Journal of Data Efficiency
Volume 1 · Number 1 · 2026

More Examples, More Competitors: The Reference-Photo Budget in Closed-Set Face Matching

George Wells[1] · Henry Collins[2] · Ryan King[3]

[1] Sweden · [2] Finland · [3] Ireland

Original publication: October 24, 2026 | Article 587

Abstract

Adding reference photographs helps a nearest-gallery matcher see more of each person, but also gives competing identities more opportunities to look similar. On one fixed Olivetti split, increasing the gallery from one to three and five photographs per identity raised correctness from 56.0% to 75.5% and 85.5%. The five-photo policy corrected sixty formerly wrong queries and made one formerly correct query wrong. This transparent pixel-based experiment measures the value of a reference budget within a known roster. It does not estimate the number of images needed by a learned identity-preserving generator.

Public-data research draft. Conditional benchmark analysis; final scientific and publication review is pending.

More Examples, More Competitors: The Reference-Photo Budget in Closed-Set Face Matching

> Public-data research draft. Conditional benchmark analysis; final scientific and publication review is pending.

Abstract

Adding reference photographs helps a nearest-gallery matcher see more of each person, but also gives competing identities more opportunities to look similar. On one fixed Olivetti split, increasing the gallery from one to three and five photographs per identity raised correctness from 56.0% to 75.5% and 85.5%. The five-photo policy corrected sixty formerly wrong queries and made one formerly correct query wrong. This transparent pixel-based experiment measures the value of a reference budget within a known roster. It does not estimate the number of images needed by a learned identity-preserving generator.

1. Reference data are part of the decision rule

A gallery is often described as supporting data for an identity matcher, while the representation is treated as the main methodological choice. For a nearest-example system, that division is incomplete. The stored examples define which comparisons are possible. A new photograph can place a person's appearance closer to a query, but adding photographs for every identity also changes the competing matches. The effect of a larger gallery is therefore an empirical property of a representation, a roster, and a particular set of queries.

The research question asks how much additional reference photographs change closed-set matching correctness when the image representation is held fixed. We compare one, three, and five gallery photographs per identity on the same two hundred queries. The Olivetti archive AT&T Laboratories Cambridge, 2026 makes the comparison small enough to inspect completely. Its historical setting is linked to the face-identification work of Samaria and Harter, 1994. We use the released images as a numerical benchmark, without acquiring new face data or interpreting its limited roster as a demographic sample.

Reference sufficiency is also relevant when identity information conditions generation. Meng et al., 2026 develops stacked multi-view identity mosaic injection for subject-preserving video generation. That method motivates attention to the information carried by reference views, but the present study does not run its generator or evaluate generated videos. Nearest-gallery correctness on existing photographs is a different endpoint. Our experiment isolates a simple prerequisite question: how strongly can an unchanged comparison rule depend on how many labeled examples are available for each known identity?

Two propositions guide the analysis. At the aggregate level, a larger gallery may improve coverage of within-person variation. At the individual-query level, improvement need not be monotonic because new competing examples are added at the same time. These propositions are compatible. A higher average accuracy can coexist with a small set of newly introduced errors. Reporting only the aggregate change would conceal that distinction and encourage an overly simple interpretation of additional reference data as universally beneficial.

2. Representation and reference coverage

Classical face-recognition research provides useful context for separating representations from exemplars. Eigenface recognition Turk and Pentland, 1991 and the comparison of eigenfaces with class-specific linear projection Belhumeur et al., 1997 examine learned coordinate systems. Our implementation fits neither. Each photograph is centered by its own mean and scaled to unit Euclidean norm. A query receives the identity label of the closest stored photograph under squared Euclidean distance. The analysis consequently concerns an explicit pixel-space baseline rather than a claim about the best available recognition architecture.

Modern face embeddings define a different comparison space. FaceNet Schroff et al., 2015, ArcFace Deng et al., 2018, and CosFace Wang et al., 2018 motivate distances learned with identity-related objectives. AdaFace Kim et al., 2022 and MagFace Meng et al., 2021 further connect recognition representations with image-quality considerations. We cite these approaches to delimit the missing experiment: a reference-budget curve in a learned embedding could have a different shape. No embedding from these systems is downloaded, and no reported accuracy in this article is attributed to one of them.

Dataset choice is equally consequential. MegaFace Kemelmacher-Shlizerman et al., 2015 emphasizes recognition at scale, VGGFace2 Cao et al., 2017 organizes variation across pose and age, and Labeled Faces in the Wild Huang et al., 2008 concerns less constrained imagery. The small Olivetti experiment cannot recreate these settings. It keeps a fixed forty-person roster, with each query identity already represented in every gallery. Increasing the reference count therefore expands coverage within known identities. It never adds an unknown person to the system or tests whether the system should refuse an unfamiliar face.

FairFace Kärkkäinen and Joo, 2019 illustrates the separate need for demographic coverage in face-data evaluation. The ordinal Olivetti labels used here provide no such annotation. We cannot estimate performance differences across demographic categories or infer fairness from a single overall correctness rate. The reference-budget result should remain conditional on the observed roster. A resource allocation claim for a broader population would require an appropriately documented dataset, rather than extrapolation from the number of available photographs alone.

3. Constructing the nested galleries

The local MATLAB archive contains 400 grayscale photographs, each represented by 4,096 pixels. Its checksum matches the maintained scikit-learn mirror, and the loading procedure follows that mirror's image orientation and forty consecutive groups of ten photographs. Values receive the same archive-wide positive affine scaling before reshaping. Such common scaling does not change Euclidean rankings after the subsequent per-image centering and unit normalization. The representation uses no learned parameters, source labels, or query outcomes.

A random generator initialized with seed 20261107 permutes the ten source rows separately within every identity. The first five rows form a reference pool and the last five form the fixed queries. The one-photo gallery retains the first pool item for each identity. The three-photo gallery retains its first three, and the five-photo gallery retains all five. Total gallery sizes are consequently forty, one hundred twenty, and two hundred. Each policy evaluates exactly the same two hundred query photographs, with no gallery-query row overlap.

The prefix construction matters because the galleries are nested. A gain cannot be attributed to replacing an unfavorable one-photo gallery with a disjoint larger set. All initially stored examples remain available. Nevertheless, the comparison is one realized random ordering, not an average over possible reference

selections. The source-row lists are retained so the observed changes can be traced to the actual photos admitted at each budget. No photo is chosen because it later performs well on the query set.

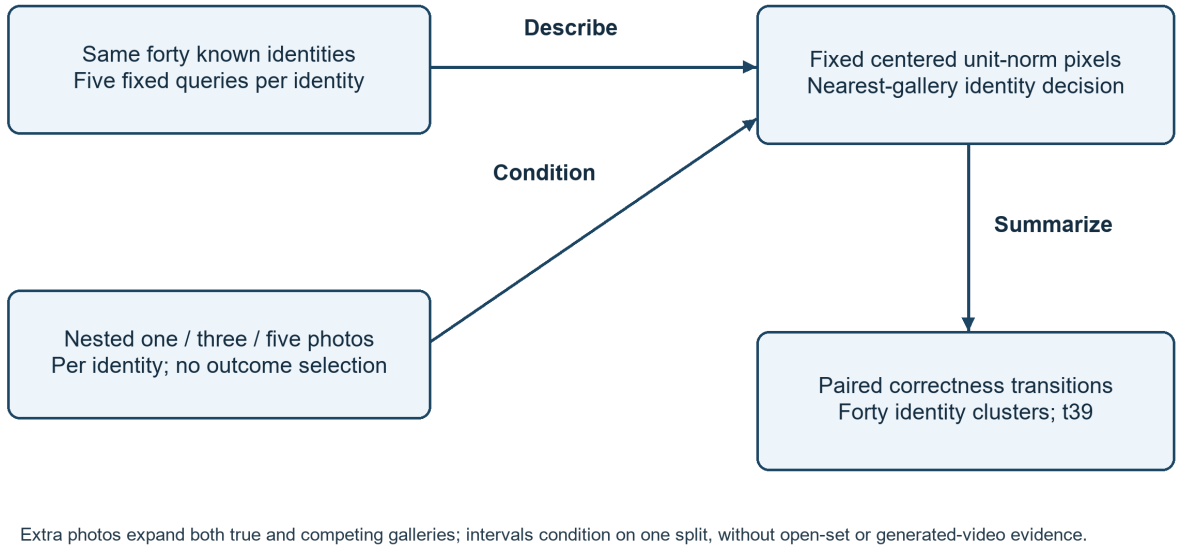


Figure 1. Nested labeled reference counts and common-query matching outcomes.

Figure 1 represents the reference budget as the changing component and the normalized pixel representation as fixed. Both the correct identity and thirty-nine competing identities receive additional stored examples. The output is a paired change in correctness for a shared query set. This diagram does not contain a training stage, learned gallery weights, or a generative identity-preservation endpoint. Those omitted mechanisms would require additional data and implementations before being placed on the empirical path.

4. Conditional comparison and uncertainty

The nearest gallery row is obtained by exhaustive squared-distance comparison, with the first minimum resolving numerical ties. Correctness equals one when that row's identity matches the released query label and zero otherwise. For each query, we subtract one-photo correctness from three-photo correctness and from five-photo correctness. Each difference takes the value negative one, zero, or positive one. Two intercept-only OLS regressions estimate their means, which are accuracy changes expressed as proportions.

Standard errors cluster on the forty query identities, with a small-sample correction and t reference using thirty-nine degrees of freedom. Robust covariance reasoning White, 1980 and the cluster-inference discussion of Cameron and Miller, 2015 motivate recording the grouping and its limitations. The intervals condition on the actual gallery selection, representation, and shared roster. They do not integrate uncertainty from obtaining another reference pool, changing the roster, or observing a different photographic collection. Shared fixed gallery competitors are part of the conditioning information.

The grouping also does not turn this split into independent-person external validation. Every query identity is known to the gallery by design. Structured validation guidance Roberts et al., 2017 helps distinguish such within-roster assessment from a test on previously unseen identities. The model-selection concerns discussed by Varma and Simon, 2006 explain why budgets and normalization should not be tuned against the same outcomes used to report gains. Here, the comparison uses three declared budgets and a single fixed representation; it offers no search for the most favorable threshold or feature setting.

Accuracy is the primary outcome because the task is a closed-set identity assignment. It is not a calibrated probability. Proper-scoring-rule theory Gneiting and Raftery, 2007 addresses a different predictive object when a system outputs probabilities or distributions. Similarly, a ROC curve Fawcett, 2006 would require a specified matching score and thresholded verification task. Our nearest-gallery decisions always assign one of the forty labels. No acceptance threshold, unknown-person rejection, or false-accept target is estimated from these accuracy contrasts.

5. Gains are substantial but not uniform

The one-photo gallery correctly identifies 112 of 200 queries, or 0.560. Three photos per identity raise that count to 151, or 0.755. Five raise it to 171, or 0.855. Relative to the one-photo policy, the three-photo gain is 0.195 with standard error 0.0354368 and 95% interval 0.1233222 to 0.2666778. The five-photo gain is 0.295 with standard error 0.0385557 and interval 0.2170138 to 0.3729862. These intervals describe conditional paired gains, not a universal sample-size requirement.

Table 1. Paired accuracy changes from expanding the reference gallery

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.1950 (0.0354)	[0.1233, 0.2667]	200
M1: fit	R-squared = 0.0000	-	-	200
M2	Mean outcome/contrast	0.2950 (0.0386)	[0.2170, 0.3730]	200
M2: fit	R-squared = 0.0000	-	-	200

Model labels: M1 = 3-versus-one gallery photographs per identity; M2 = 5-versus-one gallery photographs per identity.

Table 1 contains intercept-only OLS models, without covariate controls or fixed effects. N is two hundred shared queries. Standard errors cluster on forty query identities with a finite-sample correction, and intervals use t39. M1 compares three photographs with one per identity; M2 compares five with one. R-squared is zero for the mean-only model and does not measure recognition quality. Accuracy changes of 0.195 and 0.295 correspond to 19.5 and 29.5 percentage points.

The transition counts reveal what the averages omit. Moving from one to three photos corrects forty-two initially wrong queries, leaves forty-six wrong, preserves one hundred nine correct decisions, and introduces three errors among initially correct queries. Moving from one to five corrects sixty, leaves twenty-eight wrong, preserves one hundred eleven correct decisions, and introduces one error. Thus the net gains reflect many favorable changes with a smaller number of unfavorable changes. Additional examples do not preserve every previously correct nearest-neighbor assignment.

The five-photo accuracy equals the normalized-pixel result obtained from this same fixed five-gallery/five-query split in the companion representation analysis. That equality is an intentional consistency check on the shared data construction, not independent replication. The research contrast here changes the labeled reference count while freezing representation. It should be assessed as one question within the same finite archive, rather than combined with the companion study as though two independent face populations had been observed.

6. Interpreting the reference budget

The result supports a local coverage interpretation: the extra reference photographs create more opportunities for a query to find a nearby same-identity example. The newly introduced mistakes show the parallel competition effect. Because we add photographs to all identities together, the analysis does not identify the separate marginal contribution of true-person coverage and competitor expansion. A crossed experiment that expands one side while freezing the other would answer that additional mechanism question. We do not infer it from the aggregate gains.

The same distinction matters when discussing reference-conditioned generation. InstantID Wang et al., 2024, PhotoMaker Li et al., 2023a, and PuLID Guo et al., 2024 use identity information in generative customization, while ID-Animator He et al., 2024 and ConsisID Yuan et al., 2024 concern identity-preserving video generation. Their inputs and objectives differ from storing labeled photographs for nearest-gallery retrieval. Our result cannot tell a user how many reference views any of those systems needs. It establishes only that an explicit reference count can be a large source of variation in a transparent baseline.

The experiment also does not establish a cost-optimal gallery. Photograph acquisition, labeling expense, storage, privacy conditions, and query processing time are not measured. A larger finite gallery increases the number of available comparisons, but a measured wall-clock claim would require a separate implementation and timing protocol. Influence diagnostics Cook, 1977 could help examine whether particular identities dominate a mean gain; such a diagnostic is a follow-up rather than a reported analysis here. The retained query transitions provide a starting point without pretending that acquisition costs are known.

Nor should the three observed points be joined into a smooth learning law. They are evaluations of nested finite reference sets, and the experiment contains no budgets between or beyond those points. The ten-per-identity archive limits how many photographs can be allocated while preserving the present query set. Estimating a diminishing-return curve or claiming that five photographs are sufficient would add assumptions beyond the executed comparison. A future reference acquisition study could instead repeat gallery sampling, measure the actual acquisition burden, and retain an external query collection. That design would address the uncertainties that the present conditional contrasts deliberately leave open.

Selective classification Geifman and El-Yaniv, 2017 offers a possible future extension in which uncertain assignments are deferred. That would change the outcome from unconditional roster assignment to a coverage-error tradeoff. Dataset-shift research Ovadia et al., 2019 likewise motivates evaluating whether a fixed rule transports beyond the original collection. Neither extension is answered by the present closed-set accuracy table. The observed gains can justify examining reference sufficiency, while leaving abstention rules, open-set reliability, and domain transfer unresolved.

7. Data stewardship and conclusion

The numerical workflow retains the archive checksum, per-identity permutations, gallery source rows, query source rows, predicted labels, paired transitions, and structured regression results. AT&T Laboratories Cambridge receives source attribution. Public availability of a maintained MATLAB mirror is not treated as proof of unrestricted photo redistribution. Dataset documentation principles Gebru et al., 2021 favor recording that distinction explicitly, and FAIR data principles Wilkinson et al., 2016 support preserving the identifiers and transformations needed to reproduce the numerical analysis within applicable source terms.

On this fixed known-identity benchmark, increasing reference photographs produced sizeable average gains and a small number of new errors. The answer is therefore more informative than either “more data always helps” or a single maximum-gallery accuracy. Reference examples participate directly in the comparison rule. A report about identity matching should state how many examples each identity contributes, how they were selected, and whether the conclusion concerns known-roster retrieval or a different task. This study supplies those details for one conditional pixel-based comparison.

References

1. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.
2. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.
3. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.
4. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
5. Wang, Hao, et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.
6. Kemelmacher-Shlizerman, Ira, et al. "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale" *arXiv preprint arXiv:1512.00596*, 2015. https://arxiv.org/abs/1512.00596.
7. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
8. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
9. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
10. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.
11. Li, Zhen, et al. "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding." *arXiv preprint arXiv:2312.04461*, 2023. https://arxiv.org/abs/2312.04461.
12. Kärkkäinen, Kimmo and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age" *arXiv preprint arXiv:1908.04913*, 2019. https://arxiv.org/abs/1908.04913.
13. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.
14. Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.
15. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
16. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.
17. Wang, Qixun, et al. "InstantID: Zero-shot Identity-Preserving Generation in Seconds." *arXiv preprint arXiv:2401.07519*, 2024. https://arxiv.org/abs/2401.07519.
18. Yuan, Shenghai; Huang, Jinfa; He, Xianyi; Ge, Yunyuan; Shi, Yujun; Chen, Liuhan; Luo, Jiebo; Yuan, Li. (2024). Identity-Preserving Text-to-Video Generation by Frequency Decomposition. arXiv:2411.17440. https://arxiv.org/abs/2411.17440.
19. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
20. Samaria, F.S.; Harter, A.C.. (1994). Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. https://doi.org/10.1109/acv.1994.341300.

21. AT&T Laboratories Cambridge. (2026). Modified Olivetti faces: Sam Roweis MATLAB version [Data archive; mirror accessed 7 November 2026]. File identity and row-label layout verified against the scikit-learn maintainer loader. https://ndownloader.figshare.com/files/5976027.

22. Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

23. Guo, Zinan, et al. "PuLID: Pure and Lightning ID Customization via Contrastive Alignment." *arXiv preprint arXiv:2404.16022*, 2024. https://arxiv.org/abs/2404.16022.

24. He, Xuanhua; Liu, Quande; Qian, Shengju; Wang, Xin; Hu, Tao; Cao, Ke; Yan, Keyu; Zhang, Jie. (2024). ID-Animator: Zero-Shot Identity-Preserving Human Video Generation. arXiv:2404.15275. https://arxiv.org/abs/2404.15275.

25. Cao, Qiong, et al. "VGGFace2: A dataset for recognising faces across pose and age" *arXiv preprint arXiv:1710.08092*, 2017. https://arxiv.org/abs/1710.08092.

26. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

27. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

28. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

29. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

30. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.