Received: 21 August 2026; **Revised:** 20 October 2026; **Accepted:** 27 October 2026; **Published:** 20 November 2026
Archs Sci. (2026) Volume 76, Issue 5 Pages 121-132, Paper ID 2026512.
https://doi.org/10.68304/as/76512

A Lightweight Digital ID Solution for Grassroots Cultural Heritage Institutions: Training-Free Identity Matching for Low-Computing-Power Environments

Xilei Qin[1] and Shengli Sun[1,*]
[1]The Dunhuang Academy Archives, Dunhuang 736200, Gansu, China
Correspondence: 108739260@qq.com

Abstract Grassroots cultural heritage institutions must confirm that the item in front of a camera is the same item as the one in their register, yet they work with a few images per object, a single CPU-grade terminal, and no practical capacity to retrain a network whenever a new item is accessioned. This paper reports a digital identity solution designed around that constraint and evaluates it using a purpose-collected archival corpus rather than a public benchmark. Each object is described by a 1280-dimensional descriptor extracted using a frozen ImageNet-pretrained MobileNetV2 network with GeM pooling; the registration views of each object are summarised into an identity template, a query is matched using cosine similarity, and an acceptance threshold, together with mandatory manual review, determines whether an identity match is confirmed. Identity is stored in a structured record containing the identifier, the template, collection metadata, the model and template version, and an integrity hash. The corpus contains 42 objects and 435 images in four object groups, of which 35 objects are registered and 7 are used as unknowns, with registration and query images separated according to their acquisition conditions; a perceptual-hash audit found no near-duplicate pair across that split. Closed-set top-1 accuracy is 67.2% and top-5 accuracy is 92.7%. Two principal findings structure the discussion. First, the same objects matched under an identical acquisition condition achieve 77.7%, whereas the identical pipeline achieves only 67.2% across different acquisition conditions; this 10.5-percentage-point gap is not eliminated by backbones containing up to 85.8 million parameters, by metric-learning heads trained with triplet and relational-distillation losses, or by colour normalisation and re-ranking, so the principal limitation is the acquisition condition rather than the capacity of the model. Second, the quantised model occupies 2.49 MB and runs at a median latency of 4.14 ms with a 1.1-point accuracy loss, meeting the deployment target without any training stage. We therefore argue that, for small archival corpora, engineering effort should be directed towards acquisition protocols and record design, and that an open-set equal error rate of 33.1% requires the system to be specified as a candidate-ranking aid for human review rather than as an autonomous identity decision-making system.

Index Terms Digital identity; Cultural heritage archives; Instance-level retrieval; Training-free descriptor; Open-set rejection; Edge deployment

I. Introduction

Cultural heritage is the tangible repository of historical memory, of the evolution of social civilisation, and of shared cultural identity. Its scientific conservation, careful management and digital preservation are of considerable significance for maintaining cultural diversity, disseminating historical knowledge, and ensuring the sustainable use of cultural resources. UNESCO has emphasised that digital technology is continuously transforming the ways in which cultural resources are preserved, disseminated and made accessible, and has set out the requirements that such preservation must satisfy if digital objects are to remain usable over time [1]. With the development of computer vision, edge computing and digital information management, cultural heritage conservation is shifting from a model that relied on manual registration, static image archiving, and experience-based judgement towards one organised around sensing, digital representation, automatic recognition, and data-driven decision-making. Within that shift, a stable, reliable and traceable mechanism that links the identity of a physical object to its digital record has become a central problem, because collection registration, stocktaking, gate verification, circulation, and security auditing all require an answer to one question: is the item in front of me the item recorded under this number?

A considerable gap separates the state of the art in heritage digitisation from the technical capacity of grassroots institutions. Large museums maintain information infrastructure, specialised technical teams, and digitised resources; many

grassroots museums, memorial halls and preservation institutions do not. Their equipment consists largely of standard CPUs, low-cost industrial computers or embedded terminals, which cannot readily sustain the computational demands of complex vision models over extended periods. Their collections are characterised by numerous object types, substantial variation between objects of the same type, and very few images per object, which means that supervised methods requiring large annotated training sets may transfer poorly. Moreover, existing collection management systems establish identity through manually entered accession numbers, names, materials, and condition reports; the international reference ontology for the interchange of such information specifies the entities and relations that a record should expose [2], but it does not provide any mechanism that verifies a record against the physical object itself. Extending intelligent conservation technology to these institutions therefore requires individual identification that is accurate, efficient and scalable under limited computation and few samples, and that can be embedded within the collection record rather than appended as a separate system.

Deep learning has advanced heritage recognition considerably. Early work relied on local descriptors and traditional image matching, searching for similarity through edges, corners and local texture; such methods adapt poorly to cluttered backgrounds, scale changes, partial occlusion and viewpoint change, and cannot represent high-level relations between objects of different materials and complex structure. Convolutional networks trained end to end subsequently supported classification, artwork retrieval and instance identification. Yet heritage recognition cannot be reduced to category classification alone. For a digital identity application the model must decide not only whether an object is a ceramic, a bronze artefact or a sculpture, but whether the image corresponds to a particular registered object. Fine-grained representations that separate instances inside a category are therefore the operative requirement, and they are not the representations that a conventional categorical classifier is optimised to produce.

Instance-level artwork recognition has been advanced by datasets built for exactly that task, from collection-level benchmarks such as the Oriental Museum corpus used to trace trafficked antiquities [3] to the larger and deliberately adversarial Met Dataset. Ypsilantis et al. [4] introduced The Met Dataset, a benchmark of 224,568 artwork classes and 397,121 training images drawn from the open collection of the Metropolitan Museum of Art, which pairs standard collection photographs with query images taken by gallery visitors and with unrelated distractor images, so that the task more closely resembles open-set identity retrieval under real-world conditions. Related studies construct visual embedding spaces through supervised or self-supervised contrastive learning [5] and matches by nearest neighbour, showing that feature-space identity matching can scale to large and growing collections. The same line of work exposes the difficulty that dominates the present paper: standardised collection images and in-situ query images differ in viewpoint, background, illumination, occlusion and image quality, so descriptors drift across the two conditions, while different objects of one type share overall structure and ornamental vocabulary, which can make an incorrect match easy to generate and difficult to recognise. Global semantic similarity alone does not satisfy the reliability and rejection requirements of a records system.

Scarcity of collection imagery has motivated augmentation and synthesis. Bartkowiak et al. [6] proposed SynGallery, which renders identity-preserving multi-view images of authentic paintings inside a procedural three-dimensional gallery, narrowing the distribution gap between standard collection photography and photographs taken in situ. From 4,898 catalogue images in the Met benchmark, the method renders 24,490 views from five viewpoints under varied framing, lighting and glass; at an equal training budget, training on these renders alone raises the retrieval score from 67.18 to 73.47 GAP, and adding them to the full training set moves the published protocol from 35.97 to 38.48 GAP. The ablation findings are particularly informative: the gain comes from scene-level view variation, whereas simulating capture artefacts such as blur, sensor noise, and compression consistently reduces performance. This indicates that geometric variation carries useful training information for instance-level tasks. The benefit nevertheless depends on how well the generated scenes resemble the actual test environment, and for objects with directional inscriptions, irregular local damage or complex three-dimensional form, aggressive rotation, cropping or texture alteration can destroy precisely the evidence that distinguishes one object from another. Improvements in retrieval performance also do not necessarily translate ininto an improved ability to reject objects that were never registered.

Mobile and edge computing have made lightweight networks a practical option for resource-constrained deployment. Sandler et al. [7] introduced MobileNetV2, whose inverted residual structure, depthwise separable convolutions and linear bottlenecks reduce parameters and arithmetic while preserving representation quality. Qin et al. [8] extended the family with MobileNetV4, using a universal inverted bottleneck and hardware-aware design to improve efficiency on mobile CPUs, GPUs and accelerators, and demonstrated that measured inference cost depends not only on parameter count and theoretical floating-point operations, but also on the operator mix, memory access patterns, bandwidth, and hardware execution characteristics. In the heritage domain, Minango et al. [9] built an offline mobile application that recognises 21 artefacts across six rooms of a single museum by transfer learning on MobileNetV2 with INT8 quantisation, reporting 92.2% accuracy, 63 ms per inference on current phones and a model of 775 kB. Their error analysis is the clearest external support for the argument of Section 2.3: 47% of the failures are attributed to camera angle that hides distinguishing features and 22% to reflections and shadows on the display case, so the dominant source of error is the acquisition condition and it is not removed by augmenting the training data. Existing applications of this kind, however, are designed either to classify a small set of categories or to

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

support visitor interaction; they do not address the registration of individual identity across a large and growing collection, the separation of near-identical objects, or reliable rejection of objects absent from the database. A records system must also avoid the most consequential practical error, namely recording an object movement or condition report against an incorrect accession number.

Model compression has been studied extensively as a means of reducing computational cost. Pruning removes redundant structure [10], quantisation lowers the precision of weights and activations [11], [12], and knowledge distillation transfers the behaviour of a high-capacity teacher to a small student [13], [14]. Distillation has an easily overlooked precondition: the teacher must provide a sufficiently informative target for the specific task, rather than merely having greater general-purpose capacity. When the target is instance-level identity over a small corpus, that precondition may fail, because a high-capacity network pre-trained on object categories has no reason to represent the differences that separate two trophies awarded in the same year, and because a teacher trained on a few dozen identities cannot supply a transferable similarity structure. The consequence is that a distillation pipeline can be more expensive, harder to maintain and less accurate than simply using the pre-trained features unchanged, and the relevant empirical check is straightforward: measure the teacher on the same held-out objects as the student before investing in a distillation pipeline. This paper reports the result of that check, which directly informed the design of the proposed system.

This paper therefore asks what the least expensive architecture is that still delivers a usable digital identity for a small archival collection, and which factors primarily limit its accuracy. The principal contributions are as follows. First, we formulate object identity in an archive as an open-set instance-matching problem coupled to a record, and specify a digital identity object that carries the identifier, a visual template, collection metadata, the model and template version, and an integrity hash, so that a match is auditable rather than merely plausible. Second, we specify a pipeline that requires no task-specific training at all: a frozen ImageNet MobileNetV2 with GeM pooling yields a 1280-dimensional descriptor; registration views are summarised into a template, and matching is a cosine comparison against that template, so that a newly accessioned object enters the system simply through the addition of one template. Third, we evaluate a purpose-collected corpus of 42 objects and 435 images from a heritage-archive collection, in which registration and query images are separated by acquisition condition, and we document a perceptual-hash audit confirming that no near-duplicate pair crosses the split. Fourth, we quantify the factors limiting accuracy on such a corpus: a 10.5-point gap between same-condition and cross-condition matching that neither capacity (up to 85.8M parameters) nor metric-learning heads with relational distillation nor colour and re-ranking post-processing closes, together with the corresponding negative results and the cost of training where it does improve rejection performance. Fifth, we report measured deployment characteristics on CPU hardware, including the observation that dynamic INT8 quantisation is thirty-four times slower than static quantisation while saving only 0.19 MB, a result with direct implications for selecting an appropriate deployment configuration.

II. A Training-Free Identity Representation for Archival Objects

A. Representing the Identity of an Individual Object

Within an archive, the unit of management is the individual item: a trophy, a framed certificate, a seal, or a print. Two objects of the same type may share manufacturer, material, dimensions and even inscription template, and the distinguishing feature that matters to the register may be a single engraved line. A representation adequate for a digital identity therefore has to satisfy three conditions at once: it must separate instances of one type, it must remain sufficiently stable across the acquisition conditions under which the object may subsequently be photographed, and it must be storable and comparable on a device with no accelerator. Figure 1 summarises the representation adopted here, from the input image through convolutional stages, pooling and normalisation to a descriptor that is compared against stored templates.

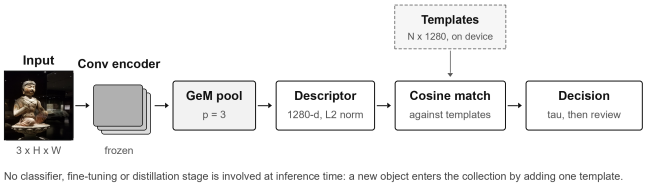


Figure 1: From image to identity descriptor. Convolutional stages increase semantic abstraction while spatial resolution decreases; GeM pooling produces one vector per image, which is L2-normalised and compared with the stored templates of registered objects. No classifier, fine-tuning or distillation stage is involved at inference time.

Let I denote an image of an object and φ the convolutional encoder. The encoder produces a feature map

$$F = \varphi(I) \in \mathbf{R}^{H' \times W' \times C}, \tag{1}$$

where H′ and W′ are the spatial dimensions of the last convolutional stage and C is the number of channels. Global average pooling, the conventional choice for classification, weights every spatial location equally, which may be inappropriate for this application: the discriminating evidence for an archival object is often a small region such as an engraved plate, a seal face or a printed seal, and averaging dilutes it. We therefore use generalised mean pooling, which interpolates between average pooling and max pooling through a single parameter p:

$$v = \frac{1}{H'W'} \left(\sum_{i=1}^{H'W'} f_i^p \right)^{1/p}, \quad p = 3, \tag{2}$$

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

with f_i the i-th spatial location of F. For p = 1 the expression reduces to average pooling, and as p grows it approaches max pooling; p = 3 was fixed before the experiments and remained unchanged throughout the evaluation, and it concentrates weight on strongly activated regions without discarding the rest of the object. The pooled vector is normalised to the unit sphere, so that the comparison of two images is the cosine of the angle between them and is unaffected by contrast scaling:

$$x = v / \|v\|_2 . \tag{3}$$

The registered images of object k form the set R_k. Their descriptors are combined into a template

$$\mu_k = \frac{\left(\sum_{i \in R_k} x_i\right)}{\left\|\sum_{i \in R_k} x_i\right\|_2}, \tag{4}$$

which is again a unit vector, so that templates constructed from different numbers of registration views remain directly comparable. Averaging within the normalised feature space is an intentional design choice: it treats the registration views as repeated measurements of one identity and suppresses the view-specific component, whereas averaging unnormalised descriptors would let a single high-contrast view dominate the template.

One design decision deserves explicit justification, because it differs from the usual supervised formulation. A classification head with a softmax over the registered identities would be the natural way to obtain discriminative features, and it is what most comparable systems train. However, such a component is unsuitable for the intended archival records application. A classifier fixes the set of identities at training time, so every accession requires retraining or at least an additional output unit, creating precisely the maintenance burden that a small institution may be unable to accommodate; it is trained to separate the classes it has seen and provides no principled response to an object it has not; and with a few dozen identities and a few images each, the classification objective is dominated by whichever classes happen to have more views. Open-set recognition literature has long treated the rejection of unknown classes as a separate problem requiring an explicit decision rule [15]. We therefore keep only the embedding, and place the decision in a threshold on the similarity between the query descriptor and the stored template, as specified in Section 2.3.

B. A Feature Encoder That Requires No Task-Specific Training

Figure 2 shows the six stages of the pipeline. Stages 2 to 4 are executed once when a device build is prepared; stage 5 is executed at every verification; stage 6 is re-executed only when either the available review budget or the target hardware changes. None of the components in this processing chain is fitted to the collection itself.

Image preprocessing is deliberately kept minimal. Each image is resized so that its shorter side measures 256 pixels and is then centre-cropped to 224 × 224, which matches the input resolution of the encoder and preserves the aspect ratio of the object without introducing padding. No colour correction, histogram equalisation, or white-balance normalisation is applied. This is a deliberate departure from the usual recommendation for cross-condition matching: in the ablation reported in Section 3.3.4, grey-world correction, per-channel stretching, Reinhard transfer and instance normalisation all reduced accuracy relative to the uncorrected images, because the colour cast introduced by a display case or mixed indoor lighting can itself carry information about the acquisition conditions, and removing that cast can also remove part of the signal that distinguishes one object's appearance from another's.

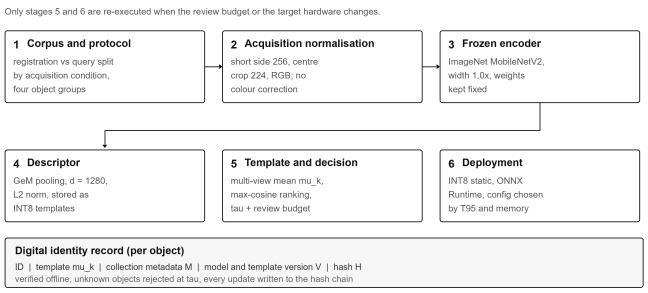


Figure 2: The proposed pipeline. Registration and query views are separated by acquisition condition; a frozen ImageNet MobileNetV2 with GeM pooling produces a 1280-dimensional descriptor; objects are matched against templates at a threshold calibrated to the available review budget; the result is written into a digital identity record that supports offline verification and traceable update.

The encoder is MobileNetV2 with a width multiplier of 1.0, pre-trained on ImageNet [16], [17] and used with its weights frozen. Freezing the encoder is not simply a compromise imposed by the limited data volume; it is an explicit design decision. A network trained to separate a thousand object categories has learned filters that respond to material, contour, surface finish and printed structure, which correspond to the visual properties used to describe an archival object, and those filters are shared by all objects rather than tied to the identities currently registered. Keeping them fixed means that adding an object changes one row of the template table and nothing else, that two institutions can compare templates produced by the same published weights, and that the model version can be recorded within the identity record and audited subsequently. Section 3.3.4 examines what is lost by not adapting the encoder, and reports that on this corpus adaptation loses more than it gains.

For an object k with registration set R_k, the similarity between a query image and a registered object is calculated as the mean of the two highest similarities between the query descriptor and the individual registration descriptors, which we write

$$s(x, k) = \frac{1}{2}(\sigma_1 + \sigma_2), \quad \sigma_1 \geq \sigma_2, \tag{5}$$

rather than the similarity to the template alone. The reason is empirical and was fixed by a held-out selection described in Section 3.3.1: a plain maximum over views is sensitive to a single registration image with an unusual background, the centroid alone loses the benefit of having several views, and the mean of the two highest scores retains much of the robustness of the maximum while requiring two views to agree. Objects with a single registration view fall back to that view's similarity.

The descriptor has 1280 dimensions, the channel count of the last MobileNetV2 stage. When stored in float32 format, a template occupies approximately 5.0 kB, and as int8 with a per-vector scale it occupies 1.25 kB, so ten thousand objects require about 12.5 MB of storage, which fits comfortably on the terminals used in the institutions we studied. The same int8 representation is what the quantised network in Section 3.3.5 produces, so templates and query descriptors are computed in the same arithmetic and compared without rescaling.

C. The Acquisition-Condition Gap and the Decision Rule

Registration and query images are produced under systematically different conditions, and the difference is not noise that more data will average away. Registration is a controlled act: the object is placed on a neutral sheet, the camera is supported, the light is fixed, and several views are taken in one session. A query is an uncontrolled act: the object is held in one hand, photographed on a shelf, seen through the glass of a display case, or captured under mixed daylight and tungsten, at whatever distance the situation allows. Formally, if P_r denotes the distribution of registration images and P_q that of query images of the same object, the two distributions share the underlying identity label but differ in acquisition-related characteristics, and a descriptor learned or selected to minimise error under P_r is not thereby minimising error under P_q. The quantity that matters operationally is the drop in accuracy when the evaluation moves from P_r to P_q, which Section 3.3.2 measures directly by holding the objects fixed and changing only the acquisition condition.

It is important to explain why this gap could not be eliminated merely by increasing model capacity, since that was the first hypothesis we tested. When the same objects are compared under the same condition, the pipeline reaches 77.7% top-1 accuracy; across conditions it reaches 67.2%. Enlarging the encoder from 2.23M to 85.8M parameters does not recover the difference, and in fact reduces accuracy at the upper end of that range; training a projection head with triplet, classification and relational-distillation objectives reduces it further; colour normalisation and k-reciprocal re-ranking also reduce it. Across these experiments, the observed pattern is consistent: the information that separates the two conditions is not a matter of representational capacity, and no objective computed on a few dozen identities supplies it. The design consequence is that the effort should be concentrated on the acquisition protocol, in the choice of how many registration views to take and from which angles, and in deciding how much human review the institution can afford, rather than in the network.

The decision rule follows from that consequence. For a query descriptor x, the system reports the object with the highest similarity and accepts the identification only if that similarity exceeds a threshold τ:

$$accept \;\Leftrightarrow\; s\,(x,\,k^{*}) \;\geq\; \tau, \qquad k^{*} \;=\; arg\,max_{k}\; s(x,\,k). \tag{6}$$

The threshold is not chosen to maximise accuracy on the registered objects, because the cost of the two error types is asymmetric and neither is a classification error in the ordinary sense. A false acceptance writes a record against the wrong object, which is the failure a museum or archive most needs to avoid; a false rejection costs one manual check. We therefore fix τ so that the fraction of known objects rejected stays within the review capacity the institution can absorb, typically between five and twenty per cent of verification attempts, and report the false acceptance rate that this implies, as in Section 3.3.3. In the proposed user interfaces, a result below τ is presented as a short candidate list rather than as a refusal, because a ranked list of five candidates is still useful to a registrar who must confirm an accession number by hand.

The identity record binds the descriptor to the administrative object. For object k we store the tuple

$$\begin{cases} D_k = (ID_k,\; \mu_k,\; M_k,\; V_k,\; H_k)\,, \\ H_k = h\big(ID_k \,\|\, \mu_k \,\|\, M_k \,\|V_k \,\|\, H_{k-1}\big), \end{cases} \tag{7}$$

where ID_k is the accession or archive number, μ_k the template, M_k the collection metadata that the institution already maintains (designation, sub-type, material, date, dimensions, storage location), V_k the identifier of the encoder and the template format, and H_k a hash chained to the previous record so that silent edits are detectable. Two properties follow which matter more than accuracy in an archival setting. The record is self-describing, so a verification performed years later can be interpreted even if the encoder has since been replaced, because V_k names the encoder that produced μ_k; and the record is verifiable offline, because matching requires only the stored templates and a local forward pass, with no network, no central server and no external service. When a device build is upgraded, templates are recomputed and V_k is incremented, so an audit trail exists of which version produced which verification.

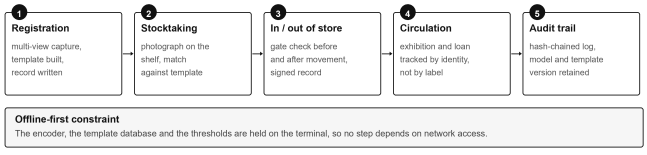


Figure 3: Lifecycle of a digital identity in archival work. Registration, stocktaking, gate verification, circulation and audit all reduce to a comparison against the stored template, and all are executable on a local terminal without network access.

Figure 3 places the record in the archival workflow it is meant to serve. The same template supports stocktaking against the shelf, verification at the store door, tracking during exhibition and loan, and the audit trail that follows; each of these is a comparison against μ_{k}, so the marginal cost of adding a workflow step is a user interface rather than a model.

III. Experiments

A. Objectives and Evaluation Design

The experiments answer four questions, in the order in which they matter to an institution that has to decide whether to adopt the method. First, how accurately can a registered object be identified from a query image taken under a different condition? Second, how much of the error is attributable to that difference in condition rather than to the objects themselves? Third, how reliably are objects that were never registered rejected? Fourth, what does the method cost to run on the hardware that such an institution is likely to operate? A fifth question, whether task-specific training improves any of the above, is treated as an ablation, because its answer determines whether the pipeline needs a training stage at all.

Three protocols are used, and it matters which number comes from which. The closed-set protocol registers all 35 known objects with their registration views and then identifies each of the 177 query views against the full set of 35 candidates; every accuracy in Tables 2 to 4 and in Figures 5 to 7 refers to this protocol, whereas Table 5 and the trained bars of Figure 8 use the five-fold protocol described below. The holdout protocol splits the query set by parity of file order into two halves, selects the template aggregation rule on the first half and reports the result on the second, which is the only safeguard available here against tuning on the evaluation set; with 177 queries, a single query represents approximately 0.56 percentage points, so differences below about two points should not be read as meaningful. The object-level five-fold protocol is used only where a trained component must be compared with an untrained one: in each fold 28 objects are used for training and the remaining 7 are held out, the candidate pool remains all 35 objects, and the frozen pipeline is scored on the same held-out objects. This last restriction is essential: a comparison that omits it scores trained and untrained variants over different object sets and is void, because the held-out objects then differ between the arms being compared.

For the open-set protocol the 65 query images of the 7 unknown objects are added to the candidate pool as objects that must not be accepted, and rejection is scored by the false acceptance rate, the false rejection rate, the equal error rate and the area under the receiver-operating characteristic curve. The threshold is swept in steps of 0.001 over the score range and the equal error rate is reported as the mean of the two error rates at the sampled threshold at which they are closest, so the crossing need not be solved exactly. Latency is the median and the 95th percentile of the end-to-end time of one image, from decoded pixels to a similarity score, measured over 50 runs after 8 warm-up runs on a 24-core Intel CPU with ONNX Runtime 1.28, using either one thread or four and reporting which. Memory is the peak resident set of the process, which includes about 50 MB of runtime baseline that any Python deployment would carry.

B. Corpus Construction and Audit

The corpus was built from the archive of a heritage conservation institution and records the categories its archive actually holds rather than categories chosen to make the task convenient. Thirty-five objects were registered and photographed repeatedly; seven further objects were photographed but deliberately withheld from registration to serve as unknowns. Each object was photographed under two conditions in separate sessions. Registration views were taken with the object placed on a neutral sheet under fixed light, with the camera supported; query views were taken as the object is ordinarily encountered, handheld or on a shelf, sometimes through the glass of a display case, at whatever distance and angle the situation allowed, in mixed daylight and interior lighting. Figure 4 shows illustrative objects under each acquisition condition and makes the difference concrete: the same trophy is a flat-lay on white paper in registration and a vertical object against a door in query, and the same seal is a flat ink impression on paper in registration and a three-dimensional block seen at an oblique angle in query.

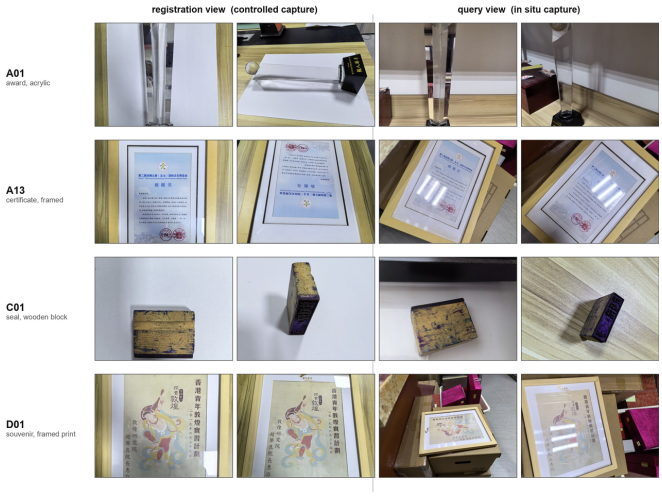


Figure 4: The same objects photographed under the two acquisition conditions. Left: registration views, object on a neutral sheet under fixed light. Right: query views, handheld or on a shelf under mixed light. A01 award, A13 certificate, C01 wooden seal, D01 framed print.

Table 1 summarises the composition. Four groups are represented: honorary awards, which dominate the corpus and comprise trophies, wall plaques, certificates and medals; seals, both wooden and rubber; prints and souvenirs, both paper-based and porcelain; and one object proper. The dominance of the honorary group is not a design choice but a consequence of access: approval to photograph objects classified as cultural relics could not be obtained within the project period, and display cases restrict the angles available for what is exhibited,

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

so the corpus is weighted towards documentary and honorary material that the archive can handle directly. Section 3.3 reports accuracy per group as well as in aggregate, so that this imbalance remains visible instead of being averaged away. The single object proper is a modern reproduction and is labelled as such.

Table 1: Composition of the corpus. Registration and query images were acquired in separate sessions under different conditions; the seven unknown objects contribute query images only.

Group	Sub-type	Objects	Materials	Registration	Query
A Honorary awards	trophies	7	acrylic, composite, glass	42	35
	wall plaques	5	metal, composite	25	26
	certificates	6	composite, framed paper	29	30
	medals	5	metal	29	24
C Seals	wooden and rubber seals	5	wood, rubber	27	24
D Prints and souvenirs	prints, framed souvenirs, porcelain	6	paper, porcelain	35	32
F Object proper	painted pottery (reproduction)	1	earthenware	6	6
Known total		35		193	177
Unknown	held out from registration	7	mixed	0	65
Total		42		193	242

All images are 3072 $\times$ 4096 or 4096 $\times$ 3072 pixels, in portrait or landscape orientation as the object required, so the 224 $\times$ 224 network input is a downsampling by a factor of fourteen to eighteen and no image is upsampled. Two audits were run before any experiment, because the usual way for an instance-retrieval result to be inflated is leakage between the registration and query halves. First, every pair of images was compared by an 8 $\times$ 8 average hash, with a Hamming distance of four or less taken as a near-duplicate; across the 94,395 pairs formed by the 435 images, no near-duplicate pair crossed the registration-query boundary for the same object and no pair crossed between objects. Second, sharpness was estimated by the variance of the Laplacian; the median is 605, and the distribution has a lower tail of eight images between 43 and 79, all of them registration views of small seals photographed at minimum focus distance with a shallow depth of field. Those views were retained, because a registrar photographing a 4 cm seal with a handheld camera will produce exactly that kind of image, and removing them would flatter the result.

We do not report results on the public benchmarks discussed in Section 1. The Met Dataset, for instance, contains roughly two orders of magnitude more classes than our corpus and is designed for a scale of evaluation we could not run within the resources of this study; reporting a reduced subset of it would invite exactly the comparison that the subset makes invalid. The claims in this paper are therefore claims about a small archival corpus of the kind a grassroots institution holds, and they are scoped accordingly.

C. Results

1) Identification of Registered Objects

With the pooling and the template rule fixed as described in Section 2, the pipeline identifies 67.2% of the 177 query views correctly at rank one and 92.7% within the first five candidates. The holdout check behaves as it should: the mean-of-top-two rule was selected on the first half of the query set, where it gave 64.0%, and reproduced 70.5% top-1 and 92.0% top-5 on the second half, which the selection never saw. The six-point spread between the halves is a reminder of how much a corpus of this size moves with individual images, and is the reason we report per-group figures rather than a single number [18].

Table 2 breaks the aggregate down by group and records the number of query views behind each figure.

Table 2: Closed-set identification accuracy over 177 query views and 35 registered objects. Group and sub-type rows share the same 35-object candidate pool; n is the number of query views.

Group / sub-type	Objects	n	Top-1 (%)
A Honorary awards	23	115	61.7
trophies	7	35	48.6
wall plaques	5	26	53.8
certificates	6	30	70.0
medals	5	24	79.2
C Seals	5	24	54.2
D Prints and souvenirs	6	32	90.6
F Object proper	1	6	100.0
All	35	177	67.2
All, top-5	35	177	92.7
Holdout half B, unseen during selection	35	88	70.5

The group-specific results offer more insight than the aggregate score alone. Prints and souvenirs reach 90.6% and the single object proper is identified in all six of its query views, whereas seals reach 54.2% and trophies 48.6%. The ordering follows the amount of view-invariant structure the object offers. A framed print has a large, flat, high-contrast graphic surface that survives a change of viewpoint, and its descriptor is dominated by the print itself. A seal is a few centimetres across, its registration view is a flat ink impression on paper while its query view is a three-dimensional block seen at an angle, and the two have almost no shared image structure; the block and its impression are, in effect, two different visual objects. Trophies fail for a related reason: they are tall, thin and specular, their registration views are flat-lays while their query views are vertical, and several carry the same engraved competition text with only the year and the placing differing. This is the same-type confusion quantified in Section 3.3.2, and it is where a records application is most at risk, since confusing two trophies of one series is precisely the error that produces an incorrect accession entry.

Figure 5 shows three query cases with their three highest-ranked candidates. Two deserve comment. In row (b) a wall plaque is matched to a different plaque of the same series at 0.857, followed by a second plaque of that series at 0.850, while the correct object does not appear among the three highest-ranked candidates; the plaques differ in the awarding body and the year, which occupy a small part of the frame, and the descriptors are dominated by the shared layout. This is the failure mode of greatest practical relevance: not a failure to recognise a type, but a failure to separate instances within a type. In row (c) the correct seal is ranked first at 0.722 with the runner-up at 0.722 as well, so the margin is effectively zero and the outcome is decided by the fourth decimal place. A

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

registrar reading such a result should not be told that the object has been identified; the interface must show the candidate list and the margin, which is why the record design of Section 2.3 keeps the ranking rather than only the top label.

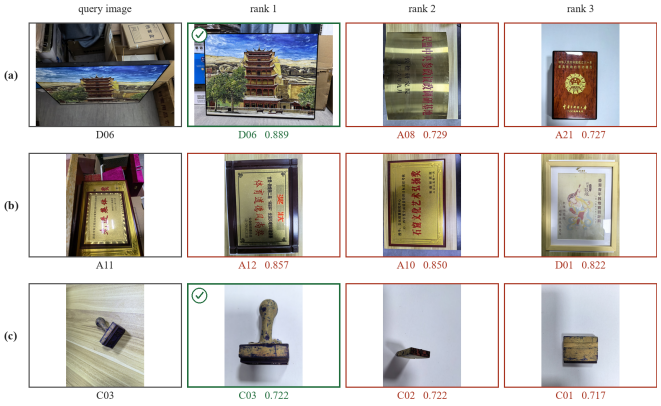


Figure 5: Query view, three highest-ranked candidates and cosine similarity. A green frame marks the correct object, a red frame an incorrect one. Row (b) is a same-type confusion between two plaques of one series; row (c) is a correct match whose margin over the runner-up is zero to three decimal places.

2) The Cost of the Acquisition Condition

The central measurement of this paper is obtained by holding the object set fixed and varying only the acquisition condition. Each registration view is in turn used as a probe and removed from its own template, so that no probe can match itself, while the remaining views of every object are scored with the top-two rule of Eq. (5), exactly as queries are; all 35 objects compete. The pipeline then reaches 77.7% top-1 and 97.9% top-5. Applied to the query views, the same pipeline with the same templates reaches 67.2% and 92.7%. The difference of 10.5 points at rank one and 5.2 points at rank five is the measured effect of changing the acquisition condition: the objects, the encoder, the pooling, the scoring rule and the templates are identical, and only the way the photograph was taken differs. Figure 6 reports the two conditions side by side. The conclusion does not depend on the aggregation rule: if the maximum over the remaining registration views replaces the top-two mean, the same-condition baseline rises to 81.9% and 98.4%, and the gap to 14.7 and 5.7 points, which bounds the effect from above rather than altering its interpretation.

A second view of the same effect is obtained by subsampling the templates. With one registration view per object, accuracy is 48.5%; two views give 57.1%, three 61.3% and four 64.9%, against 67.2% with all views, whose mean is 5.5 per object (Table 3).

The accuracy curve rises steeply initially and begins to flatten after the third registration view, which has a direct operational reading: a single photograph per object, which is what many small collections hold, is not enough to support identity verification even with a perfect encoder, whereas four to five views taken in one session recover most of what is available. The standard deviation falls from 3.5 to 2.3 points over the same range, so additional views also make the outcome less dependent on which views happen to be available. This is the most actionable result for an institution: the marginal return on a third and fourth photograph is much larger than the marginal return on a larger network.

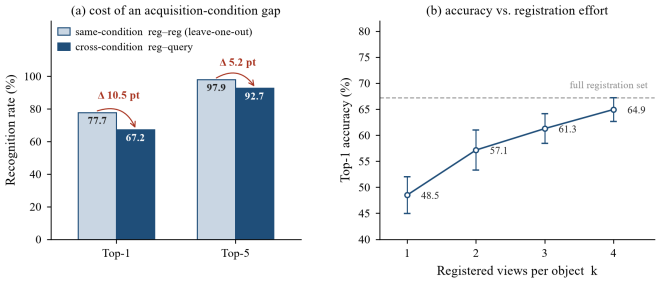


Figure 6: Identification over identical objects under different acquisition conditions. Left: each registration view used as a probe against the other registration views of the same objects, scored with the top-two rule. Right: query views scored against the same templates. The 10.5-point difference at rank one is attributable to the acquisition condition, since nothing else changes.

Table 3: Accuracy as a function of the number of registration views per object. Views are sampled at random without replacement from each object's registration set; mean and standard deviation over five draws.

Registration views per object	Top-1 (%)	s.d. (points)
1	48.5	3.5
2	57.1	3.8
3	61.3	2.9
4	64.9	2.3
all (mean 5.5)	67.2	—

3) Rejection of Unregistered Objects

The 65 query views of the seven unknown objects were mixed into the candidate pool. The similarity of a known query to its own template averages 0.814 with a standard deviation of 0.048; the highest similarity of an unknown object to any template averages 0.777 with a standard deviation of 0.040. The two distributions overlap heavily, the equal error rate is 33.1% at a threshold of 0.795, and the area under the curve is 0.730. Table 4 gives the operating points that matter for a review workflow.

If the institution can absorb a 10% manual re-check rate, the threshold still admits 64.6% of unknown objects, and tolerating a 20% re-check rate reduces that to 41.5%.

Figure 7 shows the trade-off curve and the two score distributions. The overlap is the honest headline of this section: a 1280-dimensional descriptor from a frozen network, compared by cosine similarity, does not by itself decide whether an object belongs to the collection. Two properties of the

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

experiment make this number pessimistic relative to a deployment figure, and both should be stated. The unknown objects come from the same institution as the registered ones, so they share materials, manufacturers and lighting with them and are harder to reject than objects from another collection would be; and the unknown set contains several awards of the same type as the registered awards, which is the adversarial case. Against a genuinely external set the same threshold would produce a lower false acceptance rate, although we cannot say by how much without such a set.

Table 4: Open-set operation. Thresholds are placed on the highest object score of Eq. (6), that is on the largest of the 35 values of s(x, k); 177 known and 65 unknown query views. The equal-error row is the sampled threshold at which the two rates lie closest to each other, and the equal error rate quoted in the text, 33.1%, is defined as the mean of the two rates at that threshold (33.9% and 32.3%), the crossing itself falling between two sampled thresholds.

Operating point	Threshold τ	FRR (%)	FAR (%)
Equal-error point	0.795	33.9	32.3
Review budget 5%	0.736	5.1	86.2
Review budget 10%	0.751	10.2	64.6
Review budget 20%	0.771	20.3	41.5

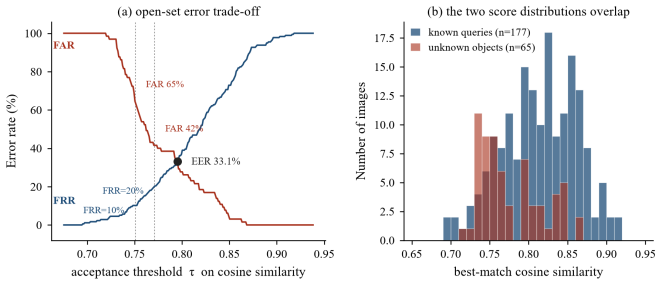


Figure 7: Open-set rejection. (a) False acceptance and false rejection rates as the acceptance threshold varies, with the equal error point and the two review-budget operating points marked. (b) Distribution of the score of Eq. (6) for known and unknown query views.

The operational conclusion is a specification rather than a number. The system must therefore be presented as a candidate-ranking aid: it returns the best match, the runner-up, the margin between them and a flag when that margin is small, and the registrar confirms. A threshold calibrated to the available review budget therefore serves a useful purpose, because it concentrates human attention on the cases that need it, which is the same logic as a triage rule in a reading room. It does not do the work of deciding identity, and describing the system as fully automatic identity verification would be misleading at this observed error level.

4) Training-Based Adaptation and Post-Processing

Because the usual way to improve a descriptor is to train it, we trained one under the same protocol used to score the frozen pipeline. A projection head of one linear layer to 128 or 256 dimensions with batch normalisation was placed on the frozen 1280-dimensional descriptor and trained on the 28 training objects of each fold with a batch-hard triplet loss at margin 0.3, following the standard metric-learning formulation of [19]; further variants added a softmax cross-entropy loss over the training identities, an additive angular margin loss following [20], and a relational distillation term from a ResNet-50 teacher, using the distance-wise and angle-wise formulation of [13]. The results are in Table 5.

Table 5: Trained adaptation compared with the frozen pipeline under an identical object-level five-fold protocol: each fold trains on 28 objects and is scored on the 7 held-out objects against a 35-object candidate pool; the frozen pipeline is scored on the same held-out objects.

Variant	Top-1 (%)	s.d. (points)	EER (%)	AUROC
Frozen descriptor, no training	66.4	8.6	33.1	0.730
+ triplet head, 128-d	58.9	9.1	29.3	0.781
+ classification and triplet head	57.7	5.8	27.7	0.755
+ classification, triplet and RKD from ResNet-50	56.6	6.8	27.7	0.761
ResNet-50 teacher alone, frozen	58.7	4.9	36.8	0.661

Every trained variant is worse at identification than leaving the descriptor alone, by between 7.5 and 9.8 points, and the variant with the most machinery is the worst. The teacher is itself worse than the frozen baseline, at 58.7% against 66.4%, which means the distillation target lay below the level the student could reach without it, so the distillation term could only pull the student down. There is a real and quantifiable benefit in the other direction: the equal error rate improves from 33.1% to 27.7% and the area under the curve from 0.730 to 0.755 in the variant that combines the classification and triplet heads, which is the lowest equal error rate in the table, so training does make the similarity scores more separable between known and unknown, at the cost of reduced identification accuracy among known objects. The best area under the curve, 0.781, is reached by the triplet head alone, at an equal error rate of 29.3%; the two criteria are not maximised by the same variant. That trade is defensible for a triage system and indefensible for an identification system, and it is the reason the pipeline reported here contains no trained component: with 35 identities there is little for a metric-learning objective to learn that the pre-trained features do not already encode, and the 28 identities available in a fold are too few to estimate a margin.

The capacity scan in Table 6 makes the same point from a different direction. Under one identical protocol, 224-pixel input, global average pooling and a maximum-over-views template, accuracy rises from 48.6% for ShuffleNetV2 at 1.37M parameters, through 49.2% for MobileNetV3-Small [21], to 62.7% at 11.2M, and then falls again, to 57.6% at 23.5M and 58.2% at 85.8M.

Two mechanisms are plausible and this data cannot separate them: larger ImageNet classifiers such as ResNet-50 [22] are trained for semantic category separation, which is not the objective here, and efficient scaling rules [23] distribute representational capacity over many dimensions that are irrelevant to the identity of one trophy. Either way, the practical

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

conclusion is that the mobile-scale network is not a concession to the hardware but the right choice on accuracy grounds, which is a more comfortable position for a lightweight paper than the usual apology.

Table 6: Capacity scan under one identical protocol: frozen ImageNet weights, 224-pixel centre crop, global average pooling, maximum-over-views template, 35 candidates and 177 query views. Parameters are counted for the network that produced the descriptor, with the ImageNet classifier removed where the implementation exposes it as a separate module; ShuffleNetV2 is the exception, since its released definition evaluates the 1000-way layer inside the forward pass, so the 1.37M figure includes that layer. MobileNetV2 appears at 59.3% here because this protocol uses global average pooling rather than the GeM pooling of the proposed pipeline, which accounts for the difference from Table 2.

Encoder	Parameters (M)	Descriptor	Top-1 (%)	Top-5 (%)
ShuffleNetV2 ×0.5	1.37	1000	48.6	75.1
MobileNetV3-Small [21]	1.53	576	49.2	81.4
MobileNetV2, proposed encoder [7]	2.23	1280	59.3	85.3
ResNet-18 [22]	11.18	512	62.7	88.1
ResNet-50 [22]	23.51	2048	57.6	91.0
ConvNeXt-T	27.82	768	62.2	87.6
ViT-B/16	85.80	768	58.2	89.8

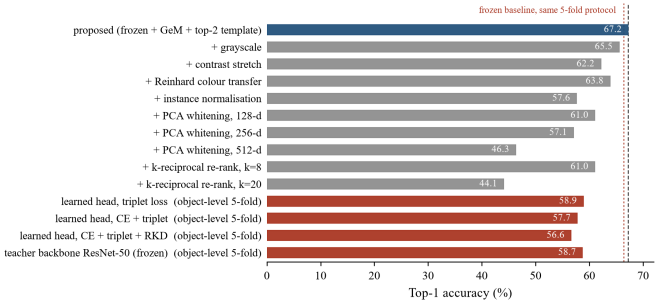


Figure 8: Post-processing and training ablations against the proposed pipeline. Grey: fixed transformations applied to the descriptor. Red: components trained under the object-level five-fold protocol, to be compared with the frozen baseline at 66.4% on the same protocol (dash-dot line) rather than with the 67.2% full-protocol result (dashed line).

Figure 8 collects the post-processing and training ablations. Colour normalisation, PCA whitening at any dimensionality and k-reciprocal re-ranking all reduce accuracy, and the reductions are large: whitening to 512 dimensions costs 20.9 points and re-ranking with k = 20 costs 23.1. These negative findings are reported because the techniques are commonly recommended for cross-domain retrieval, and on this corpus they fail for an intelligible reason. Each estimates a transformation from the available data, and 35 templates are not enough to estimate a whitening matrix of rank 512 or a neighbour graph of degree 20; the estimate is dominated by the sample and removes as much identity information as condition information. The same argument explains why the learned heads of Table 5 lose accuracy, and it is one explanation rather than three: at this scale, anything that has to be estimated from the collection identifies less accurately than the fixed alternative, although it can still sharpen the known-unknown separation.

5) Deployment Cost

The pipeline was exported to ONNX and measured on the CPU of an ordinary workstation, with one thread and four, in float32 and in two integer formats (Table 7).

Table 7: Deployment measurements on a 24-core Intel CPU with ONNX Runtime 1.28. Latency is the end-to-end time per image from decoded pixels to similarity score, median over 50 runs after 8 warm-up runs; peak memory is the process resident set including about 50 MB of runtime baseline, so differences between rows are the meaningful part.

Configuration	Size (MB)	Median (ms)	T95 (ms)	Peak RSS (MB)	Top-1 (%)
FP32, 1 thread	8.46	6.28	10.83	80.6	67.2
INT8 dynamic, 1 thread	2.30	139.75	142.82	86.6	63.8
INT8 static, 1 thread	2.49	4.14	7.14	70.8	66.1
FP32, 4 threads	8.46	3.72	3.79	—	67.2
INT8 dynamic, 4 threads	2.30	79.42	94.49	—	63.8
INT8 static, 4 threads	2.49	2.39	2.74	—	66.1

In float32 format, the model occupies 8.46 MB, takes 6.28 ms at the median and 10.83 ms at the 95th percentile with one thread, and 3.72 ms at the median with four. Static quantisation to INT8, with activations calibrated on 120 registration images, reduces the size to 2.49 MB and the median latency to 4.14 ms, a speed-up of 1.5 times, at the cost of 1.1 points of top-1 accuracy and 0.8 points of equal error rate, which moves from 33.1% to 33.9% while the area under the curve changes from 0.730 to 0.737. Dynamic quantisation, which quantises weights only and computes activations in floating point, produces a slightly smaller file of 2.30 MB but takes 139.75 ms at the median, thirty-four times the latency of static INT8 and twenty-two times that of float32, because activation quantisation is then performed per inference on the CPU with no optimised kernel behind it. Dynamic quantisation is thus dominated on both counts and is not a candidate for deployment.

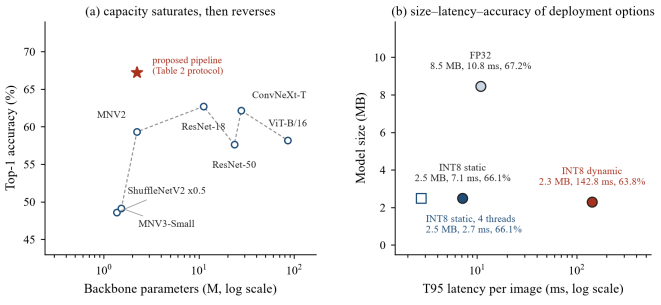


Figure 9: Deployment configuration. (a) The capacity scan of Table 6: accuracy against backbone size, with the proposed pipeline marked. (b) Accuracy against model size and 95th-percentile latency for the deployment configurations of Table 7.

Figure 9 places the configurations on the accuracy-size-latency plane together with the capacity scan. The useful

region is sharply bounded: static INT8 at 2.49 MB and 4.14 ms loses one point of accuracy, and no configuration we measured improves on it in every coordinate at once. For the terminals described in Section 1, where verification is an occasional operation and 200 ms is imperceptible during stocktaking, the float32 model would already meet the computational requirement; the quantised model matters because it lets the template database and the runtime sit comfortably within the memory a low-cost industrial computer leaves free, and because a 2.5 MB artefact can be shipped and versioned together with the record.

D. Discussion

Three findings seem to us to transfer beyond this corpus. The first is that the acquisition condition, not the model, is the dominant term in the error budget at this scale: the same pipeline loses 10.5 points when only the way the photograph was taken changes, whereas moving from 2.2M to 85.8M parameters changes accuracy by less than six points and in the wrong direction. The second is that every estimate drawn from a 35-object collection costs identification accuracy, whether it is a whitening matrix, a neighbour graph for re-ranking or a trained projection head, because the sample dominates the estimate. Training is not uniformly harmful: it lowers the equal error rate from 33.1% to as little as 27.7% and raises the area under the curve from 0.730 to as much as 0.781, so it does separate known from unknown objects better, but it pays 7.5 to 9.8 points of top-1 for that, which is the wrong trade for an identification system. The third is that the number of registration views is the most effective single operational variable available to an institution: going from one to four views recovers 16.4 points, which no architectural change in our experiments achieved. Taken together, these findings reverse the usual allocation of effort in a lightweight recognition project: the acquisition protocol, the registration procedure and the record design deserve more attention than the network, and the network can be a published, frozen, quantised model that never sees the collection at all.

The study limitations are similarly specific and important. The corpus has 35 registered objects from one institution, of which 23 are honorary awards and one is an object proper, so the per-group figures rest on small numbers and the aggregate is dominated by a group that is atypical of archival collections. The unknown set is drawn from the same institution and therefore overstates the false acceptance rate relative to cross-institution deployment. Photographs were taken by the authors, and although the registration and query sessions were separated in time and in procedure, one photographer's habits may reduce the condition gap relative to images contributed by different staff over several years. Latency was measured on one CPU with one runtime: the relative ordering of the quantisation schemes is likely to hold on other x86 hardware, the absolute values are not. Labels are administrative and were not independently verified object by object, a limitation shared by every study that builds a corpus from a working collection. Finally, no attempt is made here to explain individual decisions by localisation [24]; for a review workflow that is the natural next component, since a registrar is more likely to accept a candidate that comes with a highlighted region than one that comes with a number.

Two extensions follow directly and are the subject of current work. An external unknown set obtained from the Digital Dunhuang resource platform of the Dunhuang Academy, consisting of mural and sculpture units that share materials and palette with the archive but belong to no registered object, would allow the false acceptance rate to be estimated under a genuinely cross-collection condition. A study in which several staff photograph the same objects over a period of months would separate the condition gap into a component due to session and lighting and a component due to photographer, which is the decomposition an institution needs in order to write a capture protocol.

IV. Conclusion

This paper set out to determine what the least expensive architecture is that still yields a usable digital identity for a small archival collection, and what limits its accuracy. The resulting answer is more narrowly defined than the initial research question might suggest. A frozen ImageNet MobileNetV2 with GeM pooling, a mean-of-top-two similarity rule and an acceptance threshold yield 67.2% top-1 and 92.7% top-5 accuracy over 177 query views of 35 registered objects, at 2.49 MB and 4.14 ms after static INT8 quantisation, with no training stage, no collection-specific parameters and no retraining when an object is accessioned. The same pipeline reaches 77.7% when the acquisition condition is held constant, so 10.5 points of the error budget belong to the condition rather than to the model; larger backbones, trained projection heads with relational distillation, colour normalisation and re-ranking all fail to recover them, and the teacher in the distillation experiment is itself worse than the frozen baseline.

For an institution the practical consequences are specific: take at least four registration views per object in one session; record the capture protocol alongside the object, because a template is only valid under the conditions it was built for; keep the descriptor, the metadata, the model and template version and an integrity hash in one record, so that a verification can be interpreted and audited years later; and specify the system as a candidate-ranking aid with an explicit review budget, because, at an equal error rate of 33.1%, it can narrow a registrar's search but cannot reliably make the final decision. The principal contribution is not a newly developed network but a documented negative result about where the effort belongs, together with a record design that makes a training-free matcher usable in archival practice.

Author Contributions

Both authors contributed equally to this research and approved the final version for publication.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Funding

This research was supported by the Fan Jinshi Fund of the Dunhuang Grottoes Conservation and Research Foundation of China, through the project entitled "Research on the Construction of a Standard System for the 'Activation and Utilization' of Dunhuang Grottoes Archives in the Digital Age" (Project No. 2025-FJS-YB-2).

Data Availability

The corpus described in Section 3.2 was collected from the archive of a heritage conservation institution and cannot be released publicly, because the images identify individual accessioned objects held in a closed archive. The processing scripts, the template construction and the evaluation protocols are available from the corresponding author on reasonable request; a reproducibility package would include the feature extraction pipeline, the data-split definitions, and the ONNX export and quantisation steps, which would enable the evaluation procedures in Section 3 to be reproduced on another corpus assembled using the same protocol.

References

[1] National Library of Australia. (2003). *Guidelines for the preservation of digital heritage* (CI-2003/WS/3). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000130071

[2] International Organization for Standardization. (2023). *Information and documentation—A reference ontology for the interchange of cultural heritage information (ISO 21127:2023)*. https://www.iso.org/standard/85100.html

[3] Winterbottom, T., Leone, A., & Al Moubayed, N. (2022). A deep learning approach to fight illicit trafficking of antiquities using artefact instance classification. *Scientific Reports, 12*, Article 13468. https://doi.org/10.1038/s41598-022-15965-2

[4] Ypsilantis, N.-A., Garcia, N., Han, G., Ibrahimi, S., van Noord, N., & Tolias, G. (2021). The Met dataset: Instance-level recognition for artworks. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1). https://cmp.felk.cvut.cz/met/

[5] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 119, pp. 1597–1607). PMLR. https://proceedings.mlr.press/v119/chen20j.html

[6] Bartkowiak, P., Markil, J., Kotrys, B., Michels, D. L., Pirk, S., & Pałubicki, W. (2026). *SynGallery: A synthetic gallery of real paintings for instance-level artwork recognition* [Preprint]. arXiv (arXiv:2607.18907). https://arxiv.org/abs/2607.18907

[7] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE. https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

[8] Qin, D., Leichner, C. H., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., & Howard, A. (2025). MobileNetV4: Universal models for the mobile ecosystem. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision—ECCV 2024* (Lecture Notes in Computer Science, Vol. 15098, pp. 78–96). Springer. https://doi.org/10.1007/978-3-031-73661-2_5

[9] Minango, P., Zambrano, M., Huerta Suarez, C. I., & Minango, J. (2026). Deep-learning-based mobile application for real-time recognition of cultural artifacts in museum environments. *Applied Sciences, 16*(9), Article 4064. https://doi.org/10.3390/app16094064

[10] Fang, G., Ma, X., Song, M., Mi, M. B., & Wang, X. (2023). DepGraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16091–16101). IEEE. https://openaccess.thecvf.com/content/CVPR2023/html/Fang_DepGraph_Towards_Any_Structural_Pruning_CVPR_2023_paper.html

[11] Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization, and Huffman coding. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1510.00149

[12] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2704–2713). IEEE. https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html

[13] Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3967–3976). IEEE. https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html

[14] Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* [Preprint]. arXiv (arXiv:1503.02531). https://arxiv.org/abs/1503.02531

[15] Bendale, A., & Boult, T. E. (2016). Towards open-set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1563–1572). IEEE. https://openaccess.thecvf.com/content_cvpr_2016/html/Bendale_Towards_Open_Set_CVPR_2016_paper.html

[16] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large-scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[18] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE, 14*(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365

[19] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823). IEEE. https://openaccess.thecvf.com/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

[20] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4690–4699). IEEE. https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html

[21] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314–1324). IEEE. https://openaccess.thecvf.com/content_ICCV_2019/html/Howard_Searching_for_MobileNetV3_ICCV_2019_paper.html

[22] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[23] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 97, pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[24] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html