Hard Negatives without Uncontrolled Identity Confounds

Timothy Marshall (NA), Wesley Foster (RU), Spencer Anderson (GQ)

Abstract

A difficult negative comparison can test whether an identity evaluator distinguishes similar subjects, but its difficulty can also come from mismatched cameras, clothing, or references. This article develops a protocol for constructing and auditing hard negatives in subject-preserving video studies. ARGUS motivates the need to separate reference resemblance from identity evidence. No negative examples are generated or scored here. The proposed design records how negatives are selected and uses matched comparisons to make a future discrimination result interpretable.

Difficulty needs a specified source

An identity benchmark can become harder by adding people who look similar, by reducing reference quality, or by introducing unfamiliar output conditions. These changes do not test the same capability. A negative selected because a particular embedding finds it close may expose that instrument's weakness, while a negative selected by human judgment can depend on presentation and familiarity. Calling both hard does not establish that they represent a common scientific challenge.

ARGUS conditions generated videos on structured references (Meng et al., 2026b). In this setting, an evaluator needs to distinguish preserving the referenced subject from reproducing incidental appearance that also occurs in another person's images. This article proposes a negative-comparison study rather than reporting recognition results. Its question is how to make difficult comparisons informative about subject discrimination without allowing uncontrolled acquisition differences or a convenient selection metric to determine the outcome in advance.

Building a comparison record before scoring

A proposed collection would state the source population, the rule used to select candidate negatives, and the information available during selection. A search based on one embedding should retain the search parameters and candidate pool. Human-selected cases should preserve the task instructions and uncertainty in their selection judgments. These records explain what kind of difficulty has been introduced and prevent the final comparison from appearing broader than its construction supports.

The study would then document observable nuisance conditions for positive and negative examples. Matching or stratification can address measured differences such as reference resolution or visible viewpoint, provided these variables are defined consistently. It cannot guarantee that all relevant appearance factors have been controlled. A useful design would include both the selected hard negatives and a separately sampled comparison group. Their results should remain distinguishable, since a deliberately difficult set does not estimate the frequency of confusions in ordinary use.

Selection and evaluation should use separate evidence where feasible. If the same instrument chooses the most confusing cases and supplies the only final outcome, the result describes an adversarial challenge to that instrument. It does not automatically establish how humans or another representation would behave. An independent assessment can reveal whether difficulty transfers. Any tuning of the selection rule belongs in a development stage, with the final negative set fixed before the principal method comparison.

Connecting difficult negatives to generated clips

Video comparison adds opportunities for confounding. A negative clip may differ in motion, lighting, or framing as well as subject. A proposed paired design would hold the requested scene and reference budget as similar as the workflow permits, while recording residual differences. It should retain clips that cannot be generated or assessed. Excluding those cases after seeing their difficulty can change the target population and obscure a method's practical limitations.

Outcomes should distinguish false acceptance of a negative from rejection of a true subject under an unfamiliar view. A threshold that reduces one error may increase the other. Reporting the operating point and uncertainty is therefore more useful than presenting a single favorable similarity difference. The relevant unit also matters: many frame comparisons involving one difficult identity do not provide the same evidence as independently selected people across a broader population.

Finally, the label negative requires adequate reference evidence. An ambiguous identity assignment should not become a hard negative merely because its score is inconvenient. A documented adjudication process can separate confirmed different-subject cases from uncertain provenance. The study can retain both categories for different purposes, but they should not share an unqualified error-rate denominator. This boundary keeps a demanding benchmark from becoming a collection of comparisons whose presumed truth cannot be independently checked.

Hard negatives selected independently of the outcome

For the proposed comparison, the key observation is same-subject and different-subject comparison. A studio shoot creates highly similar portraits of different people in matching clothes and lighting. This hypothetical case poses a question about false acceptance at the prespecified operating point. Record random different-subject comparison and nuisance-matched different-subject comparison separately. Selecting only cases that one embedding finds close makes the final outcome a test of selection by that embedding. Use a separate selection rule and preserve the candidate pool, including negatives not selected.

Choose difficult negatives using a rule that can be reconstructed before the final evaluator scores them. Clothing, camera and pose can be matched independently of facial similarity. A second set can deliberately use another embedding to find visually close candidates. Keep that selection mechanism visible, as it creates a different test population. Assess false acceptance at a fixed operating point on each subset, along with the observed score ranges. Avoid adjusting the threshold separately to make the final comparison look favorable.

An initial table can record the eligibility and observation status of same-subject and different-subject comparison, before any performance statistic is calculated. For the same-subject and different-subject comparison, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for same-subject and different-subject comparison. To evaluate false acceptance at the prespecified operating point, keep paired observations together during resampling, since independent resampling would lose their shared source of variation. When examining nuisance-matched different-subject comparison, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

Use the comparison with random different-subject comparison to identify which parts of the finding survive a reasonable change in the decision threshold. The evaluation of random different-subject comparison can distinguish an incorrect subject from a permitted change in expression, pose or scene. A

transfer study of false acceptance at the prespecified operating point would need fresh cases chosen independently of the observed failures.

Samples and evidence

Pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled. Related work tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). Earlier research defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). In practical terms, the size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description.

Earlier research provides face observations with substantial variation in pose and age (Cao et al., 2017). In an evaluation, an existing collection provides a starting point for auditing subgroup coverage while leaving the deployment distribution to be specified separately. Earlier research constructs a face attribute dataset with demographic balance (Kärkkäinen and Joo, 2019). Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow.

Measurement and error

One study uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). A cosine-margin baseline helps distinguish rejection due to the representation from rejection caused by the threshold. A methodological precedent uses a cosine margin for discriminative face recognition (Wang et al., 2018). During validation, angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture.

Related work combines detection and association considerations in tracking assessment (Luiten et al., 2020). During validation, a relevant contribution reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017). Report the detection and association components as well as the aggregate so that tracking continuity can be interpreted alongside object availability. Report coverage beside error rates. Rejecting difficult cases can improve conditional accuracy while leaving fewer cases with an answer.

A methodological precedent assesses probability calibration and studies temperature scaling (Guo et al., 2017). For a prospective comparison, one published study learns an embedding space for face verification and recognition (Schroff et al., 2015). After fitting a decision threshold, reliability curves on an independent split can show whether reported confidence still matches observed success frequency. Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity.

Related work learns aligned visual and language representations through contrastive training (Radford et al., 2021). Representation changes can alter relative prompt scores, so the encoder checkpoint and text preprocessing belong in the comparison record. For implementation, a whole-image embedding reflects scene description and background alongside subject appearance. A local check can distinguish these contributions. One study investigates scaling in contrastive language-image learning (Cherti et al., 2022).

One approach formulates classification decisions with explicit costs for different errors (Elkan, 2001). One approach contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated. In a controlled test, low

prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts.

Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential. Earlier research learns facial representations with a quality-related feature magnitude (Meng et al., 2021). In a controlled test, a quality-sensitive representation can be used to define inspectable strata before comparing acceptance rates across views with very different visibility. Earlier research constructs a full-reference metric with structure and texture components (Ding et al., 2020).

Using a representation learned without explicit class labels gives a useful comparison when a prompt-derived score may reward the wrong visual feature. For a prospective comparison, separating uncertainty in the measured inputs from variability in the underlying property is important when the same noisy proxy also defines the strata. Earlier research learns visual representations through self-supervision (Caron et al., 2021). Related work investigates statistical consequences of imperfectly measured variables (Carroll et al., 2006).

Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable. Feature magnitude and recognition confidence can help identify low-quality crops, but agreement with an independent quality annotation should be checked first. One published study compares choices of forecast error scale across series (Hyndman and Koehler, 2006). For a prospective comparison, a methodological precedent uses feature norms to adapt training margins to face image quality (Kim et al., 2022).

Systems and comparators

One approach adds spatial controls to pretrained text-to-image diffusion (Zhang et al., 2023). Separate facial structure and texture checks can identify whether a convincing local detail conceals changes in global appearance over time. One approach separates global facial structure and local detail in identity-conditioned video (Yuan et al., 2024). For a prospective comparison, fixed spatial controls offer a way to vary prompts while holding a geometric scaffold stable, or to vary the scaffold under common language.

Keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait. An appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. In practical terms, a relevant contribution adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). Earlier research uses appearance information in online association (Wojke et al., 2017).

Analysis and interpretation

Matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. In a controlled test, an existing method describes estimation and implementation of linear mixed-effects models (Bates et al., 2015). One published study reviews matching as a design step for comparing observational units (Stuart, 2010). Mixed-effects models account for repeated observations of the same unit, while the choice of grouping still needs a design rationale.

Linking evidence to the evaluation

Related work asks dataset creators to record collection, intended use and limitations (Gebru et al., 2021). Earlier research provides object annotations and standardized visual recognition tasks (Lin et al., 2014). A relevant contribution shows bias from reusing validation for selection and error estimation (Varma and Simon, 2006). Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. Reusing a validation split for repeated configuration selection can make the final error estimate optimistic; retain a separate assessment for the chosen rule. Documenting construction and intended use clarifies which exclusions occurred during collection, which arose from a model failure and which were imposed during analysis.

Conclusion

Hard negatives become scientifically useful when their difficulty and selection history are explicit. The proposed protocol preserves candidate pools, nuisance conditions, independent assessment, and uncertain labels. It reports no new discrimination performance. Its aim is a future evaluation in which a difficult comparison reveals a defined limitation of subject preservation or its measurement, rather than an unknown mixture of identity similarity, acquisition differences, and selection by the scoring instrument itself.

References

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.

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.

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.

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.

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.

Yuan, Shenghai, et al. "Identity-Preserving Text-to-Video Generation by Frequency Decomposition." *arXiv preprint arXiv:2411.17440*, 2024. https://arxiv.org/abs/2411.17440.

Luiten, Jonathon, et al.. "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking" *arXiv preprint arXiv:2009.07736*, 2020. https://arxiv.org/abs/2009.07736.

Gebru, Timnit, et al.. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

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.

Cherti, Mehdi, et al. "Reproducible scaling laws for contrastive language-image learning." *arXiv preprint arXiv:2212.07143*, 2022. https://arxiv.org/abs/2212.07143.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

Meng, Zijie, et al. "Argus: Stacked multi-view identity mosaic injection for subject-preserving video generation." *arXiv preprint arXiv:2606.11670* (2026).

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.

Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

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.

Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Wojke, Nicolai, et al. "Simple Online and Realtime Tracking with a Deep Association Metric." *arXiv preprint arXiv:1703.07402*, 2017. https://arxiv.org/abs/1703.07402.

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.

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.

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.

Ruiz, Nataniel, et al.. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

Zhang, Lvmin, et al. "Adding Conditional Control to Text-to-Image Diffusion Models." *arXiv preprint arXiv:2302.05543*, 2023. https://arxiv.org/abs/2302.05543.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.

Lin, Tsung-Yi, et al. “Microsoft COCO: Common Objects in Context.” *Lecture Notes in Computer Science*, 2014, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.