

# Training on Mass Surveillance Datasets Provides No Empirical Benefits

A Pranav\*, Anne Lauscher

Trustworthy AI Lab, University of Hamburg, Germany

## Abstract

Mass surveillance technologies rely on person re-identification (ReID), the task of matching images of the same person by appearance across changes in pose, viewpoint, time, and camera. ReID datasets are built from footage of pedestrians, students, and shoppers, recorded by surveillance cameras in public space without their consent. These datasets have drawn sustained ethical criticism, yet they remain openly hosted and widely used, defended by the claim that the models trained on them are useful for surveillance deployment. We test that claim on the tasks the footage was collected to enable, and it does not hold. We find that models finetuned on ReID datasets do not deliver their claimed deployment benefits: retrieval accuracy degrades outside the training corpus, tracking does not improve over a general pretrained encoder across eight architectures and three pedestrian-tracking benchmarks, and having more data does not change this. We further find that the representations this training produces encode stereotypes: clothing colour becomes far more decodable after finetuning, finetuned encoders are less likely to distinguish between different people wearing same clothing and saliency for the matching decision localises on clothing. We add empirical depth to the ethical case against ReID datasets, which until now made on the consent violation alone. We conclude that these datasets are built from non-consensual data and have no significant benefit in downstream deployment, thus these datasets should not be built.

## 1 Introduction

Advances in artificial intelligence have driven demand for computer vision in mass surveillance (Browne 2015). In research, most computer vision papers extract data about human bodies, and the patents that cite them mostly lead to surveillance applications (Kalluri et al. 2023). A major component of surveillance is *person re-identification* (ReID), *the task of matching cropped images of the same person by appearance, across changes in pose, viewpoint, time, and camera*. Research on ReID has been published continuously at top computer vision conferences (CVPR, ICCV, and ECCV) since the early 2010s, with dedicated surveys,

\*Corresponding author email: pranav (dot) agrawal (at) uni (dash) hamburg (dot) de  
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

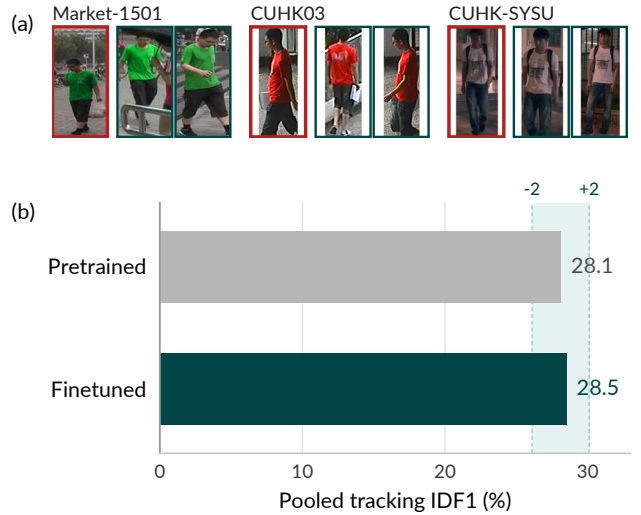


Figure 1: **ReID datasets are built from non-consensual surveillance footage, and finetuning on them does not improve tracking.** (a) Query (red frame) and matching gallery images (teal frames) from MARKET-1501, CUHK03, and CUHK-SYSU, three ReID datasets built from footage of people recorded without consent. (b) Models finetuned on these datasets do not significantly outperform pretrained ones on tracking: mean IDF1 across eight architectures and three benchmarks, with the shaded band marking the  $\pm 2$  pp equivalence margin around the PRETRAINED mean.

toolkits, and benchmark suites (Ye et al. 2021; Zhou and Xiang 2019). ReID models are trained on datasets extracted from passers-by recorded on surveillance cameras at university campuses, subway stations, and other public sites, often without any consent (Zheng et al. 2015; Loy, Xiang, and Gong 2009; Prabhu and Birhane 2020). Most of these corpora remain openly hosted on academic and community platforms (Harvey 2021). Researchers continue to train and publish on these corpora on the premise that the resulting models deliver identity-preserving matching at deployment in mass-surveillance settings (Wang 2013). In this paper, we counter this argument and empirically demonstrate that models trained on ReID datasets do not deliver the deploy-

ment benefits claimed for them.

We list the openly available ReID datasets in Table 1, with their collection sites, identity counts, and documentation status. One of the datasets is GRID, which consists of CCTV recordings from a London Underground station, supplied to Queen Mary, University of London by the UK Home Office under the i-LIDS programme (Loy, Xiang, and Gong 2009). For MARKET-1501, six cameras were set up in front of a campus supermarket at Tsinghua University, and students going to the classes were recorded without consent (Zheng et al. 2015). DUKEMTMC-REID draws on eight surveillance cameras on the Duke University campus, recording students and visitors moving through public spaces (Zheng, Zheng, and Yang 2017). In 2019, the dataset was retracted after reporting traced its use to mass surveillance research used in public, and a subsequent Duke review found that it had been distributed beyond the scope of the original IRB approval (Harvey 2021). Even so, it has continued to appear in published work and remains available on Kaggle, in torrents, and on GITHUB (Peng, Mathur, and Narayanan 2021). The ReID literature defends these corpora on two grounds.

**Claim 1. Finetuning on ReID corpora improves tracking-based surveillance** as ReID re-associates a person through occlusion, pose change, and camera handover, and tracking which helps in public-safety policing, missing-persons recovery, and retail analytics (Wang 2013; Ye et al. 2021; Zhou and Xiang 2019).

**Rebuttal.** ReID is a retrieval task: given a query crop from one camera, the model is asked to find the same person in a gallery from other cameras, and performance is reported as Rank-1 accuracy and mean average precision (mAP). These metrics measure retrieval, not tracking. Retrieval gains therefore carry no evidence about the extrinsic task, and the tracking benefit the claim is about has never been measured.

The public-safety framings in Claim 1 all involve identifying individuals across cameras without their knowledge, and scholars have documented that ReID datasets and models get deployed in policing, defence research, and surveillance industry products, including products sold by companies sanctioned for human-rights abuses (Harvey 2021; Kalluri et al. 2023; Whittaker et al. 2018).

**Claim 2. An encoder trained on footage from one collection site transfers to deployment at unseen but similar sites, and collecting larger corpora would help in improving this transfer** (Lan, Zhu, and Gong 2019; Lin, Li, and Kot 2020; Wei et al. 2018).

**Rebuttal.** Retrieval accuracy degrades as soon as the test corpus changes, and the literature concedes this cross-domain degradation while recommending the models for deployment regardless (Song et al. 2019).

This paper tests whether the usage of ReID datasets is justified. Vision-based mass surveillance, as the ReID literature describes it, consists of tracking pedestrians within a camera and re-identifying them across cameras, and Claims 1

and 2 defend building these datasets by their contribution to these two tasks. We therefore measure what the justification claims: cross-corpus retrieval, tracking, and probing the finetuned representations. We develop our argument from the empirical critique of facial recognition technology (FRT). There, scholars have shown that models tend to rely on ‘shortcuts’ like skin tone and illumination rather than stable identity, which result in errors on marginalised people (Buolamwini and Gebru 2018; Stark 2019; Stark and Hutson 2022). Geirhos et al. (2020) name this pattern *shortcut learning*: a feature that predicts the target inside the benchmark but does not transfer to deployment. In ReID, each identity is recorded at one location over a short window, which could result in shortcuts to appear like clothing colour or body shape. We anticipate the same pattern in ReID and extend the empirical critique of FRT to test it: we measure cross-domain retrieval between the corpora themselves, evaluate ReID-finetuned encoders on tracking benchmarks, and probe the features they encode. Our experiments return two findings.

**Finding 1: We find that finetuning on ReID datasets do not deliver its claimed deployment benefits** (Sections 4–6). Encoders finetuned on one ReID corpus lose most of their retrieval accuracy on every other corpus. On tracking, across eight architectures and three benchmarks, FINETUNED encoders have insignificant gain over the equivalent PRETRAINED baseline. Concatenating all five corpora into a single training set also does not show any improvement in tracking.

**Finding 2: We find that the representations produced by ReID finetuning encode stereotypes, particularly about clothing** (Sections 7–9). From our probing experiments, we find that ReID finetuning encodes shortcuts like clothing color to match people. FINETUNED representations are less likely to distinguish between two different people wearing same clothing color; and finetuning gradients tend to localise more on the shortcuts.

These findings ask a harder question to researchers and policymakers: *why are non-consensual ReID datasets still being built?* We find no empirical basis for continued surveillance ReID research, and we oppose continued research on surveillance ReID on that basis.<sup>1</sup>

## 2 Background And Related Work

In this section we provide the necessary background on the technologies used in mass-surveillance research, and an overview of the critical literature on facial recognition technology from which we extend our arguments.

### 2.1 What is involved in mass surveillance?

Tracking pedestrians across a network of surveillance cameras combines three components running in sequence: *an object detector, a tracker, and a re-identification model*.

<sup>1</sup>Surveillance covers a wide range of practices, including data retention, communications interception, and financial monitoring. In this paper we scope the term in its most immediate, visual sense: computer vision systems that track and re-identify people through cameras.

- **An object detector** like YOLOX processes each frame and returns axis-aligned bounding boxes around every visible pedestrian (Ge et al. 2021).
- **A tracker** like BYTETRACK (Zhang et al. 2022) or DEEPSORT (Wojke, Bewley, and Paulus 2017) associates those boxes within each camera into per-identity trajectories, using a motion model alongside appearance similarity to handle short occlusions and crossings.
- **Person re-identification (ReID)** supplies the appearance signal for both association roles: re-associating a person within one camera through occlusion and crossings, and linking trajectories across cameras whose fields of view do not overlap (Zheng, Yang, and Hauptmann 2016; Ye et al. 2021). Trackers such as DEEPSORT train this appearance model on ReID corpora (Wojke, Bewley, and Paulus 2017).

Each cropped person box is mapped to a fixed-length feature vector. Under the training objective, two crops of the same person yield nearby vectors while two crops of different people yield distant ones, including when the crops come from different cameras at different times. Pedestrian tracking is evaluated against benchmarks of densely populated public-space footage with annotated identity tracks. The MOTCHALLENGE suite covers everyday street scenes (Dendorfer et al. 2020a), MOT20 contributes highly crowded sequences (Dendorfer et al. 2020b), CROWD-TRACK adds difficult real-world scenes with frequent occlusion (Fu et al. 2025), and PERSONPATH22 draws from surveillance, sports, and broadcast footage (Shuai et al. 2022). IDF1 is the F1 score of the bipartite matching between predicted and ground-truth identities across a sequence (Ristani et al. 2016). IDF1 drops when one identity is split across multiple tracks, or two identities are merged into one track. This makes it more sensitive to the appearance encoder than detection-only metrics are.

## 2.2 What datasets are ReID models being trained on?

ReID models are trained on corpora of cropped pedestrian images built from public-space CCTV footage of people who were never asked: pedestrians, transit riders, students on a campus, shoppers in a store. Table 1 lists the five corpora used in our controlled comparison, their collection sites, and the documentation status of each introducing paper. A typical corpus contains tens of thousands of images organised by identity across non-overlapping cameras. MARKET-1501 comprises 32,668 images of 1,501 identities filmed by six cameras outside a supermarket at Tsinghua University (Zheng et al. 2015); DUKEMTMC-REID comprises 36,411 images of 1,812 identities filmed by eight cameras on the Duke campus (Ristani et al. 2016); MSMT17 comprises 126,441 images of 4,101 identities filmed by fifteen cameras at a university campus (Wei et al. 2018). The systems trained on these corpora are intended for deployment in the same public spaces, against the same kinds of publics from whom the training data was taken. Evaluating such a system only for accuracy or fairness reproduces the harm rather than naming it: it accepts the sys-

tem’s premises and sets aside the prior question of how it redistributes power between those who watch and those who are watched (Kalluri 2020). Kalluri et al. (2023) show that people-targeting tasks are the single largest downstream use of academic vision data, and that benchmark research is consistently cited in operational surveillance work. The same pattern of non-consensual collection runs across vision corpora (Prabhu and Birhane 2020), and ReID corpora are the acute case because the downstream task is itself surveillance. Harvey (2021) identifies the cameras and populations behind specific corpora; the retraction of DUKEMTMC did not stop circulation of derived weights and datasets (Peng, Mathur, and Narayanan 2021).

## 2.3 Why do facial recognition systems have flawed deployment?

Facial recognition technology (FRT) recognises individuals from images of their faces. The critical literature on FRT is the closest model for an audit of ReID: both technologies are trained on data taken without consent, and both are deployed in the same surveillance settings.

The FRT critical literature has documented that facial-analysis systems respond to surface attributes rather than the constructs they claim to measure. Arcas, Todorov, and Mitchell (2018) reanalysed a widely-reported deep-learning classifier that claimed to identify sexual orientation from face photographs, and showed that the reported accuracy was carried by grooming, facial expression, and self-presentation; once those surface cues were accounted for, the classifier no longer separated the two classes above chance. Scheuerman, Paul, and Brubaker (2019) found that commercial gender classifiers respond to hair length and other surface cues: error rates track how closely a face matches the binary cisgender presentation dominating the training set, and the classifiers fail systematically on transgender and non-binary subjects. Both cases talk about shortcut learning: a decision rule that separates the training data through superficial cues rather than the construct the benchmark is taken to measure, and that fails once the testing conditions change (Geirhos et al. 2020). Linear probing tests which attributes a frozen representation encodes: a linear classifier is trained on the frozen features to predict the attribute, and its accuracy measures how linearly decodable the attribute is (Alain and Bengio 2018; Belinkov 2021). Decodability shows that the attribute is present in the representation, not that a system built on that representation relies on it (Belinkov 2021). A representation that separates a benchmark through superficial cues fails when those cues shift, and ReID has documented this: cross-domain transfer between academic benchmarks degrades sharply (Song et al. 2019), with the analogous benchmark-to-benchmark gap also established for image classification (Recht et al. 2019).

## 3 Methodology and Experimental Rationale

The community’s case for surveillance ReID corpora rely on the claim that finetuning on these corpora produces encoders that improve identity matching in deployment. We test this claim three ways on the same architecture pool: by measur-

| Dataset                          | Collection site            | Identities | Ethics review | Consent procedure |
|----------------------------------|----------------------------|------------|---------------|-------------------|
| MARKET-1501 (Zheng et al. 2015)  | Tsinghua campus            | 751        | ×             | ×                 |
| CUHK03 (Li et al. 2014)          | CUHK campus                | 767        | ×             | ×                 |
| CUHK-SYSU (Xiao et al. 2016)     | Hong Kong urban streets    | 10,740     | ×             | ×                 |
| ETHZ (Schwartz and Davis 2009)   | Zurich pedestrian streets  | 146        | ×             | ×                 |
| GRID (Loy, Xiang, and Gong 2009) | London Underground station | 125        | ×             | ×                 |

Table 1: ReID training corpora used in the controlled comparison: collection site, training-set identity count, and documentation status of each corpus’s introducing paper. None of the five papers reports an ethics review (×) or a consent procedure (×) for the filmed individuals.

ing cross-domain retrieval between the ReID corpora themselves (Section 3.1), by measuring tracking on three pedestrian benchmarks (Section 3.1), and by probing what is encoded in the frozen features through linear attribute decoding, pair-based identity verification, and saliency attribution (Section 3.2).

**Experiment controls.** Figure 2 summarises the experimental design. We test whether finetuning on a surveillance ReID corpus improves tracking over the matched PRE-TRAINED condition. We hold the architecture, the finetuning recipe, the optimiser, schedule, and augmentation pipeline constant across conditions, and vary only the source of training data, so that the comparison measures the effect of the training corpus rather than the effect of the recipe. Each of eight architectures appears in both a PRETRAINED variant and a FINETUNED variant, with the latter trained separately on each of five ReID corpora. This design lets a difference in tracking performance be read as a consequence of the training corpus.<sup>2</sup>

**Encoder architectures.** The eight encoders cover three architectural categories. The first is two general-purpose Convolutional Neural Network (CNN) backbones widely used as off-the-shelf ReID feature extractors: ResNet-50 (Residual Network, 50 layers) (He et al. 2016) and DenseNet-121 (Densely Connected Convolutional Network, 121 layers) (Huang et al. 2017). The second is three CNN architectures designed specifically for person re-identification: OSNet ×1.0 (Omni-Scale Network), which aggregates features across receptive fields of multiple scales within each layer (Zhou et al. 2019); OSNet-AIN ×1.0, an OSNet variant with adaptive instance normalisation that targets cross-domain transfer (Zhou et al. 2021); and MLFN (Multi-Level Factorisation Net), which decomposes features into latent factors at multiple semantic levels (Chang, Hospedales, and Xiang 2018). These five CNNs cover the architectural choices the ReID community currently uses or recommends, as catalogued in the field’s standard surveys (Ye et al. 2021). The third category is three transformer encoders that represent the current state of the art among general-purpose vi-

<sup>2</sup>Data contamination is not a concern here. IMAGENET (Deng et al. 2009), the ReID corpora, and the tracking benchmarks (Dendorfer et al. 2020b) were assembled for distinct purposes (object classification, cross-camera identity matching, and multi-object tracking evaluation). They have different sources and were collected in different eras; the three data pools do not overlap.

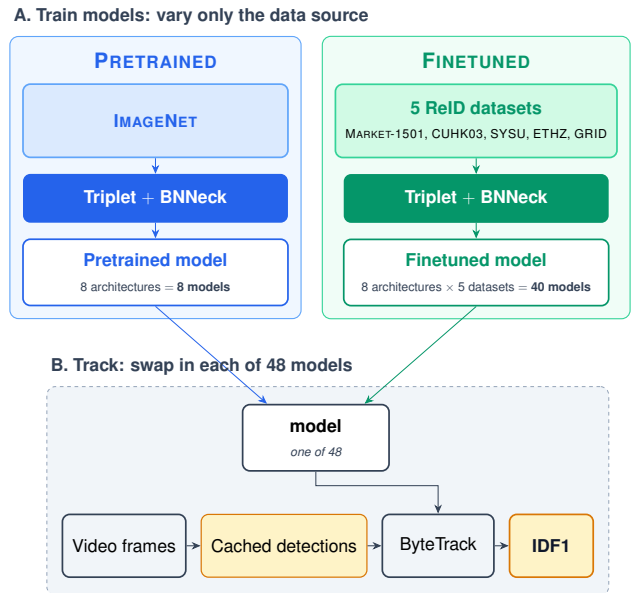


Figure 2: Experimental design for Failure 2: architecture, recipe, and detections are held constant.

sion backbones: ViT-B/16 (Vision Transformer, base,  $16 \times 16$  patch size) (Dosovitskiy et al. 2021), SwinV2-B (Swin Transformer V2, base, with shifted-window attention) (Liu et al. 2022), and DINOv3 ViT-B (Siméoni et al. 2025), a self-supervised vision foundation model. All three are among the strongest general-purpose vision encoders currently available, with state-of-the-art reported performance on standard image-classification benchmarks such as IMAGENET-1k linear evaluation. We do not view these transformer encoders as realistic surveillance deployments: their parameter counts and inference cost place them above what is typically used in the field. We include them so that the comparison covers more than the older or lighter CNN designs the ReID literature has historically favoured, and so that any null tracking result cannot be set aside on the grounds that the encoders involved are weaker than current general-purpose backbones.

**The standard ReID finetuning recipe.** ReID encoders are trained with a triplet loss (Hermans, Beyer, and Leibe 2017): for each anchor crop, the embedding distance to a same-identity crop must be smaller than the distance to any

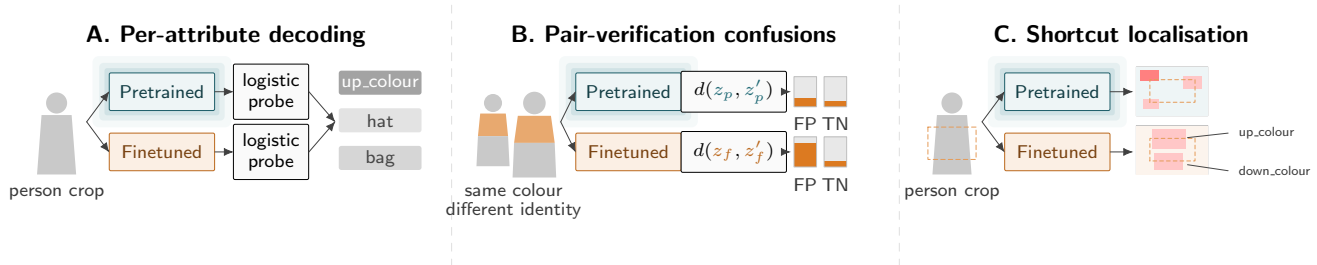


Figure 3: Experimental design for Failures 4–6: Three probes on frozen PRETRAINED and FINETUNED encoders. Per-attribute decoding (A, Section 7), pair-verification confusions (B, Section 8), and shortcut localisation (C, Section 9).

different-identity crop by at least a fixed margin  $m$ . We use the batch-hard variant, which, for each anchor in a mini-batch, supplies the same-identity crop at maximum distance and the different-identity crop at minimum distance, concentrating updates on the triples that currently violate the margin. A Batch-Normalisation neck between the backbone and the classification head separates the embedding used for retrieval from the logits used for an auxiliary classification loss, and has been the default ReID configuration since Luo et al. (2019). All eight architectures share this triplet + BN-Neck head with margin  $m = 0.3$ , the default in Luo et al. (2019) and in the torchreid library (Zhou and Xiang 2019); holding the head and margin constant fixes the loss surface across conditions. The PRETRAINED condition carries the same triplet + BNNeck head and the same embedding extraction path as the FINETUNED variants, and the tracker computes the same cosine distance on the same embedding layer under both conditions, so the two conditions differ in the source of training data and in nothing else.

**Training-data variants.** The five FINETUNED variants are produced by triplet finetuning each of the eight architectures on MARKET-1501, CUHK03, CUHK-SYSU, ETHZ, and GRID, and compared against PRETRAINED. Table 1 characterises each corpus by collection site, training-set identity count, and the documentation status of its introducing paper. We use five corpora to span diversity in collection site (campus, urban streets, public transit), camera placement (fixed CCTV, vehicle-mounted, handheld), and identity count (146 to 10,740). We additionally test whether scaling the training data helps by pooling the five corpora into a single training set of 12,529 identities, which we refer to as COMBINED-5; the per-corpus results are reported in Section 5 and the COMBINED-5 results in Section 6.

### 3.1 Evaluation setup

**Cross-domain retrieval.** We measure retrieval on the field’s own protocol before measuring tracking. Encoders are finetuned on MARKET-1501, CUHK03, and MSMT17, and each finetuned encoder is evaluated single-query on all three corpora, with MSMT17 evaluated under the MARKET-1501-style protocol, reporting Rank-1 accuracy and mean average precision (mAP). Training corpus and evaluation corpus are crossed, so each cell of the resulting matrix is either in-domain (train and test from the same

corpus) or cross-domain. Six of the eight architectures enter this comparison; the two OSNet variants are excluded. Cell values are means over the six architectures.

**Tracking.** We use BYTETRACK (Zhang et al. 2022) as the tracker: it predicts each track’s next position with a Kalman motion model and associates detections to tracks in two stages using spatial overlap (intersection-over-union, IoU) and appearance similarity. We plug the encoder pool into the appearance branch of the first stage and fix the IoU/appearance mixing weight at  $\alpha = 0.5$  for every run.  $\alpha = 0.5$  gives the appearance signal equal weight with motion, so the comparison runs in a regime where the manipulated component can affect the outcome.

**Detector.** We pair BYTETRACK with YOLOX (Ge et al. 2021) for detection. YOLOX is an anchor-free single-stage object detector, and the original BYTETRACK release includes joint YOLOX weights tuned for pedestrian MOT, which we use unchanged. We cache the detections once and reuse them bit-identically across every encoder run, so only the appearance feature can change tracking output between conditions.

**Benchmarks.** We evaluate the encoder pool on three multi-object tracking (MOT) benchmarks for pedestrian footage. MOT20 (Dendorfer et al. 2020b) contains very dense pedestrian crowds in transit halls and public squares. CROWDTRACK (Fu et al. 2025) extends the dense-crowd setting to a wider range of urban scenes. PERSONPATH22 (Shuai et al. 2022) contributes longer surveillance-style videos with persistent identities. The three benchmarks cover the deployment conditions ReID is claimed to support, so a consistent pattern across them cannot reduce to one benchmark’s idiosyncrasies.

**Metrics.** We report IDF1 (Ristani et al. 2016) as the standard identity-preservation metric for these benchmarks. IDF1 matches every predicted track one-to-one with a ground-truth identity and reports the harmonic mean of identity precision and identity recall, so it captures the identity persistence ReID training is supposed to add over a generic encoder. IDF1 has displaced detection-weighted metrics such as MOTA for this purpose because MOTA is dominated by detection errors and moves little under association changes (Ristani et al. 2016). The identity splits and merges

that lower IDF1 are the error types that defeat the uses in Claim 1: a split identity is a person lost across the camera network, and a merged identity is two people conflated into one.

**Statistical tests.** We pair every FINETUNED condition against its matched PRETRAINED condition by architecture and sequence, so the within-architecture difference is the unit of analysis. We test paired FINETUNED-vs-PRETRAINED differences with the Wilcoxon signed-rank test, report 95% confidence intervals from a percentile bootstrap with  $B = 10,000$  resamples, and certify equivalence with two one-sided tests (TOST) (Lakens 2017) at a  $\pm 2$  pp IDF1 margin. Published work that adds an appearance-association branch to a tracker reports IDF1 gains from one to ten points (Zhang et al. 2022), so a  $\pm 2$  pp margin sits at the low end of the smallest such gains. The repeated family is the eight within-architecture comparisons, one per architecture, so eight tests run on the same benchmark data and the chance of one spurious per-architecture finding grows with each; the Holm step-down procedure controls the family-wise error across them.

### 3.2 Interpretability

The tracking evaluation reports whether ReID finetuning improves downstream identity preservation, not why. Figure 3 summarises the three interpretability probes we run on the frozen encoder pool: attribute probing, pair-based identity verification, and saliency attribution. A probe can decode attributes that play no role in identity matching (Belinkov 2021), so we run all three rather than rely on any single signal. All three use the MARKET-1501 attribute annotations of Lin et al. (2019). The FINETUNED pool contains the five single-corpus variants of Section 3, four of which were never trained on MARKET-1501, so decoding MARKET-1501 attributes from those encoders is a cross-corpus measurement rather than a within-benchmark one, and the colour result cannot reduce to one benchmark’s collection conditions.

**Attribute probing.** We use attribute probing to identify which pedestrian attributes are linearly decodable from the frozen features. A linear probe is a logistic regression fit on the frozen encoder features with the backbone held fixed, so its accuracy measures how directly an attribute is recoverable from the representation rather than how well the whole network could be trained to predict it. For each attribute annotated by Lin et al. (2019), we fit this probe and report chance-normalised accuracy,  $(\text{acc} - \text{acc}_{\text{majority}})/(1 - \text{acc}_{\text{majority}})$ , so 0 is the majority-class baseline and 1 is perfect. The attribute set covers clothing colour (up.colour, down.colour) and non-colour controls (hat, backpack, bag, gender, age).<sup>3</sup> We split train and test by identity using GroupShuffleSplit so no identity appears in both partitions, run five seeds per probe, and pair encoders within each architecture, then compute within-architecture FINETUNED-minus-PRETRAINED

<sup>3</sup>Gender and age in these annotations are inferred from appearance by Lin et al. (2019). We feel that this approach is speculative and we leave those results out of this paper.

deltas per seed with 95% bootstrap CIs.

**Pair-based identity verification.** We classify same-identity and different-identity pairs of crops by cosine distance on the frozen features, choosing an optimal threshold per encoder and seed, to measure how often the two crops in a cross-identity error share an attribute. Different-identity pairs are sampled under five regimes: a random regime, and four hard-negative regimes that constrain the two crops to share one attribute (hat, backpack, lower-body colour, upper-body colour). The same pairs are reused across encoders within a seed, so comparisons are paired. Per regime we report attribute-sharing lift on the verification false positives. With  $p_{\text{FP}}(a)$  the rate at which the two crops in a false-positive pair share attribute  $a$  and  $p_{\text{TN}}(a)$  the same rate among true negatives,

$$\text{lift}(a) = \frac{p_{\text{FP}}(a)}{p_{\text{TN}}(a)},$$

so a lift of 1 means errors share  $a$  at the chance rate and above 1 means they are enriched for  $a$ .

**Saliency attribution.** In addition to the probes, we localise where in the crop the matching decision is concentrated. For the convolutional encoders and SwinV2-B we apply GradCAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017), which weights the final spatial feature map by gradients of the matching score to produce a coarse heatmap over the input. This requires positions in the feature map to correspond to spatial regions of the input, which holds for the stride-32 CNN feature grids and the stride-32 windowed-attention output of SwinV2-B but not for the global self-attention in the final block of ViT-B/16 (Dosovitskiy et al. 2021) and DINOv3 ViT-B. For these we apply the gradient-weighted attention rollout of Chefer, Gur, and Wolf (2021), which composes per-block attention contributions  $\bar{A} = \text{ReLU}(A \odot \partial \mathcal{L} / \partial A)$  multiplicatively across blocks and extracts the classification (CLS) token row over the  $16 \times 8$  patch grid. We report the fraction of class activation map (CAM) argmax positions falling inside the torso bounding box, defined as the middle vertical band of the  $256 \times 128$  pedestrian crop, and test against the matched PRETRAINED pair with a Fisher exact one-sided test.

## 4 Failure 1: ReID Models Do Not Transfer Across Domains

We find that encoders finetuned on one ReID corpus lose most of their retrieval accuracy on every other corpus. Figure 4 reports mean Rank-1 accuracy over six architectures for every pairing of training and evaluation corpus, under the retrieval protocol of Section 3.1. Cross-domain, Rank-1 falls 2.7 to 81.1 pp below each corpus’s in-domain value. The ReID literature has reported this degradation and continued to recommend the models for deployment (Song et al. 2019; Lan, Zhu, and Gong 2019; Lin, Li, and Kot 2020). We argue that this recommendation is dubious on the field’s own evidence. If Rank-1 drops between two academic benchmarks, a comparable drop is expected between a benchmark and the unseen cameras of a deployment site.

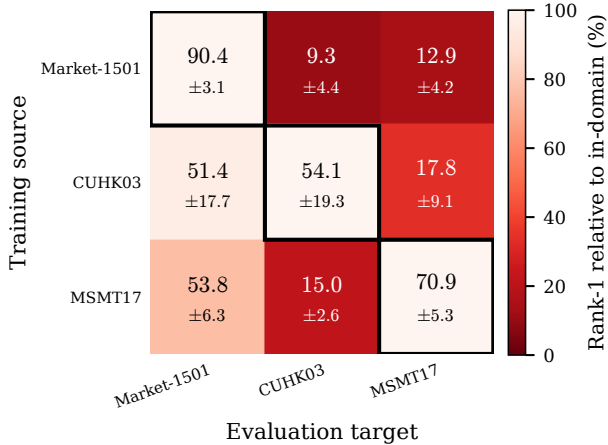


Figure 4: **ReID retrieval does not transfer across corpora.** Mean Rank-1 (%)  $\pm$  standard deviation over six architectures; rows are the training corpus, columns the evaluation corpus. Cell colour encodes Rank-1 relative to the same row’s in-domain value, so darker red marks stronger degradation. Outlined diagonal cells are in-domain.

## 5 Failure 2: ReID Finetuning Does Not Improve Tracking

| Encoder      | PRETRAINED | FINETUNED | Equiv. |
|--------------|------------|-----------|--------|
| RESNET-50    | 25.75      | 25.97     | ✓      |
| DENSENET-121 | 26.05      | 26.46     | ✓      |
| OSNET        | 26.09      | 26.72     | ✓      |
| OSNET-AIN    | 26.12      | 26.63     | ✓      |
| MLFN         | 26.30      | 26.10     | ✓      |
| VIT-B/16     | 31.99      | 33.17     | ✓      |
| SWINV2-B     | 32.29      | 33.30     | ✓      |
| DINOv3 ViT-B | 30.17      | 31.77     | ✓      |

Table 2: **ReID finetuning does not improve tracking.** Mean IDF1 (%) per architecture under PRETRAINED and FINETUNED (mean across the five single-corpus variants), on MOT20, CROWDTRACK, and PERSONPATH22. ✓: the within-architecture difference is within  $\pm 2$  pp under two one-sided tests. Macro mean across the three benchmarks  $\Delta = +0.44$  pp (95% CI  $[+0.23, +0.65]$ ).

Following the methodology in Section 3, we find that ReID-finetuned encoders do not improve tracking IDF1 over their PRETRAINED counterparts. Across the three benchmarks MOT20, CROWDTRACK, and PERSONPATH22, the macro mean of the within-architecture FINETUNED-minus-PRETRAINED difference is  $+0.44$  pp (95% CI  $[+0.23, +0.65]$ ), statistically equivalent to zero at the  $\pm 2$  pp margin (Holm-corrected  $p < 0.001$  on every architecture). This shows that **ReID datasets provide no empirical benefit on the tracking task they are claimed to enable.**

For each of the eight architectures, we ran BYTE-

TRACK with six encoders on each of the 24 sequences: a PRETRAINED encoder, and that same encoder additionally finetuned on each of five ReID corpora (MARKET-1501, CUHK03, CUHK-SYSU, ETHZ, and GRID). Architecture, loss, augmentation pipeline, and the cached YOLOX detection set were held constant across runs, and the IoU/appearance mixing weight stayed at  $\alpha = 0.5$ , so the only thing that varied between any two runs of an architecture was the encoder’s training data.

Table 2 reports the per-architecture PRETRAINED mean and the FINETUNED mean across the five single-corpus variants. The within-architecture FINETUNED-vs-PRETRAINED difference falls between  $-0.20$  and  $+1.60$  percentage points, every value inside the  $\pm 2$  pp margin. Pooled over all paired sequences the difference is  $+0.67$  pp, and the macro mean across the three benchmarks is  $+0.44$  pp (95% CI  $[+0.23, +0.65]$ ), with per-benchmark values of  $-0.44$  pp on MOT20 ( $n = 160$ ),  $+1.48$  pp on CROWDTRACK ( $n = 400$ ), and  $+0.29$  pp on PERSONPATH22 ( $n = 400$ ). Two one-sided tests reject any true effect larger than  $\pm 2$  pp at  $p < 0.001$  (Lakens 2017). None were statistically significant after Holm-Bonferroni correction; all of them pass the equivalence test at  $\pm 2$  pp.

## 6 Failure 3: More ReID Data Does Not Improve Tracking

| Condition  | CNNs  | Transformers |
|------------|-------|--------------|
| PRETRAINED | 26.06 | 31.48        |
| FINETUNED  | 26.38 | 32.75        |
| COMBINED-5 | 26.14 | 25.76        |

Table 3: **More ReID training data does not improve tracking.** Mean IDF1 (%) by architecture family. COMBINED-5 pools the five corpora into one training set (12,529 identities); FINETUNED averages the five single-corpus variants.

One might think: if we collect more surveillance data, the resulting tracking would improve. We find that scaling the ReID training data does not change the Failure 2 result. We pooled all five corpora into a single training set of 12,529 identities, which we call COMBINED-5, and reran the tracking comparison of Section 5 unchanged. Table 3 reports mean IDF1 by architecture family across the three training conditions. For the five convolutional architectures ( $n = 120$ ) the gap to the PRETRAINED baseline is  $+0.07$  pp (95% CI  $[-0.13, +0.27]$ ), inside the equivalence margin and indistinguishable from zero. For the architectures that converged, the equivalence holds at COMBINED-5’s 12,529 identities just as at MARKET-1501’s 751.

## 7 Failure 4: Finetuning ReID Encodes Stereotypes, Particularly About Clothing

In this section, we ask which features ReID models encode when they are finetuned on surveillance ReID datasets. To understand this, we run the attribute probe of Section 3.2

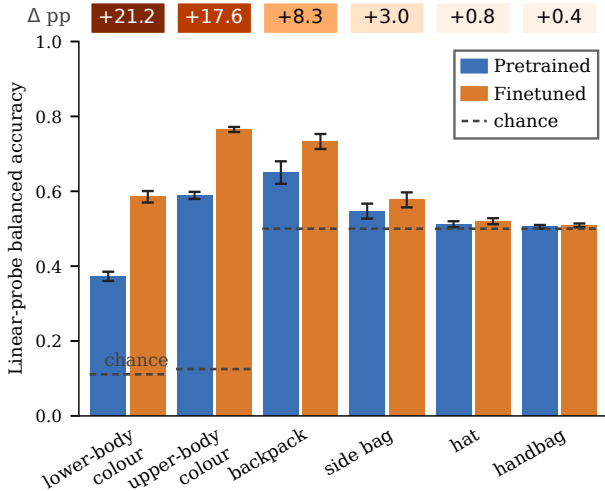


Figure 5: **Clothing-colour decodability rises after ReID finetuning.** Chance-normalised linear-probe accuracy on frozen PRETRAINED and FINETUNED representations on MARKET-1501, per architecture, for clothing-colour attributes (up\_colour, down\_colour) and non-colour controls (hat, backpack, bag). Bars are the mean over five seeds; error bars are 95% bootstrap confidence intervals. FINETUNED bars pool the five single-corpus ReID variants within architecture.

on the frozen encoder pool, paired by seed; Figure 5 shows chance-normalised accuracy per architecture and attribute.

The results show that colour decodability rises: down\_colour by +0.21 and up\_colour by +0.18 on average across the pool. This lift appears in every architecture, every 95% CI excludes zero, and  $p_{\text{Holm}} < 0.001$ . The other cues rise slightly (backpack +0.08, bag +0.03, hat +0.01), but none of these increases is significant after Holm correction in any architecture.

We find that finetuning amplifies clothing colour more than any other attribute. *From this experiment, we observe that ReID models rely on stereotypical cues like clothing colour to track people, and these cues are amplified when they are finetuned.*

## 8 Failure 5: ReID Encoders Confuse Different People In The Same Colours

We hypothesise that a representation organised around clothing colour places different people who dress alike close together, and crowded public footage contains many such pairs. In this section, we ask whether ReID models confuse different people who wear the same clothing colour. To understand this, we run the pair-based verification of Section 3.2 on the same eight architectures and report the colour-sharing lift on its false positives; Figure 6 shows the per-architecture lift.

The results show that FINETUNED false positives pair people who wear the same clothing colour far more often

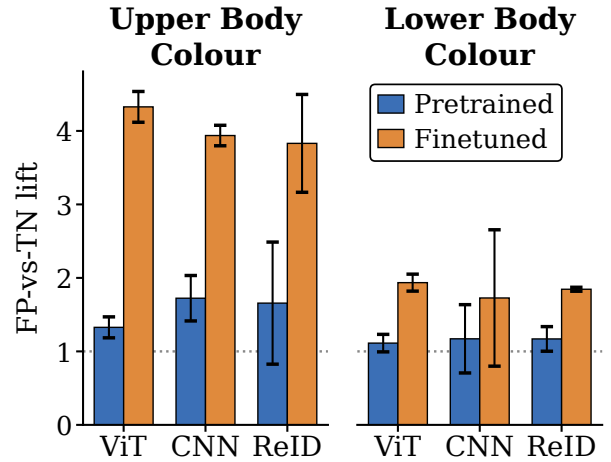


Figure 6: **False-positive pairs share clothing colour at three to five times the rate of true-negative pairs.** False-positive vs. true-negative upper-body-colour-sharing lift per architecture on MARKET-1501. A lift of 1 (vertical dashed line) means false-positive and true-negative pairs share colour at the same rate; a lift of 4 means false-positive pairs share colour four times as often as true-negative pairs. Bars are the mean over five seeds; error bars are 95% bootstrap confidence intervals. FINETUNED bars pool the five single-corpus ReID variants within architecture.

than PRETRAINED false positives do, and that non-colour attributes show no such pattern. For 8 of 8 architectures, the false-positive pairs of the FINETUNED probe share upper-body colour 3 to 5 times as often as its true-negative pairs ( $p_{\text{Holm}} < 0.001$ ). Under PRETRAINED, the same ratio never exceeds 2.1. The transformer encoders sit at the top of that range, with FINETUNED upper-body-colour lifts of 4.72 for ViT-B/16, 4.52 for SWINv2-B, and 4.47 for DINOv3 ViT-B. Lower-body colour shows the same 8-of-8 pattern. By contrast, the non-colour attributes show no comparable lift under either training regime: hat lift stays inside [0.9, 1.1] in 7 of 8 architectures, and the FINETUNED-minus-PRETRAINED difference on backpack lift stays within 0.3 at every architecture.

We find that FINETUNED models rely on clothing colour to match people, and so cannot distinguish two people who dress alike. *From this experiment, we observe that ReID models tend to confuse different people who share the same clothing colour, producing false positives.*

## 9 Failure 6: ReID Saliency Localises On Clothing

In this section, we ask which parts of a person’s image ReID models use to match identities. To understand this, we run the saliency attribution of Section 3.2 on the matched PRETRAINED and FINETUNED pairs, scoring for each MARKET-1501 crop whether the highest-attribution pixel lies inside the torso strip. Figure 7 shows the maps for

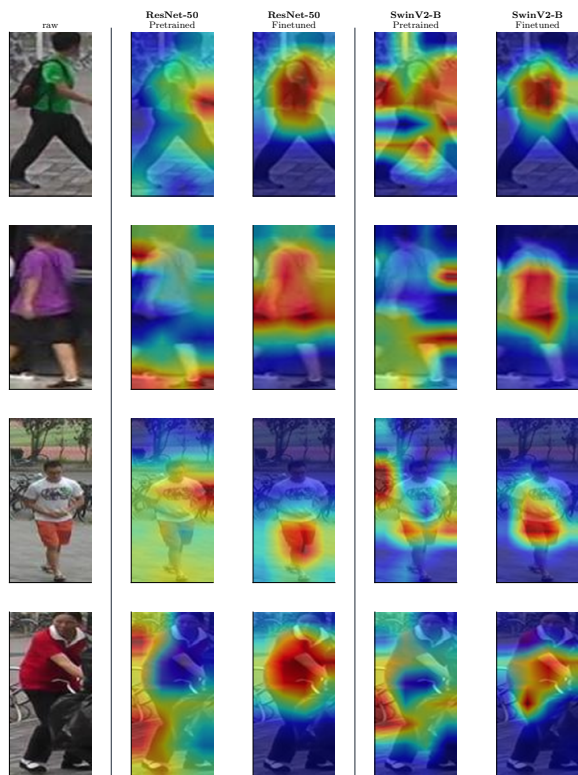


Figure 7: **Saliency localises on clothing under FINE-TUNED encoders.** Saliency maps for four MARKET-1501 gallery crops under the PRETRAINED and FINETUNED variants of RESNET-50 and SWINV2-B, computed by GRAD-CAM on the final spatial feature map. The query image is fixed within each row and only the encoder weights vary between columns. Under FINETUNED the attribution localises on the locally most distinctive shortcut feature: the colour-saturated garment in rows 2 and 4, the backpack region in row 1, and the red shorts (not the white t-shirt) in row 3.

#### RESNET-50 and SWINV2-B.

The results show that FINETUNED attribution localises on a few shortcut regions, the clothing colour and the backpack, whereas identity cues are spread over the whole body (Figure 7). The highest-attribution pixel lies inside the torso strip on 63 of 84 FINETUNED maps (75.0%) and on 46 of 84 PRETRAINED maps (54.8%), pooling 7 matched architecture pairs  $\times$  12 maps per condition; Fisher’s exact one-sided test gives  $p = 4.7 \times 10^{-3}$ , odds ratio 2.48. In the PRETRAINED columns of Figure 7 the attribution is spread over the head, limbs, and surrounding scene. In the FINETUNED columns it localises on the locally most distinctive shortcut feature: the colour-saturated garment in rows 2 and 4, the backpack region in row 1, and the red shorts, not the white t-shirt, in row 3.

We find that as models are finetuned, attribution shifts off the head and limbs and onto the clothing. *From this experiment, we observe that ReID models rely on clothing regions, not on the face, posture, or body shape, to match identities.*

## 10 Discussion

Person re-identification research is built on surveillance footage recorded without the consent of the people in it: recordings of students, shoppers, and commuters across university and commercial camera networks. The stated justification for collecting and keeping this footage is that the models it trains deliver identity matching in deployment. In this paper, we tested this justification and provide two major findings.

**Finding 1: ReID finetuning does not improve tracking performance in downstream evaluation due to cross-domain degradation.** (Sections 4–6). Encoders finetuned on one ReID corpus fail to generalise: they lose most of their retrieval accuracy on every other corpus (Section 4). On tracking, FINETUNED encoders gain only +0.44 pp in IDF1 over the matched PRETRAINED baseline, a difference statistically equivalent to zero, across eight architectures and three benchmarks (Section 5). Training on all five corpora at once, 12,529 identities in total, shrinks the gain further, to +0.07 pp (Section 6). When appearance and motion are weighted equally in the tracker, pooled IDF1 for every encoder is lower than for motion alone (Section 5). Finetuning on ReID corpora does not improve tracking on the corpora, the architectures, or the benchmarks the field itself uses.

**Finding 2: The representations produced by ReID finetuning encode stereotypes, particularly about clothing** (Sections 7–9). All three interpretability experiments find the same reliance on clothing colour. After finetuning, clothing colour becomes the most decodable attribute in the representations (Section 7). Different people who wear the same colours are falsely matched as the same person (Section 8). Saliency for the matching decision localises on clothing rather than on the person (Section 9). Non-colour attributes such as hat, backpack, and bag show minor but insignificant reliances in all three experiments. This colour shortcut follows from how the corpora are collected: each person is recorded at one site over a short window, in the same clothes, so clothing colour predicts identity within the corpus better than any cue specific to the person. The shortcut also explains Finding 1: clothing colour stops predicting identity as soon as the site, cameras, or clothing change, so the finetuned encoders neither transfer nor improve tracking. The benefit claimed to justify collecting this footage is absent, and its privacy cost remains.

**ReID for mass surveillance is a flawed premise.** ReID benchmarks are recorded by fixed surveillance cameras, the same hardware that supplies a deployed tracking system with its input. A fixed camera recording a person at ten metres or more returns a body crop that is small, low in resolution, and frequently occluded or motion-blurred. Such low resolution leaves facial detail and fine texture unresolvable, and coarse clothing colour is the appearance cue that remains. This colour dependence is what the probes in Sections 7 through 9 recovered, and it is why ReID training produced no tracking gain in Section 5: training cannot add an identity signal that the recorded input never carried. The limit originates in the recording conditions, and enlarging the backbone or the training corpus does not remove it. The decision to attempt pedestrian identification on such inputs

is therefore a policy choice that improvements in models do not change.

**ReID for mass surveillance constitutes security theatre and data colonialism.** ReID datasets do not deliver the tracking improvement that would justify their collection (Section 5). They are collected, curated, and deployed anyway, a pattern we characterise as *security theatre*. The deployments project an image of safer public space while producing false positives on colour-matched pairs (Section 8). The footage is collected without consent and framed as a contribution to research progress, and the models trained on it are then used to surveil the very publics from which it was taken. We argue that this dynamic reflects data colonialism (Couldry and Mejias 2019): public life appropriated as raw material for research, with the capacity for identification held entirely by the collectors. The people in the footage carry the privacy cost of these collections, and the measurements show no benefit in return.

**The ethical arguments should be in the foreground.** We argue that algorithms which surveil public space and harm marginalised communities should not be built, whatever their accuracy. Our rationale is grounded in consent and disproportionate harm, and the experiments add that the promised capability is also missing. Mass surveillance is a sociotechnical problem that our paper cannot resolve, and addressing it requires sustained engagement between researchers, policymakers, and the communities subject to these systems. Our contribution to that engagement is the empirical finding that training on surveillance corpora neither improves tracking nor yields the identity representations that would justify collecting them. Even if future models were to deliver the missing tracking gain, the footage was still recorded without consent. ReID datasets such as DUKEMTMC-REID and MARKET-1501 consist of footage of students and pedestrians who had no opportunity to refuse when the campus cameras recorded them. The surveillance systems built from that footage bear disproportionately on already-policed populations (Browne 2015). For researchers, surveillance use is inherent to these corpora from the moment of collection, and a dataset assembled from non-consenting pedestrians cannot be retroactively justified by a tracking benefit the experiments do not show. For policymakers, our results weaken the public-safety case for collecting non-consensual surveillance footage, as we show that collecting such datasets does not improve tracking performance.

**Conclusion.** Encoders finetuned on ReID datasets lose most of their retrieval accuracy outside their training corpus, and across eight architectures and three pedestrian-tracking benchmarks they do not improve tracking over the matched PRETRAINED baseline; increasing the training data does not change this. Instead, when these models are finetuned, they tend to rely on superficial cues, especially clothing colour and accessories: colour becomes more decodable, false positives cluster on colour-matched pairs, and saliency localises on clothing. These false positives amplify the harm for marginalised people at deployment. These datasets were built on the promise of ‘AI progress’ on surveillance and tracking; however, our experiments show that training on

these datasets does not provide any empirical benefits. What remains is footage of students, shoppers, and commuters recorded without their consent, built into surveillance infrastructure that targets vulnerable and marginalised people.

## Acknowledgments

We thank our anonymous reviewers and base.camp team for their feedback. The work of Trustworthy AI lab is funded under the Excellence Strategy of the German Federal Government and States.

## References

- Alain, G.; and Bengio, Y. 2018. Understanding intermediate layers using linear classifier probes. *ArXiv:1610.01644 [stat]*.
- Arcas, B. A. y.; Todorov, A.; and Mitchell, M. 2018. Do algorithms reveal sexual orientation or just expose our stereotypes?
- Belinkov, Y. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *ArXiv:2102.12452 [cs]*.
- Browne, S. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.
- Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. In *CVPR*, 2109–2118.
- Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.
- Couldry, N.; and Mejias, U. A. 2019. Data colonialism: Rethinking big data’s relation to the contemporary subject. *Television & new media*, 20(4): 336–349.
- Dendorfer, P.; Ošep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; and Leal-Taixé, L. 2020a. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *ArXiv:2010.07548 [cs]*.
- Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020b. MOT20: A benchmark for multi object tracking in crowded scenes. *ArXiv:2003.09003 [cs]*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. ISSN: 1063-6919.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houslsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv:2010.11929 [cs]*.
- Fu, T.; Chen, Y.; Chen, Z.; Zhao, M.; Li, B.; and Xue, X. 2025. CrowdTrack: A Benchmark for Difficult Multiple Pedestrian Tracking in Real Scenarios. *arXiv preprint arXiv:2507.02479*.

- Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.
- Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11): 665–673. ArXiv:2004.07780 [cs].
- Harvey, A. 2021. Exposing.ai: The WILDTRACK Seven-Camera HD Dataset.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.
- Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Kalluri, P. 2020. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815): 169.
- Kalluri, P. R.; Agnew, W.; Cheng, M.; Owens, K.; Soldaini, L.; and Birhane, A. 2023. The surveillance AI pipeline. *arXiv preprint arXiv:2309.15084*.
- Lakens, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4): 355–362.
- Lan, X.; Zhu, X.; and Gong, S. 2019. Universal person re-identification. *arXiv preprint arXiv:1907.09511*.
- Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.
- Lin, S.; Li, C.-T.; and Kot, A. C. 2020. Multi-domain adversarial feature generalization for person re-identification. *IEEE Transactions on Image Processing*, 30: 1596–1607.
- Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving Person Re-identification by Attribute and Identity Learning. *Pattern Recognition*, 95: 151–161. ArXiv:1703.07220 [cs].
- Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.
- Loy, C. C.; Xiang, T.; and Gong, S. 2009. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1988–1995. IEEE.
- Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; and Gu, J. 2019. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. *arXiv preprint arXiv:1906.08332*.
- Peng, K.; Mathur, A.; and Narayanan, A. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. ArXiv:2108.02922 [cs].
- Prabhu, V. U.; and Birhane, A. 2020. Large image datasets: A pyrrhic win for computer vision? ArXiv:2006.16923 [cs].
- Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? ArXiv:1902.10811 [cs].
- Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, 17–35. Springer.
- Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW): 144:1–144:33.
- Schwartz, W. R.; and Davis, L. S. 2009. Learning discriminative appearance-based models using partial least squares. In *2009 XXII Brazilian symposium on computer graphics and image processing*, 322–329. IEEE.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Shuai, B.; Bergamo, A.; Buechler, U.; Berneshawi, A.; Boden, A.; and Tighe, J. 2022. Large scale real-world multi-person tracking. In *European Conference on Computer Vision*, 504–521. Springer.
- Siméoni, O.; Vo, H. V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.
- Song, J.; Yang, Y.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2019. Generalizable Person Re-identification by Domain-Invariant Mapping Network. In *CVPR*, 719–728.
- Stark, L. 2019. Facial recognition is the plutonium of AI. *XRDS*, 25(3): 50–55.
- Stark, L.; and Hutson, J. 2022. Physiognomic Artificial Intelligence. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 32(4): 922.
- Wang, X. 2013. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1): 3–19.
- Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 79–88.
- Whittaker, M.; Crawford, K.; Dobbe, R.; Fried, G.; Kaziunas, E.; Mathur, V.; West, S. M.; Richardson, R.; Schultz, J.; Schwartz, O.; et al. 2018. AI now report 2018.
- Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. ArXiv:1703.07402 [cs].
- Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2016. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2): 4.
- Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. ArXiv:2001.04193 [cs].
- Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. ByteTrack:

Multi-Object Tracking by Associating Every Detection Box. ArXiv:2110.06864 [cs].

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. In *ICCV*, 3754–3762.

Zhou, K.; and Xiang, T. 2019. Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch. ArXiv:1910.10093 [cs].

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-Scale Feature Learning for Person Re-Identification. In *ICCV*, 3702–3712.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2021. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5056–5069.