Identity Coverage before Identity Scores: An Audit of the LFW Verification Pairs

Nathaniel Peterson (PY), Jordan Cooper (MU), Nathan Stewart (MW)

Abstract

The released LFW verification list contains 6,000 pairs but 7,701 distinct images, so pair counts should not be read as counts of independent photographs. An audit of the ten prescribed folds finds 4,281 identities and no identity or image shared across folds. Within folds, 3,217 images participate in multiple pairs. Identity exposure is also uneven: its concentration corresponds to about 2,413 equally represented identities. These findings support the original separation between folds while showing why uncertainty within a fold needs care. For subject-preserving video generation, identity coverage and exposure should be documented before a single similarity score is interpreted.

1. The reference set is part of an identity claim

A generated subject can look recognizable from one viewpoint and become ambiguous when the camera turns. Evaluating that behavior requires decisions about who is represented, which reference observations are available, and how often the same observation contributes to a reported result. Those decisions are distinct from selecting a face embedding. A benchmark may contain many pair comparisons while providing substantially fewer independent reference images. It may also be well separated across evaluation folds despite reusing observations within a fold.

ARGUS addresses identity preservation in video through stacked multi-view identity mosaic injection (Meng et al., 2026b). Its reported mechanism organizes selected identity evidence into a compact mosaic and supplies that evidence during generation. The paper's stress tests address changes in viewpoint and occlusion. This motivates asking what the reference observations actually cover. The present study does not run ARGUS or measure its identity fidelity. It examines a much older static verification protocol to make the units of identity evidence explicit, then develops the implications for a future video evaluation.

The empirical object is the public ten-fold pair list for Labeled Faces in the Wild (Huang et al., 2008). Its relevance is methodological rather than a claim that static web photographs reproduce the conditions of generated video. The pair protocol provides an unusually transparent opportunity to distinguish images, identities, comparison slots, and folds. All of those units can be counted without downloading facial images or applying a recognition system.

2. What was downloaded and counted

The input is the pairs.txt file distributed through the official scikit-learn LFW loader's documented mirror. The loader identifies the ten-fold file separately from the development training and test lists. Its source code and the downloaded bytes were retained with hashes. Scikit-learn provides a reproducible route to established dataset protocols, but its software interface is not a new source of human identity labels (Pedregosa et al., 2011). The original LFW paper remains the source for the protocol's scientific purpose.

The file header specifies ten folds with 300 matched pairs and 300 mismatched pairs per fold. A matched row names one identity and two image indices; a mismatched row names two identities and an image index for each. The parser verifies the resulting 6,000 rows. An image identifier combines its identity label with its image index, because an index alone is reused across people. A fold is assigned from the prescribed row blocks rather than generated through a fresh random split.

An identity slot is one appearance on one side of a pair. There are therefore 12,000 identity slots. The same person can occupy many slots, and an image can be compared with more than one other image. Counts are summarized anonymously in the manuscript; no person's name is needed to demonstrate the weighting issue. The analysis does not infer demographic attributes, assess recognition accuracy, or decide whether any identity annotation is correct.

Datasheets distinguish the provenance and intended uses of a dataset from the questions a later analyst might ask of it (Gebru et al., 2021). That distinction is especially useful here. The public pair list supports a protocol audit; it does not supply trajectories, occlusion measurements, consent histories, or video generation outputs. Preserving those limits allows a small analysis to be informative without turning metadata into evidence about unobserved behavior.

3. Separation between folds survives direct checking

For each fold, the set of identities and the set of image identifiers are compared with the union of the other nine folds. All ten identity intersections are empty. All ten image intersections are also empty. Thus the audit finds no cross-fold reuse of identities or images in the analyzed pair file. This is a positive property of the prescribed organization, not an issue to be repaired by randomly repartitioning pair rows.

Fold sizes are equal at 600 pairs, but the number of identities ranges from 397 to 451. Distinct image counts range from 737 to 813. Equal pair counts therefore do not imply identical identity coverage. An evaluator who averages fold accuracies gives each equally sized pair set the same influence, while an identity-macro summary would give people equal influence. These are different targets even when both follow the correct file.

Quantity in the pair file	Value	Unit	Interpretation
Verification pairs	6,000	Comparisons	Prescribed evaluation rows
Distinct identities	4,281	People labels	Coverage of this list
Distinct images	7,701	Photographs	Unique identity-index combinations
Reused images	3,217	Photographs	Images appearing in multiple pairs

The largest number of pair appearances for any image is six. Reuse occurs within folds, not across them. It therefore concerns dependence among comparisons rather than leakage across the prescribed fold boundary. Structured validation should follow the data's grouping (Roberts et al., 2017); replacing the original fold assignment with a row-level split could destroy precisely the separation confirmed by the audit. Model selection performed on evaluation folds would introduce a different problem, which requires its own protected validation boundary (Varma and Simon, 2006).

4. Coverage and concentration describe different things

The 4,281 observed identities occupy their 12,000 slots unevenly. The largest identity share is 0.167 percent, and the ten largest shares together account for 1.417 percent. Concentration is therefore not an extreme case in which one person dominates the entire list. Nevertheless, unequal exposure matters for how an identity-level summary would be interpreted.

To describe concentration, each person's slot count is divided by 12,000. The reciprocal of the sum of squared shares is 2,413.35. This is the number of equally exposed identities that would have the same concentration. It is not an estimated sample size for a recognition accuracy, a count of independent people

in a population, or a confidence-interval correction. No prediction errors were observed, so the analysis cannot estimate the dependence of errors within identity or image.

A separate calculation uses only matched pairs and gives every matched comparison one identity contribution. Its concentration-equivalent count is 1,016.72. This number answers a different question from the all-slot calculation: it describes the distribution of within-person comparisons rather than the identities appearing in both positive and negative examples. Reporting the two without defining their denominators would create an apparent inconsistency. Stating the unit resolves it.

Repeated-image error dependence could be positive, negative, or small for a particular system; the pair file alone does not reveal it. Cluster-robust inference requires a grouping structure compatible with the errors actually analyzed (Cameron and Miller, 2015). A heteroskedasticity correction is not a general substitute for that dependence structure (White, 1980). For dyadic comparisons, grouping by just the identity on the left side is particularly arbitrary. A future accuracy analysis should retain both identities and both image identifiers and justify its variance calculation accordingly.

5. An identity embedding does not define the evaluation population

FaceNet learns an embedding intended to support face verification and related tasks (Schroff et al., 2015). ArcFace and CosFace develop margin-based objectives for discriminative face representations (Deng et al., 2018) (Wang et al., 2018). Their relevance is that a numeric identity score is produced by a learned measurement procedure. A protocol still determines which images enter that procedure, how a threshold is selected, and whose errors receive weight. A strong embedding cannot retrospectively turn repeated photographs into independent observations.

Dataset design changes the range of identity challenges. VGGFace2 explicitly expands variation in pose and age (Cao et al., 2017), while MegaFace increases the scale of distractor-based evaluation (Kemelmacher-Shlizerman et al., 2015). Neither property is established by simply increasing the number of comparisons within the same reference set. A future ARGUS evaluation should distinguish additional viewpoints of the same person, additional people, additional motion trajectories, and additional distractor identities. Each expands evidence along a different dimension.

FairFace addresses representation across annotated demographic groups (Kärkkäinen and Joo, 2019). The present audit does not attach those attributes to LFW identities and cannot evaluate demographic balance. Identity-frequency balance should therefore not be described as demographic fairness. A causal account of performance differences would need information about acquisition conditions, selection, and model training that these pair labels do not contain (Greenland et al., 1999). Keeping a descriptive concentration measure separate from a fairness or causal claim is a substantive part of the analysis.

6. From static references to subject-preserving video

Subject personalization introduces another layer between the reference set and the generated output. DreamBooth studies subject-driven generation through adaptation of a text-to-image model (Ruiz et al., 2022), and IP-Adapter develops image prompting compatible with text conditioning (Ye et al., 2023). ARGUS instead emphasizes a structured multi-view identity memory within video generation. These are different mechanisms for using reference evidence. A comparison should record the number and type of reference observations available to each method rather than letting an input advantage disappear inside a final score.

A reference mosaic can improve the availability of visual evidence without proving that a generated trajectory preserves identity. The evaluator must observe the output under the relevant pose and occlusion

conditions. VBench decomposes video generation quality into multiple dimensions (Huang et al., 2023). In that spirit, identity fidelity should be reported alongside temporal consistency, motion, and prompt compliance. A single frontal reference comparison can miss the very viewpoints motivating a multi-view design.

Generic visual representations also have distinct purposes. CLIP connects image and language representations (Radford et al., 2021), whereas DINO investigates self-supervised visual features (Caron et al., 2021). Either may be useful as one component of an assessment, but neither automatically constitutes a calibrated person-identity measurement across generated video. Structural similarity (Wang et al., 2004), learned perceptual distance (Zhang et al., 2018), and video-distribution metrics (Unterthiner et al., 2018) add other views of visual behavior. They should not be silently substituted for evidence that viewers can recognize the same subject across a trajectory.

The LFW findings suggest a concrete recording rule for a video benchmark: keep the identities, reference images, generated clips, and frame identifiers separately linked. Two methods might share a reference image but generate different clips, or produce multiple trajectories for one person. An analyst should be able to choose an identity-level, clip-level, or frame-level target without reconstructing those relationships from filenames after evaluation.

7. What a subsequent score analysis should report

First, define the primary average before examining model rankings. An identity-macro average asks how a uniformly selected represented person fares; a clip-weighted average asks how a uniformly selected released clip fares. Exposure is part of the question. Report the other weighting as a sensitivity analysis when practical. No weighting can recover people or motion conditions absent from the evaluation set.

Second, preserve uncertainty at the appropriate level. Threshold selection and classifier calibration should be separated from final measurement (Guo et al., 2017). Confidence can deteriorate under changes in the data distribution (Ovadia et al., 2019), including a move from photographs to generated video. If a system or judge declines ambiguous identity decisions, its abstention rate should accompany its accuracy among accepted cases (Geifman and El-Yaniv, 2017). None of those quantities is estimated by the present metadata audit.

Third, compare measurements rather than assuming equivalence. Automated identity similarity and human judgments may correlate while differing systematically under profile views or partial occlusion; method-comparison analysis makes that possibility explicit (Bland and Altman, 1986). Many viewpoint, subject, and time-window comparisons create multiplicity, so a future inferential study should specify primary tests and handle exploratory comparisons transparently (Benjamini and Hochberg, 1995). Persistent identifiers and documented transformations support reusable evaluation (Wilkinson et al., 2016), while tracking model and data dependencies helps prevent a scoring pipeline from changing unnoticed (Sculley et al., 2015).

8. Conclusion

The prescribed LFW pair file separates identities and images cleanly across its ten folds. Its 6,000 comparison rows nevertheless reuse photographs within folds and give different identities different exposure. Those statements are compatible: a sound split does not imply independent rows or a unique population average. Subject-preserving video evaluations can retain the same discipline by documenting reference coverage and grouping before interpreting similarity scores. The audit supplies no new recognition result, but it identifies the records required for such a result to have a clear meaning.

References

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

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.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

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.

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.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Ye, Hu, et al.. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

Huang, Ziqi, et al.. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.

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

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

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

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.

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

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.