The Met Dataset: Instance-level Recognition for Artworks supplementary material

Nikolaos-Antonios Ypsilantis
VRG, Faculty of Electrical Engineering
Czech Technical University in Prague

Noa Garcia
Institute for Datability Science
Osaka University

Guangxing Han
DVMM Lab
Columbia University

Sarah Ibrahimi
Multimedia Analytics Lab
University of Amsterdam

Nanne van Noord
Multimedia Analytics Lab
University of Amsterdam

Giorgos Tolias
VRG, Faculty of Electrical Engineering
Czech Technical University in Prague

The supplementary material includes an appendix with implementation details for the baseline methods, information about dataset hosting, maintenance, and licensing, attribution to the Flickr users, additional results, and extra details about the dataset including image examples. It additionally includes a datasheet for the Met dataset.

A Appendix

A.1 Implementation details

All methods are implemented in PyTorch [3] and use FAISS [1] 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 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 set to 10^{-3}. The training with 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 of the parametric classifiers is also tuned according to validation GAP once the training is finished.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

A.2 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 code to provide 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.3 License

The annotations 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.

We, the authors of this paper and creators of the dataset, bear all responsibility in case of violation of rights.

A.4 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.5 Additional results

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

Local descriptors: We evaluate the kNN classifier where the image-to-image similarity is computed with HOW local descriptors [5] (ECCV2020 R18 trained model) and ASMK [4]. It achieves 25.3 GAP, 47.6 GAP$^-$ and 50.9 ACC, which is the highest performance for this backbone 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 the paper, 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.

Mini dataset: We additionally create a smaller version of the database (training set) that contains all images from the classes that constitute 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.

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 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.

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 Arcface loss as the reference method, the following methods are additionally used for training. *Class weighting*: The contribution of

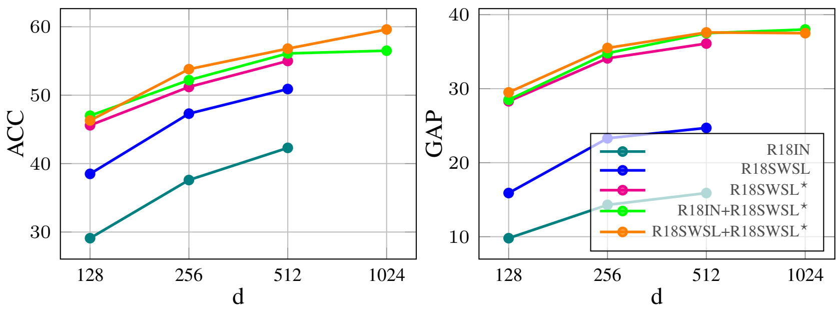

Figure 1: Performance with a kNN classifier versus dimensionality for different backbones. Two approaches are combined by simple representation concatenation before PCAw and is denoted by "+". ⋆: Contrastive *Syn+Real-Closest* training on the Met dataset.

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 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.

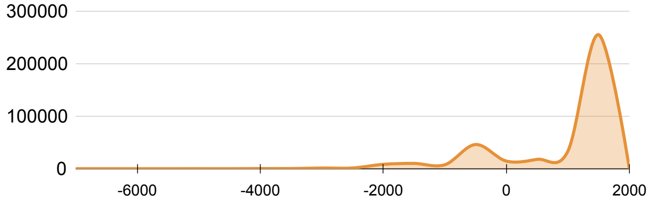

Figure 2: Number of exhibit images per time period.

each sample in the loss function is weighted by the inverse of its class frequency. *Class-balanced sampling*: The mini-batch samples 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*. [2]. 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.

A.6 Dataset extras

Figure 2 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.

The number of photographers versus the Met queries that belong to them is shown in Figure 3.

We present examples of Met queries and training images from the same class in Figures 4 - 8. Finally, we show examples of distractor queries from the other-artwork and non-artwork category in Figure 9 and Figure 10, respectively.

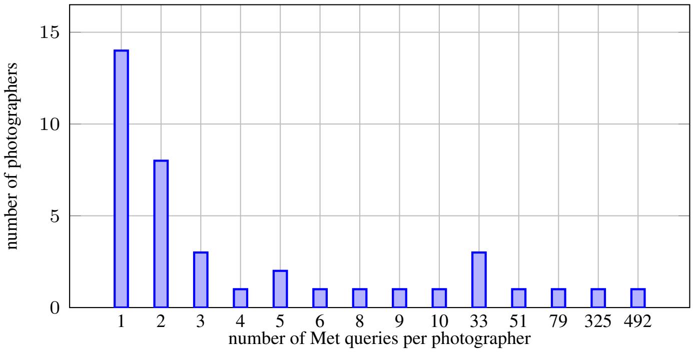

Figure 3: The number of photographers versus the Met queries that belong to them.

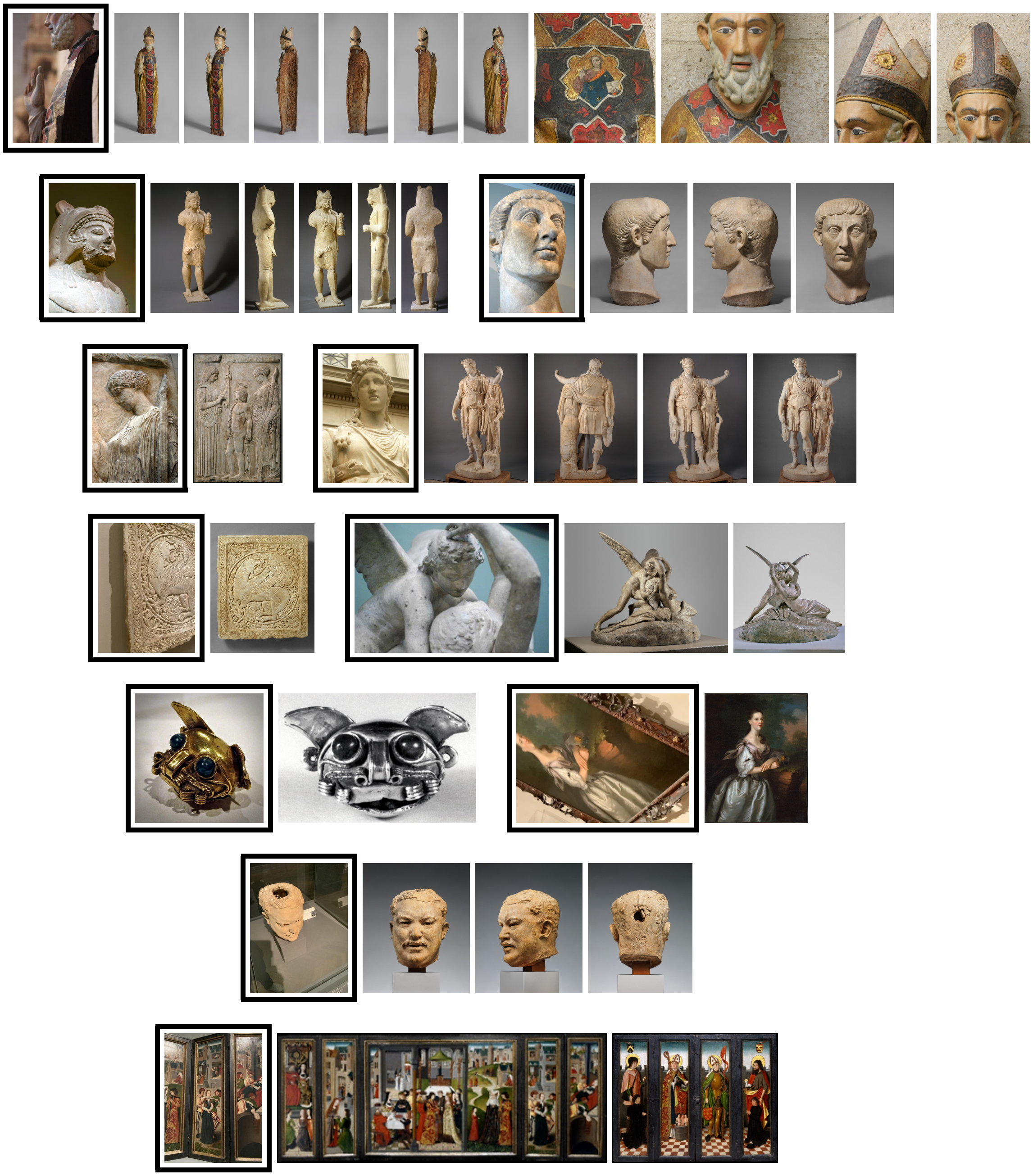

Figure 4: Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.

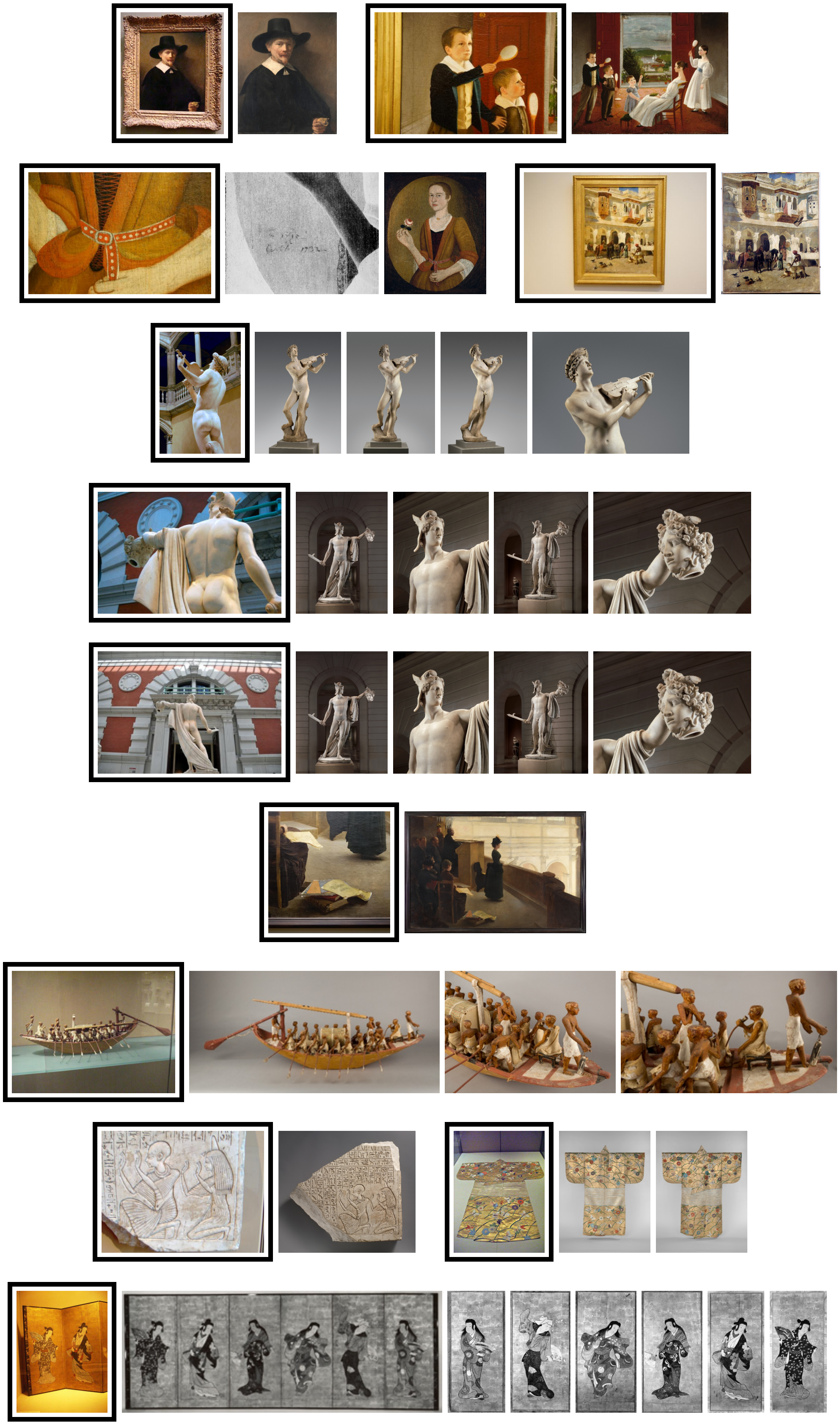

Figure 5: Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.

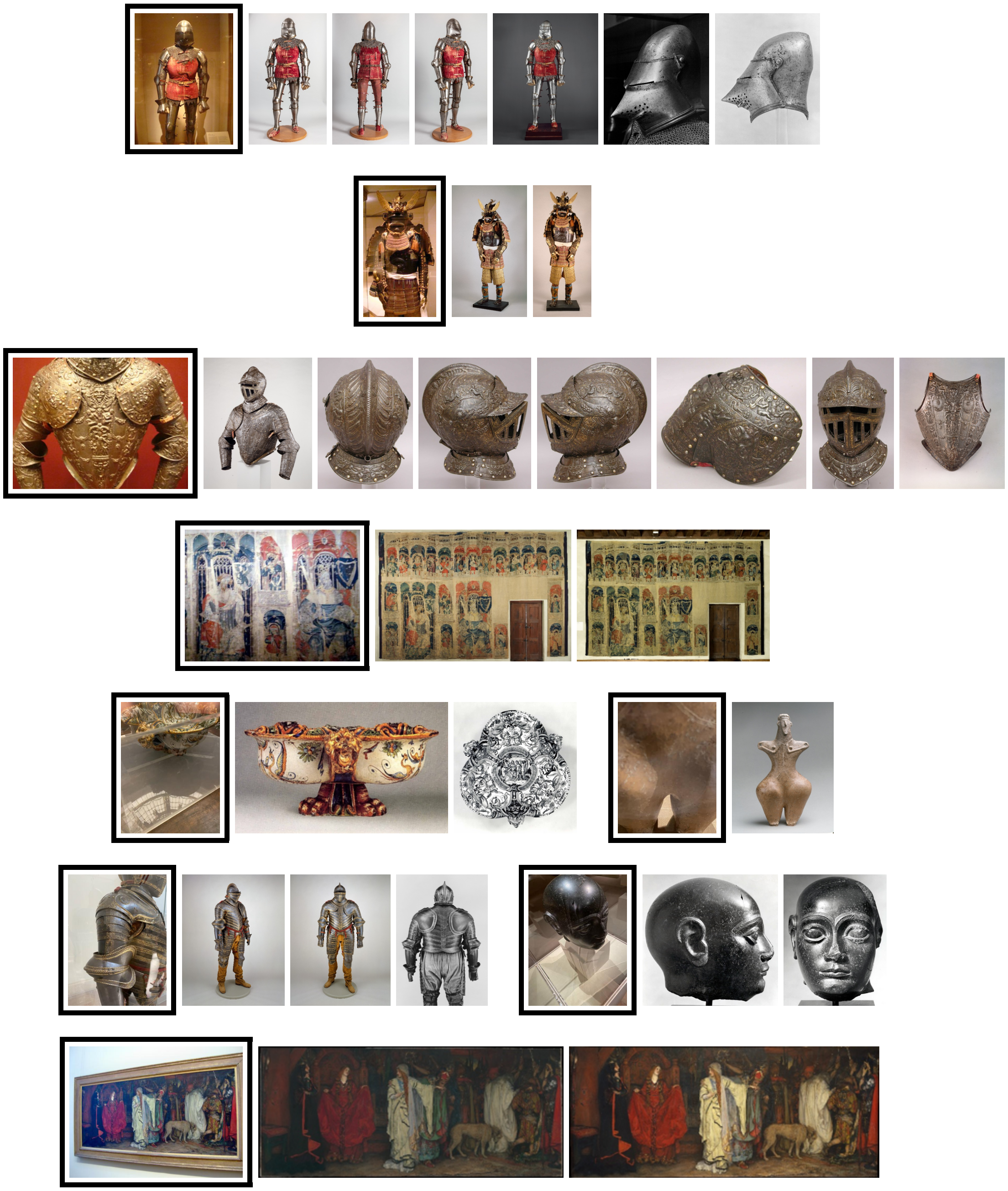

Figure 6: Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.

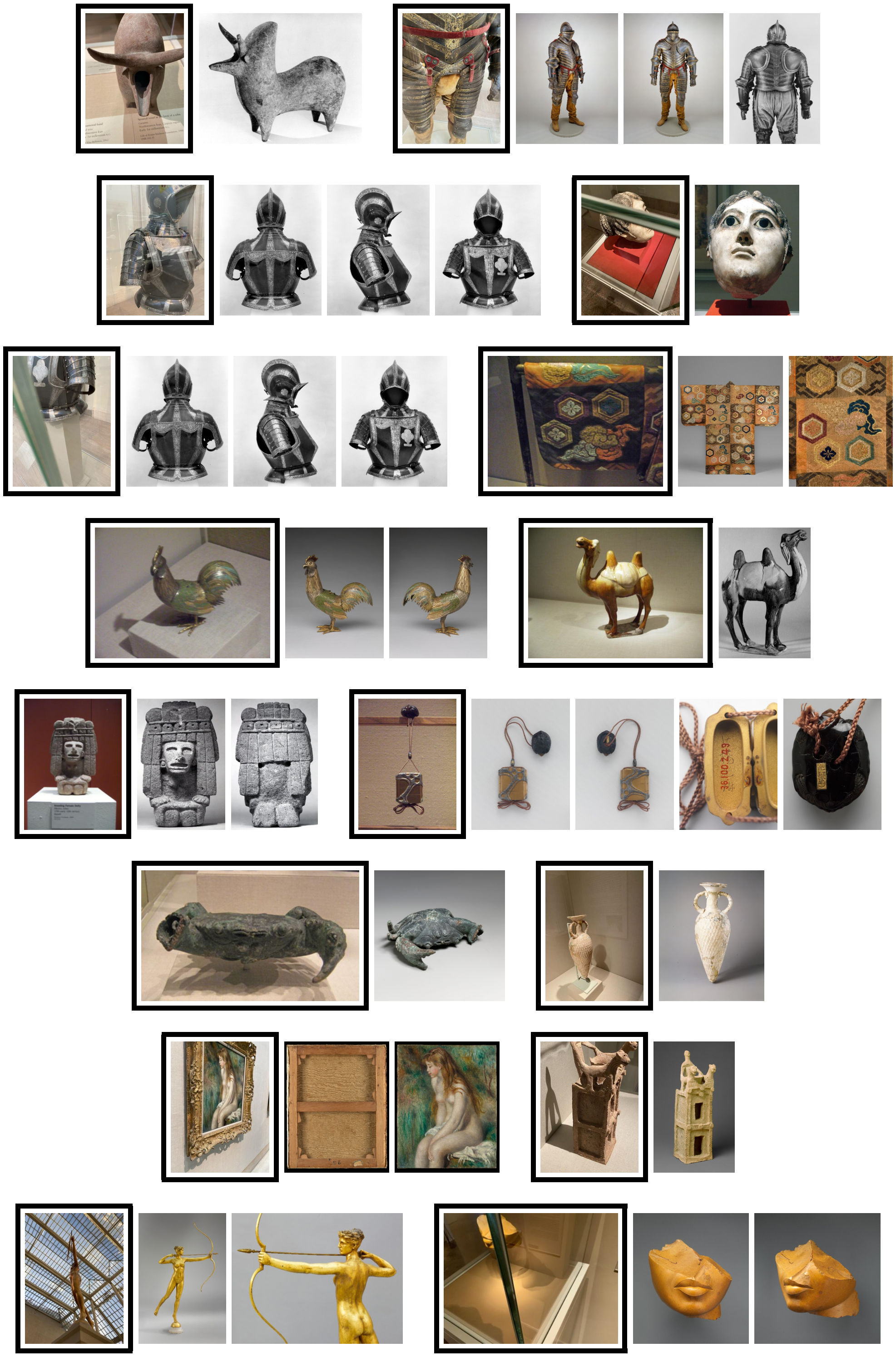

Figure 7: Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.

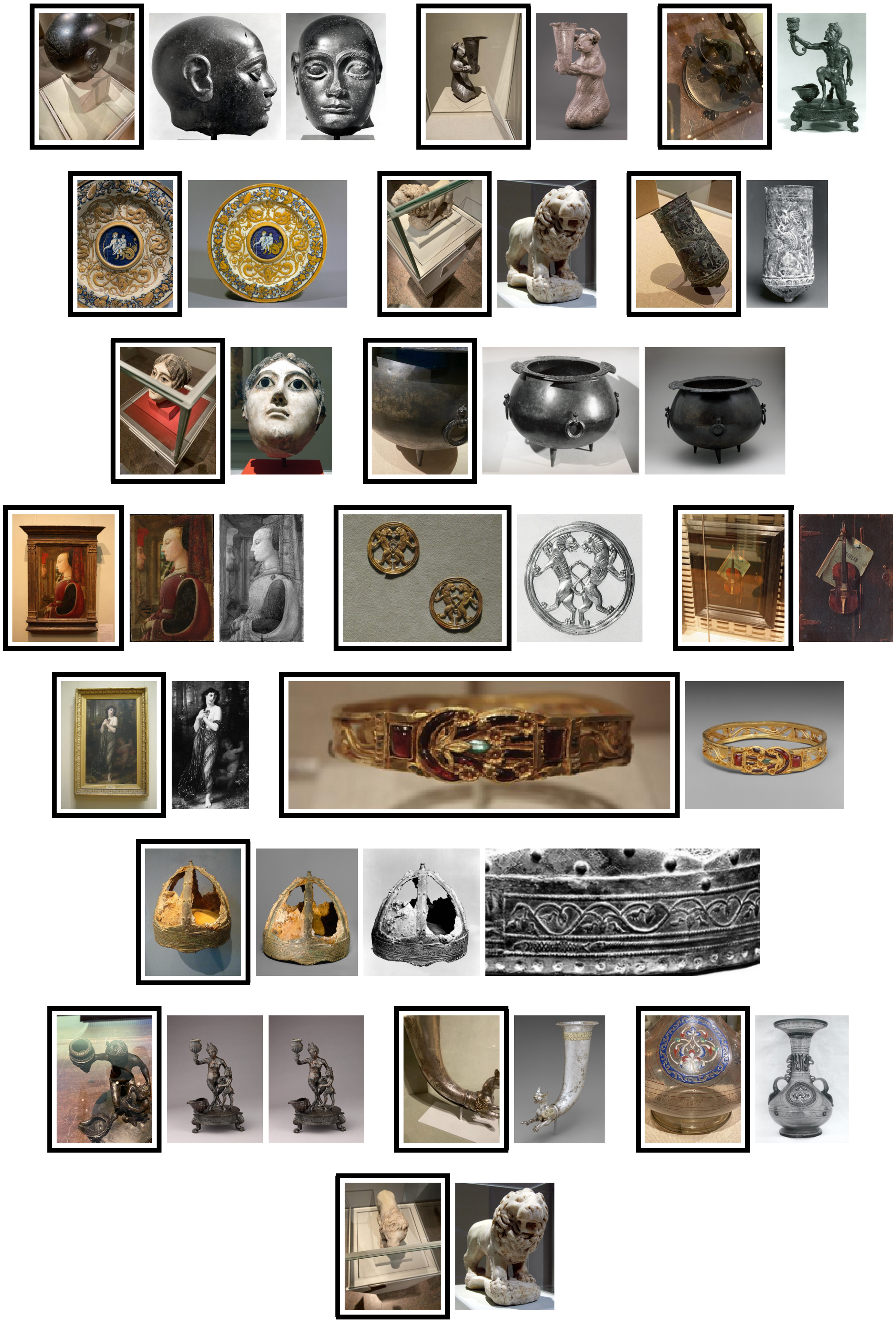

Figure 8: Examples of Met query images and training (exhibit) images of the corresponding Met class. Query images are shown in black border.

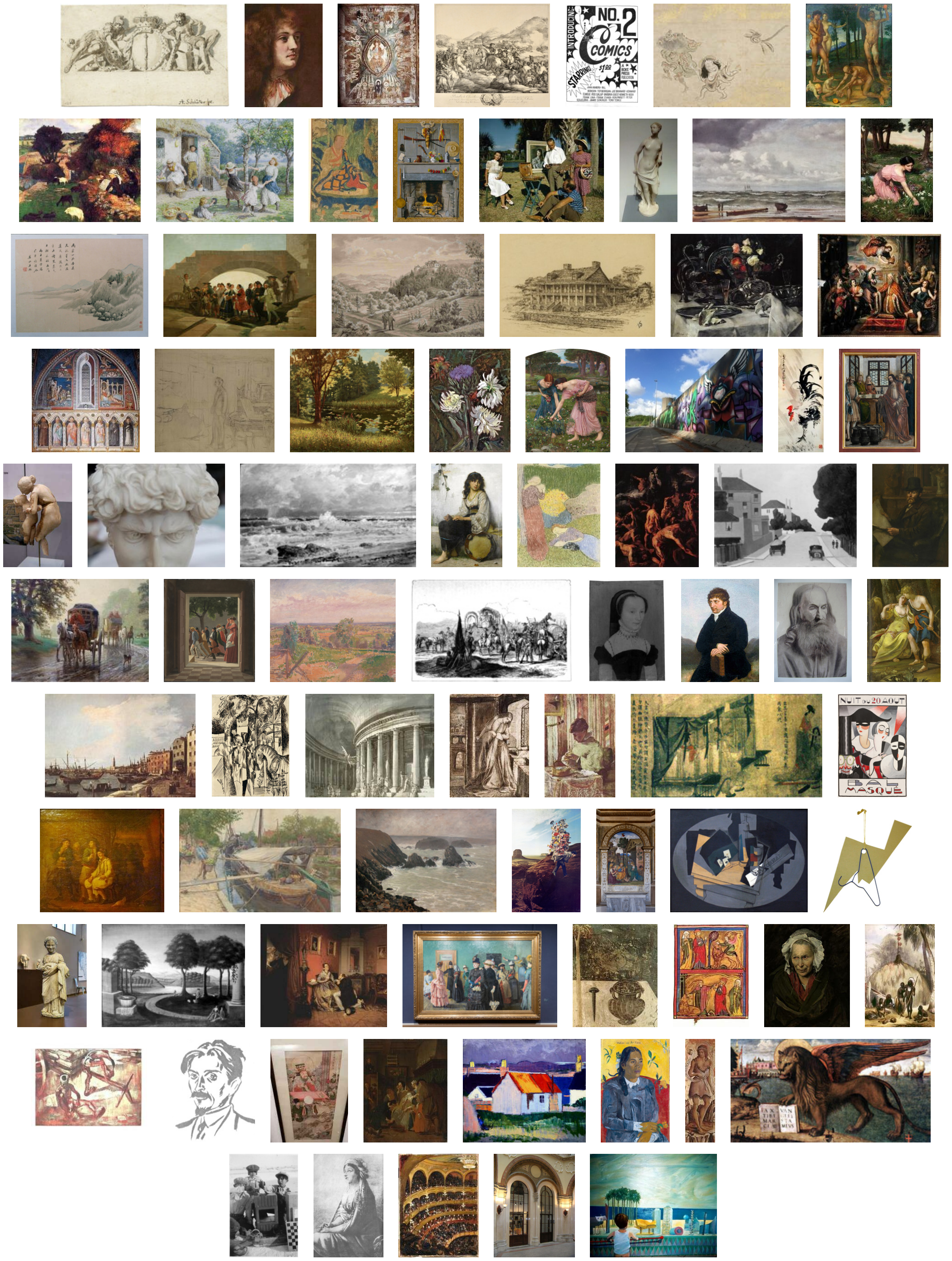

Figure 9: Examples of other-artwork (distractor) queries.

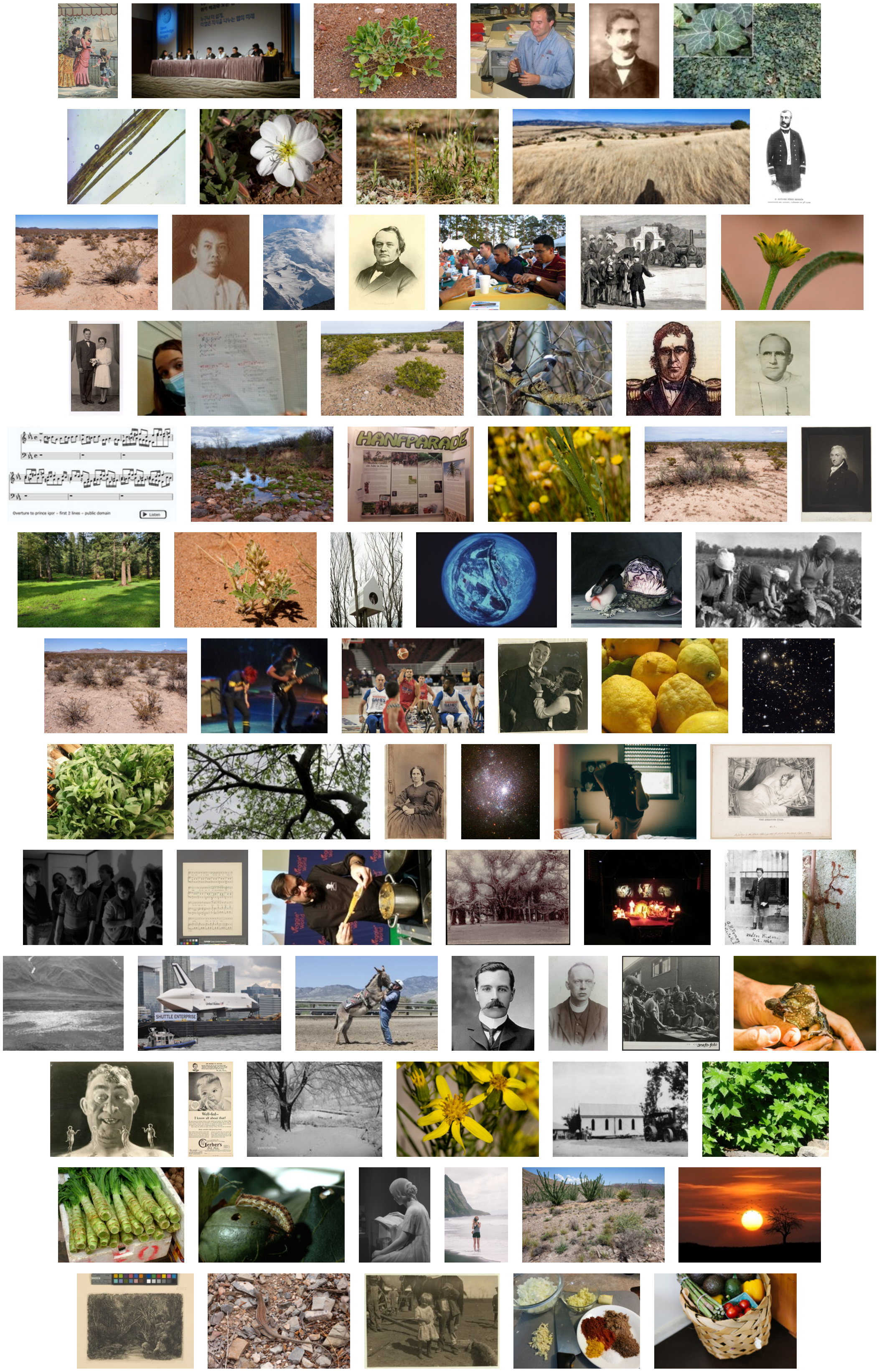

Figure 10: Examples of non-artwork (distractor) queries.

B Datasheet

MOTIVATION

For what purpose was the dataset created?

To our knowledge this is the only ILR dataset at this scale that does not include any noise in the ground-truth and is fully publicly available. Existing datasets are either significantly smaller, or created the ground-truth via crowd-sourcing which resulted in noisy labels, or are not fully publicly available. ILR has many application domains with corresponding real world applications, and existing datasets include the domains of landmarks, clothing, or products in general. The Met dataset is the first ILR dataset in the artwork domain. The dataset is created for the specific task of matching user photos of artworks to a database of images of artworks with known metadata, with the aim of identifying the artwork in the user photo.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Nikolaos-Antonios Ypsilantis (Czech Technical University in Prague), Noa Garcia (Osaka University), Guangxing Han (Columbia University), Sarah Ibrahimi (University of Amsterdam), Nanne van Noord (University of Amsterdam), Giorgos Tolias (Czech Technical University in Prague).

What support was needed to make this dataset?

N.A.

Any other comments?

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Each instance in the dataset is an image depicting either artworks or non-artwork content. There are *exhibit* images that form the training set, and *query* images that form the test set. The exhibit images are images from the open-access Met Catalog, made available by the Met museum through their online platform. There are two types of query images, the ones that depict an item from Met and the ones that do not. The latter are called distractor queries. The non-distractor query images consist of user photos taken at the Met museum by visitors depicting any of the items shown in the exhibit images of the dataset. A portion of these were taken by the creators of the dataset, and others were collected from Flickr. The distractor query images are images taken from Wikimedia Commons and might depict both artwork (other-art) and non-artwork (non-art) content.

How many instances are there in total (of each type, if appropriate)?

The number of images in the Met dataset are as follows:

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

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

All types of instances are samples from a larger set. For the Met catalog images that were not open-access or those that had a highly skewed aspect ratio were excluded, additionally a maximum of (the first) 10 images was selected per exhibit of the same artwork - to reduce overrepresentation of exhibits. The query images were selected from online sources based on open-access availability.

In the case of Met museum query images an additional selection criterion was the availability of identifying metadata so that we can establish ground-truth correspondence with any of the Met exhibits.

What data does each instance consist of?
Each instance is an image in JPEG format with a maximum resolution of 500×500.

Is there a label or target associated with each instance?
Each distinct Met exhibit included in the set of exhibit images of the dataset forms its own class. Query images are assigned to one of these classes, if the exhibit is depicted, or to the distractor class if no Met exhibit is depicted.

Is any information missing from individual instances?
Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
The relationships between exhibit images from the Met catalog and user query images are made explicit via discrete class labels. Additional relationships between user images, such as being captured by the same photographer, are made explicit via the metadata for the query images.

Are there recommended data splits (e.g., training, development/validation, testing)?
The dataset is divided into a training, validation, and test split. 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.

Are there any errors, sources of noise, or redundancies in the dataset?
We have performed multiple rounds of automated, semi-automated, and manual verification of the ground-truth and filtering of errors to minimize the chance of included errors. The accuracy of the final ground-truth has been verified by two different annotators. The open collection of the Met includes duplicate entries which we spotted (identical images) and removed.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?
No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No

Does the dataset relate to people?
A subset of the artworks depict persons (not always in a photorealistic manner, and the persons depicted might be fictive). The query images (particularly those from Wikimedia Commons) may contain depictions of persons, but the dataset nor the metadata contain information about these persons depicted (the source for the image Wikimedia Commons may have identifying information). Specifically, this dataset does not directly concern persons, nor does it contain data to identify any persons.

Does the dataset identify any subpopulations (e.g., by age, gender)?

No demographic information is included with the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Persons depicted in artworks can be identified via museum metadata. However, images collected from Flickr that are taken by museum guests do not depict individuals in an identifiable way; we have removed those images.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

The dataset does not contain sensitive data, as all images were collected from open-access online sources.

Any other comments?

COLLECTION

How was the data associated with each instance acquired?

Each Met exhibit forms its own Met class. Each exhibit image is labeled to a Met class according to the Met metadata. We label query images with their corresponding Met class, if any. Met queries taken by our team are annotated based on exhibit information, whereas Met queries downloaded from Flickr are annotated manually. 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. Finally, 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 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.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The majority of images were collected using software to crawl the Met catalog, Flickr, and Wikimedia Commons, the hardware used to taken these images varies significantly. The images collected by the team were taken with an iPhone 11 pro max.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Sampling was done based on availability and adherence to selection criteria, no specific (statistical) sampling strategy was used.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

All data collection and curation was performed by the paper authors themselves.

Over what timeframe was the data collected?

The dataset was constructed between September 2020 and September 2021. Images included in the dataset from public sources might have been captured before this timeframe.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No

Does the dataset relate to people?

A subset of the artworks depict persons (not always in a photorealistic manner, and the persons depicted might be fictive). The query images (particularly those from Wikimedia Commons) may contain depictions of persons, but the dataset nor the metadata contain information about these persons depicted (the source for the image Wikimedia Commons may have identifying information). Specifically, this dataset does not directly concern persons, nor does it contain data to identify any persons.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was crawled from open-access collections online (Met catalog, Flickr, Wikimedia Commons). Photos we collected ourselves were taken such as to avoid capturing other museum visitors.

Were the individuals in question notified about the data collection?

No, all data used was already public and available under an open-access license or does not contain persons.

Did the individuals in question consent to the collection and use of their data?

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Any other comments?

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The data was processed according to the following steps:

1. Gathered raw images from Flickr: The images were collected as described in the collection section.
2. Filtering: images that contain visitor faces, images not depicting exhibits, or images with more than one exhibit were discarded.
3. Annotation: query images were annotated with the corresponding Met class, similarly distractor images were discarded if they corresponded to a Met exhibit.
4. Verification: the label for each image was verified by two different annotators.
5. Rescaling: all images were resized to a maximum resolution of 500×500, preserving aspect ratio.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The 'raw' data is available from online sources, where relevant the metadata contains reference to the source image data.

Is the software used to preprocess/clean/label the instances available?

No, this process mainly consisted of manual effort with small specific scripts to automate simple tasks.

Any other comments?

USES

Has the dataset been used for any tasks already?

Yes, the paper has been used for Instance-level Recognition of artworks. See [PAPER] for details.

Is there a repository that links to any or all papers or systems that use the dataset?

No, we do not intend to collect all such links. We will ask future research publications that use the dataset to cite our paper. In such way, it should be possible to track its usage.

What (other) tasks could the dataset be used for?

The dataset could potentially be used for other Computer Vision tasks related to artistic images, such as attribute prediction, additionally given the domain shift between the exhibit and the query images the dataset could be used for domain adaptation.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

The dataset was collected and constructed with the ILR task in mind, because of this there might be limitations for future uses. Additionally, certain applications within the artistic domain rely on high resolution images, for this dataset the images have been downscaled, which might inhibit such applications.

Are there tasks for which the dataset should not be used?

Any other comments?

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset is available for download from `http://cmp.felk.cvut.cz/met/`. The website is under construction. A simple version is offered to provide access to reviewers, and a complete version will become available before publication.

The supporting code for evaluation and reproducing some of the baseline in the paper is provided in `https://github.com/nikosips/met`.

When will the dataset be distributed?

The dataset is already publicly available through the corresponding webpage.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The ownership of all images in the dataset are with their original publishers (e.g., the Met, Flickr users, and Wikimedia Commons users), however as all images are either licensed using a Creative Commons License or are in the public domain there are no limitations on the distribution and use of the dataset. We are providing attribution for all Flickr images by mentioning the creator and the corresponding Flickr url.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
No

Any other comments?

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is hosted at the Czech Technical University in Prague. Long-term administrator access is guaranteed for Giorgos Tolias.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Questions and comments about the dataset can be sent to Giorgos Tolias: giorgos.tolias@cmp.felk.cvut.cz

Is there an erratum?
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
In the unlikely event (see above for our effort to remove errors) that a number of errors are spotted in the future, the dataset will be updated and the relevant baselines scores will be updated too.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?
The dataset does not relate to people.

Will older versions of the dataset continue to be supported/hosted/maintained?
In the unlike even that spotted errors will trigger a dataset update, the older version (instances, and ground-truth) will remain publicly available.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
There is no specified mechanism but we are willing to update the dataset webpage by adding links to any useful extensions.

Any other comments?

References

[1] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. In *arXiv*, 2017.

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

[3] 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, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *arXiv*, 2019.

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

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