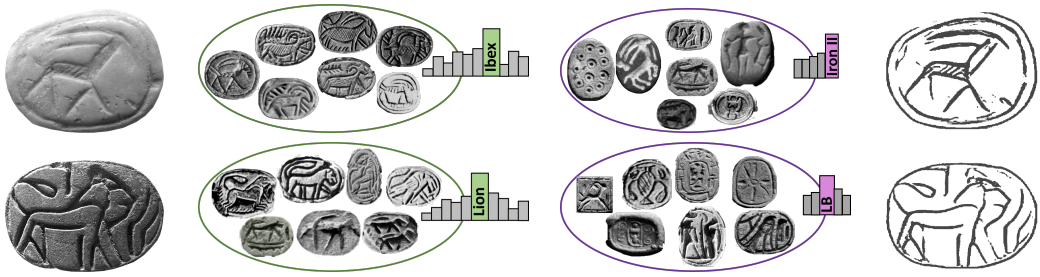
ArcAid: Analysis of Archaeological Artifacts using Drawings

Offry Hayon
Technion
offryh@gmail.com

Stefan Münger
U. of Bern
stefan.muenger@unibe.ch

Ilan Shimshoni
U. of Haifa
ishimshoni@is.haifa.ac.il

Ayellet Tal
Technion
ayellet@ee.technion.ac.il



(a) Input image (b) Classification by shape (c) Classification by period (d) Generation

Figure 1. **Analyzing archaeological artifacts.** Images of archaeological artifacts are extremely challenging to analyze, since they are eroded, broken and stained (a). Our model manages to classify these artifacts not only by shape (b), but also by historical period (c). In this figure, each ellipse bounds artifacts from the same class, demonstrating the great diversity within a class. Furthermore, our model generates a drawing of the given artifact (d), which is a standard way of archaeological documentation. The top row shows an *Ibex* dated to the *Late Iron*, whereas the bottom row shows a *Lion* dated to the *Late Bronze*.

Abstract

Archaeology is an intriguing domain for computer vision. It suffers not only from shortage in (labeled) data, but also from highly-challenging data, which is often extremely abraded and damaged. This paper proposes a novel semi-supervised model for classification and retrieval of images of archaeological artifacts. This model utilizes unique data that exists in the domain—manual drawings made by special artists. These are used during training to implicitly transfer the domain knowledge from the drawings to their corresponding images, improving their classification results. We show that while learning how to classify, our model also learns how to generate drawings of the artifacts, an important documentation task, which is currently performed manually. Last but not least, we collected a new dataset of stamp-seals of the Southern Levant. Our code[1] *and dataset*[2] *are publicly available.*

1. Introduction

Archaeology benefits society through understanding of the past and is acknowledged worldwide as a major research field. Advances in computer vision may be harnessed to the task in order to automatize some aspects of the study of archaeological findings (artifacts). For instance, a core task in the domain is to look for similar artifacts, which may reveal relations, commerce and connections between countries and cultures. The current practice is to leaf through thousands of pages in site reports. Instead, performing this task (and others) utilizing vision methods could take minutes. These methods are essential not only because the number of artifacts is large and the number of experts is small, but also because datasets are distributed all over the world.

This, however, is an intriguing task, as the archaeological domain exposes the limits of current computer vision techniques, due to several unique properties. First, there is a shortage of labeled data, since labelling must be performed by archaeological experts. Second, many archaeological artifacts are preserved in poor state of condition, eroded or broken, which differs from that of standard natural images. Third, since the artifacts are hand crafted, the consistency between different items of the same class is relatively weak.

A major task in archaeology is to classify artifacts by different criteria. The few classification methods in the domain exhibit good results, however they mostly focus on classes that have small variety within each class (for instance, similar coins with varied state of preservation) [2, 6, 10, 12, 15, 34, 38]. Our goal is broader: classi-

[1] https://github.com/offry/Arc-Aid
[2] https://cgm.technion.ac.il/arcaid/

fying a given artifact, where images in each class may vary greatly. This is either because they were produced in different periods or by different artists (e.g. considerably different lions and ibexes shown in Figure 1(b)) or because they are clustered by periods, even though the appearance of the shapes from the same period inherently differ (Figure 1(c)).

Since we do not have access to the real artifacts, which are often kept in store rooms of archaeological services, we focus on the visual documentation of these artifacts. One obvious such documentation is images, which is our input. Oftentimes, the images of the artifacts are accompanied by illustrations, made by trained drafts persons. Though the drawings and the photos are not necessarily aligned and the drawings are not exact depictions of the images, some features of the artifacts look clearer and more enhanced in the drawings. Thus, we propose to use them during training.

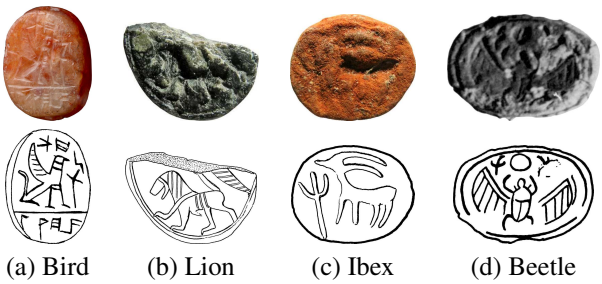
We introduce a novel semi-supervised approach for classifying archaeological artifacts. During training we utilize unlabeled pairs of drawings and images, together with a smaller number of labeled pairs. At inference, however, only an image is given as input. Our approach addresses the unique domain's challenges—a small dataset, the poor state, the lack of consistency between artifacts in the same class, and similar objects in different classes (which require expertise). Furthermore, we show that our approach is general and enables us to classify the same artifact into various classification types. Specifically, as demonstrated in Figure 1(b)-(c), our model classifies both by shape and by period. In these examples, the given image (at inference) is not ideal in terms of quality. Yet, our model classifies the images correctly according to both classification types.

Our method is based on a key observation that although the drawings are not exact edge detections of the photos, utilizing them during training is beneficial. This is so since they both represent the same main features of the artifact, so the global features found in both are similar, but clearer to detect in a drawing due to the state of the artifact (Figure 2). Forcing similarity between the embeddings of images and drawings contributes to the representation learning.

During training we solve an additional task—drawing generation—with unlabeled image-drawing pairs. Currently, since drawing generation requires special artists, it is done for very few selected artifacts, rather than to the whole data. We present SoTA results in classification, retrieval, and image-to-drawing generation in our domain.

Last but not least, we present a novel dataset, *Corpus of the Stamp-seals of the Southern Levant (CSSL)*. It contains scarabs and other seals from Egypt and the Southern Levant (1750-330 BCE), classified by experts both by shape and period. This is an important contribution, since archaeological datasets are rare in general, and in particular datasets of paired images and drawings.

Hence, this paper makes three contributions.



(a) Bird (b) Lion (c) Ibex (d) Beetle

Figure 2. **Input to training.** The edges of the drawings are clear and complete, comparable to their counterpart in the images. Moreover, the details differ and the pairs are misaligned.

1. It presents a semi-supervised method for multi-modal learning of paired drawings & photos in archaeology.
2. It introduces a model for image classification and retrieval of scarabs, with respect to the shape or the period, jointly with the ability for drawing generation.
3. We collected a new dataset of images and drawings of decorated scarabs, together with classifications and retrieval benchmarks according to both shape and period.

2. Related work

Computer vision for archaeology. Most of the papers that develop vision techniques for the archaeological domain address one of three tasks: (1) documentation, where the goal is to either generate 3D curves [31, 32] or to extract reliefs [13, 14, 35, 57]; (2) restoration, where the attempt is to restore the way the artifact looked before it was damaged, such as in hole completion [33, 45, 47], or in reassembly [9, 17, 36, 51]; (3) location, where the attempt is to locate the artifact in time and space through classification [2,4,6,15,27,34,38,44] or retrieval [3,25,29,38,44,46].

Our focus in this paper is on classification and retrieval of images of archaeological artifacts. In [2] a new architecture, *CoinNet*, is presented, which performs classification by shape (decoration). Results are presented on a Roman coins dataset. In [6] the *GlyphNet* architecture is introduced, which presents classification results on Hieroglyphs. In [38] classification is done both by period and by site, utilizing multiple CNNs and inferring with a voting ensemble approach. Their dataset contains images of archaeological tools and artifacts, mostly found in good preservation conditions. We compare our results to those of the recent models of [2, 6] and to those of the backbone model of [38], which work on 2D data and made their code available.

To generate drawings from 3D archaeological data, in [22, 30, 53] the sought-after curves are mathematically defined. Based on this definition nice results are produced for a few available 3D shapes. Since 3D archaeological data

is even scarcer than 2D data, we are the first to address the problem in 2D.

2D archaeological datasets. A dataset of $4,310$ grayscale images of hieroglyphs is presented in [21]. The artifacts are not well-preserved, however the variety in each class is relatively low. A dataset that comprises photographs of $6,770$ artifacts of different types is presented in [38], most are relatively well-preserved, with high variation within each class. The released version of the dataset is in a much lower resolution than that of the original, therefore we do not experiment on it. The dataset published in [2] contains $18,225$ images of ancient Roman coins. The coins are well-preserved and the variety within each class is low. We present a new dataset, which is not only the first to contain labeled pairs of images and drawings, but also contains challenging artifacts in terms both of preservation and of class variety.

Semi-supervised multi-modal learning. This task aims to integrate information from multiple modalities and learn shared knowledge, making use of both labeled and unlabeled data. Approaches such as pseudo-labeling [48, 56], teacher-student distillation [11,26], or co-training [8,37,54] do not fully leverage the nature of our paired dataset. Our work is somewhat related to teacher-student, as we aim to transfer knowledge from one encoder to another. The difference however is in the quality of the input of the teacher and the student, which use drawings & images, respectively.

There are also works which combine sketches and images in various ways [7, 18, 19, 41, 42, 52, 55, 58].

3. Method

Our goal is to design a model which, given only an image of a 3D decorated archaeological artifact, will output an embedding vector representation that can be used for analysis applications, in particular classification or retrieval. Such images differ from natural images in several manners. First, they are monochromatic. Second, the quality of the artifacts is poor, missing some contours, while others are created due to noise. Finally, the photos are often in poor condition.

Sometimes, these images are associated with drawings, created by special archaeological artists, as shown in Figure 2. These drawings consist of clear edges and their quality is superior to the quality of the corresponding photographs, for a couple of reasons: Due to the artists' vast experience, they can draw parts of the contours, even if they are abraded. In addition, the artist may have access to the real artifact when drawing it and can see the 3D features that are unclear in the image. Hence, these drawings encapsulate important domain knowledge in them. However, drawings and images are not necessarily aligned to each other, in terms of the actual geometric alignment of the edges, additional edges, missing edges, and missing damages. We will show that despite these drawbacks, when paired images and drawings are available during training, image representation is improved, in comparison to learning only from images.

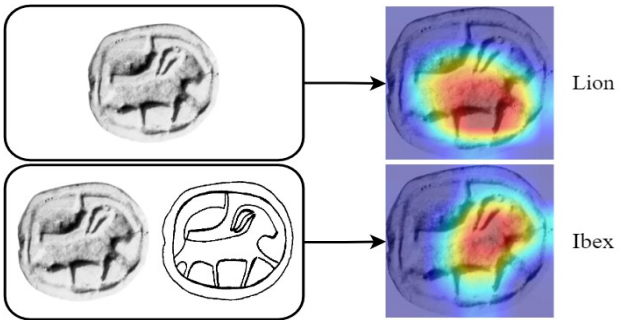


Figure 3. **Training with and without drawings.** When training only using images (top), the model focuses on the torso of the ibex, which leads to misclassification as a lion. Conversely, thanks to the drawing, our model focuses on the horns and the head (bottom), and classifies the image correctly.

We propose to utilize the drawings during training to improve image representation. During inference, however, images are the sole input, since in most cases drawings are unavailable. Intuitively, a drawing of an object can be considered as an augmentation of the photo, one which expresses the object's shape as an edge map and makes it easier to extract features that are difficult to obtain from the image. This is demonstrated in Figure 3, where the drawing enables our model to differentiate between engravings of ibexes and lions. Using only images, the localization map [43] focuses on a region (torso) in which the ibex and the lion engravings are hardly distinguishable, whereas when trained with our method, the model's focus is on the horns, which distinguishes between the two classes. This intuition is reinforced in our experiments in Section 6. Interestingly, since training is performed also for drawings, it enables our model to be used for a generative task—drawing an artifact, i.e. *image-to-drawing* in the archaeological domain.

Since in archaeology labeled data is scarce, we propose a training process that utilizes both labeled and unlabeled data, which exists in much larger numbers. We show that the mere existence of image-drawing pairs, even when unlabeled, helps. In particular, Section 6 shows that by training in this semi-supervised manner, we achieve better results than by using just the labeled data. Only classes that are unknown to the model appear in the unlabeled data.

The model. As shown in Figure 4, both labeled and unlabeled pairs of images and drawings are received as input, randomly in each batch, where most of the pairs are naturally unlabeled. The key idea is to optimize image embedding for classification, by maximizing the similarity between paired images and drawings. This is since training a network for drawing classification is easier than for images, as the shape is much clearer in drawings. Thus, we assume that the feature map of a drawing, represented by

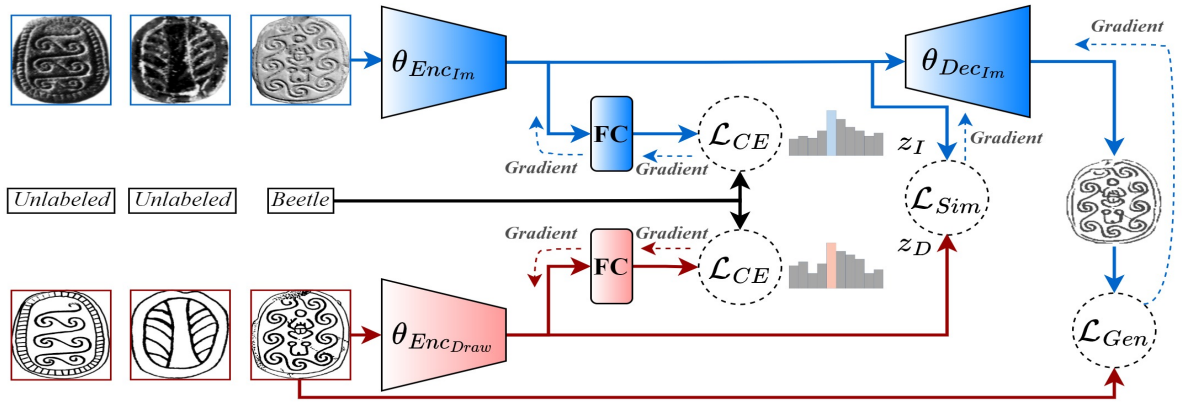


Figure 4. **Model.** This figure illustrates the processing of a batch of image-drawing pairs (3 in this example), where some of them are unlabeled and some are. The image and its corresponding drawing are encoded, where $\theta_{Enc_{Draw}}$ and $\theta_{Enc_{Im}}$ represent the parameters of their encoders, respectively. The FC components represent the classifiers. The image decoder, whose parameters are $\theta_{Dec_{Im}}$, generates the reconstructed drawing. The loss $\mathcal{L}$ consists of three components: $\mathcal{L}_{CE}$ for classification, $\mathcal{L}_{Gen}$ for image-to-drawing generation and $\mathcal{L}_{S}im$, whose goal is to maximize the similarity between pair embeddings. In case of an unlabeled pair, the classification component is ignored. Thus, we freeze $\theta_{Enc_{Draw}}$ and update the other components. In case of a labeled pair, $\theta_{Enc_{Draw}}$ is updated due to $\mathcal{L}_{CE}$; $\theta_{Enc_{Im}}$ is updated due to all the components of the loss function; and $\theta_{Dec_{Im}}$ is updated via $\mathcal{L}_{Gen}$.

an embedding vector, is more informative than that of its corresponding image. Under this assumption, making the image embedding similar to its paired drawing embedding, contributes significantly to improve image representation.

Our model, which realizes this idea, consists of two encoders, Enc_{Draw} and Enc_{Im}, whose inputs are drawings, D, and the corresponding images, I. Two fully-connected layers of the same structure are trained for classification. The output of the image encoder is fed into a decoder, Dec_{Im}, whose goal is to reconstruct drawings from images. Let us denote the parameters of the drawing encoder by $\theta_{Enc_{Draw}}$ and of the image encoder and image-to-drawing decoder by $\theta_{Enc_{Im}}$ and $\theta_{Dec_{Im}}$, respectively.

Our method is general and may train using many types of encoder/decoder backbone architectures. We will show in Section 5 that our method significantly improves the results irrespective of the chosen backbone.

Losses. Two losses use the outputs of the encoders, the *similarity loss* ($\mathcal{L}_{Sim}$) whose goal is to maximize the embedding similarity between the drawing and the matching image, and the *cross-entropy loss* ($\mathcal{L}_{CE}$) whose goal is to minimize classification errors. The output of the image decoder, jointly with the original drawing, are used for yet another loss function, the *generation loss* ($\mathcal{L}_{Gen}$).

In the unsupervised case, when a label is not given, we would still like to train the two encodings to be similar, by modifying the image encoding to be similar to the drawing encoding, which is kept fixed. In addition, the reconstructed drawing should be close to the original drawing. Thus, the gradients of $\mathcal{L}_{Sim}$ and $\mathcal{L}_{Gen}$ are used to update $\theta_{Enc_{Im}}$, the gradient of $\mathcal{L}_{Gen}$ is used to update $\theta_{Dec_{Im}}$, and $\theta_{Enc_{Draw}}$ is frozen. Freezing $\theta_{Enc_{Draw}}$ means that drawings affect and improve images embedding, but not the opposite, in a way that might harm drawings embedding.

In the supervised case, the classifier components are also used. Thus, in addition to the above, the gradients of the two $\mathcal{L}_{CE}$ losses update their respective encoders. Only in this case $\theta_{Enc_{Draw}}$ is updated.

Our image encoder loss is a sum of the three losses:

$$\mathcal{L} = \gamma_1 \cdot \mathcal{L}_{Sim} + \gamma_2 \cdot \mathcal{L}_{CE} + \gamma_3 \cdot \mathcal{L}_{Gen}. \tag{1}$$

The weights, γ_i, are hyper-parameters, chosen by trial and error and their values are $\gamma_1 = 0.8$, $\gamma_2 = 0.15$ and $\gamma_3 = 0.05$. We hereby elaborate on the losses.

To maximize the similarity in the latent space between the image embedding z_I and its corresponding drawing embedding z_D, we minimize a negative cosine similarity loss:

$$\mathcal{L}_{Sim} = -\frac{z_I \cdot z_D}{\|z_I\|_2 \cdot \|z_D\|_2}. \tag{2}$$

Thus, by maximizing the similarity, we force the image embedding to be closer to the drawings embedding, updating only $\theta_{Enc_{Im}}$ (and not $\theta_{Enc_{Draw}}$).

We train the FC layers for classification with a CE loss, $\mathcal{L}_{CE}$, one for drawings and the other for images. The goal is to improve both drawing and image embedding by updating $\theta_{Enc_{Draw}}$ and $\theta_{Enc_{Im}}$ for classification. The CE loss is computed only for the labeled pairs in the batch.

The generation loss takes into account the distance between the reconstructed drawing and the original drawing, using both the ℓ_2 norm and a perceptual loss, $\mathcal{L}_P$, similarly to [28]. While ℓ_2 aims to minimize mismatches be-

	(a) Ankh	Anthrop.	Bands	Beetle	Bird	Circles	Cross	Ibex	Lion	Sa
(b)	152	130	81	152	95	89	71	101	120	75

Figure 5. **Shape classes.** (a) shows a single instance of an image-drawing pair for each of the 10 classes. (b) is the number of labeled pairs.

tween pixels of a drawing and a generation, the use of the perceptual loss encourages a similar feature representation. Because the pairs of drawings and images are misaligned, a full pixel-wise match is not possible. Hence, the combination of ℓ_2 and $\mathcal{L}_P$ enables the model to learn a better visually-similar generation. $\mathcal{L}_{Gen}$ is thus defined as

$$\mathcal{L}_{Gen} = \alpha \cdot \left\| \tilde{D} - D \right\|_2 + \beta \cdot \mathcal{L}_P. \tag{3}$$

Here, $\tilde{D}$ is the reconstructed drawing, D is the original drawing, and α & β are weights. In our implementation, $\alpha = 0.3$ and $\beta = 0.7$; they were chosen via grid search.

4. Our new dataset - CSSL

A major contribution of our paper is a novel dataset of pairs of images & drawings of ancient Egyptian scarabs, called the *Corpus of the Stamp-seals of the Southern Levant (CSSL)*. This is the first dataset that contains paired images and drawings of any class of archaeological artifacts. The data was collected by seven different archaeologists for their own archaeological research. Thus, images might be centered and aligned differently between the archaeologists. Each artifact was classified by an expert archaeologist and was drawn by a trained drafts person. Despite the archaeological significance of these findings to their owners, we managed to get permission from all the involved parties to make this data available to the computer vision community.

CSSL contains $6,636$ pairs, out of which $1,020$ pairs are classified into 10 classes of scarab shapes and $5,616$ pairs are unclassified. The classes and the number of objects per class are shown in Figure 5. The supplemental material contains additional images of the various classes.

This dataset has a secondary classification into three periods and five sub-periods. In particular, 955 pairs are labeled into Middle Bronze (MB), Late Bronze (LB) and Iron ages and 296 of the labeled data do not have a period label. The Middle Bronze age and the Iron age are divided into two sub-periods each. Thus, 820 pairs are labeled into sub-periods. Figure 6 illustrates the difficulty of classification into periods, as lion-decorated scarabs, which are relatively similar to a non-expert, are related to different periods.

	(a) M.B.	M.B.	L.B.	Iron	Iron
		Early		Early	Late
(b)		Late			
(c)	144	374	217	43	44

Figure 6. **Sub-period classes.** All the findings are decorated by lions. They differ in shape and are dated to different periods (a) and sub-periods (b). The bottom row (c) shows the number of labeled pairs in each sub-period (not just for lions).

Figures 5-6 demonstrate why Archaeological datasets might be very challenging for computer vision techniques. The artifacts are eroded, causing the shapes to be hard to discern. In addition, the artifacts were made over a period of hundreds of years, which naturally resulted in the shapes of the objects from the same class to vary significantly.

5. Experimental results

Experimental setup. We evaluated our results using a variety of backbones as encoders: Resnet101 [23], DenseNet161 [24], EfficientNetB3 [50] (used in [38]), CoinNet [2], pretrained on ImageNet [16], and Glyphnet [1, 6]. The latter two are designed for archaeological data. For each encoder, we implemented a decoder, such that the setup is similar to that of UNet [39]. We show that for each backbone and task, our method significantly improves the results, demonstrating its generality.

We present the average result obtained for a two-fold cross-validation process, 50% train and 50% test. We use the accuracy measure for classification and *mean average precision (mAP)* for retrieval, both are the most-commonly used measures. For mAP, we evaluate P@1 and P@10. Ad-

Model	Class.	P@1	P@10
DenseNet161 [24]	80.8%	0.78	0.77
DenseNet161 [24]+ours	90.5%	**0.90**	**0.89**
Resnet101 [23]	82.5%	0.78	0.77
Resnet101 [23]+ours	89.6%	0.86	0.86
EfficientNetB3 [50]	72.9%	0.64	0.61
EfficientNetB3 [50]+ours	87%	0.85	0.84
Architectures for Archaeology			
CoinNet [2]	79.6%	0.77	0.76
CoinNet [2]+ours	**90.8%**	0.88	0.88
Glyphnet [6]	55.4%	0.43	0.43
Glyphnet [6]+ours	73.7%	0.65	0.64

Table 1. **Shape classification and retrieval on our dataset.** Our method is general and can utilize a variety of backbones. It outperforms all previous methods when compared on the same backbone. The best classification and retrieval results are obtained when using our method on top of DenseNet161 and CoinNet respectively.

Model	3 Periods	5 Periods
DenseNet161 [24]	81.3%	71.3%
DenseNet161 [24]+ours	84.0%	**72.7%**
Resnet101 [23]	81.1%	68.9%
Resnet101 [23]+ours	83.8%	72.4%
EfficientNetB3 [50]	77.3%	66.6%
EfficientNetB3 [50]+ours	82.4%	71.6%
Architectures for Archaeology		
CoinNet [2]	83.6%	70.2%
CoinNet [2]+ours	**84.5%**	72.2%
Glyphnet [6]	72.2%	56.6%
Glyphnet [6]+ours	75.6%	65.6%

Table 2. **Period classification on our dataset.** Our method improves all models. Best results are obtained when using our method on top of DenseNet for 3-period and CoinNet for 5-period.

ditional metrics and Confusion matrices, are given in the supplemental material.

Classification and retrieval. Table 1 shows that our method indeed improves the shape classification accuracy of each of the five models on CSSL by 7.1%-18.3%, and the mAP score by 0.08-0.23.

Table 2 shows the results of classification by period. We trained our model first on shape classification and then fine-tuned it for periods. For each backbone, our method improves the results of 3-period classification by 0.9%-5.1% and of 5-period by 1.4%-9.0%. The best results are obtained when using our method on top of CoinNet for 3-period (84.5%) and DenseNet161 for 5-period (72.7%).

Next, we experimented with the hieroglyphs dataset of [21]. The artifacts are not well-preserved, similarly to our dataset. Thus, we assumed that paired drawings (during training) could improve classification. However, this

Model	Dataset	Classification
Glyphnet [6]	Full [6]	97.6%
Glyphnet [6]	Released [21]	99.2%
Glyphnet [6]+ours	Released [21]	**99.4%**

Table 3. **Hieroglyphs classification on [21].** Our results are better than those reported in [6] on the full dataset and those attained on the released subset, when using the same backbone.

Model	Classification
Resnet50 [23]	94.5%
Resnet50 [23]+Ours	95.0%
Densnet161 [24]	96.0%
Densnet161 [24]+Ours	96.5%
Glyphnet [6]	96.2%
Glyphnet [6]+Ours	**96.7%**

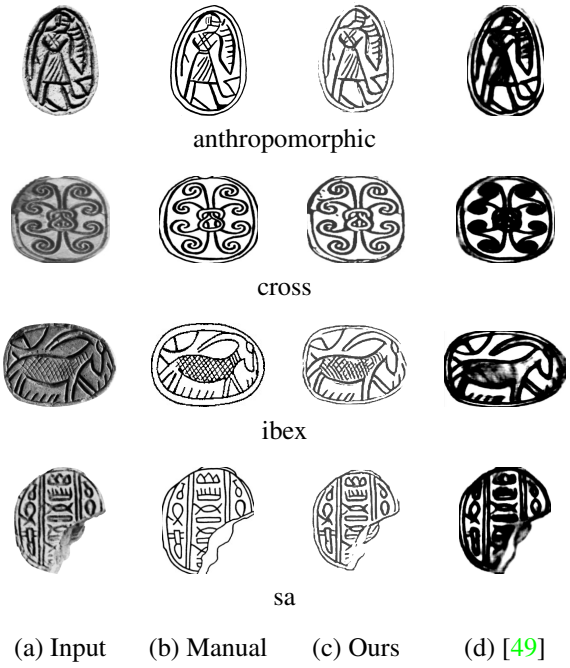
Table 4. **Hieroglyphs classification on [21].** Our method outperforms all backbones. All the backbones are trained from scratch; similar results are obtained for pre-trained backbones.

dataset does not contain drawings. Instead, we utilized general illustrations of the hieroglyph types from [5, 20, 40]. During training we paired each image to a random sample of hieroglyph drawing of the same type; they differ in style and small features. We used 2-7 drawings for each class. In contrast to our dataset, here all pairs are labeled.

We compare our results to those of [6]. They use a train, test and validation split of 70%, 15% and 15% respectively. Most of the dataset is released, but not all, thus we re-split the released subset with the same ratios. Table 3 compares the results presented in [6] on the full dataset to the results attained when training it on the released sub-set, with and without our model. It is shown that our results outperform [6]'s. Thus, we show that even though the drawings are not accurate matches to the images, using them during training improves the classification results. We note that the diversity within each class in this dataset is low, which makes classification easier than it is for our dataset.

Since this dataset is relatively easy, to further evaluate the benefit of our model, we applied it to a small training set. We split the released dataset into 20% train and 80% test. Table 4 compares the results on several backbones. The classification accuracy improves by 0.5% by our method, which is quite significant for this specific dataset.

Image-to-drawing generation. We strive to create an informative drawing from a given image of an archaeological artifact. This task differs from that of classical edge detection for a couple of reasons. First, unlike natural images, artifacts might be eroded and highly noisy, and edges are expected to be also in the eroded parts. Second, the drawings are not fully aligned with the image. Recall that the sec-



(a) Input (b) Manual (c) Ours (d) [49]

Figure 7. **Image-to-drawing generation.** Our results are more similar to the manual drawings than those of [49] in generating delicate textures and in completing faint edges. Additional results are given in the supplemental materials.

ond challenge is handled by $\mathcal{L}_P$, which aims to maximize similar feature maps, rather than full pixel-wise match.

Figure 7 demonstrates our results qualitatively. Since no prior work that generates drawings in our domain exists, to allow comparison we trained a SoTA supervised edge detector, *DexiNed* [49] on CSSL, considering paired drawings as edge maps. A main challenge is to deal with noise, existing in images but not in drawings, caused as a result of the quality of the images, erosion of the artifacts, and the ability of the artist to complete the missing lines. In addition, each artifact has defects, some of which also present in our generations. *DexiNed*'s results miss the delicate textures and drawing features, such as the texture on the ibex's torso. DexiNed also generates thick edges, in accordance with the image itself. Conversely, our method generates the delicate textures, nicely completes faint edges, and uses edge thickness as learned from the manual drawings. A couple of archaeologists we consulted with consider the generated drawings as useful.

Table 5 quantitatively compares our results to DexiNed [49]'s, using common edge detecting metrics: *F-measure* of *Optimal Dataset Scale (ODS)*, *Optimal Image Scale (OIS)* and *Average Precision (AP)*. In this experiment we consider the manual drawings as ground truth edge maps. Our results outperforms DexiNed's in all metrics.

Training method	ODS	OIS	AP
Our drawings	**0.38**	**0.39**	**0.29**
DexiNed [49]'s drawings	0.28	0.28	0.25

Table 5. **Image-to-drawing generation quantitative results.** Our model outperforms DexiNed [49]'s in common metrics for edge detection, considering the manual drawings as ground truth.

Input type / size	Full set	$1/2$ set	$1/4$ set
Ours: omitted unlabeled	89.6%	86.5%	78.0%
Ours: omitted unused	89.6%	85.8%	77.8%
Photos only	82.5%	72.0%	60.3%

Table 6. **Training with different inputs.** The classification results, when training with paired images & drawings, outperform the results when using only images during training. Furthermore, the less data available, the more important it is to use these pairs.

We note that our model is even able to generate accurate drawings from photos of objects it was not trained on, such as artistic reliefs, sculptures and a variety of archaeological artifacts. We show qualitative results in the supplementary.

6. Ablation study

This section evaluates the benefits of the different components of our method. In particular, it evaluates the contribution of the drawings for the training, of jointly solving classification and drawing generation, of training with both labeled and unlabeled data and of training two separate encoders. We also examine the benefit of our method when the amount of labeled data is low, which is likely to be the case for future archaeological datasets. For that sake, we use three different sizes of labeled pairs during 2-fold cross validation training: full dataset (507,513), half dataset (256,259) and a quarter of the dataset (129,132). The size and split of the test set remains the same, 507 & 513 pairs. In all experiments we use the Resnet101 backbone. Similar results are obtained for other backbones.

The contribution of the drawings for the training. We trained our model with and without drawings. Recall that the input at inference is always an image, as this is the prevalent available data. Table 6 shows that as expected, training with pairs is indeed preferable. Moreover the more data, the better. Finally, it shows that the less available data during training, the more important it is to use paired data. This is tested in two cases, when the omitted examples are not used at all and when the omitted examples are considered as unlabeled pairs.

The benefit of jointly solving all tasks. We evaluate the impact on classification of image-drawing similarity and of drawing generation. Toward this end, we checked the impact of $\mathcal{L}_{Sim}$ and $\mathcal{L}_{Gen}$. Specifically, we trained

Training set size	Full set	1/2 set	1/4 set
Full method	89.6%	86.5%	78.0%
w/o $\mathcal{L}_{Sim}$	82.2%	72.1%	61.8%
w/o $\mathcal{L}_{Gen}$	87.8%	84.1%	73.9%

Table 7. **Impact of $\mathcal{L}_{Sim}$ and $\mathcal{L}_{Gen}$ on classification.** Both $\mathcal{L}_{Sim}$ $\mathcal{L}_{Gen}$ are crucial for the accuracy of solving classification.

Training	Full set	1/2 set	1/4 set
Ours: labeled+unlabeled	89.6%	86.5%	78.0%
Ours: only labeled	83.8%	75.8%	64.8%

Table 8. **Supervised vs. semi-supervised training.** This experiment shows that unlabeled data is beneficial and improves the results. Since we do not expect to have much labeled data in this domain, this is very important.

the classification model without $\mathcal{L}_{Gen}$ and then a model that solves generation without forcing embedding similarity $\mathcal{L}_{Sim}$ (Equation 1). Table 7 shows that, as expected, $\mathcal{L}_{Sim}$ is crucial for classification. Furthermore, adding $\mathcal{L}_{Gen}$ improves classification as well, while providing drawings that are important for documentation.

Semi-supervised vs. fully supervised. This experiment studies the impact of using unlabeled data. We trained two models, one only with the available labeled data (fully supervised) and the other also with the unlabeled data. In the first case we used $1,020$ labeled pairs, and in the second we used the additional $5,616$ unlabeled pairs.

Table 8 shows that for all sizes of training sets, the results achieved by the semi-supervised training outperforms those of the supervised training. Thus, additional input, even if unlabeled, should be used. Furthermore, the fewer the labeled data is, the more beneficial semi-supervision is. This is so since when having fewer labeled pairs, but the same number of unlabeled pairs, their influence grows.

Shared encoder vs. separate encoders. Our approach employs 2 encoders and achieves accuracy of 89.6%. If instead we used a shared encoder the accuracy decreases to 85.15%.

Hyper-parameters. These are quite robust. Specifically, in the loss function, if we change the most important γ_1 (similarity), which is 0.8, by ± 0.1 (at the expense of γ_2), the accuracy change will be limited to 0.5%. Increasing α over β affects the generation; however changing their values from $(0.3, 0.7)$ to $(0.4, 0.6)$ for instance, the impact will be almost invisible. More details are given in the supplementary.

Limitations. Figure 8 shows cases where our model fails to classify the objects correctly. In these cases, the artifacts are worn out and are erroneously classified into related classes.

Figure 9 shows cases where our generated drawings do not succeed to generate the delicate textures. It can be seen though, that our drawings are still better than those of [49].

(a) lion (b) beetle (c) ibex

Figure 8. **Classification limitation.** These three worn-out artifacts are classified erroneously: the lion as as anthropomorphic (a), the beetle as ankh (b) and the ibex as lion (c).

Input Manual Ours [49]

Figure 9. **Image-to-drawing limitation.** Our method might fail to draw some fine details, such as the buckles on the anthropomorphic belt and some decorations on the lion's body and tail. Still, our results preserve most of the original details, compared to [49].

7. Conclusions

In archaeology, artifacts are studied using their photographs. For a subset of them, drawings are made by trained drafts people. These are experts in their field and are able to complete features that are not visible due to the artifacts eroded condition. The challenge addressed in this paper is how such data can be used to solve classification in the case of a small damaged dataset. We show that implicit knowledge obtained from a drawing is transferred to an image, guiding it to the more important features. Furthermore, we show that performing the training in a semi-supervised way, takes advantage of unlabelled image-drawing pairs.

In addition, our model generates from the image a drawing of the object. This is challenging since the image and drawing are not exactly aligned and what is transferred is the approximate position of the image features. The resulting model is able to mimic with high accuracy the knowledge and the drawing expertise of the artist, as well as the knowledge of the archaeologist. Our method can be generalized to objects of reliefs, either archaeological or artistic.

Last but not least, we created a relatively large and challenging dataset, which can be used in future research.

ACKNOWLEDGMENTS. We gratefully acknowledge the support of the Israel Science Foundation (ISF) 1083/18 and Ministry Science and Technology (MOST) 3-17513.

References

[1] Anton Alekseev. alexeyev/glyphnet-pytorch: GlyphNet, PyTorch implementation. https://github.com/alexeyev/glyphnet-pytorch, 2021. 5

[2] Hafeez Anwar, Saeed Anwar, Sebastian Zambanini, and Fatih Porikli. Deep ancient roman republican coin classification via feature fusion and attention. *Pattern Recognition*, 114:107871, 2021. 1, 2, 3, 5, 6

[3] Hafeez Anwar, Serwah Sabetghadam, and Peter Bell. An image-based class retrieval system for roman republican coins. *Entropy*, 22(8):799, 2020. 2

[4] Ognjen Arandjelović. Automatic attribution of ancient roman imperial coins. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1728–1734, 2010. 2

[5] Google Arts and Culture. Google arts fabricius workbench dataset. https://github.com/googleartsculture/workbench, 2022. 6

[6] Andrea Barucci, Costanza Cucci, Massimiliano Franci, Marco Loschiavo, and Fabrizio Argenti. A deep learning approach to ancient egyptian hieroglyphs classification. *Ieee Access*, 9:123438–123447, 2021. 1, 2, 5, 6

[7] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4247–4256, 2021. 3

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 3

[9] Virginio Cantoni, Mauro Mosconi, and Setti Alessandra. Javastylosis: a tool for computer-assisted chromatic and semantics based anastylosis of frescoes. In *Proceedings of the 21st International Conference on Computer Systems and Technologies' 20*, pages 208–214, 2020. 2

[10] Mario Canul-Ku, Rogelio Hasimoto-Beltran, Diego Jiménez-Badillo, Salvador Ruiz-Correa, and Edgar Román-Rangel. Classification of 3d archaeological objects using multi-view curvature structure signatures. *IEEE Access*, 7:3298–3313, 2018. 1

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 3

[12] Aladine Chetouani, Sylvie Treuillet, Matthieu Exbrayat, and Sébastien Jesset. Classification of engraved pottery sherds mixing deep-learning features by compact bilinear pooling. *Pattern Recognition Letters*, 131:1–7, 2020. 1

[13] Ye-Chan Choi, Sheriff Murtala, Beom-Chae Jeong, and Kang-Sun Choi. Relief extraction from a rough stele surface using svm-based relief segment selection. *IEEE Access*, 9:4973–4982, 2020. 2

[14] Ye-Chan Choi, Sheriff Murtala, Beom-Chae Jeong, and Kang-Sun Choi. Deep learning-based engraving segmentation of 3-d inscriptions extracted from the rough surface of ancient stelae. *IEEE Access*, 9:153199–153212, 2021. 2

[15] Jessica Cooper and Ognjen Arandjelović. Learning to describe: A new approach to computer vision based ancient coin analysis. *Sci*, 2(2):27, 2020. 1, 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[17] Niv Derech, Ayellet Tal, and Ilan Shimshoni. Solving archaeological puzzles. *Pattern Recognition*, 119:108065, 2021. 2

[18] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2179–2188, 2019. 3

[19] Nishanth Dikkala, Sankeerth Rao Karingula, Raghu Meka, Jelani Nelson, Rina Panigrahy, and Xin Wang. Sketching based representations for robust image classification with provable guarantees. *Advances in Neural Information Processing Systems*, 35:5459–5470, 2022. 3

[20] Lauren Fay. Fayrose middle egyptian dataset. https://github.com/fayrose/MiddleEgyptianDataset, 2021. 6

[21] Morris Franken and Jan C van Gemert. Automatic egyptian hieroglyph recognition by retrieving images as texts. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 765–768, 2013. 3, 6

[22] Ayelet Gilboa, Ayellet Tal, Ilan Shimshoni, and Michael Kolomenkin. Computer-based, automatic recording and illustration of complex archaeological artifacts. *Journal of Archaeological Science*, 40(2):1329–1339, 2013. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5, 6

[25] Estíbaliz Iglesias-Franjo and Jesús Vilares. Tir over egyptian hieroglyphs. In *2016 27th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 198–203, 2016. 2

[26] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems*, 34:3584–3595, 2021. 3

[27] Barak Itkin, Lior Wolf, and Nachum Dershowitz. Computational ceramicology. *arXiv preprint arXiv:1911.09960*, 2019. 2

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[29] Stephan Karl, Peter Houska, Stefan Lengauer, Jessica Haring, Elisabeth Trinkl, and Reinhold Preiner. Advances in digital pottery analysis. *it-Information Technology*, 2022. 2

[30] Michael Kolomenkin, George Leifman, Ilan Shimshoni, and Ayellet Tal. Reconstruction of relief objects from archeological line drawings. *Journal on Computing and Cultural Heritage (JOCCH)*, 6(1):1–19, 2013. 2

[31] Michael Kolomenkin, Ilan Shimshoni, and Ayellet Tal. On edge detection on surfaces. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2767–2774. Ieee, 2009. 2

[32] Kai Lawonn, Erik Trostmann, Bernhard Preim, and Klaus Hildebrandt. Visualization and extraction of carvings for heritage conservation. *IEEE transactions on visualization and computer graphics*, 23(1):801–810, 2016. 2

[33] Stefan Lengauer, Reinhold Preiner, Ivan Sipiran, Stephan Karl, Elisabeth Tinkl, Benjamin Bustos, and Tobias Schreck. Context-based surface pattern completion of ancient pottery. In *20th Eurographics Workshop on Graphics and Cultural Heritage: GCH 2022*, 2022. 2

[34] Yuanyuan Ma and Ognjen Arandjelović. Classification of ancient roman coins by denomination using colour, a forgotten feature in automatic ancient coin analysis. *Sci*, 2(2):37, 2020. 1, 2

[35] Gil Melnik, Yuval Yekutieli, and Andrei Sharf. Deep segmentation of corrupted glyphs. *J. Comput. Cult. Herit.*, 15(1), jan 2022. 2

[36] Xiaolei Niu, Qifeng Wang, Bin Liu, and Jianxin Zhang. An automatic chinaware fragments reassembly method framework based on linear feature of fracture surface contour. *ACM Journal on Computing and Cultural Heritage*, 16(1):1–22, 2022. 2

[37] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 3

[38] Abraham Resler, Reuven Yeshurun, Filipe Natalio, and Raja Giryes. A deep-learning model for predictive archaeology and archaeological community detection. *Humanities and Social Sciences Communications*, 8(1):1–10, 2021. 1, 2, 3, 5

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[40] Serge Rosmorduc. Jsesh dataset. https://github.com/rosmord/jsesh, 2022. 6

[41] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023. 3

[42] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8504–8513, 2021. 3

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[44] Xi Shen, Ilaria Pastrolin, Oumayma Bounou, Spyros Gidaris, Marc Smith, Olivier Poncet, and Mathieu Aubry. Large-scale historical watermark recognition: dataset and a new consistency-based approach. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6810–6817. IEEE, 2021. 2

[45] Ivan Sipiran. Completion of cultural heritage objects with rotational symmetry. In *Proceedings of the 11th Eurographics Workshop on 3D Object Retrieval*, pages 87–93, 2018. 2

[46] Ivan Sipiran, Patrick Lazo, Cristian Lopez, Milagritos Jimenez, Nihar Bagewadi, Benjamin Bustos, Hieu Dao, Shankar Gangisetty, Martin Hanik, Ngoc-Phuong Ho-Thi, et al. Shrec 2021: Retrieval of cultural heritage objects. *Computers & Graphics*, 100:1–20, 2021. 2

[47] Ivan Sipiran, Alexis Mendoza, Alexander Apaza, and Cristian Lopez. Data-driven restoration of digital archaeological pottery with point cloud analysis. *International Journal of Computer Vision*, 130(9):2149–2165, 2022. 2

[48] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 3

[49] X. Soria, E. Riba, and A. Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1912–1921, Los Alamitos, CA, USA, mar 2020. IEEE Computer Society. 7, 8

[50] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5, 6

[51] Haiping Wang, Yufu Zang, Fuxun Liang, Zhen Dong, Hongchao Fan, and Bisheng Yang. A probabilistic method for fractured cultural relics automatic reassembly. *J. Comput. Cult. Herit.*, 14(1), jan 2021. 2

[52] Xinggang Wang, Xiong Duan, and Xiang Bai. Deep sketch feature for cross-domain image retrieval. *Neurocomputing*, 207:387–397, 2016. 3

[53] Josef Wilczek, Fabrice Monna, Ahmed Jébrane, Catherine Labruère Chazal, Nicolas Navarro, Sébastien Couette, and Carmela Chateau Smith. Computer-assisted orientation and drawing of archaeological pottery. *Journal on Computing and Cultural Heritage (JOCCH)*, 11(4):1–17, 2018. 2

[54] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. Deep co-training with task decomposition for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8906–8916, 2021. 3

[55] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019. 3

[56] Jeongbeen Yoon, Dahyun Kang, and Minsu Cho. Semi-supervised domain adaptation via sample-to-sample self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1978–1987, 2022. 3

[57] Matthias Zeppelzauer, Georg Poier, Markus Seidl, Christian Reinbacher, Samuel Schulter, Christian Breiteneder, and Horst Bischof. Interactive 3d segmentation of rock-art by enhanced depth maps and gradient preserving regularization. *Journal on Computing and Cultural Heritage (JOCCH)*, 9(4):1–30, 2016. 2

[58] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1105–1113, 2016. 3