Can a US-Trained Object Detector Transfer to German Streets? A Controlled 2×3 Zero-Shot, US-Fine-Tuned, and In-Domain Study

Ian Constantijn Ronk*

July 2026

Abstract

Public training data for driving-scene object detection is geographically concentrated, so a practitioner with US data but a European deployment faces a concrete question: does a US-trained detector transfer, and how much performance is left on the table? A 2024 course project asked this — it trained a YOLOv3 on US autonomous-driving data, evaluated on European footage, saw a large drop, and concluded that European data "transfers poorly." That conclusion was not identifiable: the backbone was frozen and only the head was trained for six epochs (two of four classes at zero precision *on the US set as well*), only precision was reported, the European images were pre-downscaled to the US resolution, and there was no held-out split — so "fails to transfer" could not be separated from "never learned to detect." We re-run the study as a controlled 2×3 — {zero-shot COCO, US-fine-tuned, EU-fine-tuned} $\times$ {held-out US-test, held-out target-test} — with resolution-matched inputs, seeded held-out splits, multi-seed fine-tuning, and COCO-style mAP. The target is KITTI (Karlsruhe, Germany), a substitute for the unreachable Swedish ZOD, so the honest scope is *US-urban* $\rightarrow$ *German-driving*, not "Europe." Three findings. (a) A region-neutral COCO detector already works on both regions, and the German target is in fact the *easier* set, so the original's raw drop was mostly never-trained-plus-resolution, not geography. (b) Fine-tuning on US data does not help the held-out German target (shared-class mAP gain $+0.001$), whereas fine-tuning on in-domain German data nearly doubles it ($+0.153$): for German deployment, US fine-tuning $\approx$ off-the-shelf COCO, and you need in-domain data. (c) Read as the 2×2 interaction term of {zero-shot, fine-tuned}$\times$region — a difference-in-differences — fine-tuning specialises toward its own dataset: the US fine-tune shifts the gap toward the US ($\Delta = +0.077 \pm 0.007$ shared-class mAP, 3/3 seeds; $+0.073$ on a YOLOv3 architecture check), while the in-domain German fine-tune shifts it the *other* way, and by *more* ($\Delta = -0.169$; -0.174 on the YOLOv3 check); the sign survives a near-threshold-free re-evaluation. Because there is one dataset per region, "US-favoring" is collinear with "Udacity-favoring": this is symmetric narrow-fine-tune specialisation, not evidence of a specifically *US* bias. None of this is a new phenomenon: it re-confirms textbook domain shift and the source-only/oracle gap of domain-adaptive detection. The contribution is methodological — a reproducibility case study that corrects a specific "catastrophic transfer" claim, plus a clean quantification with an in-domain control that makes the residual gap interpretable.

*Independent researcher, Amsterdam. ianronk0@gmail.com. This preprint is an independent PyTorch reimplementation and methodological correction of a 2024 Bocconi University computer-vision course project (20600), originally carried out with Gianfranco Salocchi, whose contribution to the original study is gratefully acknowledged. He is not a co-author of this preprint.

1 Introduction

Public training data for driving-scene object detection is geographically lopsided. The large autonomous-driving (AV) benchmarks are recorded at a handful of collection sites — Waymo and nuScenes in US and Singapore street grids (Sun et al., 2020; Caesar et al., 2020), BDD100K across US roads (Yu et al., 2020), KITTI and Cityscapes on German streets (Geiger et al., 2012; Cordts et al., 2016), ZOD in Sweden (Alibeigi et al., 2023). A practitioner who has US data but wants to deploy in Europe therefore faces a concrete, practical question: does a US-trained detector transfer to European streets, and how much performance is left on the table? This is a transfer question first — a deployment decision — and only secondarily a question about *why* any gap exists. The standard "why" is dataset bias: Torralba and Efros (2011) showed that benchmarks carry a recognizable "signature" a model latches onto, so a detector can look strong on its own distribution and weaker on another. We treat that mechanism as downstream of the transfer measurement, not as the headline.

We are explicit up front that the transfer *phenomenon* is not a discovery. Cross-geography degradation for driving detectors is textbook: it is the exact axis of Wang et al. (2020), "Train in Germany, Test in the USA," and a special case of the source-only/oracle gap that domain-adaptive detection has studied for years (Chen et al., 2018; Hasan et al., 2021). What this paper contributes is not the existence of a gap but two controlled, falsifiable pieces around it: (i) a reproducibility case study that re-analyses a specific prior "catastrophic transfer" claim under fair training, showing it was a training/resolution artifact rather than geography; and (ii) a clean 2×3 quantification with an *in-domain* fine-tuning arm, which is what lets us say how much of the residual gap US data can and cannot close.

The prior study and its ambiguity. This paper grew out of a 2024 university course project that asked exactly the transfer question. It trained a YOLOv3 (Redmon and Farhadi, 2018) on a US AV dataset, evaluated on European dashcam footage, saw a large drop, and reported that European data transfers poorly. The intent was right; the implementation could not support the conclusion. The detector was never trained into a working state — its Darknet-53 backbone was frozen and only the head was trained, for six epochs, with the authors noting they "did not have enough data to train." Two of the four classes scored 0.00 precision *on the US data itself*; the paper reported precision only (a near-useless summary for a model that barely fires); the European images were pre-downscaled to the US resolution before inference; and there was no held-out test split. A model that cannot detect on its own training distribution will, trivially, also fail on a new one. The original could not tell "fails to transfer to Europe" apart from "never learned to detect." This is a concrete instance of a documented pitfall: under-trained source-only baselines systematically *overstate* the domain gap (Kay et al., 2025).

Scope: German-driving, not "Europe." We are equally explicit about scope. Our transfer target is KITTI (Geiger et al., 2012), recorded in and around Karlsruhe, Germany. It is a substitute for the originally intended Zenseact Open Dataset (ZOD; Swedish footage) (Alibeigi et al., 2023), whose only ungated download route was unreachable in our environment. "Europe" here therefore means *one German city*, and every claim below is scoped to *US-urban → German-driving (KITTI)*, never to Europe at large.

Design. We keep the question and discard the design. We run a 2×3 experiment: three model conditions — a zero-shot COCO-pretrained detector (region-neutral, seen neither region), the same detector fine-tuned on US data, and the same detector fine-tuned on in-domain German (KITTI) data — each evaluated on a held-out US test set and a held-out German test set. The zero-shot arm is the region-neutral control; the EU-fine-tuned arm is the in-domain (oracle-style) control. Two questions fall out. (Q1) *Can US data substitute for in-domain data?* Compare, on

the German target, the gain from US fine-tuning against the gain from in-domain fine-tuning. (Q2) *Does fine-tuning specialise toward its own training region?* Take the 2×2 sub-grid {zero-shot, fine-tuned}×{US, EU} and read its interaction term — a difference-in-differences on the US→EU gap — for the US-fine-tuned *and* the in-domain-fine-tuned model, so the two directions can be compared.

Contributions.

- **A reproducibility case study of a specific prior claim.** Five confounds in the original study (frozen/under-trained model, a resolution confound, precision-only reporting, the train-versus-transfer confound, and no holdout) are each mapped to an explicit fix (§3). Under fair training the raw "catastrophic" drop largely evaporates and the German target is in fact the *easier* set — a signed correction, not a re-confirmation, in the reproducibility lineage of Musgrave et al. (2020) and the under-trained-baseline finding of Kay et al. (2025). The corrected target is the authors' own unpublished course project, so the contribution is methodological (a worked correction others can reuse), not a critique of a published result.
- **A deployment corollary that confirms known domain-adaptation results.** On the held-out German target, US fine-tuning buys essentially nothing over off-the-shelf COCO (+0.001 shared-class mAP), while in-domain fine-tuning nearly doubles performance (+0.153). This is the practitioner's version of the source-only/oracle gap of Chen et al. (2018) and Hasan et al. (2021): for German deployment you need German data (§6).
- **A clean interaction-term quantification.** A region-neutral zero-shot control plus a 2×2 factorial interaction term (a difference-in-differences, reported with seeded CIs on the primary model and reproduced on a single-seed second architecture) isolates how fine-tuning moves the US→EU gap net of the static scene/resolution confound (§5).
- **Honest scoping of the mechanism.** The in-domain arm shows the German target had ample headroom, so US fine-tuning's failure there is substantive, not a ceiling artifact. It also supplies the symmetric control: applying the *same* interaction-term rule to the in-domain fine-tune yields a larger, opposite Δ (-0.169 vs. $+0.077$), so the mechanism is narrow-fine-tune specialisation — each fine-tune favours its own dataset — and, with one dataset per region, is not separable from a specifically *US* bias (Kumar et al., 2022). We report both directions and do not overclaim the region-specific reading (§7–§8).

The headline effect is small, and the phenomenon is old. The contribution is turning a non-identified "Europe transfers poorly" into a controlled, signed, cross-checked statement about what US fine-tuning does — and does not do — to a German target.

2 Related Work

Dataset bias and shortcut learning. Torralba and Efros (2011) is the canonical statement that vision benchmarks are biased: a classifier can identify which dataset an image came from, and cross-dataset generalization is far worse than within-dataset performance. The same phenomenon, seen from the optimization side, is *shortcut learning* (Geirhos et al., 2020): networks preferentially fit dataset-specific cues that do not generalize. Its *geographic* face is documented too: DeVries et al. (2019) show object recognizers work far worse on images from under-represented regions, and Shankar et al. (2017) trace this to the geographic concentration of open datasets — the same lopsidedness our source data exhibits. Our framing — that fine-tuning on one region may buy in-domain accuracy at a transfer cost — is a detection-flavoured instance of this, and we are careful (§7) not to claim the cost is necessarily *region*-specific rather than a generic narrow-fine-tune effect.

Cross-geography AV transfer. The most directly comparable work is Wang et al. (2020), "Train in Germany, Test in the USA," which studies exactly the KITTI(Germany)↔USA axis. It works in *3D from LiDAR*, finds the cross-geography gap is largely a correctable car-*size* artifact, and is therefore *more optimistic* than we are on the same axis. We differ in modality (2D RGB detection), in adding an explicit in-domain fine-tuning control, and in using the setup to re-analyse a prior claim as a reproducibility case study rather than to propose an adaptation method. Geographic distribution shift is also a headline case in the WILDS benchmark suite (Koh et al., 2021).

Domain-adaptive object detection (DAOD) and the source-only/oracle gap. Cross-domain detection is a developed subfield, surveyed by Oza et al. (2023). Chen et al. (2018) adapt Faster R-CNN across domains (e.g. Cityscapes to Foggy Cityscapes) with feature-level alignment, Saito et al. (2019) refine this with strong-weak alignment, and Hasan et al. (2021) show that a general pedestrian detector often beats a specialized source-trained one on new domains — i.e. you need target data. All report the same source-only-vs-oracle structure our deployment corollary reproduces. We deliberately use *no* domain-adaptation method: the point is to measure the bare transfer gap and how fine-tuning moves it, not to close it; DAOD is the natural follow-up.

Reproducibility and under-trained baselines. Our correction sits in the empirical-hygiene lineage of Musgrave et al. (2020), whose "reality check" showed that once baselines are trained and evaluated fairly, an apparently large method effect shrinks dramatically. The detection-specific version is Kay et al. (2025) (ALDI), who show that weak or under-trained source-only baselines inflate the apparent domain gap and that fair training recovers a large fraction of it. This is the general principle behind our re-analysis; our instance is a worked correction of an *unpublished* prior study (the authors' own course project), with the direction of the correction (the gap flips sign under fair training) made explicit — a reproducibility case study rather than a critique of a published result.

Fine-tuning and out-of-distribution robustness. Kumar et al. (2022) show that full fine-tuning can *distort* pretrained features and underperform out-of-distribution relative to the pretrained model — precisely the mechanism behind our "US fine-tuning ≈ off-the-shelf COCO on the German target" result, and behind our hedge that the penalty may be generic rather than US-specific. Robust fine-tuning methods such as WiSE-FT (Wortsman et al., 2022) exist to mitigate exactly this and are an obvious follow-up we do not attempt. Concurrent cross-dataset YOLO transfer matrices on overlapping driving datasets (Song et al., 2025; Chakraborty et al., 2026) report the same own-data-helps / cross-data-does-not diagonal; our added element is the zero-shot region-neutral control and the symmetric interaction-term reading.

Detectors and metrics. We use the YOLO family (Redmon et al., 2016; Redmon and Farhadi, 2018) via the Ultralytics implementation (Jocher et al., 2023), all models COCO-pretrained (Lin et al., 2014). Evaluation uses COCO-style mean average precision through TorchMetrics (Detlefsen et al., 2022) on PyTorch (Paszke et al., 2019).

Placement. No single work above subsumes our exact contribution, but together they make the transfer phenomenon textbook. We do not claim it as new. What we add is the controlled reproducibility case study of a specific claim, and a 2×3 quantification whose in-domain arm makes the residual gap interpretable.

3 The Original Study and Its Confounds

The 2024 project trained a YOLOv3 with a frozen Darknet-53 backbone — only the detection head was updated, for six epochs — on a US AV dataset, evaluated on European footage, and reported a large US→Europe drop in precision. Five design choices, taken together, make that drop uninterpretable. We list each with its fix; Table 1 is the summary.

Table 1: The five confounds in the original study and the corresponding fix in this redo.

Original flaw	Fix in this redo
Model never trained in-domain (frozen backbone, 6 head-only epochs; 2/4 classes at 0.00 precision on US too)	COCO-pretrained init, backbone unfrozen; all 4 classes $\subset$ COCO, so detection works from epoch 0
Resolution confound (Europe pre-downscaled to US size; small objects lost on EU only)	Identical letterbox to 416 for both datasets at model input
Precision-only metric (uninformative for a near-non-detecting model)	mAP@0.5, mAP@0.5:0.95, per-class AP, recall
Transfer/training confound (cannot separate "fails to transfer" from "never trained")	Region-neutral zero-shot COCO control + US- and EU-fine-tuned arms (the 2×3)
No holdout split	Seeded, disjoint train/val/test per region; the transfer target is never used to pick an epoch

1. **The model was never trained into a working state.** A frozen backbone plus six head-only epochs, on a subset the authors themselves called too small, produced a detector with 0.00 precision on two of four classes *on the US set*. *Fix:* initialize from COCO-pretrained weights, unfreeze the backbone, and rely on the fact that all four target classes are COCO classes, so the model detects from epoch zero. This is the concrete form of the "under-trained baselines overstate the gap" pitfall (Kay et al., 2025).
2. **Resolution confound.** The European images were downscaled to roughly the US resolution before inference, destroying small objects on the European side only. *Fix:* an identical aspect-preserving letterbox to 416×416 for both datasets at model input.
3. **Precision-only reporting.** Precision alone is close to meaningless for a model that rarely fires. *Fix:* report mAP@0.5, mAP@0.5:0.95, per-class AP, and recall.
4. **Train-versus-transfer confound.** With only an under-trained US model, there is no way to separate "does not transfer" from "never learned." *Fix:* add a region-neutral zero-shot COCO control and read the effect as a difference-in-differences against it (§5).
5. **No holdout.** The original had no held-out split. *Fix:* seeded, disjoint US-train / US-val / US-test splits, with the German target held out as an eval-only transfer target (and, for the in-domain arm, its own disjoint train/val/test).

4 Data

We harmonize both datasets to a single four-class vocabulary [`car`, `person`, `bicycle`, `traffic light`], chosen because all four are COCO classes and so are detectable zero-shot by a COCO-pretrained model. Every record is a unified manifest entry: an image path, a list of `xyxy` boxes, and integer class labels in $\{0, 1, 2, 3\}$.

US (source). The US data is the Udacity/CrowdAI self-driving-car dataset, distributed on Kaggle (`alincijov/self-driving-cars`). Its raw labels map to our vocabulary as {`car`, `truck`}→`car`, `pedestrian`→`person`, `bicyclist`→`bicycle`, `light`→`traffic light`. From the in-vocabulary images we draw seeded, disjoint splits: 2,500 train / 300 validation / 500 test. Validation is held out for best-epoch selection during fine-tuning; test is used only for final scoring. US image size is 480×300 as distributed on Kaggle (a downscaled release of the original 1920×1200 Udacity frames).

Target (German-driving, KITTI). The transfer target is KITTI (Geiger et al., 2012), recorded in and around Karlsruhe, Germany, accessed through Kaggle (`xavierlermusieaux/kitti-yolo`). Its `data.yaml` vocabulary is [`Car`, `Truck`, `Pedestrian`, `Cyclist`], which we map

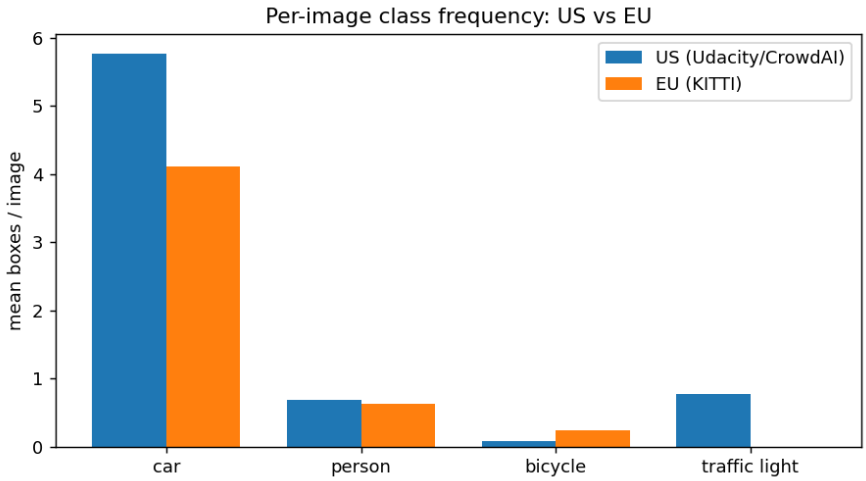

Figure 1: Per-image class frequency, US (Udacity/CrowdAI) vs. target (KITTI). `car` dominates both; `bicycle` is very sparse (US ≈ 0.08 boxes/image); KITTI has no `traffic light` ground truth.

as {`Car`, `Truck`}→`car`, `Pedestrian`→`person`, `Cyclist`→`bicycle`. We use 1,200 held-out eval images throughout; for the in-domain (EU-fine-tuned) arm we additionally carve a disjoint 2,500-image train split (matched to the US train size) and a 300-image validation split from later indices, so the 1,200-image test set stays identical — and its zero-shot / US-fine-tuned numbers stay comparable — across all conditions. KITTI native size is $\approx 1242 \times 375$. As noted, KITTI substituted for the unreachable ZOD (Alibeigi et al., 2023): the consequence is that "Europe" here is specifically German driving footage, and KITTI has *no traffic-light annotations*. All cross-region comparisons are therefore computed over the three classes shared by both datasets (`car`, `person`, `bicycle`); `traffic light` appears only on the US side of the overall tables.

Resolution handling. Both datasets pass through the *same* aspect-preserving letterbox to 416×416 at model input — there is no Europe-only pre-downscaling, which removes the original's resolution confound at the input stage. It does not equalize native scene scale: KITTI's wide frames contain larger, more prominent cars than the small, crowded objects in the US urban frames. We state this residual scene-scale confound rather than hide it; the interaction-term design (§5) is what neutralizes the part of it that is constant across conditions.

Class statistics (EDA). Figure 1 compares per-image class frequency (from `notebooks/eda.ipynb`). `car` dominates both datasets (US ≈ 5.8, EU ≈ 4.1 boxes/image); `bicycle` is extremely sparse, especially on the US side (≈ 0.08 boxes/image), which echoes the original project's observation that bicycles are underrepresented and explains the near-zero bicycle AP under the zero-shot and US-fine-tuned conditions we report later. Figure 2 shows bounding-box centre heatmaps per class per region: objects concentrate on the horizon band in both datasets, with the empty EU traffic-light panel making the missing class explicit.

5 Method

5.1 The 2×3 design and the difference-in-differences estimand

The experiment crosses three model conditions with two evaluation regions (Table 2). The conditions are **zero-shot** (the COCO-pretrained detector, region-neutral, no training), **US-fine-tuned** (fine-tuned on US-train), and **EU-fine-tuned** (fine-tuned on the in-domain KITTI train split). The regions are US-test and EU-test, both held out; the transfer target is never used to select an epoch.

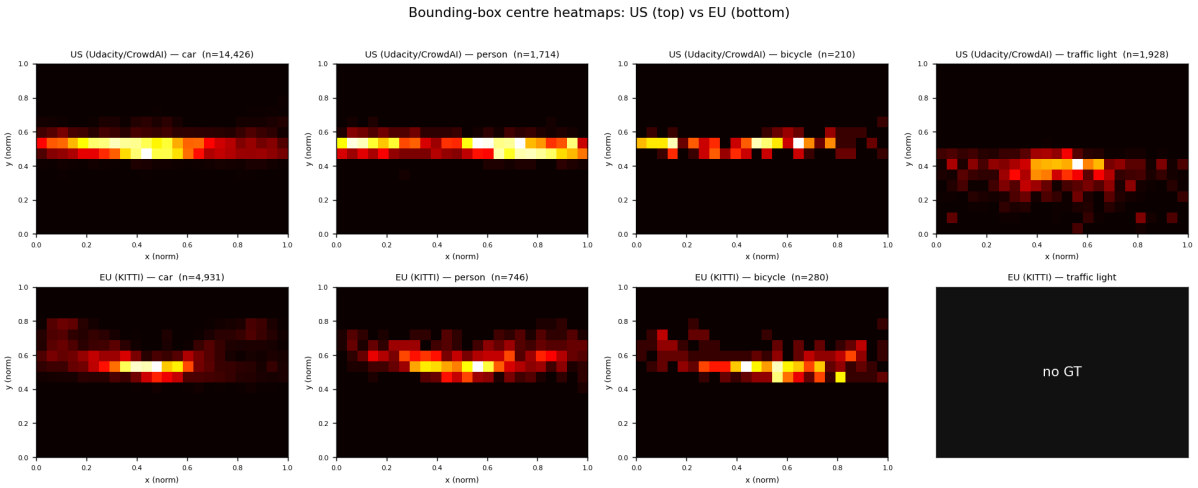

Figure 2: Bounding-box centre heatmaps (normalized image coordinates), US (top) vs. target (bottom), per class. Objects cluster on the horizon band in both regions. KITTI has no `traffic light` ground truth (rightmost EU panel empty).

Table 2: The 2×3 design. Each region's test split is held out from that region's training.

Condition	Train data	Eval: US-test	Eval: EU-test (KITTI)
Zero-shot (COCO, region-neutral)	none	✓	✓
US-fine-tuned	US-train	✓	✓
EU-fine-tuned (in-domain control)	KITTI-train	✓	✓

Q1: can US data substitute for in-domain data? Let g_c^r be the shared-class mAP@0.5:0.95 of condition c on test region r, and read the *gain over zero-shot*, $g_c^r - g_{\mathsf{zs}}^r$. If, on the German target, the in-domain-fine-tune gain far exceeds the US-fine-tune gain, then German performance needs German data and US data is not a substitute. This is the deployment question.

Q2: does fine-tuning specialise toward its own region? Restrict to a 2×2 sub-grid {zero-shot, fine-tuned}×{US, EU}. Let $m_{c,r}$ denote shared-class mAP@0.5:0.95 for condition c and region $r \in \{\mathsf{US}, \mathsf{EU}\}$. For a fine-tuned condition $f \in \{\mathsf{US\text{-}ft}, \mathsf{EU\text{-}ft}\}$ define the per-condition US→EU gap and the difference-in-differences:

$$G_c = m_{c,\mathsf{US}} - m_{c,\mathsf{EU}}, \qquad \Delta_f = G_f - G_{\mathsf{zs}} = \big(m_{f,\mathsf{US}} - m_{f,\mathsf{EU}}\big) - \big(m_{\mathsf{zs},\mathsf{US}} - m_{\mathsf{zs},\mathsf{EU}}\big). \tag{1}$$

Δ_f is arithmetically the *interaction term of a* 2×2 *(condition × region) factorial*; "difference-in-differences" is a descriptive label for it, not a new estimand. $\Delta_{\mathsf{US\text{-}ft}} > 0$ means US fine-tuning *widened* the gap toward the US (the model improved more, or degraded less, on its training region than on the German target); by symmetry $\Delta_{\mathsf{EU\text{-}ft}} < 0$ means the in-domain fine-tune widened it toward Germany. **We report both**, because their comparison is what tells the two accounts apart. A *US-specific-bias* account predicts an asymmetry (US fine-tuning specialises, in-domain fine-tuning is neutral or benign); a *generic narrow-fine-tune* account predicts a symmetric pattern (each fine-tune favours its own dataset). Crucially, with exactly one dataset per region, "US-favoring" is collinear with "Udacity-favoring": neither Δ_f can isolate a *geographic* bias from a dataset-identity effect, so a symmetric result is the honest reading and we do not claim more. $\Delta_f \approx 0$ would instead mean fine-tuning moved both regions together — the original "transfer failure" being training failure, not geography. The differencing in Eq. (1) cancels any region effect that is *constant* across the two conditions, including the bulk of the static scene/resolution confound (§4); what it cannot cancel is a confound that itself *interacts* with fine-tuning (e.g. a differential ceiling), a caveat we return to in §8 and probe with the in-domain arm.

5.2 Models

We use two COCO-pretrained detectors from Ultralytics (Jocher et al., 2023). **`yolov8s` is the primary detector**: it is fast enough to run three seeds per condition, so it carries the confidence intervals. `yolov3u`, a modernized YOLOv3 that matches the original study's architecture family (Redmon and Farhadi, 2018), serves as a single-seed cross-architecture check. The backbone is unfrozen in fine-tuning. In the zero-shot path the detector emits COCO's 80 classes and we read out the four relevant channels via the fixed map `COCO_TO_OURS`= {2,7,5 → 0 (car), 0 → 1 (person), 1 → 2 (bicycle), 9 → 3 (traffic light)} before scoring; no training occurs. The COCO `truck`(7) and `bus`(5)→`car` merge (fix M1) *mirrors* the ground-truth merges (US {`car`,`truck`}, KITTI {`Car`,`Truck`}), so the zero-shot read-out is not penalised for correctly calling a GT-"car" a truck/bus — a penalty that would otherwise differ by region (US-urban vs. KITTI-highway truck fractions) and leak into Δ. The reported Δ is computed *after* this correction. In the fine-tune path the head emits our classes directly, so no COCO remapping is needed at evaluation.

5.3 Training protocol

Fine-tuning uses input size 416, batch size 8, the Ultralytics auto-optimizer (SGD or AdamW) with initial learning rate 0.01, and a fixed budget of 12 epochs. Early stopping is disabled (patience ≥ epochs) so every seed trains the same number of epochs — early stopping on the small, noisy validation split made best-epoch selection vary across seeds, a protocol artifact rather than genuine seed variance — while best-epoch weights are still selected by held-out validation mAP. The primary model (`yolov8s`) is run for three seeds $\{42, 123, 7\}$ in *each* of the US- and EU-fine-tuned conditions; `yolov3u` is run for a single seed (42) in each. We report each difference-in-differences Δ_f as the *mean* ± *standard deviation over the three* ***`yolov8s`*** *seeds*, with, for the US-ft direction, the count of seeds for which $\Delta_{\text{US-ft}} > 0$. All runs are on a single Apple-silicon (MPS) laptop.

5.4 Metrics

Evaluation uses TorchMetrics' COCO-style `MeanAveragePrecision` (Detlefsen et al., 2022) through one shared code path for both models, reporting mAP@0.5, mAP@0.5:0.95, per-class AP, and recall. Two honest details:

- **Confidence threshold.** Predictions are thresholded at confidence 0.25, so absolute mAP values read *lower* than a standard COCO evaluation (which uses confidence ≈ 0.001) and should *not* be compared to published COCO numbers. The threshold is applied uniformly across all six cells, so the within-experiment gap comparison — which is all Eq. (1) uses — remains valid. We verify this is not an artifact of the cut by re-evaluating threshold-free (§6, robustness).
- **Recall** is TorchMetrics' MAR@100 (mean average recall at 100 detections).

For fairness to KITTI's three-class vocabulary, predictions of any class absent from *all* ground truth in a dataset are dropped before scoring, so EU traffic-light predictions cannot create false positives. The shared-class gap then compares like with like.

6 Results

Table 3 gives the full 2×3 in overall metrics; Table 4 reduces it to the shared-class mAP@0.5:0.95 used by Q1 and Q2. (Overall = mean AP over classes present in that region's ground truth: four for US including `traffic light`, three for KITTI; the German column is therefore identical between the two tables.)

Table 3: The 2×3 overall results. `yolov8s` fine-tuned cells are seed means (n=3); `yolov3u` fine-tuned cells are single-seed (n=1). Absolute mAPs use a confidence threshold of 0.25 and are *not* comparable to published COCO numbers (§5).

		US-test		**EU-test (KITTI)**	
Model	Condition	mAP@.5	mAP@.5:.95	mAP@.5	mAP@.5:.95
`yolov8s` (primary)	zero-shot COCO	0.215	0.110	0.324	0.167
`yolov8s` (primary)	US-fine-tuned (n=3)	0.407	0.210	0.335	0.168
`yolov8s` (primary)	EU-fine-tuned (n=3)	0.183	0.091	0.523	0.321
`yolov3u` (arch chk)	zero-shot COCO	0.247	0.126	0.378	0.200
`yolov3u` (arch chk)	US-fine-tuned (n=1)	0.374	0.192	0.326	0.163
`yolov3u` (arch chk)	EU-fine-tuned (n=1)	0.184	0.084	0.534	0.332

Table 4: Shared-class (`car`/`person`/`bicycle`) mAP@0.5:0.95, the quantity used in Q1/Q2. Gain columns are relative to the zero-shot row of the same model/region.

Model	Test region	Zero-shot	US-fine-tuned (gain)	EU-fine-tuned (gain)
`yolov8s`	US-test	0.136	0.214 (+0.078)	0.121 (−0.015)
`yolov8s`	EU-test (KITTI)	0.167	0.168 (+0.001)	0.321 (+**0.153**)
`yolov3u`	US-test	0.154	0.190 (+0.036)	0.112 (−0.042)
`yolov3u`	EU-test (KITTI)	0.200	0.163 (−0.037)	0.332 (+**0.132**)

The zero-shot German target is the easier set. Zero-shot, the German numbers are *higher* than the US numbers for both models (e.g. `yolov8s` EU mAP@0.5 = 0.324 vs. US 0.215; shared mAP@0.5:0.95 EU 0.167 vs. US 0.136). On a region-neutral control, KITTI is the *easier* set — its highway cars are large and prominent rather than small and crowded — so the original study's raw US→EU drop was largely an artifact of never-trained-plus-resolution, not geography.

Q1: for German deployment, US fine-tuning ≈ off-the-shelf COCO. Table 4 is the practitioner's answer. On the held-out German target, fine-tuning on US data moves shared-class mAP by +0.001 (`yolov8s`) and −0.037 (`yolov3u`) — i.e. US fine-tuning is, at best, indistinguishable from the off-the-shelf COCO model on KITTI, and can be worse. Fine-tuning on *in-domain* German data instead lifts the same number by +0.153 and +0.132, nearly doubling zero-shot performance (EU mAP@0.5 rises to 0.523/0.534). The message is blunt: for German deployment, US data is not a substitute for German data — you need in-domain data. This is the source-only/oracle gap of DAOD (Chen et al., 2018; Hasan et al., 2021) in practitioner form.

Q2: fine-tuning specialises toward its own dataset — symmetrically. Table 5 gives the interaction terms of Eq. (1) in both directions. The US-fine-tuned model shifts the gap toward the US ($\Delta_{\text{US-ft}} = +0.077 \pm 0.007$ for the primary `yolov8s`, 3/3 seeds positive; +0.073 for the `yolov3u` check). But applying the *identical* rule to the in-domain German fine-tune shifts the gap the other way, and by more than twice as much: $\Delta_{\text{EU-ft}} = -0.169 \pm 0.010$ (`yolov8s`) and −0.174 (`yolov3u`). The effect is therefore *symmetric* — each fine-tune favours its own dataset, and the in-domain one specialises harder. Because there is exactly one dataset per region, "US-favoring" cannot be separated from "Udacity-favoring", so this is narrow-fine-tune specialisation, *not* evidence of a specifically *US* geographic bias; reading $\Delta_{\text{US-ft}} > 0$ alone as US-bias would over-claim. The two architectures land on essentially the same pair of Δs; we read this convergence as a sanity check on the evaluation pipeline, not as an independent significance claim.

Table 5: Shared-class US→EU gap and the *symmetric* difference-in-differences (Eq. (1)). Gap $G_c = m_{c,\mathsf{US}} - m_{c,\mathsf{EU}}$ in mAP@0.5:0.95; $\Delta_f = G_f - G_{\mathsf{zs}}$; ± is SD over seeds. The US-ft and EU-ft Δs have opposite signs and comparable-to-larger magnitude — each fine-tune favours its own dataset — so neither is evidence of a region-specific (as opposed to dataset-specific) bias.

	Zero-shot	**US-fine-tuned**		**EU-fine-tuned (in-domain)**	
Model	gap G_{zs}	gap G_f	$\Delta_{\mathsf{US\text{-}ft}}$	gap G_f	$\Delta_{\mathsf{EU\text{-}ft}}$
`yolov8s` (primary)	−0.031	$+0.046 \pm 0.007$	$\mathbf{+0.077 \pm 0.007}$	-0.200 ± 0.010	$\mathbf{-0.169 \pm 0.010}$
`yolov3u` (arch chk)	−0.046	+0.027	**+0.073**	−0.220	**−0.174**

Robustness to the confidence threshold. Because the conf-0.25 cut could in principle truncate the precision–recall curve differently for zero-shot vs. fine-tuned heads (and by region), we re-evaluated threshold-free at confidence 0.001 (seed 42). The sign and magnitude survive: $\Delta = +0.078$ (`yolov8s`) and +0.081 (`yolov3u`), with the zero-shot gap negative (−0.029, −0.039) and the US-fine-tuned gap positive (+0.049, +0.042) in both cases. The gap-sign is therefore not an artifact of the confidence threshold.

Per-class breakdown and the bicycle caveat. Table 6 gives per-class AP. `car` carries the signal, as expected from its dominance. `bicycle` AP is at or near zero in the zero-shot and US-fine-tuned conditions — a direct consequence of the bicycle sparsity in Figure 1, not a pipeline failure — so in exactly the two conditions that define Δ, the three-class mean is effectively a `car`/`person` result. Notably, in-domain fine-tuning is the one condition that *does* learn bicycles (EU `bicycle` AP rises to 0.164/0.170), which reinforces rather than weakens the "you need in-domain data" conclusion. Reading `car` and `person` AP directly on the German target: zero-shot 0.322/0.180, US-fine-tuned 0.327/0.171 (barely moved), in-domain 0.541/0.257 (both sharply up) for `yolov8s`.

To confirm $\Delta_{\mathsf{US\text{-}ft}}$ is not an artifact of the near-zero bicycle class, we recompute it on class subsets: dropping bicycle gives $\Delta_{\mathsf{US\text{-}ft}} = +0.072$ (`yolov8s`) / +0.109 (`yolov3u`) on `car`+`person`, and +0.121 / +0.140 on `car` alone — the same sign, comparable-or-larger magnitude, so the interaction term is genuinely carried by the well-populated classes, not by bicycle noise. Decomposing the +0.153 in-domain German gain by class (`yolov8s`): `car` +0.219, `person` +0.077, `bicycle` +0.164 — every shared class improves, and roughly a third of the gain is the recovery of the cyclist class from AP ≈ 0. That last point has a safety reading: on German streets the zero-shot and US-fine-tuned detectors are effectively blind to cyclists (the vulnerable road users), and only in-domain data restores them.

7 Discussion

What this says for deployment. For a practitioner holding US data and targeting Germany, the result is a caution with a clear recipe. A region-neutral COCO detector already carries over to both regions, so "German footage is intrinsically hard" is not the obstacle — on KITTI it is in fact easier. The obstacle is the recipe: fine-tuning on the US data you have adds essentially nothing on the German target beyond the off-the-shelf COCO model (+0.001 shared mAP for the primary model), whereas fine-tuning on in-domain German data nearly doubles performance. The decision-relevant number is the held-out German one; the US number moves in the opposite, reassuring-looking direction and should not be reported alone.

The in-domain arm refutes a pure ceiling reading. A tempting deflation of $\Delta > 0$ is that KITTI simply starts higher zero-shot, leaves less headroom, and so mechanically resists improvement. The EU-fine-tuned control rules this out: in-domain fine-tuning lifts the German target from 0.167 to 0.321 shared mAP, so the target had ample headroom. US fine-tuning's

Table 6: Per-class AP@0.5:0.95 (zero-shot single-seed; fine-tuned seed-mean). "–" marks KITTI's absent `traffic light` class.

Model	Condition	Region	car	person	bicycle	traffic light
yolov8s	zero-shot	US	0.292	0.116	0.000	0.031
yolov8s	zero-shot	EU	0.322	0.180	0.000	–
yolov8s	US-fine-tuned	US	0.418	0.131	0.095	0.198
yolov8s	US-fine-tuned	EU	0.327	0.171	0.007	–
yolov8s	EU-fine-tuned	US	0.217	0.080	0.066	0.000
yolov8s	EU-fine-tuned	EU	0.541	0.257	0.164	–
yolov3u	zero-shot	US	0.323	0.138	0.000	0.041
yolov3u	zero-shot	EU	0.383	0.215	0.001	–
yolov3u	US-fine-tuned	US	0.434	0.136	0.000	0.197
yolov3u	US-fine-tuned	EU	0.354	0.135	0.000	–
yolov3u	EU-fine-tuned	US	0.208	0.058	0.069	0.000
yolov3u	EU-fine-tuned	EU	0.565	0.259	0.170	–

failure to move it is therefore substantive — US data does not transfer to the German target — not an artifact of a ceiling.

The German-side drop is the clean half. Of the two components of the US-fine-tuned gap, the US-test number shares images with the training region (and US-val drives model selection), while the German set is a true holdout. The fact that the held-out German number does not improve under US fine-tuning is the part of the result that needs the fewest assumptions: it stands on the held-out side alone, independent of any selection effect on the US side.

Why an interaction term and not a level comparison. A level comparison of German mAP before and after fine-tuning conflates the region effect with everything fine-tuning changes. By subtracting the US→EU gap of the region-neutral control, Eq. (1) removes the constant scene/ resolution advantage KITTI enjoys and isolates the *change* fine-tuning induces — which is why a positive Δ reads as a fine-tuning-induced shift rather than as KITTI simply being easier (already netted out in G_{zs}).

What the effect is, and is not. The honest reading is that narrow fine-tuning traded transfer for in-domain accuracy — a detection-level instance of the dataset-bias / feature-distortion picture (Torralba and Efros, 2011; Geirhos et al., 2020; Kumar et al., 2022). Whether the priors it learned are specifically *US-scene* priors or simply *whatever-it-trained-on* priors that would hurt transfer to any new domain, the *symmetry* of the interaction term argues for the latter. Applying the same difference-in-differences to the in-domain fine-tune (Table 5) yields $\Delta_{\mathsf{EU\text{-}ft}} = -0.169$, opposite in sign and *larger* in magnitude than the US-ft $\Delta_{\mathsf{US\text{-}ft}} = +0.077$: the German fine-tune specialises toward Germany even harder than the US fine-tune specialises toward the US. Because each region contributes exactly one dataset, a US-specific-bias account and a generic narrow-fine-tune account are collinear here, and the observed symmetry is what the generic account predicts. We therefore claim only the dataset-neutral statement — fine-tuning specialises toward its own training distribution, symmetrically — and explicitly *do not* read $\Delta_{\mathsf{US\text{-}ft}} > 0$ as US-specific geographic bias. Robust-fine-tuning baselines (Wortsman et al., 2022) are the obvious mitigation, and cleanly separating dataset identity from geography would need multiple datasets per region (§8).

Positioning. None of the three findings is a new phenomenon. The transfer gap re-confirms textbook domain shift (Wang et al., 2020; Torralba and Efros, 2011); the deployment corollary re-confirms the DAOD source-only/oracle gap (Chen et al., 2018; Hasan et al., 2021); the correction is a concrete instance of the under-trained-baseline pitfall (Musgrave et al., 2020; Kay et al., 2025). The value is in doing all three cleanly, as a reproducibility case study on a specific prior claim, with the controls that make the numbers mean what they say.

8 Limitations

We list the constraints plainly; several bound what the result can mean.

1. **"Europe" is KITTI, i.e. Karlsruhe, Germany** — not a broad European sample, and not the intended ZOD (Swedish) footage, which was unreachable in our environment. Every claim is US-urban → German-driving, not US → Europe at large.
2. **KITTI has no traffic-light ground truth**, so all cross-region comparisons use the three shared classes (`car`, `person`, `bicycle`); `traffic light` appears only on the US side.
3. **The fine-tune is light** (12 epochs on a single MPS laptop, ∼2.5k images per region). A stronger or longer fine-tune would, if anything, *increase* specialization and widen the gap, so the reported Δ is conservative.
4. **The static scene/resolution confound is netted out only insofar as it is constant.** The interaction term cancels the additive part; a confound that *interacts* with fine-tuning is not fully ruled out by Eq. (1) alone. We check the most obvious such interaction — a differential ceiling — with the in-domain arm (§7), which shows the German target had headroom.
5. **Bicycle AP ≈ 0 under zero-shot and US fine-tuning** because bicycle ground truth is extremely sparse (Figure 1), so in the two conditions that define Δ the shared-class mean is effectively `car`/`person`. In-domain fine-tuning does recover bicycle AP.
6. **Dataset identity and geography are collinear.** Each region contributes exactly one dataset (US = Udacity, target = KITTI), so a US-specific *geographic* bias cannot be separated from a Udacity-specific *dataset* effect (Kumar et al., 2022). The symmetric interaction term ($\Delta_{\mathsf{EU\text{-}ft}} = -0.169$ vs. $\Delta_{\mathsf{US\text{-}ft}} = +0.077$) is what a generic narrow-fine-tune cost predicts; we do not claim the cost is region-bound. Decomposing the two would require multiple datasets per region.
7. **Possible frame-level leakage on the US side.** The US (Udacity) frames come from driving *video*, and our train/val/test split is a seeded random partition of *frames*, not of drives — so near-duplicate adjacent frames could straddle the split and inflate the US-test number, and with it the in-domain US gain and the US-ward $\Delta_{\mathsf{US\text{-}ft}}$. The KITTI object benchmark is a curated set of non-consecutive frames, so its held-out target is at low risk; and Q1's central result — that US fine-tuning does not help the *held-out German* target — lives entirely on that KITTI holdout and is immune to any US-side leakage. A drive-disjoint US split is the clean fix and a planned follow-up.
8. **Compute and scope.** Apple-silicon MPS; three seeds for the primary model and one for the architecture check; no domain-adaptation or robust-fine-tuning methods and no resolution-controlled ablation — deliberate out-of-scope follow-ups.
9. `yolov3u` **is a modernized YOLOv3**, not a from-scratch Darknet reimplementation; it matches the original's architecture family but is not bit-identical to it, and is run at a single seed, so its Δ carries no confidence interval.

9 Conclusion

A 2024 course project asked a good and practical question — can a US-trained driving detector be deployed in Europe? — with a design that could not answer it, because an under-trained, precision-only, resolution-confounded, holdout-free model conflates "cannot transfer" with "cannot detect." We re-ran the study as a controlled 2×3 with a region-neutral COCO control and an in-domain fine-tuning arm, scoped honestly to US-urban → German-driving (KITTI). The answers

are small and clean on the primary model, and reproduce on a single-seed second architecture. A region-neutral detector already transfers to both regions (KITTI is in fact the *easier* set), so the original's raw drop was largely an artifact. For German deployment, fine-tuning on US data buys essentially nothing over off-the-shelf COCO (+0.001 shared mAP), while in-domain German data nearly doubles performance (+0.153): you need in-domain data. And read as a 2×2 interaction term, fine-tuning specialises toward its own dataset *symmetrically*: the US fine-tune shifts the US→EU gap toward the US ($\Delta_{\text{US-ft}} = +0.077 \pm 0.007$ shared-class mAP, 3/3 seeds; +0.073 on the YOLOv3 check) while the in-domain fine-tune shifts it the other way and more strongly ($\Delta_{\text{EU-ft}} = -0.169$), a pattern that survives a threshold-free re-evaluation — so with one dataset per region the effect is narrow-fine-tune specialisation, not a specifically *US* bias. None of these is a new phenomenon — they re-confirm textbook domain shift, the DAOD source-only/oracle gap, and the under-trained-baseline pitfall. The contribution is a reproducibility case study that corrects a specific "catastrophic transfer" claim and a clean quantification with an in-domain control, not the magnitude of any effect.

Reproducibility Statement

All code, manifests, and figures are in the project repository. The pipeline is deterministic given the seeds and the Kaggle credentials (`~/.kaggle/kaggle.json`). The command sequence is:

```
PYTHONPATH=.  python3 data/get_us.py
PYTHONPATH=.  python3 data/get_kitti.py
python3 run_zeroshot.py
bash run_matrix.sh   # 2×3 US+EU fine-tune sweep
python3 robustness_conf.py -conf 0.001 -seed 42
python3 aggregate.py
```

Each (model, domain, seed) run in `run_matrix.sh` is independent and resumable (`finetune_one.py` skips a run whose result JSON already exists), so an interrupted sweep can simply be re-launched; `aggregate.py` renders `results/transfer_table.md` and `results/metrics.json`, the source of every number in §6. The EDA figures come from `notebooks/eda.ipynb`. Evaluation uses a fixed confidence threshold of 0.25 (with a threshold-free 0.001 robustness pass) and TorchMetrics COCO-style mAP/recall, applied identically to all cells of the 2×3.

References

Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20178–20188, 2023.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

Sourav Chakraborty et al. Towards robust cross-dataset object detection generalization under domain specificity. *arXiv preprint arXiv:2601.09497*, 2026.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3339–3348, 2018. arXiv:1803.03243.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Achel Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics – measuring reproducibility in PyTorch. *Journal of Open Source Software*, 7(70):4101, 2022.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 52–59, 2019. arXiv:1906.02659.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:2003.08799.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO. https://github.com/ultralytics/ultralytics, 2023. Software.

Justin Kay, Timm Haucke, Suzanne Stathatos, Siqi Deng, Erik Young, Pietro Perona, Sara Beery, and Grant Van Horn. Align and distill: Unifying and improving domain adaptive object detection. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2403.12029.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2012.07421.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.10054.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, pages 681–699, 2020. arXiv:2003.08505.

Poojan Oza, Vishwanath A. Sindagi, Vibashan Vishnu Murali, and Vishal M. Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(4):4018–4040, 2023. arXiv:2105.13502.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019. arXiv:1812.04798.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NeurIPS Workshop on Machine Learning for the Developing World*, 2017. arXiv:1711.08536.

Zhaoxin Song et al. Synthetic dataset evaluation based on generalized cross validation. *arXiv preprint arXiv:2509.11273*, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. Train in Germany, test in the USA: Making 3D object detectors generalize. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2005.08139.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2109.01903.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020.