Telescope:
Learnable Hyperbolic Foveation for Ultra-Long-Range Object Detection

Parker Ewen[1], Dmitriy Rivkin[1], Mario Bijelic[1,2], Felix Heide[1,2]

[1]Torc Robotics, [2]Princeton University

Abstract

Autonomous highway driving, especially for long-haul heavy trucks, requires detecting objects at long ranges beyond 500 meters to satisfy braking distance requirements at high speeds. At long distances, vehicles and other critical objects occupy only a few pixels in high-resolution images, causing state-of-the-art object detectors to fail. This challenge is compounded by the limited effective range of commercially available LiDAR sensors, which fall short of ultra-long range thresholds because of quadratic loss of resolution with distance, making image-based detection the most practically scalable solution given commercially available sensor constraints. We introduce Telescope, a two-stage detection model designed for ultra-long range autonomous driving. Alongside a powerful detection backbone, this model contains a novel re-sampling layer and image transformation to address the fundamental challenges of detecting small, distant objects. Telescope achieves 76% *relative improvement in mAP in ultra-long range detection compared to state-of-the-art methods (improving from an absolute mAP of 0.185 to 0.326 at distances beyond 250 meters), requires minimal computational overhead, and maintains strong performance across all detection ranges. Our project page is available at* `https://light.princeton.edu/telescope`.

1. Introduction

Autonomous driving requires perceptual understanding of the surrounding scene [48, 55]. Existing datasets and benchmarks focus heavily on city driving [3, 6, 10, 15, 20, 41, 47], where low vehicle speeds require short-range perception.

Highway driving, and in particular heavy-duty trucking, presents a fundamentally different challenge. At highway speeds, a fully-loaded truck requires on the order of 150–$200\,\mathrm{m}$ to come to a complete stop [16]. Existing benchmarks often provide limited sensing horizons around 80–$100\,\mathrm{m}$, which is insufficient for safe braking and strategic maneuvers such as merging or lane changes. As a result, safe highway autonomy requires reliable perception at hundreds of meters, and up to the kilometer scale for full visual scene understanding [16]. We study ultra-long range object detection in TruckDrive [16] which provides annotations up to $1\,\mathrm{km}$. From this dataset, we identify three major challenges for ultra-long range detection. First, active LiDAR and radar measurements become increasingly sparse and low in signal at long distances —fundamental to the sequential scanning and quadratic intensity falloff of diffuse reflections—making cameras the only sensing modality that can provide dense spatial coverage at hundreds of meters. Second, ultra-long range objects project to extremely small bounding boxes. In these datasets, objects beyond $250\mathrm{m}$ frequently occupy only tens of pixels. Consequently, the number of resolvable visual features on distant objects is directly limited by the input image resolution. Third, there is an extreme scale imbalance between objects, where nearby objects and background regions dominate the pixel budget and ultra-long range objects contribute only a negligible fraction of tokens or patches (Figure 1).

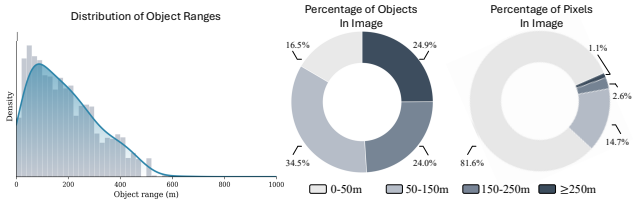

Figure 1. **Long-range Objects in Driving Datasets.** Analysis of the TruckDrive [16] dataset shows the distribution of object distances and the breakdown of the pixel-wise composition of objects at each distance. While all object ranges are equally represented in images, the proportion of pixel area disproportionately favors nearby objects, with long ($150 - 250$m) and ultra-long (≥ 250m) objects occupying only a small fraction of image pixels.

Figure 1 illustrates this imbalance, where images in TruckDrive contain both nearby and distant objects. Notably, distant objects represent a fraction of the image features compared to nearby objects. This distribution imbal-

ance highlights that ultra-long range detection is not only a long-tail data problem, but also a fundamental representational problem caused by extreme scale disparity within individual images.

Detection methods must therefore efficiently process high-resolution images with minimal latency and memory usage in order to capture objects at long range. We argue the detection mechanism should therefore avoid the quadratic complexity of standard self-attention, which becomes prohibitive at high resolutions and dilutes attention [5, 60]. Finally, the model requires an explicit resampling mechanism that magnifies distant objects while shrinking nearby ones, normalizing object scale to facilitate learning [22].

To address these requirements, we introduce Telescope, a two-stage ultra-long range detection method with an application to autonomous highway driving (Figure 2). In the first stage, we learn an image transformation called the hyperbolic foveation, which magnifies salient image regions. This transformation interpolates between the Poincaré disk projection [40] and the identity function, inducing a Riemannian space [24] where bounding boxes can be parameterized and learned, then re-projected to image space with machine precision and no warping artifacts. Notably, this transform is inspired by biological vision [21]. By learning the foveation parameters on down-sampled images, stage one incurs minimal computational overhead.

The second stage applies the learned transformation to full-resolution images, which are then processed by a pre-trained foundation model encoder [2, 4, 33, 34, 39] and lightweight Deformable DETR detection head [60]. The pre-trained encoder enables fast convergence and robustness across scales, while sparse sampling avoids the quadratic cost and attention dilution of standard transformers, making it well-suited for high-resolution inputs. While this two-stage approach is applied to the context of ultra-long range object detection for highway driving, the proposed foveated transform is general and can be applied to existing image-based approaches, including Vision Language Models (VLMs).

Evaluated on long-range autonomous driving benchmarks, Telescope consistently improves detection performance across all distance ranges and achieves up to 76% improvement in mAP at ultra-long ranges over existing state-of-the-art approaches (improving from an absolute mAP of 0.185 to 0.326 at distances greater than $250\mathrm{m}$), with detections extending to $1\,\mathrm{km}$. In summary, the contributions of this paper are:

- An analysis of ultra-long range object detection and the identification of several critical model requirements.
- A systematic ablation of foundation model image encoders, detection heads, and training schemes for ultra-long range object detection, providing practical insights into which backbone representations and optimization strategies are most effective under constrained fine-tuning budgets.
- A novel learnable and invertible hyperbolic foveation image transform and associated Riemannian bounding box reparameterization for ultra-long range domain scaling.
- A state-of-the-art ultra-long range object detection model, Telescope, for highway driving, which improves object detection performance by 76% at ultra-long ranges compared to existing methods (increasing absolute mAP from 0.185 to 0.326 at distances beyond 250 meters).

2. Related Work

Object detection has been a cornerstone application of modern deep learning-based computer vision over the past decade [58, 61]. The autonomous driving domain, in particular, has spurred innovations in detection methods tailored to the unique challenges of on-road perception [29, 36]. We present a review of relevant work for object detection for autonomous driving, small object detection, and learned spatial transformations.

Object Detection for Autonomous Driving. Autonomous driving has helped push the development of specialized object detection benchmarks and methods. Prominent datasets include KITTI [15], nuScenes [3], Waymo Open Dataset [41], CityScapes [10], and Argoverse [6, 47], which provide multi-modal sensor data primarily catered towards low-speed and city driving scenarios. Early methods adapted general-purpose detectors like Faster R-CNN [7, 36] and YOLO [35] to automotive contexts. More recent approaches leverage transformer-based architectures, including DETR [5] and its variants [11, 12, 25], which formulate detection as a set prediction problem. Deformable DETR [60] introduced sparse spatial sampling [50] to improve efficiency and convergence for high-resolution inputs. Methods like CenterNet [59] and FCOS [43] explore anchor-free detection paradigms better suited to the wide range of object scales in driving scenes. Despite these advances, most benchmarks emphasize urban driving scenarios at limited ranges [10, 41], leaving ultra-long range highway detection under-explored.

Small Object Detection. Small object detection presents fundamental challenges stemming from limited pixel support [31], poor signal-to-noise ratios [23], and severe class imbalance in high-resolution images [30, 46]. Standard detection architectures struggle when objects occupy few pixels due to metrics such as IoU scaling poorly to small bounding boxes [9, 18]. Methods addressing small objects typically employ multi-scale feature pyramids [17, 26, 53], specialty losses [45, 52], super-resolution preprocessing [1, 27, 32], or attention mechanisms to enhance fine-grained

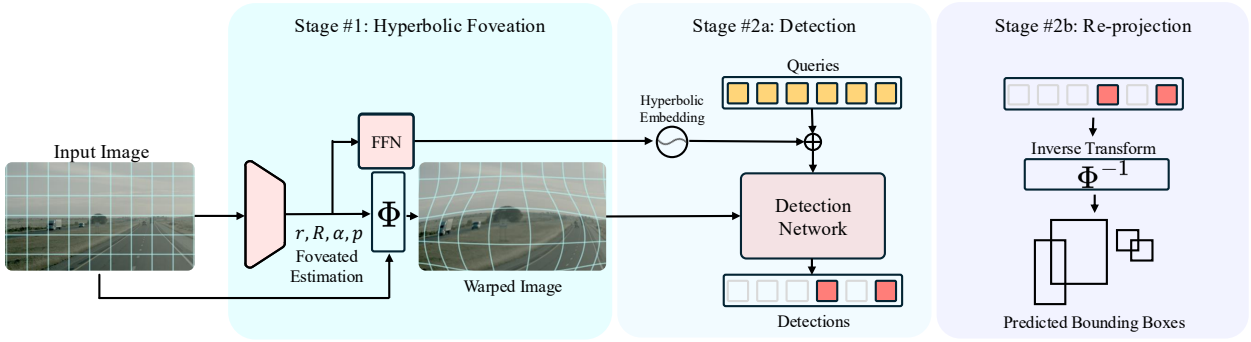

Figure 2. **Telescope.** We propose a two-stage ultra-long range detection model. Stage one uses a down-sampled image to estimate the hyperbolic foveation image transformation parameters. This transformation enlarges distant objects at the center of the transform while shrinking nearby objects at the periphery. Stage two uses this transformed image alongside learned hyperbolic embeddings to detect objects at distances of up to 1km.

representations [7, 8, 18, 44].

Datasets like TinyPerson [54], AI-TOD [51], and DOTA [49] focus specifically on small object scenarios, primarily in aerial imagery and crowd surveillance contexts. However, these datasets either contain only small objects [49, 51] or tend to have limited variance in object sizes within individual images [54]. In contrast, object sizes vary significantly in autonomous driving, where nearby vehicles may occupy orders-of-magnitude more pixels than distant vehicles [16].

Learned Spatial Transformations. While some approaches focus on network architectures and loss functions tailored towards the small object domain [13, 17, 18, 45, 56], other methods re-sample the image directly to magnify or warp high salience regions. Spatial Transformer Networks [22] introduced learnable geometric transformations to warp input images for improved spatial invariance. More recently, FOVEA [42] extends this concept to autonomous driving with a learned foveation that magnifies distant regions for long-range detection. However, FOVEA requires full-resolution images to estimate transformation parameters and maintains axis-aligned bounding box representations. Inspired by biological visual systems [21], our proposed hyperbolic foveation differs by estimating parameters from low-resolution images by leveraging strong object detection priors from the network encoder. Furthermore, the proposed method directly estimates the warped bounding boxes in the local Riemannian coordinate frame without requiring a rectilinear coordinate frame. This reparameterization enables more natural object representations in the transformed domain without the geometric constraints of axis-aligned boxes under non-linear transformations.

3. Ultra-Long-Range Detection

In this section, we describe the proposed ultra-long range detection method, Telescope, as shown in Figure 2. We first introduce the hyperbolic foveated transform in Sec. 3.1 as a means of normalizing object sizes across scales. This transform magnifies distant objects while compressing nearby ones and ensures minimal computational overhead. In Sec. 3.2, we describe the parameterization of the bounding boxes in this transformed space. Finally, we describe a network architecture in Sec. 3.3 to efficiently operate on these high-resolution transformed images, enabling object detection at distances of up to 1km.

3.1. Hyperbolic Foveated Transform

Let $\mathcal{M} = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ denote the Poincaré disk equipped with the standard metric tensor ds^2. Directly projecting an image with finite domain onto $\mathcal{M}$ is undesirable since the metric diverges as $\|x\| \to 1$, causing unbounded distortion and numerical instability near the image boundary.

To overcome this issue, we define a pseudo-Riemannian projection which radially interpolates between a Poincaré-like contraction and the identity transform. Let normalized image coordinates be $x \in [-1, 1]^2$ and let $o \in [-1, 1]^2$ be the Poincaré origin projected into Euclidean coordinates. Define $r = \|x - o\|$ as the offset between the center of the image and the origin of the Poincaré disk. The Poincaré projection is given as

$$h(x; o) = o + \frac{\tanh(\alpha r)}{r}(x - o), \tag{1}$$

where $\alpha > 0$ is the hyperbolic contraction strength.

The hyperbolic foveated transform is then defined as

$$\Phi(x) = (1 - w(r))\, x + w(r)\, h(x), \tag{2}$$

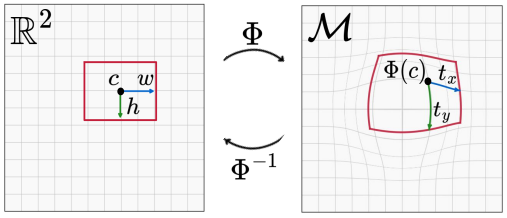

Figure 3. **Hyperbolic Foveated Transform.** The transformation coefficients enable the re-parameterization of the bounding box in the induced Riemannian space where the box center and tangent vector magnitudes fully describe the box location and shape.

where $w(r) = (1-\min(r/R, 1))^p$ is the radial interpolation coefficient, $p > 0$ is the fixed blending exponent, and $R > 0$ is the radial scale of the Poincaré disk. For $r \ll R$, Φ behaves as a hyperbolic contraction around o, while for $r \geq R$ it smoothly approaches the identity mapping. This enables a numerically stable projection of images onto the induced Riemannian manifold without unbounded distortion, meaning objects near the boundaries of the image are still visible.

While the exact inverse of this transform, $\Phi^{-1}(x)$, cannot be computed explicitly, its existence is provable[1]. Furthermore, we can approximate the inverse differentiably, up to numerical precision, and with convergence guarantees[1] via the Newton-Raphson algorithm. Given $y = \Phi(x)$, the inverse $\Phi^{-1}(y)$ is obtained numerically via

$$x^{(k+1)} = x^{(k)} + \eta\big(y - \Phi(x^{(k)})\big), \tag{3}$$

initialized with $x^{(0)} = y$ and step size $\eta \in (0, 1]$.

3.2. Hyperbolic Box Parameterization

When projecting images onto the induced Riemannian manifold, bounding boxes become warped, and axis-aligned, recti-linear boxes can no longer be used. Akin to the 4-parameter bounding box parameterization in axis-aligned image-space coordinates, We propose estimating the local coordinates of these warped boxes directly and provide a 4-parameter parameterization for boxes in the Riemannian manifold induced by Eq. (2).

Let a box be defined in Euclidean image coordinates by center $c \in \mathbb{R}^2$, width $w \in \mathbb{R}$, and height $h \in \mathbb{R}$ such that $b = [c_x, c_y, w, h]$. The re-parameterized box in the induced Riemannian space is then $b' = [\Phi_x(c), \Phi_y(c), \|t_x\|, \|t_y\|]$, where the first two components are the projected box center and the last two components are the tangent vector magnitudes of the local coordinates at $\Phi(c)$. Notably, only the tangent vector magnitudes are needed as the tangent vectors are fully defined given the transform parameters. Figures 3 and 5 provides a visualization of this re-parameterization.

[1]See Appendix for Theorem and Proof.

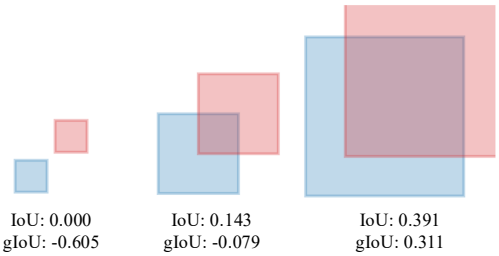

Figure 4. **gIoU for Multi-Scale Training.** The gIoU metric is a better training loss for multi-scale objects as it provides a gradient even when bounding boxes are non-overlapping.

Figure 5. **Learned hyperbolic foveated transform on the TruckDrive dataset.** The original image (left) and the foveated image (right) are shown, together with the percentage increase in object bounding-box area. Both views are cropped to the same image region, highlighting the local magnification induced by the foveation transform. The proposed transform is effective for both isolated targets and dense, busy scenes.

The tangent vectors of the bounding box are computed as

$$t_x = J_\Phi(c)\,[w, 0]^\top \tag{4}$$

$$t_y = J_\Phi(c)\,[0, h]^\top \tag{5}$$

where $J_\Phi(c)$ is the Jacobian of Φ at c.

All the terms in the re-parameterization depend only on the original box parameterization and the transform parameters (α, R, p, o), and can therefore be evaluated analytically and vectorized efficiently during training. This parameterization is fully determined by (R, α, p, o) through J_Φ and uniquely specifies the local box geometry in the induced Riemannian space. Furthermore, the original bounding box parameterization can be recovered using (3) and by inverting (4) and (5).

3.3. Detection Network Architecture

High-resolution images are needed to keep ultra long range objects resolvable. As such, detection networks must scale favorably to image dimensions. To this end, we propose a model architecture leveraging pre-trained foundation models with existing efficiencies baked in for optimal scaling

performance.

Foundation model encoders have been shown to be strong object detection priors [2, 33, 39]. A variety of foundation model encoders were tested with both DETR and Deformable DETR detection heads, and we find the optimal model combination to be the SAM3 image encoder and a Deformable DETR detection head.

Notably, the SAM3 image encoder uses windowed attention and sparse global attention to minimize the computational burden of processing high-resolution images [4]. The Deformable DETR head shows fast convergence across all encoders and better evaluation metrics than DETR heads. This is because DETR requires full self-attention at each decoder layer [5], which distributes attention across the exponentially growing image patch features while Deformable DETR leverages sparse sampling [60].

4. Implementation

Foveated Transform Parameters. The goal of the hyperbolic foveation is to magnify distant objects in the scene. To this end, we must first estimate where these objects are. We train a small FFN using the output of the image encoder to estimate the center, o, and radius, R, of the transform. A low-resolution image is used (i.e., 256×256 or 512×512) such that the parameter estimation incurs minimal computational overhead.

To set the other foveation parameter, we empirically determine which values maximize the transformed bounding boxes via grid search. For TruckDrive, the optimal parameters are set as $\alpha = 2.0$ and $p = 2.0$.

To compute the hyperbolic embeddings, we additionally train another FFN to project the set of foveation parameters into the same dimension as the object queries. This provides the detection network with information about the image-specific hyperbolic transformation such that the hyperbolic box parameterization, b', can be estimated. Examples of this learned transform are shown in Figure 5.

Training Losses. The hyperbolic box parameterization represents bounding boxes in the induced Riemannian space. Unfortunately, defining distance and area in this space is non-trivial, making computing the losses between the target and predicted boxes challenging in Riemmanian space.

To overcome this challenge, the predicted boxes are projected into Euclidean space where vanilla L1 and gIoU losses are computed. Notably, the iterative inverse (3), (4), and (5) are all differentiable and are thus amenable to backpropagation. The gIoU metric [37] is better suited for detection of ultra-long range objects as it provides a gradient even when matched boxes do not overlap [9]. See Figure 4 for a visualization of this phenomenon.

Table 1. **Ablation Experiments on the Components of Telescope**. A de-noising training scheme helps improve model performance and the hyperbolic foveation improves mAP at far, long, and ultra-long ranges while slightly reducing performance for nearby objects. All rows use SAM3 encoder.

Method	COCO				
	mAP	$mAP_{0\text{-}50}$	$mAP_{50\text{-}150}$	$mAP_{150\text{-}250}$	mAP_{250+}
Deformable DETR	0.32	0.52	0.34	0.24	0.17
+ De-noising	0.50 (**+0.18**)	0.69 (**+0.17**)	0.48 (**+0.14**)	0.32 (**+0.08**)	0.29 (**+0.12**)
+ Hyperbolic Foveation	0.50 (**+0.00**)	0.61 (**-0.08**)	0.50 (**+0.02**)	0.34 (**+0.02**)	0.33 (**+0.03**)

Following [56], we apply the de-noising training scheme, where ground truth boxes are noised and concatenated to the prediction queries to more effectively learn box alignment. In this case the projected Riemannian boxes are used as the ground truth anchors to align them with the predictions in Riemmanian space.

5. Experimental Validation

We evaluate the proposed ultra-long range detection framework on the TruckDrive dataset [16], and conduct controlled ablations to isolate the impact of backbone encoders, detection heads, and the proposed hyperbolic foveated transform. Our evaluation focuses in particular on performance at long and ultra-long ranges, where existing detection pipelines degrade most severely.

Dataset and Metrics. As we are concerned with long and ultra-long range object detection, we use the TruckDrive dataset which is currently the only dataset with annotations at ultra-long range distances. Since these ranges exceed the reliable operating regime of LiDAR and radar, object distances are estimated from bounding box height, camera intrinsics, and class-specific average object heights. An image resolution of 1024×1024 is used for all experiments. We follow standard object detection protocols and report COCO-style mean average precision (mAP), together with distance-wise mAP computed over four distance bins ($mAP_{0\text{–}50m}$, $mAP_{50\text{–}150m}$, $mAP_{150\text{–}250m}$, and $mAP_{>250m}$). We additionally report PASCAL-style mAP at IoU thresholds of 0.5 and 0.75.

5.1. Architecture Ablation Experiments

This study provides an analysis of which foundation model backbones and training schemes are most effective for ultra-long range detection.

Table 1 presents the ablation of the proposed Telescope pipeline. Absolute improvements are highlighted in bold. Notably, the de-noising training scheme significantly improves overall accuracy across all distance bins, confirming the utility of this approach for fine-tuning foundation models [28]. While the foveated transform slightly degrades performance at short distances, it consistently im-

Table 2. **Distance-wise Ablation Experiments of Encoder and Head**. All models perform well for objects close to the ego vehicle, but have degraded performance as distance increases. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, and denoising-based training to provide optimal performance, especially at ultra-long range distances and across all classes.

Method	COCO					PASCAL	
	mAP	$mAP_{0\text{-}50}$	$mAP_{50\text{-}150}$	$mAP_{150\text{-}250}$	mAP_{250+}	mAP_{50}	mAP_{75}
DINOv2 + Deformable DETR	0.186	0.432	0.208	0.108	0.042	0.375	0.161
DINOv3 + Deformable DETR	0.212	0.467	0.250	0.137	0.059	0.419	0.190
SAM3 + Deformable DETR	0.317	0.523	0.344	0.244	0.171	0.521	0.329
SAM3 + Denoising [56]	0.501	0.692	0.483	0.321	0.292	0.758	0.545

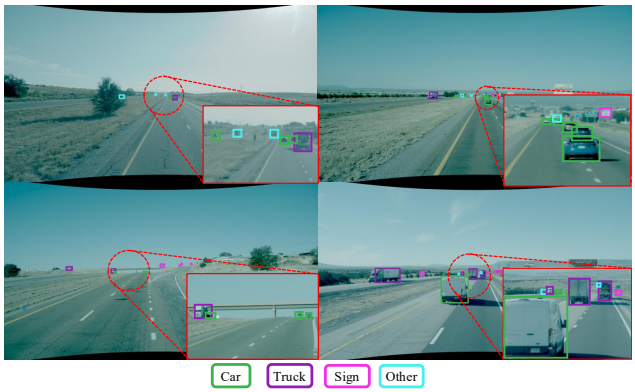

Figure 6. **Qualitative Visualization.** Detections from Telescope on the TruckDrive [16] dataset. Telescope consistently detects and localizes distant vehicles that occupy only a few pixels, while preserving accurate predictions for nearby objects. These examples highlight the effect of the proposed hyperbolic foveated transform in magnifying ultra-long range regions and improving sensitivity to objects in these ranges. Cut-outs provide a high-resolution, zoomed-in region to more clearly visualize objects.

Table 3. **Class-wise Ablation Experiments for Backbone and Head**. Results for the TruckDrive dataset for each class for existing state-of-the-art models as well as foundation model ablations. Leveraging foundation models as pre-trained image encoders provides a strong prior for object detection comparable to state-of-the-art specialized models. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, and de-noising training scheme to provide optimal performance across all classes.

Method	COCO mAP					
	Person	Bike	Sign	Car	Truck	Debris
DINOv2 + Deformable DETR	0.154	0.302	0.175	0.302	0.301	0.092
DINOv3 + Deformable DETR	0.149	0.304	0.222	0.345	0.334	0.120
SAM3 + Deformable DETR	0.106	0.236	0.415	0.459	0.442	0.245
SAM3 + Denoising [56]	0.330	0.436	0.617	0.631	0.596	0.451

proves performance in the medium, long, and ultra-long regimes. Importantly, the proposed transform significantly reduces the performance gap between near and far distance bins, producing a more balanced detector across spatial scales. This behavior reflects the design goal of hyperbolic foveation, which magnifies distant objects and compresses nearby ones.

We next study the influence of the image encoder, detection head, and training scheme on model performance. We evaluate five foundation model encoders, SAM3 [4], DINOv2 [33], DINOv3 [39], and Perception Encoder [2], in combination with both DETR [5] and Deformable DETR [60] heads. To ensure a controlled comparison, all encoders are frozen and only the detection heads are trained from scratch for 12 epochs using identical hyperparameters. The results are summarized in Table 2. DETR-based models and models using the Perception Encoder consistently fail to converge and are therefore omitted from further ablation analysis. In contrast, Deformable DETR yields stable training and substantially better performance across all distance ranges. Interestingly, the Perception Encoder backbone, despite being used internally by SAM3, also fails to converge for both DETR and Deformable DETR heads when used directly under the same training protocol.

Among the tested encoders, SAM3 provides the strongest overall performance. This reflects the strong spatial inductive biases inherited from large-scale segmentation pre-training. While DINOv2 and DINOv3 achieve competitive results for several categories, their overall accuracy remains lower under the same compute budget. Table 3 further shows the class-dependent behavior across foundation encoders. DINO-based encoders tend to perform slightly better on pedestrians and bicycles, whereas SAM-based encoders favor vehicles and traffic signs. As the application of this ultra-long range object detection network is autonomous highway driving, this motivates the use of the SAM3 backbone as it aligns with the application domain. Based on these results, we adopt the SAM3 encoder with a Deformable DETR head in all subsequent experiments.

We evaluate the denoising training scheme [56] and associated losses using the SAM3 image encoder. As shown in Tables 1 2, and 2, de-noising improves detection accuracy across all distance bins and classes and is therefore adopted within the Telescope model.

Figure 6 presents qualitative examples showing that the full 2-stage Telescope model accurately detects objects at long and ultra-long ranges.

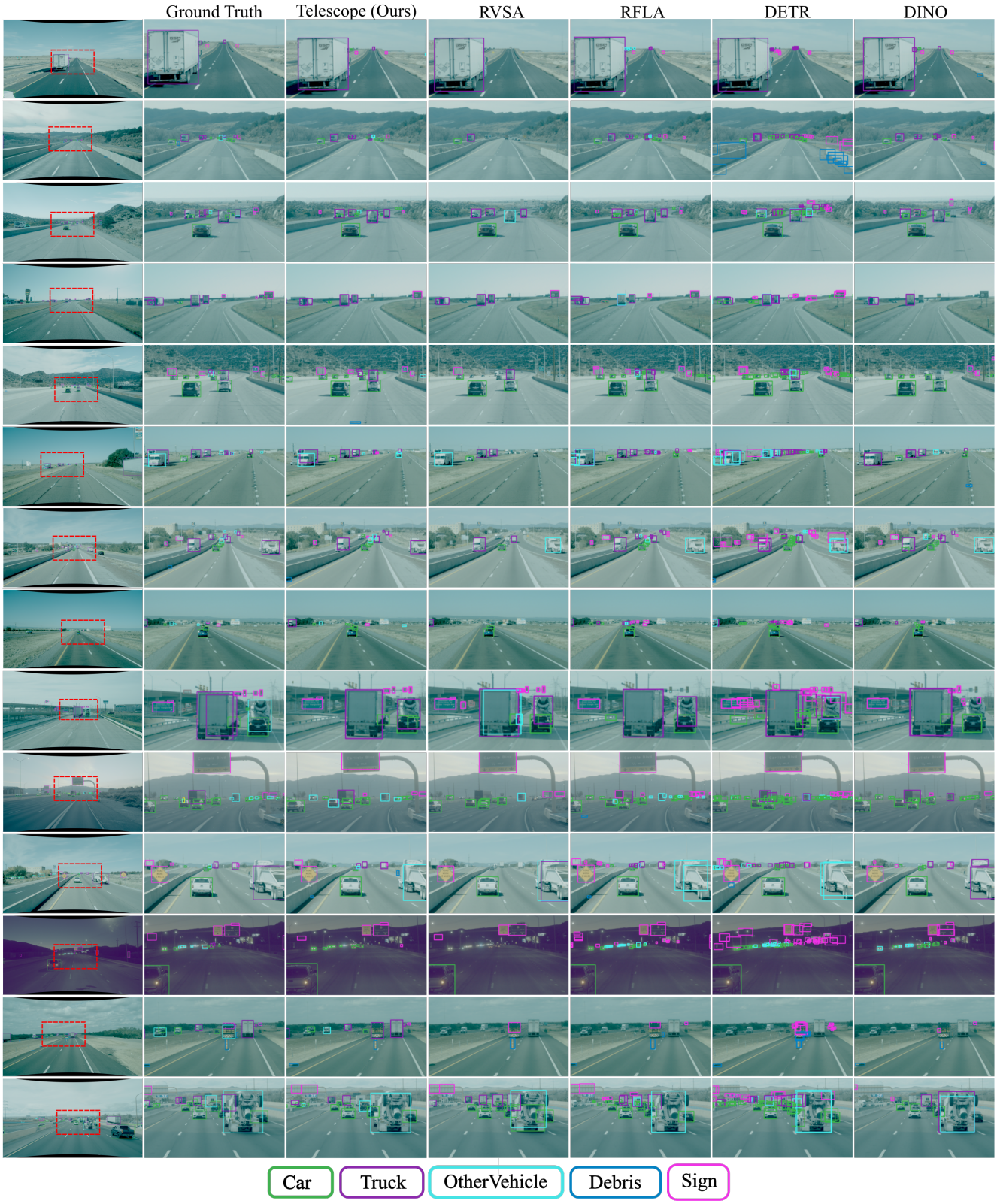

Figure 7. **Qualitative Comparison.** Qualitative comparison between the proposed method, Telescope, and state-of-the-art baselines. Both RVSA [44] and RFLA [52] are specialized for small object detection while DETR [5] and DINO [56] are strong general object detectors, but perform worse in long an ultra-long range object detection. Ground truth annotations are shown on the left. Zoomed-in views corresponding to the red rectangles are provided to highlight detections at long and ultra-long range, where some objects reach up to 1km. All baselines are fine-tuned on the TruckDrive [16] dataset.

Table 4. **Distance-Wise Ultra-long Range Object Detection Evaluation on TruckDrive dataset**. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, de-noising training, and TeleScope re-sampling layer provides optimal performance, especially at ultra long range distances, where it significantly improves over previous methods.

Method	COCO					PASCAL	
	mAP	$mAP_{0\text{-}50}$	$mAP_{50\text{-}150}$	$mAP_{150\text{-}250}$	mAP_{250+}	mAP_{50}	mAP_{75}
DETR [5]	0.166	0.396	0.178	0.081	0.072	0.335	0.147
Grounding DINO [28]	0.286	0.376	0.262	0.147	0.156	0.417	0.296
FOVEA [42]	0.113	0.169	0.086	0.008	0.005	0.189	0.115
YOLO11x [19]	0.266	0.421	0.218	0.134	0.117	0.510	0.195
DINO [56]	0.222	0.335	0.239	0.189	0.179	0.371	0.226
QueryDet [53]	0.248	0.449	0.286	0.199	0.094	0.415	0.257
UniverseNet [38]	0.305	0.518	0.334	0.236	0.145	0.474	0.318
RVSA [44]	0.325	0.502	0.298	0.233	0.183	0.488	0.351
RFLA [52]	0.306	0.501	0.320	0.239	0.185	0.512	0.317
Telescope (Ours)	0.497	0.608	0.507	0.335	0.326	0.801	0.494

Table 5. **Class-Wise Ultra-Long Range Detection Evaluation on TruckDrive dataset**. Leveraging foundation models as pre-trained image encoders provides a strong prior for object detection comparable to state-of-the-art specialized models. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, de-noising training, and TeleScope re-sampling layer provides optimal performance across all classes.

Method	COCO mAP					
	Person	Bike	Sign	Car	Truck	Debris
DETR [5]	0.222	0.327	0.179	0.299	0.247	0.083
Grounding DINO [28]	0.141	0.174	0.472	0.591	0.317	0.024
FOVEA [42]	0.028	0.043	0.060	0.377	0.056	0.107
YOLO11x [19]	0.172	0.400	0.370	0.330	0.376	0.049
DINO [56]	0.059	0.819	0.264	0.431	0.334	0.226
QueryDet [53]	0.034	0.222	0.329	0.469	0.349	0.185
UniverseNet [38]	0.165	0.240	0.429	0.538	0.427	0.230
RVSA [44]	0.069	0.297	0.429	0.551	0.467	0.203
RFLA [52]	0.085	0.209	0.436	0.560	0.434	0.225
Telescope (Ours)	0.454	0.620	0.568	0.651	0.595	0.397

5.2. Ultra Long Range Object Detection

We next compare Telescope against state-of-the-art 2D object detectors in Tables 4 and 5. All baselines are initialized from the best publicly available checkpoints and fine-tuned on TruckDrive for 12 epochs following [16]. Telescope is trained following the same protocol used in the ablation study described in Section 5.1. Notably, QueryDet [53], UniverseNet [38], RVSA [44], FOVEA [42], and RFLA [52] are all designed for small-object detection, while Grounding DINO [28] is a state-of-the-art visual-language model.

Baselines in Tables 4 and 5 rely on backbones pre-trained on relatively modest datasets (e.g., ImageNet [14]) and require small-object–specific losses, data augmentations, and specialized training strategies. In contrast, Tables 2 and 3 show that simply initializing from a stronger foundation encoder (SAM3), trained at much larger scale, and applying standard training on modest dataset already matches or exceeds these specialized methods. Incorporating de-noising and foveation further yields substantial gains, clearly outperforming the strongest existing approaches by a wide margin as seen in Table 4.

These results demonstrate that explicitly re-balancing object scales through hyperbolic foveation, together with a pre-trained image encoder and de-noising training, improves sensitivity to distant objects with an mAP increase of 76% for ultra-long range (increasing it from 0.185 to 0.326 for distances greater than 250m) without sacrificing overall detection quality.

A qualitative comparison between the proposed method, the strongest small-object detection baselines, and widely used general object detectors is shown in Figure 7. DETR, DINO, and RFLA tend to over-predict the number of objects in a scene, whereas RVSA is more conservative and produces fewer detections. Telescope offers a middle ground, reducing false positives relative to DETR and DINO while maintaining higher recall than RVSA.

Additional medium and long-range ($< 250\text{m}$) experiments on the Argoverse [6, 47] dataset are presented in the Appendix.

6. Conclusion

We present Telescope, a two-stage algorithm for ultra-long range object detection that explicitly addresses the extreme scale imbalance inherent in autonomous highway driving scenarios. In the first stage, we introduce a learnable hyperbolic foveated transform that magnifies distant regions while compressing nearby ones, normalizing object scales and reducing the dominance of large, nearby objects. In the second stage, we combine this transformation with a high-resolution detection architecture built on a foundation model image encoder and Deformable DETR detection head, enabling efficient processing and training without the quadratic cost of standard self-attention.

Experiments on the long-range TruckDrive dataset demonstrate that the proposed foveation consistently improves detection accuracy for distant objects and reduces the performance gap between near and far ranges. In particular, Telescope achieves a 53% relative improvement in overall performance (increasing overall mAP from 0.325 to 0.497) but most notably achieves up to a 76% relative improvement in mAP over the strongest existing baselines at ultra-long distances, increasing absolute mAP from 0.185 to 0.326 at distances greater than $250\mathrm{m}$.

We note that the proposed hyperbolic foveated transform is architecture-agnostic and invertible, and can be readily integrated into existing high-resolution perception pipelines, including future multi-modal and vision–language detection systems. We believe this work establishes a principled and extensible foundation for addressing the representational challenges of simultaneous perception at close surroundings with up beyond hundreds of meters to kilometer-scale distances.

7. Limitations and Scope

Telescope is a research contribution to one component of a broader autonomous perception system. Deployment in safety-critical applications would require integration with complementary sensing modalities, system-level validation, and compliance with applicable regulatory frameworks. The results reported here reflect performance on the TruckDrive dataset and should not be interpreted as a guarantee of real-world system performance.

Acknowledgements

Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award and a Amazon Science Research Award. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a cofounder of Cephia AI.

References

[1] Bai, Y., Zhang, Y., Ding, M., Ghanem, B.: Finding tiny faces in the wild with generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–30 (2018)

[2] Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception Encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)

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

[4] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A.e.a.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)

[6] Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3D tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8748–8757 (2019)

[7] Chen, C., Liu, M.Y., Tuzel, O., Xiao, J.: R-CNN for small object detection. In: Asian conference on computer vision. pp. 214–230. Springer (2016)

[8] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)

[9] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: Survey and benchmarks. IEEE transactions on pattern analysis and machine intelligence **45**, 13467–13488 (2023)

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

[11] Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic DETR: End-to-end object detection with dynamic attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2988–2997 (2021)

[12] Dai, Z., Cai, B., Lin, Y., Chen, J.: UP-DETR: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1601–1610 (2021)

[13] De Plaen, H., De Plaen, P.F., Suykens, J.A., Proesmans, M., Tuytelaars, T., Van Gool, L.: Unbalanced optimal transport: A unified framework for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3198–3207 (2023)

[14] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[15] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)

[16] Ghilotti, F., Palladin, E., Brucker, S., Sigal, A., Bijelic, M., Heide, F.: TruckDrive: Long-range autonomous highway driving dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)

[17] Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., Han, Z.: Effective fusion factor in FPN for tiny object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1160–1168 (2021)

[18] Guo, G., Chen, P., Yu, X., Han, Z., Ye, Q., Gao, S.: Save the tiny, save the all: Hierarchical activation network for tiny object detection. IEEE transactions on circuits and systems for video technology **34**(1), 221–234 (2023)

[19] Hidayatullah, P., Syakrani, N., Sholahuddin, M.R., Gelar, T., Tubagus, R.: YOLOv8 to YOLO11: A comprehensive architecture in-depth comparative review. arXiv preprint arXiv:2501.13400 (2025)

[20] Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., Yang, R.: The Apolloscape dataset for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 954–960 (2018)

[21] Jabbireddy, S., Sun, X., Meng, X., Varshney, A.: Foveated rendering: Motivation, taxonomy, and research directions. arXiv preprint arXiv:2205.04529 (2022)

[22] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)

[23] Lee, G., Hong, S., Cho, D.: Self-supervised feature enhancement networks for small object detection in noisy images. IEEE signal processing letters **28**, 1026–1030 (2021)

[24] Lee, J.M.: Introduction to Riemannian manifolds, vol. 2. Springer (2018)

[25] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR training by introducing query denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13619–13627 (2022)

[26] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

[27] Liu, J., Zhang, J., Ni, Y., Chi, W., Qi, Z.: Small-object detection in remote sensing images with super-resolution perception. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 15721–15734 (2024)

[28] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)

[29] Mao, J., Shi, S., Wang, X., Li, H.: 3D object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision **131**(8), 1909–1963 (2023)

[30] Mirzaei, B., Nezamabadi-Pour, H., Raoof, A., Derakhshani, R.: Small object detection and tracking: A comprehensive review. Sensors **23**(15), 6887 (2023)

[31] Nguyen, N.D., Do, T., Ngo, T.D., Le, D.D.: An evaluation of deep learning methods for small object detection. Journal of electrical and computer engineering **2020**(1), 3189691 (2020)

[32] Noh, J., Bae, W., Lee, W., Seo, J., Kim, G.: Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9725–9734 (2019)

[33] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[34] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L.e.a.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

[35] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)

[36] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

[37] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)

[38] Shinya, Y.: USB: Universal-scale object detection benchmark. arXiv preprint arXiv:2103.14027 (2021)

[39] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)

[40] Stanoyevitch, A., Stegenga, D.A.: The geometry of Poincaré disks. Complex Variables and Elliptic Equations **24**(3-4), 249–265 (1994)

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

[42] Thavamani, C., Li, M., Cebron, N., Ramanan, D.: FOVEA: Foveated image magnification for autonomous navigation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15539–15548 (2021)

[43] Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)

[44] Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L.: Advancing plain vision transformer toward remote sensing foundation model. IEEE transactions on geoscience and remote sensing **61**, 1–15 (2022)

[45] Wang, J., Xu, C., Yang, W., Yu, L.: A normalized Gaussian Wasserstein distance for tiny object detection. arXiv preprint arXiv:2110.13389 (2021)

[46] Wei, W., Cheng, Y., He, J., Zhu, X.: A review of small object detection based on deep learning. Neural Computing and Applications **36**(12), 6283–6303 (2024)

[47] Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)

[48] Wong, K., Gu, Y., Kamijo, S.: Mapping for autonomous driving: Opportunities and challenges. IEEE Intelligent Transportation Systems Magazine **13**(1), 91–106 (2020)

[49] Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)

[50] Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022)

[51] Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.S.: Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 79–93 (2022)

[52] Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.S.: Rfla: Gaussian receptive field based label assignment for tiny object detection. In: European conference on computer vision. pp. 526–543. Springer (2022)

[53] Yang, C., Huang, Z., Wang, N.: QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 13668–13677 (2022)

[54] Yu, X., Gong, Y., Jiang, N., Ye, Q., Han, Z.: Scale match for tiny person detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1257–1265 (2020)

[55] Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access **8**, 58443–58469 (2020)

[56] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)

[57] Zhao, Y., Zhu, F., Mi, Y., Chen, D., Xiong, G.: Simple-FPN: An image anomaly detection and localization network based on SimpleNet and feature pyramid. In: 2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI). pp. 417–422. IEEE (2024)

[58] Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)

[59] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)

[60] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

[61] Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3), 257–276 (2023)

Appendix

Section A reports details on image-based object distance estimation as well as distance-based statistics and information regarding the Argoverse 2 [47] autonomous driving dataset. Section B provides an additional evaluation of the proposed network and several baselines on the Argoverse dataset. Section C provides additional details regarding the hyperbolic foveation and its inverse, including computation times, existence of the inverse, and convergence guarantees. Lastly, Section D provides additional details regarding the proposed network, Telescope, architecture while Section E discusses training details.

A. Dataset Analysis

This section discusses the Argoverse 2 dataset [47]. This dataset contains objects in the near ($\leq$ 50m), far ($50-150$m) and long ($150-250$m) ranges, however no ultra-long ($\geq$ 250m) objects are present. In Sec. A.1 we demonstrate how object ranges are computed using the object bounding box and camera parameters. Next, we analyze the Argoverse dataset and the distribution of object ranges in Sec. A.2.

A.1. Distance Approximation from Bounding Boxes

Objects in the TruckDrive [16] dataset are annotated at distances extending up to $1\,\mathrm{km}$, where reliable LiDAR or radar measurements are often unavailable. Therefore, to estimate object distances, we follow [16] and approximate depth using the apparent size of the object in the image together with the camera intrinsics and an average class height prior.

Let h_p denote the height of the detected bounding box in pixels, f the camera focal length in pixels, and H_c the average real-world height of the object class c given in Table 6. For the TruckDrive dataset, the focal length is $f = 3304$. Under the pinhole camera model, the object distance d can be approximated as

$$d \approx \frac{f H_c}{h_p}. \tag{6}$$

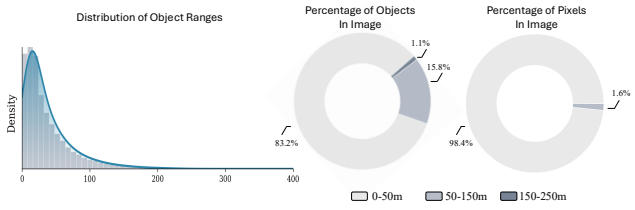

Figure 8. **Long-range Objects in Argoverse Driving.** Analysis of the Argoverse [47] dataset shows the distribution of object distances and the breakdown of the pixel-wise composition of objects at each distance. Multiple object ranges are represented in images, but nearby objects are disproportionately favored in terms of pixel area, with far (50-150m) and long (150-250m) range objects occupying only a small fraction of image pixels.

A.2. Argoverse Dataset

Similar to TruckDrive, an analysis of the Argoverse [47] dataset is provided in Figure 8. Notably, the Argoverse dataset only contains objects up to long range (e.g. $<$ 250m). Nevertheless, we provide an ablation on Telescope and baseline performance to demonstrate the generalizability of the the proposed approach in Tables 7 and 8.

B. Evaluation on Argoverse Dataset

To demonstrate the generalizability of the proposed object detection model, Telescope, we perform further evalautions using the Argoverse dataset. While this dataset does not contain ultra-long range objects, it does contain objects up to $<$ 250m. To align with the evaluation in Section 5.2, object distances are estimated from bounding box height, camera intrinsics, and class-specific average object heights denoted in Table 6. The camera focal length in pixels for the Argoverse 2 dataset is 1682.

As in Section 5.2, an image resolution of 1024×1024 is used for all experiments. We follow standard object detection protocols and report COCO-style mean average precision (mAP), together with distance-wise mAP computed over three distance bins ($\text{mAP}_{0\text{–}50\text{m}}$, $\text{mAP}_{50\text{–}150\text{m}}$, and $\text{mAP}_{150\text{–}250\text{m}}$). We additionally report PASCAL-style mAP at IoU thresholds of 0.5 and 0.75.

Based on the findings of Table 4 and 5, we train and evaluate the *three best-performing baselines*, Universenet [38], RVSA [44], and RFLA [52]. Baselines and Telescope are all fine-tuned for 5 epochs.

As reported in Tables 7 and 8, we find that the proposed model, Telescope, outperforms all baselines across all distance ranges and classes. Evaluations are computed in the same manner as in Tables 4 and 5 with TruckDrive. This reflects the findings from Section 5.2, confirming the efficacy and generalizability of Telescope across both the TruckDrive [16] and Agoverse [47] datasets. Qualitative results for the Argoverse dataset are presented in Figure 9, along with additional qualitative results from the TruckDrive dataset in Figure 10.

C. Hyperbolic Foveation Computation

We first prove the existence of the inverse of the hyperbolic foveated transform. We then show that this inverse

Table 6. **Average Class Heights**. Average heights across the class used to compute the approximate object distance given bounding box height and camera focal length.

Class	Person	Bike	Car	Sign	Truck	Debris
Average Height [m]	0.70	0.71	1.89	1.26	2.90	0.41

Table 7. **Distance-Wise Object Detection Evaluation on Argoverse 2 Dataset**. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, de-noising training, and TeleScope re-sampling layer provides optimal performance across all distances.

Method	COCO				PASCAL	
	mAP	$mAP_{0\text{-}50}$	$mAP_{50\text{-}150}$	$mAP_{150\text{-}250}$	mAP_{50}	mAP_{75}
UniverseNet [38]	0.123	0.156	0.042	0.016	0.271	0.100
RVSA [44]	0.121	0.150	0.052	0.026	0.260	0.098
RFLA [52]	0.106	0.131	0.036	0.023	0.250	0.070
Telescope (Ours)	0.232	0.268	0.104	0.036	0.502	0.177

Table 8. **Class-Wise Object Detection Evaluation on Argoverse 2 Dataset**. Leveraging foundation models as pre-trained image encoders provides a strong prior for object detection comparable to state-of-the-art specialized models. The proposed model using the SAM3 image encoder backbone, Deformable DETR detection head, de-noising training, and TeleScope re-sampling layer provides optimal performance across all classes. Shown are results for the 7 most common classes in Argoverse 2.

Method	COCO mAP						
	Regular Vehicle	Pedest.	Bollard	Const. Barrel	Stop Sign	Bicycle	Wheeled Device
UniverseNet [38]	0.396	0.251	0.069	0.353	0.209	0.151	0.098
RVSA [44]	0.381	0.183	0.099	0.192	0.120	0.151	0.091
RFLA [52]	0.352	0.233	0.044	0.269	0.153	0.124	0.070
Telescope (Ours)	0.562	0.412	0.154	0.417	0.316	0.355	0.259

is approximated using the Newton-Raphson algorithm with guarantees on convergence.

Theorem 1 (Existence of the Inverse). Assume $\alpha, p, R > 0$. Then Φ is a diffeomorphism on $\mathbb{R}^2$.

Proof. For $r > 0$, the map in (2) is a smooth radial deformation centered at o with strictly positive radial derivative $\partial_r \|\Phi(x) - o\| > 0$ since both the hyperbolic contraction $\tanh(\alpha r)$ and the interpolation weight $w(r)$ are monotone in r.

The Jacobian of Φ is everywhere non-singular, implying local invertibility. Global injectivity follows from strict radial monotonicity, and surjectivity follows from $\Phi(x) = x$ for $r \geq R$. Hence Φ is a diffeomorphism. □

Theorem 2 (Convergence of the inverse approximation). Let $y = \Phi(x^\star)$. If the Jacobian $J_\Phi(x^\star)$ is non-singular, then the Newton–Raphson iteration $x^{(k+1)} = x^{(k)} - J_\Phi(x^{(k)})^{-1}(\Phi(x^{(k)}) - y)$ converges locally and quadratically to $x^\star$.

Proof. Since Φ is continuously differentiable and $J_\Phi(x^\star)$ is invertible by Theorem 1, the standard Newton–Raphson convergence theorem applies, yielding local quadratic convergence. □

Table 9. Network Details of SAM3 + DINO 2-Stage Model.

Component	Sub-Component	Layer	Parameters
SAM3 Backbone	ViT (ViTDet)	Patch Embed Transformer	patch: 14, dim: 1024 depth: 32, heads: 16, mlp: 4.625×
	FPN Neck	ConvTranspose2d Conv2d	×2 upsample per scale (1024, 256), k=1; (256, 256), k=3
Foveation Estimation	Head	AdaptiveAvgPool2d MLP	output: 1×1 (256, 128, 4), ReLU, Sigmoid/Softplus
	Foveation Embed	MLP	(4, 64, 256), ReLU
DINO Transformer	Input Proj	Conv2d + GN	(256, 256), k=1, GN(32)
	Encoder ×6	MSDeformAttn FFN	heads: 8, levels: 3, points: 4 (256, 2048, 256), ReLU
	Decoder ×6	MSDeformAttn FFN	heads: 8, levels: 3, points: 4 (256, 2048, 256), ReLU
	Query Selection	Top-k	k=300 proposals from encoder
Detection Head	Class Embed Bbox Embed Label Embed (DN)	Linear MLP Embedding	(256, C) (256, 256, 4), 3 layers (C+1, 256)

Transform Runtime We also analyze the computation time for the forward and backwards transform. We randomly initialize 100 boxes, use a batch size of 4, and run 50 transformations. The Euclidean to Riemannian (forward) transformation takes 1.86 ± 0.08ms. For the Riemannian to Euclidean (backwards) transformation ir takes approximately 8 Newton-Raphson iterations to achieve an error tolerance of $< 1e-06$, which combined take 16.6 ± 2.73ms.

D. Network Architecture Details

We leverage the SAM3 image encoder for this work. This encoder is derived from the Perception Encoder [2] and uses a ViT (Vision Transformer) with 32 layers, an embedding dimension of 1024, and 14×14 patches. Windowed local self-attention is used with global self-attention every 7th layer, alongside 2D RoPE and a SimpleFPN neck [57] that produces 256-dimensional feature maps at $4\times$, $2\times$, and $1\times$ resolutions. The ViT backbone and FPN are frozen for all experiments.

The proposed model illustrated in Figure 2 of the main manuscript, Telescope, consists of two stages. In the first stage, the input image is down-sampled to 512×512 and passed through the SAM3 image encoder. The $1\times$ feature resolution outputs are then flattened and run through a 3-layer MLP which estimates the four foveation parameters. These parameters are the foveation center, $[c_x, c_y]$, and the foveation radius, R_x and R_y. For simplicity, the maximum radius is used (i.e. $\max(R_x, R_y)$.

In the second stage, the hyperbolic foveated transform is applied to the original resolution image. The transformed image is then down-sampled to 1024×1024 and passed through the SAM3 image encoder. All three feature resolutions are then used as inputs to a Deformable DETR [60] detection head consisting of a 256-d deformable encoder and decoder with 4 sampling points per level. Two-stage refinement is used from encoder proposals. A 3-layer MLP head then estimates the Riemannian bounding box parameters as discussed in Section 3.3. See Table 9 for details on the network parameters.

Figure 9. **Additional Qualitative Comparison on Argoverse Dataset.** Qualitative comparison between the proposed method, Telescope, and state-of-the-art baselines specialized for small object detection. Ground truth annotations are shown on the left. Notably, there are many target boxes which represent occluded objects (rows 1, 2, 3, 6, and 7). All methods are fine-tuned on the Argoverse [47] dataset.

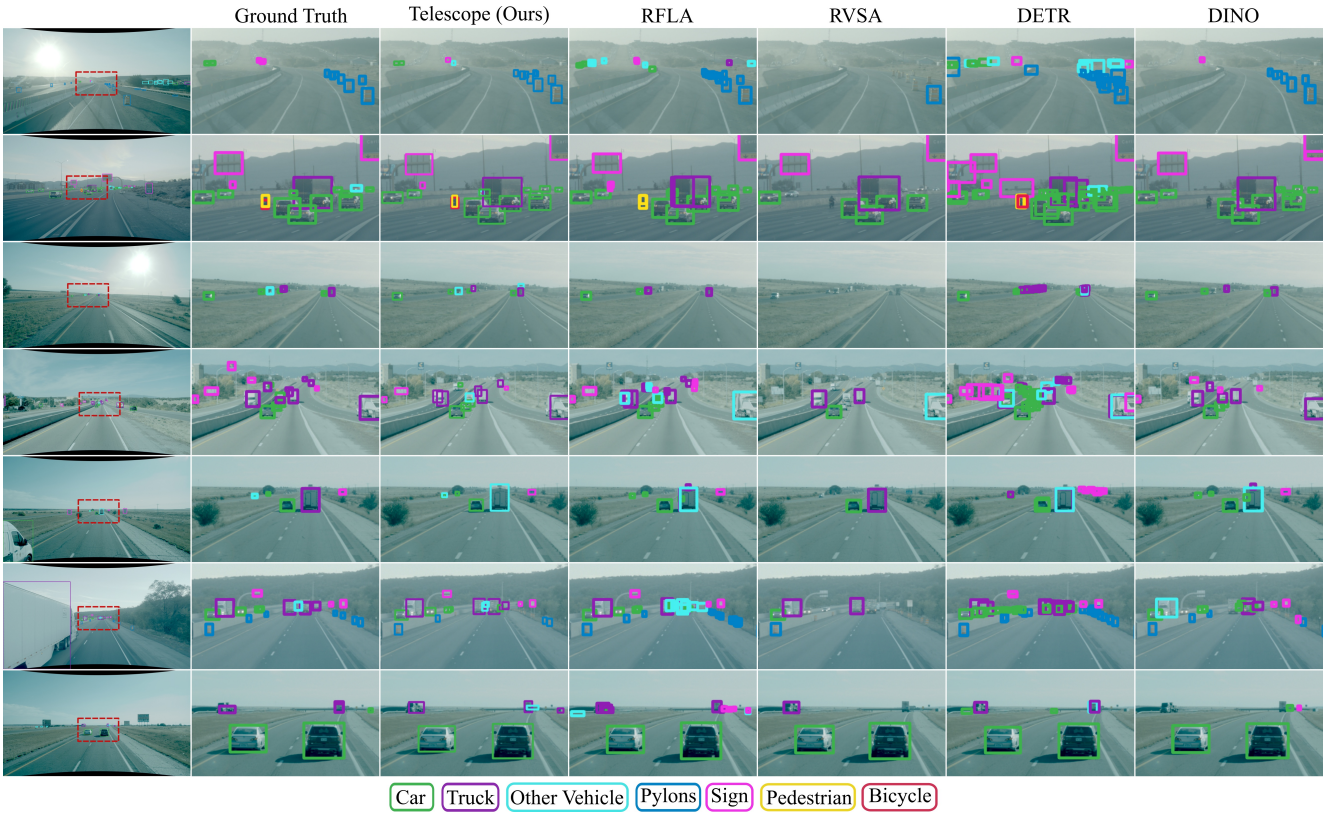

Figure 10. **Additional Qualitative Comparison on TruckDrive Dataset.** Qualitative comparison between the proposed method, Telescope, and state-of-the-art baselines. Both RVSA [44] and RFLA [52] are specialized for small object detection while DETR [5] and DINO [56] are strong general object detectors, but perform worse in long an ultra-long range object detection. Ground truth annotations are shown on the left. Zoomed-in views corresponding to the red rectangles are provided to highlight detections at long and ultra-long range, where some objects reach up to 1km. All baselines are fine-tuned on the TruckDrive [16] dataset.

E. Training Details

In our training, we use the denoising proposed in [56], where ground truth bounding box parameters are first noised, concatenated with the object queries, and then denoised by the detection head. This helps stabilize the matching process during learning and provides an early training signal to the detection head for bounding box localization. For Telescope, the Euclidean ground truth boxes are first noised, then projected to the Riemannian space via (2) and appended to the queries. This ensures that these boxes remain in the same space as the network predictions.

For all Telescope experiments, we used a learning rate of $1e-04$, a lambda learning rate schedule with 1 warm-up epoch, a batch size of 4, 300 decoder queries, and trained across 2 A100 GPUs. For baselines, the default training parameters specified in the publicly available repos were used.