Selective Identity Retrieval from Candidate-Level Separability

Joseph Quadrino
joseph.quadrino@yourhilbert.space

Abstract

A retrieval system ranks candidates. A system that attaches an identifier, links a record, or merges two entries must also decide whether the top candidate *is* the queried item, and should decline when the evidence does not support that. We study this decision as *selective identity retrieval*: after ranking, a gate either commits to one candidate identity or abstains. The hypothesis is that a true match tends to appear as a candidate *separated* from the rest of the list, not merely as the candidate with the highest score. We test it on a public benchmark of 60,502 images of 11,316 products (59,716 leave-one-image-out queries, each scored once with its identity in the gallery and once with it removed) and on two small private catalogs, comparing thirteen confidence scores with bootstrap intervals: closed-form separability measures, local-density scores, temperature-scaled softmax confidence, a small learned gate, and singleton-mode gating by mean shift (included because it was a natural clustering operationalisation of the hypothesis, not because we treat it as the preferred rule). The hypothesis holds on the benchmark: every separability score is a much better gate than the top similarity (for the top-two margin, 10.3% correct identifications at a 1% false-accept rate against 6.5%, and AUROC 0.866 against 0.712). A distance-ratio test and the learned gate are the strongest single-channel scores; the margin, a fixed-bandwidth density score and mean shift are nearly interchangeable, and mean shift adds nothing beyond the margin (rank correlation of critical bandwidth with the margin 0.983; no stable bandwidth interval). List-scaled scores lose a few points of accuracy within a collection but transfer: thresholds set on the benchmark for 5% false accepts produce 18–29% on the other collections and with another encoder for raw-scale scores, against 1–5% for list-scaled ones. With several channels, agreement rules are best at strict false-accept rates; a Pareto mean-shift rule becomes competitive only once a third channel is added, and a learned multi-channel gate matches or exceeds both. On one private catalog, in which the same image is often filed under several identities, no gate works. The evidence supports separability, not maximal similarity, as the basis for an automated identity decision.

1 Introduction

Retrieval research has mostly asked how to order candidates. Hybrid methods such as reciprocal rank fusion [9] and weighted combinations of dense and sparse scores improve the order, and ranking metrics measure the improvement. Many deployed systems need a second answer that a ranking does not contain. A system that attaches a catalog identifier to an incoming image, resolves a document to a canonical record, or links an extracted entity to a database row takes an action that is costly to reverse. For such a system the relevant question is whether the top candidate is the queried item, and a false commitment usually costs more than a refusal to commit.

We separate the two decisions explicitly (Figure 1). Retrieval and ranking produce an ordered list of candidate identities with scores. A *gate* then inspects the scores and either commits to one candidate or abstains, in which case the ranked list is returned unchanged or routed to review. The gate uses only quantities that retrieval has already computed, so it requires no change to the

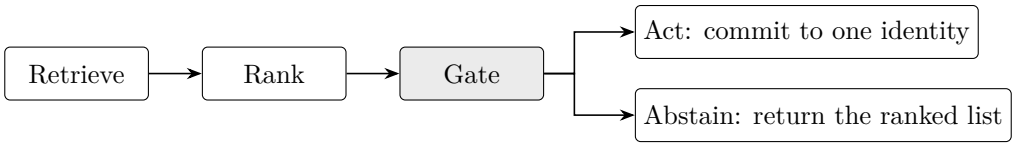


Figure 1: Selective identity retrieval. The gate reads the candidate scores produced by retrieval and ranking and decides whether they justify committing to a single identity.

encoder, the index, or the ranker. We call the task *selective identity retrieval*, by analogy with selective classification [7, 12].

Hypothesis. The simplest gate thresholds the top similarity. This paper starts from a different view of what the scores say about identity: an automated identity decision should require evidence that one candidate is *separable* from the alternatives, and a high score alone is not such evidence. A list whose best candidate scores 0.94 with the next candidates near 0.75 identifies an item; a list whose best candidates score 0.96, 0.955 and 0.951 does not, although its top score is higher. Whether true matches do appear as separated candidates is an empirical property of an encoder and a collection, and we treat it as a hypothesis to be tested, not as a general property of embeddings.

Scores under test. Any quantity derived from the list can serve as a confidence score. We compare thirteen, spanning closed-form separability (margin, distance ratio, list-scaled variants), local density, temperature-scaled softmax confidence, a small learned gate trained on identity-disjoint folds, and singleton-mode gating by mean shift [13, 8]. Mean shift is included because clustering the scores and accepting only a singleton top mode is a natural operationalisation of "separated", not because the study is organised around it. It raises its own questions—modes exist only relative to a bandwidth; a singleton top mode resembles a large top-two margin; with several channels, Pareto-undominated modes become rarer as channels are added—and Section 4 formalises them so that the comparison can be fair.

Contributions. We formulate selective identity retrieval with an evaluation protocol in which every query is scored once with its identity present and once with its identity removed, so that false commitments on unknown items are measured directly (Section 2). We define a family of list-derived confidence scores, including the mean-shift gate through its *critical bandwidth* (which makes it comparable with any other score), with an analytical relation to the margin for two candidates and a multi-channel formalisation with a diagonal bandwidth matrix and an analysis of abstention under independent channels (Section 4). We compare the thirteen scores on 59,716 queries from a public benchmark and on two private catalogs, with risk–coverage and open-set identification curves, bootstrap intervals, bandwidth and list-size sensitivity, and transfer of operating points across collections and encoders (Sections 5 and 6).

Findings. Separability scores are far better gates than the top similarity on the benchmark and on one of the two private catalogs. Among single-channel scores, a distance-ratio test and a small learned gate lead; the margin, fixed-bandwidth density and mean shift are nearly interchangeable, and mean shift is empirically a margin test whose threshold varies between about $2h$ and $3.3h$ with local density. No range of bandwidths gives stable decisions: the bandwidth is an operating point with the same sensitivity as a margin threshold. A threshold in raw similarity units does not carry over to another collection or encoder; a threshold expressed in units of each list's spread does, at a

measurable cost in accuracy. With several channels, agreement rules are best at strict false-accept rates and a Pareto mean-shift rule is competitive at moderate ones once a third channel is added; a learned multi-channel gate matches or exceeds both. On the second private catalog the image does not determine the identity, and every gate fails. We report these outcomes as they are, including the ones that count against the clustering rule that originally motivated the work.

2 Problem Setting

Lists and gates. A gallery contains items, each labelled with an *identity*; an identity may own several items (for example, several photographs of one product). For a query q, retrieval returns a list of N distinct candidate identities $c_1, \dots, c_N$. Candidate c_i carries a score vector $\boldsymbol{x}_i \in \mathbb{R}^k$ with one coordinate per *channel* (for example, similarity under two encoders). The score of an identity in a channel is the maximum over the items it owns, and the list consists of the N best identities under a designated primary channel. For $k = 1$ we write the scores in decreasing order as $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(N)}$.

A gate is a map from the list to $\{1, \dots, N\} \cup \{\perp\}$: it proposes one candidate or abstains. Every gate considered here has the form "propose candidate $i^\star$ if a confidence score u is at least θ, otherwise abstain", so a gate is specified by a proposal rule, a confidence score, and a threshold. For single-channel gates the proposal is always the top-ranked candidate.

Outcomes. Let $y(q)$ be the true identity of the query. The query is *present* if $y(q)$ is in the gallery and *absent* otherwise. A decision is correct if and only if the gate commits and the proposed candidate is $y(q)$. Three kinds of error follow: committing to a wrong candidate although the true one is in the list, committing although the true identity was not retrieved, and committing although the true identity is not in the gallery at all. The gate cannot tell the second from the third, and both matter in practice, since a deployed system rarely knows that every incoming item is already catalogued.

Metrics. We evaluate each query twice, once in each condition (Section 5), and report three families of metrics.

On present lists we report *selective risk at fixed coverage* [12, 14]: the lists are sorted by confidence, the most confident share c is accepted, and the risk is the share of accepted lists whose proposal is wrong.

For the open-set setting we borrow the watch-list measures of biometric identification [20]. The *false-accept rate* $\mathrm{FAR}(\theta)$ is the share of absent lists on which the gate commits. The *detection and identification rate* $\mathrm{DIR}(\theta)$ is the share of present lists on which the gate commits to the correct identity. We report DIR at the threshold whose FAR is 1%, 5% or 10%. These measures do not depend on the unknown proportion of absent queries in deployment.

As a single-number summary we also report, on the equal mixture of present and absent lists, the area under the risk–coverage curve (AURC) and the area under the ROC curve (AUROC) of the confidence score as a predictor of a correct decision. In this mixture at least half of all decisions would be wrong at full coverage, so absolute AURC values are high; they are useful for comparing scores and not as operating figures.

3 Related Work

Selective prediction and open-set recognition. Classification with a reject option goes back to Chow [7]; the risk–coverage formulation is due to El-Yaniv and Wiener [12] and was applied to deep networks by Geifman and El-Yaniv [14]. Open-set recognition [22] studies classifiers that must reject inputs from unknown classes, and maximum softmax probability is the standard baseline for detecting misclassified and out-of-distribution inputs [16]. Retrieval differs in that there is no fixed label set and no classifier head: the evidence is a list of similarities to a changing gallery. Calibration methods [21, 15] map scores to probabilities when labelled data are available, and conformal risk control [3, 1] turns a held-out sample into a threshold with a finite-sample guarantee; both are complementary to the choice of score studied here.

Biometric identification. Open-set (watch-list) identification asks the question of this paper for faces and voices, and its evaluation methodology [20] is the source of our DIR and FAR measures. Cohort-based score normalisation [2] standardises a match score against the scores of other enrolled identities, which is the z-score baseline below.

Separation of the best match. The ratio between the distances to the nearest and second-nearest neighbour is the classical test for a distinctive local-feature match [17]. Query-performance prediction estimates retrieval quality from the score distribution of the result list, for instance its clarity or dispersion [10, 23]; it targets ranking quality for a query, whereas we target a per-list commitment to one identity. The local outlier factor [6] scores how isolated a point is relative to the density around its neighbours.

Mean shift and modality. Mean shift moves points uphill on a kernel density estimate until they reach its modes [13, 8]. For a Gaussian kernel the number of modes is non-increasing in the bandwidth, which led Silverman [24] to define a critical bandwidth as a test statistic for multimodality. The critical bandwidth of Section 4.2 is the analogous quantity for the event that the top candidate has a mode to itself. A mixture of two equal Gaussians is bimodal exactly when the means are more than two standard deviations apart [11, 4], which gives the two-candidate case in closed form. The number of Pareto-undominated points among random vectors was analysed by Bentley et al. [5]; we use it as a reference for multi-channel abstention.

4 Separability Scores

All scores below are computed from one list and are oriented so that larger values indicate more confidence that the proposed candidate is the queried identity. Sections 4.1–4.3 concern a single channel and Section 4.4 several.

4.1 Closed-form scores

We group the single-channel scores by what they need to know about the scale of the similarity axis.

Absolute. The *top score* $s_{(1)}$.

Raw scale. The *margin* $s_{(1)} - s_{(2)}$. The *distance ratio* $1 - (1 - s_{(1)})/(1 - s_{(2)})$, which is Lowe's ratio test [17] applied to cosine distances. The *density at the top with a fixed bandwidth*, $-\log \sum_{j \geq 2} \exp\{-(s_{(1)} - s_{(j)})^2/2h^2\}$ up to a constant: the less kernel mass the other candidates place on the top candidate, the more isolated it is. The bandwidth h is the same for every list and is selected on held-out queries.

List-scaled. These scores divide by a scale estimated from the list itself and are therefore invariant to affine changes of the similarity axis. The *cohort z-score* $(s_{(1)} - \bar{s})/\hat{\sigma}$, with mean and standard deviation taken over the other $N-1$ candidates [2]. The *margin over MAD*, $(s_{(1)} - s_{(2)})/\hat{\sigma}_{\mathrm{MAD}}$ with $\hat{\sigma}_{\mathrm{MAD}} = 1.4826 \cdot \mathrm{median}_i |s_i - \mathrm{median}_j s_j|$. The *margin over local spacing*, which divides the margin by the mean of the five gaps below the runner-up. The *local outlier factor* of the top candidate with five neighbours [6]. The *density at the top with a list-scaled bandwidth*: as above, with the bandwidth given by Silverman's rule of thumb applied to the other candidates.

Fitted temperature. The *softmax probability* p_1 of the top candidate under $p_i \propto \exp(s_i/T)$ and the *negative entropy* of p. The temperature T minimises the negative log-likelihood of the true identity's position on held-out queries [15].

4.2 Singleton-mode gating and the critical bandwidth

Let $\hat{f}_h(s) = \sum_i \exp\{-(s - s_i)^2/2h^2\}$ be the Gaussian kernel density estimate of the list with bandwidth h. Mean shift iterates $y \leftarrow \sum_i w_i s_i / \sum_i w_i$ with $w_i = \exp\{-(y - s_i)^2/2h^2\}$ from each s_i; the iteration converges to a mode of $\hat{f}_h$, and candidates that reach the same mode form one cluster. The gate at bandwidth h *accepts* if the cluster of the highest mode has exactly one member, and then proposes that member, which is the top-ranked candidate.

For a fixed list, acceptance depends on h. When h is small every candidate is its own mode and the gate accepts; when h is large the whole list is one mode and it abstains. We define the *critical bandwidth*

$$h^\star = \sup\{h : \text{ the gate accepts at bandwidth } h\}$$

and use it as the confidence score of the gate. If acceptance is monotone in h, the gate at bandwidth h and the rule $h^\star > h$ are the same decision, so sweeping a threshold on $h^\star$ sweeps the bandwidth. Monotonicity is not guaranteed, because a runner-up can leave the top candidate's basin as h grows; Section 6.2 reports that this happens in under 0.1% of lists. The definition converts a rule with a hyperparameter into a score, which is what allows a like-for-like comparison with the margin and every other score in this paper at any operating point.

We consider two parameterisations. With a *fixed* bandwidth, h is in raw similarity units and is the same for all lists. With a *scaled* bandwidth, $h = \kappa\, \hat{\sigma}_{\mathrm{MAD}}$ for a dimensionless multiplier κ, and the score is the critical multiplier $\kappa^\star$. The scaled form is invariant to affine changes of the similarity axis; the fixed form is not.

4.3 Relation to the margin

Proposition 1. *For a list of two candidates, the gate with a Gaussian kernel and bandwidth h accepts if and only if $s_{(1)} - s_{(2)} > 2h$. Equivalently, $h^\star = (s_{(1)} - s_{(2)})/2$.*

Proof. The estimate $\hat{f}_h$ is an equal mixture of two Gaussians with standard deviation h, which is bimodal if and only if the means differ by more than $2h$ [11, 4]. If it is bimodal, the antimode is the midpoint by symmetry, each candidate lies on its own side of it, and mean shift takes each to a

different mode, so the highest mode is a singleton. If it is unimodal, both candidates converge to the single mode, whose cluster has two members. □

With more candidates there is no closed form, but the direction of the effect is clear. Candidates below the runner-up add mass to the lower mode. A kernel placed on a slope of that mass needs a larger separation to remain a mode of its own, so the margin required for acceptance rises above $2h$ as the list becomes denser just below the runner-up. Singleton-mode gating in one dimension is therefore a margin test with a threshold that depends on the local density. Proposition 1 and this argument give no reason to expect that the dependence improves decisions, and they do not exclude it. Section 6.2 measures both the size of the dependence and its effect.

4.4 Several channels

With k channels each candidate is a point $\boldsymbol{x}_i \in \mathbb{R}^k$. Early fusion collapses the channels into one score before any decision is made, and a weighted sum can then prefer a candidate that one channel rates highly while another channel prefers a different candidate. The multi-channel rule keeps the channels as coordinates until the decision.

Bandwidth. Channels are not commensurable in general: similarities under two encoders, or an image similarity and a text similarity, differ in location and spread. We therefore use a diagonal bandwidth matrix $\boldsymbol{H} = \mathrm{diag}(h_1^2, \ldots, h_k^2)$ and the kernel $K_{\boldsymbol{H}}(\boldsymbol{u}) = \exp(-\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{H}^{-1}\boldsymbol{u})$, with two choices of the diagonal. The *scaled* choice sets $h_c = \kappa\,\hat{\sigma}_c$, where $\hat{\sigma}_c$ is the MAD scale of channel c over the candidates of the list; a single multiplier κ then has the same meaning in every channel and every list. The *raw* choice sets $h_c = h$ for all channels and is appropriate only when the channels share a scale. Mean shift with $\boldsymbol{H}$ is ordinary mean shift with unit bandwidth on the rescaled points $(x_{i1}/h_1, \ldots, x_{ik}/h_k)$.

Decision rule. Let the modes have centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$. Mode a is *dominated* if some other centre is at least as large in every channel and larger in at least one, and *undominated* otherwise. The gate accepts if and only if exactly one mode is undominated and its cluster has one member, which it proposes. It abstains for one of two reasons: several modes are undominated (*conflict*: the channels disagree about which group of candidates is best), or the single undominated mode has several members (*shared mode*: the channels agree on a group but cannot separate a candidate within it). For $k = 1$ the rule reduces to that of Section 4.2. The critical multiplier $\kappa^\star$ is defined as before and serves as the confidence score. In several dimensions acceptance is often not monotone in κ, because increasing κ first merges dominated competitors, which removes conflicts, and only later merges the top candidate into a group; we therefore also report the rule at fixed values of κ.

How fast does abstention grow with the number of channels? Two reference calculations bound the answer. Their proofs are short; the second is due to Bentley et al. [5].

Proposition 2. *Suppose the N score vectors of a list are exchangeable and their k coordinates are mutually independent and continuous.* (i) *The probability that some candidate dominates all others is N^{1-k}.* (ii) *The expected number of undominated candidates is $E_k(N) = \sum_{i=1}^{N} E_{k-1}(i)/i$ with $E_1 \equiv 1$; in particular $E_2(N) = 1 + \frac{1}{2} + \cdots + \frac{1}{N}$ and $E_k(N) \sim (\ln N)^{k-1}/(k-1)!$.*

Proof of (i). A candidate dominates all others exactly when it is the maximum in every channel. The positions of the k channel maxima are independent and uniform on the N candidates, so they coincide with probability $N \cdot N^{-k}$. □

Table 1: Collections. "Shared" counts images that have a near-exact copy filed under a different identity; they are removed before evaluation. Top-1 and in-top-20 refer to the primary channel on present lists.

Collection	Images	Identities	Shared	Queries	Query ident.	Top-1 (%)	In top-20 (%)
S (public benchmark)	60,502	11,316	71	59,716	11,062	49.8	70.3
A (private catalog)	301	41	46	254	40	59.8	92.1
B (private catalog)	514	181	350	135	56	9.6	66.7

For $N = 20$ this gives a single dominant candidate with probability 5%, 0.25% and 0.0125% for $k = 2, 3, 4$, and 3.6, 7.3 and 11.0 undominated candidates on average. Independent channels are a pessimistic reference: agreement is then pure coincidence, and a rule that needs agreement almost never accepts. If instead each channel ranks the true identity first with probability p_c and the channels err independently, all channels agree on the true identity with probability $\prod_c p_c$. Real channels are positively dependent, so observed agreement lies above these references, and the amount by which agreement on present lists exceeds agreement on absent lists is precisely the identity evidence that a multi-channel gate can use. Section 6.5 reports the observed values. Smoothing with $\kappa > 0$ trades conflict for shared modes, since merging removes competing modes but also absorbs the top candidate.

Alternatives. We compare the Pareto rule with four other ways of using the same channels. *Fuse, then margin*: channels are robustly standardised within the list (median and MAD), averaged, and the margin of the fused score is thresholded; the proposal is the fused top candidate. *Unanimous margin*: the gate requires the same top candidate in every channel and thresholds the smallest standardised margin across channels. *Per-channel mean shift, intersected*: the single-channel gate with a scaled bandwidth is applied to each channel separately, and the gate accepts when all channels accept the same candidate; the confidence is the smallest critical multiplier. This addresses the option of clustering each channel independently before combining the evidence. *Learned gate*: described next.

4.5 Learned reference

To indicate how much identity evidence the lists contain beyond any single hand-written score, we train a gradient-boosted tree classifier to predict whether the top-ranked candidate is correct. In one dimension its inputs are the ten highest raw scores and seven of the scores of Section 4.1. In several dimensions its inputs are the five highest standardised scores in each channel, the standardised lead of the top candidate over the best other candidate in each channel, and the three highest raw primary scores. It is trained with two-fold cross-fitting over disjoint sets of query identities within one collection. It is a reference for what labels can buy in the best case, and it is not a fair comparison for the transfer experiments, in which no labels from the target are available.

5 Experimental Setup

Collections. The main benchmark, S, is the test split of Stanford Online Products [18]: 60,502 photographs of 11,316 products in twelve coarse categories, with about five photographs per product. It is public, about two orders of magnitude larger than the private catalogs, and its categories allow a leave-one-category-out analysis. Two private catalogs, A and B, serve as small stress tests from a different source; they are described only by the statistics in Table 1, and no identifying content is

used or released. Catalog A has 301 images of 41 identities and Catalog B has 514 images of 181 identities. An identity is a product in all three collections.

Channels. All encoders are frozen and none was trained on these collections. The primary channel E_1 is cosine similarity under a publicly released SigLIP ViT-B/16 image–text encoder [25] fine-tuned on product images. E_2 is cosine similarity under DINOv2 ViT-B/14 [19], a self-supervised encoder with no text supervision. C is the Bhattacharyya coefficient between HSV colour histograms ($12 \times 4 \times 4$ bins). On the private catalogs a text channel T is available: the cosine similarity between the query *image* embedding and the embedding of the candidate's catalog title under the text tower of the E_1 model. T never uses text associated with the query, so it cannot leak the query's label. To check that conclusions do not depend on the primary encoder, Appendix G repeats the single-channel analysis with E_2 as the primary channel.

Protocol. Evaluation is leave-one-image-out, and every eligible image is a query. For the *present* list the query image is withheld from the gallery, together with any near-exact copy of it under the same identity, and the remaining images of its identity stay. For the *absent* list the query's identity is deleted from the candidate list, which is what retrieval would return had the identity never been indexed; any commitment on an absent list is an error. Lists contain the $N = 20$ best identities under the primary channel unless stated otherwise; Appendix E varies N from 10 to 200. On S this gives 59,716 queries from 11,062 identities and twice as many lists. The true identity is ranked first in 49.8% of present lists and appears among the twenty candidates in 70.3%; the remaining present lists are cases in which the item is catalogued but not retrieved, and any commitment on them is also an error.

Duplicate images. An image that appears, as a near-exact copy, under two identities cannot be assigned to either. We flag an image as shared if some image of another identity has cosine similarity at least 0.99 under E_1 and at least 0.98 under E_2, and remove shared images from both queries and gallery. This removes 71 images from S (0.1%), 46 from A (15.3%) and 350 from B (68.1%). The last figure already indicates that in Catalog B an image often does not determine an identity. Appendix H repeats the private-catalog analysis without this step.

Mean shift. Fixed bandwidths range over $h \in \{0.002 \cdot 2^{j/2}\}_{j=0}^{14}$ (from 0.002 to 0.256 in similarity units) and multipliers over $\kappa \in \{0.125 \cdot 2^{j/2}\}_{j=0}^{14}$ (from 0.125 to 16). The critical bandwidth is the largest accepting grid value, refined by eight bisection steps inside the bracketing grid cell, and is set to zero for lists on which the gate never accepts; on S this is the case for 30.4% of lists, all of which lie outside the operating range of interest. Appendix B gives the convergence settings and a check of the iterative implementation against a direct evaluation of the density estimate.

Statistics. Confidence intervals are 95% percentile intervals from 1,000 bootstrap replicates that resample query *identities*, so that the several queries of one identity stay together. Differences from the margin are computed within each replicate, which gives paired intervals. The softmax temperature, the bandwidth of the fixed-bandwidth density score and the learned gate are cross-fitted on two halves of the queries with disjoint identities; the density bandwidth selected was $h = 0.01$ on both halves of S. No other quantity is tuned. All numbers in the text and tables are generated from the result files by the same script that draws the figures.

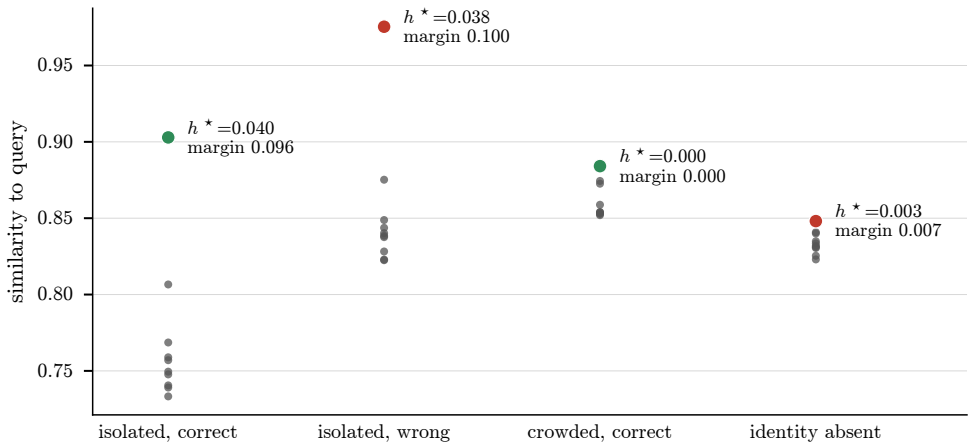


Figure 2: Four lists from S (ten highest scores each), selected by fixed rules and not by hand: the median list by critical bandwidth among present lists in the top decile of $h^\star$ with a correct and with a wrong top candidate, the median among correct lists in the bottom quartile, and the median absent list. The large marker is the top candidate, green if it is the queried identity. The second list has the highest top score of the four and a clear margin, and is wrong; 4.5% of present lists in the top decile of $h^\star$ are of this kind.

Table 2: Single-channel gates on S ($N = 20$, 59,716 queries, each evaluated with its identity present and absent). All gates propose the top-ranked candidate and differ in the confidence score. Brackets give 95% bootstrap intervals; those of AURC have a half-width of at most 0.6 points and are omitted. The last column is the paired difference in DIR at 1% FAR from the margin, in percentage points.

	Mixed		Present: risk (%) at coverage		Open set: DIR (%) at FAR		ΔDIR at 1%
Score	AUROC	AURC ↓	10% ↓	25% ↓	1% ↑	10% ↑	vs. margin
Absolute							
Top score $s_{(1)}$	0.712	59.0	13.5 [12.3, 14.9]	24.4 [23.4, 25.6]	6.5 [5.9, 7.0]	19.7 [19.1, 20.3]	−3.9 [−4.9, −2.7]
Raw scale							
Margin $s_{(1)} - s_{(2)}$	0.866	49.5	4.0 [3.5, 4.6]	6.1 [5.6, 6.7]	10.3 [9.3, 11.3]	33.5 [32.8, 34.2]	—
Distance ratio	0.878	48.3	3.2 [2.6, 3.8]	5.0 [4.5, 5.6]	14.6 [13.0, 16.2]	35.5 [34.9, 36.2]	+4.3 [+3.2, +5.4]
Density at top, fixed h	0.878	49.1	4.1 [3.5, 4.6]	6.2 [5.6, 6.7]	10.4 [9.3, 11.3]	33.6 [32.9, 34.3]	+0.0 [+0.0, +0.1]
Mean shift, fixed h	0.868	49.4	4.5 [3.9, 5.2]	6.4 [5.9, 6.9]	9.6 [8.8, 10.6]	33.7 [33.0, 34.4]	−0.7 [−1.1, −0.3]
List-scaled							
Cohort z-score	0.850	50.7	4.4 [3.9, 5.0]	7.8 [7.3, 8.5]	9.3 [8.5, 10.2]	30.6 [30.0, 31.3]	−1.1 [−1.8, −0.3]
Margin / MAD	0.842	51.1	4.8 [4.1, 5.4]	8.3 [7.7, 9.0]	9.1 [8.3, 9.7]	30.3 [29.6, 31.0]	−1.3 [−2.0, −0.5]
Margin / local spacing	0.793	54.4	7.3 [6.5, 8.0]	13.4 [12.7, 14.2]	6.7 [6.2, 7.2]	24.3 [23.7, 25.0]	−3.7 [−4.5, −2.8]
Local outlier factor	0.814	53.4	7.5 [6.8, 8.3]	12.8 [12.0, 13.5]	6.8 [6.3, 7.2]	26.5 [25.9, 27.1]	−3.6 [−4.4, −2.7]
Density at top, list-scaled h	0.850	50.9	4.9 [4.3, 5.5]	8.1 [7.5, 8.7]	9.0 [8.2, 9.7]	30.5 [29.9, 31.2]	−1.4 [−2.1, −0.7]
Mean shift, scaled h	0.843	51.2	5.2 [4.6, 5.9]	8.7 [8.1, 9.3]	8.4 [7.6, 9.2]	30.2 [29.5, 30.9]	−2.0 [−2.7, −1.1]
Fitted temperature							
Softmax probability	0.876	49.3	4.4 [3.8, 5.0]	6.6 [6.1, 7.2]	9.8 [9.0, 10.9]	33.3 [32.6, 34.0]	−0.5 [−0.8, −0.1]
Negative entropy	0.868	49.7	4.4 [3.8, 5.1]	7.0 [6.4, 7.6]	9.8 [9.0, 10.8]	32.6 [31.8, 33.2]	−0.6 [−0.9, −0.1]
Supervised							
Learned gate	0.891	47.8	2.8 [2.3, 3.4]	4.8 [4.3, 5.3]	15.0 [13.5, 16.3]	36.0 [35.4, 36.7]	+4.6 [+3.6, +5.6]

6 Results

Sections 6.1–6.5 use the public benchmark S; Section 6.6 reports the private catalogs.

6.1 Separability against absolute similarity

Table 2 and Figure 3 compare the fourteen scores. The top score is the weakest by a wide margin. At a 1% false-accept rate it identifies 6.5% of present queries and the margin identifies 10.3%; at

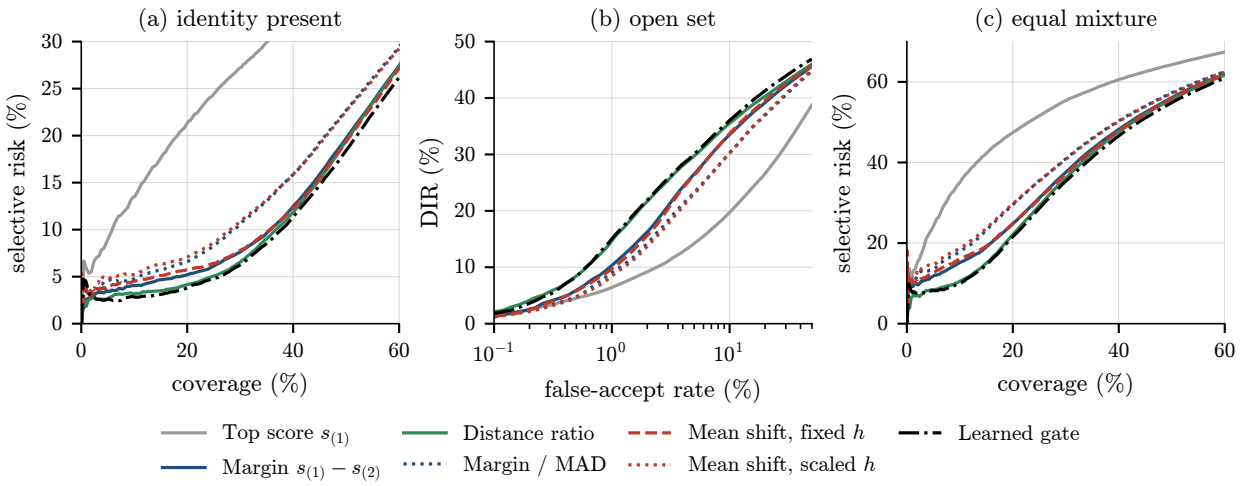


Figure 3: Single-channel gates on S. (a) Risk–coverage on present lists. (b) Detection and identification rate against false-accept rate. (c) Risk–coverage on the equal mixture of present and absent lists.

10% the figures are 19.7% and 33.5%. On present lists, the most confident tenth by top score contains 13.5% errors, against 4.0% for the margin. The mixed-protocol AUROC rises from 0.712 to between 0.793 and 0.878 for the hand-written separability scores. Figure 2 shows why: correct top candidates occur at ordinary similarity levels when they stand apart, and lists with uniformly high similarity are common among absent queries. On this benchmark the hypothesis of Section 1 holds: separation from the alternatives carries much more identity evidence than the level of the best score.

Among separability scores the ordering is as follows. The distance ratio is the best closed-form score, with 14.6% DIR at 1% FAR, a gain of +4.3 points over the margin (interval +3.2 to +5.4). It multiplies the margin by a factor that grows as the runner-up approaches the query, so it uses the level of the scores as well as their separation. Its advantage depends on the encoder: with E_2 as the primary channel it is better than the margin at 10% FAR and not at 1% (Appendix G). The learned gate is slightly better still (15.0%), which puts what a richer function of one channel's list can add at about five points. The raw-scale margin, fixed-bandwidth density and fixed-bandwidth mean shift are within one percentage point of each other on every risk and DIR figure. The softmax probability with a fitted temperature is equivalent to the margin for practical purposes. All list-scaled scores are worse than their raw-scale counterparts within this collection: dividing by the spread of the list discards the information that a gap of a given absolute size is more convincing than the same gap in a list that happens to be tightly packed. The local outlier factor and the margin over local spacing, which normalise by the spacing of only a few neighbours, are the weakest separability scores and at 1% FAR are no better than the top score.

Even the best gate commits rarely at strict operating points: at 1% FAR about one present query in seven is identified, and at 10% about one in three, while the true identity is ranked first in half of them. This is a statement about how much identity evidence a single frozen embedding provides on this benchmark, and it sets the scale against which the differences between gates should be read.

6.2 Mean shift against the margin and against density at the top

The critical bandwidth is nearly a function of the margin: on S their rank correlation is 0.983, and the ratio of margin to $h^\star$ has median 2.46 and lies between about 2 and 3.3 for almost all lists

(Appendix C, Proposition 1). Paired differences versus the margin are small and of mixed sign (-0.71 points of DIR at 1% FAR; $+0.21$ at 10%). Where the two disagree at a matched acceptance rate, the lists accepted only by mean shift are the less precise ones. A fixed-bandwidth density score that never iterates to modes matches the margin as well (DIR at 1% FAR 10.4%). Mean shift does not outperform either construction; the useful signal is the separation of the top candidate. Full equivalence tables and the margin–$h^\star$ scatter are in Appendix C.

6.3 Bandwidth sensitivity

Modes exist only relative to a bandwidth, so we check whether that bandwidth is a free structural parameter or an ordinary threshold. It is the latter (Appendix D). For all but 0.07% of lists the decision is a single step in h; each $\sqrt{2}$ step on the grid changes between 3.1% and 10.6% of decisions; the critical-bandwidth distributions of correct and incorrect lists overlap; and there is no plateau of stable acceptance. Precision among accepted present lists stays high while acceptance falls, exactly as with a margin threshold that is raised. At matched thresholds the margin rule tracks the gate across the grid, and a $\sqrt{2}$ mis-set of the operating point moves FAR from 1% to 2.5% for mean shift against 2.3% for the margin. The fixed-bandwidth gate is also insensitive to list length from $N = 10$ to $N = 200$ (Appendix E); longer lists do not give it a cluster-mass advantage over the margin.

6.4 Transfer of an operating point

An appeal of a structural rule is the prospect that it needs no calibration of the similarity scale. For the fixed-bandwidth gate this does not hold: h is a length on the similarity axis, and the proportion of lists with a critical bandwidth above h depends on the encoder and the collection as much as the proportion with a margin above δ does. Table 3 measures this by setting each threshold for 5% FAR on a source and applying it elsewhere.

Across the categories of S, all separability scores transfer acceptably. The realised FAR stays within a factor of two of the target in 11 of twelve categories for the fixed-bandwidth gate (range 1.2–8.9%) and in all twelve for the scaled gate (2.5–7.6%), while the top score ranges from 0.9% to 12.3% (Figure 4).

Across collections and encoders the raw-scale scores fail. The margin threshold and the bandwidth that give 5% FAR on S give 18% and 20% on Catalog A, 23% and 29% on Catalog B, and 26.8% and 29.3% when the encoder is exchanged; in the opposite direction of the encoder exchange they give 0.4% and 0.2%. The list-scaled scores are much closer to the target: 5.3% and 4.6% for the scaled gate in the two directions of the encoder exchange, 3% on Catalog B and 1% on Catalog A. The transfer is not exact. On Catalog A the scaled thresholds are conservative, with correspondingly low DIR, and a threshold set on the 254 lists of Catalog A and applied to S gives 13.5% for the scaled gate, against 1.2% for the fixed-bandwidth gate, where 5% was intended. The margin over MAD, the cohort z-score and the list-scaled density behave as the scaled gate does throughout, so this property belongs to list scaling and not to mean shift.

The two parameterisations thus trade against each other. Within a collection, raw-scale scores are more accurate by about three points of DIR at 10% FAR (Table 2). A list-scaled operating point survives a change of encoder and is at least of the right order on a new collection. Neither removes the need to check the false-accept rate on data from the deployment setting.

Table 3: Realised operating point when the threshold is set for 5% FAR on a source and applied unchanged to a target (all values in %). Left: each of the twelve categories of S in turn is the target and the other eleven are the source. Centre: S is the source and a private catalog the target. Right: the threshold from E_1 lists is applied to lists built with E_2 as the primary channel on S. The upper block contains scores in raw similarity units and the lower block list-scaled scores.

	Held-out category of S (12 folds)			S → A		S → B		Encoder swap
Score	FAR median	FAR range	Within 2×	FAR	DIR	FAR	DIR	FAR
Top score $s_{(1)}$	5.3	0.9–12.3	9/12	1	6	15	1	3.9
Margin $s_{(1)} - s_{(2)}$	4.7	1.4–8.7	11/12	18	35	23	1	26.8
Distance ratio	5.0	3.1–8.6	12/12	8	30	33	2	12.1
Mean shift, fixed h	4.7	1.2–8.9	11/12	20	39	29	2	29.3
Cohort z-score	5.0	2.7–7.5	12/12	3	5	4	0	5.2
Margin / MAD	5.3	2.5–7.5	12/12	2	4	3	0	5.3
Density at top, list-scaled h	5.0	2.7–7.4	12/12	3	6	4	0	5.4
Mean shift, scaled h	5.2	2.5–7.6	12/12	1	4	3	0	5.3

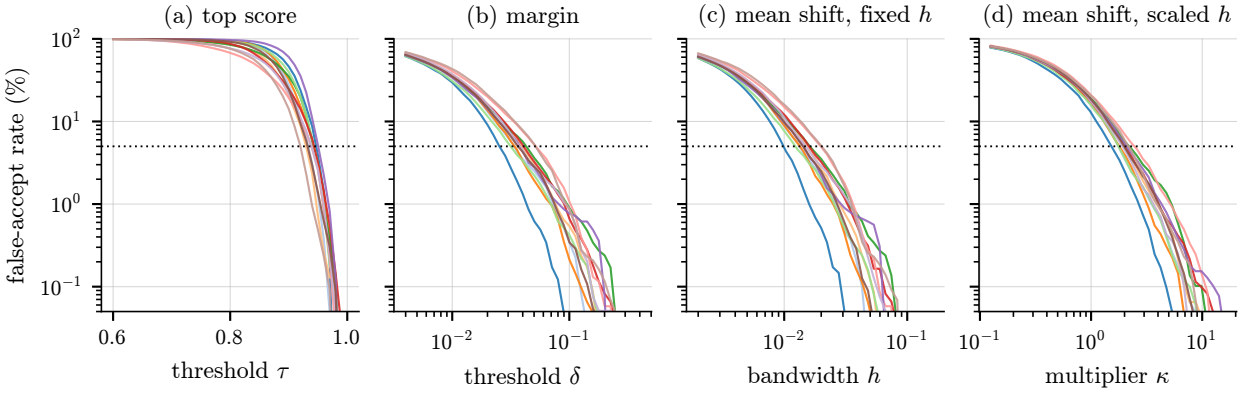


Figure 4: False-accept rate as a function of the gate parameter, one curve for each of the twelve categories of S. The dotted line marks 5%. A parameter value transfers between categories to the extent that the curves coincide where they cross the line.

6.5 Several channels

Agreement between channels shrinks as channels are added, more slowly than under independence and much faster on absent lists (Appendix F). With three channels a single dominant candidate is 9.4 times as frequent on present lists as on absent ones, and when it occurs on a present list it is the queried identity 96% of the time. Smoothing with a Pareto mean-shift rule trades conflict for shared modes as the bandwidth grows (Appendix F).

Table 4 and Figure 5 compare the rules. With two channels (E_1 and E_2), neither form of the Pareto rule improves on the single-channel margin, and unanimity gains little: two image encoders are too strongly dependent. With the colour channel added, the multi-channel rules improve and separate by operating point. At 1% FAR the unanimity rules are best among those without training (19.8% and 19.9%), close to the learned gate (21.0%); the scaled Pareto rule is not distinguishable from the single-channel margin. At 10% FAR the unanimity rules have exhausted the lists they can accept, and the Pareto rule with one raw bandwidth reaches 46.6%, equal to the learned gate. No untrained rule is best at both operating points; the learned gate is best or equal at both. The benefit depends on the added channel: colour alone ranks the queried identity first in 57.5% of

Table 4: Multi-channel gates on S. "Ever acc.": share of present lists (%) that the rule can accept at any threshold. DIR in % with 95% intervals. A rule that cannot accept enough absent lists to reach the stated FAR is reported at its most permissive setting.

	E_1+E_2 ($k = 2$)			E_1+E_2+C ($k = 3$)		
Rule	Ever acc.	DIR@1%	DIR@10%	Ever acc.	DIR@1%	DIR@10%
Margin, primary channel only	100.0	10.3 [9.3, 11.3]	33.5 [32.8, 34.2]	100.0	10.3 [9.3, 11.3]	33.5 [32.8, 34.2]
Fuse channels, then margin	100.0	9.6 [8.8, 10.5]	34.6 [33.8, 35.3]	100.0	11.8 [10.6, 12.9]	41.2 [40.3, 42.0]
Unanimous margin	50.4	10.1 [9.1, 11.1]	36.4 [35.7, 37.2]	39.4	19.8 [18.2, 21.4]	37.7 [37.1, 38.3]
Per-channel mean shift, intersected	42.8	9.7 [8.8, 10.6]	36.5 [35.6, 37.2]	31.7	19.9 [18.0, 21.3]	30.8 [30.2, 31.4]
Pareto mean shift, scaled h	67.2	8.7 [8.0, 9.4]	33.3 [32.5, 34.0]	58.1	10.9 [10.0, 11.7]	40.8 [39.9, 41.6]
Pareto mean shift, one raw h	64.1	7.2 [6.5, 7.8]	33.7 [32.8, 34.6]	54.4	14.1 [12.7, 15.1]	46.6 [45.8, 47.4]
Learned gate, all channels	100.0	15.9 [14.5, 17.4]	40.3 [39.5, 41.0]	100.0	21.0 [19.1, 23.0]	46.6 [45.9, 47.2]

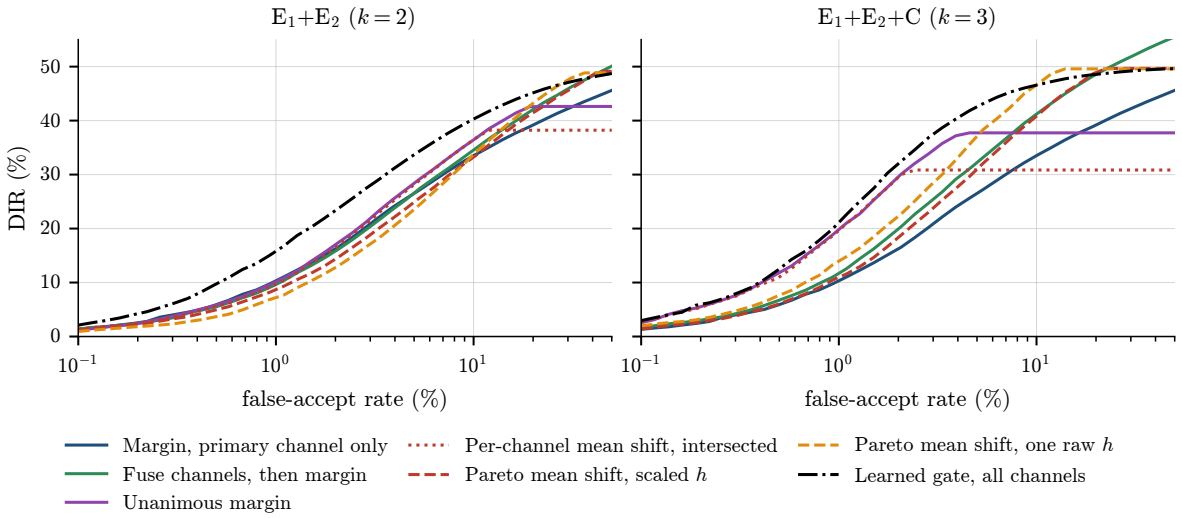


Figure 5: Multi-channel gates on S: detection and identification rate against false-accept rate with two and three channels. Curves of rules that require agreement end where the rule has accepted every list it can accept.

present lists on this benchmark, more than either encoder, because photographs of one product often share a setting—a property that should not be expected of colour in general.

6.6 Private catalogs

The private catalogs are too small to rank methods; we use them to check whether S is contradicted.

Catalog A. The pattern of S recurs with wide intervals (Table 5): top-score AUROC 0.66, other scores between 0.75 and 0.78, critical bandwidth tracking the margin (rank correlation 0.994). Thresholds taken from S at 5% FAR (Appendix H) accept a substantial share of present lists at a realised FAR well above the target for raw-scale scores and almost nothing for list-scaled ones.

Catalog B. Every gate fails. Top-1 accuracy is only 9.6%; AUROCs lie between 0.34 and 0.62; at 10% FAR no score identifies more than 0.7% of present queries. Of the 14 present lists with the largest margins, 14 propose a wrong identity, typically at high similarity to the query under both encoders while falling short of our duplicate thresholds—the pattern expected when near-copies are

Table 5: Single-channel gates on the private catalogs ($N = 20$; A: 254 queries from 40 identities; B: 135 queries from 56 identities). Risk on present lists at 25% coverage and DIR at 10% FAR, in %, with 95% intervals.

	Catalog A			Catalog B		
Score	AUROC	Risk@25% ↓	DIR@10% ↑	AUROC	Risk@25% ↓	DIR@10% ↑
Top score $s_{(1)}$	0.663	19 [10, 34]	24 [16, 33]	0.474	94 [78, 100]	0 [0, 4]
Margin $s_{(1)} - s_{(2)}$	0.769	10 [2, 19]	24 [17, 33]	0.530	94 [84, 100]	0 [0, 0]
Distance ratio	0.772	8 [1, 18]	32 [21, 40]	0.484	97 [88, 100]	0 [0, 0]
Density at top, fixed h	0.779	8 [2, 16]	25 [18, 35]	0.518	88 [77, 100]	0 [0, 1]
Mean shift, fixed h	0.777	6 [2, 14]	26 [19, 36]	0.546	91 [80, 100]	0 [0, 0]
Cohort z-score	0.760	9 [3, 17]	23 [14, 32]	0.600	85 [73, 97]	0 [0, 2]
Margin / MAD	0.760	13 [5, 21]	21 [13, 34]	0.587	91 [79, 100]	1 [0, 2]
Density at top, list-scaled h	0.769	11 [4, 19]	23 [14, 33]	0.587	88 [76, 97]	1 [0, 2]
Mean shift, scaled h	0.768	13 [5, 20]	22 [14, 35]	0.611	89 [75, 100]	1 [0, 3]
Learned gate	0.751	16 [6, 28]	27 [19, 35]	0.545	91 [74, 100]	1 [0, 2]

filed under different identities. A separability gate measures whether the scores isolate a candidate; it cannot detect that the isolation is unrelated to identity. Without duplicate removal, and with multi-channel rules including text, the failure persists (Appendix H).

7 Discussion

What the evidence supports. On the benchmark, and on the private catalog in which images identify items, a commitment to an identity is far better justified by the separation of the top candidate than by its similarity. A gate built on separation adds a decision layer to an existing retrieval system at negligible cost, since it reads at most a few hundred numbers that retrieval has already produced. Among ways to measure that separation, mean shift is not privileged: in one dimension it is a margin test with a mildly density-dependent threshold, and that dependence is neutral to slightly harmful; a fixed-bandwidth density score is equivalent to both; where they differ at strict operating points the margin is the more precise.

Practical guidance. For a single channel, the distance ratio and the margin are the best closed-form scores in our experiments and cost one line of code each; the ratio was clearly ahead with the primary encoder and ahead only at looser operating points with the second. When labelled queries from the target collection exist, a small learned gate on the list is better and should be preferred; a threshold with a statistical guarantee can then be obtained by conformal risk control [1]. When a threshold must be carried to a collection or encoder without labels, list-scaled scores are the safer starting point, at a cost of a few points of accuracy, and the realised false-accept rate still has to be verified. With several channels that carry different information, keeping them separate until the decision is worthwhile: requiring unanimity is effective at strict false-accept rates and a Pareto mode criterion at moderate ones. Before any gate is deployed, the collection should be checked for images that occur under several identities, since no score-based gate can compensate for labels that the channels cannot see.

A note on clustering the scores. The multi-channel Pareto mean-shift rule is the one setting in which clustering did something the closed-form alternatives did not: with three channels at a 10% false-accept rate it matched a learned gate without training, and each abstention is auditable

as a conflict between channels or as a shared mode. Whether that advantage persists on other benchmarks and channel combinations is an open question; it is not a reason to prefer mean shift in one dimension.

8 Limitations

The main results come from one public benchmark of product photographs and one query modality. Text retrieval, cross-modal retrieval and collections with very different near-duplicate structure were not evaluated, and the private catalogs are too small to do more than fail to contradict the benchmark.

The encoders are frozen and general-purpose. The primary encoder was fine-tuned on a product domain that differs from much of the benchmark; the second encoder gives the same conclusions (Appendix G), but encoders trained for instance-level matching may produce score lists with different structure.

Absent queries are simulated by deleting the query's identity from its list. Unknown items in deployment may differ systematically from known ones, for example by belonging to product lines that are entirely missing, which would change false-accept rates.

Leave-one-image-out evaluation leaves other photographs of the queried product in the gallery, often taken in the same session. This is favourable to all gates and particularly to the colour channel.

Identity labels are taken as given apart from the duplicate filter, whose thresholds were set by inspection and not tuned; Catalog B shows that labels can be inconsistent with visual identity in ways such a filter does not remove.

The learned gate is cross-fitted within a collection and was not tested under transfer. Multi-channel operating points were not tested under transfer either. The softmax temperature and the density bandwidth are selected on held-out identities of the same collection, which favours those baselines slightly relative to a deployment without labels.

Critical bandwidths are resolved on a grid with a lower limit of 0.002; lists below it are tied at zero, which affects only coverage levels above about 70% on the mixed protocol.

We provide no guarantee on the false-accept rate of any gate. Lists with several correct identities are outside the scope of the task as defined.

9 Conclusion

Selective identity retrieval is the decision, after ranking, whether to commit to one candidate identity or to abstain. Under a controlled evaluation the right evidence for that decision is the separation of the best candidate from the alternatives, not its absolute similarity, and a gate built on separation can sit behind any ranker. Among single-channel measurements of separation, a distance ratio or a small learned gate is more accurate than the top-two margin; mean shift reduces in practice to a margin test whose bandwidth is just another operating point. List-scaled variants give up some accuracy for an operating point that survives a change of encoder. With several channels, postponing fusion is useful when the channels carry different information, and the choice between unanimity, a Pareto mode criterion and a learned gate depends on the false-accept rate that the application tolerates. The decision layer between ranking and action deserves evaluation in its own right, with absent queries, with accepted counts reported next to precision, and with collections on which it fails.

A Algorithm

```
gate(X, h):                        # X: N x k scores, h: k bandwidths
    Z      = X / h                 # rescale each channel by its bandwidth
    Y      = mean_shift(Z)         # Gaussian kernel, unit bandwidth
    modes  = group(Y)              # candidates with a common limit
    front  = [m for m in modes if centre(m) is not dominated]
    if len(front) == 1 and size(front[0]) == 1:
        return member(front[0])  # commit
    return ABSTAIN

critical_bandwidth(X, scale):    # scale = 1 (raw) or per-channel MAD
    g* = largest g on the grid with gate(X, g * scale) != ABSTAIN
    refine g* by bisection inside its grid cell
    return g*                      # confidence score; 0 if never accepted
```

For $k = 1$, "not dominated" means "highest centre". A deployed gate commits when the critical bandwidth is at least a threshold set for a target false-accept rate on held-out lists; evaluating the gate once at that bandwidth gives the same decision in one dimension, apart from the rare non-monotone lists of Table 6.

B Implementation Details

Mean shift. Fixed-point iterations run until the largest move is below 10^{-5} bandwidths, with at most 2,000 iterations. Converged points are grouped by single linkage at 0.1 bandwidths. As a check that is independent of the iteration, we evaluated the kernel density estimate on a grid of 4,096 points and tested directly whether a local minimum separates the two highest scores. On 10,335 decisions (the first 300 present lists of each collection at all fifteen fixed bandwidths) the two procedures disagreed 0 times.

Scores. The local outlier factor uses five neighbours on the score line. The list-scaled density uses the bandwidth $0.9 \min(\hat{\sigma}, \mathrm{IQR}/1.349)(N-1)^{-1/5}$ computed from the candidates other than the top one. The fixed density bandwidth is selected from $\{0.005, 0.01, 0.02, 0.04, 0.08\}$ by AUROC on the other fold. The softmax temperature is selected from forty values between 0.002 and 0.5. The learned gates are histogram-based gradient-boosted trees with 200 iterations, learning rate 0.05, at most 15 leaves and at least 20 samples per leaf.

Metrics. Ties in a confidence score are treated as one threshold step; risk at a coverage that falls inside a tie group is interpolated, which corresponds to random tie-breaking. DIR at a given FAR uses the most permissive threshold whose FAR does not exceed the target.

C Mean Shift against the Margin and Density at the Top

The critical bandwidth is close to a function of the margin (Table 6, Figure 6). On S their rank correlation is 0.983. The ratio of margin to $h^\star$ has median 2.46 and lies between 2.00 and 3.24 for 98% of lists, so the gate at bandwidth h behaves as a margin test with a threshold between $2h$ and about $3.3h$, as Section 4.3 anticipated. The lower end is the two-candidate bound of Proposition 1;

Table 6: Critical bandwidth against the corresponding margin statistic. "Reference / critical bandwidth" is the ratio of the margin (or margin over MAD) to $h^\star$ (or $\kappa^\star$) over lists with a positive critical bandwidth. "Non-monotone" is the share of lists on which acceptance switches from abstain to accept somewhere along the bandwidth grid. "Max. decision disagreement" is the largest share of lists, over the grid, on which the gate and a margin rule accepting the same number of lists differ.

Collection	Pair	Rank corr.	Reference / critical bandwidth median	1st–99th pct.	Ratio < 2 (%)	Non-monotone (%)	Max. decision disagreement (%)
S	fixed h vs. margin	0.983	2.46	2.00–3.24	0.40	0.07	2.7
	scaled h vs. margin/MAD	0.995	2.40	2.00–3.23	0.34	0.08	2.8
A	fixed h vs. margin	0.994	2.32	2.00–3.05	0.00	0.00	3.9
	scaled h vs. margin/MAD	0.989	2.38	2.00–3.05	0.00	0.00	3.9
B	fixed h vs. margin	0.992	2.28	2.00–3.04	0.00	0.00	5.9
	scaled h vs. margin/MAD	0.988	2.35	2.00–3.06	0.96	0.00	3.7

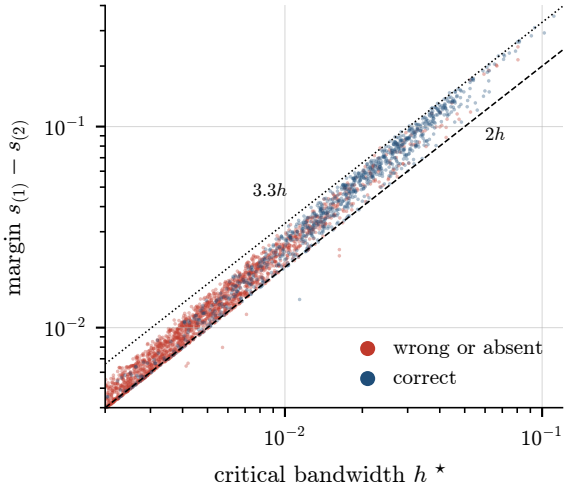


Figure 6: Margin against critical bandwidth for a random sample of 4,000 lists from S (present and absent). The dashed line is the two-candidate identity of Proposition 1, margin $= 2h^\star$. Lists lie in a band between $2h^\star$ and about $3.3h^\star$; the position within the band reflects the density of candidates just below the runner-up. Correct decisions (blue) concentrate at large values of both quantities and are not separated from errors by position within the band.

it is not a strict bound for longer lists, and 0.40% of lists fall below it. Acceptance is non-monotone in h for 0.07% of lists. The scaled bandwidth stands in the same relation to the margin over MAD (rank correlation 0.995), and the same holds on both private catalogs and with the second encoder as the primary channel (Appendix G).

The density-dependent part of the threshold does not help. The paired differences between fixed-bandwidth mean shift and the margin are small and of mixed sign: -0.71 points of DIR at 1% FAR (interval -1.08 to -0.25), $+0.21$ points at 10% FAR ($+0.07$ to $+0.37$), and $+0.47$ points of risk at 10% coverage ($+0.19$ to $+0.75$). The benchmark is large enough to resolve these differences, and they amount to a slight disadvantage for mean shift at strict operating points and a slight advantage at loose ones. A direct comparison shows the same pattern. When a margin rule is set to accept the same number of lists as the gate, the two differ on at most 2.7% of lists. At $h = 0.0226$ each rule accepts 795 lists that the other rejects; 81.8% of the lists accepted only by the margin are correct, against 68.9% of those accepted only by mean shift. At $h = 0.008$, a much more permissive setting, the order reverses (21.4% against 30.9%). Where precision matters, then, the lists that mean shift accepts and the margin rejects, which are lists with a sparse neighbourhood below the runner-up, are the less reliable ones.

Density at the top. A simpler hypothesis than "the top candidate is a singleton mode" is that the top candidate lies in a region to which the other candidates contribute little density. The fixed-bandwidth density score tests it directly, without iterating to modes. It performs as the margin

does (DIR at 1% FAR 10.4%, difference +0.01 points with interval −0.05 to +0.09; AUROC 0.878), because at the selected bandwidth the kernel sum is dominated by the runner-up. With a list-scaled bandwidth the density score matches the scaled mean-shift gate and the margin over MAD (30.5%, 30.2% and 30.3% DIR at 10% FAR). Mean shift does not outperform a simpler density measure or the margin in either parameterisation. The useful signal is the separation of the top candidate, and the three constructions are different readings of it.

D Bandwidth Sensitivity

Table 7: The mean-shift gate on S at each grid bandwidth, with counts. "Accepted" and "Precision" refer to the 59,716 present lists, FAR to the absent lists (all in %). "Changed" is the share of all lists whose decision differs from the decision at the previous (smaller) bandwidth. The right half shows the margin rule at the threshold that corresponds to h through the median ratio of Table 6.

	Mean shift at bandwidth h					Margin at $\delta = 2.46\,h$			
h	Accepted	(%)	Precision	FAR	Changed	Accepted	(%)	Precision	FAR
0.0028	41,604	69.7	65.8	50.9	9.3	40,448	67.7	66.9	48.0
0.0040	36,836	61.7	71.6	38.4	10.3	35,873	60.1	72.5	36.4
0.0057	31,641	53.0	78.3	25.9	10.6	31,040	52.0	78.7	25.0
0.0080	26,447	44.3	85.0	15.6	9.5	26,233	43.9	84.9	15.5
0.0113	21,360	35.8	89.9	8.6	7.8	21,607	36.2	89.7	8.9
0.0160	16,037	26.9	93.1	4.4	6.6	16,748	28.0	92.9	4.8
0.0226	10,354	17.3	94.5	2.1	5.9	11,746	19.7	94.9	2.5
0.0320	5,319	8.9	95.6	0.83	4.9	6,954	11.6	95.8	1.1
0.0453	1,955	3.3	96.5	0.24	3.1	3,172	5.3	96.6	0.42
0.0640	461	0.8	95.9	0.06	1.3	1,090	1.8	97.2	0.13
0.0905	52	0.1	100.0	0.00	0.4	256	0.4	98.0	0.03
0.1280	0	0.0	—	0.00	0.0	19	0.0	100.0	0.00

Table 8: Distribution of the critical bandwidth on S by outcome. A list is accepted at every bandwidth below its critical value.

		Percentile of the critical bandwidth				
Bandwidth	Lists	5th	25th	50th	75th	95th
$h^\star$	Present, top candidate correct	0.0000	0.0082	0.0161	0.0266	0.0486
	Present, top candidate wrong	0.0000	0.0000	0.0026	0.0053	0.0136
	Identity absent	0.0000	0.0000	0.0029	0.0058	0.0150
$\kappa^\star$	Present, top candidate correct	0.21	0.91	1.80	3.17	6.55
	Present, top candidate wrong	0.00	0.16	0.37	0.75	1.86
	Identity absent	0.00	0.17	0.39	0.78	1.97

Because modes are defined only relative to a bandwidth, a singleton top mode could be an artefact of the bandwidth chosen. We address four questions.

How often does a decision change as the bandwidth varies? For all but 0.07% of lists the decision is a single step: accept below $h^\star$, abstain above. The number of decisions that change between two bandwidths is therefore the number of lists whose critical bandwidth lies between them. Each step between adjacent grid values, which differ by a factor $\sqrt{2}$, changes between 3.1% and 10.6% of all

Table 9: Sensitivity of the operating point on S. The threshold θ gives 1% FAR; the remaining columns show FAR and DIR when it is too small or too large by a factor $\sqrt{2}$, and the share of all lists whose decision changes between those two settings.

		FAR (%) at threshold			DIR (%) at threshold			Decisions
Score	θ	$\theta/\sqrt{2}$	θ	$\sqrt{2}\,\theta$	$\theta/\sqrt{2}$	θ	$\sqrt{2}\,\theta$	flipped (%)
Margin $s_{(1)} - s_{(2)}$	0.082	2.31	1.00	0.38	17.8	10.3	4.7	7.9
Mean shift, fixed h	0.030	2.46	1.00	0.30	17.8	9.6	3.8	8.5
Margin / MAD	10.8	2.23	1.00	0.47	15.3	9.1	4.6	6.6
Mean shift, scaled h	4.0	2.31	1.00	0.40	15.1	8.4	3.8	7.0

decisions for grid values from 0.004 to 0.045 (Table 7).

Is there an interval of bandwidths over which decisions are stable? There is none. The critical bandwidths of correct lists and of lists that should be rejected overlap (Table 8): the median for correct present lists is 0.0161 and a quarter of them fall below 0.0082, while 5% of absent lists exceed 0.0150. The number of accepted lists falls smoothly across the whole grid, with no plateau. What is stable is the precision among accepted present lists, which stays between 93.1% and 96.5% for $0.016 \le h \le 0.045$ while the accepted share falls from 26.9% to 3.3% and FAR from 4.4% to 0.24%. The bandwidth is an operating-point parameter of the same kind as a threshold. It is possible to obtain an arbitrary acceptance rate by choosing it, exactly as with a margin threshold, and the remedy is the same: it must be set on held-out data for a target false-accept rate and reported with the number of accepted lists.

Does the gate outperform a margin test across the range? It does not (right half of Table 7). At matched thresholds the precision of the margin rule is within one and a half points of that of the gate at every bandwidth with more than a hundred acceptances, and the margin rule retains somewhat more lists at large bandwidths. The gate is also slightly more sensitive to a misplaced operating point (Table 9): a bandwidth that is too small by a factor $\sqrt{2}$ raises FAR from 1% to 2.5%, against 2.3% for the margin threshold.

Does it work with longer lists? Appendix E addresses this.

E List Size

Table 10: Effect of the number of candidates N on S, on a fixed random subsample of 2,000 queries (AUROC on the mixed protocol; DIR in % at 10% FAR).

	AUROC at list size N					DIR (%) at 10% FAR at list size N				
Score	10	20	50	100	200	10	20	50	100	200
Margin $s_{(1)} - s_{(2)}$	0.875	0.875	0.875	0.875	0.875	33.6	33.6	33.6	33.6	33.6
Mean shift, fixed h	0.877	0.877	0.877	0.877	0.877	34.1	34.1	34.1	34.1	34.1
Cohort z-score	0.827	0.856	0.870	0.866	0.855	27.9	31.9	32.5	32.7	32.0
Margin / MAD	0.819	0.847	0.856	0.857	0.857	28.4	31.6	32.2	32.2	32.2
Density at top, list-scaled h	0.832	0.858	0.868	0.868	0.868	27.6	31.7	32.6	32.6	32.4
Mean shift, scaled h	0.818	0.849	0.859	0.857	0.858	27.8	31.4	31.8	32.7	32.0

Table 10 varies the list size from 10 to 200 on a fixed subsample of 2,000 queries. The margin does not depend on N by construction. The fixed-bandwidth gate does not depend on it in practice

either: at the bandwidths of interest, candidates far below the runner-up contribute negligible kernel weight, and its figures are unchanged to three decimals. The gate therefore works with 50, 100 or 200 candidates, but longer lists do not give it the additional cluster mass that might have distinguished it from the margin. List-scaled scores depend on N through the scale estimate. They improve from $N = 10$ to $N = 50$ (scaled gate: AUROC 0.818 to 0.859) and are flat beyond, and they remain below the raw-scale scores at every list size.

F Multi-Channel Agreement and Pareto Bandwidth

Table 11: Agreement between channels before any smoothing ($N = 20$). Left: share of lists (%) in which one candidate dominates all others, that is, every channel ranks the same candidate first; "& correct" adds the requirement that this candidate is the queried identity; $\prod_c p_c$ and N^{1-k} are the independence references of Section 4.4. Right: mean number of Pareto-undominated candidates, with the independence reference $E_k(N)$ of Proposition 2.

			One candidate dominates all others (%)					Undominated candidates (mean)		
Coll.	Channels	k	Present	& correct	$\prod_c p_c$	Absent	N^{1-k}	Present	Absent	Indep.
S	E_1	1	100.0	49.8	49.8	100.0	100.0	1.00	1.00	1.00
	E_1+E_2	2	50.4	42.6	26.5	19.3	5.00	2.02	2.79	3.60
	E_1+E_2+C	3	39.4	37.7	15.2	4.2	0.25	3.46	5.81	7.27
A	E_1	1	100.0	59.8	59.8	100.0	100.0	1.00	1.00	1.00
	E_1+E_2	2	56.3	46.1	35.3	30.7	5.00	1.75	2.46	3.60
	E_1+T	2	36.6	31.1	24.5	15.7	5.00	2.20	2.89	3.60
	E_1+E_2+C	3	40.2	35.4	18.5	15.0	0.25	2.76	4.33	7.27
	E_1+E_2+T	3	26.4	24.4	14.5	7.9	0.25	3.36	4.84	7.27
	E_1+E_2+C+T	4	20.1	19.7	7.6	4.3	0.01	4.61	7.06	11.03
B	E_1	1	100.0	9.6	9.6	100.0	100.0	1.00	1.00	1.00
	E_1+E_2	2	40.0	6.7	1.4	38.5	5.00	2.01	2.13	3.60
	E_1+T	2	16.3	3.7	1.2	14.8	5.00	2.98	3.02	3.60
	E_1+E_2+C	3	21.5	6.7	0.3	20.7	0.25	3.19	3.33	7.27
	E_1+E_2+T	3	7.4	2.2	0.2	5.9	0.25	4.57	4.75	7.27
	E_1+E_2+C+T	4	4.4	2.2	0.0	3.0	0.01	5.99	6.30	11.03

Table 12: The Pareto mean-shift rule on S at fixed multipliers κ of the per-channel MAD scale. "Conflict": more than one undominated mode. "Shared mode": one undominated mode with several members. Accept, conflict and shared mode sum to 100%.

		Identity present (%)				Identity absent (%)	
Channels	κ	Accept	Precision	Conflict	Shared mode	Accept	Conflict
E_1+E_2 ($k = 2$)	0.5	54.0	80.9	37.1	8.9	25.3	59.6
	1	49.9	82.9	11.2	38.9	20.1	16.1
	2	30.4	91.4	1.8	67.8	6.3	2.1
	4	10.9	95.2	0.2	88.8	1.2	0.2
E_1+E_2+C ($k = 3$)	0.5	42.0	94.5	57.2	0.8	6.1	93.1
	1	49.4	89.4	28.5	22.0	13.4	46.0
	2	33.3	93.7	5.2	61.5	5.0	7.2
	4	11.7	96.1	0.8	87.5	1.0	0.8

The discussion in Section 6.5 summarises these tables. Smoothing is not monotone in κ: for

20.0% of lists with two channels and 17.7% with three the decision switches from abstain to accept somewhere along the grid.

G Second Encoder as the Primary Channel

Table 13: Single-channel gates on S with E_2 as the primary channel (top-1 accuracy 55.0%). Layout as in Table 2.

	Mixed		Present: risk (%) at coverage		Open set: DIR (%) at FAR		ΔDIR at 1%
Score	AUROC	AURC ↓	10% ↓	25% ↓	1% ↑	10% ↑	vs. margin
Absolute							
Top score $s_{(1)}$	0.728	56.5	17.1 [15.8, 18.8]	22.7 [21.7, 23.8]	4.6 [4.1, 5.1]	20.2 [19.5, 20.8]	−4.7 [−5.8, −3.7]
Raw scale							
Margin $s_{(1)} - s_{(2)}$	0.844	47.8	3.8 [3.2, 4.4]	7.0 [6.5, 7.6]	9.3 [8.4, 10.3]	33.2 [32.4, 34.0]	—
Distance ratio	0.867	46.6	6.4 [5.5, 7.5]	7.2 [6.6, 7.8]	8.9 [7.7, 10.2]	37.4 [36.6, 38.2]	−0.4 [−1.5, +0.9]
Density at top, fixed h	0.850	47.6	3.8 [3.2, 4.4]	7.0 [6.5, 7.6]	9.3 [8.4, 10.3]	33.2 [32.4, 34.0]	+0.0 [+0.0, +0.0]
Mean shift, fixed h	0.845	47.8	4.0 [3.4, 4.7]	7.4 [6.9, 8.1]	9.4 [8.4, 10.2]	33.2 [32.3, 34.0]	+0.1 [−0.4, +0.4]
List-scaled							
Cohort z-score	0.828	48.9	3.9 [3.4, 4.5]	7.7 [7.1, 8.4]	7.4 [6.6, 8.4]	30.7 [29.9, 31.4]	−1.8 [−2.5, −1.1]
Margin / MAD	0.823	49.3	4.9 [4.3, 5.7]	8.4 [7.8, 9.0]	7.1 [6.5, 7.9]	30.1 [29.3, 30.8]	−2.1 [−2.8, −1.5]
Margin / local spacing	0.778	52.3	6.4 [5.7, 7.1]	11.3 [10.6, 12.1]	5.8 [5.3, 6.4]	25.2 [24.4, 25.9]	−3.5 [−4.2, −2.7]
Local outlier factor	0.796	51.5	7.3 [6.5, 8.1]	12.0 [11.2, 12.7]	5.8 [5.2, 6.4]	26.5 [25.8, 27.2]	−3.5 [−4.3, −2.8]
Density at top, list-scaled h	0.832	48.9	4.5 [3.9, 5.3]	8.1 [7.5, 8.7]	7.3 [6.6, 8.0]	30.5 [29.6, 31.2]	−2.0 [−2.7, −1.4]
Mean shift, scaled h	0.823	49.4	5.7 [5.0, 6.4]	8.8 [8.2, 9.5]	6.7 [6.0, 7.4]	29.9 [29.0, 30.6]	−2.6 [−3.3, −2.0]
Fitted temperature							
Softmax probability	0.843	48.0	3.9 [3.3, 4.5]	7.7 [7.1, 8.4]	9.4 [8.4, 10.3]	32.5 [31.7, 33.2]	+0.2 [−0.3, +0.4]
Negative entropy	0.825	48.9	4.0 [3.4, 4.6]	8.1 [7.5, 8.8]	9.4 [8.4, 10.2]	31.5 [30.7, 32.2]	+0.1 [−0.3, +0.4]
Supervised							
Learned gate	0.878	45.9	4.2 [3.6, 4.8]	5.9 [5.3, 6.5]	9.9 [8.7, 11.0]	38.0 [37.2, 38.8]	+0.6 [−0.3, +1.4]

Table 13 repeats Table 2 with lists retrieved and scored by E_2. The conclusions are unchanged. Separability scores exceed the top score by a wide margin, the critical bandwidth tracks the margin (rank correlation 0.996, median ratio 2.38), fixed-bandwidth mean shift and the margin are indistinguishable (difference in DIR at 1% FAR +0.1 points, interval −0.4 to +0.4), and list-scaled scores are lower. The distance ratio behaves differently with this encoder than with E_1. It is better than the margin at 10% FAR (37.4%, a difference of +4.2 points), no better at 1% FAR (−0.4 points), and worse on present lists at 10% coverage (risk 6.4%, +2.5 points). Its ranking among closed-form scores is therefore encoder-dependent.

H Private Catalogs: Additional Results

Thresholds transferred from S. Raw-scale thresholds from S overshoot the 5% FAR target on both catalogs; list-scaled ones are conservative.

Without duplicate removal. When shared images are kept (Table 15), top-1 accuracy falls to 49.3% on Catalog A and 2.6% on Catalog B, because the nearest gallery image of many queries is a copy of the query filed under another identity. On Catalog B the top score then has AUROC 0.15: a higher similarity makes a correct decision less likely. The margin-type scores retain some ranking ability on the mixed protocol, because they still separate present from absent lists, but none reaches a usable risk. These numbers show how strongly an evaluation of identity gates depends on the handling of duplicated content.

Table 14: Private catalogs with thresholds taken from S at 5% FAR, as counts. A decision on a present list is correct if the proposed identity is the queried one.

		Identity present			Identity absent	
Catalog	Score (threshold from S)	Accepted	Correct	Precision (%)	Accepted	FAR (%)
A ($n = 254$)	Top score $s_{(1)}$	17	14	82	3	1
	Margin $s_{(1)} - s_{(2)}$	101	88	87	46	18
	Distance ratio	85	77	91	21	8
	Mean shift, fixed h	113	98	87	50	20
	Margin / MAD	13	10	77	5	2
	Mean shift, scaled h	10	9	90	3	1
B ($n = 135$)	Top score $s_{(1)}$	21	2	10	20	15
	Margin $s_{(1)} - s_{(2)}$	30	2	7	31	23
	Distance ratio	42	3	7	44	33
	Mean shift, fixed h	37	3	8	39	29
	Margin / MAD	3	0	0	4	3
	Mean shift, scaled h	3	0	0	4	3

Table 15: Private catalogs without removal of images shared across identities. Layout as in Table 5.

Catalog	Score	AUROC	Risk@25% ↓	DIR@10% ↑
A (300 queries, top-1 49.3%)	Top score $s_{(1)}$	0.524	72 [43, 95]	0 [0, 6]
	Margin $s_{(1)} - s_{(2)}$	0.720	33 [19, 47]	3 [1, 19]
	Mean shift, fixed h	0.715	35 [18, 51]	4 [1, 20]
	Mean shift, scaled h	0.720	31 [18, 43]	9 [5, 16]
B (506 queries, top-1 2.6%)	Top score $s_{(1)}$	0.152	100 [100, 100]	0 [0, 0]
	Margin $s_{(1)} - s_{(2)}$	0.705	94 [91, 98]	0 [0, 0]
	Mean shift, fixed h	0.688	95 [90, 98]	0 [0, 0]
	Mean shift, scaled h	0.688	96 [91, 99]	0 [0, 0]

Table 16: Multi-channel gates on Catalog A. Layout as in Table 4.

	E_1+T ($k = 2$)			E_1+E_2+C+T ($k = 4$)		
Rule	Ever acc.	DIR@1%	DIR@10%	Ever acc.	DIR@1%	DIR@10%
Margin, primary channel only	100	0 [0, 16]	24 [17, 33]	100	0 [0, 16]	24 [17, 33]
Fuse channels, then margin	100	1 [0, 13]	24 [13, 37]	100	2 [0, 8]	26 [16, 39]
Unanimous margin	37	2 [0, 15]	25 [15, 36]	20	7 [0, 21]	19 [12, 26]
Per-channel mean shift, intersected	27	2 [0, 16]	24 [15, 34]	13	7 [0, 16]	11 [6, 17]
Pareto mean shift, scaled h	58	0 [0, 15]	25 [15, 34]	45	2 [0, 9]	29 [17, 42]
Pareto mean shift, one raw h	56	0 [0, 18]	27 [14, 38]	42	2 [0, 11]	31 [16, 43]
Learned gate, all channels	100	4 [0, 13]	27 [14, 35]	100	6 [0, 17]	34 [23, 43]

Several channels. Tables 16 and 17 report the multi-channel rules with the text channel T and with all four channels. On Catalog A the point estimates of the Pareto rule and the learned gate with four channels are above the single-channel margin at 10% FAR, and the intervals overlap it. On Catalog B all rules remain near zero. Table 11 includes the agreement statistics for both catalogs; on Catalog B channels agree on a single dominant candidate about as often on absent lists as on present ones, which is the multi-channel form of the failure described in Section 6.6.

Table 17: Multi-channel gates on Catalog B. Layout as in Table 4.

	E_1+T ($k = 2$)			E_1+E_2+C+T ($k = 4$)		
Rule	Ever acc.	DIR@1%	DIR@10%	Ever acc.	DIR@1%	DIR@10%
Margin, primary channel only	100	0 [0, 0]	0 [0, 0]	100	0 [0, 0]	0 [0, 0]
Fuse channels, then margin	100	0 [0, 0]	2 [0, 5]	100	0 [0, 1]	3 [0, 8]
Unanimous margin	16	2 [0, 6]	2 [0, 6]	4	1 [0, 5]	2 [0, 6]
Per-channel mean shift, intersected	4	2 [0, 6]	2 [0, 6]	1	0 [0, 0]	0 [0, 0]
Pareto mean shift, scaled h	35	0 [0, 1]	2 [0, 5]	18	0 [0, 1]	4 [0, 8]
Pareto mean shift, one raw h	31	0 [0, 0]	0 [0, 4]	19	0 [0, 2]	4 [0, 12]
Learned gate, all channels	100	0 [0, 0]	0 [0, 0]	100	0 [0, 0]	0 [0, 0]

I Reproducibility and Data Handling

The pipeline has five scripts: embedding (`embed_bundles.py`), list construction (`build_lists.py`), mean-shift sweeps (`run_ms.py`), analysis (`analyze.py`) and generation of tables, figures and in-text numbers (`make_artifacts.py`). The primary encoder checkpoint is the publicly available Marqo-FashionSigLIP (ViT-B/16); the second encoder is the `timm` model `vit_base_patch14_dinov2.lvd142m` at 224×224 input. The benchmark is the public test split of Stanford Online Products and can be re-embedded with the first script. For the private catalogs the embedding step writes only numerical arrays (embeddings, colour histograms, integer identity labels); file names, titles and all other identifying strings are discarded at that point and appear in no released artefact.

Relation to the first draft. A first draft of this work reported results on the two private catalogs only. Its evaluation used one query per identity, did not remove images shared across identities, used a text channel computed from the query's own catalog title, which reveals the label, and fixed bandwidths after looking at the evaluated lists. The present evaluation replaces it entirely, and the figures of the first draft should not be cited.

References

[1] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1–3):42–54, 2000.

[3] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

[4] J. Behboodian. On the modes of a mixture of two normal distributions. *Technometrics*, 12(1):131–139, 1970.

[5] J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson. On the average number of maxima in a set of vectors and applications. *Journal of the ACM*, 25(4):536–543, 1978.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000.

[7] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[8] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[9] G. V. Cormack, C. L. A. Clarke, and S. Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

[10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[11] I. Eisenberger. Genesis of bimodal distributions. *Technometrics*, 6(4):357–363, 1964.

[12] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[13] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[14] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[15] 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 (ICML)*, 2017.

[16] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[20] P. J. Phillips, P. Grother, and R. Micheals. Evaluation methods in face recognition. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*. Springer, 2005.

[21] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[22] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[23] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems*, 30(2), article 11, 2012.

[24] B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B*, 43(1):97–99, 1981.

[25] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.