Can we recognize Cars we've never seen? A Journey Through Zero-Shot Learning in Vehicle Recognition

Adriana-Victorița Miu (Pavel)
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*—Vehicle Make and Model Recognition (VMMR) is vital for intelligent transportation and surveillance but remains challenged by subtle inter-class visual similarities, intra-class diversity, and the emergence of unseen vehicle types. While deep learning methods perform well on known classes, they struggle to generalize at scale. This paper surveys recent advances in zero-shot learning (ZSL) and vision-language models (VLMs) such as CLIP, LiT, and PaLI, which enable recognition without direct supervision. We further propose a hybrid pipeline combining prompt engineering, LoRA-based tuning, and InfoNCE regularization to address these limitations and support scalable, fine-grained vehicle classification.**

***Index Terms*—Vehicle Recognition, Zero-Shot Learning, Unseen Classes, Fine Grained, Vision Language**

I. Introduction

Vehicle Make and Model Recognition (VMMR), an emerging research area within Intelligent Transportation Systems (ITS), has recently gained significant attention [1]. In automotive terminology, "make" refers to the manufacturer of a vehicle (e.g., Honda, BMW, Nissan), whereas "model" denotes a particular vehicle type produced by that manufacturer (e.g., Civic, 3 Series, Altima). VMMR systems are widely utilized in various applications such as the identification of suspicious vehicles [2], traffic management [3], surveillance [4], anomaly detection [5], traffic analytics [6], autonomous vehicles [7], and the identification of suspicious vehicles [1] and law enforcement [5]. Despite substantial advancements in License Plate Recognition (LPR) technologies, which currently dominate vehicle detection and identification within security and surveillance systems, reliance exclusively on license plates remains problematic. License plates can be intentionally concealed, altered, or replaced, significantly impairing system reliability. Knowledge of a vehicle's make and model allows the retrieval of extensive data such as vehicle weight, dimensions, production year, maximum speed, engine horsepower, and capacity. Integrating this detailed vehicle information with LPR can substantially enhance the accuracy and efficacy of security and surveillance systems. VMMR represents a type of fine-grained intra-category classification, which is a specialized area within the broader field of fine-grained object recognition. Fine-grained classification aims to distinguish among highly similar objects within the same category. The large number of existing classes, combined with significant inter-class variations and subtle intra-class differences, makes fine-grained classification tasks, including VMMR, particularly challenging. These minimal visual distinctions can be extremely difficult for human observers to discern in certain instances.

Front-view vehicle images are considered the most effective choice for VMMR for two primary reasons. First, vehicles with identical body styles (sedan, SUV, MPV, truck, etc.) can appear highly similar from specific angles. [9] These similarities result from industry standards, such as aerodynamic requirements, complicating accurate recognition. Second, certain vehicle models from the same manufacturer share similar platforms, with only minor variations present primarily in their frontal sections, including headlights, grills, and bumpers. Consequently, front-view imagery provides the most discriminative visual features essential for reliable recognition.

One major challenge in developing deep learning-based VMMR systems is the scarcity of comprehensive, high-quality datasets suitable for network training. This challenge is intensified by the continual introduction of new vehicle models and the retirement of older ones, necessitating regular dataset updates to maintain relevance and accuracy. Additionally, real-world operational environments introduce further complexities. Variations in lighting conditions, obstructions, and environmental factors such as fog, snow, and dust significantly affect the reliability and precision of VMMR systems [8]. To address these challenges and enhance traffic security and monitoring efficiency, research in VMMR increasingly emphasizes the development of robust and accurate recognition methods using front-view vehicle images. This requires the availability of detailed and regularly updated datasets, which include extensive annotations for critical vehicle components such as headlights, grills, and bumpers. This detailed annotation supports targeted preprocessing and robust classification,

2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025

significantly enhancing the overall effectiveness and reliability of VMMR solutions. Moreover, the rise of unseen vehicle models highlights the limits of supervised learning. Zero-shot learning (ZSL) offers a practical solution by enabling recognition of classes not present during training, using semantic cues like text or attributes. This approach supports scalable and adaptable VMMR without constant data updates.

II. Related work in vehicle Recognition

Fine-grained vehicle make and model recognition (VMMR) is crucial in the broader context of Intelligent Transportation Systems (ITS) and Advanced Driver Assistance Systems (ADAS), aimed at enhancing safety, surveillance, and automated traffic management. This task, however, is inherently challenging due to issues such as inter-class similarity, where different vehicle models or makes exhibit highly similar visual features, and intra-class diversity, characterized by substantial variations within a single vehicle model due to generational changes, customization, or varied imaging conditions.

Traditional deep learning methods, particularly convolutional neural networks (CNNs), have significantly advanced VMMR by automating the extraction of discriminative features directly from extensive image datasets. Among the notable developments in this area is the DeepCar 5.0 framework proposed by Amirkhani and Barshooi (2023) [9], which employs a multi-agent ensemble learning system. This framework uses individual CNNs trained separately on distinct vehicle parts such as headlights, grills, and bumpers, thereby enhancing accuracy in challenging conditions including partial occlusion and variable illumination.

Complementing this, the Two-Branch Two-Stage architecture introduced by Lyu et al. (2022) [10] effectively mitigates classification ambiguity by segmenting the recognition task into separate branches for vehicle make and model identification. This structural approach significantly reduces confusion between closely related vehicle categories, consequently improving recognition precision. Furthermore, Lu et al. (2023) [11] proposed a part-level feature optimization strategy within CNN frameworks, optimizing and aggregating local features without additional manual annotations. This approach notably enhances performance in fine-grained tasks and provides computational efficiency. Another innovative method in CNN-based vehicle recognition is the Bag of Expressions (BoE) model presented by Jamil et al. (2020) [12]. BoE builds upon traditional Bag-of-Words methods by integrating neighborhood contextual information with Histogram of Oriented Gradients (HOG) descriptors, coupled with multi-class linear Support Vector Machines (SVMs). This model adeptly manages issues like scale variance and occlusions, demonstrating significant robustness and flexibility in practical applications.

More recent developments in deep learning have seen attention mechanisms and transformer architectures emerge as powerful tools in the domain of vehicle recognition. Attention mechanisms, as utilized by Amirkhani and Barshooi (2023) [9], allow for selective focus on critical vehicle regions, boosting fine-grained accuracy and robustness. Similarly, Taki

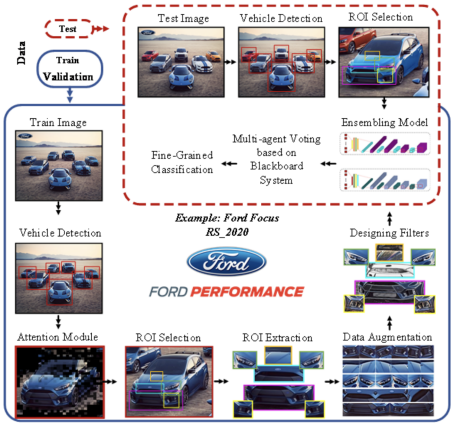

Fig. 1: Flowchart of the proposed approach [9] with multi-agent ensemble learning system

and Zemmouri (2023) [13] applied Vision Transformer (ViT) models to vehicle image classification tasks, demonstrating superior performance compared to traditional CNN methods, especially when processing small datasets with low-resolution images. Transformers inherently capture extensive relational information between image patches, providing strong advantages for complex visual tasks.

Data augmentation and clustering techniques have also been pivotal in enhancing CNN robustness. Nafzi et al. [14] effectively applied data augmentation alongside hierarchical clustering to address both inter-class similarity and intra-class diversity, significantly improving CNN performance. This methodological synergy provides a practical framework for continuously refining recognition models, essential for maintaining high performance in dynamic real-world scenarios.

In parallel to traditional deep learning approaches, alternative learning paradigms such as zero-shot learning (ZSL) and few-shot learning (FSL) have gained considerable attention due to their ability to handle new vehicle categories with limited or no prior training examples. Zero-shot learning specifically addresses the recognition of previously unseen vehicle models by leveraging semantic embeddings. Approaches such as CLIP, RegionCLIP, and SigLIP utilize visual-semantic mappings to generalize across unseen classes based on learned semantic relationships, thus providing a crucial solution for recognizing emergent vehicle models without explicit visual training.

Few-shot learning, exemplified by RelationNet++ (Kezebou et al.) [15], addresses scenarios where only minimal visual examples per new vehicle model are available. RelationNet++ effectively generalizes recognition capabilities using relational reasoning, surpassing conventional CNN approaches even with minimal training samples. This methodology demonstrates practical scalability and adaptability, essential for environments with frequently introduced new vehicle models.

Additionally, open-set recognition (OSR) represents another

crucial advancement, distinct from zero-shot learning. OSR approaches, such as the framework proposed by Vázquez-Santiago et al. [16], dynamically detect novel vehicle categories during real-time operation without pre-existing semantic information. Using probabilistic models and clustering techniques, OSR methods not only reject unknown instances but simultaneously identify and integrate them into the model, enabling continuous learning and adaptation. This is particularly beneficial in practical applications, where encountering unrecognized vehicle types is commonplace.

III. Zero-Shot techniques for unseen vehicle classes

Modern computer vision systems have achieved exceptional performance through supervised deep learning approaches trained on large-scale labeled datasets such as ImageNet (Deng et al., 2009) [17]. However, this progress is limited by the cost of manual annotation and the inability to anticipate all visual categories in advance. In fine-grained tasks such as vehicle make and model recognition, these limitations are particularly problematic: vehicle variants evolve rapidly, and data collection cannot keep pace with the emergence of new models (Dong et al., 2024; Semiromizadeh et al., 2025) [18], [19]. Zero-Shot Learning (ZSL) addresses this challenge by enabling the classification of unseen categories using side information such as semantic attributes or natural language descriptions. One foundational method, proposed by Zhang and Saligrama (2015), introduced Semantic Similarity Embedding (SSE): it maps both source domain data (attributes) and target domain data (visual features) into a shared semantic space where the similarity is defined via inner product on mixture histograms of seen class proportions. Their max-margin formulation learns class-dependent feature transformations that generalize well to unseen classes while remaining robust to noise. To extend ZSL beyond traditional attribute-based approaches, models like CLIP (Radford et al., 2021) [20] and LiT (Zhai et al., 2023) [21] leverage vision language contrastive pretraining, aligning text and image modalities into a shared embedding space. CLIP, trained on 400M image-text pairs, computes zero-shot classifier weights from textual prompts, enabling remarkable generalization across downstream tasks. LiT improves on this by "locking" the image encoder and tuning only the text side, reducing training time while maintaining performance. Yet, standard ZSL methods often fail under the Generalized Zero-Shot Learning (GZSL) setting, where models must correctly classify both seen and unseen classes. To overcome the bias toward seen classes, Zhang and Zhang (2024) [22] proposed a Semantic Feedback (SF) module integrated into a f-VAEGAN architecture. This model uses instance-level contrastive learning and semantic consistency constraints to ensure that generated features are semantically aligned with both seen and unseen class distributions, improving discriminative capacity under GZSL. Transformer architectures have also shifted the landscape of representation learning. The Vision Transformer (ViT) (Dosovitskiy et al., 2020) [23] replaced convolutions with self-attention, enabling scalable and interpretable image modeling. Extensions such as Tokens-to-Token ViT (Yuan et al., 2021) [24] and DINOv2 (Oquab et al., 2024) [25] optimize token aggregation and self-supervised pretraining to learn robust, transferable features. Notably, DINOv2 showed that dense patch-level representations are essential for localization and grounding—properties valuable for tasks like vehicle detection and fine-grained recognition. For fine-tuning and adaptation, prompt-based learning has become a lightweight yet powerful alternative. Zhou et al. (2022) [26] introduced Context Optimization (CoOp) for CLIP-like models, replacing handcrafted prompts with learnable context tokens. CoOp achieves over 15In the domain of vehicle recognition, Dong et al. (2024) [18] enhanced ViT-based classification with multi-scale patch fusion and inter-class attention, achieving superior performance over CNNs on BoxCars116k. Similarly, Semiromizadeh et al. (2025) [19] proposed using 3D spatial attention modules to capture geometric distinctions across makes and models, demonstrating the advantage of integrating spatial reasoning in zero-shot settings. This survey consolidates current progress in zero-shot learning with multimodal transformers, with a focus on loss functions, transformer evolution and applications to vehicle make and model recognition. By synthesizing advances across CLIP, PaLI, SigLIP, BLIP, and GZSL methods, we present a unified view of how generalization to unseen categories can be achieved without direct supervision—an essential requirement for scalable, real-world visual understanding systems.

A. Zero-Shot Learning: Conventional vs. Generalized

Zero-Shot Learning (ZSL) enables the classification of instances from classes unseen during training by leveraging auxiliary semantic information—such as attributes, word embeddings or textual descriptions—to bridge the semantic gap between seen and unseen categories (Zhang and Saligrama, 2015; Zhang et al., 2017) [27], [28]. In its conventional form, ZSL assumes that only unseen classes appear during testing, while the more realistic Generalized Zero-Shot Learning (GZSL) setting includes both seen and unseen classes at inference time, often leading to a strong bias toward seen categories (Zhang and Zhang, 2024) [22]. Core ZSL methods project data into a shared semantic space via visual-to-semantic, semantic-to-visual, or joint embedding strategies, with foundational models like Semantic Similarity Embedding (SSE) representing unseen classes as mixtures of seen class histograms (Zhang and Saligrama, 2015) [27]. However, challenges such as the hubness problem, domain shift and noisy or incomplete semantic descriptors limit ZSL performance, especially in fine-grained tasks (Zhang et al., 2017; Rezaei and Shahidi, 2020; Kim et al., 2022) [28] [29] [30] . To overcome these issues, recent approaches employ vision-language models like CLIP to extract rich semantic embeddings (Radford et al., 2021) [20] and integrate multi-source semantic fusion mechanisms (Yang et al., 2025) [31]. Additionally, generative models like f-VAEGAN synthesize visual features for unseen classes using semantic input, with semantic feedback modules incorporating

contrastive and consistency-based learning to enhance feature realism and class separability (Zhang and Zhang, 2024) [22].

B. Transformer Architectures in Vision and Multimodal Models

Transformer-based models have revolutionized visual learning by replacing convolutions with self-attention, starting with the Vision Transformer (ViT) (Dosovitskiy et al., 2020) [23], which processes images as token sequences. To address its data efficiency limits, variants such as as T2T-ViT (Yuan et al., 2021) [24] and DINOv2 (Oquab et al., 2024) [25] introduced hierarchical token modeling and self-supervised dense patch-level training, respectively. These advances improved representation quality and generalization, critical for zero-shot tasks. For efficient adaptation, LoRA (Hu et al., 2021) [32] enabled lightweight fine-tuning through low-rank updates. Transformers also have advanced vision language pretraining, where models such as CLIP (Radford et al., 2021) [20] use contrastive learning on image-text pairs to align modalities in a shared embedding space. LiT (Zhai et al., 2023) [21] reduces the computation by tuning only the text encoder, while SigLIP and SigLIP2 (Tschannen et al., 2025) [33] adopt sigmoid loss to enhance semantic alignment and avoid softmax competition. Joint models like PaLI-3 (Chen et al., 2023) [34] further unify image and text processing in a shared transformer, enabling strong zero-shot performance with fewer parameters. Prompt-tuning approaches, such as CoOp (Zhou et al., 2022) [35], optimize CLIP's context vectors, enhancing performance with minimal supervision. These models collectively enable scalable, generalizable and multimodal zero-shot learning, particularly useful in fine-grained domains like vehicle recognition.

C. Loss functions in Zero-Shot Learning (ZSL) and Vision-Language Models (VLMs)

Loss functions are central to aligning visual and semantic modalities in zero-shot and vision-language models. The widely used contrastive loss, introduced in CLIP (Radford et al., 2021) [20], maximizes the similarity between matching image-text pairs while pushing apart mismatched ones using softmax normalization. While effective, this approach introduces competition between pairs, which can hinder dense alignment and generalization. To address this, SigLIP and SigLIP2 (Zhai et al., 2023; Tschannen et al., 2025) [33] [36] propose a sigmoid-based contrastive loss that treats each image-text pair independently, improving semantic consistency, localization, and robustness key for tasks like fine-grained vehicle recognition. In generative ZSL, models like f-VAEGAN (Zhang and Zhang, 2024) [22] synthesize unseen class features using semantic embeddings. Their loss combines variational, adversarial, and semantic feedback terms to ensure discriminative and semantically coherent outputs. Meanwhile, prompt tuning methods like CoOp [35]learn task-specific context vectors using cross-entropy loss, allowing CLIP to adapt efficiently with minimal supervision. Together, these diverse loss formulations drive the generalization ability and flexibility of modern zero-shot models.

IV. A Comprehensive VMMR Pipeline: Proposed Methodology

The proposed approach integrates data hygiene, prompt engineering, lightweight adaptation, and cascade-style inference, grounded in recent advances in vision–language models and fine-grained recognition literature.

A. Multi-Source Data Consolidation and Duplicate Removal

We merge **CompCars**, **VeRi-776**, **VehicleID**, **VERI-Wild**, and a custom dataset into a unified collection using a common ontology: *brand*, *model*, *body type*, and *year range*. To avoid inflated performance due to data redundancy, we remove near-duplicate images using **perceptual hashing (pHash)** with a *Hamming distance threshold of* ≤ 4, following best practices to preserve visual diversity without redundancy.

The **custom dataset** was created by crawling vehicle images from online sources. Annotation was fully automated using a *two-stage LLM-based pipeline*: first, a primary LLM (e.g., GPT-4) generated structured labels (make, model, year) and textual descriptions; second, a secondary LLM verified the annotations using consistency checks and confidence scoring. Such dual-LLM pipelines have been shown to match or even exceed human level annotation in several domains [37].

B. Attribute-Rich Prompt Generation

For each class, we generate **three descriptive prompts** using GPT-4o, incorporating attributes such as color, style, era, and distinctive visual traits. For example:

> *"a 2012 lime-green hatchback Mazda2 with swept-back head-lamps"*

Inspired by prompt-augmentation strategies in transformer-based classifiers, these prompts are carefully tuned to vehicle-specific visual semantics, enriching vision-language alignment.

C. Hybrid Training with LoRA and InfoNCE

LoRA-Based Fine-Tuning: We insert **Low-Rank Adaptation (LoRA)** modules into the final four vision transformer blocks of the **PaLI-Gemma-3B** model, yielding an efficient adaptation path with just **9M trainable parameters** [**?**].

Supervised CLIP Loss: We use a **CLIP loss** to align labelled vehicle images with their corresponding prompts, facilitating strong visual-text representation learning.

Self-Supervised Multi-View InfoNCE Regularization: We also incorporate **unlabelled Re-ID vehicle images** through a **multi-view InfoNCE loss**, encouraging viewpoint invariance without requiring explicit supervision [**?**]. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}} + \frac{1}{2}\,\mathcal{L}_{\text{InfoNCE}} \tag{1}$$

This hybrid objective enables training that maintains fine-grained accuracy while leveraging unlabelled data and avoids catastrophic forgetting common in full fine tuning.

D. Two-Stage Cascade Inference

Stage 1: DenseNet-201 Make Classifier: We use a **DenseNet-201** model trained on make classes to serve as a coarse filter, reducing inter-brand confusion in a *2-Branch 2-Stage (2B–2S)* strategy [10].

Stage 2: Prompt-Based Model Recogniser: From the top-K predicted brands ($K = 5$), the input image is compared to all three prompts of each candidate model via cosine similarity. The model with the highest average similarity score is selected. This strategy leverages prompt diversity.

E. Cross-Modal Explainability

We integrate **Grad-CAM++** saliency from the vision model with token-level prompt attention to provide interpretable, cross-modal justifications. This hybrid explanation enhances transparency compared to visual only saliency [38].

REFERENCES

[1] R. S. Feris et al., "Large-scale vehicle detection, indexing, and search in urban surveillance videos," IEEE Trans. Multimedia, vol. 14, no. 1, pp. 28–42, Feb. 2012, doi: 10.1109/TMM.2011.2170666.

[2] R. S. Feris et al., "Large-scale vehicle detection, indexing, and search in urban surveillance videos," IEEE Trans. Multimedia, vol. 14, no. 1, pp. 28–42, Feb. 2012, doi: 10.1109/TMM.2011.2170666.

[3] M. Papageorgiou et al., "ITS and traffic management," in Handbooks in Operations Research and Management Science. Amsterdam, The Netherlands: North-Holland, 2007, pp. 715–774.

[4] T. Celik and H. Kusetogullari, "Solar-powered automated road surveillance system for speed violation detection," IEEE Trans. Ind. Electron., vol. 57, no. 9, pp. 3216–3227, Sep. 2010, doi: 10.1109/TIE.2009.2038395.

[5] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 6, pp. 1257–1272, Nov. 2012, doi: 10.1109/TSMCC.2012.2215319.

[6] C. T. Barba, M. A. Mateos, P. R. Soto, A. M. Mezher, and M. A. Igartua, "Smart city for VANETs using warning messages, traffic statistics and intelligent traffic lights," in Proc. IEEE Intell. Vehicles Symp., Jun. 2012, pp. 902–907.

[7] M. Biglari, A. Soleimani, and H. Hassanpour, "A cascaded part-based system for fine-grained vehicle classification," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 273–283, Jan. 2018, doi: 10.1109/TITS.2017.2749961.

[8] X. Ma and A. Boukerche, "An AI-based visual attention model for vehicle make and model recognition," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2020, pp. 1–6.

[9] A. Amirkhani and A. H. Barshooi, "DeepCar 5.0: Vehicle Make and Model Recognition Under Challenging Conditions," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 1, pp. 541-553, Jan. 2023, doi: 10.1109/TITS.2022.3212921.

[10] Lyu, Yangxintong Schiopu, Ionut Cornelis, Bruno Munteanu, Adrian. (2022). Framework for Vehicle Make and Model Recognition—A New Large-Scale Dataset and an Efficient Two-Branch–Two-Stage Deep Learning Architecture.

[11] Lu, Lei Cai, Yancheng Huang, Hua Wang, Ping. (2023). An efficient fine-grained vehicle recognition method based on part-level feature optimization.

[12] Jamil, A.A.; Hussain, F.; Yousaf, M.H.; Butt, A.M.; Velastin, S.A. Vehicle Make and Model Recognition using Bag of Expressions. Sensors 2020, 20, 1033. https://doi.org/10.3390/s20041033.

[13] Taki, Youssef. (2023). Vehicle Image Classification Method Using Vision Transformer.

[14] Nafzi, Mohamed Brauckmann, Michael Glasmachers, Tobias. (2020). Data Augmentation and Clustering for Vehicle Make/Model Classification.

[15] L. Kezebou, V. Oludare, K. Panetta and S. Agaian, "Few-Shots Learning for Fine-Grained Vehicle Model Recognition," 2021 IEEE International Symposium on Technologies for Homeland Security (HST), Boston, MA, USA, 2021, pp. 1-9, doi:10.1109/HST53381.2021.9619823.

[16] Vázquez-Santiago, Diana-Itzel Acosta-Mesa, Héctor Mezura-Montes, Efrén. (2023). Vehicle Make and Model Recognition as an Open-Set Recognition Problem and New Class Discovery.

[17] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[18] Dong, Xinlong Shi, Peicheng Tang, Yueyue Yang, Li Yang, Aixi Liang, Taonian. (2024). Vehicle Classification Algorithm Based on Improved Vision Transformer.

[19] N. Semiromizadeh, O. N. Manzari, S. B. Shokouhi and S. Mirzakuchaki, "Enhancing Vehicle Make and Model Recognition with 3D Attention Modules," 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, Islamic Republic of, 2024, pp. 087-092, doi: 10.1109/ICCKE65377.2024.10874749.

[20] Radford, Alec, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, arXiv, 26 Feb. 2021, https://doi.org/10.48550/arXiv.2103.00020.

[21] X. Zhai et al., "LiT: Zero-Shot Transfer with Locked-image text Tuning," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 18102-18112, doi: 10.1109/CVPR52688.2022.01759.

[22] L. Zhang and G. Zhang, "Semantic Feedback for Generalized Zero-Shot Learning," in 2024 4th International Conference on Neural Networks, Information and Communication (NNICE), Guangzhou, China: IEEE, Jan. 2024, pp. 298–302. doi: 10.1109/NNICE61279.2024.10499130.

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., 38; Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[24] Yuan, Li Chen, Yunpeng Wang, Tao Yu, Weihao Shi, Yujun Jiang, Zihang Tay, Francis Feng, Jiashi Yan, Shuicheng. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. 538-547. 10.1109/ICCV48922.2021.00060.

[25] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," Feb. 02, 2024, arXiv: arXiv:2304.07193. doi: 10.48550/arXiv.2304.07193.

[26] Zhou, Kaiyang Yang, Jingkang Loy, Chen Change Liu, Ziwei. (2022). Conditional Prompt Learning for Vision-Language Models. 16795-16804.10.1109/CVPR52688.2022.01631.

[27] Z. Zhang and V. Saligrama, "Zero-Shot Learning via Semantic Similarity Embedding," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile: IEEE, 2015, pp. 4166–4174. doi: 10.1109/ICCV.2015.474.

[28] L. Zhang, T. Xiang, and S. Gong, "Learning a Deep Embedding Model for Zero-Shot Learning," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, 2017.

[29] M. Rezaei and M. Shahidi, "Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review," Intelligence-Based Medicine, vol. 3–4, p. 100005, 2020.

[30] Kim, K. Shim, and B. Shim, "Semantic Feature Extraction for Generalized Zero-Shot Learning," AAAI, vol. 36, no. 1, pp. 1166–1173, Jun. 2022.

[31] G. Yang, W. Sun, X. Liu, Y. Liu, and C. Wang, "Semantic Fusion and Contrastive Generation for Generalized Zero-Shot Learning," Int J Multimed Info Retr, vol. 14, no. 3, 2025, doi: 10.1007/s13735-025-00372-w.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in International Conference on Learning Representations (ICLR), 2022.

[33] M. Tschannen et al., "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," Feb. 20, 2025, arXiv: arXiv:2502.14786. doi: 10.48550/arXiv.2502.14786.

[34] X. Chen et al., "PaLI-3 Vision Language Models: Smaller, Faster, Stronger," Oct. 17, 2023, arXiv: arXiv:2310.09199. doi: 10.48550/arXiv.2310.09199.

[35] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," Int J Comput Vis, vol. 130, no. 9, pp. 2337–2348, 2022, doi: 10.1007/s11263-022-01653-1.

[36] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023. doi: 10.1109/iccv51070.2023.01100.

[37] Horych, Tomáš Mandl, Christoph Ruas, Terry Greiner-Petter, André Gipp, Bela Aizawa, Akiko Spinde, Timo. (2024). The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection. 10.48550/arXiv.2411.11081.

[38] Zhang, Xiaofeng Shen, Chen Yuan, Xiaosong Yan, Shaotian Xie, Liang Wang, Wenxiao Gu, Chaochen Tang, Hao Ye, Jieping. (2024). From Redundancy to Relevance: Enhancing Explainability in Multimodal Large Language Models. 10.48550/arXiv.2406.06579.