Positive and Negative Pair Exposure in LFW Identity Verification

Derek Mason (IQ), Charles Wallace (SG), Kyle Butler (IT)

Abstract

The same number of matched and mismatched LFW pairs does not imply the same identity coverage. The released ten-fold list contains 1,532 identities in positive pairs and 3,736 in negative pairs, with 987 shared between those roles. Another 545 appear only in positive pairs and 2,749 only in negative pairs. This metadata analysis shows why class balance and subject balance are different properties. For subject-preserving video evaluation motivated by ARGUS, the reference population and the comparison roles should be recorded separately before a similarity average is interpreted.

A balanced label count can hide different roles

A verification protocol asks whether two observations refer to the same identity. A matched pair contributes one represented identity, while a mismatched pair contributes two. The resulting identity populations can differ even when the two pair classes have equal numbers of rows. A person may be represented through within-person appearance variation, through distractor comparisons, or through both. These roles pose different measurement questions and should not be inferred from the total pair count alone.

ARGUS addresses subject-preserving generation using structured multi-view identity evidence (Meng et al., 2026b). This motivates asking which subjects are represented as the intended target and which appear as alternatives or distractors. The present study does not generate video or measure ARGUS recognition performance. It audits the positive and negative roles in a public static verification list. The connection is a design principle: a subject benchmark needs an account of how identities enter each comparison, not simply a claim that its pair labels are balanced.

The counts recovered from the prescribed file

The analyzed LFW file is the ten-fold assessment list identified in the documented software loader. It contains 3,000 matched and 3,000 mismatched pairs. Identity labels are read from the prescribed rows, without downloading images or inferring any demographic attributes. For matched pairs, the identity is counted once per pair when describing positive-role exposure. For mismatched pairs, the two identity fields contribute their respective appearances. This definition preserves the different structures of the two comparison types.

There are 1,532 distinct positive-role identities and 3,736 negative-role identities. Their intersection contains 987. The positive-only set therefore contains 545 identities, and the negative-only set contains 2,749. Their union is 4,281, consistent with the independently calculated total identity count for the complete list. The maximum positive pair count for one identity is nine; the maximum number of negative pair-side appearances is seven. These maxima have different units and should not be compared as though they were the same exposure measure.

What the asymmetry supports and does not support

The result shows a substantial difference in the breadth of identities occupying the two roles. It does not show that any identity was selected unfairly, that a classifier performs differently across the roles, or that the protocol's balanced classes are inappropriate. No prediction scores or errors are observed. The audit instead identifies the population structure that a subsequent accuracy analysis would need to retain. An identity-macro positive endpoint would not describe the same set of people as a negative-role macro endpoint.

For a generated-subject benchmark, a similar distinction can arise when a small set of reference subjects is compared with many distractors. A high rate of rejecting alternatives need not establish fidelity across the subject's own viewpoints. Conversely, strong within-subject similarity does not establish discrimination among visually similar alternatives. A useful evaluation would identify target subjects, distractor subjects, available reference observations, and the rules by which the two are paired. Their denominators should remain recoverable in per-case results.

The composition of the distractor set also matters. Repeatedly using easy alternatives can increase the number of negative comparisons without adding a difficult distinction. Using very similar alternatives can make a different task that requires careful annotation and independent interpretation. The LFW labels analyzed here do not supply a newly measured hardness score, so neither interpretation is assigned to its negative pairs. The scientific value of the audit is to separate an observed role count from unobserved claims about visual difficulty.

An additional challenge is how uncertainty is calculated when identities reappear. Two comparisons can share a person without sharing the same photograph. Their errors may therefore be dependent in ways a simple class-balanced row count does not reveal. A future analysis should preserve both identities and both image keys rather than choose an arbitrary left-side grouping. The present metadata counts are exact for the parsed file and do not provide an estimated error correlation or a correction factor for a classifier's confidence interval.

Exposure as a property of pair construction

For the proposed comparison, the key observation is identity and pair role. A person can appear frequently in between-person comparisons and never contribute a within-person comparison. This hypothetical case poses a question about overlap of identity exposure across the two comparison roles. Record participation in matched pairs and participation in nonmatching pairs separately. Pair balance can coexist with unequal evidence about same-person variation across the underlying identities. Rebuild the identity-by-role incidence table and retain counts of unique photographs as a separate column.

A two-column incidence table can identify subjects contributing positive comparisons only, negative comparisons only or both. These categories can be calculated from the released pairs without inferring how many photographs existed outside the protocol. Pair count and person count should appear together. If a person contributes many comparison slots, it may be because of repeated use of a few portraits. A follow-up verification study could sample within each exposure category, document image eligibility and check whether the coverage difference persists after counting unique source photographs.

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

Store conditioning views, generated frames and visibility annotations separately for identity and pair role. To evaluate overlap of identity exposure across the two comparison roles, keep paired observations together during resampling, since independent resampling would lose their shared source of variation. When examining participation in nonmatching pairs, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

A sensitivity display can show where the conclusion about overlap of identity exposure across the two comparison roles changes as assumptions about the eligible observations change. The evaluation of participation in matched pairs can distinguish an incorrect subject from a permitted change in expression, pose or scene. Set aside examples of identity and pair role for independent assessment, rather than testing transfer by revisiting the cases that motivated the protocol.

Samples and evidence

Another study provides face observations with substantial variation in pose and age (Cao et al., 2017). Pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled. For implementation, earlier research defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow.

The size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description. An existing collection provides a starting point for auditing subgroup coverage while leaving the deployment distribution to be specified separately. For implementation, a relevant contribution tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). Another study constructs a face attribute dataset with demographic balance (Kärkkäinen and Joo, 2019).

Measurement and error

Earlier research compares choices of forecast error scale across series (Hyndman and Koehler, 2006). A general visual representation can provide an independent similarity score whose errors may differ from those of a face or prompt-specific model. One approach learns general-purpose visual features from curated image data (Oquab et al., 2023). Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable.

A cosine-margin baseline helps distinguish rejection due to the representation from rejection caused by the threshold. A quality-sensitive representation can be used to define inspectable strata before comparing acceptance rates across views with very different visibility. A methodological precedent learns facial representations with a quality-related feature magnitude (Meng et al., 2021). During validation, earlier research uses a cosine margin for discriminative face recognition (Wang et al., 2018).

Another study compares the local structure of a pair of images (Wang et al., 2004). An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated. An existing method formulates classification decisions with explicit costs for different errors (Elkan, 2001). During validation, local structural agreement can be inspected alongside task metrics when pixel changes are expected but geometry or regional coherence should remain stable.

A methodological precedent investigates scaling in contrastive language-image learning (Cherti et al., 2022). For a prospective comparison, after fitting a decision threshold, reliability curves on an independent split can show whether reported confidence still matches observed success frequency. Related work assesses probability calibration and studies temperature scaling (Guo et al., 2017).

Representation changes can alter relative prompt scores, so the encoder checkpoint and text preprocessing belong in the comparison record.

Angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture. One study learns an embedding space for face verification and recognition (Schroff et al., 2015). Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity. For implementation, related work uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018).

An existing method distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). A relevant contribution combines detection and association considerations in tracking assessment (Luiten et al., 2020). Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings. Report the detection and association components as well as the aggregate so that tracking continuity can be interpreted alongside object availability.

Another study jointly localizes faces and facial landmarks (Deng et al., 2019). Using a representation learned without explicit class labels gives a useful comparison when a prompt-derived score may reward the wrong visual feature. In an evaluation, one published study learns visual representations through self-supervision (Caron et al., 2021). Landmark localization provides a way to identify crop failures before measuring resemblance and to retain a denominator for faces that were never detected.

Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential. Related work provides object annotations and standardized visual recognition tasks (Lin et al., 2014). Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. For implementation, another study constructs a full-reference metric with structure and texture components (Ding et al., 2020).

Systems and comparators

One approach uses appearance information in online association (Wojke et al., 2017). One study separates global facial structure and local detail in identity-conditioned video (Yuan et al., 2024). In practical terms, an appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. Separate facial structure and texture checks can identify whether a convincing local detail conceals changes in global appearance over time.

An existing method uses tuning-free identity customization with attention to preserving image editability (Guo et al., 2024). An adapter can hold the type of external condition constant when comparing how faithfully different systems honor an edge, depth or pose specification. One published study introduces adapters to convey spatial or visual conditions to image diffusion (Mou et al., 2023). For implementation, a paired edit test can separate matching the reference subject from following a permitted change to composition, clothing or surrounding scene.

Analysis and interpretation

A family of endpoints should be declared before evaluation so that reported discoveries reflect every comparison eligible to appear in the result. In an evaluation, a methodological precedent builds a stacked identity representation from input portraits (Li et al., 2023a). A stacked portrait input permits a test of view coverage, provided the evaluator retains an independent portrait and records near-duplicate input views. One published study controls the false discovery rate for a declared family of hypotheses (Benjamini and Hochberg, 1995). A relevant contribution reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017).

An existing method allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). In an evaluation, single-reference conditioning provides a budget-matched comparison for identifying which errors disappear when extra views are made available to the generator. Serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points. A methodological precedent conditions identity-preserving image synthesis on a single reference (Wang et al., 2024).

Conclusion

Positive and negative LFW pairs expose different identity populations despite equal class counts. The distinction is directly observable in the public protocol and requires no new face analysis. For subject-preserving generation, it motivates separate records of target and distractor coverage, comparison roles, and repeated identities. These records cannot prove a generator's fidelity, but they define which claim a later score can support. Class balance is a useful property whose meaning becomes clearer when the identities behind each class remain visible.

References

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.

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.

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.

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.

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.

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.

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.

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.

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Deng, Jiankang, et al. "RetinaFace: Single-stage Dense Face Localisation in the Wild." *arXiv preprint arXiv:1905.00641*, 2019. https://arxiv.org/abs/1905.00641.

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

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

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.

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.

Oquab, Maxime, et al. "DINOv2: Learning Robust Visual Features without Supervision." *arXiv preprint arXiv:2304.07193*, 2023. https://arxiv.org/abs/2304.07193.

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.

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.

Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Mou, Chong, et al. "T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models." *arXiv preprint arXiv:2302.08453*, 2023. https://arxiv.org/abs/2302.08453.

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.

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

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Wang, Qixun, et al. "InstantID: Zero-shot Identity-Preserving Generation in Seconds." *arXiv preprint arXiv:2401.07519*, 2024. https://arxiv.org/abs/2401.07519.

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.

Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

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.

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.

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

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.

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.