Zero-Shot Learning for Advanced Vehicle Recognition

Adriana-Victorița Miu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
adrianamiuvictoria55@gmail.com

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—Zero-Shot Learning (ZSL) is an innovative approach in computer vision that enables models to recognize classes they have not encountered during training. Unlike traditional supervised methods, ZSL leverages semantic relationships, such as attributes or embeddings, to enable inference about novel categories. This capability is especially useful in fine-grained tasks like vehicle recognition, where identifying specific car makes and models requires generalization with minimal labeled data. Vehicle recognition in a ZSL context faces significant challenges, including subtle inter-class similarities and considerable intra-class variability, often influenced by environmental factors. Recent advancements address these challenges by employing multimodal models and vision-language representations, such as CLIP, RegionCLIP, and PaLI-GEMMA. These models use both visual and textual data, establishing joint embedding spaces for improved classification of unseen classes. Additionally, Large Vision-Language Models (LVLMs) provide enhanced multimodal input fusion, incorporating context from Large Language Models (LLMs) to generate detailed, context-rich vehicle descriptions. This study proposes a combined approach utilizing CLIP, RegionCLIP and PaLI-GEMMA to embed vehicles within a shared visual-semantic space, complemented by LLM-generated descriptions. Our approach demonstrates enhanced ZSL performance, improving recognition of unseen car makes and models.**

***Index Terms*—Zero-shot learning (ZSL), Vehicle recognition, Multimodal models, Object detection, Vehicle re-identification**

I. Introduction

In traditional supervised learning, models rely on large, labeled datasets to learn specific classes, limiting their ability to generalize beyond the trained categories. ZSL [4] addresses this limitation by transferring knowledge from known classes to unknown ones, leveraging semantic relationships through attributes or embeddings. This approach is particularly valuable for fine-grained recognition tasks, such as vehicle recognition, where models must identify specific car makes and models despite limited or absent labeled data.

The need for ZSL in vehicle recognition is underscored by challenges in acquiring diverse, high-quality labeled data for every possible make and model. Vehicles often exhibit subtle design differences, resulting in high inter-class similarity across models and substantial intra-class variability within a single model. These visual subtleties are further complicated by environmental factors, including variations in lighting, angles, and occlusion, which can impair the model's ability to generalize effectively [2]. Consequently, ZSL [5] offers a promising solution by enabling models to infer class relationships from shared characteristics, for example in allowing accurate recognition of unseen vehicle categories.

Recognizing specific car makes and models in a ZSL setting introduces unique challenges. Cars from different manufacturers or models often share visual characteristics, making inter-class similarity a prominent issue. Furthermore, cars of the same model can vary based on modifications, year or color, adding intra-class variability that complicates classification. These factors demand a robust recognition system capable of distinguishing fine details without reliance on exhaustive labeled data. The visual ambiguity in vehicle recognition requires ZSL models that can accurately represent semantic relationships to generalize across varied vehicle categories.

Recent ZSL models incorporate multimodal and vision-language frameworks to address the complexities of objects recognition. Key approaches include models like CLIP [1] and RegionCLIP [3], trained on extensive image-text datasets to create joint embedding spaces. CLIP enables zero-shot classification by aligning images with textual prompts, while RegionCLIP enhances this by focusing on region-specific features, allowing for finer visual-text alignment crucial for detailed object detection. Vision-Language Models (VLMs), such as PaLI-GEMMA [6], extend these capabilities by integrating both vision and language processing to create a contextual understanding across modalities, vital for distinguishing nuanced features, like vehicles attributes. Large Vision-Language Models (LVLMs) are pivotal for combining visual and textual data streams in a unified framework. These models typically include a vision encoder ViT, a text tokenizer, a projection layer like an MLP for aligning visual and textual embeddings, and a fusion mechanism that merges these embeddings before inputting them into a Large Language Model (LLM). This setup allows for robust multimodal input processing, with text outputs providing a grounded context that aids in vehicle recognition. The integration of LLMs further enhances this process by generating contextually rich prompts, enabling ZSL models to refine embeddings for visually similar classes with greater accuracy.

This study introduces a combined approach using CLIP,

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

RegionCLIP, and PaLI-GEMMA for ZSL-based vehicle recognition. By embedding vehicle features within a shared visual-semantic space, these models facilitate recognition through both image embeddings and descriptive prompts. An added car detection module enables precise localization of vehicles in images, while LLM-generated descriptions provide contextual depth. This approach leverages multimodal embeddings to enrich vehicle recognition under a ZSL framework, demonstrating improved model generalization for unseen car makes and models.

II. Related Work

A. Attribute-Based Models

Attribute-based models are a foundational approach in Zero-Shot Learning (ZSL), where the focus lies on representing classes through a set of human-defined or learned attributes. For objects, common attributes include ”color,” ”shape,” and ”size,” while in the context of vehicle recognition, attributes like ”body style,” ”engine type,” and ”wheel design” are used to describe different car models. These attributes play a crucial role in enabling models to map unseen classes to known classes by leveraging shared features between them. The concept of using attributes for ZSL was first introduced by Lampert et al. (2009) [8], who laid the groundwork for using such representations in unseen class recognition tasks. Over time, this concept has evolved, with significant advancements such as Xian et al. (2018) [5], who developed a framework for applying attribute-based information to generalized ZSL tasks.

B. GAN-Based Models

Generative Adversarial Networks (GANs) [9] [10] have been applied in ZSL to generate synthetic data for unseen classes. The key idea is that GANs can learn the distribution of known class features and then generate plausible instances of unseen classes by sampling from this learned space. This has been crucial in tasks where labeled data for new categories is challenging. In [11] is demonstrated that GANs can be used to generate synthetic images of unseen categories by learning from the attributes of seen classes using text descriptions. This technique can be extended to vehicle recognition by generating images of cars with unseen features or makes. GANs have been especially effective when coupled with attribute-based methods, where the model generates images conditioned on the attributes of an unseen vehicle class.

C. Traditional VMMR system

VMMR systems are responsible for identifying the specific make, model and potentially the year of manufacture of vehicles within an image, following the initial detection phase. According to Boukerche and Ma (2021) [12], traditional VMMR systems rely on deep learning methodologies, particularly Convolutional Neural Networks (CNNs), to extract and classify vehicle features. The VMMR process is a subcategory of fine-grained recognition, where subtle inter-class similarities and significant intra-class variances pose unique challenges. For example, two different car models may look very similar (high inter-class similarity), while the same model may vary due to color, modifications, or different model years (high intra-class variability). The authors note that environmental factors such as lighting, camera angles, and occlusions significantly impact recognition accuracy. Additionally, modifications to vehicles (e.g., added decals or accessories) and dataset limitations are critical challenges for robust VMMR performance. Traditional methods may incorporate License Plate Recognition (LPR) as an adjunct to VMMR; however, issues like damaged or obscured plates limit LPR’s reliability, making visual recognition methods more critical.

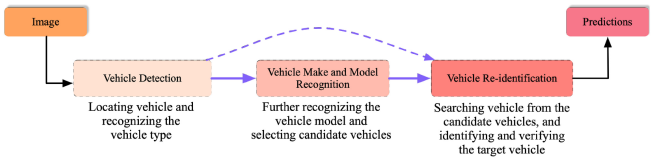

Fig. 1: Overview of the coarse to fine vehicle recognition procedure from [12]

III. Methodology

A. Challenges Specific to ZSL in Vehicle Recognition based on VLM

The approach for Zero Shot Learning in vehicle recognition presented in this article leverages vision-language models (VLMs), with a particular emphasis on the capabilities of the PaLI-GEMMA model. By integrating visual and language processing, PaLI-GEMMA facilitates a deeper contextual understanding necessary for identifying and distinguishing between different vehicle makes and models, even under zero-shot scenarios where labeled data for certain classes may be unavailable. This model’s structure allows for the creation of rich multimodal embeddings, enhancing the system’s ability to generalize and accurately recognize unseen vehicle categories. PaLI-GEMMA is an advanced, open vision-language model that extends the PaLI series by integrating it with the Gemma family of language models. It builds on PaLI’s lineage, which began with models like the original PaLI using a classification-pretrained Vision Transformer (ViT) and mT5 language model. PaLI-GEMMA strikes a balance between model size and performance, combining a 400M SigLIP vision encoder with a 2B Gemma language model into a compact sub-3B architecture. The architecture’s integration of SigLIP for visual representation and the auto-regressive decoder-only Gemma model for language processing makes PaLI-GEMMA versatile and efficient for a wide range of vision-language tasks. This includes standard benchmarks like COCO captions and VQAv2, as well as specialized challenges like remote-sensing visual question answering (VQA) and referring expression segmentation. For vehicle recognition, PaLI-GEMMA’s architecture is particularly beneficial due to its ability to effectively process and align visual and textual data. It can handle complex scenes with multiple vehicles, extracting detailed attributes and generating descriptive outputs that identify the

Fig. 2: Example of vehicle recognition output from the PaLI-GEMMA model. The model accurately identifies the car's make, model, and color from an image, responding with 'black golf'.

make, model and colour of the car. The model's smaller, optimized design allows it to perform high-level recognition while maintaining computational efficiency, making it well-suited for applications that require comprehensive vehicle analysis in diverse and potentially dense visual contexts. Other VLM tested and pretrained for vehicle recognition is CLIP, developed by OpenAI, that uses a dual-encoder architecture that aligns images and text in a shared embedding space, making it highly effective for zero-shot tasks by enabling prompts such as *"blue sports car"* or *"sedan with a specific logo"* to retrieve relevant images and classify vehicle attributes without specific prior training on them. Its ability to map images and textual descriptions allows CLIP to be leveraged for tasks like identifying the make and model of a vehicle by using tailored prompts. Another model utilised is BLIP-2, an evolution of BLIP, that enhances vision-language alignment by incorporating a transformer-based vision encoder with a text decoder that excels in generating detailed captions and understanding visual context. This helps in scenarios where recognizing not only the make and model but also unique series information is needed by generating text outputs that describe the car in detail, supported by the model's capability to process complex multimodal inputs. Llava, a recent vision-language model, is designed for image question-answering and multimodal dialogue, utilizing a highly refined alignment between its visual encoder and language model to respond to specific queries about vehicle details like *"What car make and model is shown in the image?"*. Its conversational capability allows for iterative refinement and clarification, useful for complex vehicle recognition tasks. PaLI-GEMMA, an advanced model designed for tasks involving multilingual and multimodal understanding, employs a mix of transformers optimized for text and image cross-modal tasks and is capable of processing dense scenes with multiple vehicles to identify their make, model, and series through sophisticated localization and attribute extraction. Each of these models brings its unique architectural strengths: CLIP's robust embedding alignment for quick zero-shot matching, BLIP-2's detailed image-caption generation, Llava's interactive Q&A for specific detail extraction, and PaLI-GEMMA's capacity to handle complex, scene-dense tasks.

B. Current Datasets utilized in training

- **Stanford Cars Dataset** [13]: It contains 16,185 images of 196 different car models across 10 categories. Each image in the dataset is labeled with the car's make, model, and year, making it a comprehensive dataset for both training and evaluation of vehicle recognition models.
- **CompCars Dataset** [14]: The CompCars dataset is a large-scale dataset designed for car recognition and comparison. It contains over 136,000 images of cars from 1,687 different car models, categorized into multiple types such as sedans, SUVs, and coupes. The dataset provides detailed annotations not only for the car make and model but also for car parts such as the front, side, and rear views. This diversity in view points is beneficial for training models *to recognize vehicles from different angles and under varying conditions*, which is critical for real-world applications.
- **VehicleID Dataset**: The VehicleID dataset focuses specifically on vehicle re-identification tasks, but it is also widely used for make and model classification. It includes over 220,000 images of more than 13,000 vehicles, distributed across 2,500 different car models. This dataset is designed to simulate the challenging task of identifying a specific vehicle across different cameras and viewpoints.
- **VMMR Dataset** [15] : The VMMR dataset is another large-scale collection specifically tailored for car make and model recognition tasks. It includes over 120,000 images of vehicles from *1,000 different car makes and models*. The dataset is structured to support both supervised and zero-shot learning tasks, making it valuable for training conventional machine learning models and fine-tuning Vision-Language Models for recognizing vehicles in unseen scenarios.

5. **Car Connection Dataset**: The Car Connection dataset is unique in that it focuses not only on vehicle make and model but also includes attributes such as vehicle features, trim level and market region. This dataset offers a more granular level of detail compared to traditional car recognition datasets and is particularly useful for fine-

tuning VLM. The dataset contains over 10,000 labeled images.

REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), 2021.

[2] Manzoor, M.A.; Morgan, Y. Vehicle Make and Model classification system using bag of SIFT features. In Proceedings of the 2017 IEEE 7th Annual IEEE Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 9–11 January 2017.

[3] Zhong, Yiwu, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[4] Palatucci, M. , Pomerleau, D. , Hinton, G. E. , & Mitchell, T. M. . (2009). Zero-shot Learning with Semantic Output Codes. Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. DOI:10.1184/R1/6476456.V1

[5] Y. Xian, C. H. Lampert, B. Schiele and Z. Akata, "Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2251-2265, 1 Sept. 2019

[6] Beyer, Lucas, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, Xiaohua Zhai, et al. PaliGemma: A versatile 3B VLM for transfer. In arXiv preprint arXiv:2407.07726, 2024.

[7] Yu, X., Aloimonos, Y. (2010). Attribute-Based Transfer Learning for Object Categorization with Zero/One Training Example. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6315. Springer, Berlin, Heidelberg.

[8] C. H. Lampert, H. Nickisch and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 951-958.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018.

[10] J. Dong, B. Xiao, B. Ding and H. Wang, "GT-GAN: A General Transductive Zero-Shot Learning Method Based on GAN," in IEEE Access, vol. 8, pp. 147173-147184, 2020.

[11] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng and A. Elgammal, "A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[12] Pavel, Monirul & Tan Esther, Siok Yee & Abdullah, Azizi. (2022). Vision-Based Autonomous Vehicle Systems Based on Deep Learning: A Systematic Literature Review. Applied Sciences. 12. 6831. 10.3390/app12146831.

[13] J. Krause, M. Stark, J. Deng and L. Fei-Fei, "3D Object Representations for Fine-Grained Categorization," 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, Australia, 2013, pp. 554-561, doi: 10.1109/ICCVW.2013.77.

[14] Yang, Linjie, Ping Luo, Chen Change Loy, and Xiaoou Tang. "A Large-Scale Car Dataset for Fine-Grained Categorization and Verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] Tafazzoli, S., Mirza, M., & Sadeghi, M. (2017). A Large and Robust Dataset for Fine-Grained Vehicle Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.