Future Technology
Open Access Journal
ISSN 2832-0379

February 2027| Volume 06 | Issue 01 | Pages 51-66

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.6.1.5

Future Publishing LLC

Article

An explainable deep learning framework for intelligent documentation of silk textile heritage: cross-cultural pattern recognition and shortcut auditing on open museum collections

Yanrong Li[1], Lin Zhu[1]*, Penpisut Sikakaew[2]

[1]Department of Creative Design, Chinese International College, Dhurakij Pundit University, Bangkok, Thailand
[2]Department of Eastern Languages, Faculty of Humanities, Kasetsart University, Bangkok, Thailand

ARTICLE INFO

Article history:
Received 25 July 2026
Received in revised form
30 October 2026
Accepted 21 November 2026

Keywords:
Explainable artificial intelligence, Deep learning, Grad-CAM, Transfer learning, Cultural heritage, Silk textiles

*Corresponding author
Email address:
lin.zhu@dpu.ac.th

DOI: 10.55670/fpll.futech.6.1.5

ABSTRACT

Open museum collections allow deep learning to support the documentation of historical textiles, but photographs taken over decades carry conventions that a model can learn in place of the objects. This study presents an explainable deep learning framework for the intelligent documentation of silk textile heritage and applies it to 11,388 open silk textile images from The Metropolitan Museum of Art and the Cleveland Museum of Art. Images of nine cultural origins were represented with frozen features from a residual network (ResNet-50), a vision transformer (ViT-B/16), and two foundation models, DINOv2 and CLIP. Ridge multinomial classifiers were evaluated with five repetitions of cross-validation grouped by object and near-duplicate photograph, and the ridge estimates for the main settings and backbones were reproduced in R. Gradient-weighted class activation mapping (Grad-CAM), supported by a deletion test, showed that the initial model attended in some classes to accession labels, color calibration charts, and the black-and-white format of archive photographs. After restricting the data to color photographs and trimming image borders, ResNet-50 features recognized origin with a balanced accuracy of 0.577 (95% CI 0.561 to 0.593) (chance 0.111). DINOv2 and CLIP features raised this to 0.655 and 0.641, respectively, and partial fine-tuning of ResNet-50 to 0.624. Baselines built on date, museum, image size, or background reached at most 0.389, and weaving technique was recognized with a balanced accuracy of 0.722 across four classes (0.709 after the same audit). Explainability auditing should therefore precede the use of open museum images in heritage recognition and design tools.

1. Introduction

Historical silk textiles record centuries of technical skill and of exchange along the trade routes that linked East Asia, the Islamic world and Europe. Many museums now publish these objects as open digital collections, and computer vision makes it possible to organize and compare them at a scale that curatorial work alone cannot reach [1,2]. Deep learning has been applied to paintings and artistic metadata [3-5], ethnic textile patterns in China [6-8], handloom fabrics in India [9,10], and the woven ikat of Japan and Indonesia [11,12]. Open records released under Creative Commons Zero (CC0) terms [13,14] let anyone rebuild a training set, but the photographs were taken over decades, with changing conventions that a model may learn alongside the objects. Silk heritage is a demanding test case. Song brocade and other Chinese traditions have been studied in detail [15], and Persian, Ottoman and Italian silks developed their own motif vocabularies through trade. The European SILKNOW project predicted production place, period, and technique for European silk records [16-18], yet recognition across Asian and European silks in open collections has received far less attention. Deep networks also exploit shortcuts, features that correlate with the label but carry no meaning for the task [19]. Pneumonia detectors learned to recognize the hospital system that produced an image [20], COVID-19 detectors relied on text markers [21], and melanoma classifiers reacted to skin markings [22]. Museum photographs contain similar cues, such as accession numbers, color charts, mounts, and the photographic process itself, and gradient-weighted class activation mapping (Grad-CAM) can reveal them when used

as an audit step [23-25]. This study develops and evaluates an explainable deep learning framework for the intelligent documentation of silk textile heritage, built on two open museum collections. Its objectives are as follows.

- To assemble a curated, reproducible set of labeled silk textile images from nine cultural origins, with evaluation grouped by object and near-duplicate photograph.
- To audit model decisions with Grad-CAM, qualitatively and quantitatively, and to measure how far photographic conventions inflate apparent accuracy.
- To estimate how well cultural origin and weaving technique are recognized once these shortcuts are removed, with confidence intervals, and to test these estimates against confounds of date, museum, image properties and object type.
- To compare convolutional, transformer, self-supervised and vision-language features and several classifiers under one protocol, and to reproduce the main ridge estimates in two independent software environments.

2. State of the art and problem analysis

2.1 Computer vision for textile and art heritage

Machine learning for cultural heritage covers classification and retrieval [1,2], fine-tuned networks classify paintings by artist and style [3], and benchmarks such as the Rijksmuseum Challenge and OmniArt opened large art collections to model development [4,5]. Textile work is a smaller strand. Tujia brocade images were segmented with Gaussian mixture clustering [6], Zhuang brocade patterns were generated automatically [7], and Yi patterns were remodeled with a generative adversarial network [8]. Indian handloom fabrics have been recognized from images [9,10]. Outside Asia, SILKNOW predicted the place, period, and technique of European silks [16] and combined image, text, and tabular data to fill missing museum metadata [18]. Kurume Kasuri ikat was graded with an EfficientNet classifier [11]. Most of these studies treat a single tradition, and few examine which parts of an image drive the prediction.

2.2 Pattern retrieval and foundation-model features

Retrieval is the second main use of image features in textile work. Hierarchical search over learned binary codes was proposed for large fabric databases [26], fine-grained similarity was defined for fabric retrieval in production [27], descriptors with an auxiliary classification loss were used to search European silk collections [17], and a retrieval network for Indonesian woven fabrics outperformed seven pretrained networks [12].

text pairs [28], and DINOv2 learns them through self-supervision [29]. Both have been applied to artwork classification [30] and retrieval [31], and in decorative-art similarity search their activation maps often fell on background and framing instead of the decoration [32]. Whether these features separate silk traditions better than ImageNet features, and whether they inherit the same photographic shortcuts, has not been tested.

2.3 Transfer learning, explainability and shortcut learning

Off-the-shelf convolutional features are a strong baseline [33,34], ImageNet accuracy predicts transfer performance [35,36], and residual networks [37] now compete with vision transformers [38]. Explainable artificial intelligence aims to make model behavior inspectable [25,39]. Class activation mapping showed that classification networks also localize the evidence for each class [40], and Grad-CAM generalized the idea through gradients of the class score [23]. Post hoc explanations can mislead [41], so heat maps are best paired with a quantitative faithfulness test such as deleting the pixels they rank highest [42]. Explanation methods have uncovered predictors that relied on watermarks [24], and dataset bias [43] and metadata shortcuts [44] are well documented. Results are also inflated when near-identical images fall into both training and test sets [45] or when models are selected on the test folds [46], and class imbalance calls for balanced accuracy [47].

2.4 Problems addressed in this study

Three gaps follow. Recognition of silk textiles across Asian and European traditions has not been tested on open museum collections with object-grouped evaluation. No study has examined, with qualitative and quantitative evidence, whether museum photographic conventions act as shortcuts for heritage textile classification. The value of convolutional, transformer, self-supervised, and vision-language features, and the roles of date, museum, and image properties as confounders, also remain untested for this domain.

3. Materials and methods

Figure 1 summarizes the framework. Open museum records were curated into labeled image sets. The images were converted into several variants, encoded with frozen pretrained networks and classified under grouped cross-validation. Grad-CAM heat maps were used to audit the decisions, and the audit informed the image variants used for the final estimates.

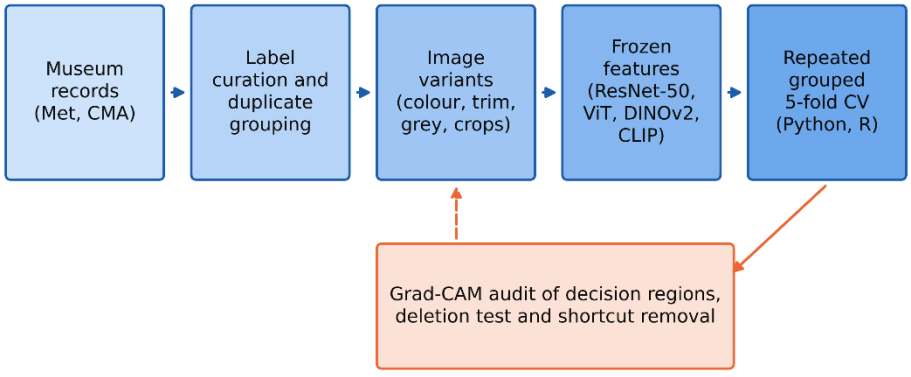


Figure 1. Workflow of the explainable deep learning framework for silk textile documentation. CMA, Cleveland Museum of Art; CV, cross-validation

The Metropolitan Museum of Art (Met) file downloaded on 24 September 2026 contained 484,956 object records [13], of which 13,404 were public domain textiles with silk named in the medium. The Cleveland Museum of Art (CMA) file contained 68,778 records under a CC0 dedication [14], of which 1,227 were silk textiles with a web image. Both files are identified by their SHA-256 checksums. Images came from the museums' public image services. Of the 11,426 study records, 11,420 returned a usable image. The selection rule does not exclude objects that also contain other fibers. In the nine-class set, 1,320 records name cotton, linen, wool, hemp, ramie, or rayon in the medium besides silk, mostly silk embroidery or patterning on another ground, linen and silk fragments from Egypt, or wool and silk carpets. They were kept in the main analysis and removed in a sensitivity run. Museum object names were also grouped into four types, namely fragments and samples, garments and dress accessories, furnishings and liturgical textiles, and other objects such as tassels and small cases.

3.2 Label definitions and curation

Cultural origin was assigned from the museums' culture and country fields, so the labels are catalog categories that mix geography, culture, and period. China, Japan, India, Italy, France, and Spain are country names used for the place of production. Iran also covers records cataloged as Persian. The Ottoman or Turkey class joins both attributions, because the catalogs use both names for the same centers, such as Bursa and Istanbul. The Byzantine or Egypt class joins Byzantine records with records from Egypt, most of them fragments from burial sites of the late Roman, Byzantine, and early Islamic periods. A record was labeled only when exactly one study origin was named, and no other region appeared. Records naming two regions (524), uncertain attributions such as “possibly” or “in the style of” (566), other origins (2,105), and sample book pages (10) were excluded. Central Asia, with 32 images, was too small and was left out, so the main task used nine classes. Weaving techniques were grouped into woven patterned silks, embroidery, velvet, and tapestry for a second task with 10,466 images. Figure 2 shows two color examples of each origin. Several include color bars, accession numbers, or mounts at the edges.

3.3 Photographic condition and image variants

Many museum images are scans of older black-and-white photographs. An image was classed as black-and-white when its mean summed absolute difference between the three-color channels was below 3 on a 0 to 255 scale. Figure 3 shows why this value was chosen. The distribution has two well-separated modes, one at 0 for scans saved in a color format and one near 90 for color photographs. Only 11 images fall between 2 and 5, so any threshold in that range gives almost the same split. In total, 3,725 of the 11,388 images (32.7%) were black-and-white, with shares that varied widely by origin (Table 1). The rule cannot catch slightly toned prints. A visual check of all 212 color-classed images with values below 20 found about 30 such prints, all with values below 12 and together equal to 0.4% of the color set.

Five image settings were defined (Table 2). The original setting used every image, and the color setting kept only the 7,663 color photographs. The trimmed setting kept the central 70% of each image's width and height, removing the outer 15% band where labels, rulers, and color charts usually appear. The band width was chosen after the Grad-CAM audit of the initial model showed responses to labels, charts, and mounts in that band (Section 4.2). Because cropping also removes genuine patterns, central crops of 60% and 80% and a random 70% crop placed anywhere in the image were run as checks on the first repetition of the folds. The two greyscale settings converted trimmed images to luminance, for all images or for color photographs only. All images were first downscaled to 320 pixels on the shorter side.

3.4 Duplicate grouping and evaluation design

Parts of one object were first grouped by accession number. Near-duplicate images were then detected as pairs whose ResNet-50 feature vectors (Section 3.5) had a cosine similarity above 0.97 and merged into shared groups by a union-find pass. The threshold was set by inspecting pairs on either side of it. Almost all pairs above 0.97 were the same photograph filed under two records or a second photograph of the same object. Pairs between 0.95 and 0.97 were mostly separate objects of one design photographed in the same way, such as matching chairs or a pair of cushions. At 0.97, 263 image pairs exceeded the threshold, and 179 merges reduced 11,232 accession groups to 11,053 image groups. Thresholds of 0.95 and 0.99 and grouping by accession number alone were tested on the first repetition of the folds. Performance was estimated with five-fold cross-validation stratified by class and grouped by image group, repeated with five fold assignments (seeds 42 to 46). The same 25 splits were used for every model, setting, and software environment, and point estimates are the mean over the 25 test folds.

3.5 Feature extraction

Four backbones pretrained without silk images served as fixed feature extractors. ResNet-50 [37], trained on ImageNet [36], used the torchvision weights IMAGENET1K_V2 (file resnet50-11ad3fa6.pth) and gave 2,048 features. The vision transformer ViT-B/16 [38] used the torchvision weights IMAGENET1K_V1 (file vit_b_16-c867db91.pth) and gave a 768-dimensional class token. DINOv2 ViT-B/14 (vit_base_patch14_dinov2.lvd142m) [29] and the image encoder of CLIP ViT-B/16 with the weights released by its authors [28] were loaded through the PyTorch Image Models library [48], each giving 768 features. All four received the same preprocessing (resizing to 232 pixels, a central 224-pixel crop, and each model's own normalization) in PyTorch 2.6.0 [49]. To test whether adaptation to silk images adds to frozen features, the last residual stage of ResNet-50 (layer4) and a new output layer were fine-tuned on the trimmed color photographs of each training fold of the first repetition. Training used four epochs of AdamW with a one-cycle schedule (peak 0.0001 for layer4 and 0.001 for the output layer), class-weighted cross-entropy and frozen batch-normalization statistics.

3.6 Classifiers and two-package verification

The main classifier was a multinomial logistic regression with a ridge penalty. For image i with standardized features xi and class k, it minimizes the weighted penalized loss in Eq. (1).

$$L = -\sum_i w_i \log p(y_i \mid x_i) + \| W \|^2/(2C) \quad (1)$$

Here w_i is a class weight inversely proportional to class frequency, W holds the coefficients, and C sets the penalty strength. A linear classifier on frozen features measures how much task information a representation already holds, and its coefficients serve as the output layer for Grad-CAM. C was chosen from 0.0003, 0.001, 0.003, and 0.01 by a grouped three-fold search inside each training fold [46].



Figure 2. Two-color examples of each origin from the study set, cropped to a square. Labels give the museum and record numbers. CMA, Cleveland Museum of Art. Byz./Egypt, Byzantine and Egyptian silks.

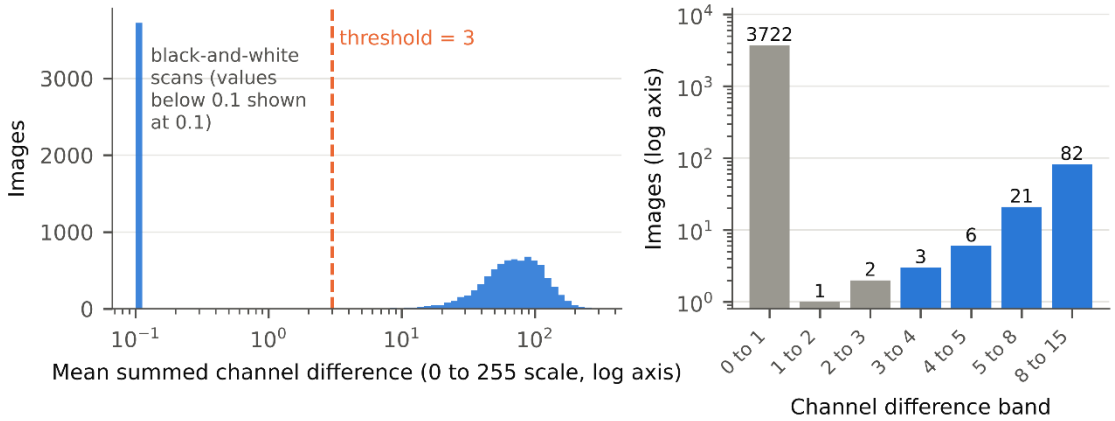


Figure 3. Distribution of the mean summed absolute difference between color channels for the 11,388 nine-class images (left, logarithmic axis) and the number of images in narrow bands around the black-and-white threshold of 3 (right). Grey bars mark images classed as black-and-white.

Table 1. Composition of the nine-class image set by origin and photographic format. The color images form the color and trimmed settings [13,14]

Origin	All images	Black-and-white	Black-and-white share (%)	Color images
Japan	3,677	368	10.0	3,309
France	2,313	1,338	57.8	975
Italy	2,239	829	37.0	1,410
China	1,509	727	48.2	782
Spain	439	196	44.6	243
Iran	362	59	16.3	303
Byzantine and Egypt	319	81	25.4	238
Ottoman	310	98	31.6	212
India	220	29	13.2	191
Total	11,388	3,725	32.7	7,663

Four additional classifiers were fitted on the ResNet-50 features in the main setting, using the folds from the first repetition and the same inner search. A random forest [50] with 500 trees was tuned over leaf size (1 or 5), depth (unlimited or 20), and class weighting (whole sample or each bootstrap sample). A radial-kernel support vector machine [51] was tuned over C (1 or 10), a distance-weighted k-nearest-neighbor classifier [52] over k (5, 15 or 31), and a perceptron with one hidden layer of 256 units over its weight decay (0.001 or 0.01). An untuned forest was also fitted to every setting (Table 2). The ridge models, the support vector machine, the nearest-neighbor classifier and the random forests were fitted in both Python, with scikit-learn 1.8.0 [53], and R 4.3.3, with glmnet 4.1-8 [54], ranger 0.16.0 [55], e1071 1.7-14 and FNN 1.1.4. The perceptron was fitted in Python only. The settings chosen in Python for each fold were passed to R, and agreement between the two environments served as a check on the implementation.

3.7 Baselines and confound checks

Several baselines used the same folds. One used 148 hand-crafted color and texture statistics. A second used only the object's start and end dates with their squares and cubes, and a combined model added these terms to the image features. Others used only the source museum, only image properties (width, height, aspect ratio, and file size per pixel), or only the mean color, brightness, and contrast of the outer border band. Balanced recall of the image model was also computed within five date bands, within each museum and within each object type, and within each date band the date-only model was evaluated on the same images.

3.8 Grad-CAM visual explanation

Grad-CAM was applied to the final convolutional block of ResNet-50 [23], with a ridge classifier trained on all images of the setting as the output layer. For class c and feature map A^k, the channel weight is the spatial mean of the gradient of the class score y^c in Eq. (2). The heat map is the rectified weighted sum in Eq. (3).

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

$$L^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{3}$$

Heat maps were produced for the most confident correct images of each origin in the original and main settings, the most confident errors of the main model, and the correct predictions and main confusions of the technique model. For the initial model, the share of the top 10% of heat-map mass in the outer 15% band was computed for 450 correct images, 50 per origin, and compared with the band's share of the input. For both models, a deletion test [42] replaced the 10% of pixels with the highest heat-map values with the image mean and recorded the drop in the probability of the true class. Two controls deleted the same region, shifted to a random position, or deleted the 10% of pixels with the lowest values. Nearest-neighbor retrieval served as a second diagnostic, recording for every image the share of its five nearest images outside its own group that shares its origin.

3.9 Evaluation metrics and statistical tests

Balanced accuracy was the primary metric because class sizes ranged from 191 to 3,309 images [47]. It is the mean of per-class recall in Eq. (4), where K is the number of classes, TP_k is the number of correctly classified images of class k, and Nk is the number of images of class k. Within subsets, such as one museum or one date band, the same mean over the classes present is reported as balanced recall.

$$\mathrm{BA} = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{TP}_k}{N_k} \tag{4}$$

Chance balanced accuracy is 0.111 for nine classes. Confidence intervals came from a grouped percentile bootstrap [56] that resampled whole image groups 2,000 times. Two settings or models evaluated on the same images were compared with a paired bootstrap and, for comparisons over all five repetitions, with the corrected resampled t-test [57], which allows for overlapping training sets. The area under the receiver operating characteristic curve (AUC) was computed for each class against the rest, also in R. Calibration was summarised with the expected calibration error (ECE) in Eq. (5), with M = 10 equal-width bins B_m of the predicted probability.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{5}$$

Because class weighting and the ridge penalty pull probabilities toward the center, probabilities were recalibrated inside each training fold of the first repetition by temperature scaling [58] and by isotonic regression for each class [59], both fitted on grouped out-of-fold predictions. A power law in Eq. (6), with asymptote a, was fitted to a learning curve on 10% to 100% of the training groups [60]. Label noise was simulated [61] by randomly reassigning 5% or 10% of the training labels before refitting.

$$\mathrm{BA}(n) = a - b\, n^{-c} \tag{6}$$

4. Results and discussion

4.1 Dataset composition

The nine-class set contained 11,388 images, led by Japan (3,677), France (2,313), Italy (2,239) and China (1,509). The black-and-white share ranged from 10.0% for Japan to 57.8% for France (Table 1). This pattern reflects how each collection was photographed and gives a model an easy cue unrelated to the textile. Most color photographs show fragments and samples (84%), and the Met supplied 89% of them. The technique set contained 6,771 woven patterned silks, 2,040 embroideries, 1,129 velvets, and 526 tapestries, with tapestries concentrated in the Byzantine and Egyptian group.

4.2 Grad-CAM audit of the initial model

The initial ResNet-50 model trained on all original images reached a balanced accuracy of 0.585 (95% CI 0.571 to 0.598). Several of its heat maps concentrate on regions outside the textile (Figure 4), such as a painted accession number on an Ottoman fragment, a color calibration strip at the edge of a French sample, and plain borders and mounts in Japanese and Italian examples. Across 450 correct images, the outer 15% band held 31% of the top 10% of heat-map mass, close to its 32% share of the 224-pixel network input. For Italian and Japanese silks, however, it held 49% and 52%, whereas for Chinese silks and the Byzantine and Egyptian fragments it held only 12% and 9%. Deleting the top 10% of heat-map pixels lowered the probability of the true class by 0.146, compared with 0.061 for the shifted region and 0.005 for the lowest-ranked pixels (Table 3).

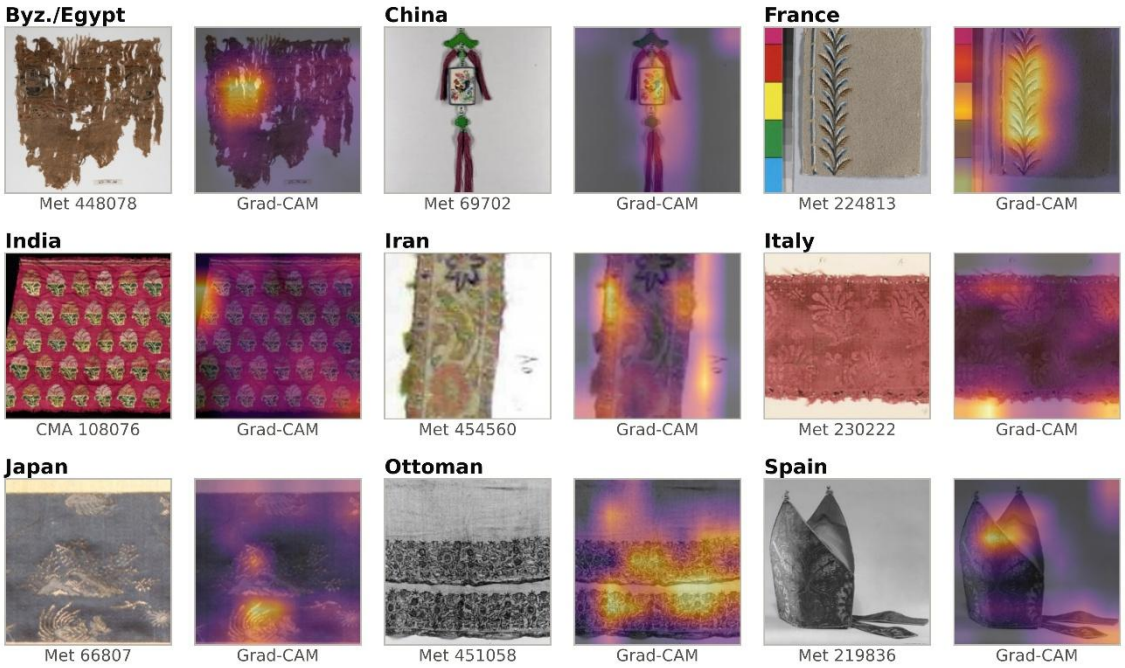


Figure 4. Grad-CAM heat maps of the initial model on original images, one high-confidence correct prediction for each origin. Brighter overlay marks the regions that most increased the class score. CMA, Cleveland Museum of Art. Byz./Egypt, Byzantine and Egyptian silks.

The predictions point the same way. Black-and-white images were assigned to France in 35.0% of cases and to China in 20.1%, the two classes with the highest black-and-white shares. Recall for Japanese textiles fell from 0.864 in color to 0.343 in black-and-white, while recall for Chinese textiles rose from 0.479 to 0.735. Byzantine and Egyptian fragments were recognized about equally well in both formats (0.803 and 0.835), probably because their tapestry bands and damaged outlines remain visible without color. Photographic format therefore acted as a shortcut [19], much as markers and borders affected clinical models [20-22].

4.3 Recognition after shortcut removal

Restricting the data to color photographs gave a balanced accuracy of 0.600 (95% CI 0.584 to 0.615), and trimming the borders lowered it to 0.577 (0.561 to 0.593) (Table 2). The paired difference of 0.023 (95% CI 0.009 to 0.036, corrected t-test $p = 0.011$) is consistent with labels, charts, and mounts near the edge carrying some class information. Greyscale conversion lowered balanced accuracy to 0.5207 (0.505 to 0.537) for color photographs and 0.5211 (0.508 to 0.534) for all images. The greyscale drop on the same color photographs, 0.056 (0.043 to 0.070, $p < 0.001$), is expected because dye color is a genuine cultural attribute, yet accuracy remained 4.7 times the chance level. The hand-crafted statistics reached 0.400, well below the deep features. The trimmed color setting is taken as the main result because it removes the identified shortcuts while keeping color, and its five repetitions ranged from 0.574 to 0.578. On the first repetition, where the central 70% crop gave 0.574, a random 70% crop that often keeps part of the border gave 0.582 (0.564 to 0.599), a difference of +0.007 (-0.007 to 0.022, $p = 0.264$), and central crops of 60% and 80% gave 0.569 and 0.592. Point estimates thus rose with the share of the image kept, and keeping border material did not significantly change accuracy.

This is consistent with much of the loss after trimming coming from the removal of image content as such, part of which is genuine pattern near the edges. The trimmed estimate is therefore conservative, and the drop of 0.023 cannot be attributed to the border shortcut alone. Figure 5 breaks these results down by class. Trimming lowered recall most for India (by 0.091), Italy (by 0.051), and Japan (by 0.048), whereas Ottoman silks gained 0.019. Greyscale conversion cost Ottoman silks 0.137 and Iranian silks 0.079, and French silks changed by only 0.008. Class-level changes of this size lie within the confidence intervals of Figure 5 and are best read as descriptive.

4.4 Class-level behavior and visual evidence

Figure 6 shows the confusion matrix for the main result, with the 95% confidence interval for each class recall. Recall was highest for Japan (0.797), the Byzantine and Egyptian group (0.761), and France (0.672), and lowest for Spain (0.385) and India (0.430). The intervals are widest for the classes with fewer than 250 color images, namely India, Ottoman silks, the Byzantine and Egyptian group, and Spain, where they span 12 to 13 points of recall, against 3 for Japan. Spanish silks were most often confused with Italian silks (0.18), and Indian textiles with Iranian (0.12) examples. The model also ranks the correct origin highly when its top choice is wrong, with a macro-averaged AUC of 0.891 in Python and 0.891 in R. Chinese textiles reached a recall of 0.546, and their most frequent confusion was with Japan (0.17). Several explanations fit this pattern. Chinese and Japanese weavers shared techniques and motifs, so the groups may be genuinely close, or the confusion could come from the collections, since 82% of the Chinese and 97% of the Japanese photographs come from the Met. Within Met photographs alone, 17% of Chinese images were assigned to Japan, the same share as overall (17%), so the mix of museums does not explain the pattern.

Table 2. Origin and technique recognition with ResNet-50 features under each image setting. Balanced accuracy of the ridge classifier is the mean over 25 test folds (five repetitions of grouped five-fold cross-validation) with its 95% grouped bootstrap confidence interval. The R columns and the untuned random forest use the first repetition.

Setting	Images	Ridge, Python	95% CI	Ridge, R	Forest, Python	Forest, R
All images, original	11,388	0.5845	0.571 to 0.598	0.5854	0.359	0.373
Color photographs, original	7,663	0.5996	0.584 to 0.615	0.5968	0.301	0.349
Color photographs, trimmed (main)	7,663	0.5769	0.561 to 0.593	0.5744	0.255	0.321
Color photographs, greyscale and trimmed	7,663	0.5207	0.505 to 0.537	0.5200	0.223	0.298
All images, greyscale and trimmed	11,388	0.5211	0.508 to 0.534	0.5268	0.302	0.330
Hand-crafted color and texture statistics	11,388	0.3996	0.387 to 0.413	0.3966	0.368	0.381
Weaving technique, all images (four classes)	10,466	0.7220	0.706 to 0.737	0.7286	0.359	0.438

Note: Balanced accuracy is given to four decimals so that settings with close values can be told apart. Image counts per origin are given in Table 1. CI, confidence interval. Chance-balanced accuracy is 0.111 for origin and 0.25 for technique.

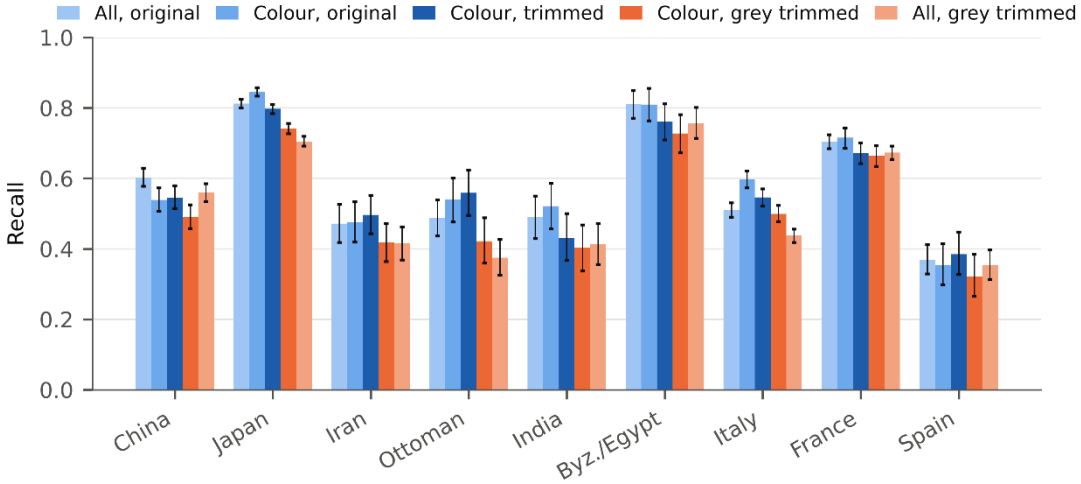


Figure 5. Recall of each origin under the five image settings, with 95% bootstrap confidence intervals. Byz./Egypt, Byzantine and Egyptian silks.

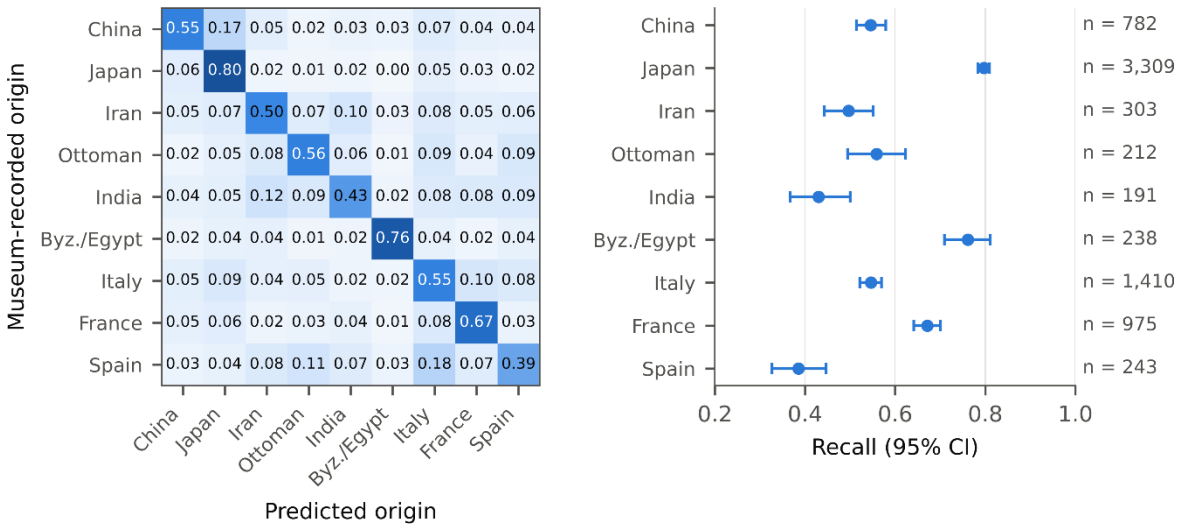


Figure 6. Row-normalized confusion matrix for color photographs with trimmed borders (left) and the recall for each origin, with its 95% bootstrap confidence interval and number of images (right). Byz./Egypt, Byzantine and Egyptian silks.

Within fragments and samples, the share rose to 32%, so the confusion is concentrated in small pieces, where layout and object form offer fewer clues. Photographic conditions shared within one museum cannot be separated from the objects in this way, so the result is best read as evidence that small Chinese and Japanese fragments are close in this feature space, without establishing why. In the main model, the strongest heat-map responses lie on motifs and pattern repeats, such as the medallions of an Ottoman silk, floral sprigs on French silks, and the lattice of a Japanese woven fragment (Figure 7). Deleting the top 10% of heat-map pixels lowered the probability of the true class by 0.118, compared with 0.041 for the shifted region and -0.005 for the lowest-ranked pixels, so the maps point to regions the model relies on.

The border-mass share does not apply to the main model because the band is removed before the network sees the image. The most confident errors give a fuller picture (Figure 8). A Japanese patchwork of brocade squares was assigned to China and a Chinese woven fragment to Japan, which fits the overlap above. Darkened and torn fragments from China, Spain and Iran were assigned to the Byzantine and Egyptian group with probabilities above 0.95, and their heat maps lie on discolored and damaged areas. An Italian tapestry with a pastoral scene was assigned to France. Condition and degradation therefore act as a residual cue that trimming cannot remove, because they belong to the object itself. Nearest-neighbor retrieval shows how much cross-cultural signal the frozen feature space already holds. With ResNet-50 features, 58.7% of the five nearest neighbors of an image shared its origin, compared with 25.2% expected by chance, and the mean across classes was 40.8%, compared with 11.1% expected by chance.

The share rose to 59.9% with ViT-B/16, 62.4% with CLIP, and 64.9% with DINOv2, in the same order as their balanced accuracy. In Figure 9, an Ottoman silk with circular medallions retrieves two Ottoman medallion silks, and then Iranian and Japanese silks with similar repeats, and a Song-period Chinese tapestry retrieves three Chinese silks after two Japanese ones. The feature space thus groups silks by layout and palette first, which suits similarity search for design. Before recalibration, the model was under-confident, with an ECE of 0.149. Temperature scaling, with temperatures between 0.64 and 0.65, reduced the ECE to 0.023 without changing any prediction, and isotonic regression reduced it to 0.038 at the cost of 3.1 points of balanced accuracy. After temperature scaling, predictions with a probability of at least 0.8 covered 2,718 images (35.5%) and were correct in 90.0% of cases, and those of at least 0.9 covered 1,736 images with 92.5% accuracy. A documentation system could accept these suggestions and route the rest to experts.

4.5 Robustness to confounds, duplicates and label noise

Table 3 collects the checks on confounds and design choices. The source museum alone gave a balanced accuracy of 0.148, image size and file properties 0.216, and the color and contrast of the border band 0.242. Each lies above chance, because museums, photographic campaigns and mounting styles are unevenly spread across origins, but far below the image model. Within museums, the main model reached 0.577 on Met photographs and 0.493 on CMA photographs, the latter from only 868 images. Balanced accuracy was 0.565 for fragments and samples, lower for garments and dress accessories (0.373) and for furnishings and liturgical textiles (0.429), and 0.478 for other objects.

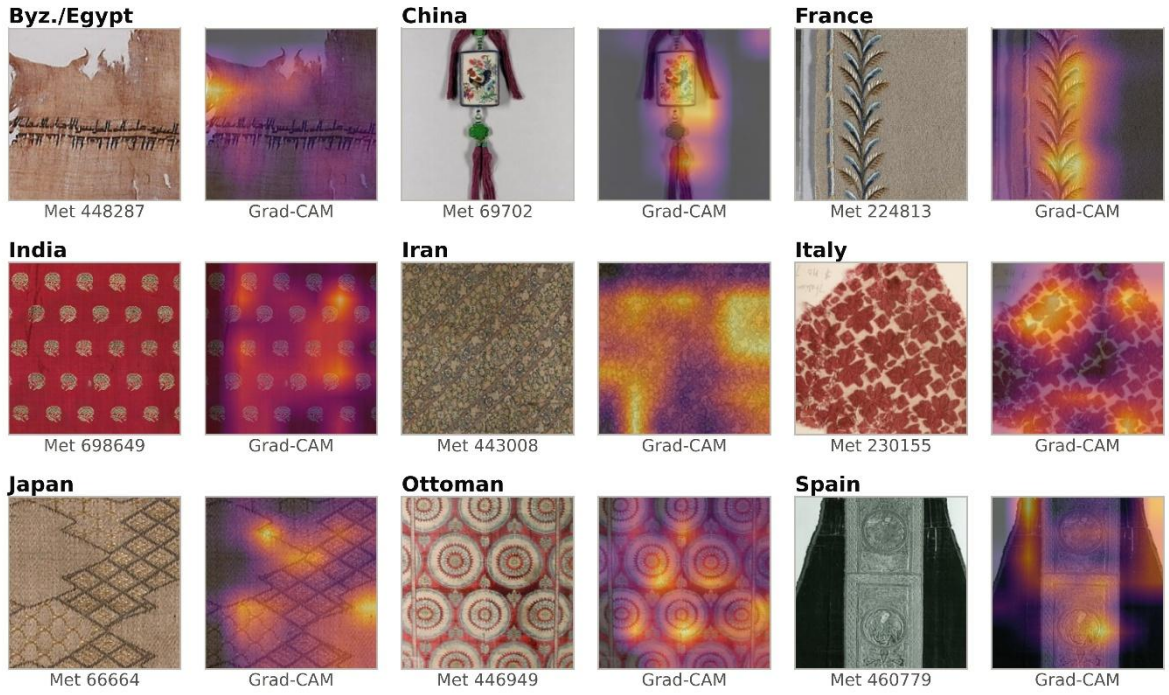


Figure 7. Grad-CAM heat maps of the main model on color photographs with trimmed borders, one high-confidence correct prediction for each origin. CMA, Cleveland Museum of Art. Byz./Egypt, Byzantine and Egyptian silks.

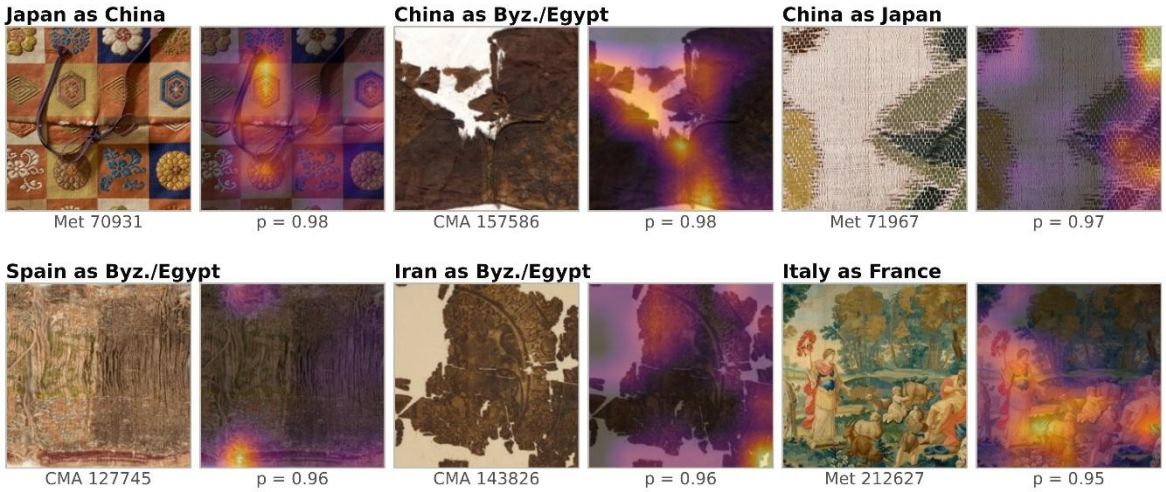


Figure 8. Grad-CAM heat maps for the six most confident misclassifications of the main model, each from a different pair of true and predicted origins. The titles give the true and predicted origins, and p is the predicted probability. Each heat map is shown for the predicted class. CMA, Cleveland Museum of Art.

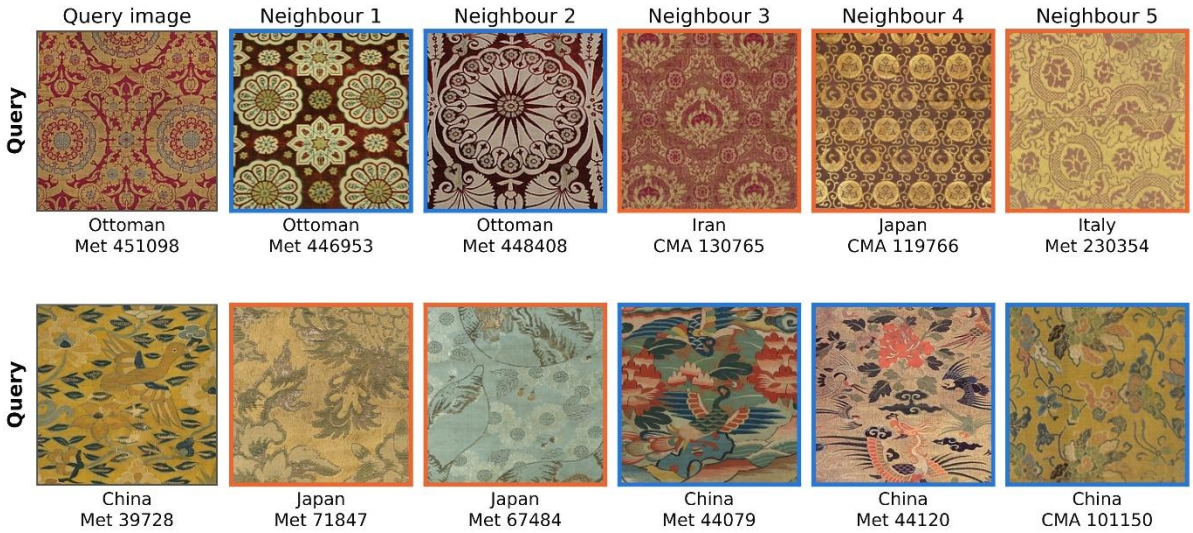


Figure 9. Nearest-neighbor retrieval in the frozen ResNet-50 feature space for two query images. Neighbors from the query's own image group are excluded. Blue frames mark neighbors of the same origin as the query, and orange frames mark neighbors of another origin. CMA, Cleveland Museum of Art.

Removing the 803 color photographs of mixed-fiber objects gave 0.508 on the first repetition. This removed 205 of the 238 Byzantine and Egyptian photographs, and the 33 that remained were rarely recognized (recall 0.152), while recall for the other origins changed by at most 0.04, except for Ottoman silks, which lost a third of their images. Fiber mixture thus helps define the Byzantine and Egyptian group, and its removal changed the results for the other traditions little. The duplicate threshold matters little. Grouping at 0.95 made 208 merges and gave 0.580, and grouping at 0.99 made 154 merges and gave 0.581, against 0.574 at 0.97 ($p = 0.346$ and $p = 0.171$). Grouping by accession number alone gave 0.582, a paired difference of 0.008 (95% CI -0.004 to 0.019, $p = 0.196$).

This difference lies within fold-to-fold variation, so near-duplicate leakage had little effect on these data, although the grouping is kept because leakage can only raise an estimate. Random label noise of 5% to 10% has a modest effect. With 5% of training labels reassigned at random, balanced accuracy fell from 0.574 to 0.567, and with 10% it fell to 0.547, so errors in the training labels cost little at these rates. The learning curve rose from 0.405 with about 617 training images per fold to 0.574 with the full 6,130 and was still rising. The fitted power law in Eq. (6) has an asymptote of 0.710, and it reaches 0.60 at about 11,436 training images, 1.9 times the present size. A balanced accuracy of 0.65 would need about 64,555 training images under the same fit, 11 times the present size. More open data, especially for the smaller classes, would therefore help, with diminishing returns.

Table 3. Checks on confounds, design choices, and the heat maps for the main setting (trimmed color photographs, ResNet-50 features, ridge classifier unless stated).

Check	Images	Balanced accuracy	95% CI or detail
Main result	7,663	0.577	0.561 to 0.593
Source museum only	7,663	0.148	0.141 to 0.155
Image size and file properties only	7,663	0.216	0.205 to 0.228
Border color and contrast only	7,663	0.242	0.230 to 0.254
Date only	7,663	0.389	0.379 to 0.400
Image features plus date	7,663	0.606	0.591 to 0.623
Main model, Met photographs	6,795	0.577	0.559 to 0.596
Main model, CMA photographs	868	0.493	0.451 to 0.531
Main model, fragments and samples	6,402	0.565	0.547 to 0.583
Main model, garments and dress accessories	373	0.373	0.308 to 0.445
Main model, furnishings and liturgical textiles	547	0.429	0.364 to 0.492
Main model, other objects	341	0.478	0.411 to 0.545
Mixed-fiber objects removed	6,860	0.508	0.483 to 0.530
Random 70% crop	7,663	0.582	0.564 to 0.599
Central 60% crop	7,663	0.569	0.551 to 0.586
Central 80% crop	7,663	0.592	0.574 to 0.609
Grouping by accession number only	7,663	0.582	0.566 to 0.598
Near-duplicate threshold 0.95	7,663	0.580	0.563 to 0.598
Near-duplicate threshold 0.99	7,663	0.581	0.564 to 0.599
5% of training labels reassigned	7,663	0.567	clean value 0.574
10% of training labels reassigned	7,663	0.547	clean value 0.574
Deletion test, initial model	450	not applicable	drop 0.146 (shifted 0.061, lowest 0.005)
Deletion test, main model	450	not applicable	drop 0.118 (shifted 0.041, lowest -0.005)

Note: Values are means over five repetitions except the mixed-fiber, crop, grouping and label-noise rows, which use the first repetition (main result 0.574). Deletion rows give the mean drop in the probability of the true class when the top 10% of heat-map pixels, the same region shifted, or the lowest 10% are replaced by the image mean.

4.6 Date as a confounder

The date-only model reached a balanced accuracy of 0.389 (0.379 to 0.400). Adding the date terms to the image features gave 0.606 (0.591 to 0.623), a paired change of +0.030 (0.023 to 0.036, $p < 0.001$). Date therefore carries some information about origin that the image features do not capture. This is expected, because museum dates also draw on documents, provenance and comparison with other objects. Museum dates and attributions are often set by the same curatorial judgment, so the date-only result is best read as an upper bound on what period knowledge alone contributes. Figure 10 shows balanced recall within date bands, defined by the midpoint of each object's date range, together with the origins present in each band. Recall was 0.427 for 1000 to 1499, 0.511 for 1500 to 1699, and 0.489 for 1700 to 1799, bands that hold most images and contain six to eight origins with at least ten images each. Recall was 0.483 before 1000, in 114 images dominated by Byzantine and Egyptian silks and China, and 0.449 from 1800 onward, where images are dominated by Japan and France. Within every band, the image model also stayed well above the date-only model evaluated on the same images, which reached only 0.120 to 0.211. The residual performance is therefore not explained by the period and origin mix of each band.

4.7 Feature extractors and classifiers

Table 4 compares the four backbones on the 7,663 trimmed color photographs under identical preprocessing and folds, together with the further classifiers and the fine-tuned network. With the ridge classifier, ViT-B/16 features reached 0.598, 0.021 above ResNet-50 ($p = 0.009$), and the two foundation models did clearly better, with 0.655 for DINOv2 and 0.641 for CLIP, gains of 0.078 and 0.064 over ResNet-50 ($p < 0.001$ for both).

DINOv2 improved recall for every origin, most for India (by 0.15), Iran (0.11), and France (0.10), while Spain remained the hardest class with every backbone. For most single origins, the confidence intervals of the backbones overlap, so these class-level gains are descriptive. With frozen features and a linear classifier, these results show how much each pretrained representation already holds about silk traditions, and the ranking could change if the transformer backbones were adapted to silk images. On ResNet-50 features, the ridge model remained the strongest frozen-feature classifier. On the same folds of the first repetition, where the ridge model gave 0.574, the support vector machine reached 0.480, the nearest-neighbor classifier 0.467, the perceptron 0.538, and the tuned random forest 0.439, all clearly below the ridge model. Partial fine-tuning of the last residual stage reached 0.624 (0.607 to 0.641) on the first repetition, 0.050 above the frozen ResNet-50 model on the same folds (0.035 to 0.065, $p < 0.001$). Adapting even part of the network to silk images therefore adds clearly to what frozen ImageNet features hold, and its gain is of the same order as that of the foundation-model features. Fuller fine-tuning of the stronger backbones is a natural next step.

4.8 Weaving technique recognition

Technique was recognized with a balanced accuracy of 0.722 (0.706 to 0.737) across four classes from all original images, against chance at 0.25. Recall was 0.777 for woven patterned silks, 0.759 for velvet, 0.717 for embroidery, and 0.635 for tapestry. Because technique labels come from the same catalog fields as the silk selection and are linked to origin, the technique model was audited like the origin model. On trimmed color photographs, it reached 0.709, and its balanced recall was 0.735 for color and 0.685 for black-and-white photographs, so neither borders nor format account for much of the result.

Table 4. Comparison of frozen backbones and classifiers on all 7,663 trimmed color photographs. Backbones use the ridge classifier and five repetitions. Classifiers use ResNet-50 features and the first repetition, and the ridge value for that repetition is given for comparison.

Model	Balanced accuracy	95% CI	Difference from ResNet-50 ridge	Paired p
ResNet-50, ridge (five repetitions)	0.577	0.561 to 0.593	reference	not applicable
ViT-B/16, ridge	0.598	0.583 to 0.613	+0.021	0.009
DINOv2 ViT-B/14, ridge	0.655	0.639 to 0.670	+0.078	< 0.001
CLIP ViT-B/16, ridge	0.641	0.626 to 0.656	+0.064	< 0.001
ResNet-50, ridge (first repetition)	0.574	0.557 to 0.592	reference	not applicable
ResNet-50, support vector machine	0.480	0.462 to 0.497	-0.095	< 0.001
ResNet-50, k-nearest-neighbour classifier	0.467	0.452 to 0.483	-0.107	< 0.001
ResNet-50, multilayer perceptron	0.538	0.521 to 0.556	-0.036	< 0.001
ResNet-50, tuned random forest	0.439	0.425 to 0.454	-0.135	< 0.001
ResNet-50, last stage fine-tuned	0.624	0.607 to 0.641	+0.050	< 0.001

Note: CI, confidence interval from 2,000 grouped bootstrap resamples. Paired p-values are from the corrected resampled t-test for the backbones and from the paired bootstrap for the classifiers, each against the ResNet-50 ridge model on the same images and folds.

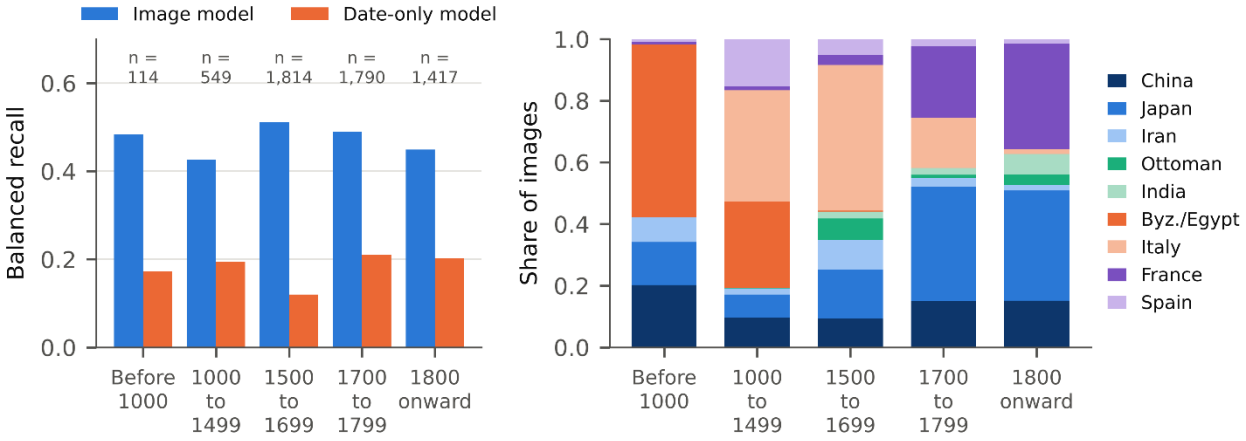


Figure 10. Balanced recall within date bands for the main image model and the date-only model on the same images (left), and the share of each origin within each date band (right).

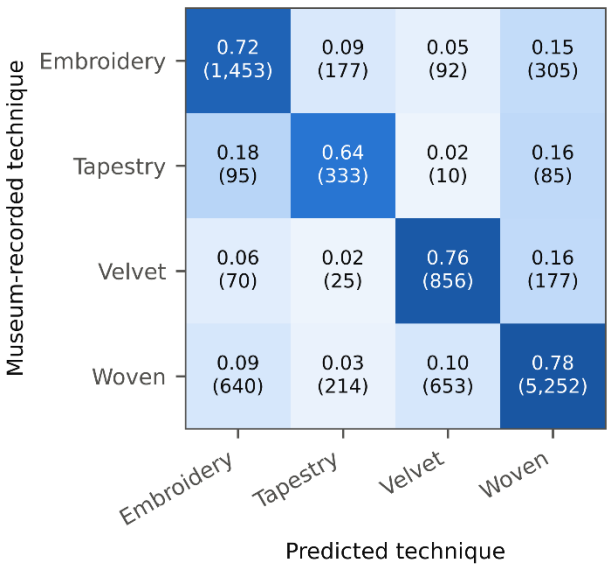


Figure 11. Confusion matrix for the weaving technique on all original images, with row proportions and image counts (summed over the five repetitions and divided by five).

Without the Byzantine and Egyptian group, where tapestry is concentrated, balanced accuracy was 0.711, and within Japan alone it was 0.473, so part of the overall accuracy may come from links between technique and origin. Balanced accuracy was 0.719 for fragments and samples and 0.523 for garments and dress accessories. The most frequent errors assigned tapestry to embroidery (0.182) and velvet to woven silks (0.157) (Figure 11). Figure 12 shows heat maps of the technique model trained on trimmed color photographs for correct predictions and for the same two confusions, which occurred in 17% and 18% of cases, respectively. Correct velvet maps lie on the pile surface and its cut pattern, and correct embroidery maps on raised stitched motifs. Tapestry fragments with fine slit-woven motifs resemble embroidered motifs on a plain ground, and voided or brocaded velvets look like patterned weaves when the pile is low or worn. These pairs differ at a scale of a few threads, which images of 320 pixels cannot resolve, so higher-resolution photographs or detail views of the weave would probably reduce these errors more than a different classifier would.

4.9 Two-package verification

The ridge classifier agreed between Python and R to within 0.0019 in every setting of the first repetition, including the ViT-B/16, DINOv2 and CLIP features, the data models, and

the task, and in most settings to within 0.0001. Table 2 lists the R values for that repetition beside the Python means over five repetitions. The untuned random forest was weaker and less consistent, with balanced accuracy in scikit-learn between 0.223 and 0.368 and differences from ranger of up to 0.079, because the two packages apply class weights inside the trees in different ways. The tuned forest gave 0.439 in Python and 0.342 in R on the same folds. The settings carry over only approximately, because ranger applies class weights only in the split criterion and uses the minimum node size as a limit for splitting a node, so the tuned forest did not meet the agreement check. The support vector machine and the nearest-neighbor classifier agreed between the two environments to within 0.001 and 0.000.

4.10 Implications for heritage documentation and design

The results extend the SILKNOW line of work [16-18] across Asian and European traditions and match the background effects reported for decorative-art images [32]. Accuracy measured on unaudited images can overstate what a model has learned about the objects, and a routine heat-map audit exposed this in a single pass.

Frozen pretrained features already separate nine silk traditions with accuracy well above chance, and foundation-model features such as DINOv2 separate them even better, making cataloging support and similarity search feasible for collections without large training budgets. For designers, the confusion and retrieval results map shared motif families, such as those of Chinese and Japanese woven silks. Table 5 turns the protocol used here into a checklist for teams building similar systems on other open museum collections.

4.11 Limitations

Several limitations qualify these conclusions. Only the last residual stage of ResNet-50 was fine-tuned, on one repetition, while the transformer backbones served as frozen feature extractors, and fuller adaptation might change both accuracy and ranking. Origin labels come from museum catalogs, and the simulation covers only random errors in the training labels. The data come from two museums, and several classes are small. The heat-map checks used one explanation method and were run with ResNet-50, so whether the DINOv2 and CLIP features rely on the same photographic cues remains to be tested. The black-and-white rule misses about 30 toned prints, and condition and damage remain a residual cue.

Table 5. Recommended practice for recognition models built on open museum photographs.

Stage	Recommended step
1. Before training	Group records by object and merge near-duplicate photographs, keeping each group inside one fold.
2. Before training	Record the photographic format of each image and report its share by class.
3. Training	Fit a regularized linear classifier on frozen features, tune it inside the training folds, and report balanced accuracy.
4. Audit	Inspect Grad-CAM maps of confident correct and incorrect cases and measure heat-map mass on borders, labels and mounts.
5. Audit	Rerun on color photographs, trimmed images, and greyscale images, and test the differences with paired statistics.
6. Controls	Fit baselines on date, museum, image properties, and background, and report accuracy within museums and object types.
7. Reporting	Recalibrate probabilities before using thresholds, and release identifiers, folds, features and scripts.

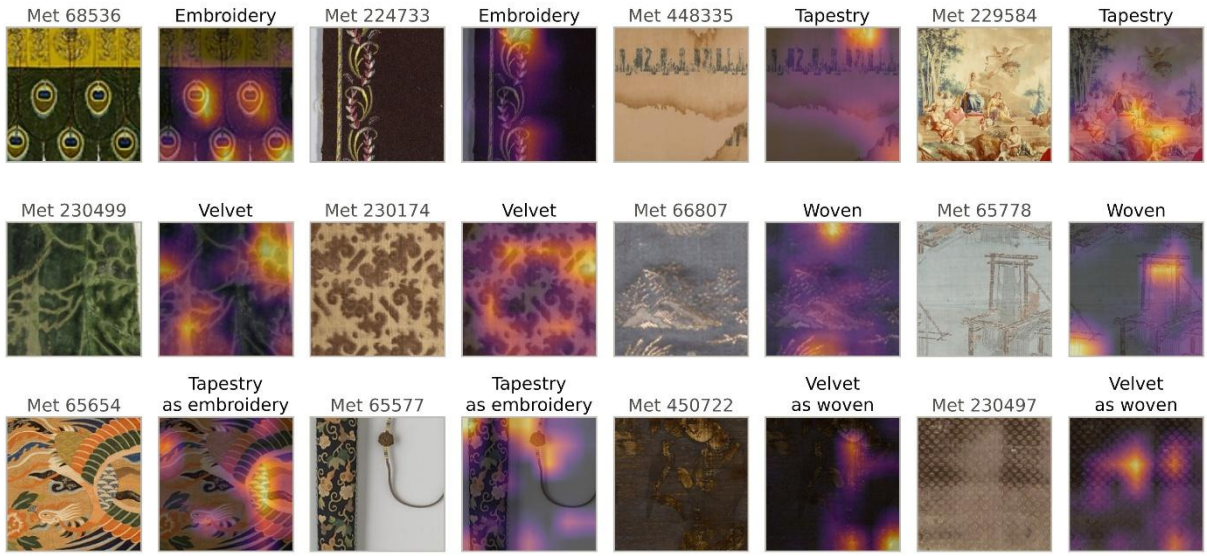


Figure 12. Grad-CAM heat maps of the technique model on trimmed color photographs. The upper rows show high-confidence correct predictions for each technique, and the lower rows show the two most frequent confusions: tapestry predicted as embroidery and velvet predicted as woven patterned silk, with the heat map for the predicted class. CMA, Cleveland Museum of Art.

5. Conclusion

This study developed an explainable deep learning framework for the intelligent documentation of silk textile heritage and applied it to 11,388 open museum images from nine cultural traditions. Grad-CAM auditing, supported by quantitative border and deletion analysis, showed that an initial model relied on accession labels, color charts, and the black-and-white format of archive photographs. After removing these shortcuts and grouping near-duplicates, frozen ResNet-50 features recognized cultural origin with a balanced accuracy of 0.577 (95% CI 0.561 to 0.593) and weaving technique with 0.722 (0.709 on trimmed color photographs). Foundation-model features raised origin recognition to 0.655 with DINOv2, and partial fine-tuning of ResNet-50 raised it to 0.624, so the audited ResNet-50 estimate is a lower bound on what such systems can reach. These figures lie far above chance and above baselines built on date, museum, image properties, or background. The ridge estimates were reproduced in Python and R. Future work should adapt the backbones to silk images more fully, add collections from Asian museums, and extend the analysis to finer weave structures, so that the learned motif families can support cross-cultural textile design.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request. The museum records and images analyzed in this study are openly available from The Metropolitan Museum of Art [13] and the Cleveland Museum of Art [14]. The exact file versions are identified by the SHA-256 checksums de617b9c947458e426111207f81a65bd1379a151c0077d3ce29cfc22fc0b9183 (MetObjects.csv) and 28abe722b33ec1c0cc397583ec4ef1ef59c0c9676c407168ab0ebb44fb871b9b (data.csv). The list of record identifiers with their labels, the image-group and fold assignments, the extracted feature matrices, and the Python and R analysis scripts are deposited at https://doi.org/10.5281/zenodo.24395779

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, "Machine learning for cultural heritage: A survey," Pattern Recognition Letters, vol. 133, pp. 102-108, 2020. DOI: https://doi.org/10.1016/j.patrec.2020.02.017

[2] E. Cetinic and J. She, "Understanding and creating art with AI: Review and outlook," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 2, pp. 1-22, 2022. DOI: https://doi.org/10.1145/3475799

[3] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," Expert Systems with Applications, vol. 114, pp. 107-118, 2018. DOI: https://doi.org/10.1016/j.eswa.2018.07.026

[4] T. Mensink and J. van Gemert, "The Rijksmuseum Challenge: Museum-centered visual recognition," in Proceedings of International Conference on Multimedia Retrieval, 2014, pp. 451-454. DOI: https://doi.org/10.1145/2578726.2578791

[5] G. Strezoski and M. Worring, "OmniArt: A large-scale artistic benchmark," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 14, no. 4, pp. 1-21, 2018. DOI: https://doi.org/10.1145/3273022

[6] S. He, "Research on a segmentation algorithm for the Tujia brocade images based on unsupervised Gaussian mixture clustering," Frontiers in Neurorobotics, vol. 15, Art. no. 739077, 2021. DOI: https://doi.org/10.3389/fnbot.2021.739077

[7] M. Ni, Q. Huang, N. Ni, H. Zhao, and B. Sun, "Research on the design of Zhuang brocade patterns based on automatic pattern generation," Applied Sciences, vol. 14, no. 13, Art. no. 5375, 2024. DOI: https://doi.org/10.3390/app14135375

[8] K. Meng, H. Li, Y. Shu, M. Chen, X. Han, and L. He, "Database construction and remodeling method on traditional Yi nationality patterns of China with GAN model," npj Heritage Science, vol. 13, no. 1, Art. no. 181, 2025. DOI: https://doi.org/10.1038/s40494-025-01707-w

[9] L. B. Mahanta, D. R. Mahanta, T. Rahman, and C. Chakraborty, "Handloomed fabrics recognition with deep learning," Scientific Reports, vol. 14, no. 1, Art. no. 7974, 2024. DOI: https://doi.org/10.1038/s41598-024-58750-z

[10] A. Das, A. Deka, K. Medhi, and M. J. Saikia, "Deep learning to authenticate traditional handloom textile," Information, vol. 15, no. 8, Art. no. 465, 2024. DOI: https://doi.org/10.3390/info15080465

[11] K. Arai, J. Shimazoe, and M. Oda, "A method for maintaining a unique Kurume Kasuri pattern of woven textile classified by EfficientNet by means of LightGBM-based prediction of misalignments," Information, vol. 15, no. 8, Art. no. 434, 2024. DOI: https://doi.org/10.3390/info15080434

[12] S. Tena, R. Hartanto, and I. Ardiyanto, "Content-based image retrieval for traditional Indonesian woven fabric images using a modified convolutional neural network method," Journal of Imaging, vol. 9, no. 8, Art. no. 165, 2023. DOI: https://doi.org/10.3390/jimaging9080165

[13] The Metropolitan Museum of Art, "MetObjects.csv," The Metropolitan Museum of Art Open Access data set, file version downloaded 24 September 2026, SHA-256 de617b9c947458e426111207f81a65bd1379a151c0077d3ce29cfc22fc0b9183. [Online]. Available: https://github.com/metmuseum/openaccess

[14] The Cleveland Museum of Art, "data.csv," The Cleveland Museum of Art Open Access data set, file version downloaded 24 September 2026, SHA-256 28abe722b33ec1c0cc397583ec4ef1ef59c0c9676c407168ab0ebb44fb871b9b. [Online]. Available: https://github.com/ClevelandMuseumArt/openaccess

[15] K. Chen, D. Lu, Z. Jin, M. Su, and J. Jin, "Song brocade in the Ming and Qing dynasties," Clothing and Textiles Research Journal, vol. 38, no. 4, pp. 285-297, 2020. DOI: https://doi.org/10.1177/0887302X20932657

[16] M. Dorozynski, D. Clermont, and F. Rottensteiner, "Multi-task deep learning with incomplete training samples for the image-based prediction of variables describing silk fabrics," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-2/W6, pp. 47-54, 2019. DOI: https://doi.org/10.5194/isprs-annals-IV-2-W6-47-2019

[17] M. Dorozynski and F. Rottensteiner, "Deep descriptor learning with auxiliary classification loss for retrieving images of silk fabrics in the context of preserving European silk heritage," ISPRS International Journal of Geo-Information, vol. 11, no. 2, Art. no. 82, 2022. DOI: https://doi.org/10.3390/ijgi11020082

[18] L. Rei, D. Mladenic, M. Dorozynski, F. Rottensteiner, T. Schleider, R. Troncy, J. S. Lozano, and M. Gaitán Salvatella, "Multimodal metadata assignment for cultural heritage artifacts," Multimedia Systems, vol. 29, no. 2, pp. 847-869, 2023. DOI: https://doi.org/10.1007/s00530-022-01025-2

[19] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, no. 11, pp. 665-673, 2020. DOI: https://doi.org/10.1038/s42256-020-00257-z

[20] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLOS Medicine, vol. 15, no. 11, Art. no. e1002683, 2018. DOI: https://doi.org/10.1371/journal.pmed.1002683

[21] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nature Machine Intelligence, vol. 3, no. 7, pp. 610-619, 2021. DOI: https://doi.org/10.1038/s42256-021-00338-7

[22] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle, "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," JAMA Dermatology, vol. 155, no. 10, pp. 1135-1141, 2019. DOI: https://doi.org/10.1001/jamadermatol.2019.1735

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626. DOI: https://doi.org/10.1109/ICCV.2017.74

[24] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," Nature Communications, vol. 10, no. 1, Art. no. 1096, 2019. DOI: https://doi.org/10.1038/s41467-019-08987-4

[25] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," Proceedings of the IEEE, vol. 109, no. 3, pp. 247-278, 2021. DOI: https://doi.org/10.1109/JPROC.2021.3060483

[26] J. Xiang, N. Zhang, R. Pan, and W. Gao, "Fabric image retrieval system using hierarchical search based on deep convolutional neural network," IEEE Access, vol. 7, pp. 35405-35417, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2898906

[27] J. Xiang, R. Pan, and W. Gao, "An efficient retrieval system framework for fabrics based on fine-grained similarity," Entropy, vol. 24, no. 9, Art. no. 1319, 2022. DOI: https://doi.org/10.3390/e24091319

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning, PMLR, vol. 139, 2021, pp. 8748-8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[29] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," Transactions on Machine Learning Research, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[30] M. V. Conde and K. Turgutlu, "CLIP-Art: Contrastive pre-training for fine-grained art classification," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3951-3955. DOI: https://doi.org/10.1109/CVPRW53098.2021.00444

[31] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Exploiting CLIP-based multi-modal approach for artwork classification and retrieval," in The Future of Heritage Science and Technologies: ICT and Digital Heritage, Communications in Computer and Information Science, vol. 1645, Cham: Springer, 2022, pp. 140-149. DOI: https://doi.org/10.1007/978-3-031-20302-2_11

[32] A. Toumarkine and C. Vidal-Gorène, "Beyond accuracy: Investigating vision model perception on 19th-century decorative arts," in Computational Humanities Research 2025, Anthology of Computers and the Humanities, vol. 3, 2025, pp. 1544-1562. DOI: https://doi.org/10.63744/k4Jy3Lr4rdK5

[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 512-519. DOI: https://doi.org/10.1109/CVPRW.2014.131

[34] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717-1724. DOI: https://doi.org/10.1109/CVPR.2014.222

[35] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2656-2666. DOI: https://doi.org/10.1109/CVPR.2019.00277

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255. DOI: https://doi.org/10.1109/CVPR.2009.5206848

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778. DOI: https://doi.org/10.1109/CVPR.2016.90

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[39] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921-2929. DOI: https://doi.org/10.1109/CVPR.2016.319

[41] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019. DOI: https://doi.org/10.1038/s42256-019-0048-x

[42] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proceedings of the British Machine Vision Conference (BMVC), 2018. [Online]. Available: http://bmvc2018.org/contents/papers/1064.pdf

[43] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in CVPR 2011, 2011, pp. 1521-1528. DOI: https://doi.org/10.1109/CVPR.2011.5995347

[44] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley, "Deep learning predicts hip fracture using confounding patient and healthcare variables," npj Digital Medicine, vol. 2, no. 1, Art. no. 31, 2019. DOI: https://doi.org/10.1038/s41746-019-0105-1

[45] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023. DOI: https://doi.org/10.1016/j.patter.2023.100804

[46] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, no. 1, Art. no. 91, 2006. DOI: https://doi.org/10.1186/1471-2105-7-91

[47] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in 2010 20th International Conference on Pattern Recognition, 2010, pp. 3121-3124. DOI: https://doi.org/10.1109/ICPR.2010.764

[48] R. Wightman, "PyTorch image models," software repository, version 1.0.15 used, 2019. [Online]. Available: https://github.com/huggingface/pytorch-image-models

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, 2019, pp. 8024-8035. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[50] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324

[51] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. DOI: https://doi.org/10.1007/BF00994018

[52] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967. DOI: https://doi.org/10.1109/TIT.1967.1053964

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825-2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[54] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," Journal of Statistical Software, vol. 33, no. 1, pp. 1-22, 2010. DOI: https://doi.org/10.18637/jss.v033.i01

[55] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," Journal of Statistical Software, vol. 77, no. 1, pp. 1-17, 2017. DOI: https://doi.org/10.18637/jss.v077.i01

[56] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman and Hall/CRC, 1994. DOI: https://doi.org/10.1201/9780429246593

[57] C. Nadeau and Y. Bengio, "Inference for the generalization error," Machine Learning, vol. 52, no. 3, pp. 239-281, 2003. DOI: https://doi.org/10.1023/A:1024068626366

[58] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning, PMLR, vol. 70, 2017, pp. 1321-1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[59] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694-699. DOI: https://doi.org/10.1145/775047.775151

[60] T. Viering and M. Loog, "The shape of learning curves: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7799-7819, 2023. DOI: https://doi.org/10.1109/TPAMI.2022.3220744

[61] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845-869, 2014. DOI: https://doi.org/10.1109/TNNLS.2013.2292894



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).