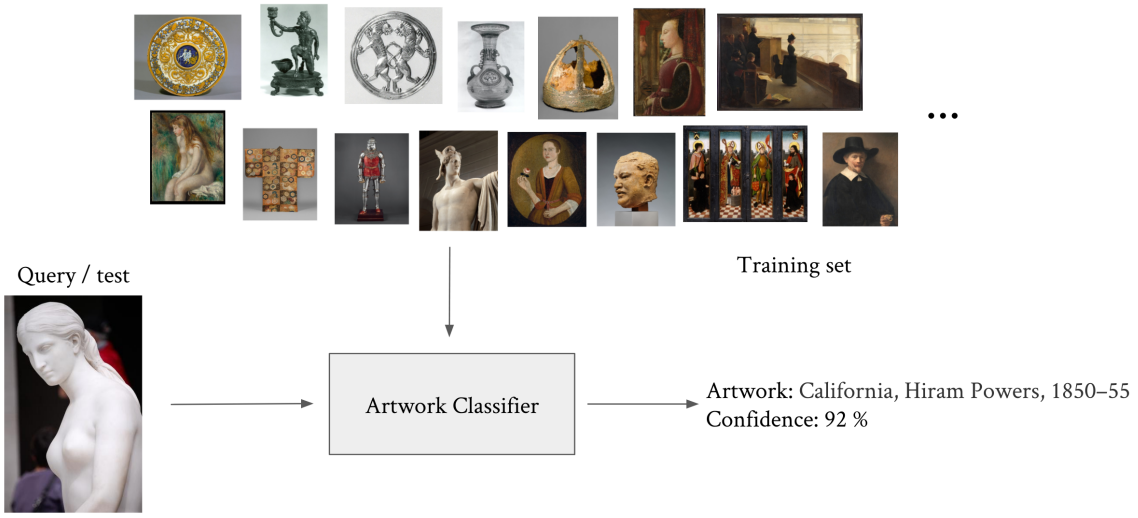

Figure 3. *Σχεδιάγραμμα του προβλήματος ILR. Υπάρχει ένας ταξινόμητης έργων τέχνης (artwork classifier) ο οποίος έχει εκπαιδευτεί σε ένα μεγάλο σύνολο εκπαίδευσης (training set) από έργα τέχνης. Στην φάση της αξιολόγησης, κατά την οποία δίνεται ως είσοδος στον ταξινομητή μια εικόνα ερωτήματος (query image), ο ταξινομητής θα πρέπει να μπορεί να αναγνωρίζει το εικονιζόμενο έργο τέχνης, εάν απεικονίζεται κάποιο, και επίσης να παρέχει ένα μέτρο εμπιστοσύνης (confidence measure) για την πρόβλεψή του.*

Τα μεταδεδομένα για την πρόβλεψη χαρακτηριστικών λαμβάνονται από μουσεία και βάσεις δεδομένων που καθιστούν αυτές τις πληροφορίες ελεύθερα διαθέσιμες. Αυτό καθιστά βολική τη διαδικασία δημιουργίας των συνόλων δεδομένων, που όμως καταλήγουν να είναι συχνά πολύ θορυβώδη λόγω της αραιότητας αυτών των πληροφοριών [15, 17]. Μια άλλη γνωστή εργασία είναι η γενίκευση ή η προσαρμογή πεδίου (domain generalization and adaptation) όπου τα μοντέλα αναγνώρισης ή ανίχνευσης αντικειμένων εκπαιδεύονται σε φυσικές εικόνες και η γενίκευσή τους ελέγχεται σε έργα τέχνης [19]. Ένα πολύ δύσκολο πρόβλημα είναι η ανακάλυψη μοτίβων [20, 21] που προορίζεται ως εργαλείο για ιστορικούς τέχνης και στοχεύει στην εύρεση κοινών μοτίβων μεταξύ έργων τέχνης. Σε αυτή την εργασία εστιάζουμε στο πρόβλημα του ILR για έργα τέχνης (βλ. Εικόνα 3 για ένα σχεδιάγραμμα του προβλήματος) που συνδυάζει τις προαναφερθείσες προκλήσεις του ILR, σχετίζεται με εφαρμογές με θετικό αντίκτυπο, όπως στην εκπαίδευση και δεν έχει προσελκύσει ακόμη πολλή προσοχή στην ερευνητική κοινότητα.

0.2 Συνεισφορές εργασίας

0.2.1 Το σύνολο δεδομένων Met και το συνοδευτικό benchmark

Αρχικά, δημιουργούμε ένα νέο σύνολο δεδομένων μεγάλης κλίμακας (large-scale dataset) για το πρόβλημα του ILR, το οποίο στηρίζεται στην συλλογή ανοιχτής πρόσβασης του The Metropolitan Museum of Art (The Met) στη Νέα Υόρκη (βλ. Εικόνα 4 για μια επισκόπηση του).

Αποτελείται από δύο είδη εικόνων, τις εικόνες εκθεμάτων (exhibit images), και τις εικόνες ερωτήματος (query images). Οι εικόνες εκθεμάτων απαρτίζουν το σύνολο εκπαίδευσης

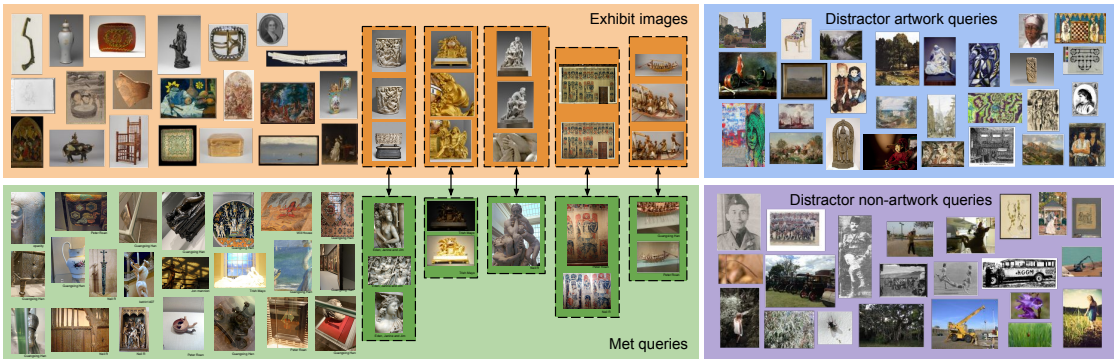

Figure 4. *Παραδείγματα εικόνων εκθεμάτων (exhibits) και εικόνων ερωτημάτων (queries) από το σύνολο δεδομένων Met. που επιδεικνύουν την ποικιλομορφία στην γωνία θέασης, του φωτισμό και το θέμα. Οι εικόνες εκθεμάτων και οι εικόνες ερωτημάτων από την ίδια κατηγορία βρίσκονται σε διακεκομμένες γραμμές. Παρουσιάζονται και οι εικόνες περισπασμού (distractor queries), που είναι υποσύνολο του συνόλου αξιολόγησης, και οι οποίες δεν απεικονίζουν εκθέματα του μουσείου.*

(training set) το οποίο αποτελείται από περίπου 400.000 εικόνες προερχόμενες από περισσότερες από 224.000 κλάσεις, με έργα τέχνης παγκόσμιας γεωγραφικής κάλυψης και από περίοδους που χρονολογούνται μέχρι την Παλαιολιθική εποχή. Κάθε έκθεμα του μουσείου αντιστοιχεί σε ένα μοναδικό έργο τέχνης και ορίζει τη δική του κλάση. Το σύνολο εκπαίδευσης παρουσιάζει κατανομή μακριάς ουράς (long-tail distribution) με περισσότερες από τις μισές κλάσεις να αντιπροσωπεύονται από μια εικόνα, καθιστώντας το πρόβλημα μια ειδική περίπτωση μάθησης με λίγα παραδείγματα (few-shot learning). Παρουσιάζει επίσης υψηλή οπτική ομοιότητα μεταξύ των κλάσεων.

Οι εικόνες ερωτήματος είναι εικόνες που πρέπει να κατηγοριοποιηθούν από τον ταξινομητή, αποτελώντας ουσιαστικά το σύνολο αξιολόγησης. Χωρίζονται στις εικόνες ερωτήματος Met (Met queries) και στις εικόνες ερωτήματος περισπασμού (distractor queries). Έχουμε δημιουργήσει τις επισημειώσεις (labels) για περισσότερες από 1.100 Met queries, που αποτελούν φωτογραφίες που τραβήχτηκαν από επισκέπτες του μουσείου και απεικονίζουν εκθέματα του μουσείου. Υπάρχει μια μετατόπιση κατανομής (distribution shift) μεταξύ αυτών και των εικόνων εκπαίδευσης, οι οποίες έχουν τραβηχτεί υπό συνθήκες στούντιο. Συμπεριλαμβάνουμε στις εικόνες ερωτημάτος επιπλέον ένα μεγάλο σύνολο εικόνων περίσπασης (distractor queries) που δεν απεικονίζουν εκθέματα του μουσείου Met, δηλαδή αποτελούν εικόνες εκτός κατανομής (Out-Of-Distribution,OOD) [22, 23]. Τις χρησιμοποιούμε ώστε να προσομοιώθούν ρεαλιστικές συνθήκες αναγνώρισης και χρησιμεύουν για τον έλεγχο της ευρωστίας ενός ταξινομητή. Μέρος τους αποτελείται από έργα τέχνης που δεν προέρχονται από το The Met, προκειμένου να αυξηθεί περαιτέρω ο αριθμός των προκλήσεων του συνόλου δεδομένων Met, ενώ οι υπόλοιπες προέρχονται από γενικές κατηγορίες. Το σύνολο δεδομένων Met ακολουθεί το πρωτόκολλο αξιολόγησης του πρόσφατου συνόλου δεδομένων Google Landmarks (GLD) [6] με σκοπό να ενθάρρυνθούν καθολικές προσεγγίσεις για το ILR, οι οποίες θα ισχύουν σε ένα ευρύτερο φάσμα πεδίων (domains). Σε αντίθεση όμως με το GLD, η επισημειώσεις δεν περιλαμβάνουν θόρυβο, και από όσο γνωρίζουμε αυτό είναι το μοναδικό σύνολο δεδομένων ILR σε αυτήν την κλίμακα, που δεν περιλαμβάνει θόρυβο στις

επισημειώσεις και είναι πλήρως διαθέσιμο στο κοινό.

Μαζί με το σύνολο δεδομένων Met δημιουργούμε και το αντίστοιχο benchmark. Για την αξιολόγηση μιας προτεινόμενης μεθόδου σε αυτό μετράμε δύο τυπικές μετρικές για το ILR, την ακρίβεια ταξινόμησης (ACC) και το Global Average Precision (GAP). Η ακρίβεια ταξινόμησης μετριέται μόνο στα Met queries, ενώ το GAP μετριέται σε όλα τα queries. Το τελευταίο λαμβάνει υπόψιν του εκτός από την πρόβλεψη και την εμπιστοσύνη της πρόβλεψης (prediction confidence) που πρέπει να παρέχει ο ταξινομητής. Το GAP χρησιμοποιεί την εμπιστοσύνη της πρόβλεψης ως τρόπο ανίχνευσης ερωτημάτων περισπασμού και εσφαλμένα ταξινομημένων Met ερωτημάτων. Επιτρέπει τη συμπερίληψη ερωτημάτων περισπασμού στην αξιολόγηση χωρίς την ανάγκη να συμπεριλαμβάνονται στην διαδικασία της εκμάθησης, καθώς ο ταξινομητής δεν προβλέπει ποτέ την κλάση περισπασμού. Για την επίτευξη βέλτιστου GAP απαιτείται, εκτός από σωστές προβλέψεις για όλες τις εικόνες ερωτημάτων Met, όλες οι εικόνες ερωτήματων περισπασμού να πάρουν μικρότερη εμπιστοσύνη πρόβλεψης από όλες τις εικόνες ερωτήματων Met.

0.2.2 Πειραματική αξιολόγηση σχετικών μεθόδων

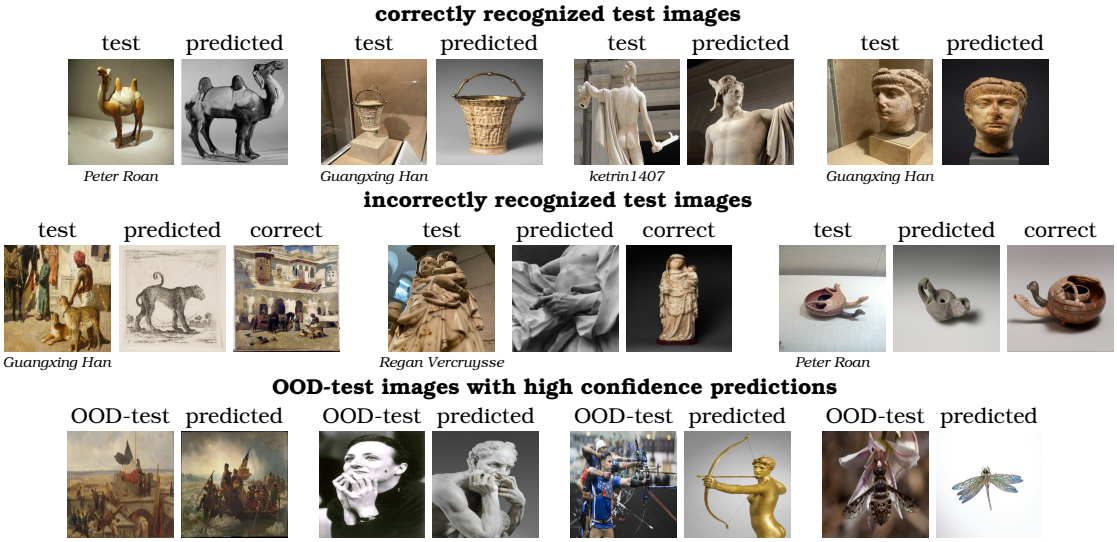

Figure 5. *Απαιτητικά παραδείγματα αναγνώρισης από το σύνολο αξιολόγησης του Met, με χρήση του ταξινομητή κοντινότερου γείτονα για την προσέγγιση με την κορυφαία απόδοση. Οι εικόνες ερωτήματος (query/test images) παρουσιάζονται δίπλα στον πλησιέστερο γείτονά τους από τις εικόνες εκθεμάτων του Met του οποίου η κλάση αποτελεί και την πρόβλεψη. Επάνω σειρά: σωστές προβλέψεις. Μεσαία σειρά: λανθασμένες προβλέψεις. Παρουσιάζουμε επίσης μια εικόνα από τη σωστή (ground-truth) κλάση. Κάτω σειρά: προβλέψεις για εικόνες περίσπασης (OOD-test) οι οποίες έχουν λάβει υψηλό μέτρο εμπιστοσύνης από τον ταξινομητή. Ο σκοπός για αυτές είναι να αποκτήσουν χαμηλή εμπιστοσύνη.*

Πραγματοποιούμε πειραματική αξιολόγηση (experimental evaluation) της απόδοσης σχετικών μεθόδων στο Met benchmark, προκειμένου να προσφέρθει μια βάση για συγκρίσεις και για να προταθούν μελλοντικές κατευθύνσεις. Για την αναπαράσταση των εικόνων σε όλα τα πειράματα χρησιμοποιούμε global αναπαραστάσεις που έχουμε εξάγει από εκπαιδευμένα

πλήρως συνελικτικά δίκτυα (fully-convolutional networks, FCNs), τα οποία παράγουν για κάθε εικόνα ένα διάνυσμα-περιγραφέα (descriptor). Αρχικά, πραγματοποιούμε μία σύγκριση μεταξύ παραμετρικών και μη-παραμετρικών ταξινομητών. Δείχνουμε ότι οι μη παραμετρικοί ταξινομητές (και συγκεκριμένα ένας ταξινομητής κοντινότερου γείτονα που έχουμε αναπτύξει) αποδίδουν πολύ καλύτερα από τους παραμετρικούς. Μπορούν να χειριστούν καλύτερα τα ερωτήματα περισπασμού και είναι πιο κατάλληλοι για αναγνώριση μακριάς ουράς.

Έπειτα πραγματοποιούμε σύγκριση μεταξύ διαφόρων τρόπων προεκπαίδευσης (pretraining) των πλήρως συνελικτικών δικτύων που εξάγουν την αναπαράσταση των εικόνων, ώστε να μελετήσουμε την ικανότητα τους για μεταφορά μάθησης (transfer learning) στο σύνολο δεδομένων μας. Τα δίκτυα, τα οποία χρησιμοποιούν όλα την ίδια αρχιτεκτονική για δίκαιη σύγκριση, συγκρίνονται με ένα δίκτυο αναφοράς, το οποίο είναι προεκπαιδευμένο για το πρόβλημα της γενικής κατηγοριοποίησης στο ImageNet. Παρατηρούμε ότι η προεκπαίδευση σε διαφορετικού είδους προβλήματα, όπως η πρόβλεψη χαρακτηριστικών, αλλά από το ίδιο πεδίο (έργα τέχνης), δεν είναι αναγκαστικά ωφέλιμη για το πρόβλημα του ILR σε έργα τέχνης, μάλιστα ρίχνει την απόδοση σε σχέση με το δίκτυο αναφοράς. Από την άλλη, η προεκπαίδευση για συγγενικά προβλήματα όπως η εκμάθηση μετρικής (metric learning), αλλά σε διαφορετικά πεδία, όπως τα αξιοθέατα, προσφέρει βελτιώσεις σε σχέση με το δίκτυο αναφοράς. Επίσης, βλέπουμε ότι η εκμάθηση αναπαράστασης χωρίς επίβλεψη (unsupervised representation learning) ως το κομμάτι της προεκπαίδευσης ενός δικτύου προσφέρει καλή γενίκευση, βελτιώνοντας λίγο την επίδοση. Τέλος, τα καλύτερα αποτελέσματα τα παίρνουμε από ένα δίκτυο το οποίο έχει προεκπαιδευθεί σε ενα πολύ μεγάλο όγκο εικόνων, επιβεβαιώνοντας τα ωφέλη της προεκπαίδευσης μεγάλης κλίμακας.

Η βελτίωση της αναπαράστασης των εικόνων καθίσταται απαραίτητη με τη χρήση μη παραμετρικών ταξινομητών. Για το σκοπό αυτό, πραγματοποιούμε εκμάθηση αναπαράστασης (representation learning) στο σύνολο εκπαίδευσης του Met dataset. Δείχνουμε ότι οι πρόσφατες μέθοδοι αυτοεποπτευόμενης (self-supervised) εκμάθησης που βασίζονται μόνο σε επαυξήσεις εικόνων (image augmentations) είναι ωφέλιμες, αλλά οι διαθέσιμες επισημειώσεις που συνοδεύουν το σύνολο δεδομένων δεν θα πρέπει να παραβλέπονται. Συνθέτουμε μία συνδυασμένη προσέγγιση αυτοεποπτευόμενης και εποπτευόμενης συγκριτικής (supervised contrastive) μάθησης που εκπαιδεύει το δίκτυο που εξάγει την αναπαράσταση με την χρήση ζευγών εικόνων και επιτυγχάνει την καλύτερη απόδοση στο benchmark μας, υποδεικνύοντας υποσχόμενες μελλοντικές κατευθύνσεις. Παραθέτουμε παραδείγματα αναγνώρισης αυτής της μεθόδου στην εικόνα 5. Αυτά δείχνουν την δυσκολία της κατηγοριοποίησης στο σύνολο αξιολόγησης του συνόλου δεδομένων μας.

Ο στόχος αυτού του συνόλου δεδομένων είναι να καθιερωθεί στα τυπικά benchmarks για το ILR. Αναμένουμε να επωφελήσει την έρευνα όχι μόνο για το ILR στο πεδίο των έργων τέχνης αλλά σε όλα τα πεδία που εφαρμόζεται το ILR σε συνδυάσμό με τα άλλα υπάρχοντα σύνολα δεδομένων.

Chapter 1

Introduction

Classification of objects can be done with categories defined at different levels of granularity. For example, the piece of art shown in Figure 1.1 is classified as "Clytie" by William Henry Rinehart, as sculpture, or artwork, from the point of view of instance-level recognition [1], fine-grained recognition [2], or generic category-level recognition [3], respectively. Instance-level recognition (ILR) is the visual recognition task that aims to recognize specific instances of objects and not only their semantic class. It is applied to a variety of domains such as products, landmarks, urban locations, and artworks. Representative examples of real world applications are place recognition [4, 5], landmark recognition and retrieval [6], image-based localization [24, 25], street-to-shop product matching [7, 8, 9], and artwork recognition [10].

There are several factors that make ILR a challenging task. It is typically required to deal with a large category set, whose size reaches the order of 10^6, with many classes represented by only a few or a single example, while the small between class variability further increases the hardness (see Figure 1.2 for challenging examples of ILR). Due to these difficulties the choice is often made to handle instance-level recognition, or equivalently instance-level classification, as an instance-level retrieval task [11]. Particular applications, *e.g.* in the product or art domain require dynamic updates of the category set; images from new categories are continuously added. Therefore, ILR is a form of open set recognition [12].

Figure 1.1. *Clytie by William Henry Rinehart. Image source:* https://www.metmuseum.org/art/collection/search/11922

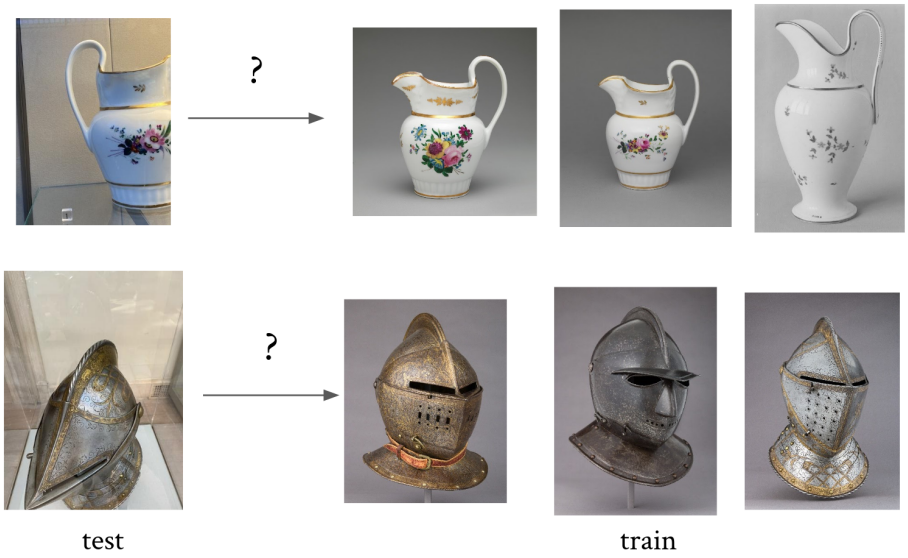

Figure 1.2. *Examples of instance-level recognition on the dataset proposed in this work, demonstrating the difficulty of the task. A proposed model has to match the test image to the one from the train images that belongs to the same class (in this case, depicts the same exhibit). The low-inter class variability requires models to attend to a high-level of detail in order to correctly classify such cases.*

Despite the many real-world applications and challenging aspects of the task, ILR has attracted less attention than category-level recognition (CLR) tasks, which are accompanied by large and popular benchmarks, such as ImageNet [13], that serve as a testbed even for approaches applicable beyond classification tasks. A major cause for this is the lack of large-scale datasets. Creating datasets with accurate ground truth at large scale for ILR is a tedious process. As a consequence, many datasets include noise in their labels [1, 10, 6]. In this work, we fill this gap by introducing a dataset for instance-level classification in the artwork domain.

The art domain has attracted a lot of attention in computer vision research. A popular line of research focuses on a specific flavor of classification, namely attribute prediction [14, 15, 16, 17, 18]. In this case, attributes correspond to various kinds of metadata for a piece of art, such as style, genre, period, artist and more. The metadata for attribute prediction is obtained from museums and archives that make this information freely available. This makes the dataset creation process convenient, but the resulting datasets are often highly noisy due to the sparseness of this information [15, 17]. Another known task is domain generalization or adaptation where object recognition or detection models are trained on natural images and their generalization is tested on artworks [19]. A very challenging task is motif discovery [20, 21] which is intended as a tool for art historians, and aims to find shared motifs between artworks. In this work we focus on the task of ILR for artworks (see Figure 1.3 for a schematic of the task) which combines the aforementioned challenges of ILR, is related to applications with positive impact, such as educational applications, and has not yet attracted attention in the research community.

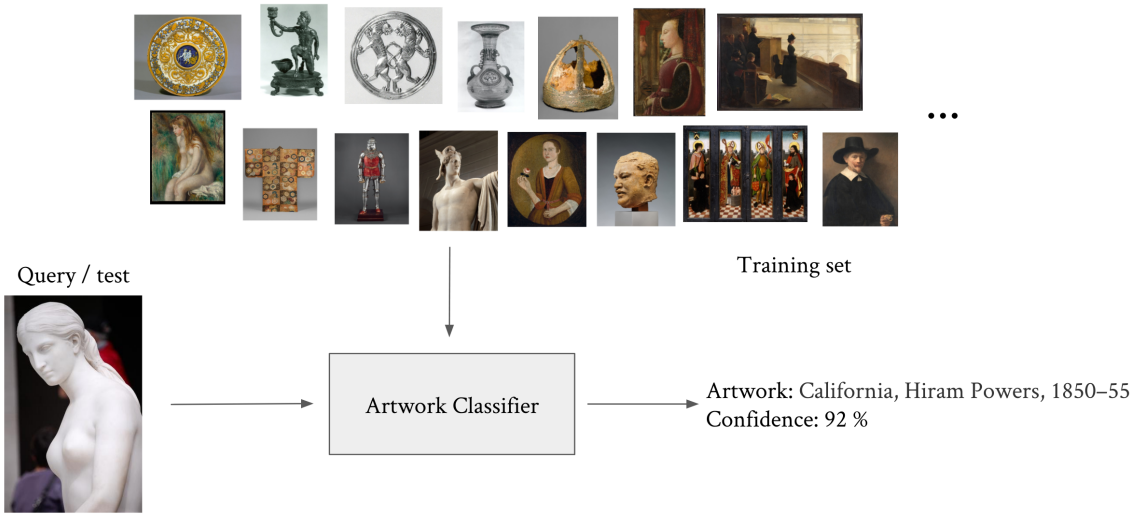

Figure 1.3. *Schematic of the artwork recognition task. There exists a classifier that has been trained on a large training set. At test time, given a query (test) image, the classifier should be able to recognize the depicted artwork, if any, and also provide a confidence measure for its prediction.*

1.1 Contributions

The contributions of this work are two-fold: Firstly, we introduce a new large-scale dataset for instance-level classification by relying on the open access collection from the Metropolitan Museum of Art (The Met) in New York (see Figure 1.4 for an overview). The training set consists of about 400k images from more than 224k classes, with artworks of world-level geographic coverage and chronological periods dating back to the Paleolithic period. Each museum exhibit corresponds to a unique artwork, and defines its own class. The training set exhibits a long-tail distribution with more than half of the classes represented by a single image, making it a special case of few-shot learning. We have established ground-truth for more than 1,100 images captured by museum visitors, which form the query set. Note that there is a distribution shift between this query set and the training images which are created in studio-like conditions. We additionally include a large set of distractor images not related to The Met, which form an Out-Of-Distribution (OOD) [22, 23] query set. The dataset follows the paradigm and evaluation protocol of the recent Google Landmarks Dataset (GLD) [6] to encourage universal ILR approaches that are applicable in a wider range of domains. Nevertheless, in contrast to GLD, the established ground-truth does not include noise. To our knowledge this the only ILR dataset at this scale, that includes no noise in the ground-truth and is fully publicly available.

On top of that, the introduced dataset is accompanied by performance evaluation of relevant approaches. We show that non-parametric classifiers perform much better than parametric ones. Improving the visual representation becomes essential with the use of non-parametric classifiers. To this end, we show that the recent self-supervised learning methods that rely only on image augmentations are beneficial, but the available ILR labels should not be discarded. A combined self-supervised and supervised contrastive learning approach is the top performer in our benchmark indicating promising future directions.

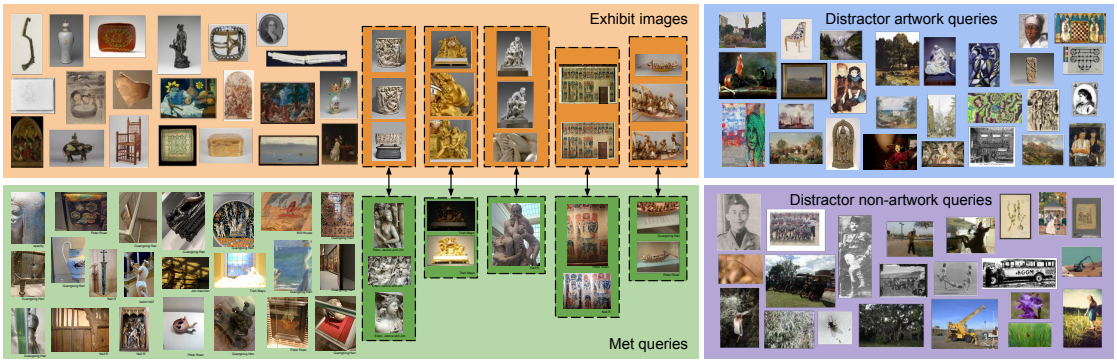

Figure 1.4. *Samples from the Met dataset of exhibit and query (Met and distractor) images, demonstrating the diversity in viewpoint, lighting, and subject matter of the images. Exhibit images and queries from the same Met class are indicated by dashed lines.*

1.2 Publications

This thesis builds on the results previously published in the following publications:

[26] Nikolaos Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord και Giorgos Tolias. *The Met Dataset: Instance-level Recognition for Artworks. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021

1.3 Structure of the Thesis

The rest of the manuscript is organized as follows. Chapter 2 discusses the related work of this thesis. Chapter 3 presents the dataset, its collection process, the suggested evaluation metrics, and a comparison with existing relevant datasets. Chapter 4 includes details about the approaches that are part of our evaluation whose results are presented in Chapter 5.

1.4 Authorship

I hereby certify that the results presented in this thesis were achieved during my own research, in cooperation with my supervisor Giorgos Tolias and my collaborators Noa Garcia, Guangxing Han, Sarah Ibrahimi and Nanne van Noord, published in [26]. More specifically, in the work presented in this thesis, I and all the aforementioned researchers were involved in the creation of the proposed dataset. The methods used in the experimental evaluation on the respective benchmark were set up and evaluated by me under the guidance of my supervisor Giorgos Tolias.

Chapter 2

Background and related work

The background and related work chapter provides the necessary context for the methods evaluated in this thesis, and is divided into two sections. The first section takes a look at the task of instance-level image retrieval. Core components of methods developed to solve this task are directly applicable to the addressed task of instance-level classification, and some are being used in this work. In the second section, we take a brief look at contrastive learning, the deep metric learning approach used as a training method to train on the proposed dataset.

2.1 Image retrieval

Instance-level image retrieval aims at, given a query image, retrieving all the images from a database that contain the same object instance depicted in the query. The task reduces to calculating the similarity between the representation of the query image and the representations of all the database images and ranking them based on it. For that reason, the choice of the image representation, also called image descriptor, that will be used to encode the image content is crucial. There are two main types of image representations used in image retrieval, global and local ones.

2.1.1 Global and local descriptors

A global descriptor [27, 28, 29] is a mapping of an image to a vector, which serves as a high level semantic signature of the image. It is a compact representation that delivers high retrieval performance and low memory footprint, as it reduces the similarity calculation to a simple nearest neighbor search. On the downside, global descriptors lose information about spatial arrangement of visual elements and often lack the capability to retrieve images with only a partial match [30] or that contain occlusions and background clutter. Before the use of deep learning in image retrieval, global descriptors were constructed by aggreggating hand-crafted local descriptors [31, 27, 32]. With the advent of deep learning, they are produced by global pooling of feature maps of Convolutional Neural Networks (CNNs) that have been either pre-trained for other tasks [28, 33], or trained to optimize them for the task of image retrieval, using ranking [34, 29] or classification losses [35].

Local descriptors encode the visual content of an image in a number (typically in the hundreds) of short vectors. These vectors, which contain information about specific image regions are used for patch-level matching, and are shown to be important for high retrieval precision. Prior to deep learning, they have been extracted using hand-crafted keypoint detectors and descriptors, like SIFT [36] and SURF [37]. With the advent of deep learning however, many works learn local descriptors in a data-driven manner [38, 39]. A typical image retrieval system uses global descriptors for a first search phase, in order to reduce the solution space of the candidate matches, as they produce fast results. In the second phase, local descriptors are used to perform spatial verification in order to re-rank the top matches and produce a new refined list [38, 39].

2.2 Contrastive learning

Deep metric (or similarity) learning is a set of approaches that aims to train deep models to produce embeddings with the following property: embeddings of data samples that are considered "similar" will be close in the representation space according to a distance measure, while data samples considered "dissimilar" will be further away.

Contrastive learning is one of the main approaches of deep metric learning, and is essentially learning by comparing between pairs of data samples. For contrastive learning, "similar" data samples form positive pairs, while "dissimilar" data samples form negative pairs. As so, this learning task requires supervision in the level of pairs of images an and not at the level of individual samples, like in other learning tasks, *e.g.* classification. By contrasting between samples of positive and samples of negative pairs, representations of positive pairs are pulled closer while representations of negative pairs are repulsed far apart in the embedding space. Given the distance metric, contrastive learning boils down to designing proper loss functions in order to achieve this goal. Representative examples of contrastive learning losses are the contrastive loss [40], the triplet loss [41] and the lifted structure loss [42].

2.2.1 Self-supervised contrastive learning

Self-supervised representation learning falls into the unsupervised learning paradigm, meaning no human supervision in the form of annotations is needed. It relies on designing a pseudo-supervised task to be solved, called the pretext task; supervision is provided by the unlabelled data itself. The performance of a model trained to solve this task might not be of immediate interest; however training to solve it can lead to learning intermediate data representations that can be useful for other downstream tasks. Because of that, self-supervised learning is most often used as a pretraining method, before finetuning to the downstream task in hand. Examples of pretext for visual representation learning include image colorization [43], patch relatedness [44], rotation prediction [45] etc.

A particular class of methods for self-supervised learning of visual representations borrows ideas from the contrastive learning framework, and has achieved competitive results as supervised counterparts, with representative examples being SimCLR [46] and

MoCo [47]. The main idea underlying all these methods is that they learn representations that are invariant under different distortions, usually in the form of image augmentations. This is achieved by trying to minimize the distance in the representation space between different augmentations of the same image (positive pairs), while pushing every other image in the dataset far away (negative pairs).

Chapter 3

The Met dataset

In this chapter, the Met dataset is introduced. Firstly, we provide an overview of the dataset. Then, we describe the collection and annotation process of it and provide more statistics and comparisons to other relevant datasets. Last but not least, we describe how we formulate the corresponding benchmark by defining the splits and explaining the evaluation metrics used.

3.1 Dataset overview

The *Met dataset* for ILR contains two types of images, namely *exhibit images* and *query images*. Exhibit images are photographs of artworks in The Met collection taken by The Met organization under studio conditions, capturing multiple views of objects featured in the exhibits. These images form the training set for classification and are interchangeably called exhibit or training images in the following. We collect about 397k exhibit images corresponding to about 224k unique exhibits, *i.e.* classes, also called *Met classes*. Query images are images that need to be labeled by the recognition system, essentially forming the evaluation set. They are collected from multiple online sources for which ground-truth is established by labeling them according to the Met classes. The Met dataset contains about 20k query images, that are divided into the following three types:

- *Met queries*, which are images taken at The Met museum by visitors and labeled with the exhibit depicted
- *other-artwork queries*, which are images of artworks from collections that do not belong to The Met, and
- *non-artwork queries*, which are images that do not depict artworks.

The last two types of queries are referred to as *distractor queries* and are labeled as "distractor" class which denotes out-of-distribution (OOD) queries. They simulate a more realistic recognition setting, and serve to test the robustness of a classifier to OOD inputs. *Other-artwork queries* also add the extra challenge of sharing the same domain with Met queries. Examples of Met query images along with exhibit images from their corresponding Met class are shown in Figure 3.1 (more examples are presented in appendix B). The distribution shift between the training images and the queries is showcased in these

examples. Some causes of this shift are viewpoint and illumination changes and clutter. Example of other-artwork and non-artwork distractor queries are shown in Figures 3.2 and 3.3 respectively.

3.2 Dataset collection

The dataset collection and annotation process is described in the following and summarized in Figure 3.4.

3.2.1 Image sources

Exhibit images are obtained from The Met collection[1]. An example of the interface of The Met collection is shown in Figure 3.5. Only exhibits labeled as open access (public domain) are considered. A maximum of 10 images per exhibit is included in the dataset, images with very skewed aspect ratios are excluded, and image deduplication is performed. Query images are collected from different sources according to the type of query. Met queries are taken on site by museum visitors. Part of them are collected by our team, and the rest are Creative Commons (CC) images crawled from Flickr. We use Flickr groups[2] related to The Met to collect candidate images. Distractor queries are downloaded from Wikimedia Commons[3] by crawling public domain images according to the Wikimedia assigned categories. Generic categories, such as people, nature, or music, are used for non-artwork queries, and art-related categories, *e.g.* art, sculptures, painting, architecture, for other-artwork queries.

3.2.2 Annotation

We label query images with their corresponding Met class, if any. Met queries taken by our team (see Figure 3.6 for examples) are annotated based on exhibit information, whereas Met queries downloaded from Flickr are annotated in three phases, namely filtering, annotation, and verification. In the filtering phase (Figure 3.7), invalid images are discarded, *i.e.* images containing visitor faces, images not depicting exhibits, or images with more than one exhibit. In the annotation phase (Figure 3.8), queries are labeled with the corresponding Met class. To ease the task, the title and description fields on Flickr are used for text-based search in the list of titles from The Met exhibits included in the corresponding metadata. We use two text-based engines: an automatic scoring system based on bag-of-words and the manual search engine provided by The Met (Figure 3.9).[4] Queries whose depicted Met exhibit is not in the public domain are discarded. Finally, in the verification phase (Figure 3.10), two different annotators verify the correctness of the labeling per query. We additionally verify that distractor queries, especially other-artwork queries, are true distractors and do not belong to The

[1] https://www.metmuseum.org/

[2] https://www.flickr.com/groups/metmuseum/, https://www.flickr.com/groups/themet/, https://www.flickr.com/groups/mma_aaoa/

[3] https://commons.wikimedia.org/wiki/Main_Page

[4] https://www.metmuseum.org/art/collection/search

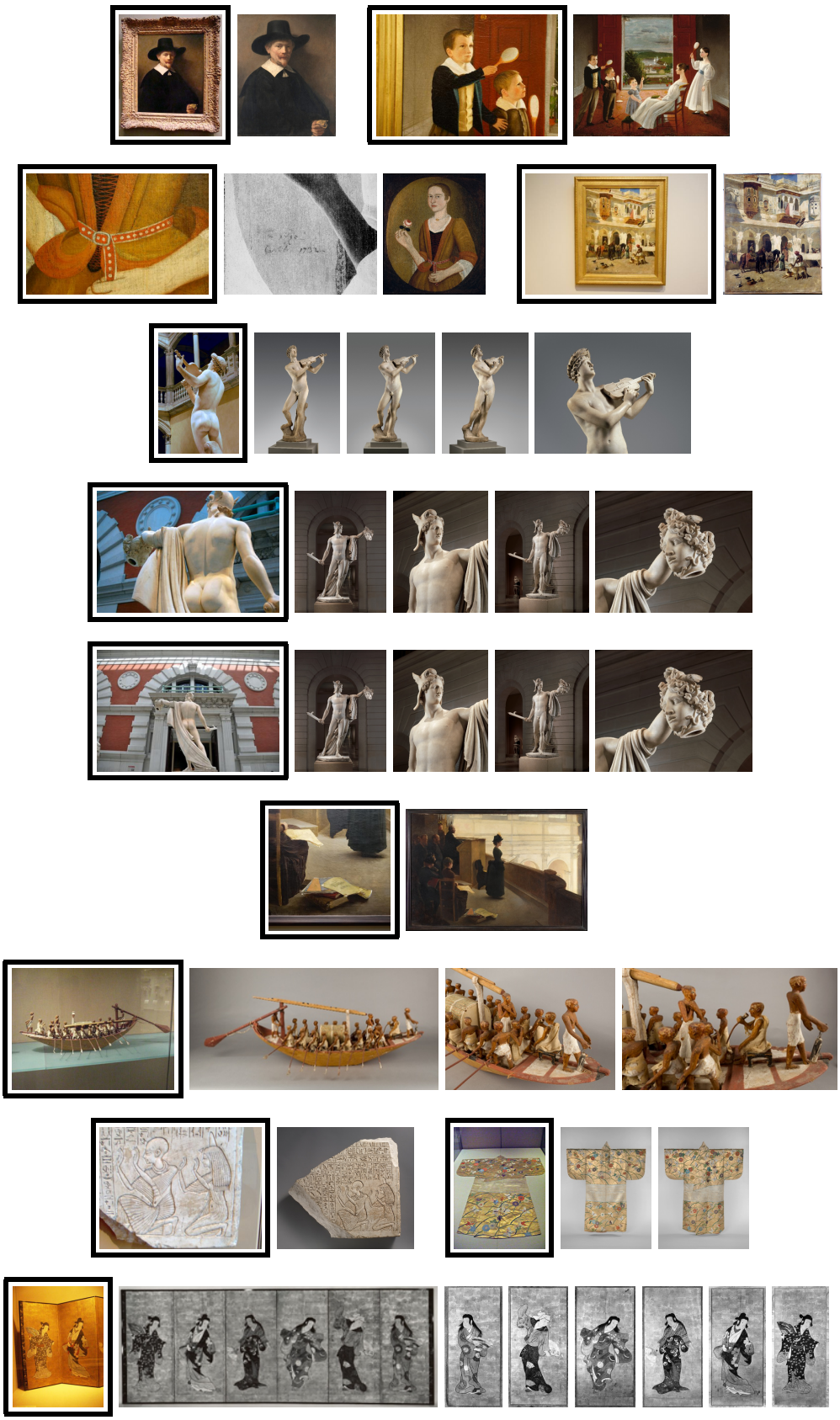

Figure 3.1. *Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.*

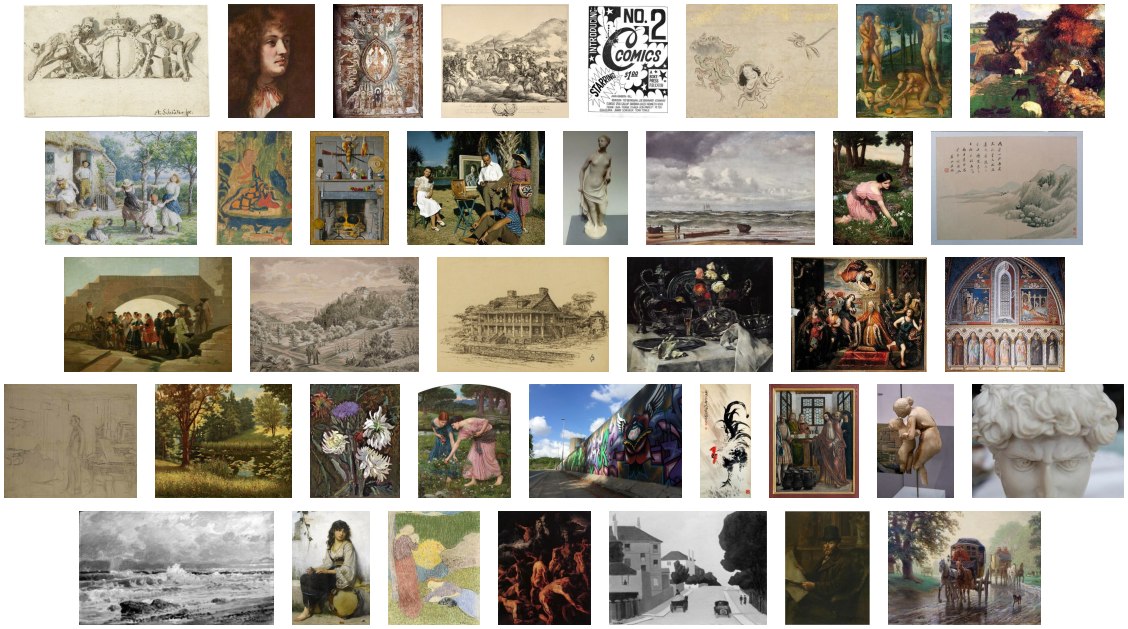

Figure 3.2. *Examples of other-artwork distractor queries. These distractor queries depict artworks that do not belong to the Met collection.*

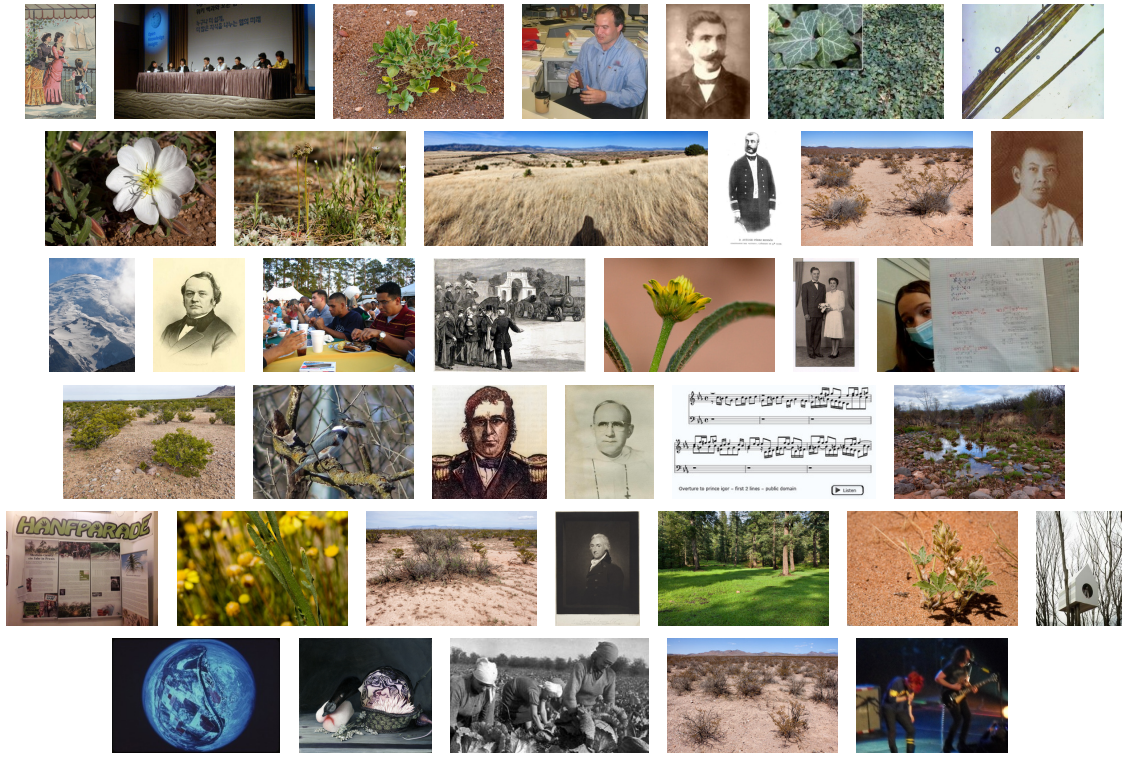

Figure 3.3. *Examples of non-artwork distractor queries. These distractor queries do not depict artworks.*

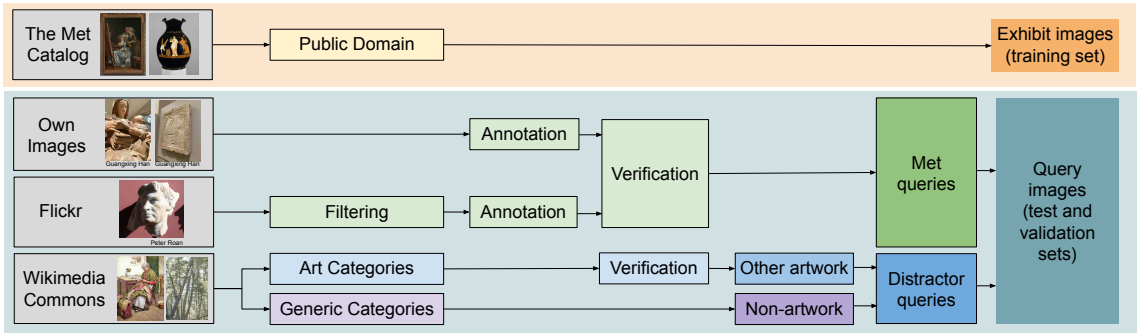

Figure 3.4. *An overview of the Met dataset collection and annotation process.*

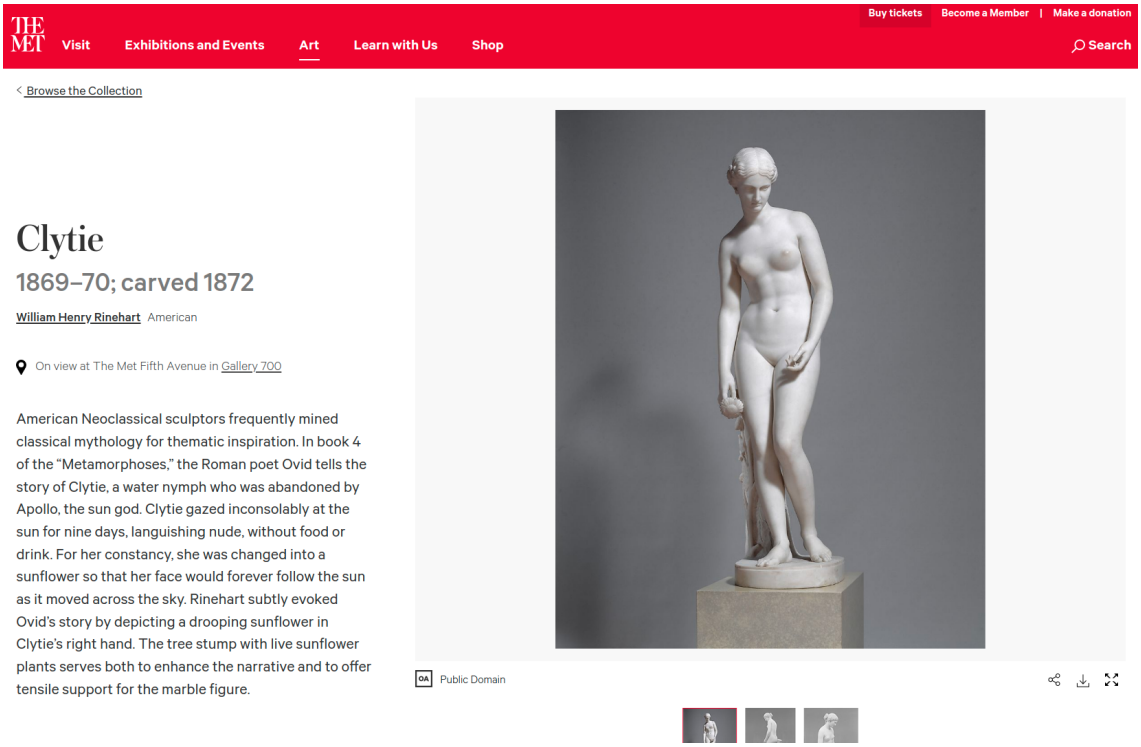

Figure 3.5. *An example of using the interface of The Met collection website.*

Figure 3.6. *Examples of Met queries captured by our team.*

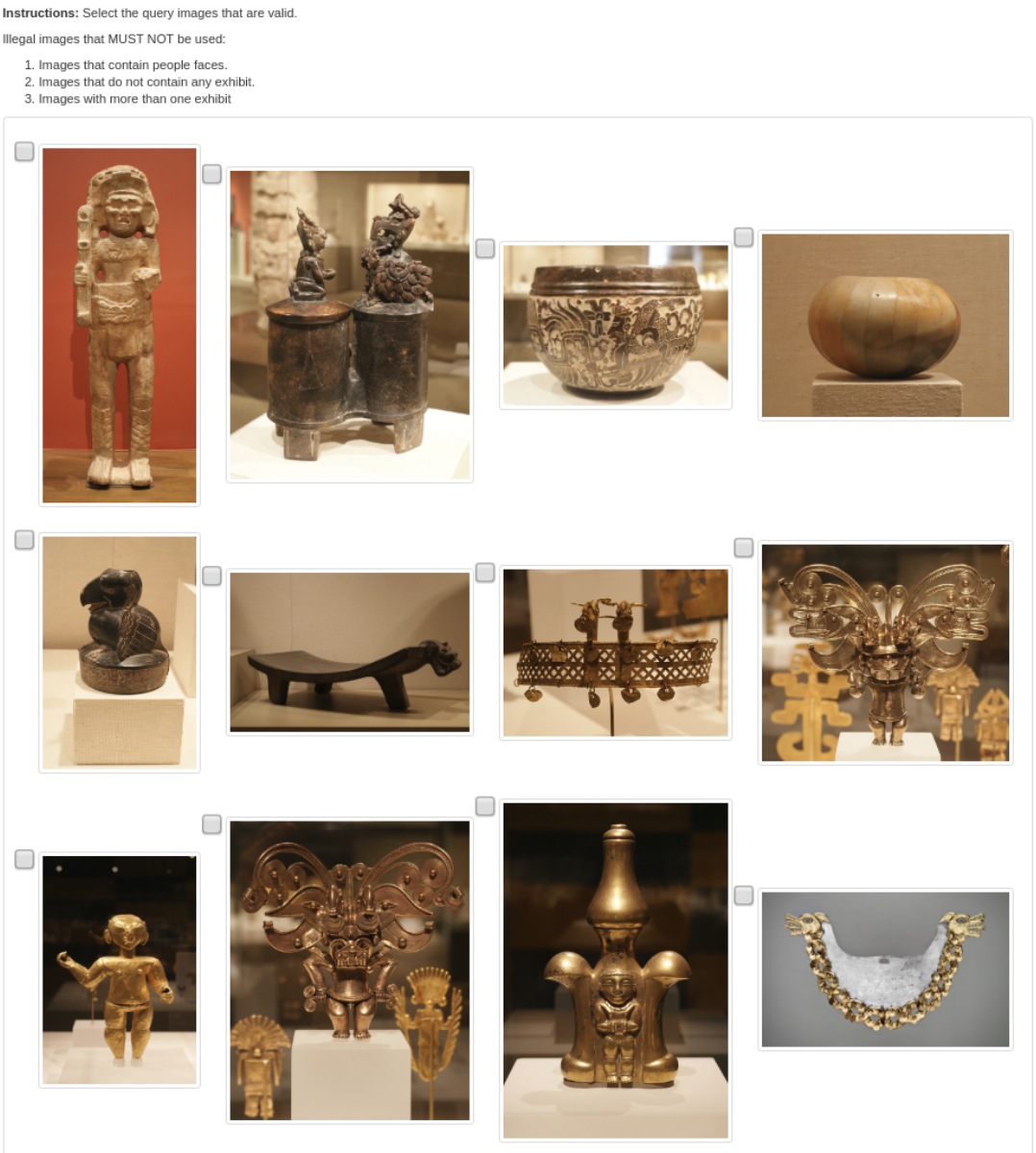

Figure 3.7. *Example annotation step during the filtering stage of the annotation process. In this phase, invalid images are discarded, i.e. images containing visitor faces, images not depicting exhibits, or images with more than one exhibit.*

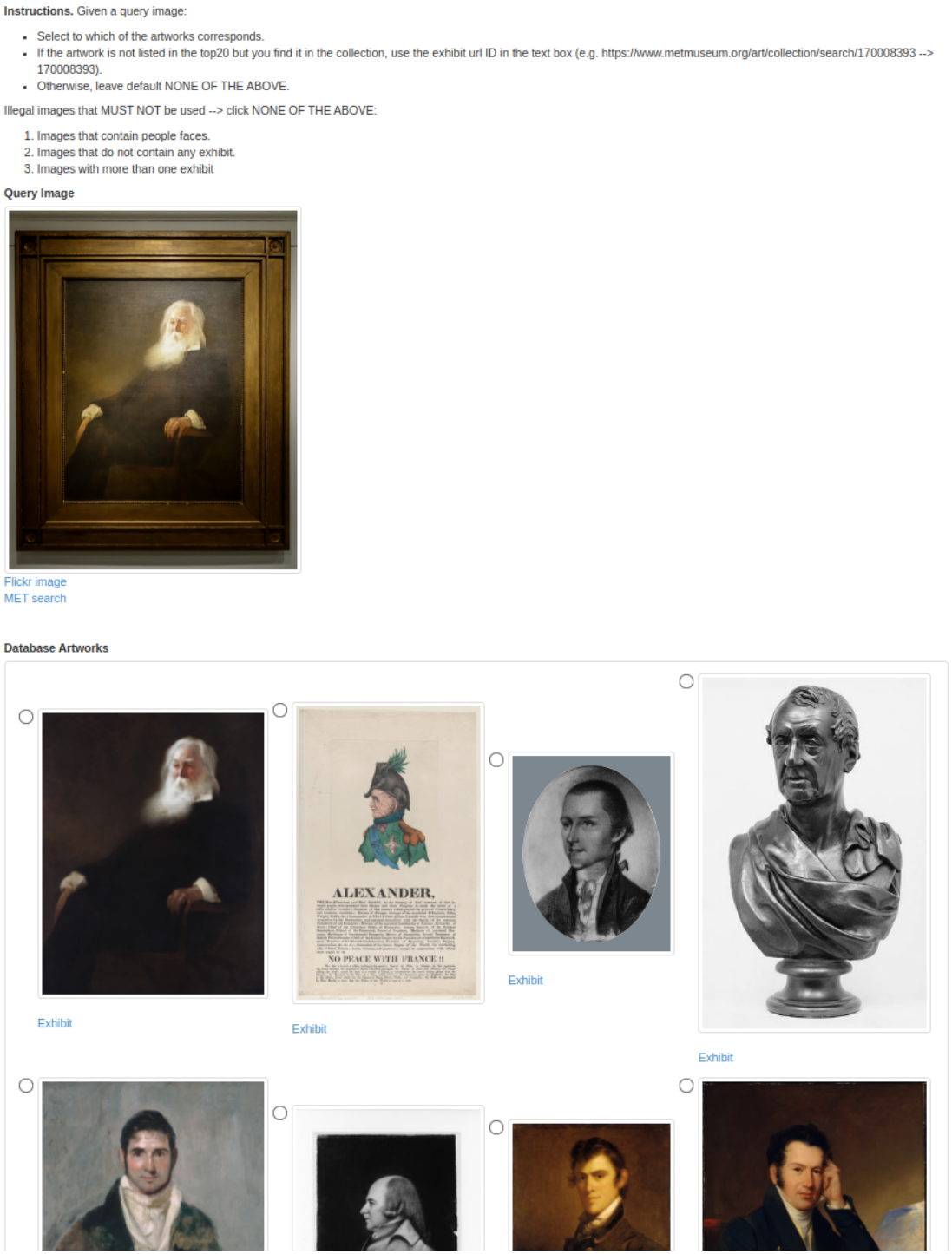

Figure 3.8. *Example annotation step during the annotation stage of the annotation process. In this phase, queries are labeled with the corresponding Met class.*

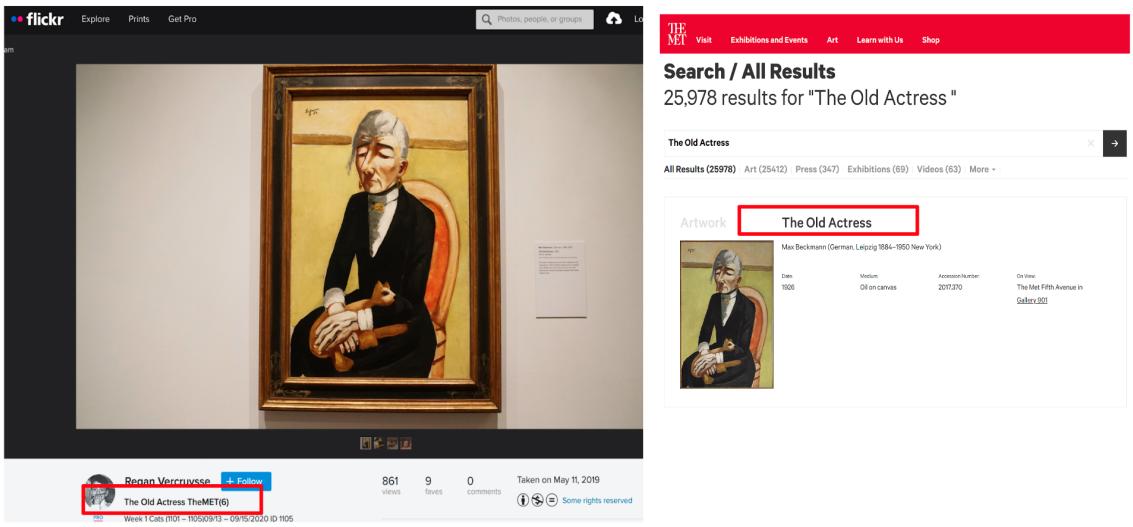

Figure 3.9. *Example of text-based search on the manual search engine provided by the Met.*

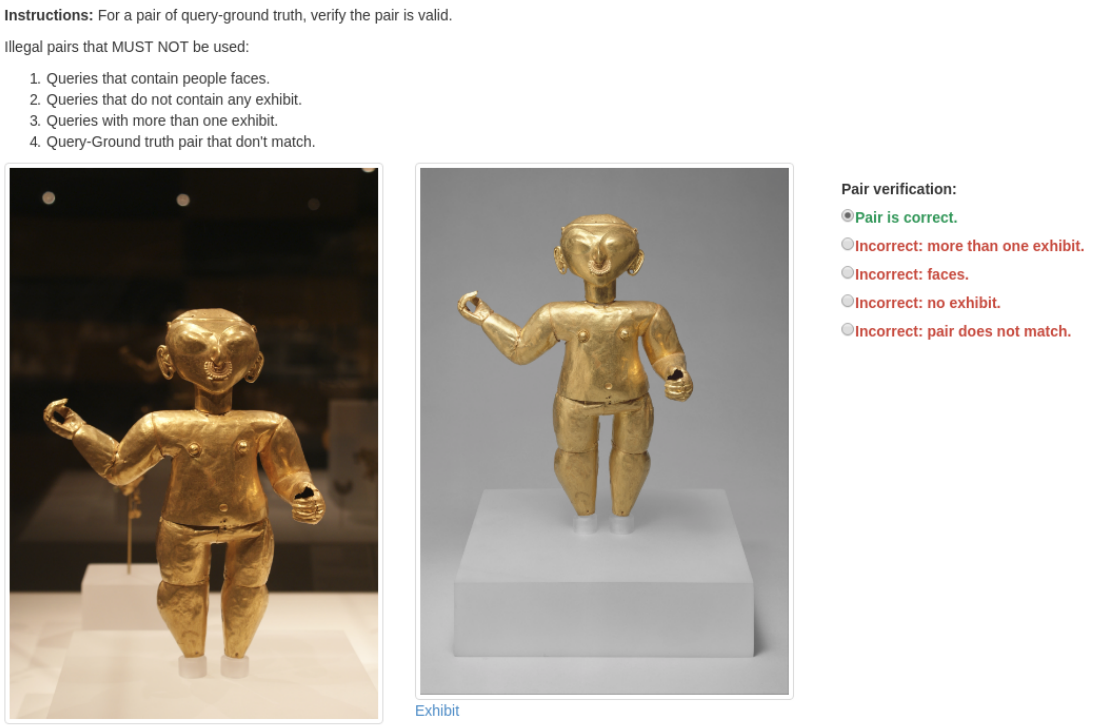

Figure 3.10. *Example annotation step during the verification stage of the annotation process. In this phase, annotators verify the correctness of the labeling per query.*

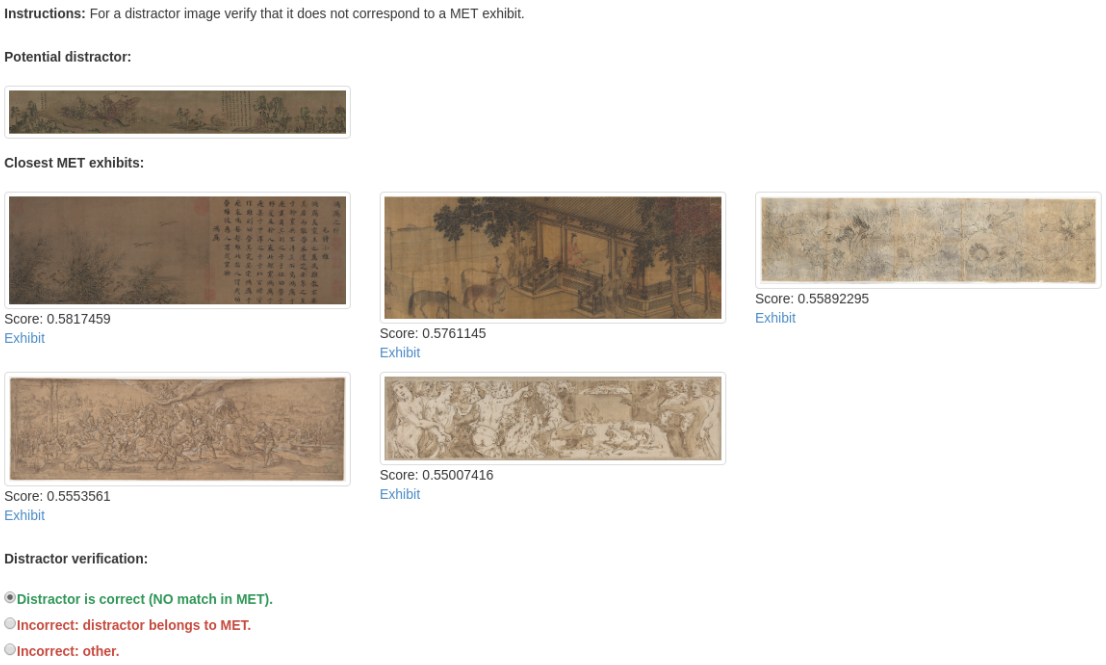

Figure 3.11. *Example annotation step during the distractor verification stage of the annotation process. In this phase, annotators verify that other-artwork distractor queries are true distractors and do not belong to The Met collection.*

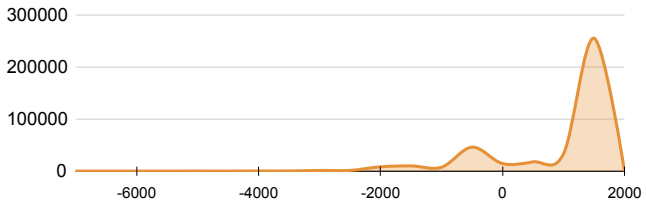

Figure 3.12. *Number of exhibit images per time period.*

Met collection. This is done in a semi-automatic manner supported by (i) text-based filtering of the Wikimedia image titles and (ii) visual search using a pre-trained deep network. Top matches are manually inspected and images corresponding to Met exhibits are removed (Figure 3.11).

3.3 Dataset statistics

The Met dataset contains artworks spanning from as far back as 240,000 BC to the current day. Figure 3.12 shows a smoothed histogram of the number of exhibit images by creation year, grouped in bins of 500 years. More than half of the exhibits were created between 1,500 AD and 1,999 AD, with a remarkable number of ancient artworks created between 500 BC and 1 BC. Figure 3.13 shows the distribution of classes and images according to The Met department. Whereas there is an imbalance for exhibits across The Met departments, queries are collected to be evenly distributed to the best of our capabilities. In this way, we aim to ensure models are not biased towards a specific type of art, i.e., developing models that only produce good results for, e.g., European paintings,

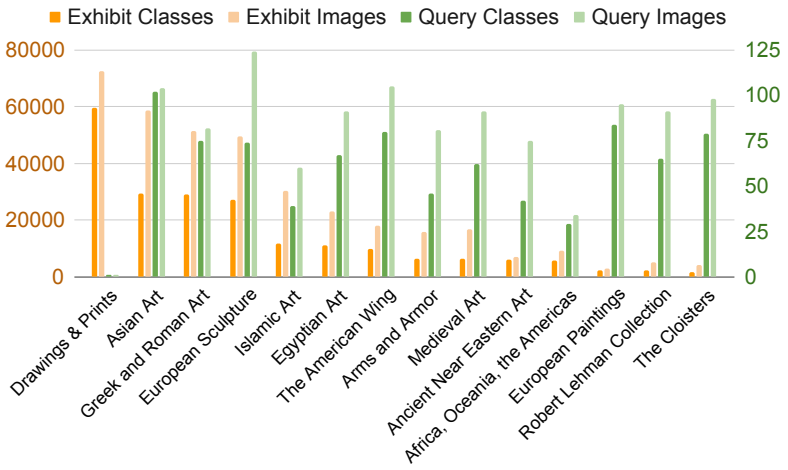

Figure 3.13. *Number of images and classes by department. Met queries are assigned to the department of their ground-truth class. Some departments that do not contain queries but contain exhibit images are not shown.*

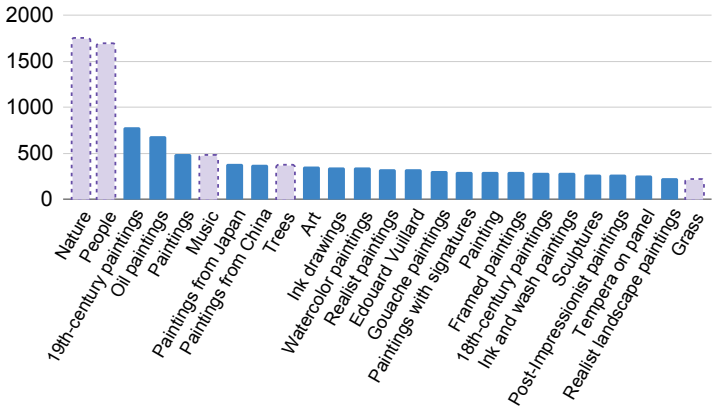

Figure 3.14. *Number of distractor images by Wikimedia category. Top categories shown: art-related categories in solid blue and generic categories in dash purple.*

will not necessarily ensure good results on the overall benchmark. Finally, Figure 3.14 shows the number of distractor query images by Wikimedia Commons categories. The class frequency for exhibit images ranges from 1 to 10, with 60.8% and 1.2% classes containing a single and 10 images, respectively (see Figure 3.15 left), indicating the long-tail distribution of the training set. Met queries are obtained from 39 visitors in total, while the maximum number of query images per class is, coincidentally, also 10. In total, 81.5% of the Met query images are the sole Met queries that depict a particular Met class (see Figure 3.15 right). The number of photographers versus the Met queries that belong to them is shown in Figure 3.16.

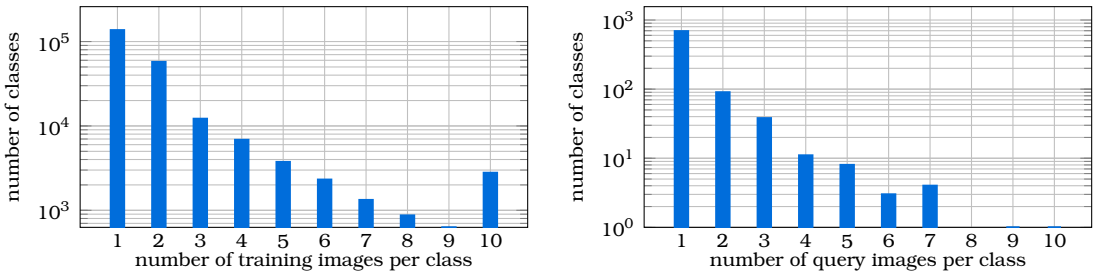

Figure 3.15. *Left: number of Met classes versus number of training images per class. Right: number of Met classes versus number of query images per class. Both vertical axes are shown in logarithmic scale, for better visualization.*

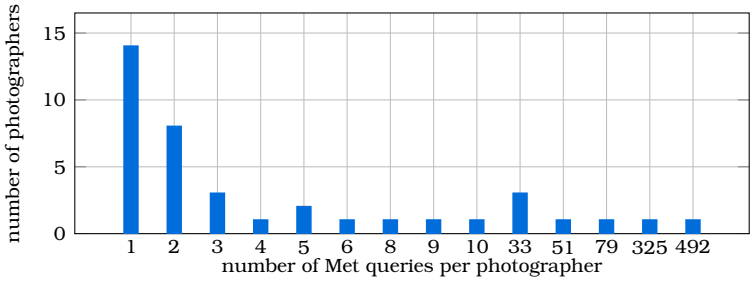

Figure 3.16. *The number of photographers versus the Met queries that belong to them.*

3.4 Comparison to other datasets

In this section, we provide a comparison of the Met dataset to other existing datasets that are relevant in terms of domain (art) or task (ILR).

3.4.1 Artwork datasets

Table 3.1 summarizes datasets in the artwork domain for various tasks. Most of the artwork datasets [14, 15, 16, 17, 18] focus on attribute prediction (AP), containing multiple types of annotations, such as author, material, or year of creation, usually obtained directly from the museum collections. Other datasets [48, 19, 18, 49] are focused on category-level recognition (CLR), aiming to recognize object categories, such as animals and vehicles, in paintings. From the artwork datasets, Open MIC [50] and NoisyArt [10] are the only ones with instance-level labels. Compared to the Met dataset, the Open MIC is smaller, with significantly less classes and mostly focuses on domain adaptation (DA) tasks. NoisyArt has a similar focus to ours, but is significantly smaller, and has noisy labels.

3.4.2 ILR datasets

In Table 3.2 we compare the Met dataset with existing ILR datasets in multiple domains. ILR is widely studied for clothing [8, 9], landmarks [6], and products [7, 1]. The

Art datasets	Year	Domain	# Images	# Classes	Type of annotations	Task	Image source
PrintArt [48]	2012	Prints	988	75	Art theme	CLR	Artstor
VGG Paintings [19]	2014	Paintings	8,629	10	Object category	CLR	Art UK
WikiPaintings [14]	2014	Paintings	85,000	25	Style	AP	WikiArt
Rijksmuseum [16]	2014	Artwork	112,039	†6,629	Art attributes	AP	Rijksmuseum
BAM [18]	2017	Digital art	65M	†9	Media, content, emotion	AP, CLR	Enhance
Art500k [15]	2017	Artwork	554,198	†1,000	Art attributes	AP	*Various*
SemArt [51]	2018	Paintings	21,383	21,383	Art attributes, descriptions	Text-image	Web Gallery of Art
OmniArt [17]	2018	Artwork	1,348,017	†100,433	Art attributes	AP	*Various*
Open MIC [50]	2018	Artwork	16,156	866	Instance	ILR (DA)	*Authors*
iMET [49]	2019	Artwork	155,531	1,103	Concepts	CLR	The Met
NoisyArt [10]	2019	Artwork	89,095	3,120	Instance (noisy)	ILR	*Various*
The Met (Ours)	2021	Artwork	418,605	224,408	Instance	ILR	*Various*

Table 3.1. *Comparison to art datasets.* † *For datasets with multiple kinds of annotations, the task with the largest number of classes is reported.*

ILR datasets	Year	Domain	# Images	# Classes	Type of annotations	Image source
Street2Shop [8]	2015	Clothes	425,040	204,795	Category, instance	*Various*
DeepFashion [9]	2016	Clothes	800, 000	33,881	Attributes, landmarks, instance	*Various*
GLD v2 [6]	2019	Landmarks	4.98M	200,000	Instance (noisy)	Wikimedia
AliProducts [1]	2020	Products	3M	50,030	Instance (noisy)	Alibaba
Products-10K [7]	2020	Products	150,000	10,000	Category, instance	JD.com
The Met (Ours)	2021	Artwork	418,605	224,408	Instance	*Various*

Table 3.2. *Comparison to instance-level recognition datasets.*

Met dataset resembles ILR datasets in those domains in that the training and query images are from different scenarios. For example, in Street2Shop [8] and DeepFashion [9] queries are taken by customers in real-life environments, whereas training images are studio shots. Getting annotations for ILR, however, is not easy, and some datasets contain a significant number of noisy annotations from crawling from the web without verification [1, 10, 6]. In that sense, the Met is the largest ILR dataset in terms of number of classes, which have been manually verified. Overall, the Met dataset proposes a large-scale challenge in a new domain, encouraging future research on generic ILR approaches that are applicable in a universal way to multiple domains.

3.5 Benchmark and evaluation protocol

3.5.1 Splits

The structure and evaluation protocol for the Met dataset follows that of the Google Landmarks Dataset (GLD) [6]. All Met exhibit images form the training set, while the query images are split into test and validation sets. The test set is composed of roughly 90% of the query images, and the rest is used to form the validation set. To ensure no leakage between the validation and test split, all Met queries are first grouped by user and then assigned to a split. Additionally, we enforce that there is no class overlap between the splits. As a result, 25 (14) users appear only in the test (validation) split, respectively. Image and class statistics for the train, val, and test sets are summarized in Table 3.3. The intended use of the validation split is for hyper-parameter tuning. All images are resized to have maximum resolution 500×500.

Split	Type	# Images			# Classes
		Met	other-art	non-art	
Train	Exhibit	397,121	-	-	224,408
Val	Query	129	1,168	868	111 + 1
Test	Query	1,003	10,352	7,964	734 + 1

Table 3.3. *Number of images and classes in the Met dataset per split. Met exhibits images are from the museum's open collection, while Met query images are from museum visitors. Query images contain distractor images too (denoted by the +1 class) while the rest of val/test classes are subset of the train classes.*

3.5.2 Metrics

For the evaluation of a proposed method on the proposed benchmark we measure classification performance with two standard ILR metrics, namely average classification accuracy (ACC), and Global Average Precision (GAP). The average classification accuracy is measured only on the Met queries, as there is no explicit modelling of distractor class. It is equal to the ratio of the correctly classified Met queries to the total number of Met queries. GAP, also known as Micro Average Precision (μAP) [6], is measured on all queries taking into account both the predicted label and the prediction confidence. In order to calculate it, all queries (Met + distractor) are ranked according to their assigned prediction confidence in descending order, and then average precision is estimated on this ranked list; predicted labels and ground-truth labels are used to infer correctness of the prediction, while distractors are always considered to have incorrect predictions. It is given by

$$GAP = \frac{1}{M} \sum_{i=1}^{T} p(i) r(i), \tag{3.1}$$

where $p(i)$ is the precision at position i, $r(i)$ is a binary indicator function denoting the correctness of prediction at position i, M is the number of the Met queries, and T is the total number of queries. An example of calculating GAP for a set of predictions and their assigned confidences is shown in the first two rows of Fig. 3.17. The GAP score is equal to the area-under-the-curve of the precision-recall curve whilst jointly taking all queries into account. We measure this for the Met queries only, denoted by GAP^{-}, and for all queries, denoted by GAP. In contrast to accuracy, this metric reflects the quality of the prediction confidence as a way to detect out-of-distribution (distractor) queries and incorrectly classified queries. It allows for inclusion of distractor queries in the evaluation without the need for distractors in the learning; the classifier never predicts "out-of-Met" (distractor) class. Optimal GAP requires, other than correct predictions for all Met queries, that all distractor queries get smaller prediction confidence than all the Met queries (last two rows of Fig. 3.17).

Figure 3.17. *First two rows: example of calculating ACC and GAP for a set of queries and their corresponding predictions and confidences. In the first row, ACC calculation is shown. It is calculated only on the Met queries, which are either correctly (green) or incorrectly classified (red). In the second row, for the GAP calculation, which takes into account distractors as well (grey), the queries are sorted by confidence in descending order. The confidence, the precision and the binary indicator of correctness at each rank of the sorted list is shown. Then, calculation of GAP is straightforward. Last two rows: Ways to incrementally achieve optimal GAP starting from the example of the top rows: 3rd row) to make as many predictions correct for the non distractor queries, 4th row) to place missclassified examples, which include distractors, in the bottom of the list*

Chapter 4

Methods

In this chapter, we present a number of existing methods that are applicable to the Met dataset, in the experimental evaluation. Firstly, we present the image representation, given a fixed - already trained - deep embedding network. Subsequently, we describe a simple non-parametric classifier that uses such image representation. Then, we list a set of networks trained for other tasks which we use to extract the image representation and test their potential for transfer learning. Finally, we present different ways of training using the Met training set, including end-to-end learning of a Deep Network classifier, self-supervised representation learning with image augmentations and contrastive learning with a Siamese architecture taking advantage of the Met labels.

4.1 Image representation

In this section we explain how we obtain the image representation, also called image descriptor, that is used to perform classification. In all of our baseline approaches, we choose to use global image representations obtained by fully-convolutional networks.

Convolutional Neural Networks (CNNs) that are used for the task of classification typically contain fully-connected layers after the convolutional layers, in order to transform the feature maps of the final convolutional layer into a vector of logits. By removing the fully-connected layers, we can obtain the fully-convolutional part of the architecture, which we call the fully-convolutional network (FCNs). The output of a FCN is a 3D tensor, which can be equivalently seen as a set of feature maps with cardinality equal to the number of filters in the last convolutional layer of the FCN. In order to map this 3D tensor to a vector, we perform global pooling of the feature maps. The global pooling operation that we use is Generalized-Mean (GeM) pooling [29], shown to be effective for representation in instance-level tasks [39]. More specifically, for a 3D tensor of shape $C \times H \times W$ produced by a FCN, where C is the number of feature maps and H, W are the spatial dimensions, the i-th component of the global descriptor produced by the GeM pooling operation is defined as:

$$f_i = \left(\frac{1}{|C_i|} \sum_{c \in C_i} c^p \right)^{\frac{1}{p}} \tag{4.1}$$

where C_i is the i-th feature map, and p is a learnable parameter of the pooling layer. Finally, we ℓ_2 normalize the global descriptor. The FCN followed by the GeM pooling

operation and ℓ_2 normalization is called the backbone. It can be equivalently seen as an embedding function $f_\partial : \mathcal{X} \to \mathbb{R}^d$ that takes an input image $x \in \mathcal{X}$ and maps it to a vector $f_\partial(x) \in \mathbb{R}^d$, equivalently denoted by $f(x)$. The backbone is parametrized by the parameter set ∂. ResNet18 (R18) and ResNet50 (R50) [52] are the FCNs comprising the backbones used in this work, producing 512D and 2048D descriptors, respectively.

4.2 kNN classification

There are some extra processing steps that we use to improve the representation used with the kNN classifier, which is described afterwards. First of all, we follow the convention [53, 39] and use an image-pyramid at test time to produce multi-scale image representations. More specifically, representation of image x, denoted by vector embedding $\mathbf{v}(x) \in \mathbb{R}^d$, is a result of aggregation of multi-resolution embeddings given by

$$\mathbf{v}(x) = \frac{\sum_{r \in R} f(x_r)}{\| \sum_{r \in R} f(x_r) \|}, \tag{4.2}$$

where x_r denotes image x down-sampled by relative factor r. We set $R = \{1, 2^{-0.5}, 2^{-1}\}$ and $R = \{1\}$ in the *multi-scale* (MS) and *single-scale* (SS) case, respectively. Also, following the standard practice in instance-level search, the image representation space is whitened with PCA whitening (PCAw) [54] learned on the representation vectors of all Met training images. Optionally, dimensionality reduction is performed by keeping the dimensions corresponding to the top components. PCAw is always performed in the rest of this work, unless stated otherwise; for simplicity we reuse notation $\mathbf{v}(x)$ for the whitened image embeddings.

Next, we describe how we perform k-Nearest-Neighbor (kNN) classification using the image representation obtained. Let $y(x)$ be the label of image x, and let q be a query image. The similarity between query q and a training image x is given by $\mathbf{v}(x)^\top \mathbf{v}(q)$, coinciding with the cosine similarity in our case, where we use ℓ_2 normalized image representations. The confidence of class c for query q is given by

$$s_c(q) = \max_{x \in \mathrm{NN}_k(q)} (\mathbf{v}(x)^\top \mathbf{v}(q)) \mathbb{1}_{y(x)=c}, \tag{4.3}$$

where $\mathrm{NN}_k(q)$ is the set of k nearest-neighbors (kNNs) of q in the d-dimensional representation space. The vector of class confidences is $\mathbf{s}(q) \in \mathbb{R}^N$ with elements $s_c(q), c \in [1, \ldots, N]$, where N is the number of training classes. Classes without any example in the top-k neighbors obtain zero confidence. An example of calculating the vector of class confidences is depicted in Figure 4.1. The predicted label $\hat{y}(q) = \arg\max_c s_c(q)$ is, according to (4.3), equivalent to the label of the closest training image. Despite label prediction requiring only $k = 1$, confidence estimation for more classes is essential for normalization and handling of OOD (distractor) queries. The normalized confidence is given by the soft-max of vector $\tau \mathbf{s}(q)$, where τ is the temperature. As it is a non-parametric classifier, it does not require training on the Met dataset; only hyper-parameters k and τ are tuned with grid search according to GAP on the validation set. A schematic of the pipeline described to

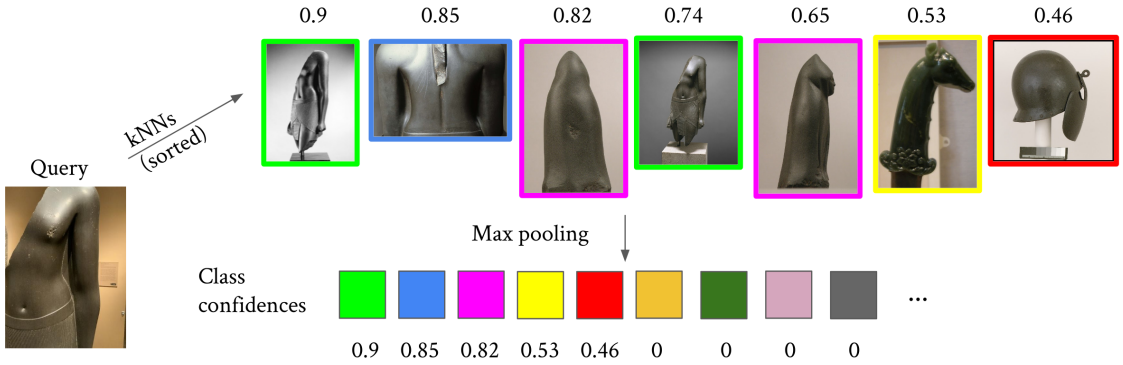

Figure 4.1. *Example of the calculation of class confidences for a query image using the proposed kNN classifier. The similarity values in this example are fictional and only exist for the purposes of the figure.*

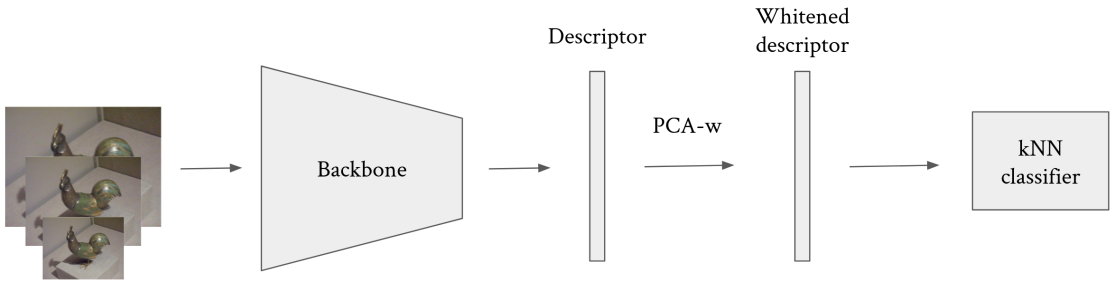

Figure 4.2. *Overview of the pipeline for kNN classification. The backbone produces the image descriptor after processing the image at multiple scales. It is subsequently whitened by PCA-whitening and used as the input to the kNN classifier.*

perform classification with the kNN classifier is shown in Figure 4.2

4.3 Pretrained models

We consider networks pretrained on different tasks and use them to obtain the image embeddings used for kNN classification. This allows us to evaluate the transfer learning potential of the pretraining task. In all the learned backbones, GeM pooling is used (replacing global average pooling of ResNet) to produce the image embedding. We list them below and give a brief overview for each one of them.

ImageNet (IN) - classification

Approach for training on ImageNet with cross-entropy loss [52], commonly used as a standard pretraining step for other tasks.

Landmarks (SfM) - metric learning

Approach for metric learning with contrastive loss on image pairs obtained from Structure-from-Motion on landmarks [29]. The image embedding uses GeM for global

pooling during the pretraining phase too.

Artwork attributes (SemArt)

Networks trained on the SemArt dataset [51] by Garcia *et al.* [55] for artwork attribute prediction. In particular, we consider variants for painting type (10 classes) or author (350 classes). Note that SemArt consists only of paintings.

StylizedImageNet (SIN)

Network trained by Geirhos *et al.* [56] on a stylized version of ImageNet to improve the texture bias of deep networks which is shown to demonstrate better transferability.

SwAV on ImageNet (IN) - self-supervision

Representation learning on ImageNet with self-supervision by instance discrimination. The resulting network has achieved good results in concept generalization [57].

Semi-weakly supervised (SWSL) on Instagram 1B + ImageNet

Teacher-student approach [58] with teacher pretrained on about 1 billion images with hashtags and student trained with teacher-generated pseudo-labels, eventually fine-tuned on ImageNet.

4.4 Training on the Met

We now turn to training on the Met dataset for the first time. Firstly, we use the training set of the Met dataset to train a Deep Network classifier for the Met classes. Better results of the kNN classifier compared to the Deep Network one (as shown in the next chapter), while using the same embeddings guides us to perform representation learning by training the backbone to obtain better image descriptors to use with the kNN classifier. Parameters θ of the backbone are initialized by the result of pretraining on ImageNet for classification, unless stated otherwise.

4.4.1 Learning with a classification objective

As the task that we are trying to solve is a classification task, we first proceed to solve it by explicitly training to minimize classification losses. Classifiers that are trained with classification losses use a weight matrix, where each row corresponds to a class. By performing multiplication of the descriptor of an input image with this matrix, a vector of logits is produced which is then fed as input to the classification loss, along with the ground truth. Below, we explain the architecture we use in order to train with two different kinds of classification losses.

Deep network (DNet) classifier with instance-level labels

The Deep Network (DNet) classifier that we use in this work consists of the backbone explained in the previous section, followed by a cosine similarity classifier. The cosine similarity classifier is a linear classifier with ℓ_2 normalized rows and no bias term. As a consequence, each row represents a learnable class prototype, and we end up essentially calculating the cosine similarity between the class prototypes and the descriptor of the input image, extracted from the backbone part of the DNet classifier (which already is ℓ_2 normalized). Such kind of classifiers have been used previously for training with imbalanced datasets [59] and for face verification tasks [60]. The beneficial post-processing step of PCA-whitening used with the kNN classifier can not be used while training this classifier end-to-end. However, as PCA-whitening is a linear operation, it can be modeled by a fully-connected (FC) layer [34] after the backbone part of the DNet classifier. This layer is set to be trainable and initialized with the result of PCA whitening learned on the training set of the Met dataset (from the pretrained descriptors), as done in [29]. The whole pipeline (see Figure 4.3 for a schematic) is differentiable, so it is trained jointly end-to-end.

We perform training by minimizing one of the two following losses. Firstly, standard Cross-Entropy (CE) loss with soft-max, used as a standard objective in classification settings. In our case, the input to the soft-max (logit vector) is equal to the cosine similarity between the backbone output and the learnable class prototypes. The logit vector is additionally multiplied by temperature γ, which is needed as the logits are bounded at the interval $[-1, 1]$, hindering the network from converging [60]. Secondly, we use the Arc-Face (AF) loss [61], which is also used in the work of Cao *et al.* [39] for instance-level recognition of landmarks. The AF loss is a more discriminative version of the CE loss, where an angular margin penalty is added between the ground truth class prototype and the input descriptor in order to boost the intra-class compactness and enhance the inter-class separability. For one sample it is given by:

$$L = -\log\left(\frac{\exp\left(\gamma \times ACS\left(w_k^T f, 1\right)\right)}{\sum_n \exp\left(\gamma \times ACS\left(w_n^T f, y_n\right)\right)}\right) \tag{4.4}$$

where w_k refers to the k-th row of the weight matrix (k-th learnable class prototype), f is the ℓ_2 normalized input descriptor, y is the ground-truth label, k is the ground-truth class index ($y_k = 1$) and γ is the temperature mentioned above. ACS denotes the Adjusted Cosine Similarity and it is calculated as:

$$ACS(s, c) = (1 - c) \times s + c \times \cos(\text{acos}(s) + m) \tag{4.5}$$

where s is the cosine similarity, m is the angular margin penalty, and c is a binary value that denotes whether it is the ground-truth category or not. Index n indexes all the Met classes.

During training with any of these two losses, two things are learned. On one side, a better descriptor for the input image, as the backbone's weights are being updated. Also,

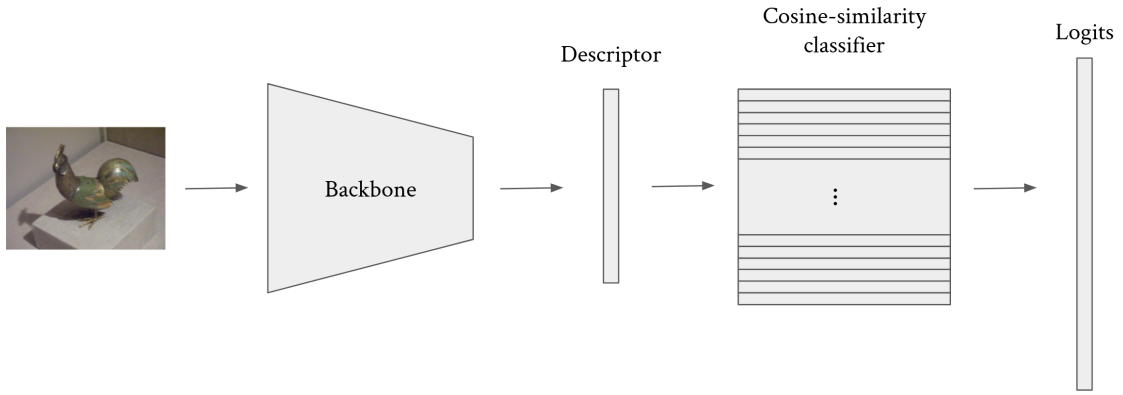

Figure 4.3. *The schematic of the DNet classifier. The backbone part of the classifier produces the image descriptor, which is subsequently fed to the cosine similarity classifier that produces the vector of logits. The latter contains the similarity of the image descriptor with every Met class. The entire pipeline is amenable to end-to-end training.*

the learnable class prototypes of the cosine similarity classifier distribute themselves on the unit hypersphere (of dimension same as the one of the descriptor) in such a way as to be representative enough of the intra-class variability, while also being maximally discriminative. During inference two options are considered. First, use the whole deep network classifier and consider its arg max and max as class prediction and confidence score, respectively. For the estimation of the confidence, we also tune a seperate temperature hyperparameter on the validation set, after training, in order to improve the GAP metric. Second, discard the cosine similarity classifier and use the backbone $f_\theta(\cdot)$ (along with the FC layer) to obtain the image representation $v(x)$ and make predictions with the kNN classifier.

4.4.2 Representation learning for the kNN classifier

We turn to representation learning to produce better image descriptors to use with the kNN classifier. For all variants described next, the optional FC layer that was described in the previous section is included in the backbone and initialized with the result of PCA whitening. The fact that most Met classes in the training set are represented by a single image make it impossible to use standard deep metric learning approaches [62], which rely on the formation of positive pairs. So, despite the availability of training labels in the considered task, a self-supervised approach becomes relevant.

Simple-siamese (SimSiam) representation learning

We apply the recent self-supervised representation learning approach by Chen and He [63] to train the backbone. Each image x is augmented twice resulting in the positive pair of x_1 and x_2. An example pair used with this method is shown in the first row of Figure 4.5. The architecture used consists of the backbone, and two Multi-Layer Perceptrons (MLPs), denoted by functions $h_{\omega_1} : \mathbb{R}^d \to \mathbb{R}^d$ and $g_{\omega_2} : \mathbb{R}^d \to \mathbb{R}^d$, called predictor and projector respectively. The loss maximizes the cosine similarity between

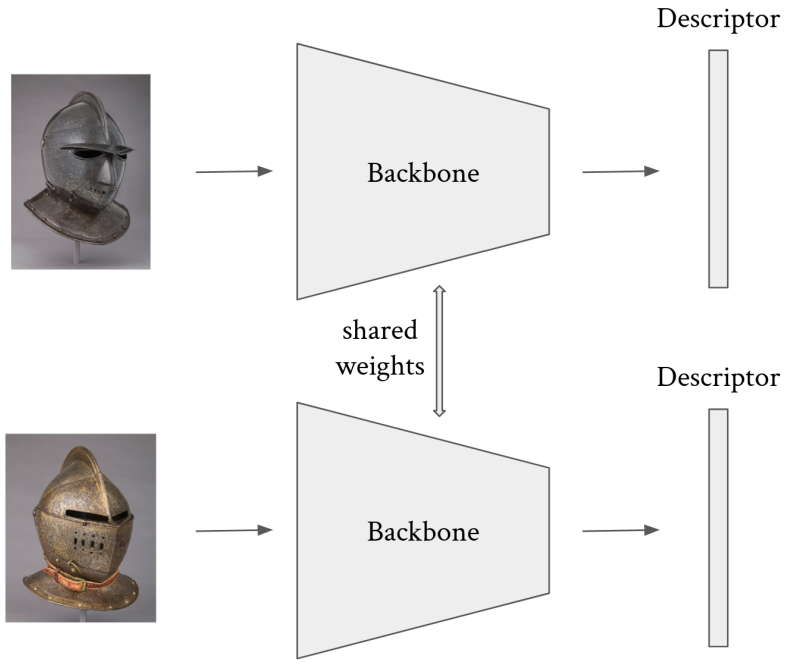

Figure 4.4. *The Siamese architecture used along with the contrastive loss. A pair of images (in this example, a negative pair) is fed into the same backbone, producing their representations. These are then fed to the contrastive loss, along with their corresponding label.*

$h_{\omega_1}(g_{\omega_2}(f_\theta(x_1)))$ and $g_{\omega_2}(f_\theta(x_2))$. The key ingredient for this method to work is to block the back-propagation of gradients in the branch related to x_2. There is no need for negatives samples, making very large batch sizes [46] and a memory bank of negative samples [47] unnecessary. During inference, only the backbone (along with the FC layer) is used to obtain the image representation and predictions are made with the kNN classifier.

Contrastive loss with synthetic/real positives and hard negatives

We now proceed with contrastive learning that makes use of label supervision that is provided with the Met dataset, by adding it in an incremental fashion. More specifically, the backbone is trained with contrastive loss [40]. This is achieved by means of a Siamese architecture [64], which is essentially two-branches of the same backbone, sharing the same parameter set (see Figure 4.4). This architecture produces the descriptors of the pair of images that are fed as its input, and feeds those descriptors to the loss function, along with the label, indicating if it is a positive (same class) or a negative pair (different class). The contrastive loss is minimized when positive pairs share the same descriptor in the representation space, while negative pairs are apart by more than a fixed margin. Formally, for one pair of images $\left(\mathbf{x}_i, \mathbf{x}_j\right)$ it is defined as:

$$\mathcal{L}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbb{1}_{[y_i=y_j]} \left\| f\left(\mathbf{x}_i\right) - f\left(\mathbf{x}_j\right) \right\|_2^2 + \mathbb{1}_{[y_i \neq y_j]} \max\left(0, \epsilon - \left\| f\left(\mathbf{x}_i\right) - f\left(\mathbf{x}_j\right) \right\|_2\right)^2 \tag{4.6}$$

where $f(\mathbf{x}_i), f(\mathbf{x}_j)$ are the descriptors of the pair of images and $\|\cdot\|_2^2$ denotes the squared Euclidean distance. The value ϵ is the margin which defines when negative pairs have

large enough distance in order to be ignored by the loss. When the images in the pair share the same label, the second part of the loss dissapears and the loss is high when their descriptors are far. When they do not come from the same class, the first part of the loss dissapears, and having descriptors that are closer than the margin produces a non-zero loss value, which is increasingly higher the closer they lie in the embedding space.

We use each training image as an anchor to form one positive and one negative pair per epoch. Hard negative mining is used to produce hard negative pairs. A hard negative pair for a given anchor is a negative pair that is closer to the anchor than most other negative pairs, therefore producing a higher loss value. As a result, hard negative pairs provide more informative supervisory signal, which speeds up convergence for the expense of extra computation at the beginning of each epoch. They are formed by randomly choosing an image among the 10 most similar images from a different class, as these are computed according to embeddings obtained with the current backbone before each epoch. All the following variants use the same hard negative pairs, as just described. They differ in the way they form the positive pair. Three different ways of forming the positive pair are tested:

- *Con-Syn*: The positive is an augmented (synthesized) version of the anchor image, as in the case of SimSiam (second row of Figure 4.5).
- *Con-Syn+Real*: The selected positive is another randomly chosen image of the same class as the anchor, or an augmented version of the anchor image. Synthetic positive or one of the real (all images in the class but the anchor) positives is chosen with equal probability which is equal to one over the number of images in the class. If the class has a single image, then augmentation is performed; note that many classes contain a single image (third row of Figure 4.5).
- *Con-Syn+Real-closest*: Same as *Con-Syn+Real* but the real positive counterpart is chosen to be the one with the most similar embedding to the anchor. This is used to avoid images that depict completely different views of the object and has previously been used in location estimation [4]. Synthetic or real positive is chosen with equal probability (50%) in this case (fourth row of Figure 4.5).

Figure 4.5. *Examples of training pairs used by the representation learning methods that are trained on the Met training set. First row: SimSiam uses only image augmentations to form positive pairs, without the use of any supervision. Second row: Con-Syn also uses image augmentations for the formation of the positive pairs, however it uses supervision from the Met labels to form hard negative pairs. Third row: Con-Syn+Real additionally picks the positive pair from a pool that contains the augmentation plus all the other images from the same class as the anchor. Fourth row: Con-Syn+Real-closest limits the pool of candidate positive pairs to either the augmentation or the closest image (as measured in the representation space) that comes from the same class as the anchor.*

Chapter 5

Experimental evaluation

In this chapter, a performance evaluation of the proposed methods is described, with GAP, GAP without considering distractors (abbr. GAP^-) and accuracy on the test queries of the Met dataset as the evaluation metrics. Training, if any, is performed on the training part of the Met dataset, while the validation queries are either used as validation set during the training, to tune the hyper-parameters of the kNN classifier or to tune the inference temperature of the deep network classifier. Multi-scale representation and PCA whitening with dimensionality reduction to 512D are used in the rest of the chapter unless otherwise stated.

5.1 Implementation details

All methods are implemented in the PyTorch [65] deep learning library and use the FAISS [66] library for nearest neighbor search. In all approaches that involve training the Adam optimizer is used, weight decay is equal to 10^{-6}, learning rate is equal to 10^{-7} for the backbone and it is decreased by a factor of 10 in the middle of the training. The image augmentations used consist of random cropping in the scale range $[0.7, 1.0]$ and resize to 500×500, color jittering with probability 0.8, and conversion to grayscale with probability 0.2. DNet is trained with a batch size of 256 images for 25 epochs with the learning rate of the classifier set to 10^{-3}. Temperature γ used with CE is set to be fixed and equal to 30, while the temperature and margin penatly for AF are set to be fixed and equal to 64 and 0.5, respectively. SimSiam is trained with a batch size of 128 images, *i.e.* 64 original images augmented twice, for 15 epochs, with the learning rates of the projector and predictor MLP are set to 10^{-3}. The training with the contrastive loss is performed for 10 epochs with the margin set to 1.8. The batch size is equal to 128 images, comprised of 64 pairs randomly sampled from the positive and negative pairs of all anchors. An epoch is finished when all training images are used as anchors once. The best epoch is chosen according to validation accuracy of corresponding (parametric or non-parametric) classifier. To speed-up the process of choosing the best epoch with the kNN classifier, single-scale representation is used without PCAw. The hyper-parameters of the kNN classifier are tuned according to GAP on the validation set with grid search on the cartesian product of the sets $\{1, 2, 3, 5, 7, 10, 15, 20, 50\}$ and $\{0.01, 0.1, 1, 5, 10, 15, 20, 25, 30, 50, 100, 500\}$ for k and τ, respectively. The temperature

ID	Net	PCAw	MS	k	τ	GAP	GAP$^-$	ACC
1	R18IN			3	15	3.7	16.7	26.8
2	R18IN	✓		7	100	10.9	28.0	33.7
3	R18IN		✓	50	10	10.5	23.8	33.5
4	R18IN	✓	✓	3	50	**15.9**	**37.5**	**42.3**
5	R18IN	✓	✓	1	-	2.9	33.6	42.3
6	R18IN†	✓	✓	3	100	14.1	36.9	42.3

Table 5.1. *Recognition performance for kNN classifier on representation obtained from ResNet18 pretrained on ImageNet. MS: multi-scale representation. †: tuning k, τ only with Met queries, and without distractor queries in the validation set.*

of the parametric classifiers is also tuned according to validation GAP once the training is finished.

5.2 Image representation and kNN classifier components

ResNet18 (R18IN) trained on ImageNet is used as backbone to extract descriptors and perform recognition with a kNN classifier. Hyper-parameters k and τ are tuned on the validation set and reported separately per experiment in Table 5.1 which shows the impact of different components. The multi-scale representation and the use of whitening are shown to be beneficial and are essential parts of main approach (ID4 vs ID1,ID2, and ID3). Fixing $k = 1$ (ID5) is equivalent to no use of soft-max normalization in the kNN classifier and has significantly lower GAP on all queries, slightly lower GAP on Met queries, and identical accuracy by definition. Confidence normalization is therefore shown to be very important for the handling of distractors and to achieve high GAP performance. Finally, we show that having distractors in the validation set boosts GAP by better hyper-parameter tuning for the kNN classifier (ID6 vs ID4).

5.3 Pretrained backbones and kNN classifier

Table 5.2 summarizes results of recognition performance with a kNN classifier for backbones pretrained on different tasks and provides a comparison to the corresponding network trained on ImageNet (R18IN, R50IN [52]). Networks for art attribute prediction (R50SemArt (author), R50SemArt (type) [55]) perform worse than the ImageNet one, verifying that the task of art attribute prediction is far from that of ILR, despite being the same domain. The network for metric learning on landmarks (R18SFM, R50SFM [29]) provides improvements; despite the domain difference (artwork vs landmarks), training for metric learning well reflects the objectives of ILR. The model that is trained to mitigate texture bias (R50SIN [56]) performs worse than the ImageNet baseline, indicating that texture might play a role in the recognition of artworks. SwAV [57] provides a performance boost, verifying the usefulness of unsupervised representation learning for better generalization. Finally, SWSL [58] is the best performing variant demonstrating the benefits of learning on a very large image corpus despite the noisy labels; we expect the training set to include

Net	GAP	GAP$^-$	ACC
R18IN [52]	15.9 (+0.0)	37.5 (+0.0)	42.3 (+0.0)
R18SFM [29]	23.2 (+7.3)	41.5 (+4.0)	45.7 (+3.4)
R18SWSL [58]	**24.7 (+8.8)**	**47.0 (+9.5)**	**50.9 (+8.6)**
R50IN [52]	22.2 (+0.0)	41.8 (+0.0)	46.4 (+0.0)
R50SFM [29]	26.6 (+4.4)	44.8 (+3.0)	48.6 (+2.2)
R50SemArt (author) [55]	1.8 (-20.4)	12.2 (-29.6)	18.0 (-28.4)
R50SemArt (type) [55]	7.9 (-14.3)	26.8 (-15.0)	31.9 (-14.5)
R50SIN [56]	15.5 (-6.7)	36.4 (-5.4)	41.7 (-4.7)
R50SwAV [57]	22.8 (+0.6)	45.0 (+3.2)	49.6 (+3.2)
R50SWSL [58]	**30.4 (+8.2)**	**52.9 (+11.1)**	**56.3 (+9.9)**

Table 5.2. *Comparison of recognition performance for kNN classifier with representation from backbone networks pretrained for different tasks. Relative improvements compared to the corresponding network trained on ImageNet are shown in parentheses.*

many artworks too.

5.4 Training on the Met dataset

Results from training on the Met dataset are shown in Table 5.3 with a parametric Deep Network classifier (DNet) and with a kNN classifier. The latter is shown to be superior, while carrying the extra cost of storing a 512-D vector per training image. AF is shown to perform better than CE, verifying prior results on ILR [39]. SimSiam improves the performance over the baseline without the use of any supervision indicating that self-supervised learning is a promising direction for ILR. Con-Syn uses the same positive pairs as SimSiam (synthetic augmentations) but further boosts the performance by incorporating supervision in the form of (hard) negative pairs. Including real positive pairs too with constrastive loss achieves the best performance but only if the positive pair is properly disambiguated (Con-Syn+Real-closest vs Con-Syn+Real). Improvements by training on the Met are confirmed starting from SWSL pretraining too. Examples where R18IN Con-Syn+Real-closest succeeds in prediction but the R18IN baseline fails are shown in Figure 5.1. These cases include challenges such as large viewpoint changes and high inter-class similarity, which are scarce in the pretraining task of ImageNet classification. In Figure 5.2 we present examples of hard negative pairs, which are generated before the first epoch using the R18IN model and the contrastive loss. These examples showcase the small inter-class variability present in the Met dataset. Finally, challenging examples from the Met dataset for the top performing approach are shown in Figure 5.3. Wrong predictions for the Met queries (2nd row) as well as high confidence predictions for OOD queries, which correspond to predictions from the same semantic class (3rd row), reveal some of the difficulties in the dataset.

Method	GAP	GAP$^-$	ACC
Parametric classification			
R18IN DNet CE	9.6	24.7	30.6
R18IN DNet AF	16.9	32.0	36.6
kNN classification			
R18IN (baseline)	15.9	37.5	42.3
R18IN DNet CE	21.6	40.4	44.7
R18IN DNet AF	23.7	43.9	47.4
R18IN SimSiam	26.8	42.3	45.6
R18IN Con-Syn	30.4	46.6	49.4
R18IN Con-Syn+Real	29.8	46.0	48.8
R18IN Con-Syn+Real-closest	**32.5**	**47.5**	**50.0**
R18SWSL (baseline)	24.7	47.0	50.9
R18SWSL Con-Syn+Real-closest	**36.1**	**52.4**	**55.0**

Table 5.3. *Performance comparison for different types of training on the Met dataset. Training starts from the result of pretraining on ImageNet (IN) or that of SWSL. Baseline: not trained on the Met.*

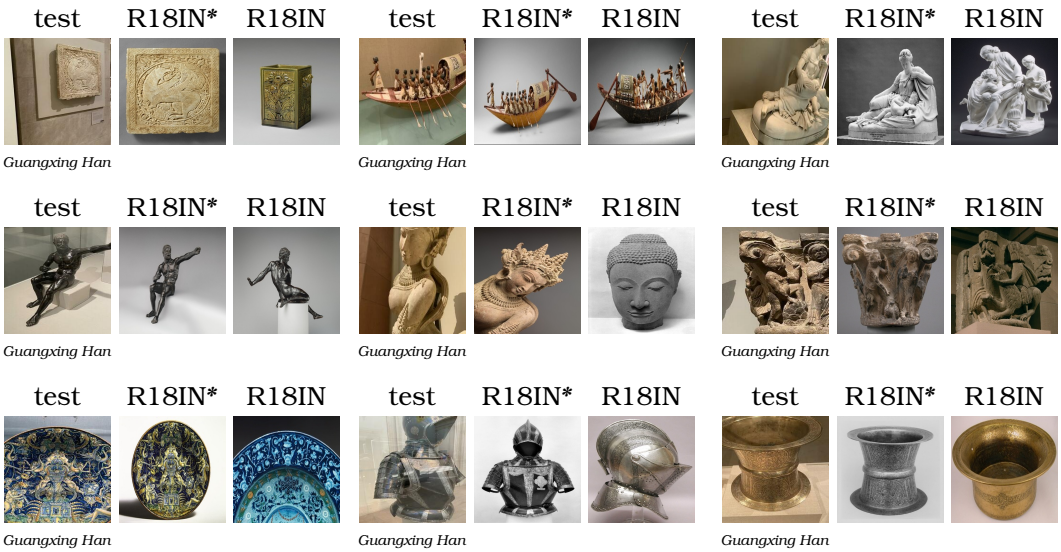

Figure 5.1. *Examples of incorrect and correct classification of test images for R18IN (baseline) and R18IN Con-Syn+Real-closest (R18IN*), respectively. The test images are shown next to their nearest neighbor from the Met exhibits that produced the respective prediction per method.*

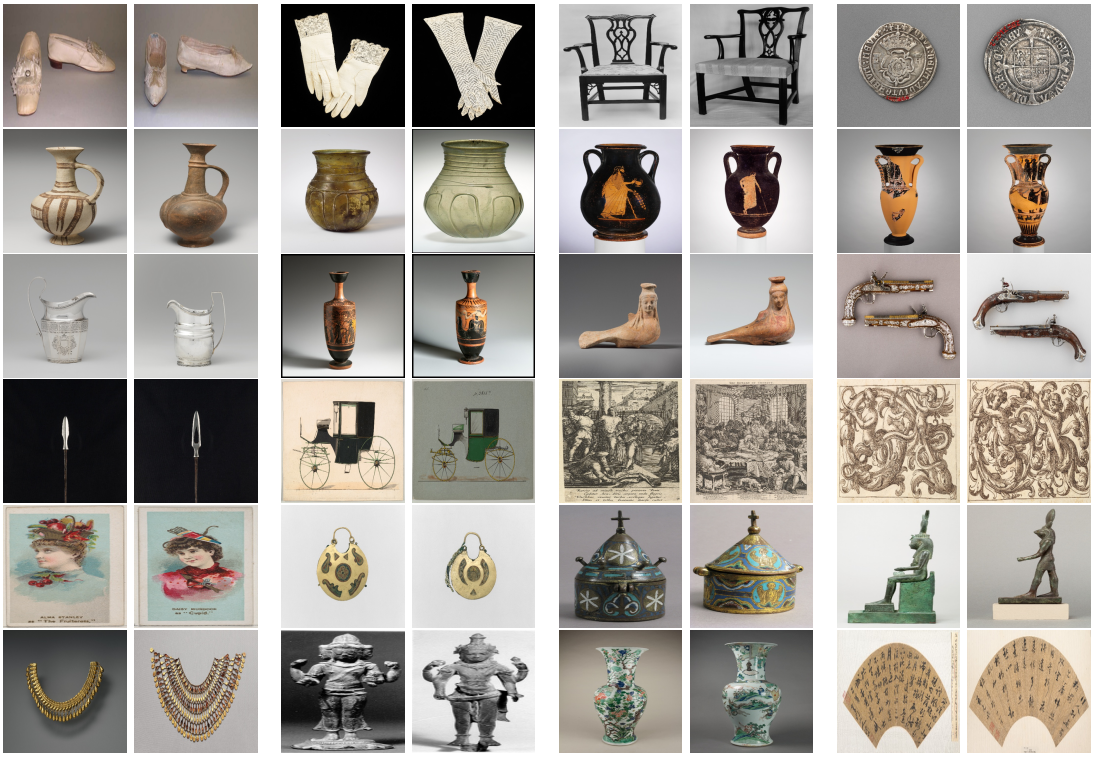

Figure 5.2. *Examples of hard negative pairs formed by the approaches that use the Contrastive loss on the Met training set. These examples additionally demonstrate the large inter-class similarity of the dataset. Images are shown as squares only for the purposes of this figure.*

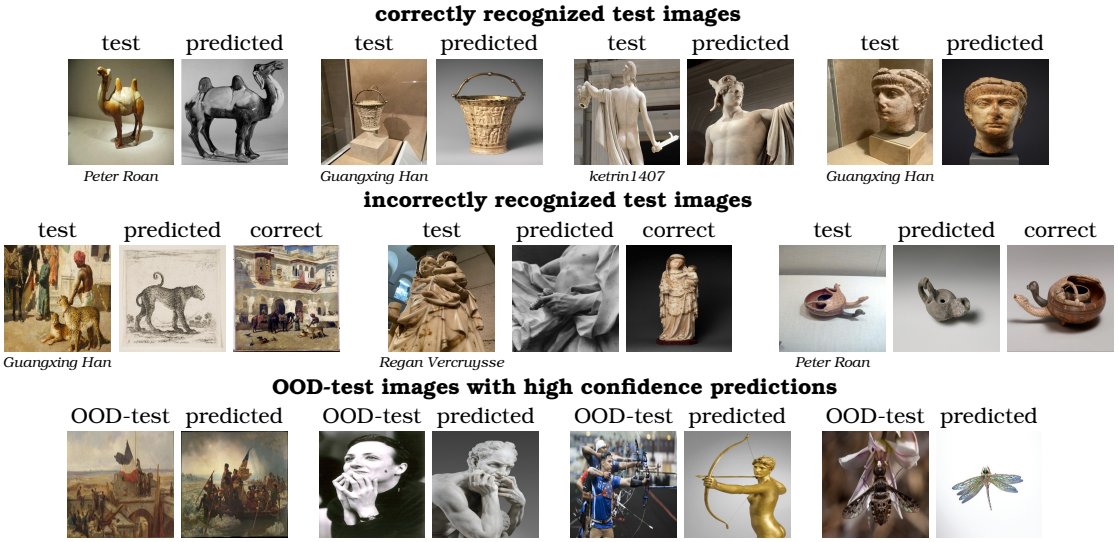

Figure 5.3. *Challenging examples from the Met dataset for the top performing approach. Test images are shown next to their nearest neighbor from the Met exhibits that generated the prediction of the corresponding class. Top row: correct predictions. Middle row: incorrect predictions; an image of the ground truth class is also shown. Bottom row: high confidence predictions for OOD-test images; the goal is to obtain low confidence for these.*

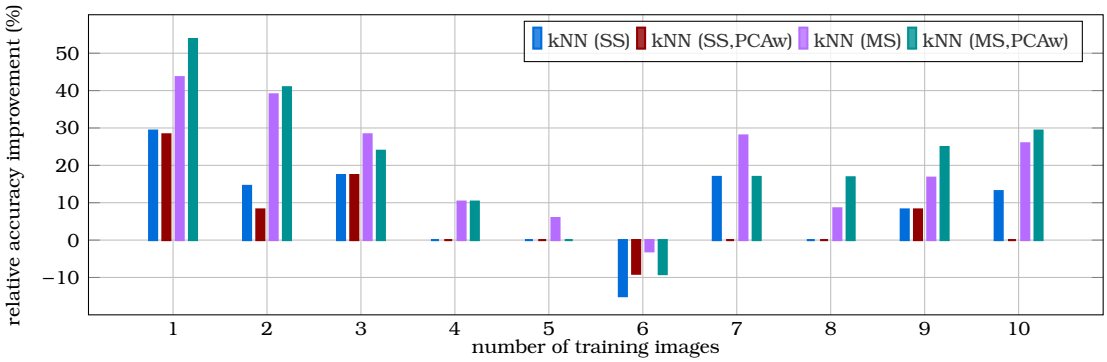

Figure 5.4. *Accuracy improvement of the kNN classifier over the parametric one for varying number of training images per class. DNet is trained with AF loss for the parametric classifier, while the embeddings learned with this setup are used for the kNN classifier. Relative improvements are reported in percentage for the different embedding variants.*

5.5 Long tail recognition and kNN classifier

We train a parametric classifier (DNet) and additionally use the resulting embeddings for the kNN classifier. A comparison is shown in Figure 5.4, where performance is reported separately according to the number of training examples per ground-truth class of each query. The kNN classifier does not only perform better than the parametric one, but is shown to be more suitable for long-tail recognition, as it achieves increasingly higher gains for more underrepresented classes.

Chapter 6

Conclusion

In this thesis, we have introduced a new large-scale dataset, the Met dataset, for the task of instance-level recognition in the domain of artworks. This dataset is the first artwork dataset to focus on this task, and is currently the only large-scale ILR dataset with clean annotations. It poses several hard challenges, as it is designed to simulate real-world conditions. It is large-scale, containing around 224k different artworks and follows a long-tail distribution, with over 60% of its classes having only one training image. Additionally, it exhibits high inter-class similarity, and distribution shift between the query and the training images, which are captured under different conditions. The query set is additionally enriched with out-of-distribution images, which puts a strong emphasis on the importance of robustness in a practical recognition system. Part of them come from the artwork domain, in order to further increase the difficulties.

Experimental evaluation on the corresponding benchmark shows that artwork related pre-training is not necessary useful, while ILR-related pre-training is more relevant. This indicates that the considered task is closer to ILR and deep representation learning than it is to popular computer vision tasks in the artwork domain, whilst including many of the same challenges. Fine-tuning the representation on The Met training set appears to be essential, as non-parametric classification is shown to be superior than its parametric counterpart. However, it is also challenging due to the training set statistics. The best performing approach is found to be a combination of self-supervised and supervised contrastive learning, leveraging the best of both worlds.

The goal of this dataset is to establish itself into the standard benchmarks for ILR. It is expected to foster research not only on ILR for artworks but also for ILR across multiple domains, when combined with other existing datasets.

Appendices

Appendix A

Dataset extras

In this appendix, information about the hosting of the Met dataset, its maintenance, and its licensing is included. Also, this appendix contains attribution to the Flickr users whose photographs have been used in the query set and more example images from the Met dataset.

A.1 Dataset hosting and maintenance

Public access and download links to the dataset are provided through the dataset webpage: http://cmp.felk.cvut.cz/met/. It contains tar files for all images and the ground truth files for evaluation. Publicly available reference code for using the dataset and computing the evaluation metrics can be found in https://github.com/nikosips/met. The code repository additionally includes code to reproduce some of the methods evaluated in the paper. The dataset is hosted at the servers of the Visual Recognition Group at the Czech Technical University in Prague.

A.2 License

The annotations of the dataset are licensed under CC BY 4.0 license. The images included in the dataset are either publicly available on the web, and come from three sources, *i.e.* the Met open collection, Flickr, and WikiMedia commons, or are created by us. The corresponding licenses for the ones that are available on the web are public domain, Creative Commons, and public domain, respectively. We do not own their copyright. For the ones created by us, we release them to the public domain. The creators of the dataset, bear all responsibility in case of violation of rights.

A.3 Flickr users

We thank the 37 following Flickr photographers whose photos with permissive license are included in the Met dataset. They appear in the form: username [real name], profile url.

- edenpictures [Eden, Janine and Jim], https://www.flickr.com/people/edenpictures

- Eric.Parker [Eric Parker], https://www.flickr.com/people/ericparker/
- semarr [Sarah Marriage], https://www.flickr.com/people/semarr/
- mharrsch [Mary Harrsch], https://www.flickr.com/people/mharrsch/
- Johnk85 [Johnk85], https://www.flickr.com/people/johnk85/
- zinetv [Lionel Martinez], https://www.flickr.com/people/zinetv/
- opacity [], https://www.flickr.com/people/opacity/
- Will.House [Will House], https://www.flickr.com/people/karloff/
- sarahstierch [Sarah Stierch], https://www.flickr.com/people/sarahvain/
- euthman [Ed Uthman], https://www.flickr.com/people/euthman/
- griannan [], https://www.flickr.com/people/griannan/
- Trish Mayo [], https://www.flickr.com/people/obsessivephotography/
- Stephen Sandoval [Stephen Sandoval], https://www.flickr.com/people/pursuebliss/
- Grufnik [], https://www.flickr.com/people/grufnik/
- smallcurio [], https://www.flickr.com/people/smallcurio/
- gtrwndr87 [Matthew Mendoza], https://www.flickr.com/people/mattmendoza/
- peterjr1961 [Peter Roan], https://www.flickr.com/people/peterjr1961/
- Stabbur's Master [Larry Syverson], https://www.flickr.com/people/124651729@N04/
- gorekun [], https://www.flickr.com/people/gorekun/
- rverc [Regan Vercruysse], https://www.flickr.com/people/rverc/
- IslesPunkFan [Neil R], https://www.flickr.com/people/islespunkfan/
- Pete Tillman [Peter D. Tillman], https://www.flickr.com/people/29050464@N06/
- squesada70 [Sergio Quesada], https://www.flickr.com/people/squesada/
- jareed [], https://www.flickr.com/people/jareed/
- stausi [], https://www.flickr.com/people/stausi/
- terryballard [Terry Ballard], https://www.flickr.com/people/terryballard/
- suetry [Susan Tryforos], https://www.flickr.com/people/stryforos/
- h-bomb [Howard Walfish], https://www.flickr.com/people/h-bomb/
- Robert Goldwater Library [The Robert Goldwater Library, The Metropolitan Museum of Art], https://www.flickr.com/people/goldwaterlibrary/
- juan tan kwon [jon mannion], https://www.flickr.com/people/jmannion/
- ctj71081 [], https://www.flickr.com/people/55267995@N04/
- ketrin1407 [], https://www.flickr.com/people/65986072@N00/
- wallyg [Wally Gobetz], https://www.flickr.com/people/wallyg/
- h_wang_02 [], https://www.flickr.com/people/7238238@N02/

- Olivier Bruchez [Olivier Bruchez], https://www.flickr.com/people/bruchez/
- JBYoder [Jeremy Yoder], https://www.flickr.com/people/jbyoder/
- jaroslavd [jerry dohnal], https://www.flickr.com/people/jaroslavd/

A.4 Extra image examples

We present more examples of Met queries and training images from the same class in Figures A.1 - A.3.

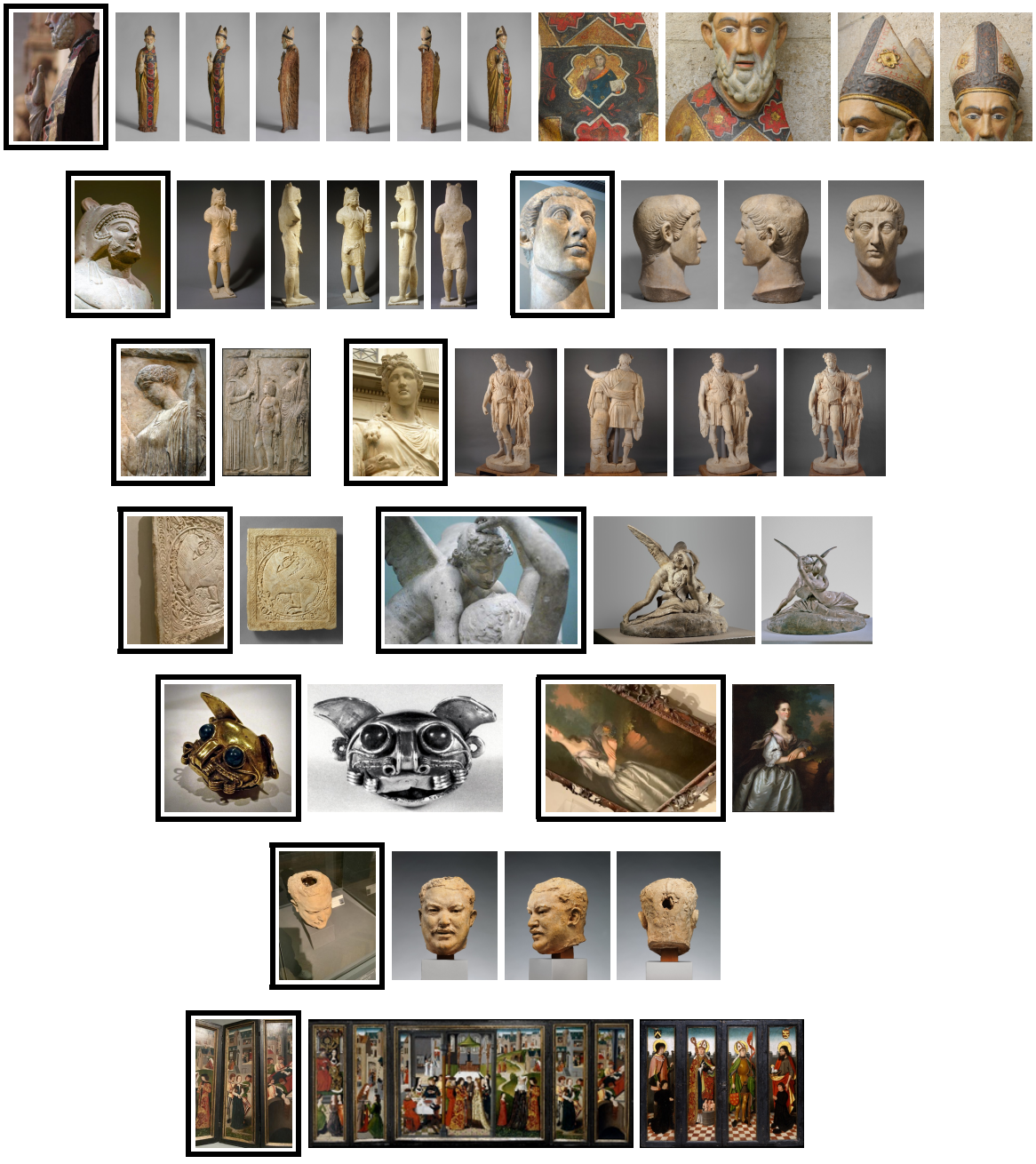

Figure A.1. *Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.*

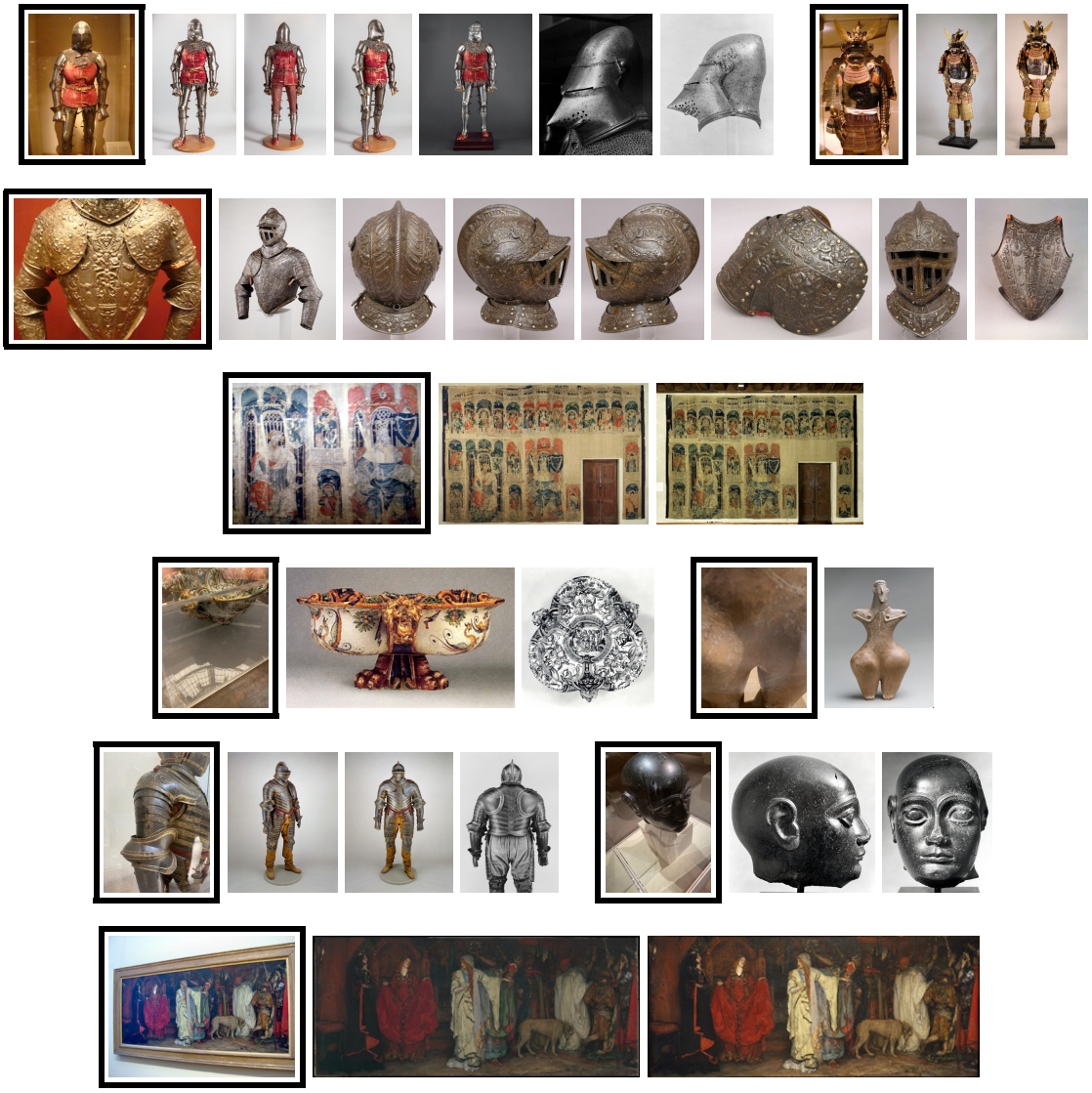

Figure A.2. *Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.*

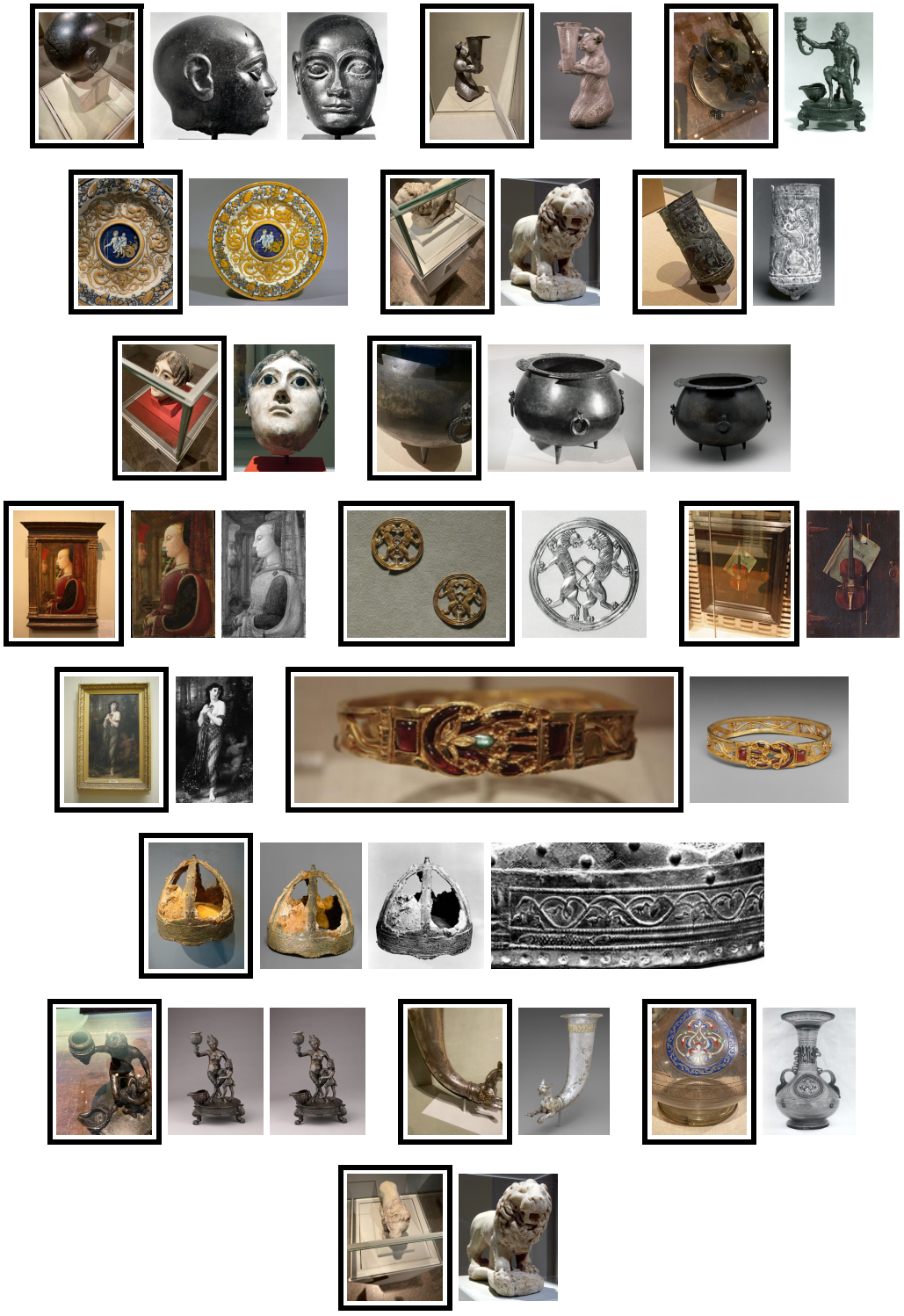

Figure A.3. *Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.*

Appendix B

Additional results

In this appendix, additional results of experimental evaluation on the proposed benchmark are included.

B.1 Descriptor dimensionality

Figure B.1 demonstrates the performance for increasing dimensionality of the image representation after PCAw. Combination by simple concatenation is shown to be effective.

B.2 Local descriptors

We evaluate the kNN classifier where the image-to-image similarity is computed with HOW local descriptors [11] (ECCV2020 R18 trained model) and ASMK [32]. It achieves 25.3 GAP, 47.6 GAP$^-$ and 50.9 ACC, which is the highest performance for this backbone (ResNet18) so far, however very close to the one achieved by the R18SWSL model and similarity with global descriptors. Note that this is a much costlier approach than all the rest in this work, which use global descriptors. The use of local descriptors trained for this task is likely to be a promising future direction especially due to the high inter-class similarities and the importance of distinctive artwork details.

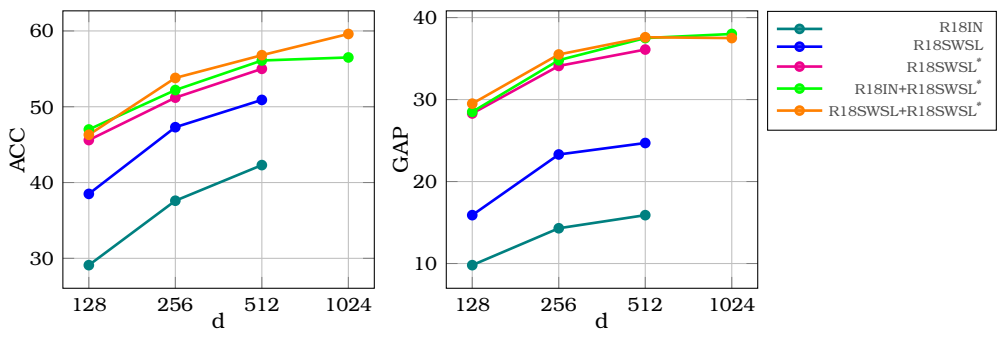

Figure B.1. *Performance with a kNN classifier versus dimensionality for different backbones. Two approaches are combined by simple representation concatenation before PCAw and is denoted by "+". *: Contrastive Syn+Real-Closest training on the Met dataset.*

B.3 Approaches for long-tail recognition

In order to mitigate the harmful effect of the imbalance of the Met training set on the learning process, we test a number of different approaches that are designed for long-tail recognition. Using the DNet classifier trained with AF as the reference method, the following methods are additionally used in training.

- *Class weighting:* The contribution of each sample in the loss function is weighted by the inverse of its class frequency.
- *Class-balanced sampling:* The samples that are contained in a training mini-batch are sampled uniformly across classes, and not across all training images.
- *Classifier retraining with class-balanced sampling:* After training the reference method, the backbone is kept frozen and only the classifier is re-initialized and trained with class-balanced sampling, as in the work of Kang *et al.* [67].

We observe no increase in accuracy with all these methods. More specifically, the reference method achieves 36.6 accuracy, class weighting achieves 35.8, class-balanced sampling achieves 33.4, and retraining achieves 35.0.

B.4 Mini dataset

We additionally create a smaller version of the database (training set) that contains all images from the Met classes present in the Met queries, plus about an extra 10% of the images from the rest of the classes of the original database. Its final size is 38,307 images from 33,501 classes. This set, along with the original query sets (test/val), form a subset of the dataset that serves as a faster way to check the validity of different training methods, before moving on to training on the entire database. This setup corresponds to an easier recognition problem than the original one. For reference, R18IN with kNN classification achieves 27.1 GAP, 49.0 GAP^{-} and 53.2 ACC on this subset.

B.5 OOD ratio

Results with and without distractors in the test set are included in the paper (GAP and GAP^{-}, respectively). We now include results, in Table B.1, for varying ratio of OOD queries in the validation set and in the test set. Results demonstrate the increasing difficulty by introducing more distractors and the fact that a small amount of validation distractors are enough for hyper-parameter tuning of the kNN classifier.

Val \ Test	0%	5%	10%	50%	100%
0%	36.9	32.9	29.7	19.9	14.1
10%	36.9	32.9	29.7	19.9	14.1
100%	37.5	33.6	30.9	21.8	15.9

Table B.1. *Performance of R18IN with kNN classification with different amount (percentage of their total number) of distractor queries in the validation (for tuning k,τ) and test set. Ratio lower than 100 is achieved by removing the appopriate amount of distractor queries.*

Bibliography

[1] Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan και Yinghui Xu. *Weakly Supervised Learning with Side Information for Noisy Labeled Images*. *ECCV*, 2020.

[2] Jonathan Krause, Hailin Jin, Jianchao Yang και Li Fei-Fei. *Fine-grained recognition without part annotations*. *CVPR*, 2015.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein και others. *Imagenet large scale visual recognition challenge*. *IJCV*, 2015.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla και Josef Sivic. *NetVLAD: CNN architecture for weakly supervised place recognition*. *CVPR*, 2016.

[5] Jan Knopp, Josef Sivic και Tomas Pajdla. *Avoiding Confusing Features in Place Recognition*. *ECCV*, 2010.

[6] Tobias Weyand, André Araujo, Bingyi Cao και Jack Sim. *Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval*. *CVPR*, 2020.

[7] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang και Wei Zhang. *Products-10K: A Large-scale Product Recognition Dataset*. *arXiv*, 2020.

[8] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg και Tamara L Berg. *Where to buy it: Matching street clothing photos in online shops*. *ICCV*, 2015.

[9] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang και Xiaoou Tang. *DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations*. *CVPR*, 2016.

[10] Riccardo Del Chiaro, Andrew D Bagdanov και Alberto Del Bimbo. *NoisyArt: A Dataset for Webly-supervised Artwork Recognition*. *VISIGRAPP (4: VISAPP)*, 2019.

[11] Giorgos Tolias, Tomas Jenicek και Ondřej Chum. *Learning and aggregating deep local descriptors for instance-level recognition*. *ECCV*, 2020.

[12] Chuanxing Geng, Sheng jun Huang και Songcan Chen. *Recent advances in open set recognition: A survey*. *PAMI*, 2020.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander

C. Berg και Li Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge. IJCV*, 115(3):211–252, 2015.

[14] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann και Holger Winnemoeller. *Recognizing image style. BMVC*, 2014.

[15] Hui Mao, Ming Cheung και James She. *Deepart: Learning joint representations of visual arts. ACM Multimedia*, 2017.

[16] Thomas Mensink και Jan Van Gemert. *The Rijksmuseum challenge: Museum-centered visual recognition. ICMR*, 2014.

[17] Gjorgji Strezoski και Marcel Worring. *Omniart: a large-scale artistic benchmark. TOMM*, 14(4):1–21, 2018.

[18] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse και Serge Belongie. *BAM! The behance artistic media dataset for recognition beyond photography. ICCV*, 2017.

[19] Elliot J Crowley και Andrew Zisserman. *The state of the art: Object retrieval in paintings using discriminative regions. BMVC*, 2014.

[20] Benoît Laurent Auguste Seguin, Carlota Striolo, Isabella di Lenardo και Frédéric Kaplan. *Visual Link Retrieval in a Database of Paintings. ECCVW*, 2016.

[21] Xi Shen, Alexei A. Efros και Mathieu Aubry. *Discovering Visual Patterns in Art Collections With Spatially-Consistent Feature Learning. CVPR*, 2019.

[22] Shiyu Liang, Yixuan Li και R Srikant. *Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. ICLR*, 2018.

[23] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon και Balaji Lakshminarayanan. *Likelihood ratios for out-of-distribution detection. NIPS*, 2019.

[24] Torsten Sattler, Bastian Leibe και Leif Kobbelt. *Fast image-based localization using direct 2d-to-3d matching. ICCV*, 2011.

[25] Vassileios Balntas, Shuda Li και Victor Prisacariu. *Relocnet: Continuous metric learning relocalisation using neural nets. ECCV*, 2018.

[26] Nikolaos Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord και Giorgos Tolias. *The Met Dataset: Instance-level Recognition for Artworks. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[27] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez και Cordelia Schmid. *Aggregating local descriptors into compact codes. PAMI*, 2012.

[28] Artem Babenko και Victor Lempitsky. *Aggregating Deep Convolutional Features for Image Retrieval. ICCV*, 2015.

[29] Filip Radenović, Giorgos Tolias και Ondřej Chum. *Fine-tuning CNN image retrieval with no human annotation. PAMI*, 41(7):1655–1668, 2019.

[30] Hui Wu, Min Wang, Wengang Zhou, Yang Hu και Houqiang Li. *Learning Token-based Representation for Image Retrieval*, 2021.

[31] Florent Perronnin και Christopher R. Dance. *Fisher Kernels on Visual Vocabularies for Image Categorization. CVPR*, 2007.

[32] Giorgos Tolias, Yannis Avrithis και Hervé Jégou. *To aggregate or not to aggregate: selective match kernels for image search. ICCV*, 2013.

[33] Giorgos Tolias, Ronan Sicre και Hervé Jégou. *Particular object retrieval with integral max-pooling of CNN activations. ICLR*, 2016.

[34] Albert Gordo, Jon Almazan, Jerome Revaud και Diane Larlus. *Deep Image Retrieval: Learning global representations for image search. ECCV*, 2016.

[35] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding και Jizhou Huang. *Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[36] David G Lowe. *Distinctive image features from scale-invariant keypoints. IJCV*, 60(2):91–110, 2004.

[37] Herbert Bay, Andreas Ess, Tinne Tuytelaars και Luc Van Gool. *Speeded-up robust features (SURF). CVIU*, 110(3):346–359, 2008.

[38] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand και Bohyung Han. *Large-scale image retrieval with attentive deep local features. Proceedings of the IEEE international conference on computer vision*, 2017.

[39] Bingyi Cao, André Araujo και Jack Sim. *Unifying deep local and global features for image search. ECCV*, 2020.

[40] Sumit Chopra, Raia Hadsell και Yann LeCun. *Learning a similarity metric discriminatively, with application to face verification. CVPR*, 2005.

[41] Florian Schroff, Dmitry Kalenichenko και James Philbin. *FaceNet: A unified embedding for face recognition and clustering. CVPR*, 2015.

[42] Hyun Oh Song, Yu Xiang, Stefanie Jegelka και Silvio Savarese. *Deep Metric Learning via Lifted Structured Feature Embedding. arXiv*, 2015.

[43] Richard Zhang, Phillip Isola και Alexei A Efros. *Colorful Image Colorization. ECCV*, 2016.

[44] Carl Doersch, Abhinav Gupta και Alexei A. Efros. *Unsupervised Visual Representation Learning by Context Prediction. ICCV*, 2015.

[45] Spyros Gidaris, Praveer Singh και Nikos Komodakis. *Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728*, 2018.

[46] Ting Chen, Simon Kornblith, Mohammad Norouzi και Geoffrey Hinton. *A simple framework for contrastive learning of visual representations. International conference on machine learning*. PMLR, 2020.

[47] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie και Ross Girshick. *Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[48] Gustavo Carneiro, Nuno Pinho Da Silva, Alessio Del Bue και João Paulo Costeira. *Artistic image classification: An analysis on the PRINTART database. ECCV*. Springer, 2012.

[49] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler και Serge Belongie. *The iMet collection 2019 challenge dataset. arXiv*, 2019.

[50] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli και Rui Zhang. *Museum Exhibit Identification Challenge for the Supervised Domain Adaptation and Beyond. ECCV*, 2018.

[51] Noa Garcia και George Vogiatzis. *How to read paintings: semantic art understanding with multi-modal retrieval. ECCV Workshops*, 2018.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep residual learning for image recognition. CVPR*, 2016.

[53] Filip Radenovic, Giorgos Tolias και Ondrej Chum. *Deep shape matching. ECCV*, 2018.

[54] Hervé Jégou και Ondřej Chum. *Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. European conference on computer vision*. Springer, 2012.

[55] Noa Garcia, Benjamin Renoust και Yuta Nakashima. *Context-Aware Embeddings for Automatic Art Analysis. Proceedings of the ACM International Conference on Multimedia Retrieval*, 2019.

[56] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann και Wieland Brendel. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR*, 2019.

[57] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski και Armand Joulin. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NIPS*, 2020.

[58] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri και Dhruv Mahajan. *Billion-scale semi-supervised learning for image classification. arXiv*, 2019.

[59] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang και Dahua Lin. *Learning a unified classifier incrementally via rebalancing. CVPR*, 2019.

[60] Feng Wang, Xiang Xiang, Jian Cheng και Alan Loddon Yuille. *Normface: L2 hypersphere embedding for face verification. Proceedings of the 25th ACM international conference on Multimedia*, σελίδες 1041–1049, 2017.

[61] Jiankang Deng, Jia Guo, Niannan Xue και Stefanos Zafeiriou. *Arcface: Additive angular margin loss for deep face recognition. CVPR*, 2019.

[62] Kevin Musgrave, Serge Belongie και Ser Nam Lim. *A metric learning reality check. European Conference on Computer Vision*. Springer, 2020.

[63] Xinlei Chen και Kaiming He. *Exploring Simple Siamese Representation Learning. CVPR*, 2021.

[64] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger και Roopak Shah. *Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems*, 6, 1993.

[65] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai και Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv*, 2019.

[66] Jeff Johnson, Matthijs Douze και Hervé Jégou. *Billion-scale similarity search with GPUs. arXiv*, 2017.

[67] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng και Yannis Kalantidis. *Decoupling representation and classifier for long-tailed recognition. Eighth International Conference on Learning Representations (ICLR)*, 2020.