

Benchmarking and Stress-Testing Off-the-Shelf Vision Embeddings for Surface Re-Identification in Open-Source Intelligence

Mohammadreza Rashidi

`mohammadreza.rashidi@ue-germany.de`

University of Europe for Applied Sciences <https://orcid.org/0009-0003-7136-7168>

Research Article

Keywords: surface re-identification; open-source intelligence; texture analysis; image retrieval; vision embeddings; CLIP; DINOv2; benchmark

Posted Date: July 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10189140/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Can Off-the-Shelf Vision Embeddings Re-Identify a Physical Surface? A Benchmark of CLIP, DINOv2, and CNN Features for OSINT Texture Matching

Mohammadreza Rashidi 

Department of Computer Science

AI and Media Analysis Lab

Berlin, Germany

mohammadreza.rashidi@ue-germany.de

Abstract—Open-source intelligence increasingly relies on matching surfaces, a granite countertop, a wood grain, a fabric weave, across photographs to link an anonymous image to a known place. A popular framing proposes that pretrained vision embeddings do this matching and cites confidence figures such as a “94.7% accuracy” without a benchmark or a task definition. We measure what these models actually do. We define surface re-identification as retrieving images of the same physical surface from a gallery of many surfaces, and evaluate the off-the-shelf backbones the framing recommends, CLIP, DINOv2, ResNet-50, and EfficientNet-B0, in seven variants including the large CLIP ViT-L/14 and DINOv2 ViT-L/14, as frozen embedders on the KTH-TIPS2-b texture database (4752 images, 44 physical surfaces, with same-material-different-sample images as hard negatives). The best model, DINOv2 ViT-L/14, returns the correct surface as its top result 98.1% of the time and verifies same versus different surfaces at AUC 0.973, yet its mean Average Precision is only 0.646 and its equal-error rate is 7.2%, and both degrade across zoom changes and on smooth surfaces. Every DINOv2 model beats every CLIP model: even the smallest DINOv2 (ViT-S/14, mAP 0.604) outranks the much larger CLIP ViT-L/14 (mAP 0.590), the model the framing names as best. We conclude that off-the-shelf embeddings are useful as lead generators for a human analyst but do not support an unqualified accuracy claim, and that a self-supervised model, not CLIP, is the better default for surface matching regardless of size. All numbers are reproduced from released code by an independent verification script.

Index Terms—OSINT, texture analysis, image retrieval, re-identification, CLIP, DINOv2, verification, open-source intelligence

I. INTRODUCTION

Background surfaces leak identity. When classified United States documents were photographed and posted online in 2023, the photographs of the documents also captured the surface they rested on, and reporting on the case described how open-source investigators connected the leaked images to a specific location through the surroundings visible in them [1]. The episode is now a standard illustration of a broader open-source intelligence (OSINT) principle: a granite countertop, a wood grain, a tile layout, or a fabric weave is a near-unique visual fingerprint, and matching it across photographs can link an anonymous image to a known place or person.

A popular framing of this idea proposes that pretrained vision models do the matching: extract an embedding from each image with CLIP, DINOv2, or a CNN, and rank surfaces by embedding similarity. The framing is appealing because these models are free, fast, and require no training. It is also frequently stated with confidence figures, for example a “94.7% texture match accuracy,” that are reported without a benchmark, a task definition, or a negative control.

This paper measures what these models actually do on the task. We define surface re-identification precisely: given a query photograph of a surface, retrieve other photographs of the *same physical surface* from a gallery that also contains many different surfaces. We evaluate the off-the-shelf backbones, the ones the popular framing recommends, as frozen feature extractors on a real texture database in which the same physical sample is photographed under controlled changes of scale, pose, and illumination. We report standard retrieval and verification metrics with confidence intervals, a paired statistical comparison between models, and a cross-scale analysis that isolates the OSINT-critical case where the query and the true match were taken at different zoom levels.

Contributions:

- A reproducible surface re-identification benchmark built on the KTH-TIPS2-b texture database [2], with the physical sample as the identity unit and same-material-different-sample pairs as hard negatives.
- An honest evaluation of CLIP [3], DINOv2 [4], ResNet-50 [5], and EfficientNet-B0 [6] as frozen embedders, with retrieval (Rank-1, Recall@5, mAP) and verification (ROC-AUC, equal-error rate) metrics, bootstrap confidence intervals, and paired Wilcoxon tests.
- A cross-scale and per-material analysis showing where surface matching is reliable and where it breaks, and what that implies for using these tools as investigative leads rather than proof.

II. BACKGROUND AND RELATED WORK

A. Pretrained embeddings as general visual features

CLIP learns image representations from image-text pairs and transfers to many tasks without fine-tuning [3]; we use

the OpenCLIP implementation and OpenAI weights [7]. DINOv2 learns features by self-supervision and is designed to give strong dense and global features without labels [4]. ResNet [5] and EfficientNet [6] are supervised ImageNet CNNs whose pre-logit activations are a long-standing general-purpose embedding. We take all four from the `timm` and OpenCLIP libraries [8] and use the pooled embedding with cosine similarity.

B. Texture datasets

Texture recognition is usually posed as material *classification*: the DTD dataset describes textures by attributes [9], and KTH-TIPS2 categorises materials under varying imaging conditions [2]. KTH-TIPS2 is built around physical *samples*: each of 11 materials has four distinct physical samples, and each sample is photographed at nine scales, three poses, and four illuminations. That sample-level structure is what makes it usable for re-identification rather than classification: it lets us ask whether a model matches the same physical surface, with other samples of the same material acting as hard distractors.

C. Retrieval and re-identification

Our protocol and metrics follow image retrieval and person re-identification [10]: a query ranks a gallery, and we score Rank-1, Recall@k, and mean Average Precision. The OSINT use is closer to geolocation by visual matching [11] than to classification, because the investigator wants to link a specific image to a specific surface, not to name the material.

D. Positioning

We are not proposing a new model or a new attack. We measure how well the exact off-the-shelf tools that OSINT writing recommends perform on a precisely defined surface re-identification task, and we report the numbers with the controls that confidence claims in that writing omit.

III. METHOD

A. Task and identity

A physical surface instance is a (material, sample) pair; KTH-TIPS2-b has 44 instances across 11 materials, 4752 images in total. Two images match iff they show the same instance. Two different samples of the same material (for example two different wool swatches) are a non-match, which makes the negatives hard and the task strict.

B. Embeddings

Each backbone is a frozen feature extractor. We resize and normalise each image with the model’s own preprocessing, take the pooled pre-logit embedding, L2-normalise it, and compare embeddings with cosine similarity. No fine-tuning and no metric learning are used; this is the off-the-shelf setting the popular framing assumes. We evaluate seven variants: CLIP ViT-B/32 and ViT-L/14, DINOv2 ViT-S/14 and ViT-L/14, ResNet-50, and EfficientNet-B0, all at a common 224-pixel input, plus DINOv2 ViT-S/14 at its native 518-pixel input to isolate the effect of resolution. The large and small

variants let us separate the effect of model family from the effect of model size. Embedding dimensions range from 384 (DINOv2 ViT-S) to 2048 (ResNet-50). The four base models were also run independently on an Apple-silicon GPU and on an NVIDIA GPU; the metrics matched to three decimals, confirming the numbers are hardware-independent.

C. Retrieval metrics

Every image is a query against the gallery of all other images. For each query we rank the gallery by cosine similarity and mark images of the same instance as relevant. We report Rank-1 accuracy (is the top result the same instance), Recall@5 (is any same-instance image in the top five), and mean Average Precision (mAP) over all queries. We also report a *cross-scale* variant in which the gallery is restricted to images photographed at a different scale than the query, so a correct match must come from a different zoom level. This is the OSINT case: the leaked photo and the social-media photo are rarely at the same magnification.

D. Verification metrics

Re-identification answers “which gallery image,” but an investigator often asks “are these two the same surface?” We sample balanced same-instance and different-instance image pairs, score each by cosine similarity, and report ROC-AUC and the equal-error rate (EER). This is the measured analogue of an unqualified “match accuracy” figure.

E. Statistics

We report 95% bootstrap confidence intervals for every aggregate (2,000 resamples for retrieval, 1,000 for AUC). We compare models with a paired Wilcoxon signed-rank test on per-query Average Precision, since every model is evaluated on the same queries. We stratify mAP by material and by the query-to-match scale gap.

F. Reproducibility

The dataset is public [2]; the image index, the embedding extraction, the metrics, the statistics, and the figures are produced by released scripts. `src/benchmark.py` writes the per-query results and verification scores; `src/analyze.py` produces every table, number, and figure; and `src/verify_numbers.py` re-derives every reported number from those files. Models run on a single GPU through PyTorch [12]; the base models were checked on both an Apple-silicon and an NVIDIA GPU with matching results.

IV. RESULTS

Table I reports every model on every metric; Fig. 1 and Fig. 2 show the retrieval scores and the verification ROC curves. Three patterns hold across all backbones.

TABLE I

SURFACE RE-IDENTIFICATION ON KTH-TIPS2-B (4752 IMAGES, 44 PHYSICAL-SURFACE INSTANCES). MAP_{XS} IS THE CROSS-SCALE MAP; AUC IS THE SAME-VS-DIFFERENT VERIFICATION ROC AREA. ALL NUMBERS ARE PRODUCED BY THE RELEASED ANALYSIS SCRIPT FROM THE PER-QUERY RESULTS.

Model	dim	mAP	Rank-1	R@5	mAP_{XS}	AUC
CLIP ViT-B/32	512	0.535	94.3%	99.3%	0.518	0.938
CLIP ViT-L/14	768	0.590	95.4%	99.6%	0.575	0.960
DINOv2 ViT-S/14	384	0.604	97.6%	99.7%	0.586	0.968
DINOv2 ViT-L/14	1024	0.646	98.1%	99.7%	0.631	0.973
DINOv2 ViT-S/14 (518px)	384	0.611	98.1%	99.9%	0.592	0.969
ResNet-50	2048	0.505	96.2%	99.4%	0.484	0.938
EfficientNet-B0	1280	0.552	97.0%	99.7%	0.531	0.950

A. Top-1 is easy, full ranking is not

Rank-1 accuracy is high for every model, from 94.3% (CLIP ViT-B/32) to 98.1% (DINOv2 ViT-L/14), and Recall@5 is above 99.4% everywhere. The single nearest neighbour is almost always another image of the same physical surface. Mean Average Precision tells a different story: it sits between 0.505 (ResNet-50) and 0.646 (DINOv2 ViT-L/14). The gap means that while the top match is reliable, the deeper ranked list mixes in the wrong surfaces, including other samples of the same material. For an investigator this is the difference between a lead and a list: the best single hit is trustworthy, an exhaustive ranking is not.

B. Family beats size: DINOv2 wins, CLIP does not

DINOv2 ViT-L/14 is the strongest model on every metric (mAP 0.646, Rank-1 98.1%, AUC 0.973, EER 7.2%). Scaling helps within a family: CLIP improves from mAP 0.535 (ViT-B/32) to 0.590 (ViT-L/14) and DINOv2 from 0.604 (ViT-S/14) to 0.646 (ViT-L/14). But the family matters more than the size. Every DINOv2 model beats every CLIP model, and the decisive comparison is that the smallest DINOv2 (ViT-S/14, 384-d, mAP 0.604) outranks the much larger CLIP ViT-L/14 (768-d, mAP 0.590), a gap that is significant ($p < 10^{-7}$, paired Wilcoxon). The popular framing names CLIP as the best texture matcher; in our measurements CLIP is the weakest family, and making it three times larger does not close the gap to the smallest DINOv2. Even the small supervised CNN EfficientNet-B0 (mAP 0.552) outranks CLIP ViT-B/32 (mAP 0.535). All 21 pairwise per-query AP differences are significant ($p < 10^{-7}$). Raising DINOv2 ViT-S/14 to its native 518-pixel input lifts mAP only from 0.604 to 0.611, so resolution is a minor factor and does not change the ranking. A self-supervised model trained without labels or text captures surface micro-structure better than text-aligned CLIP features, at any size we tested.

C. Zoom and smooth surfaces break matching

Restricting the gallery to a different photographed scale than the query, the OSINT case where the two photos are at different zoom levels, lowers mAP for every model (Fig. 3): DINOv2 ViT-L/14 falls from 0.646 to 0.631 and ResNet-50 from 0.505 to 0.484. The drop is modest here because

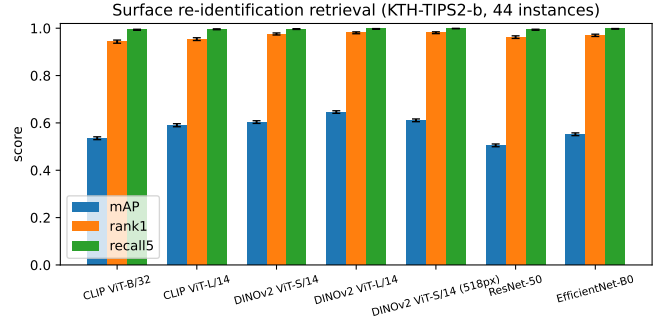


Fig. 1. Retrieval scores with 95% bootstrap confidence intervals. Rank-1 and Recall@5 are high for all models; mAP is much lower, showing that ranking every view of a surface is far harder than finding one.

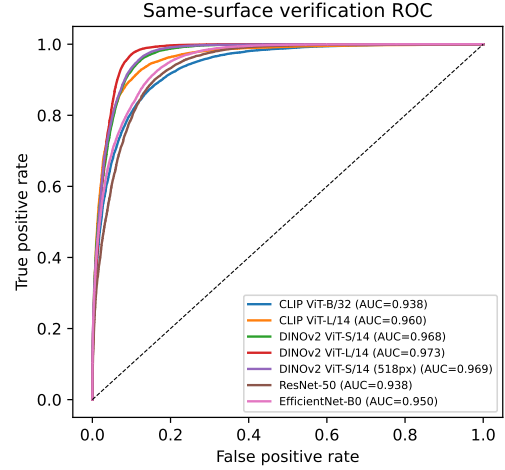


Fig. 2. Same-surface verification ROC. DINOv2 ViT-L/14 separates same from different surfaces best (AUC 0.973), but the equal-error rate is still 7.2%.

KTH-TIPS2 scales are dense, but it is consistent and in the direction that hurts real investigations. Performance also varies sharply by material. For the best model, mAP ranges from 0.489 on aluminium foil to 0.840 on lettuce leaf. The hardest materials are the smooth and specular ones (white bread, aluminium foil), whose surfaces carry little stable micro-texture; the easiest are the strongly patterned ones. A polished stone countertop sits closer to the hard end than the anecdote suggests.

V. DISCUSSION

A. What an unqualified accuracy figure hides

The framing that motivates this work cites a single “94.7% match accuracy.” Our verification AUC lands near that range (0.938 to 0.973), so the number is not invented, but accuracy alone is the wrong summary. At the balanced operating point the equal-error rate is 7.2% for the best model and up to 13.7% for the worst: between one decision in fourteen and one in seven is wrong. A reported accuracy also depends on the prior over same and different pairs and on how hard the negatives

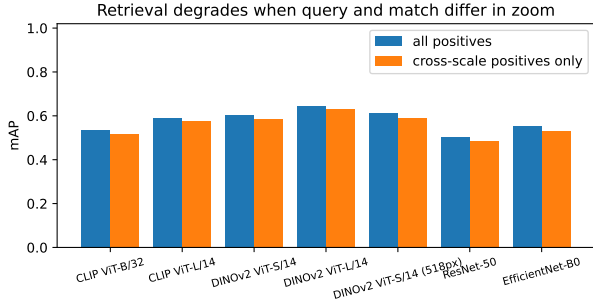


Fig. 3. mAP drops for every model when a correct match must come from a different zoom level than the query.

are; ours uses balanced pairs with same-material distractors, which is strict. An OSINT claim should state the task, the operating point, and the negative set, not a lone percentage.

B. A lead, not proof

The clearest practical message is the split between Rank-1 (up to 98.1%) and mAP (up to 0.646). Off-the-shelf embeddings are good at surfacing one strong candidate and poor at ranking an exhaustive gallery. That supports using these tools to generate a shortlist for a human analyst, with the surface match as one corroborating signal among several, and against presenting an embedding similarity as standalone identification. The cross-scale drop and the material dependence reinforce the point: the method is reliable on strongly patterned surfaces viewed similarly, and shakier on smooth surfaces or across zoom, which is exactly where a real leaked photo and a real social-media photo tend to differ.

C. Model choice

The result that every DINOv2 beats every CLIP, and that even a small supervised CNN beats CLIP, is useful and actionable. For surface matching specifically, a self-supervised feature extractor is the better default than a text-image model. Size helps second: DINOv2 ViT-L/14 is the strongest overall, but DINOv2 ViT-S/14 already beats CLIP ViT-L/14 at a fraction of the cost, so a practitioner short on compute should still prefer the small self-supervised model over a large text-image one. Practitioners copying the “use CLIP” advice are using the weakest family they were offered.

VI. LIMITATIONS

The dataset is the main one. KTH-TIPS2-b is fabrics, foods, foil, wood, and cork, not stone or granite, so it is a proxy for the countertop scenario rather than a direct test of it; a stone-instance dataset would be a better match and does not exist in this form. The images are controlled laboratory captures that vary scale, pose, and illumination cleanly, whereas a real OSINT pair differs in clutter, occlusion, perspective, sensor, and compression, all of which make the task harder than reported here, so our numbers are an optimistic bound on in-the-wild performance. We evaluate frozen off-the-shelf embeddings only; fine-tuning, metric learning, and the classical

texture descriptors the same writing mentions (GLCM, LBP, SIFT) are out of scope and could move the numbers either way. The retrieval is closed-set within one dataset, while an open-world investigation faces an unbounded and unknown gallery. We compare backbones at a common 224-pixel resolution for fairness and also test DINOv2 ViT-S/14 at its native 518 pixels, which gives only a small gain, so resolution is not a confound; we did not sweep resolution for every model. Inference is deterministic given the weights, so the numbers reproduce exactly, and the four base models gave identical metrics on two different GPUs, but we still report a single dataset and one model per family size, not a broad sweep.

VII. CONCLUSION

Off-the-shelf vision embeddings can re-identify a physical surface well enough to be useful and not well enough to be trusted blindly. On a real texture database with 44 physical surfaces, the best model finds the correct surface as its top result 98.1% of the time and separates same from different surfaces at AUC 0.973, but its mean Average Precision is only 0.646, its equal-error rate is 7.2%, and both fall further across zoom changes and on smooth surfaces. Every DINOv2 variant beats every CLIP variant, the smallest DINOv2 beats the largest CLIP, and CLIP is the weakest family, reversing the common advice that CLIP is the best texture matcher. The honest summary for OSINT practice is that these models generate leads for a human to verify, and that any single accuracy figure should be replaced by a task, an operating point, and a stated set of hard negatives.

REFERENCES

- [1] The New York Times, “Membership in online group led to arrest in leak of classified documents,” <https://www.nytimes.com/2023/04/14/us/politics/leaked-documents-jack-teixeira.html>, 2023, reporting on the Jack Teixeira case; accessed 2026-06-14.
- [2] B. Caputo, E. Hayman, and P. Mallikarjuna, “Class-specific material categorisation,” in *Tenth IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1597–1604.
- [3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021, arXiv:2103.00020.
- [4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “DINOv2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research (TMLR)*, 2024, arXiv:2304.07193.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, arXiv:1512.03385.
- [6] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning (ICML)*, 2019, arXiv:1905.11946.
- [7] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, “Reproducible scaling laws for contrastive language-image learning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, openCLIP; arXiv:2212.07143.
- [8] R. Wightman, “PyTorch image models (timm),” <https://github.com/huggingface/pytorch-image-models>, 2019, accessed 2026-06-14.
- [9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, “Describing textures in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, arXiv:1311.3618.

- [10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [11] T. Weyand, I. Kostrikov, and J. Philbin, "PlaNet: Photo geolocation with convolutional neural networks," in *European Conference on Computer Vision (ECCV)*, 2016, arXiv:1602.05314.
- [12] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [baselinesummary.csv](#)
- [perqueryglcm.csv](#)
- [perquerylbp.csv](#)
- [verifdinov2l.npz](#)
- [verifdinov2518.npz](#)
- [verifclipl.npz](#)
- [perquerydinov2l.csv](#)
- [perquerydinov2518.csv](#)
- [perqueryclipl.csv](#)
- [stats.txt](#)
- [stratamaterial.csv](#)
- [modelcompare.csv](#)
- [verificationsummary.csv](#)
- [retrievalci.csv](#)
- [summary.csv](#)
- [veriefficientnet.npz](#)
- [perqueryefficientnet.csv](#)
- [verifresnet50.npz](#)
- [perqueryresnet50.csv](#)
- [verifdinov2.npz](#)
- [perquerydinov2.csv](#)
- [verifclip.npz](#)
- [perqueryclip.csv](#)
- [run.log](#)
- [imageindex.csv](#)