Beyond Borders: Quantifying Geographic Domain Shift in Automotive Perception

Joe Murtagh MURTAGH.JOE@UL.IE
Department of Electronic & Computer Engineering
University of Limerick, Ireland

Reenu Mohandas REENU.MOHANDAS@UL.IE
Department of Electronic & Computer Engineering
University of Limerick, Ireland

Tim Brophy TIM.A.BROPHY@UL.IE
Department of Electronic & Computer Engineering
University of Limerick, Ireland

Ganesh Sistu GANESH.SISTU@UL.IE
Valeo Vision Systems, Tuam, Co. Galway, Ireland

Ciarán Eising CIARAN.EISING@UL.IE
Department of Electronic & Computer Engineering
University of Limerick, Ireland

Reviewed on OpenReview: *https://openreview.net/forum?id=go887umxL8*

Abstract

Automated driving systems rely on diverse training data, yet increasing privacy regulations restrict cross-border data sharing. This study shows that geographic separation is associated with a continuous, distance-based domain shift and progressively lower model performance, with distance acting as a proxy for latent environmental factors, challenging the assumption of binary in-domain versus out-of-domain behaviour. Using two European driving datasets, EuroCity Persons Dataset (ECP) and the Zenseact Open Dataset (ZOD), and one global dash-camera dataset OpenStreetView-5M (OSV-5M), we provide converging evidence at representation and task levels. Encoder semantic embeddings diverge systematically with geographic distance, yielding Pearson correlations of 0.476–0.632 for ECP, 0.570–0.711 for ZOD, and 0.404–0.568 for OSV-5M across three distributional metrics (Maximum Mean Discrepancy (MMD), Sliced Wasserstein Distance (SWD), Energy Distance (ED)), strengthening to 0.665–0.762 when controlling for confounding variables on ZOD. Downstream perception models exhibit corresponding performance degradation: object detection correlates at -0.512 to -0.585 on ECP and -0.563 to -0.659 on ZOD, while semantic segmentation shows effects of -0.503 to -0.726 across models. While this study is observational and does not establish causality, controlled analyses reduce confounding effects, and annotation spatial distributions show no statistically significant correlation with geographic distance, suggesting that annotation distribution differences are unlikely to be the primary driver of the observed trends. These findings suggest that geographically fragmented datasets may contribute to performance gaps when models are deployed substantially outside their training regions. We characterise geographic domain shift as continuous, with perception performance degrading progressively with increasing spatial separation rather than at a simple in-domain/out-of-domain boundary.

Keywords: Domain Shift, Automotive Perception, Regional data variability

1 Introduction

Modern machine learning typically relies on the centralised collection of data (Kairouz et al., 2021) on a server for training and subsequent deployment. However, in an era of increasing regulation surrounding the gathering and movement of data across regional boundaries (Brisimi et al., 2018), organisations' ability to collect and centralise data for training purposes is increasingly constrained. Prior work suggests that, the loss of data from a specific region cannot be compensated for by adding data from other regions due to inherent domain variation (Torralba and Efros, 2011; Koh et al., 2021; Jean et al., 2019). This study examines how geographic boundaries that restrict data sharing are associated with domain shifts and with lower model performance on regions absent from the training data. We aim to quantify this degradation and analyse its implications for continual learning frameworks.

Our work differs from prior research in that we focus on spatial rather than temporal data distribution shifts (Beery et al., 2018), and we use geographic distance as a quantifiable measure of domain divergence. Although geographic distance is often treated as a binary property in data partitioning, data are either from the same region or from an external region (Quiñonero-Candela et al., 2008; Ben-David et al., 2010); our work highlights that distance behaves as a continuous factor influencing model performance. Rather than a simple in-domain vs out-of-domain distinction, we show that performance degrades progressively with increasing geographic separation. This perspective reframes the domain shift as a spectrum rather than a discrete boundary, revealing that even moderate distances can introduce measurable divergence in visual and environmental characteristics, as conceptually illustrated in Fig.1. By quantifying this relationship, we move beyond the conventional binary framework and provide evidence that geographic distance is a meaningful predictor of generalisation gaps in automotive AI systems.

Given that models require access to diverse and representative datasets to achieve optimal performance in the regions where they are deployed (Sagawa et al., 2019), regulatory constraints pose a significant challenge. These limitations often result in models being trained on local datasets and deployed to areas not represented in the training data (Taori et al., 2020), which this study will show can adversely affect their performance. From a data-sovereignty perspective, consider a multinational company deploying automated vehicles across different continents. Regulatory restrictions prevent the centralisation of driving data from Europe, Asia, and North America (Rieke et al., 2020). As a result, models trained in one region may underperform in another due to unseen domain-specific features (Zhang et al., 2022).

Unlike prior work, we introduce a geographically stratified evaluation protocol to analyse the effects of regional domain variation. While geographic distance is not a perfect measure, it serves as a practical proxy for latent factors such as cultural norms, architectural styles, and infrastructure, which are difficult to quantify but significantly influence data distributions (Jean et al., 2019; Tachet des Combes et al., 2020). Geographic distance is intentionally used as an observable, reproducible, and scalable proxy for latent environmental variation. Before proposing solutions to this problem, it is essential to establish that the issue exists and has a measurable impact on model performance. We accomplish this

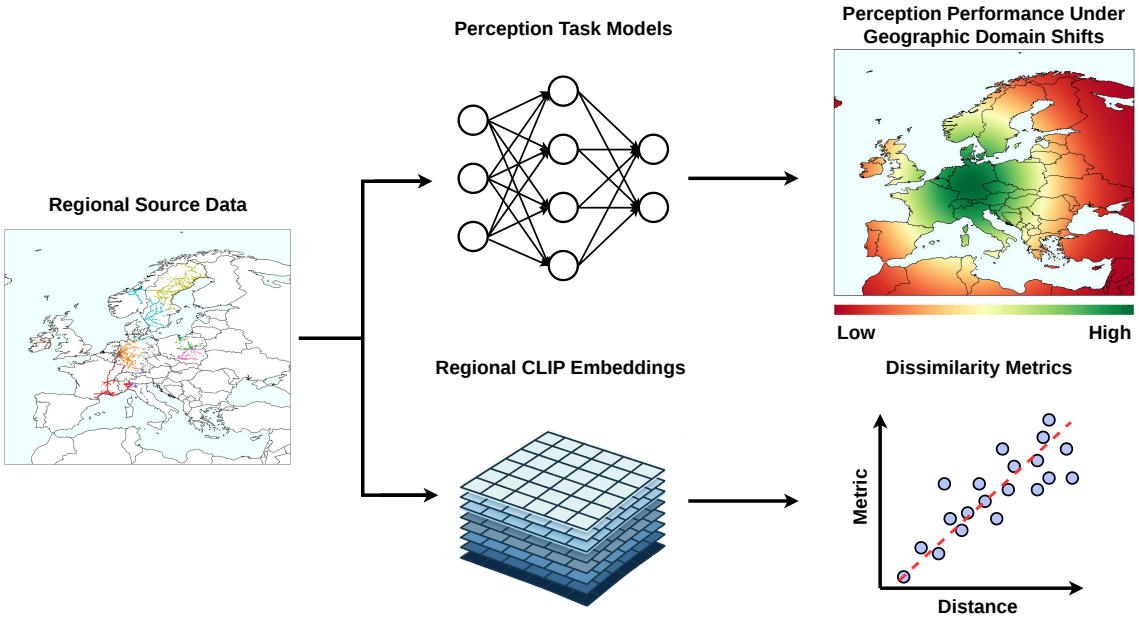


Figure 1: Conceptual relationship between geographic distance and domain shift in automotive perception. As the geographic distance between training and deployment regions increases, visual data distributions diverge, leading to greater representation dissimilarity and reduced perception performance.

by training models and demonstrating a correlation between performance degradation and the geographic distance of the validation data.

The main contributions of this study are:

1. We establish an empirical relationship between geographic distance and encoder embedding dissimilarity, showing that geographically distant data exhibit statistically significant differences in their learned representations

2. We empirically demonstrate a statistically significant correlation between geographic distance and model performance degradation in object detection and segmentation tasks.

3. We frame this challenge within data domain shifts, focusing on the setting where models must adapt to geographically induced distribution shifts.

4. We discuss the implications of our findings for model deployment in regulated environments and outline potential strategies for mitigating performance loss due to domain shift.

The remainder of the study is structured as follows: Section 2 reviews decentralised learning under regulatory constraints, and domain shifts in automotive AI systems. Section 3 describes our experimental methodology, including dataset selection criteria, ge-

ographic clustering approaches, correlation analysis methods, and the machine learning models employed for object detection and segmentation tasks. Section 4 presents our empirical findings, beginning with encoder based semantic similarity analysis across geographic distances, followed by downstream perception task performance across multiple model architectures. Section 5 discusses the implications of our findings, interpreting both representation-level and task-level correlations while examining differences between detection and segmentation tasks. Section 6 acknowledges the limitations of our observational study design and the scope of our geographic coverage. Section 7 outlines promising directions for future research, including the development of geography-aware methods and the extension of this analysis to other global regions. Finally, Section 8 concludes by summarising our key contributions and their broader implications for AI systems operating under regulatory constraints.

2 Background and Related Work

2.1 Decentralised Learning and Regulatory Constraints

In parallel with rapid advances in artificial intelligence, governments worldwide have introduced regulatory frameworks that increasingly constrain the movement of data across borders. These regulations govern not only personal data protection, but also the collection, storage, transfer, and reuse of data for AI model training, deployment, and monitoring. Major jurisdictions, including the European Union (EU), China, India, the United States, Canada, Japan, Singapore, Israel, and several Middle Eastern, East Asian, and European countries, have enacted or proposed AI-specific laws, data protection statutes, and data localisation requirements that affect cross-border data movement. Collectively, these measures signal a global shift toward jurisdictional control over AI-relevant data, posing significant challenges for centralised machine learning pipelines that rely on large-scale data aggregation.

In the EU, the General Data Protection Regulation (GDPR; Regulation (EU) 2016/679) restricts cross-border transfers of personal data and is complemented by the Artificial Intelligence Act, which introduces AI-specific data governance and lifecycle obligations (European Parliament and Council of the European Union, 2016, 2024). China's Personal Information Protection Law and Data Security Law impose consent, security assessment, and localisation requirements for exporting data used in AI applications (Standing Committee of the National People's Congress, 2021b,a). India's Digital Personal Data Protection Act authorizes transfer restrictions subject to government notification (Government of India, 2023). In North America, U.S. state privacy laws and federal instruments such as the CLOUD Act affect cross-border data access (United States Congress, 2018), while Canada's PIPEDA framework and proposed Artificial Intelligence and Data Act integrate AI accountability with data handling obligations (Government of Canada, 2000; Parliament of Canada, 2022).

Comparable constraints arise in East Asia and the Asia-Pacific region, where Japan's Act on the Protection of Personal Information governs international data transfers through adequacy and contractual mechanisms, while accompanying AI governance frameworks emphasise responsible data use (Government of Japan, 2003). Singapore's Personal Data Protection Act similarly conditions cross-border data transfers on comparable protection standards and has been complemented by AI governance instruments such as the Model

Artificial Intelligence Governance Framework (Government of Singapore, 2012; Infocomm Media Development Authority, 2019). In the Middle East and adjacent regions, Israel's Protection of Privacy Law and emerging AI regulatory initiatives impose safeguards on data export and automated decision-making (State of Israel, 1981; Israel Innovation Authority, 2022). Similar trends are reflected in Saudi Arabia's and the UAE's data protection laws (Kingdom of Saudi Arabia, 2021; United Arab Emirates, 2021), South Korea's AI Basic Act (National Assembly of the Republic of Korea, 2025), Switzerland's revised Federal Act on Data Protection, South Africa's POPIA (Swiss Confederation, 2020; Republic of South Africa, 2013), and the Council of Europe's Framework Convention on Artificial Intelligence (Council of Europe, 2024), though this is by no means a complete list.

As data privacy regulations tighten globally, restrictions on cross-border data transfers and data minimisation mandates create substantial barriers to traditional centralised machine learning pipelines. Decentralised approaches such as federated learning mitigate raw data transfer but do not inherently address temporal data evolution or heterogeneous, region-specific distributions (Li et al., 2020; Zhao et al., 2018). Continual learning directly addresses these regulatory constraints by enabling sequential adaptation to new data domains without retaining historical data, aligning with data minimisation principles (Juliussen et al., 2023). Similarly, domain adaptation (Wang and Deng, 2018) provides mechanisms for transferring knowledge between source and target distributions, but typically assumes a predefined target domain rather than the sequential emergence of new domains over time. However, none of these approaches alone provides a complete solution to the combined challenges of privacy, geographic heterogeneity, and evolving data distributions, motivating the need to investigate how models can robustly adapt to geographically induced domain shifts.

The discussion of data-sovereignty and data-protection regulations is intended primarily to motivate the study of geographic domain shift in automotive perception. Our experiments neither quantify the effects of GDPR, PIPL, DPDP, nor other regulatory frameworks, nor establish a causal link between regulation and the observed performance degradation. Instead, these regulations provide a practical example of how geographically fragmented datasets may emerge in real-world deployments. The focus of this work is therefore on empirically characterising the relationship between geographic separation, distributional shift, and perception performance, rather than assessing the impact of regulatory policies.

2.2 Domain Shifts in Automotive AI

Automotive AI systems encounter numerous sources of domain shift that affect input distributions while preserving task semantics. Weather conditions represent a primary source of variation (Kumar et al., 2025): rain, snow, fog, and bright sunlight dramatically alter visual characteristics, affecting visibility, introducing reflections, and obscuring features such as road markings and traffic signs (Kalb and Beyerer, 2023). Models trained exclusively in clear weather may exhibit degraded performance in adverse conditions (Muralidhara et al., 2025). Lighting transitions between daytime and nighttime driving constitute another significant domain shift (Wang et al., 2025). Illumination differences affect object appearance, road surface reflectance, and background characteristics. Features readily detectable during daylight may become indistinct or distorted under nighttime conditions (Sun et al., 2022).

Urban versus rural environments present distinct distributional characteristics: urban areas feature high vehicle and pedestrian density, structured infrastructure, and complex interactions, while rural regions exhibit open roads, natural obstacles, and sparser traffic patterns (Yang et al., 2024).

Sensor configurations introduce additional domain variability. Automated vehicles employ combinations of RGB cameras, LiDAR, radar, and thermal imaging, and sensor upgrades, replacements, or degradation over time create distributional shifts in the raw input space (Su et al., 2024). Traffic density varies substantially across operational contexts, with high-density scenarios involving frequent occlusions and complex multi-agent interactions, while low-density environments offer clearer visibility and simpler perceptual demands. The synthetic-to-real domain gap remains a persistent challenge, as models trained on synthetic datasets for scalability and annotation efficiency must adapt to natural data distributions during real-world deployment (Huch and Lienkamp, 2023).

While weather, lighting, and other environmental shifts have been extensively studied in the domain adaptation literature (Wilson and Cook, 2020; Pan and Yang, 2009), comparatively less attention has been paid to geographic distance as a source of domain shift. These commonly studied factors are undeniably important; however, they often occur within a single location and can, in principle, be mitigated through local data collection. In contrast, geographic distance represents an understudied but, as we will show, critical driver of performance degradation in automotive perception systems (Christie et al., 2018). Regional variations in road infrastructure (lane width, marking styles, signage conventions), traffic patterns (driving behaviours, vehicle types, road user composition), and environmental characteristics (vegetation, building architecture, terrain) create systematic distributional differences between geographically distant regions. Unlike weather or lighting changes, geographic domain changes are inherently tied to spatial separation and frequently coincide with regulatory or logistical barriers that prevent data centralisation. Despite these practical implications, the relationship between geographic distance and performance degradation has not been systematically quantified. This work addresses that gap by empirically measuring how model performance degrades as a function of spatial distance between training and evaluation regions, thereby characterising geographic separation as a measurable correlate of domain shift in automotive AI systems.

3 Methodology

3.1 Datasets

Table 1: Comparison of major machine learning driving datasets

	ECP	ZOD	Cityscapes	KITTI	Waymo	nuScenes
Splits	60-10-30[1]	80-0-20	60-10-30[1]	70-0-30[1]	69-18-13[1]	70-15-15[1]
Image Size	1920×1024	3848×2168	2048×1024	1242×375	1920×1280	1600×900
Frames/Images	47,300	100,000	25,000	15,000	200,000	40,000
Countries	12	14	3[2]	1	1	2
Sensors	Camera	Camera	Camera	Multimodal	Multimodal	Multimodal

[1] Test set labels are not publicly available.
[2] Of its 50 cities, only 2 are non-German (Strasbourg and Zurich).

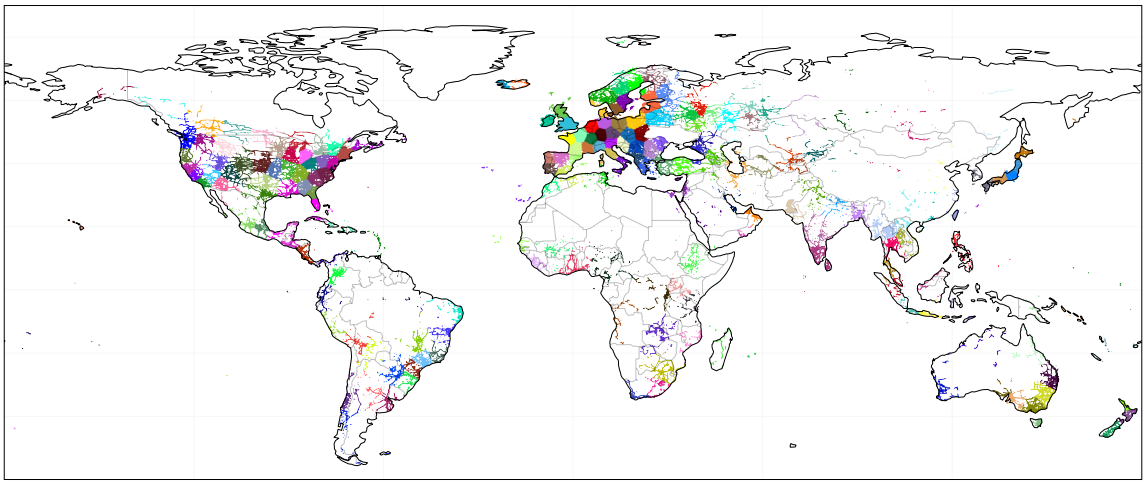


Figure 2: OSV-5M dataset clustered into 200 regions are used for embedding and comparison



(a) EuroCity Persons Dataset

(b) Zenseact Open Dataset

Figure 3: Geographical region splits used for perception training and evaluation.

The ideal dataset for this experiment would contain evenly distributed high-quality global data gathered with a singular sensor suite, detailed metadata and task-specific annotations to allow for controlled machine learning experiments. As such a dataset does not currently exist, we use three complementary datasets.

First, we experiment on the OSV-5M dataset (Astruc et al., 2024), a large-scale collection of approximately 5.1 million street-level images sourced globally via OpenStreetMap contributions. Crucially for our purposes, each image is accompanied by GPS metadata providing per-sample latitude and longitude coordinates. We use the full 5.1M from across both the train and validation splits.

OSV-5M is particularly well suited to this study for two reasons. First, its global geographic scope and per-image GPS metadata make it possible to partition samples into spatially meaningful groups and directly measure how distributional properties vary with

real-world distance. Second, unlike benchmark datasets curated within a single domain or region, OSV-5M spans a wide range of environments, architectural styles, vegetation types, and lighting conditions across multiple continents, providing the heterogeneity necessary to study distributional shift at a genuinely global scale.

Additionally, there are relatively few publicly available image datasets that simultaneously provide global-scale geographic diversity and precise geospatial metadata suitable for quantitative analysis. While many large-scale computer vision benchmarks contain substantial visual diversity, they often lack per-image latitude and longitude information or are confined to a limited set of geographic regions. OSV-5M therefore represents a uniquely suitable benchmark for investigating the relationship between geographic separation and distributional divergence.

However, OSV-5M does not provide task-specific annotations beyond geographic coordinates and therefore cannot be used directly for the downstream machine learning experiments considered in this work. To measure the impact of geographic distance on the performance of AI perception models in automated driving, we require automotive datasets that provide geographic metadata, sufficient spatial coverage across multiple regions, and a sample size large enough to allow meaningful splits, typically several thousand annotated instances per region. It is also important that the datasets allow control for confounding variables such as time of day, weather, road type, and traffic conditions.

Both the ECP (Braun et al., 2019) and the ZOD (Alibeigi et al., 2023) meet these criteria (Fig. 3). The ECP dataset, collected from 31 cities across 12 countries, provides metadata including city names and a binary classification of time of day (day vs. night) and is used for object detection. The ZOD dataset covers 14 countries and offers more detailed metadata, distinguishing day, night, and twilight, along with annotations for road conditions, road types, and weather. It supports both object detection and lane marking segmentation, enabling controlled analyses of these factors. Together, these datasets serve as valuable resources for investigating domain-incremental learning in automotive perception.

Both datasets primarily cover the Central European region and include multi-class object detection annotations. To remove the impact of class imbalance across subsets, we restrict our analysis to pedestrian detection. This approach helps control for variations in class distribution between regions, which could otherwise confound the geographic effects, and ensures a sufficient number of samples per geographic area. For lane marking segmentation in the ZOD dataset, all lane-related classes were consolidated into a single lane class before training. This decision mirrors our object detection setup, reducing class imbalance and simplifying the analysis. Other datasets were considered (Table 1) but ultimately excluded. Cityscapes (Cordts et al., 2016) and KITTI Object Detection (Geiger et al., 2013) are both high-quality datasets, but they do not contain enough data from diverse regions to support the development of meaningful region-specific models. Waymo Open (Sun et al., 2020) offers adequate data volume but lacks geographic diversity. nuScenes (Caesar et al., 2020), which includes data from only two cities, was excluded due to its limited spatial coverage. This would reduce the analysis to a binary comparison, undermining the study's goal of demonstrating a correlation between geographic distance and model performance.

3.2 Statistical Analysis

To quantify the relationship between geographic distance and model performance, we employed correlation-based approaches tailored to the structure of each dataset. For the ECP dataset, which lacks fine-grained geographic splits, we used the Pearson Correlation Coefficient (r) (Pearson, 1896). Pearson r measures the strength and direction of the linear association between performance metrics (i.e., mAP@50, IoU) and geographic distance, ranging from -1 (perfect negative) to $+1$ (perfect positive). When a significant linear relationship exists, we apply simple linear regression, using R^2 to quantify the proportion of variance explained by distance (Legates and McCabe Jr, 1999). This approach is appropriate because, in simple linear regression, $R^2 = r^2$, linking correlation strength directly to predictive utility. However, because ECP contains only four major regional splits, the resulting correlation estimates are less statistically stable than those obtained from datasets with a larger number of geographic partitions.

In contrast, the ZOD dataset was clustered into ten geographic groups and evaluated using a full cross-cluster protocol. This produced a substantially larger set of correlation estimates, making meta-analytic aggregation more appropriate. To synthesise these results, we employed Fisher's method for combining p-values (Fisher, 1970) and Fisher's z-transformation for combining correlation coefficients (Fisher, 1915). These techniques allow us to assess overall trends and robustness across multiple experiments, providing a more reliable estimate of the relationship between geographic distance and performance degradation.

For both the ECP and ZOD datasets, we report two complementary analyses. First, correlation coefficients and associated significance values are used to quantify the overall relationship between geographic distance and perception performance across experiments. Second, to account for the repeated-measures structure arising from evaluating each model on multiple target regions after training on multiple source regions, we fit mixed-effects models with crossed random intercepts for source (training) region and target (evaluation) region. A separate model is fitted for each architecture.

3.2.1 Combining p-values across experiments (Fisher's Method)

Let the p-values from k independent experiments be $p_1, p_2, \ldots, p_k$ We can compute the Fisher test statistic as:

$$X = -2 \sum_{i=1}^{k} \ln(p_i)$$

Under the null hypothesis (assuming independence), X follows a chi-square distribution with $2k$ degrees of freedom:

$$X \sim \chi^2_{2k}$$

The combined p-value is then.

$$p_{\text{combined}} = 1 - F_{\chi^2_{2k}}(X)$$

where $F_{\chi^2_{2k}}(X)$ is the cumulative distribution function (CDF) of the chi-square distribution with $2k$ degrees of freedom. Although Fisher's method assumes independence between experiments, the geographic clustering and shared model architectures may introduce residual

correlations. This could induce mild anti-conservatism in the aggregated p-values. However, given the consistent and large effect sizes observed across models, the substantive conclusions remain robust.

3.2.2 Combining Correlation Coefficients (Fisher's z-transformation)

Let the Pearson correlation coefficients from k independent experiments be $r_1, r_2, \ldots, r_k$. We can apply Fisher's z-transformation to each correlation:

$$z_i = \frac{1}{2} \ln \left(\frac{1 + r_i}{1 - r_i} \right), \quad i = 1, \ldots, k$$

and compute the mean of the transformed values:

$$\bar{z} = \frac{1}{k} \sum_{i=1}^{k} z_i$$

Transforming back obtains the combined correlation coefficient:

$$r_{\text{combined}} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1}$$

3.2.3 Mixed-effects model

To account for the repeated-measures structure of the experiments, we fitted a linear mixed-effects model with crossed random intercepts for source (training) region and target (evaluation) region. For each architecture and each source–target region pair, we recorded either object detection performance (mAP@50) or lane-segmentation performance (IoU), together with the geographic distance (km) between the centroids of the source and target regions:

$$y_{ij} = \beta_0 + \beta_1 \log(1 + d_{ij}) + u_i + v_j + \varepsilon_{ij},$$

where $u_i \sim N(0, \sigma^2_{\text{source}})$ and $v_j \sim N(0, \sigma^2_{\text{target}})$. This accounts for dependence among evaluations sharing the same source region or target region, as observations involving a common region share the corresponding random effect rather than being treated as independent. Models were fitted by maximum likelihood using `statsmodels`.

Pearson correlations were computed as descriptive summaries only and were not used for statistical inference. Statistical significance of the mixed-effects distance coefficient was instead assessed using a permutation test that preserved the crossed model–dataset structure: dataset identities were randomly shuffled within each model while maintaining the original factorial design, the mixed-effects model was refitted for each permutation (n=2000), and the permutation p-value was defined as the proportion of permuted distance coefficients whose absolute magnitude was greater than or equal to that of the observed coefficient. To determine whether the observed effect was driven solely by in-distribution evaluations, all analyses were repeated after excluding zero-distance observations, with the resulting mixed-effects coefficients, confidence intervals, and correlation statistics reported in Appendix B.

Table 2: Comparison of encoder backbones

	CLIP RN50x64	CLIP ViT-B/32	DINOv2 ViT-B/14	DINOv2 ViT-B/14 (LR)
Architecture	ResNet (CNN)	Transformer	Transformer	Transformer
Core Method	Convolutions	Self-attention	Self-attention	Self-attention
Resolution	448×448	224×224	224×224	518×518
Parameters	∼623M	∼151M	∼86M	∼86M
Output Dimension	1024	512	768	768

3.3 Image Embedding to Measure Domain Shift

Comparing subsets of image datasets presents a unique challenge due to the inherently high dimensionality of image data. Traditional dimensionality reduction techniques such as PCA (Wold et al., 1987) or t-SNE (Maaten and Hinton, 2008) often fail to preserve meaningful semantic relationships when applied to raw pixel data, making them unsuitable for nuanced comparisons (Wattenberg et al., 2016). To address this, we employed Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) encoders, which encode images into high-dimensional semantic vectors, though the relationship between encoder generalisation and train-test similarity warrants careful interpretation (Mayilvahanan et al., 2023). This approach enables comparison based on conceptual similarity rather than low-level visual features, allowing for more interpretable and contextually relevant analysis across dataset subsets (Gadre et al., 2023).

Feature representations are extracted using four pretrained vision encoders, summarised in Table 2: DINOv2 ViT-B/14, DINOv2 ViT-B/14 (higher-resolution), CLIP ViT-B/32, and CLIP RN50x64. These models were chosen due to their strong representation learning capabilities and complementary training paradigms. DINOv2 models learn visual structure through self-supervised objectives, capturing detailed spatial and semantic information without supervision, while CLIP models are trained via image-text contrastive learning and encode semantics aligned with language supervision. The selection spans both convolutional and transformer-based architectures; CLIP RN50x64 uses a CNN backbone, while the remaining three models are Vision Transformers, ensuring that the evaluation is not biased toward a single model family. Furthermore, the encoders operate across a range of input resolutions (224×224, 448×448, and 518×518), allowing us to examine whether the observed distributional structure is consistent across different levels of spatial detail.

The resulting representations form a distribution of shape (N,512), (N,1024) or (N,768), where N denotes the number of images in the subgroup. These high-dimensional vectors serve as semantic embeddings, capturing conceptual features of the images rather than raw pixel information. Each subgroup's encoded vectors are treated as a distribution representative of that region's image content. To compare these regional distributions, three metrics commonly used in the domain shift and distributional distance literature enable a robust assessment of semantic divergence between subsets. These are MMD, SWD, and ED. While all three metrics quantify distributional differences between representation subsets, they emphasise different aspects of meaning representation drift. MMD is most sensitive to subtle, high-dimensional changes in embedding distributions, making it well suited for detecting

fine-grained semantic drift, though its effectiveness depends on kernel choice. SWD provides a geometry-aware notion of drift by approximating optimal transport through random projections, capturing structural changes such as shifts, splits, or reorganisation of semantic clusters, and scaling well to high-dimensional representations. ED offers a parameter-free and interpretable baseline that directly reflects changes in between-subset separation and within-subset coherence, but is less reliable in very high dimensions due to distance concentration. Together, these metrics provide complementary perspectives on representation drift, balancing sensitivity, geometric fidelity, and interpretability.

3.3.1 Maximum Mean Discrepancy (MMD)

MMD (Gretton et al., 2012) provides a kernel-based measure of distributional difference that is particularly well-suited for high-dimensional embeddings. Given two sets of feature vectors $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ from different clusters, the squared MMD with kernel k is defined as:

$$\mathrm{MMD}_k^2(X, Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) \quad + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) \quad - \frac{2}{mn} \sum_{i,j} k(x_i, y_j)$$

We employ a radial basis function kernel to match the default configuration of the comparison metric:

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

The bandwidth parameter γ is automatically selected as the inverse of the median pairwise squared Euclidean distance in the combined dataset:

$$\gamma = \frac{1}{\mathrm{median}(\{\|x_i - x_j\|^2 : x_i \in X \cup Y, x_j \in X \cup Y\})}$$

3.3.2 Sliced Wasserstein Distance (SWD)

The SWD provides an efficient approximation to the Wasserstein distance (Rabin et al., 2011) by leveraging random projections. Given two point sets $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$ in $\mathbb{R}^d$, it is defined as:

$$\mathrm{SWD}(X, Y) = \frac{1}{L} \sum_{l=1}^{L} W_1(\mathcal{P}_{\theta_l} X,\ \mathcal{P}_{\theta_l} Y)\,,$$

where L is the number of random projections, θ_l are random unit vectors sampled from the unit sphere, $\mathcal{P}_{\theta_l}$ denotes projection onto direction θ_l, and W_1 is the 1D Wasserstein distance.

Random projection directions are generated as:

$$\theta_l = \frac{g_l}{\|g_l\|_2}$$

where $g_l \sim \mathcal{N}(0, I_d)$ are i.i.d. Gaussian vectors. The projection operation for a point set is:

$$\mathcal{P}_{\theta_l} X = \{\langle x_1, \theta_l \rangle, \ldots, \langle x_m, \theta_l \rangle\}$$

For computational efficiency, the 1D Wasserstein distance is approximated using quantile functions:

$$W_1(U,V) \approx \frac{1}{K}\sum_{k=1}^{K} |F_U^{-1}(q_k) - F_V^{-1}(q_k)|$$

where F_U^{-1} and F_V^{-1} are the inverse CDFs of the projected distributions, and $q_k = k/(K+1)$ are equally spaced quantiles.

3.3.3 Energy Distance (ED)

The ED (Székely and Rizzo, 2013) is based on the expected distance between sample points from different distributions and within the same distribution. For two distributions X and Y, it is defined as:

$$\mathcal{E}(X,Y) = 2\mathbb{E}[\|X-Y\|] - \mathbb{E}[\|X-X'\|] - \mathbb{E}[\|Y-Y'\|]$$

where X, X' are independent random variables from the first distribution and Y, Y' are independent from the second distribution.

For finite samples $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, the empirical energy distance estimator is:

$$\hat{\mathcal{E}}(X,Y) = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|x_i - y_j\| \quad - \frac{1}{m(m-1)}\sum_{i\neq j}\|x_i - x_j\| \quad - \frac{1}{n(n-1)}\sum_{i\neq j}\|y_i - y_j\|$$

The energy distance decomposes into three interpretable components:
Cross-term (between distributions):

$$D_{XY} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|x_i - y_j\|$$

Within-term for X:

$$D_{XX} = \frac{1}{m(m-1)}\sum_{i\neq j}\|x_i - x_j\|$$

Within-term for Y:

$$D_{YY} = \frac{1}{n(n-1)}\sum_{i\neq j}\|y_i - y_j\|$$

The final energy distance is:

$$\hat{\mathcal{E}}(X,Y) = 2D_{XY} - D_{XX} - D_{YY}$$

3.4 Machine Learning Models

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4070 Ti SUPER GPU (CUDA 13.0), an Intel Core i9-14900F CPU, and 32 GB DDR5 RAM (2×16 GB, 5600 MT/s), running Ubuntu 22.04.5 LTS. To ensure reproducibility, all experiments were performed using a fixed random seed of 42, which was used for weight initialisation and data-loader shuffling. Models were trained and evaluated using an input resolution of 1024×1024 pixels, except for U-Net, which used an input resolution of 512×512 pixels.

3.4.1 Object Detection Models

To test the effect of training data distance on object detection performance, we utilised models representing the three main architectural paradigms: (1) one-stage, (2) two-stage, and (3) transformer-based networks. The following list summarises the architecture of our chosen models:

1. **YOLOv10 (n, m, x)** - One-stage detectors representing nano, medium, and extra-large variants to assess the impact of model capacity on performance. We selected YOLOv10 over newer variants to ensure the models did not incorporate attention mechanisms. We used the Ultralytics implementation with 2.71M, 16.49M, and 31.66M parameters, respectively (Wang et al., 2024). All models were trained for up to 100 epochs using SGD, with a patience of 10 epochs for early stopping. The YOLOv10n and YOLOv10m models used a learning rate of 0.001 with batch sizes of 24 and 8 respectively, while YOLOv10x used a learning rate of 0.0005 and a batch size of 4.

2. **Faster R-CNN (ResNet18, ResNet50)** - Two-stage Region-based Convolutional Neural Networks using ResNet backbones of varying depths with a Region Proposal Network (RPN) that shares full-image convolutional features. We used the PyTorch implementation with 28.29M and 41.41M parameters, respectively (Ren et al., 2016). Both models were trained for up to 150 epochs using SGD with a learning rate of 0.005 and a patience of 10 epochs for early stopping. ResNet18 used batch sizes of 10 and ResNet50 used batch sizes of 12.

3. **RT-DETRv2 L** - A large transformer-based detector that provides real-time detection capabilities with attention mechanisms, representing the state-of-the-art in transformer-based object detection. We used the author's own GitHub implementation with 42.9M parameters (Zhao et al., 2024). Trained for up to 100 epochs using AdamW, with a patience of 10 epochs for early stopping and used a learning rate of 0.001 with batch sizes of 8

All object detection correlations were calculated using the pedestrian mAP@50 metric against the cluster distance.

3.4.2 Segmentation Models

Evaluations for semantic lane segmentation were conducted on the ZOD dataset. To capture the diversity of architectural approaches, we selected models spanning five primary paradigms: (1) classical encoder-decoder architectures, (2) lightweight real-time models, (3) atrous convolution-based methods, (4) transformer-based hierarchical encoders, and (5) unified detection-segmentation networks. YOLO models used the Ultralytics implementation, and all other models used the mmsegmentation implementation. The following list summarises the architecture of our chosen models:

1. **U-Net** - A classical encoder-decoder architecture with skip connections that has become a foundational model for semantic segmentation tasks with 29.07M parameters (Ronneberger et al., 2015). Training hyperparameters were based on the `unet-s5-d16_fcn_4xb4-160k_cityscapes-512x1024` configuration.

2. **Fast-SCNN** - A lightweight real-time semantic segmentation model designed for efficient inference with reduced computational requirements with 1.48M parameters (Poudel et al., 2019). Training hyperparameters were based on the `fast_scnn_8xb4-160k_cityscapes-512x1024` configuration.

3. **DeepLabv3+** - An encoder-decoder model that combines atrous spatial pyramid pooling (ASPP) with a decoder module to refine segmentation results, particularly along object boundaries, with 43.65M parameters (Chen et al., 2018). Training hyperparameters were based on the `deeplabv3plus_r50-d8_4xb2-300k_mapillay_v1_65-1280x1280` configuration.

4. **SegFormer (B0, B2)** - A transformer-based segmentation model that leverages hierarchical Transformer encoders without positional encoding with 3.8M and 24.73M parameters, respectively (Xie et al., 2021). Training hyperparameters for B0 were based on the `segformer_mit-b0_8xb1-160k_cityscapes-1024x1024` and B2 `segformer_mit-b2_8xb1-160k_cityscapes-1024x1024` configuration.

5. **Swin-UperNet** - A hierarchical vision transformer backbone (Liu et al., 2021) combined with a Unified Perceptual Parsing (UPerNet) decoder (Xiao et al., 2018) that aggregates multi-scale features, with 59.87M parameters. Training hyperparameters were based on the `swin-tiny-patch4-window7_upernet_1xb8-20k_levir-256x256` configuration.

6. **YOLOv8-seg (n, m, l)** - One-stage segmentation models representing nano, medium, and large variants that integrate detection and segmentation in a unified framework with 3.26M, 27.24M, and 45.97M parameters, respectively (Jocher et al., 2023). All models were trained for up to 100 epochs using SGD with a learning rate of 0.001 and a patience of 10 epochs for early stopping. YOLOv8n, YOLOv8m, and YOLOv8l used batch sizes of 20, 8, and 6, respectively.

7. **YOLOv11-seg (n, m, l)** - The latest iteration of the YOLO segmentation family with nano, medium, and large variants featuring enhanced spatial attention mechanisms through C2PSA blocks to focus on key image features during training with 2.84M, 22.36M, and 27.62M parameters, respectively (Khanam and Hussain, 2024). All models were trained for up to 100 epochs using SGD with a learning rate of 0.001 and a patience of 10 epochs for early stopping. YOLOv11n, YOLOv11m, and YOLOv11l used batch sizes of 20, 8, and 6, respectively.

Semantic segmentation outputs from YOLO-based models were converted to semantic predictions by merging semantic masks into a single class-wise pixel map, resolving overlaps by confidence ordering. Evaluation was then performed using standard semantic segmentation metrics. All segmentation correlations were calculated using the lane IoU metric against cluster distance.

4 Results

This section begins with the results from the comparison methods applied to both datasets, including a controlled confounder test conducted on ZOD. Performance of models trained in

various data subsets and evaluated in test sets sampled at increasing geographic distances, to measure the degradation in performance. Finally, the section concludes with an evaluation of task annotations to assess whether they are influenced by geographic distances in the same way as the models.

For all experiments, the distance between the clusters is calculated using the latitude and longitude of the clusters centroids with the Haversine distance as follows. Let two points on the Earth be given by (ϕ_1, λ_1) and (ϕ_2, λ_2), where latitude ϕ and longitude λ are expressed in radians. The Haversine distance d is defined as:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right),$$

and R is the Earth's radius ($R = 6371\,\text{km}$).

4.1 Embedded Similarity

All image data was encoded using encoder architectures listed in Table 2. The resulting embeddings were compared using three dissimilarity metrics (MMD, SWD, ED), and the outcomes were analysed relative to the geographic distance between the corresponding data regions. Appendix C and Figures 4, 5, and 6 each present results in a 3×4 grid, where rows correspond to similarity metrics (MMD, SWD, ED) and columns correspond to encoder backbones. Each subplot displays scatter points with a second-order polynomial fit. Tables 3, 5, 6, and 7 report the corresponding numerical values, including the coefficient of determination and Pearson correlation coefficient, quantifying the strength of association between geographic separation and embedding divergence.

4.1.1 OSV-5M Comparison

For the 100-cluster partition, we used size-constrained K-Means clustering (Bradley et al., 2000). For the 200-cluster partition, as shown in Fig 2, resource limitations prevented the direct application of constrained K-Means at the full dataset scale. We therefore employed a computationally efficient approximation consisting of standard K-Means centroid estimation followed by a capacity-constrained assignment step to obtain approximately balanced clusters.

Across all 24 experiment configurations (4 encoders × 3 metrics × 2 clustering granularities), distributional divergence correlates positively and significantly with log-scaled geographic distance between cluster centroids. Full Pearson and R^2 values are reported in Table 3

To assess whether this correlational structure is specific to a particular encoder or reflects a genuine property of the underlying image data, Table 4 reports Pearson correlations between the pairwise MMD matrices produced by each encoder pair (200-cluster partition). Agreement is high throughout: within-family correlations reach 0.967 for the two CLIP variants and 0.996 for the two DINOv2 variants, while cross-family correlations fall in the range [0.884, 0.896]. The near-identical DINOv2 results further confirm that input resolution has negligible influence on the downstream distributional comparison. The full pairwise scatter plots are provided in Figure 10 in Appendix A

Clusters	Encoder	Metric	R^2	Pearson
100	DINOv2 ViT-B/14	MMD	0.2084	0.456481
		SWD	0.2195	0.468522
		ED	0.1651	0.406267
	DINOv2 ViT-B/14 (LR)	MMD	0.2050	0.452727
		SWD	0.2176	0.466440
		ED	0.1636	0.404418
	CLIP ViT-B/32	MMD	0.2196	0.468632
		SWD	0.2261	0.475451
		ED	0.1777	0.421567
	CLIP RN50x64	MMD	0.2981	0.546030
		SWD	0.3110	0.557636
		ED	0.2562	0.506123
200	DINOv2 ViT-B/14	MMD	0.2254	0.474788
		SWD	0.2441	0.494097
		ED	0.1812	0.425675
	DINOv2 ViT-B/14 (LR)	MMD	0.2170	0.465789
		SWD	0.2368	0.486616
		ED	0.1740	0.417165
	CLIP ViT-B/32	MMD	0.2364	0.486167
		SWD	0.2461	0.496110
		ED	0.1923	0.438505
	CLIP RN50x64	MMD	0.3040	0.551318
		SWD	0.3226	0.567949
		ED	0.2562	0.506113

Table 3: R^2 and Pearson correlation of embedding similarity metrics across encoders and clustering levels against geographical distance (All p-values are < 0.05)

	CLIP ViT-B/32	DINOv2 ViT-B/14	DINOv2 ViT-B/14 (LR)
CLIP RN50x64	0.967	0.896	0.891
CLIP ViT-B/32	–	0.890	0.884
DINOv2 ViT-B/14		–	0.996

Table 4: Pearson correlation between the pairwise MMD matrices produced by each pair of encoders (OSV-5M, 200-cluster partition). Each matrix is symmetric with a zero diagonal, so correlations are computed over the 19,900 unique cluster pairs (200 $\times$ 199 / 2). The table lists the six unique encoder pairs.

4.1.2 ECP Comparison

The ECP dataset is pre-divided into 31 cities across 12 countries, providing a natural split into regional subgroups Fig 3a. This predefined structure allows for straightforward geographic comparisons, as the distance between subgroups can be approximated using the latitude and longitude of the respective cities. However, a key limitation of this dataset is

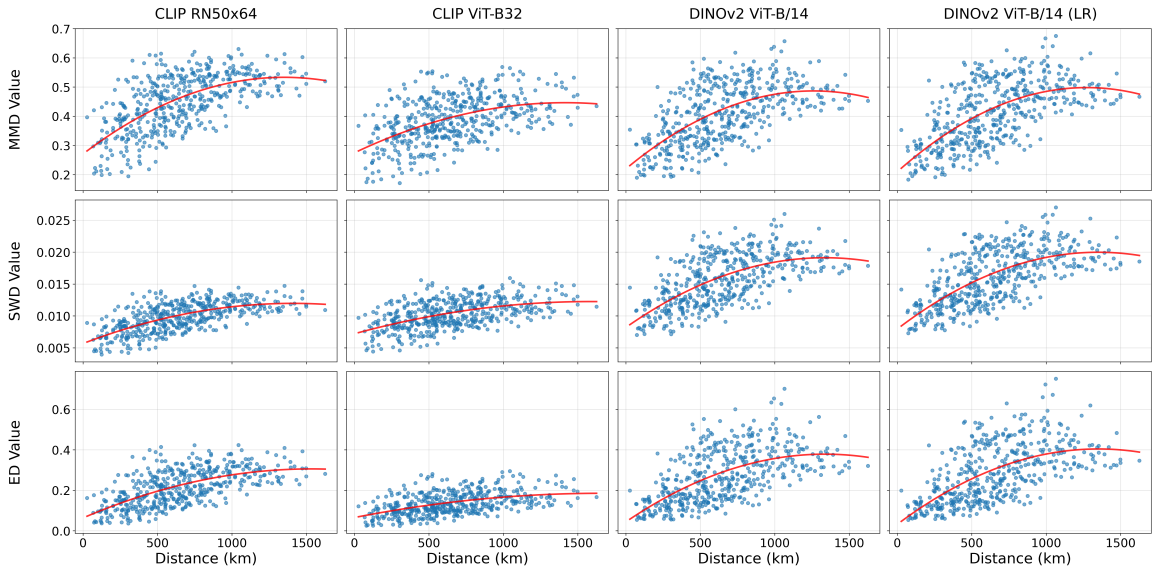


Figure 4: Geographic distance versus embedding divergence for ECP's 31 cities. Rows correspond to MMD, SWD, and ED, and columns correspond to RN50x64, ViT-B/32, ViT-B/14, and ViT-B/14 (LR). Scatter points show pairwise city comparisons and curves show second-order polynomial fits.

Table 5: Correlation between geographic distance and embedding divergence for ECP.

Encoder	Metric	R^2	Pearson
CLIP RN50x64	MMD	0.3784	0.6151
	SWD	0.3919	0.6260
	ED	0.3761	0.6133
CLIP ViT-B/32	MMD	0.2318	0.4815
	SWD	0.2544	0.5044
	ED	0.2262	0.4756
DINOv2 ViT-B/14	MMD	0.3363	0.5799
	SWD	0.3763	0.6134
	ED	0.3265	0.5714
DINOv2 ViT-B/14 (LR)	MMD	0.3583	0.5986
	SWD	0.3997	0.6322
	ED	0.3463	0.5884

Higher values indicate a stronger association between geographic separation and embedding divergence.

the imbalance in the number of samples across cities, which may introduce bias or affect the robustness of comparative analyses. The results are presented in Table 5.

4.1.3 ZOD 100 Clusters

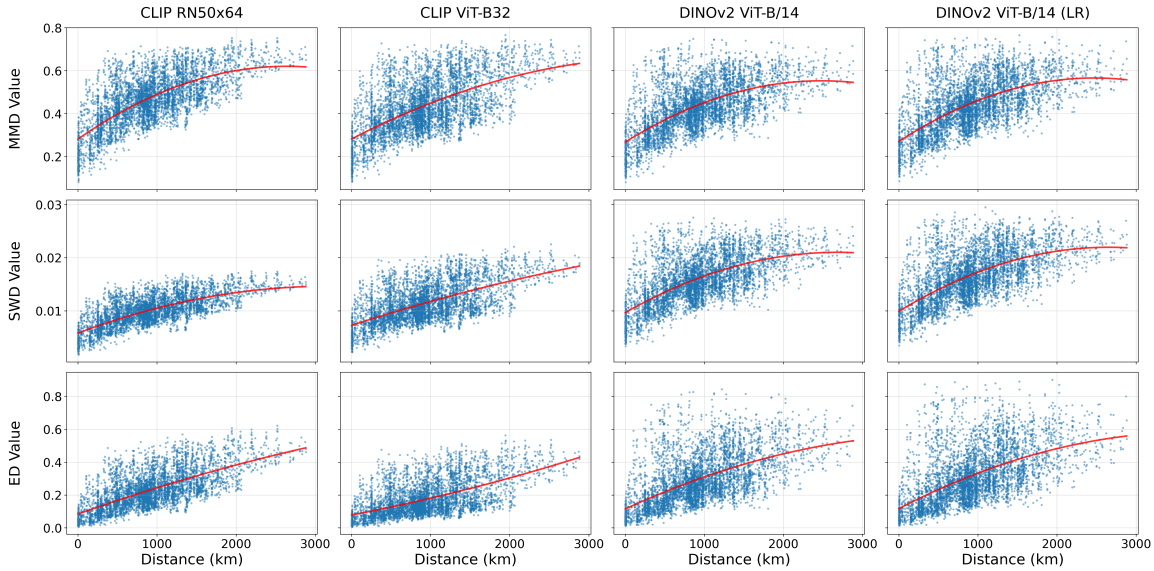


Figure 5: Geographic distance versus embedding divergence for ZOD's 100 geographic clusters. Rows correspond to MMD, SWD, and ED, and columns correspond to RN50x64, ViT-B/32, ViT-B/14, and ViT-B/14 (LR). Scatter points show pairwise cluster comparisons and curves show second-order polynomial fits.

Unlike the ECP dataset, which is organised by city, the ZOD dataset was collected across a diverse range of driving environments, including urban, rural, highway, and city settings, and does not provide clearly defined geographic boundaries for subgroup analysis. This introduces a methodological trade-off in the comparison process. Larger subgroups offer better statistical representation of the underlying data distribution, but correlation analyses benefit from a greater number of distinct comparison points to improve statistical power and confidence in observed relationships. To balance these needs, the dataset was clustered into 100 groups of 1,000 images each. This approach ensures that all images within a subgroup are spatially close to their cluster centroids, which is better for comparisons. Results are presented in Table 6

4.1.4 ZOD Control for Confounders

To further investigate the correlation between geographic distance and semantic similarity, a controlled test was conducted to reduce variance within the image subgroups. The dataset was filtered to include only daytime images captured during city driving. Additionally, any cluster containing fewer than 100 images was removed. The remaining clusters were then encoded and compared using the same methods as before. Results are presented in Table 7

Table 6: Correlation between geographic distance and embedding divergence for ZOD.

Encoder	Metric	R^2	Pearson
CLIP RN50x64	MMD	0.4703	0.6858
	SWD	0.5051	0.7107
	ED	0.4777	0.6910
CLIP ViT-B/32	MMD	0.3582	0.5986
	SWD	0.3800	0.6165
	ED	0.3449	0.5873
DINOv2 ViT-B/14	MMD	0.3671	0.6059
	SWD	0.3839	0.6196
	ED	0.3339	0.5779
DINOv2 ViT-B/14 (LR)	MMD	0.3600	0.6000
	SWD	0.3786	0.6153
	ED	0.3244	0.5696

Higher values indicate a stronger association between geographic separation and embedding divergence.

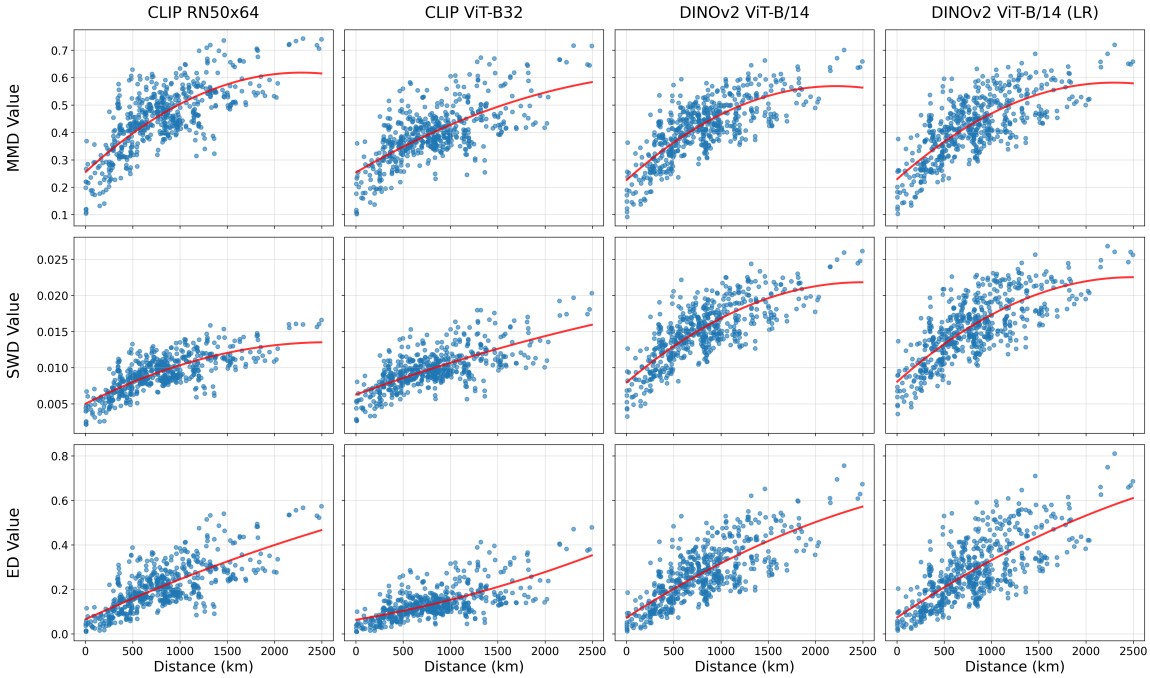


Figure 6: Geographic distance versus embedding divergence for the controlled-confounder subset of ZOD, consisting only of daytime city-driving images. Rows correspond to MMD, SWD, and ED, and columns correspond to RN50x64, ViT-B/32, ViT-B/14, and ViT-B/14 (LR). Scatter points show pairwise cluster comparisons and curves show second-order polynomial fits.

Table 7: Correlation between geographic distance and embedding divergence for ZOD controlled confounder.

Encoder	Metric	R^2	Pearson
CLIP RN50x64	MMD	0.5122	0.7157
	SWD	0.5343	0.7310
	ED	0.5290	0.7273
CLIP ViT-B/32	MMD	0.4544	0.6741
	SWD	0.4696	0.6853
	ED	0.4422	0.6650
DINOv2 ViT-B/14	MMD	0.5534	0.7439
	SWD	0.5810	0.7623
	ED	0.5721	0.7564
DINOv2 ViT-B/14 (LR)	MMD	0.5133	0.7164
	SWD	0.5445	0.7379
	ED	0.5291	0.7274

Higher values indicate a stronger association between geographic separation and embedding divergence.

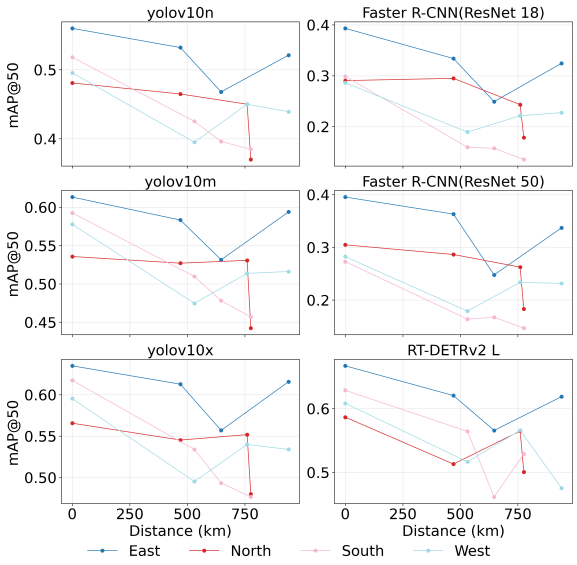
4.1.5 Annotation Spatial Variation

To further validate that geographic distance affects the underlying data distribution rather than the annotation patterns, the annotation spatial structure was examined. Heat maps of pedestrian detection and lane segmentation annotations for each training cluster were compiled and compared using cosine similarity, SWD, and MSE. Unlike the strong correlations observed in encoded embeddings, annotation distributions showed no significant relationship with geographic distance, with all metrics yielding ($p > 0.05$) across both tasks. While segmentation annotations exhibited slightly more spatial variation than object detection annotations, likely because pixel-wise labels are more tightly coupled to image appearance than bounding boxes, neither reached statistical significance. This finding implies that any performance degradation in models trained on these regions stems from geographic variation in the raw visual data itself, not from systematic differences in how annotations are distributed across regions.

4.2 Machine Learning Perception Results

4.2.1 ECP Object Detection

To test whether the domain shift observed in the image data affects model performance, models were trained on subgroups of images and evaluated across all subgroups to measure the impact on performance. For the ECP dataset, subgroups were defined to preserve country boundaries while ensuring each contained at least 5,000 training images (Fig. 3), where the colour indicates the different splits and the size of each circle conveys the amount of data available in each city. Models evaluated include Faster R-CNN (ResNet-18/50), RT-DETRv2 L, and YOLOv10 (n/m/x), with detection performance measured using mAP@50. Results are presented in Fig. 7a and Table 7b, showing Pearson correlation and correspond-



(a)

Model	Pearson	P-value	95% CI
Faster R-CNN ResNet-50	-0.521	0.039	[-0.808, -0.035]
Faster R-CNN ResNet-18	-0.512	0.043	[-0.805, -0.028]
RT-DETRv2	-0.585	0.017	[-0.838, -0.126]
YOLOv10n	-0.536	0.032	[-0.815, -0.054]
YOLOv10m	-0.525	0.037	[-0.810, -0.040]
YOLOv10x	-0.532	0.034	[-0.814, -0.050]

Negative correlations indicate declining detection performance with increasing geographic distance from the training region.

(b)

Figure 7: ECP object detection results. **(a)** mAP@50 vs. geographic distance for each model and training region. **(b)** Pearson correlation between mAP@50 and geographic distance from the training region.

ing p-values between mAP@50 and geographic distance from the training region. Mixed-effects and permutation results are given in Table 10 (Appendix B). Because ECP contains only four regions ($N = 16$ source–target evaluations per detector), the Pearson confidence intervals above treat evaluations as independent and are likely too narrow, so we regard the ECP results as exploratory.

4.2.2 ZOD Object Detection

The ZOD dataset was clustered into 10 subgroups, each containing 10,000 images, using Constrained K-Means Clustering (Bradley et al., 2000) based on the latitude and longitude metadata of the images. These clusters were randomly divided into training and testing sets using an 80/20 split (Fig. 3). Each model was trained on a single cluster and subsequently evaluated on the test data from all remaining clusters, with performance compared against the distance between cluster centroids. Models evaluated include YOLOv10 (n/m/x), RT-DETRv2 L, and Faster R-CNN (ResNet-18/50), with detection performance measured using mAP@50. Fig. 8 presents each model as a consistently colour-coded curve across cluster subplots, enabling a comparative analysis of distance robustness across architectures. Aggregate correlation results are reported in Table 8 using Fisher's z-transformation, pooled across clusters via Fisher's Method. Along with the mixed-effects model in Appendix B Table 11

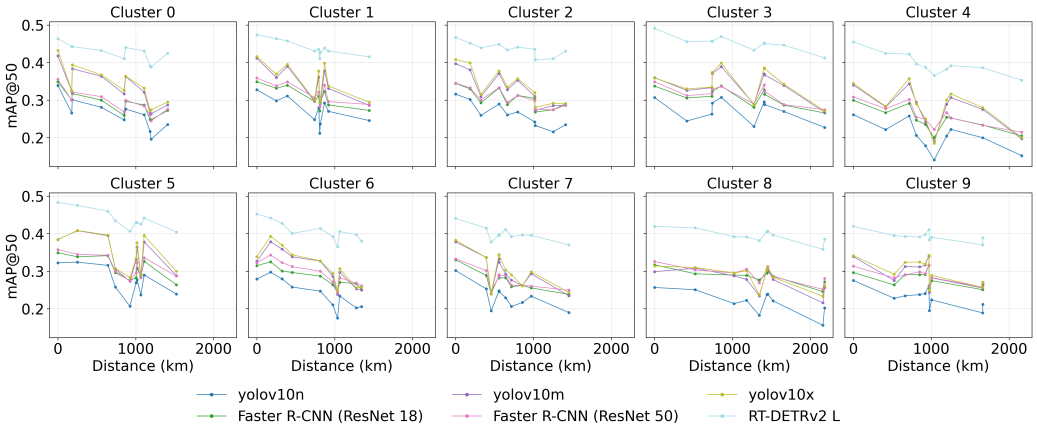


Figure 8: mAP@50 vs. geographic distance from the training cluster for each model across all ZOD clusters.

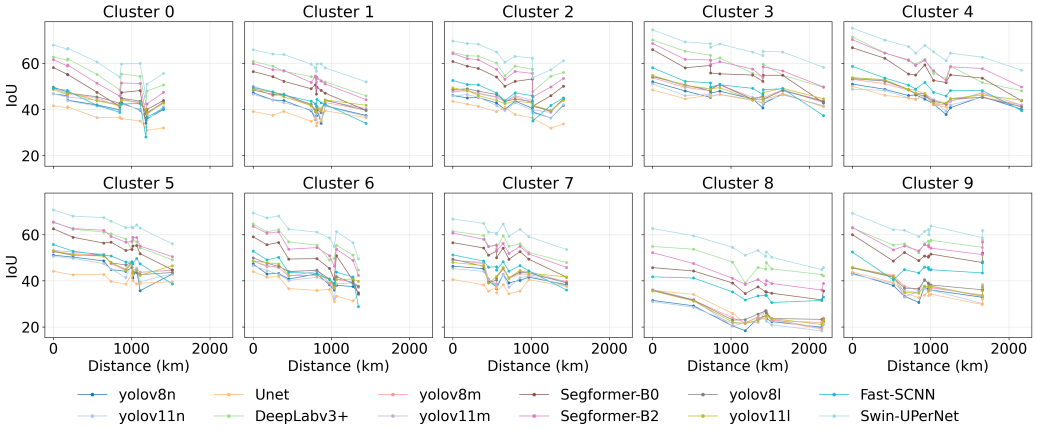


Figure 9: IoU vs. geographic distance from the training cluster for each segmentation model across all ZOD clusters.

Table 8: Detection performance correlation with geographic distance on the ZOD dataset, including and excluding zero-point observations.

	Including zero-point observations			Excluding zero-point observations		
Model	Fisher z	Combined r	95% CI	Fisher z	Combined r	95% CI
Faster R-CNN ResNet-18	-0.656	-0.575	[-0.693, -0.427]	-0.398	-0.378	[-0.543, -0.186]
Faster R-CNN ResNet-50	-0.692	-0.599	[-0.712, -0.457]	-0.436	-0.410	[-0.569, -0.222]
RT-DETRv2	-0.638	-0.563	[-0.684, -0.413]	-0.446	-0.418	[-0.576, -0.231]
YOLOv10n	-0.786	-0.656	[-0.755, -0.528]	-0.564	-0.511	[-0.650, -0.340]
YOLOv10m	-0.790	-0.659	[-0.757, -0.531]	-0.584	-0.525	[-0.660, -0.357]
YOLOv10x	-0.790	-0.659	[-0.757, -0.531]	-0.586	-0.527	[-0.662, -0.359]

Negative Fisher z values indicate a negative association between detection performance and geographic distance from the training cluster. Combined r values represent the meta-analytically combined correlations across the 10 clusters. Confidence intervals are 95% intervals for the combined correlation.

4.2.3 ZOD Segmentation

Using the same cluster split as the ZOD object detection experiments, the analysis was extended to lane marking semantic segmentation, which is also annotated in the ZOD dataset. Models evaluated include YOLOv8 (n/m/l), YOLOv11 (n/m/l), U-Net, Fast-SCNN, SegFormer (B0/B2), Swin-UPerNet, and DeepLabv3+, with segmentation performance measured using IoU. These models were trained to examine whether the domain shift associated with geographic distance extends beyond object detection to other perception tasks. The same training and evaluation methodology was applied as in section 4.2.2. Fig. 9 presents each model as a consistently colour-coded curve across cluster subplots. Aggregate correlation results are reported in Table 9 using Fisher's z-transformation pooled across clusters via Fisher's Method. Along with the mixed-effects model in Appendix B Table 12

5 Discussion

Our study demonstrates strong correlations between geographic distance and both semantic divergence in encoder embeddings and downstream perception task performance. While we present converging evidence across multiple models, datasets, and evaluation metrics, we acknowledge that our findings establish correlation rather than definitive causation. The consistent patterns across representation-level and task-level analyses, combined with our controlled experiments, provide substantial support for the hypothesis that geographic separation is associated with measurable domain shift in automotive perception systems, with distance acting as a proxy for latent environmental factors.

5.1 Multi-Encoder Geographic Structure

Across all four representation encoders and all datasets, we observe a consistent positive relationship between geographic distance and distributional divergence. 0.40-0.57 for OSV-5M, 0.48-0.63 for ECP, 0.57-0.71 for ZOD, and strengthening to 0.66-0.76 for the ZOD

Table 9: Segmentation correlation with geographic distance on the ZOD dataset, including and excluding zero-point observations.

	Including zero-point observations			Excluding zero-point observations		
Model	Fisher z	Combined r	95% CI	Fisher z	Combined r	95% CI
YOLOv8n	-0.678	-0.590	[-0.705, -0.445]	-0.608	-0.543	[-0.674, -0.378]
YOLOv8m	-0.710	-0.611	[-0.721, -0.471]	-0.651	-0.572	[-0.697, -0.414]
YOLOv8l	-0.707	-0.609	[-0.719, -0.468]	-0.654	-0.574	[-0.698, -0.417]
YOLOv11n	-0.648	-0.570	[-0.690, -0.421]	-0.601	-0.538	[-0.670, -0.372]
YOLOv11m	-0.682	-0.593	[-0.707, -0.449]	-0.619	-0.551	[-0.680, -0.388]
YOLOv11l	-0.676	-0.589	[-0.704, -0.444]	-0.621	-0.552	[-0.681, -0.389]
U-Net	-0.554	-0.503	[-0.638, -0.341]	-0.411	-0.390	[-0.552, -0.198]
Fast-SCNN	-0.724	-0.620	[-0.727, -0.482]	-0.505	-0.466	[-0.614, -0.286]
SegFormer-B0	-0.771	-0.648	[-0.749, -0.517]	-0.574	-0.518	[-0.655, -0.349]
SegFormer-B2	-0.771	-0.647	[-0.749, -0.517]	-0.556	-0.505	[-0.645, -0.333]
Swin-UPerNet	-0.811	-0.670	[-0.766, -0.545]	-0.600	-0.537	[-0.669, -0.371]
Deeplabv3+	-0.920	-0.726	[-0.807, -0.617]	-0.716	-0.614	[-0.728, -0.466]

Negative Fisher z values indicate a negative association between segmentation performance and geographic distance from the training cluster. Combined r values represent the meta-analytically combined correlations across the 10 clusters. Confidence intervals are 95% intervals for the combined correlation.

controlled confounder test. This consistency is notable because the encoders are trained using fundamentally different representation learning objectives. CLIP aligns visual representations with language supervision, whereas DINOv2 learns visual structure through self-supervision alone. Despite these differences, all encoders produce similar geographic trends, and the corresponding similarity matrices exhibit strong agreement across encoder families. These results suggest that the observed geographic structure is not merely an artefact of a particular representation model, but instead reflects an underlying property of the image data itself. The inclusion of the globally distributed OSV-5M dataset further strengthens this conclusion by demonstrating that distance-dependent representation drift is observable beyond the European automotive datasets used for downstream perception experiments

5.2 Downstream Perception Task Performance

The geographic relationships observed in encoder representation space translate to downstream task performance degradation. Both object detection and segmentation models exhibit statistically significant negative correlations between performance and geographic distance across multiple architectures and model families.

For object detection on the ECP dataset, all six tested models (Faster R-CNN variants, RT-DETRv2, and YOLOv10 variants) show Pearson correlations ranging from -0.512 to -0.585, with p-values < 0.05. Similar trends hold on the ZOD dataset with combined correlations between -0.563 and -0.659, and between -0.378 and -0.527 when excluding zero-point observations. Semantic segmentation models demonstrate particularly strong distance-performance relationships. Across 12 architectures spanning classical encoder-decoder networks (U-Net), lightweight models (Fast-SCNN), atrous convolution methods

(DeepLabv3+), and unified detection-segmentation frameworks (YOLOv8/v11-seg), with combined correlations between -0.503 and -0.726, and combined correlations between -0.39 and -0.614 when excluding zero-point observations.

The consistently stronger correlations in segmentation compared to object detection may reflect the nature of the annotations and visual features involved. Our analysis found that neither task showed strong geographic correlation in annotation spatial density ($p > 0.05$). However, segmentation exhibited slightly more spatial variation than detection. This may occur because pedestrian bounding boxes maintain relatively consistent shapes and sizes across regions, whereas pixel-wise lane marking appearance varies substantially with geographic factors such as road surface materials, marking paint types, weather-induced degradation patterns, and regional infrastructure standards. The dense, pixel-level predictions required for segmentation appear more sensitive to these geographic variations in visual appearance.

5.3 Continuous versus Binary Geographic Shift

A key question raised by our analysis is whether geographic domain shift should be interpreted as a binary in-domain versus out-of-domain phenomenon or as a continuous process. To investigate this, we repeated all perception analyses after excluding zero-distance evaluations. For the ZOD dataset, statistically significant degradation remained across both detection and segmentation models, indicating that performance declines are not driven solely by the contrast between source and target domains. Instead, the results suggest that increasing geographic separation continues to influence performance even among already out-of-domain regions.

In contrast, the ECP results do not support a continuous effect once in-domain evaluations are removed. With only four coarse regions, excluding self-evaluations leaves twelve observations per detector (four training regions and three out-of-domain targets), and the Pearson correlations fall to approximately zero ($r = -0.074$ to 0.002). The full-data ECP correlations are therefore also consistent with a binary in-domain/out-of-domain pattern in which the in-domain diagonal accounts for most of the association. Consequently, the evidence for continuous geographic degradation rests on the larger and more geographically detailed ZOD experiments, and we regard ECP as inconclusive on this question.

This interpretation is further supported by the permutation analysis. For the larger ZOD experiments, permutation tests that preserved the crossed model–dataset structure remained statistically significant across both object detection and segmentation models, confirming that the observed distance effects were unlikely to arise from the repeated-measures design alone. In contrast, permutation tests for the reduced ECP analysis were no longer significant after excluding zero-distance evaluations ($p_{\text{perm}} = 0.34$–0.65). However, the corresponding mixed-effects models continued to estimate statistically significant negative distance coefficients for all detectors, with confidence intervals consistently excluding zero. Given that the exclusion analysis contains only 12 observations (four models evaluated on three out-of-domain regions), the non-significant permutation results are most plausibly explained by limited statistical power rather than the absence of a distance effect. We therefore treat the ECP result as inconclusive regarding continuity, providing neither evidence for nor against it. Taken together, the ZOD results suggest that geographic domain

shift behaves as a continuous phenomenon, although sufficiently large and geographically diverse datasets are required to estimate its magnitude reliably.

5.4 Practical Implications for Deployment

The results have practical implications for the deployment and validation of automotive perception systems. Across multiple architectures and tasks, performance generally decreases as the geographic separation between training and evaluation regions increases. This suggests that conventional validation strategies based primarily on random train-test splits may underestimate deployment risk when models are transferred to new operating regions. In safety-critical applications such as pedestrian detection and lane understanding, even relatively modest degradations may translate into increased false negatives or reduced robustness under real-world conditions. Geographic coverage should therefore be considered an important component of dataset design, validation protocols, and data collection strategies, particularly for systems expected to operate across multiple regions or countries.

5.5 Worst-Region Performance

Because perception is safety-critical, average performance alone is insufficient: a deployed system must remain reliable across its operating regions, including regions that differ from those seen during training. For each training region, we therefore measured the performance drop from in-domain evaluation to the worst-performing target region, averaging across architectures. For ZOD lane segmentation, the mean IoU drop ranges from 11 to 16 percentage points across training clusters. For ZOD pedestrian detection, the mean mAP@50 drop ranges from 7 to 12 points, corresponding to relative declines of 22.0% to 37.1%; the largest decline for an individual model is 46.3% (YOLOv10x trained on Cluster 4). For ECP, the mean drop ranges from 10 points for North to 14 points for South, corresponding to relative declines of 21.9% to 33.2%. These substantial performance losses are safety-relevant because reduced segmentation and detection performance can increase the likelihood that critical road users or scene boundaries are missed or incorrectly localized. Consequently, strong in-region validation alone may not provide sufficient evidence of reliable perception when a system is deployed in geographically different operating conditions.

6 Limitations

6.1 Correlation vs. Causation

This study establishes a robust association between geographic separation, representation divergence, and downstream perception performance. However, the observational nature of the analysis prevents causal conclusions from being drawn. Geographic distance itself is unlikely to be the direct cause of performance degradation. Instead, it may reflect a complex mixture of latent factors including infrastructure design, traffic conventions, environmental characteristics, vehicle fleets, and cultural variations in road usage. Although the controlled ZOD experiments and annotation analyses reduce the influence of several potential confounders, unmeasured variables may still contribute to the observed relationships. Future work using causal inference techniques or controlled data collection protocols will be required to isolate the individual factors responsible for geographic domain shift.

6.2 Geographic Distance as a Proxy Variable

Geographic distance should therefore be interpreted as a practical proxy for underlying environmental and operational differences rather than as a physically meaningful causal quantity. Two regions separated by a short geographic distance may nevertheless exhibit substantial differences in road markings, architecture, vegetation, traffic behaviour, or climate, while geographically distant regions may sometimes share surprisingly similar visual characteristics. The stronger correlations observed after controlling for driving context suggest that environmental factors influence the observed relationship. Consequently, geographic distance should be viewed as a convenient and measurable surrogate for a broader set of latent variables that are difficult to quantify directly but are likely to govern real-world generalisation performance

6.3 Geographic Coverage and Generalisability

Although the representation-level analysis was extended using the globally distributed OSV-5M dataset, the downstream perception experiments are still limited to European automotive datasets. Europe provides substantial geographic diversity, but road infrastructure, driving conventions, regulatory environments, and environmental conditions are generally more homogeneous than those encountered globally. Consequently, the quantitative relationships reported in this study should not be assumed to apply universally. Evaluating geographically distributed perception datasets from Asia, Africa, North America, South America, and Australia will be necessary to determine whether the observed trends represent a broadly general property of automotive perception systems or are partially specific to European driving contexts.

7 Future Work

Our findings suggest several promising directions, broadly grouped into data-centric, model-centric, and regulatory-focused research.

7.1 Data-Centric Analysis

A natural next step is to move beyond geographic distance as a proxy and directly quantify the environmental factors that contribute to geographic domain shift. Variables such as road-marking conventions, urban density, architectural style, vegetation, weather patterns, and traffic behaviour could be measured explicitly and incorporated into predictive models of cross-region generalisation. Such analyses would help determine which factors are most responsible for performance degradation and whether they provide stronger explanatory power than geographic distance alone.

Future work could expand beyond European datasets to other continents and diverse domains, such as satellite imagery, agriculture, and urban scenes, to test whether observed distance-performance relationships generalise globally or are domain-specific. Developing a benchmark suite of geographically annotated datasets with standardised evaluation protocols would enable fair comparison. Additionally, decomposing the effect of distance into factors like infrastructure quality, traffic density, climate, and urban design, while examining how seasonal and temporal changes interact with spatial variation, would provide a

more complete understanding of geographic domain adaptation and guide targeted data collection strategies.

7.2 Model-Centric Development

The observed performance degradation suggests a need for adaptation methods specifically designed for geographically distributed deployment. Continual learning, domain adaptation, federated learning, and foundation-model fine-tuning need future research to investigate which adaptation strategies provide the best balance between performance, computational cost, data availability, and regulatory constraints in geographically fragmented settings between local and global models.

Developing geography-aware algorithms through region-prioritised replay, domain-aware regularisation, modular architectures, or meta-learning could improve adaptation across regions. Investigating why certain architectures are more robust to geographic shifts, particularly in terms of attention, multi-scale processing, and inductive biases, would inform more resilient model design. Representation-level metrics, such as encoder embeddings or distributional dissimilarity metrics (MMD, SWD), could serve as predictive tools for model performance, guiding data collection and retraining. Hybrid training strategies that combine centralised training on diverse datasets with federated fine-tuning on region-specific data may further optimise performance while respecting data sovereignty.

7.3 Regulatory and Policy Expansion

A broader mapping of global data governance and AI regulations, combined with frameworks that integrate legal constraints into training pipelines, audits, and documentation, would bridge the gap between policy and technical practice. Quantifying the performance impact of data localisation versus centralised training would inform regulatory decisions and balance privacy with effectiveness. Exploring privacy-preserving approaches, such as encrypted gradients or secure multiparty computation, could enable multi-jurisdictional model development without transferring raw data, facilitating practical deployment in diverse regulatory contexts.

8 Conclusion

This study investigates whether geographic separation is associated with domain shift in automotive perception systems. Across two automotive datasets (ECP and ZOD) and one globally distributed image dataset (OSV-5M), we find consistent evidence that increasing geographic distance is accompanied by greater representation divergence and reduced downstream perception performance. These observations were replicated across multiple representation models, including CLIP and DINOv2, multiple distributional similarity metrics, and a diverse set of object detection and semantic segmentation architectures.

At the representation level, geographic separation was consistently associated with increased distributional divergence across both image-language and self-supervised encoders. The strong agreement observed between CLIP and DINOv2 indicates that the reported trends are not specific to a single representation model and instead reflect broader structure present within the underlying visual data. The inclusion of OSV-5M further demonstrates

that distance-dependent representation drift is observable beyond the European automotive datasets used for downstream perception evaluation.

At the task level, object detection and lane-segmentation performance generally decreased as geographic distance increased. On ZOD, mixed-effects modelling and permutation-based validation indicated that this relationship remained significant after accounting for the repeated-measures structure of the experiments, and analyses excluding in-distribution evaluations showed that the effect cannot be explained solely by a binary in-domain versus out-of-domain distinction. On ECP, which contains only four regions, the association was not robust to excluding in-distribution evaluations, and we regard those results as inconclusive regarding continuity. Together, these findings support the view that geographic domain shift is better understood as a continuous phenomenon rather than a simple regional boundary.

Importantly, our results should not be interpreted as establishing geographic distance as a causal driver of performance degradation. Geographic separation is more appropriately viewed as a measurable proxy for a collection of latent environmental, infrastructural, and cultural factors that differ across regions. Nevertheless, the consistency of the observed relationships across datasets, encoders, metrics, and perception architectures suggests that geographic variation represents an important and often overlooked source of distribution shift in automotive AI systems.

From a practical perspective, these findings highlight the importance of geographic coverage when building, evaluating, and deploying perception systems. Validation protocols based solely on random train-test splits may fail to capture deployment risks associated with operating in previously unseen regions. Future work should focus on identifying the environmental factors underlying geographic domain shift, developing adaptation methods that improve cross-region robustness, and constructing geographically diverse benchmarks that enable more rigorous evaluation of real-world generalisation.

9 Acknowledgements

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094_P2 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Science Foundation Ireland Research Centre for Software (www.lero.ie)

References

M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20178–20188, 2023.

G. Astruc, N. Dufour, I. Siglidis, C. Aronssohn, N. Bouia, S. Fu, R. Loiseau, V. N. Nguyen, C. Raude, E. Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977, 2024.

S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

P. S. Bradley, K. P. Bennett, and A. Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019.

T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Council of Europe. Council of europe framework convention on artificial intelligence and human rights, democracy and the rule of law, 2024. URL `https://www.coe.int/en/web/conventions/full-list2?module=treaty-detail&treatynum=225`.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2016. Official Journal of the European Union.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`, 2024. Official Journal of the European Union.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. W. Koh, O. Saukh, A. J. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt. Datacomp: In search of the next generation of multimodal datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 27092–27112. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/56332d41d55ad7ad8024aac625881be7-Paper-Datasets_and_Benchmarks.pdf`.

A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Government of Canada. Personal information protection and electronic documents act, 2000. URL `https://laws-lois.justice.gc.ca/eng/acts/P-8.6/`. Justice Laws Website; current consolidated version.

Government of India. Digital personal data protection act, 2023. `https://www.meity.gov.in/`, 2023. Act No. 22 of 2023.

Government of Japan. Act on the protection of personal information. `https://www.ppc.go.jp/en/legal/`, 2003.

Government of Singapore. Personal data protection act 2012. `https://sso.agc.gov.sg/Act/PDPA2012`, 2012.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

S. Huch and M. Lienkamp. Towards minimizing the lidar sim-to-real domain shift: Object-level local domain adaptation for 3d point clouds of autonomous vehicles. *Sensors*, 23 (24):9913, 2023.

Infocomm Media Development Authority. Model artificial intelligence governance framework, 2019. Non-binding framework for responsible AI deployment.

Israel Innovation Authority. National artificial intelligence program, 2022. Policy framework addressing AI development and governance.

N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

G. Jocher, A. Chaurasia, and J. Qiu. Yolov8 by ultralytics, 2023. URL `https://github.com/ultralytics/ultralytics`.

B. A. Juliussen, J. P. Rui, and D. Johansen. Algorithms that forget: Machine unlearning and the right to erasure. *Computer Law & Security Review*, 51:105885, 2023.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

T. Kalb and J. Beyerer. Principles of forgetting in domain-incremental semantic segmentation in adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19508–19518, 2023.

R. Khanam and M. Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Kingdom of Saudi Arabia. Personal data protection law. `https://sdaia.gov.sa/`, 2021.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/koh21a.html`.

S. Kumar, S. Sharma, R. Asghar, R. Mohandas, T. Brophy, G. Sistu, E. M. Grua, V. Donzella, and C. Eising. Exploring sensor impact and architectural robustness in adverse weather on bev perception. *IEEE Open Journal of Vehicular Technology*, 6: 2857–2875, 2025. doi: 10.1109/OJVT.2025.3621862.

D. R. Legates and G. J. McCabe Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241, 1999.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

P. Mayilvahanan, T. Wiedemer, E. Rusak, M. Bethge, and W. Brendel. Does clip's generalization performance mainly stem from high train-test similarity? *arXiv preprint arXiv:2310.09562*, 2023.

S. Muralidhara, R. Schuster, and D. Stricker. Domain-incremental semantic segmentation for autonomous driving under adverse driving conditions. *arXiv preprint arXiv:2501.05246*, 2025.

National Assembly of the Republic of Korea. Basic act on the development of artificial intelligence and the establishment of a foundation for trustworthiness. `https://regulations.ai/regulations/RAI-KR-NA-ACTNO20-2025`, 2025. Act No. 20676, promulgated Jan. 21, 2025, effective Jan. 22, 2026.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Parliament of Canada. Artificial intelligence and data act (aida), bill c-27. `https://www.parl.ca/legisinfo/en/bill/44-1/c-27`, 2022.

K. Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.

R. P. Poudel, S. Liwicki, and R. Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Republic of South Africa. Protection of personal information act. `https://www.justice.gov.za/legislation/acts/2013-004.pdf`, 2013.

N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Standing Committee of the National People's Congress. Data security law of the people's republic of china. `https://www.npc.gov.cn/englishnpc/c2759/c23934/202112/t20211209_385109.html`, 2021a. Adopted June 10, 2021; effective Sept. 1, 2021; Presidential Order No. 84.

Standing Committee of the National People's Congress. Personal information protection law of the people's republic of china. `https://en.spp.gov.cn/2021-12/29/c_948419.htm`, 2021b. Adopted Aug. 20, 2021; effective Nov. 1, 2021.

State of Israel. Protection of privacy law, 1981. Regulates personal data processing and cross-border data transfers.

H. Su, H. Gao, X. Wang, X. Fang, Q. Liu, G.-B. Huang, X. Li, and Q. Cao. Object detection in adverse weather for autonomous vehicles based on sensor fusion and incremental learning. *IEEE Transactions on Instrumentation and Measurement*, 73:1–10, 2024.

P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21371–21382, 2022.

Swiss Confederation. Federal act on data protection (fadp). `https://www.fedlex.admin.ch/`, 2020.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

R. Tachet des Combes, H. Zhao, Y.-X. Wang, and G. J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19276–19289. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/dfbfa7ddcfffeb581f50edcf9a0204bb-Paper.pdf`.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In H. Larochelle,

M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

United Arab Emirates. Federal decree-law no. 45 of 2021 on the protection of personal data. https://u.ae/en/about-the-uae/digital-uae/data/data-protection-laws, 2021.

United States Congress. Clarifying lawful overseas use of data act (cloud act). https://www.congress.gov/bill/115th-congress/house-bill/4943, 2018. Public Law 115–141.

A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

Y. Wang, B. Li, Z. Wei, A. Rahman, D. Gummadi, H. Zhao, and V. Donzella. A survey and new perspective of sensing in the dark for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

L. Yang, Z. Luo, S. Zhang, F. Teng, and T. Li. Continual learning for smart city: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

J. Zhang, J. Huang, Z. Tian, and S. Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022.

Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16965–16974, June 2024.

A Encoder Pairwise Agreement

Figure 10 compares the pairwise MMD values produced by each encoder on the 200-cluster OSV-5M partition. Each point is one cluster pair, and the Pearson correlations shown are those reported in Table 4. Agreement is near-perfect within encoder families (CLIP: 0.967; DINOv2: 0.996) and remains high across families (0.884–0.896), indicating that which cluster pairs are far apart in embedding space is largely independent of the encoder.

B Mixed-effect models

Tables 10, 11 and 12 report the mixed-effects analysis of Section 3.2.3 for ECP object detection, ZOD object detection, and ZOD lane segmentation, respectively. N is the number of source–target evaluations per model. *Pearson* r is the pooled correlation between performance (mAP@50 or IoU) and distance across those evaluations; it is descriptive only and differs from the per-cluster Fisher-combined correlations in Tables 8 and 9. β is the fixed-effect coefficient on $\log(1 + d)$, with d in km (natural log), with 95% confidence intervals. "Excl. zero distance" removes self-evaluations (source = target). Because the log transform places zero-distance points far from all others, β under the two settings is not directly comparable in magnitude.

C OSV-5M comparison plots

Figure 11 plots embedding divergence against geographic distance (km, log scale) for OSV-5M under both clustering granularities (100 and 200 clusters), for each metric and encoder. Each point is a pair of clusters. These are the data underlying the correlations in Table 3, which are computed against log-scaled distance.

Table 10: Relationship between distance and detection performance on ECP.

Model	Data	N	Pearson r	Mixed-effect slope β [95% CI]
Faster R-CNN (R50)	All data	16	-0.521	-0.0124 [-0.0183, -0.0065]
	Excl. zero distance	12	-0.020	-0.0452 [-0.0687, -0.0218]
Faster R-CNN (R18)	All data	16	-0.512	-0.0141 [-0.0203, -0.0078]
	Excl. zero distance	12	-0.054	-0.0543 [-0.0812, -0.0274]
RT-DETRv2	All data	16	-0.585	-0.0125 [-0.0172, -0.0079]
	Excl. zero distance	12	-0.074	-0.0417 [-0.0619, -0.0215]
YOLOv10n	All data	16	-0.535	-0.0112 [-0.0166, -0.0057]
	Excl. zero distance	12	-0.052	-0.0464 [-0.0616, -0.0313]
YOLOv10m	All data	16	-0.525	-0.0103 [-0.0158, -0.0047]
	Excl. zero distance	12	0.002	-0.0388 [-0.0480, -0.0295]
YOLOv10x	All data	16	-0.533	-0.0103 [-0.0156, -0.0050]
	Excl. zero distance	12	-0.003	-0.0367 [-0.0521, -0.0213]

Each detector is evaluated on four target regions after training on each of four source regions ($N = 16$; $N = 12$ excluding zero distance). Permutation tests conducted after excluding zero-distance evaluations did not reach statistical significance, and Pearson correlations became negligible ($r = -0.074$ to 0.002). We therefore regard the ECP evidence for a distance effect beyond the in-domain contrast as inconclusive.

Table 11: Relationship between distance and detection performance.

Model	Data	N	Pearson r	Mixed-effect slope β [95% CI]
Faster R-CNN (R50)	All data	100	-0.602	-0.0075 [-0.0085, -0.0065]
	Excl. zero distance	90	-0.490	-0.0198 [-0.0232, -0.0164]
Faster R-CNN (R18)	All data	100	-0.590	-0.0072 [-0.0082, -0.0062]
	Excl. zero distance	90	-0.482	-0.0175 [-0.0212, -0.0138]
RT-DETRv2	All data	100	-0.586	-0.0069 [-0.0080, -0.0059]
	Excl. zero distance	90	-0.484	-0.0193 [-0.0229, -0.0157]
YOLOv10n	All data	100	-0.621	-0.0094 [-0.0107, -0.0080]
	Excl. zero distance	90	-0.528	-0.0244 [-0.0296, -0.0193]
YOLOv10m	All data	100	-0.607	-0.0095 [-0.0110, -0.0080]
	Excl. zero distance	90	-0.546	-0.0257 [-0.0315, -0.0199]
YOLOv10x	All data	100	-0.593	-0.0094 [-0.0110, -0.0078]
	Excl. zero distance	90	-0.533	-0.0251 [-0.0314, -0.0189]

All mixed-effects coefficients remained significant under both model-based inference and permutation testing ($p_{\text{perm}} \leq 0.003$).

Table 12: Relationship between distance and segmentation performance on ZOD.

Model	Data	N	Pearson r	Mixed-effect slope β [95% CI]
DeepLabv3+	All data	100	-0.723	-0.0161 [-0.0188, -0.0133]
	Excl. zero distance	90	-0.658	-0.0585 [-0.0663, -0.0507]
Fast-SCNN	All data	100	-0.619	-0.0140 [-0.0162, -0.0118]
	Excl. zero distance	90	-0.541	-0.0432 [-0.0502, -0.0363]
SegFormer-B0	All data	100	-0.649	-0.0169 [-0.0196, -0.0142]
	Excl. zero distance	90	-0.581	-0.0563 [-0.0644, -0.0481]
SegFormer-B2	All data	100	-0.652	-0.0168 [-0.0195, -0.0141]
	Excl. zero distance	90	-0.573	-0.0550 [-0.0628, -0.0472]
Swin-UPerNet	All data	100	-0.684	-0.0150 [-0.0176, -0.0123]
	Excl. zero distance	90	-0.608	-0.0541 [-0.0620, -0.0463]
U-Net	All data	100	-0.455	-0.0107 [-0.0126, -0.0087]
	Excl. zero distance	90	-0.373	-0.0390 [-0.0447, -0.0334]
YOLOv11l-Seg	All data	100	-0.509	-0.0123 [-0.0143, -0.0102]
	Excl. zero distance	90	-0.444	-0.0437 [-0.0499, -0.0375]
YOLOv11m-Seg	All data	100	-0.529	-0.0124 [-0.0144, -0.0103]
	Excl. zero distance	90	-0.469	-0.0416 [-0.0481, -0.0350]
YOLOv11n-Seg	All data	100	-0.507	-0.0108 [-0.0127, -0.0090]
	Excl. zero distance	90	-0.459	-0.0385 [-0.0446, -0.0325]
YOLOv8l-Seg	All data	100	-0.558	-0.0122 [-0.0142, -0.0101]
	Excl. zero distance	90	-0.503	-0.0435 [-0.0497, -0.0373]
YOLOv8m-Seg	All data	100	-0.544	-0.0121 [-0.0141, -0.0101]
	Excl. zero distance	90	-0.484	-0.0418 [-0.0480, -0.0357]
YOLOv8n-Seg	All data	100	-0.516	-0.0120 [-0.0139, -0.0100]
	Excl. zero distance	90	-0.455	-0.0417 [-0.0479, -0.0354]

All mixed-effects coefficients remained significant under both model-based inference and permutation testing ($p_{\text{perm}} \leq 0.003$).

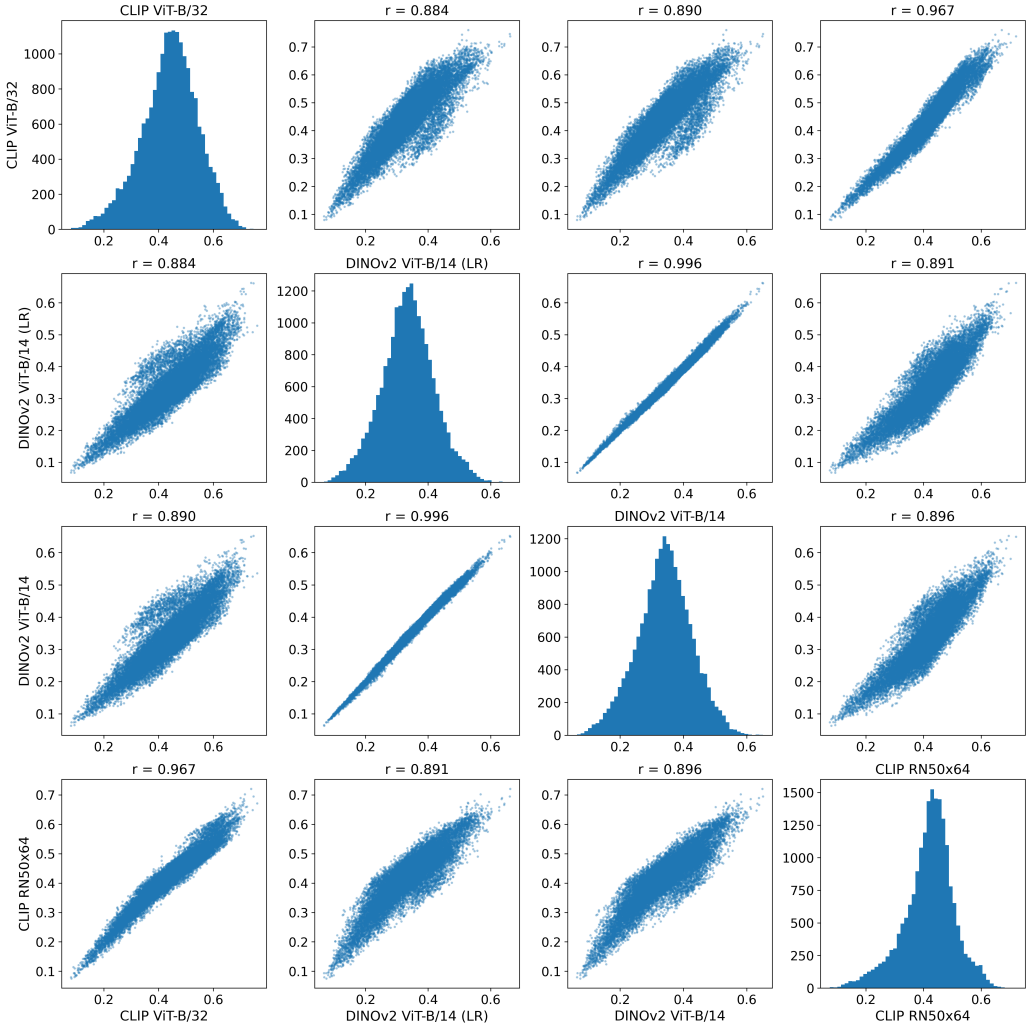


Figure 10: Agreement between encoders on the 200-cluster OSV-5M partition. Off-diagonal panels plot the pairwise MMD for the same cluster pairs under two different encoders (Pearson r in the panel title). Diagonal panels show the histogram of MMD values for a single encoder.

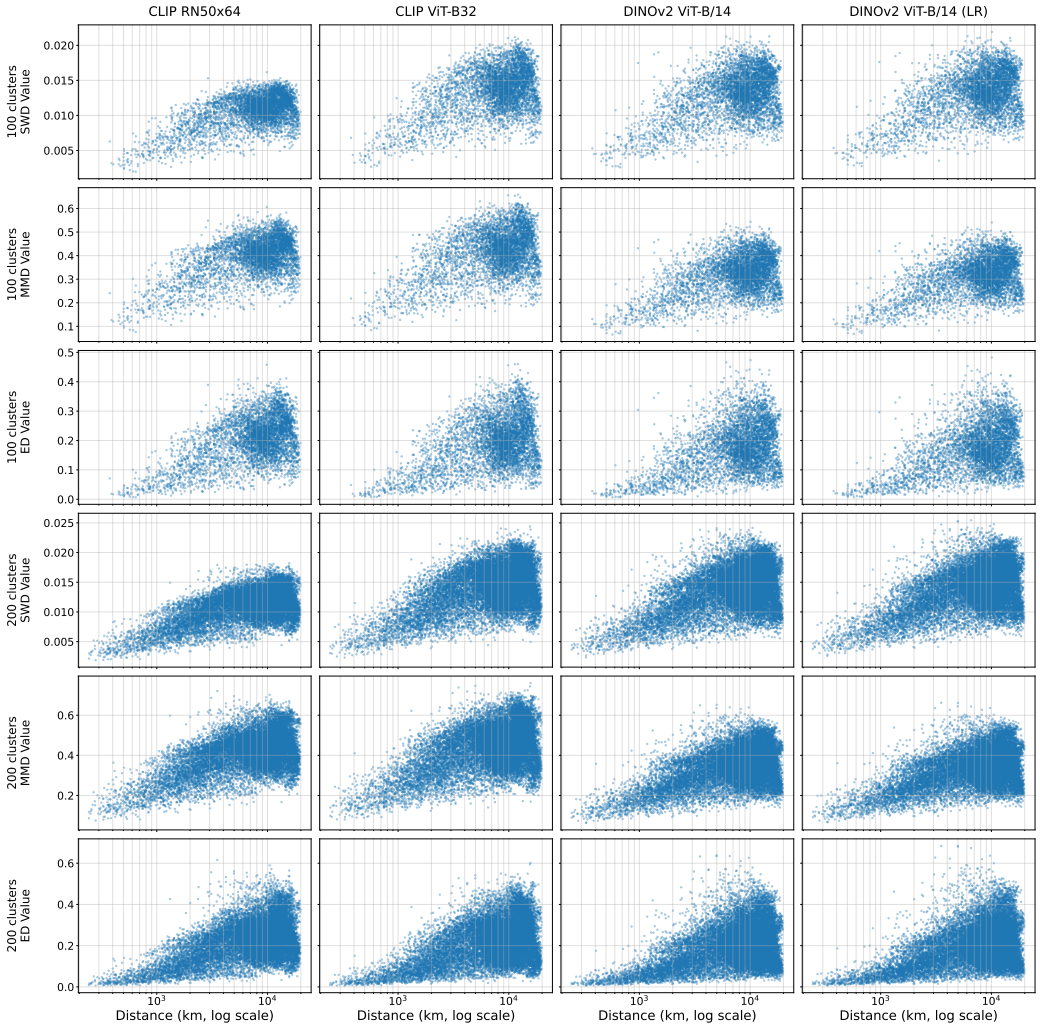


Figure 11: Geographic distance versus embedding divergence on OSV-5M. Rows show SWD, MMD and ED for the 100-cluster partition (top three) and the 200-cluster partition (bottom three); columns show the four encoders. Each point is a cluster pair, and the x-axis is log-scaled.