Subset Selection for Autonomous Driving Datasets via Fine-Tuning Vision Foundation Models

Mert Keser*,[1,2], Sinan Simsek*,[3], Deniz Karacor[3], and Alois Knoll[1]

***Abstract*—Training large-scale vision models for autonomous driving is computationally expensive and requires extensive manual annotation. While reducing dataset size could address these limitations, it typically results in degraded model performance. In this paper, we propose a novel self-supervised data selection framework that leverages vision foundation models to identify and retain high-value training samples, enabling efficient dataset curation without compromising performance. Our approach fine-tunes a foundation model's vision encoder using a contrastive objective, then perform density-based clustering in its learned embedding space to retain only those samples that maximally preserve semantic diversity. Through experiments, we show that training on our curated subset outperforms models trained on the full dataset, and exceeds random selection in semantic segmentation tasks. Additionally, our comparisons across different foundation model architectures and segmentation backbones provide insights into effective dataset curation. Our results highlight that self-supervised data selection can significantly reduce both annotation and computational overhead, providing a scalable alternative to naively expanding datasets.**

I. Introduction

Deep learning has revolutionized computer vision, particularly in autonomous driving where high-quality perception is crucial for safety-critical decisions [1]–[4]. However, the increasing complexity of autonomous driving models demands not only substantial computational resources but also massive datasets of diverse traffic scenarios, weather conditions, and road environments [5]. Training state-of-the-art perception models can require extensive computation time and significant GPU resources, making development impractical for many research teams and companies with limited resources [6]. While reducing dataset size could mitigate these computational demands, traditional sampling approaches often lead to significant performance degradation, and manual selection becomes infeasible given the scale of modern autonomous driving datasets [7], [8]. This challenge is further amplified by the inherent characteristics of autonomous driving data: they contain high-resolution multi-view images with large memory footprints, require expensive expert annotation for safety-critical tasks, and suffer from substantial redundancy due to consecutive frames capturing nearly identical scenes [9], [10]. Moreover, the common assumption that all training samples contribute equally to model learning has been increasingly challenged, suggesting that automated and intelligent data selection could maintain or even improve model performance while dramatically reducing computational and storage requirements [7], [11].

Vision Foundation Models (VFMs), have emerged as powerful tools for understanding and representing visual data through their training on vast and diverse datasets [12]–[14]. Through self-supervised learning on internet-scale data, these models have demonstrated remarkable capabilities in learning rich semantic representations that generalize across different visual domains [15]. Their ability to project diverse images into a semantically meaningful embedding space, where similar scenarios naturally cluster together, makes them particularly suitable for analyzing large-scale autonomous driving datasets. This semantic understanding, combined with their strong generalization capabilities across different data distributions, has made VFMs valuable for various computer vision tasks [12], [13].

In this work, we introduce a dataset subset selection framework that leverages the representational power of vision foundation models to address redundancy in large-scale autonomous driving datasets. As illustrated in fig. 1, our approach consists of two key components: First, we fine-tune the VFM image encoder using a contrastive learning objective [16], optimizing it to capture the nuanced similarities between driving scenarios in its latent space. Second, we leverage this learned representation through density-based clustering to identify and retain the most representative samples, effectively eliminating redundancy while preserving the semantic diversity crucial for autonomous driving tasks. Through this two-stage process, our method can automatically identify and select the most informative samples while removing redundant data points, particularly beneficial for autonomous driving datasets where consecutive frames often capture highly similar scenes. We demonstrate the efficacy of our approach on the task of semantic segmentation, where extensive experiments reveal that our intelligent sampling strategy not only outperforms random data selection but also achieves competitive—and in some cases superior—performance compared to training on the full dataset.

Our contributions are two-fold. First, we propose a novel two-stage subset selection framework that leverages self-supervised fine-tuning of VFMs and density-based clustering to mitigate redundancy in large-scale autonomous driving datasets while preserving critical semantic diversity. Second, extensive experiments on semantic segmentation tasks reveal that our approach achieves performance on par with full dataset training using only a fraction of the data, outperforming conventional random sampling.

*Authors contributed equally.
[1]Technical University of Munich, Germany mert.keser@tum.de
[2]Continental AG, Germany
[3]Baskent University, Türkiye

Manuscript 616 submitted to 2025 36th IEEE Intelligent Vehicles Symposium (IV).
Received February 6, 2025.

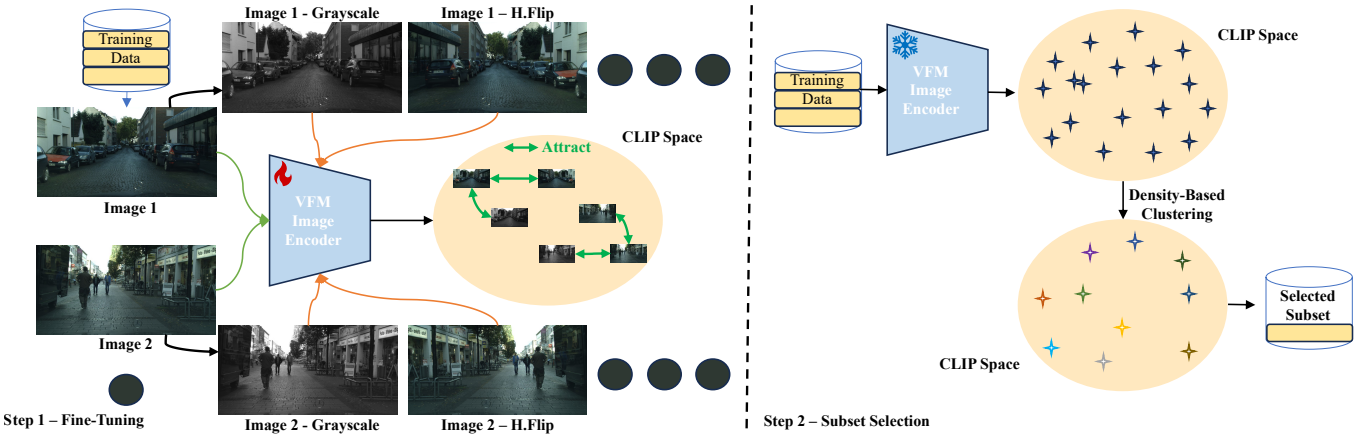

Fig. 1. Overview of our two-stage dataset subset selection framework. Left: Fine-tuning the VFM image encoder using contrastive learning, where augmented versions of input images are mapped to similar representations in CLIP space. Training pairs are created from the same base image to encourage feature consistency across different visual transformations. Right: Using the fine-tuned encoder for subset selection through density-based clustering to identify representative samples.

II. Related Work

In this section, we examine three key areas of related work. First, we review VFMs and their applications (section II-A). We then explore VFMs' role in autonomous driving (section II-B), followed by an analysis of subset selection methods (section II-C) for efficient dataset curation.

A. Vision Foundation Models

VFMs have emerged as a transformative paradigm in computer vision, analogous to large-scale language models in natural language processing. They are trained on extensive and diverse image collections—often without explicit labels—to learn a universal feature representation that can be adapted to a broad spectrum of downstream tasks with minimal fine-tuning. By capturing both low-level and high-level semantics, these models offer robust generalization, even under distributional shifts or limited labeled data.

A well-known example is CLIP [17], which employs image-text pairs to align visual and textual modalities within a shared embedding space. This alignment enables zero-shot classification and flexible retrieval, demonstrating strong resilience to style variations and out-of-distribution scenarios. Another prominent line of research, exemplified by DINO [18], exploits self-distillation to refine features without explicit labels, thereby boosting performance in tasks such as object detection and segmentation. Building on DINO's principles, DINOv2 [19] incorporates masked image modeling for increased stability and scalability, facilitating training on even larger datasets while retaining transferability. Further extending these ideas into the realm of text-conditioned object detection, Grounding DINO [20] precisely localizes image regions based on natural language prompts.

Parallel to these developments, the Segment Anything Model (SAM) [21] has risen to prominence in object segmentation. Leveraging an extensive training corpus, SAM produces high-quality segmentation masks in a zero-shot manner, underscoring its robustness across diverse tasks. Collectively, these VFMs provide a solid and flexible backbone for vision-centric pipelines, particularly in scenarios where large labeled datasets are scarce or where significant domain shifts are anticipated. Their general-purpose encoders can be easily integrated and fine-tuned, often requiring only minimal overhead, ultimately improving reliability in real-world vision applications. This adaptability and efficacy underscore the foundation role these models play in driving advances in modern computer vision.

For a comprehensive overview of the field, we refer readers to several surveys and reviews [12], [22].

B. Vision Foundation Models in Autonomous Driving

VFMs, particularly SAM [21] and DINO [19], have emerged as powerful tools in autonomous driving perception tasks. These models have demonstrated capabilities in understanding complex driving scenes and have been adapted to address various challenges specific to autonomous vehicles. In the autonomous driving domain, several pioneering works have leveraged VFMs to enhance perception capabilities. For instance, Calib-Anything [23] introduces a zero-shot approach using SAM for LiDAR-camera calibration, eliminating the need for additional training data. Shan et al. [24] conduct comprehensive studies on SAM's segmentation performance under adverse weather conditions, crucial for ensuring reliable autonomous driving systems in challenging environments. The adaptation of VFMs has also led to significant advances in semantic understanding of driving scenes. SPINO [25] demonstrates the effectiveness of DINOv2's [19] task-agnostic features for few-shot panoptic segmentation across diverse autonomous driving datasets. Furthermore, VFMs have proven valuable in enhancing 3D perception capabilities. SEAL [26] introduces self-supervised representation learning for large-scale 3D point cloud processing using SAM-inspired architectures, while Peng et al. [27] improve unsupervised domain

Manuscript 616 submitted to 2025 36th IEEE Intelligent Vehicles Symposium (IV).
Received February 6, 2025.

adaptation in 3D semantic segmentation through instance mask utilization. RadOcc [28] advances 3D scene understanding by incorporating SAM for shape priors and segment-guided affinity distillation in occupancy prediction tasks. Despite these advancements, challenges remain in adapting VFMs for autonomous driving applications, particularly in fully capturing 3D spatial information and effectively integrating multi-modal sensor data, such as LiDAR point clouds. These limitations stem from the specialized architectures required for processing diverse sensor inputs in autonomous driving systems.

C. Dataset Subset Selection

The challenge of reducing large-scale image datasets while maintaining model performance has become increasingly critical in modern machine learning workflows, particularly in autonomous driving where data volumes are substantial. Among various strategies, clustering-based approaches, notably k-means variants, have emerged as foundational methods due to their computational efficiency. Ougiaroglou et al. [29] pioneered a method utilizing k-means clustering to identify representative samples through centroid selection. While effective for simple classification datasets, this approach faces limitations in autonomous driving scenarios where scenes exhibit complex object co-occurrences. The ERHC approach [30] advanced this concept by employing k-median clustering, offering improved robustness to outliers. Further refinements by Panhalkar and Doye [31] introduced nearest-neighbor centroid selection to enhance instance diversity. However, recent research by Byerly et al. [32] challenges these methods, suggesting that samples distant from cluster centers often carry more informative features.

Alternative strategies have focused on feature space compression while maintaining the original dataset size. Birvinskas et al. [33] demonstrated the effectiveness of Discrete Cosine Transform for signal compression, while Reddy et al. [34] extensively evaluated Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for large-scale datasets. Although these dimensionality reduction methods successfully reduce computational requirements, they do not address the fundamental challenge of redundant sample retention in the training process.

Recent developments have introduced more sophisticated approaches to dataset curation. Embedding-based tools, exemplified by Labelbox [35], facilitate efficient data organization through semantic clustering, though they remain dependent on human supervision. Submodular optimization and coreset selection methods [36], [37] offer theoretical guarantees for data coverage but incur significant computational costs. Recently, Stage et al. [11] demonstrated that an AD dataset can be reduced using a similarity-based approach in the latent space of CLIP. However, the method is highly dependent on the chosen threshold.

The diverse landscape of subset selection methods underscores the ongoing challenge of balancing three critical factors: dataset size reduction, semantic diversity preservation, and computational efficiency. Notably, while these methods have been extensively validated on classification tasks with clearly defined class boundaries, their application to complex autonomous driving scenarios remains largely unexplored. In the context of autonomous driving, where dataset complexity includes high-resolution multi-view images, varying environmental conditions, and intricate urban scenes, developing methods that effectively address these trade-offs while maintaining model performance remains an active and challenging research area.

III. Methodology and Evaluation

This section first presents the vision foundation models and their corresponding backbone architectures used in our experimental evaluation (section III-A). We then detail our framework's methodology (section III-B), followed by a description of the benchmark dataset used for evaluation (section III-C). Finally, we introduce the downstream task and evaluation metrics used to assess our framework's effectiveness (section III-D).

A. Pre-trained Vision Foundation Models

We evaluate CLIP [17] with multiple backbone architectures to assess their effectiveness in dataset subset selection. Specifically, we examine both ResNet variants (RN50, RN101) [38] and Vision Transformer (ViT-B/16) [39], providing a comparative analysis between traditional convolutional architectures and modern transformer-based approaches [1].

B. Dataset Subset Selection Methodology

Our framework performs dataset subset selection through a two-stage process: (1) fine-tuning a vision foundation model to ensure semantically similar images are embedded closer together, and (2) density-based analysis in the learned feature space to identify representative samples. Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote our training dataset with N samples, where each $x_i \in \mathbb{R}^{H\times W\times 3}$ represents an RGB image.

In the first stage, we fine-tune a VFM encoder:

$$f_\theta : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d \tag{1}$$

using multi-view contrastive learning. For each image x_i, we apply a stochastic augmentation function $\mathcal{T}$ to generate k different views:

$$v_i^j = \mathcal{T}(x_i)_{j=1}^{k} \tag{2}$$

These views are then mapped to a d-dimensional embedding space by the encoder:

$$z_i^j = f_\theta(v_i^j), \quad z_i^j \in \mathbb{R}^d \tag{3}$$

We optimize the encoder parameters θ using a contrastive objective that maximizes the cosine similarity between embeddings of different views of the same image while minimizing the cosine similarity across different images:

[1]CLIP: https://github.com/openai/CLIP

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{k}\sum_{l=j+1}^{k} \log \frac{\exp((z_i^j)^T z_i^l/|z_i^j||z_i^l|/\tau)}{\sum_{m\neq i}\exp((z_i^j)^T z_m^l/|z_i^j||z_m^l|/\tau)} \tag{4}$$

where $\tau > 0$ is a temperature parameter that controls the concentration of the distribution.

In the second stage, we employ DBSCAN [40], [41] to identify representative samples in the learned feature space. For each sample x_i, we compute its ℓ_2-normalized embedding:

$$e_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}. \tag{5}$$

Let $\mathcal{E} = \{\, e_i\}_{i=1}^N$ be the set of all normalized embeddings. A point $p \in \mathcal{E}$ is a *core point* if

$$|\{\, q \in \mathcal{E} : \|e_p - e_q\|_2 \leq \epsilon\}| \;\geq\; \text{minPts}, \tag{6}$$

where $\epsilon > 0$ is the maximum neighborhood radius, and minPts is the minimum number of points required to form a dense region. Points that do not satisfy this criterion are considered outliers. In this way, DBSCAN adapts to the underlying distribution by grouping semantically similar embeddings into clusters while isolating sparse outliers.

Let $C_j \subset \mathcal{E}$ denote the j-th cluster produced by DBSCAN. Its centroid is given by

$$\mu_j = \frac{1}{|C_j|}\sum_{i\in C_j} e_i. \tag{7}$$

To form the final subset $\mathcal{S}$, we select the $\lfloor \alpha|C_j| \rfloor$ samples in C_j closest to μ_j under cosine similarity, for $\alpha \in (0, 1]$. Concretely,

$$\mathcal{S} \;=\; \bigcup_j \{\text{top } \lfloor \alpha|C_j| \rfloor \text{ samples in } C_j \text{ closest to } \mu_j\}\,.$$

This approach ensures that $\mathcal{S}$ retains the semantic diversity of the original dataset while reducing redundancy.

C. Dataset

We evaluate our framework on the Cityscapes dataset [42], which contains high-resolution (2048×1024) urban street scenes from 50 different cities. The dataset includes 2,975 training images and 500 validation images, all with fine-quality pixel-level semantic annotations across 19 classes. Each image in Cityscapes captures complex urban environments with diverse scenes including roads, buildings, vehicles, pedestrians, and various static objects.

We use the standard training set of Cityscapes ($\mathcal{D}_{train}$ with 2,975 images) to perform our subset selection. The corresponding validation set ($\mathcal{D}_{val}$ with 500 images) is used to evaluate the performance of models trained on our selected subsets.

D. Evaluation and Metrics

We evaluate our subset selection approach using DeepLabV3+ [43] as our semantic segmentation architecture. To demonstrate the generalizability of our method across different network capacities, we employ three distinct backbone networks: MobileNetV3 [44], ResNet50, and ResNet101 [38]. Let $\mathcal{S}$ denote our selected subset and $\mathcal{D}_{train}$ the full training set. For each backbone architecture f_ϕ with parameters ϕ, we train the model to minimize the cross-entropy loss:

$$\mathcal{L}_{CE}(\phi) = -\frac{1}{|\mathcal{S}|}\sum_{(x,y)\in\mathcal{S}}\sum_{c=1}^{C} y_c \log(f_\phi(x)_c) \tag{8}$$

where C is the number of Cityscapes classes and y_c represents the one-hot encoded ground truth for class c. Models are optimized using SGD with momentum of 0.9, initial learning rate of 0.01, and weight decay of $1e^{-4}$.

For evaluation, we use the mean Intersection over Union (mIoU) metric on the validation set $\mathcal{D}_{val}$. For a predicted segmentation mask $\hat{y}$ and ground truth y, the mIoU is computed as:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{|\hat{y}_c \cap y_c|}{|\hat{y}_c \cup y_c|} \tag{9}$$

where $\hat{y}_c$ and y_c represent the predicted and ground truth masks for class c, respectively.

IV. Experiments

This section introduces our experimental results. First, we examine the fine-tuning of CLIP's latent space (section IV-A) to enhance feature representation for autonomous driving scenarios. We then evaluate our subset selection methodology through comprehensive experiments (section IV-B).

A. Fine-Tuning CLIP

While CLIP [17] has demonstrated capabilities in learning visual representations from internet-scale data, its feature space may not inherently capture the nuanced semantic relationships crucial for autonomous driving scenarios. To investigate this limitation and our proposed solution, we conduct a systematic analysis of representation dynamics through contrastive fine-tuning.

Figure 1 presents a comparative study of CLIP embeddings before and after our domain-specific fine-tuning. We select two urban driving scenes from Cityscapes and apply four distinct augmentations: grayscale conversion, color jittering, horizontal flipping, and posterization. These transformations, while altering the visual appearance, preserve the semantic content relevant to AD tasks. By projecting these embeddings into 2D space, we reveal crucial insights into the representation learning process.

In the base CLIP embeddings (left), we observe a transformation-centric organization where visually similar augmentations cluster together across different base images.

Manuscript 616 submitted to 2025 36th IEEE Intelligent Vehicles Symposium (IV).
Received February 6, 2025.

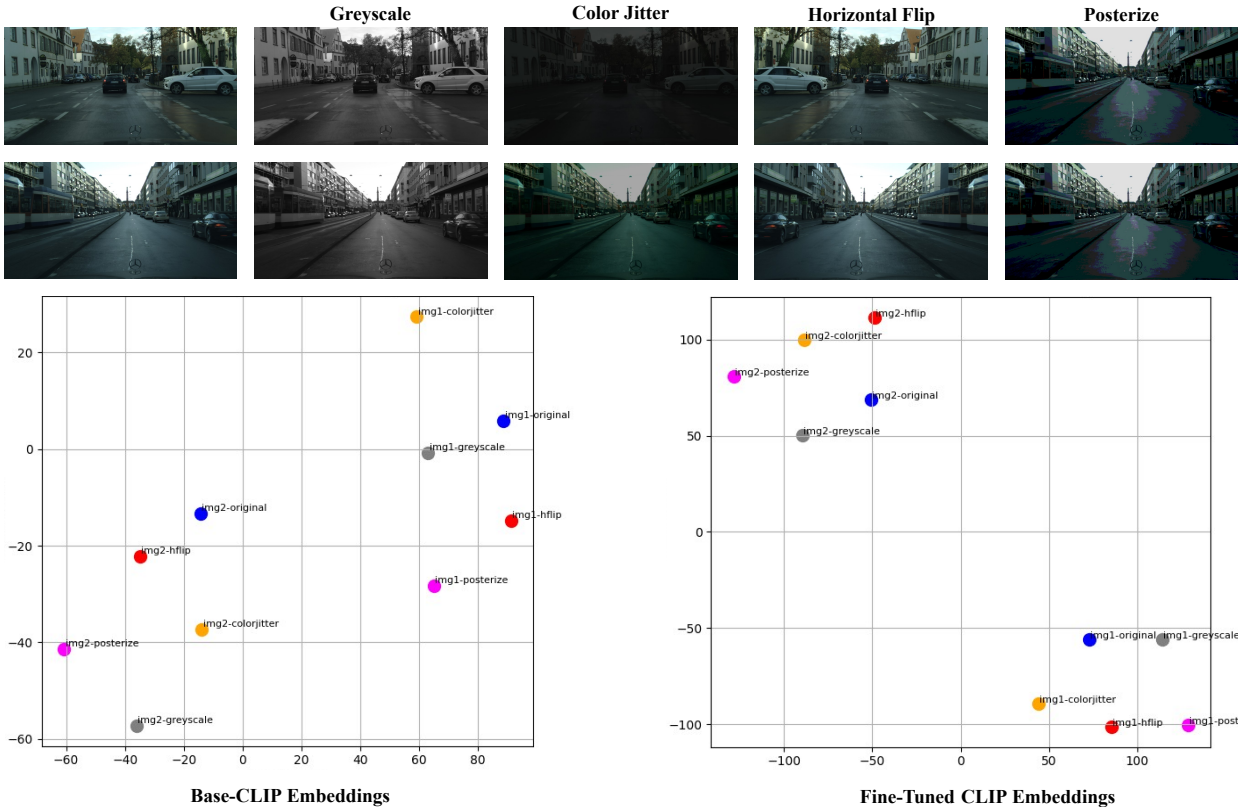

Fig. 2. Effect of CLIP fine-tuning on feature representations. Top: Two random Cityscapes images with their augmented versions (Greyscale, Color Jitter, Horizontal Flip, and Posterize). Bottom: t-SNE visualization of embeddings before (left) and after (right) fine-tuning.

For instance, all grayscale transformations appear proximally located, regardless of their source scene. This suggests that the pre-trained CLIP model exhibits a bias toward low-level visual features, potentially a consequence of its internet-scale training where style variations dominate semantic relationships.

The fine-tuned CLIP embeddings (right) demonstrate a fundamental reorganization of the feature space. Here, we observe the emergence of semantically coherent clusters where different augmentations of the same base image form tight groupings. This reorganization indicates that our fine-tuning approach successfully shifts the model's attention from low-level transformations to scene-level semantic understanding. Notably, the distance between augmented views of the same scene significantly decreases, while the separation between different urban scenes is maintained or enhanced. The CLIP [17] model now recognizes that different views of the same street scene should be considered semantically equivalent, even under various visual transformations.

B. Subset Selection Experiment

We evaluate our subset selection framework across different sampling ratios $\alpha \in \{0.1, 0.25, 0.5\}$, corresponding to 10%, 25%, and 50% of the original dataset $\mathcal{D}_{train}$. We perform experiments with three different backbone architectures for DeepLabV3+ [43], comparing their performance on our selected subsets against both random sampling and full dataset training. This setup enables us to analyze both the effectiveness of our selection method across different dataset sizes and its generalization across network architectures of varying capacities. The segmentation results for different sampled datasets are shown in Table I.

The results in Table I demonstrate that our VFM-guided selection using CLIP-ViT-B/16 consistently outperforms random sampling across all backbone architectures. Notably, with only 50% of the training data, our method achieves performance on par with or exceeding that of full-dataset training when using the MobileNetV3 backbone.

Among the CLIP variants, ViT-B/16 exhibits superior performance over its ResNet-based counterparts, particularly at lower sampling rates. This suggests that the transformer-based architecture's ability to capture rich semantic representations is more effective in identifying representative samples. Furthermore, the performance improvements remain consistent across different segmentation backbones, highlighting the robustness and generalizability of our selection strategy across architectures of varying capacity.

V. Conclusion

This work presents a framework for dataset subset selection in autonomous driving through vision foundation models. Our approach demonstrates that carefully curated subsets, selected through a combination of self-supervised fine-tuning

Manuscript 616 submitted to 2025 36th IEEE Intelligent Vehicles Symposium (IV).
Received February 6, 2025.

TABLE I
SEMANTIC SEGMENTATION PERFORMANCE (MIOU) ON CITYSCAPES VALIDATION SET USING DIFFERENT SUBSET SELECTION STRATEGIES WITH THREE DEEPLABV3+ [43] BACKBONES. BOLD NUMBERS INDICATE BEST PERFORMANCE PER COLUMN.

Segmentor Backbone	**MobileNetV3**				**ResNet50**				**ResNet101**			
Reduction Rate	**10%**	**25%**	**50%**	**100%**	**10%**	**25%**	**50%**	**100%**	**10%**	**25%**	**50%**	**100%**
Random	0.552	0.618	0.685	0.727	0.607	0.663	0.728	0.745	0.631	0.692	0.735	0.769
CLIP-RN50	0.545	0.631	0.701	0.727	0.595	0.681	0.731	0.745	0.610	0.689	0.740	0.769
CLIP-RN101	0.556	**0.646**	0.691	0.727	0.602	0.666	0.727	0.745	0.616	0.692	0.740	0.769
CLIP-ViT-B/16	**0.561**	0.639	**0.728**	0.727	**0.622**	**0.690**	**0.744**	0.745	**0.645**	**0.698**	**0.764**	0.769

and density-based clustering in the learned feature space, can achieve performance comparable to full dataset training. Our evaluation on Cityscapes reveals that models trained on our selected subsets not only consistently outperform random sampling but achieve competitive—and in some cases superior—performance using only 50% of the original data.

Several promising research directions emerge from this work. First, investigating the generalizability of our framework across different autonomous driving datasets and diverse perception tasks beyond semantic segmentation would provide valuable insights into its broader applicability. Second, adapting this methodology for online subset selection in real-time autonomous systems presents an important challenge, particularly in scenarios where data distributions evolve dynamically.

Through this work, we challenge the paradigm that larger datasets necessarily lead to better performance, demonstrating instead that subset selection can maintain model performance while significantly reducing computational overhead.

ACKNOWLEDGMENT

The research paper was written in the context of the "NXT GEN AI Methods" research project funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), The authors would like to thank the consortium for the successful cooperation

REFERENCES

[1] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2020.

[2] J. Zhang, J. Cao, J. Chang, X. Li, H. Liu, and Z. Li, "Research on the application of computer vision based on deep learning in autonomous driving technology," *arXiv preprint arXiv:2406.00490*, 2024.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[4] M. Cummings and D. Britton, "Regulating safety-critical autonomous systems: past, present, and future perspectives," in *Living with robots*. Elsevier, 2020, pp. 119–140.

[5] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] X. Wang, M. A. Maleki, M. W. Azhar, and P. Trancoso, "Moving forward: A review of autonomous driving software and hardware systems," *arXiv preprint arXiv:2411.10291*, 2024.

[7] Z. Wan, Z. Wang, C. Chung, and Z. Wang, "A survey of dataset refinement for problems in computer vision datasets," *ACM computing surveys*, vol. 56, no. 7, pp. 1–34, 2024.

[8] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[9] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2019.

[10] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: from model driven to data driven," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–34, 2022.

[11] H. Stage, L. Ewecker, J. Langner, T. S. Sohn, T. Villmann, and E. Sax, "Reducing computer vision dataset size via selective sampling," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 1422–1428.

[12] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[13] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao, and A. C. Knoll, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[14] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] J. Wu, B. Gao, J. Gao, J. Yu, H. Chu, Q. Yu, X. Gong, Y. Chang, H. E. Tseng, H. Chen *et al.*, "Prospective role of foundation models in advancing autonomous vehicles," *Research*, vol. 7, p. 0399, 2024.

[16] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," *Neurocomputing*, p. 128645, 2024.

[17] 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*. PMLR, 2021, pp. 8748–8763.

[18] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[19] 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.

[20] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[22] X. Zhou, M. Liu, B. L. Zagar, E. Yurtsever, and A. C. Knoll, "Vision language models in autonomous driving and intelligent transportation systems," *arXiv preprint arXiv:2310.14414*, 2023.

[23] Z. Luo, G. Yan, and Y. Li, "Calib-anything: Zero-training lidar-camera extrinsic calibration method using segment anything," *arXiv preprint arXiv:2306.02656*, 2023.

[24] X. Shan and C. Zhang, "Robustness of segment anything model (sam) for autonomous driving in adverse weather conditions," *arXiv preprint arXiv:2306.13290*, 2023.

[25] M. Käppeler, K. Petek, N. Vödisch, W. Burgard, and A. Valada, "Few-shot panoptic segmentation with foundation models," in *2024 IEEE*

International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 7718–7724.

[26] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] X. Peng, R. Chen, F. Qiao, L. Kong, Y. Liu, T. Wang, X. Zhu, and Y. Ma, "Sam-guided unsupervised domain adaptation for 3d segmentation," *arXiv preprint arXiv:2310.08820*, 2023.

[28] H. Zhang, X. Yan, D. Bai, J. Gao, P. Wang, B. Liu, S. Cui, and Z. Li, "Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7060–7068.

[29] S. Ougiaroglou and G. Evangelidis, "Efficient dataset size reduction by finding homogeneous clusters," in *Proceedings of the Fifth Balkan Conference in Informatics*, 2012, pp. 168–173.

[30] S. Ougiaroglou, K. I. Diamantaras, and G. Evangelidis, "Exploring the effect of data reduction on neural network and support vector machine classification," *Neurocomputing*, vol. 280, pp. 101–110, 2018.

[31] A. R. Panhalkar and D. D. Doye, "An approach of improving decision tree classifier using condensed informative data," *Decision*, vol. 47, pp. 431–445, 2020.

[32] A. Byerly and T. Kalganova, "Towards an analytical definition of sufficient data," *SN Computer Science*, vol. 4, no. 2, p. 144, 2023.

[33] D. Birvinskas, V. Jusas, I. Martisius, and R. Damasevicius, "Eeg dataset reduction and feature extraction using discrete cosine transform," in *2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation*. IEEE, 2012, pp. 199–204.

[34] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *Ieee Access*, vol. 8, pp. 54 776–54 788, 2020.

[35] "How to use embeddings to create high-quality training data," https://labelbox.com/blog/how-to-use-embeddings-to-create-highquality-training-data/, 2023, accessed: 2025-02-01.

[36] K. Wei, R. Iyer, and J. Bilmes, "Submodularity in data subset selection and active learning," in *International conference on machine learning*. PMLR, 2015, pp. 1954–1963.

[37] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[40] D. Deng, "Dbscan clustering algorithm based on density," in *2020 7th international forum on electrical engineering and automation (IFEEA)*. IEEE, 2020, pp. 949–953.

[41] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[42] 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*, 2016, pp. 3213–3223.

[43] 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)*, 2018, pp. 801–818.

[44] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.