Structural Information Guided Multimodal Pre-training for Vehicle-Centric Perception

Xiao Wang[1] , Wentao Wu[1] , Chenglong Li[1] , Zhicheng Zhao[1] , Zhe Chen[2] , Yukai shi[3] , and Jin Tang[1]

[1] Anhui University, Hefei City, 230601, Anhui Province, China
[2] La Trobe University [3] Guangdong University of Technology

https://github.com/Event-AHU

Scan for Source Code !!!

Introduction

Understanding vehicles in images is important for various applications such as intelligent transportation and self-driving system. Existing vehicle-centric works typically pre-train models on large-scale classification datasets and then finetune them for specific downstream tasks. However, they neglect the specific characteristics of vehicle perception in different tasks and might thus lead to sub-optimal performance. To address this issue, we propose a novel vehicle-centric pretraining framework called VehicleMAE, which incorporates the structural and semantic informative for effective masked vehicle appearance reconstruction. To be specific, we explicitly extract the sketch lines of vehicles as a form of the spatial structure to guide vehicle reconstruction. The more comprehensive knowledge distilled from the CLIP big model based on the similarity between the paired/unpaired vehicle image-text sample is further taken into consideration to help achieve a better understanding of vehicles. A large-scale dataset is built to pre-train our model, termed Autobot1M, which contains about 1M vehicle images and 12693 text information. Extensive experiments on four vehicle-based downstream tasks fully validated the effectiveness of our VehicleMAE.

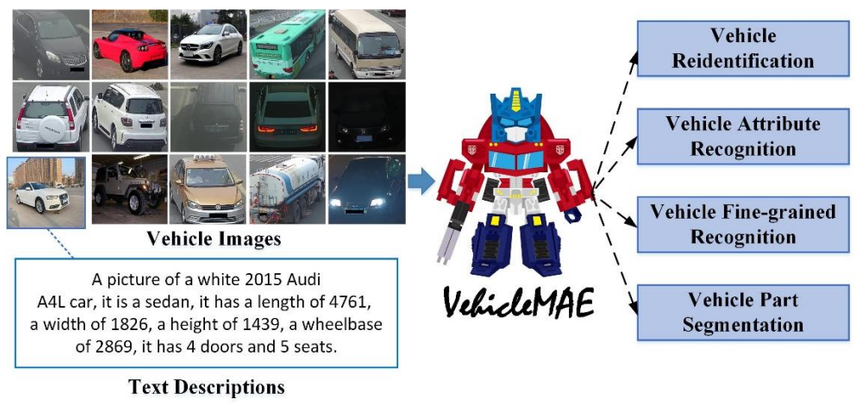

Figure 1: Our proposed pre-trained big model VehicleMAE takes the large-scale vehicle images and corresponding natural language descriptions as input and supports multiple downstream vehicle-based tasks

Method

We propose a general vehicle-centric pre-training framework that considers both vision and language description for MAE-based vehicle perception. As shown in Fig. 2, our proposed VehicleMAE contains three main modules, including the MAE, Structural Prior module, and Semantic Prior module. Specifically, we partition the given input vehicle image into non-overlapping tokens, then, mask these tokens with a high ratio. The visible tokens are fed into the Transformer encoder for feature learning, then, the masked tokens are randomly initialized and fed into the Transformer decoder for invisible token prediction. More importantly, we propose the structural prior to guide the image reconstruction procedure using extracted vehicle sketch maps. The CLIP model is also introduced to learn the semantic-aware representations using contrastive learning.

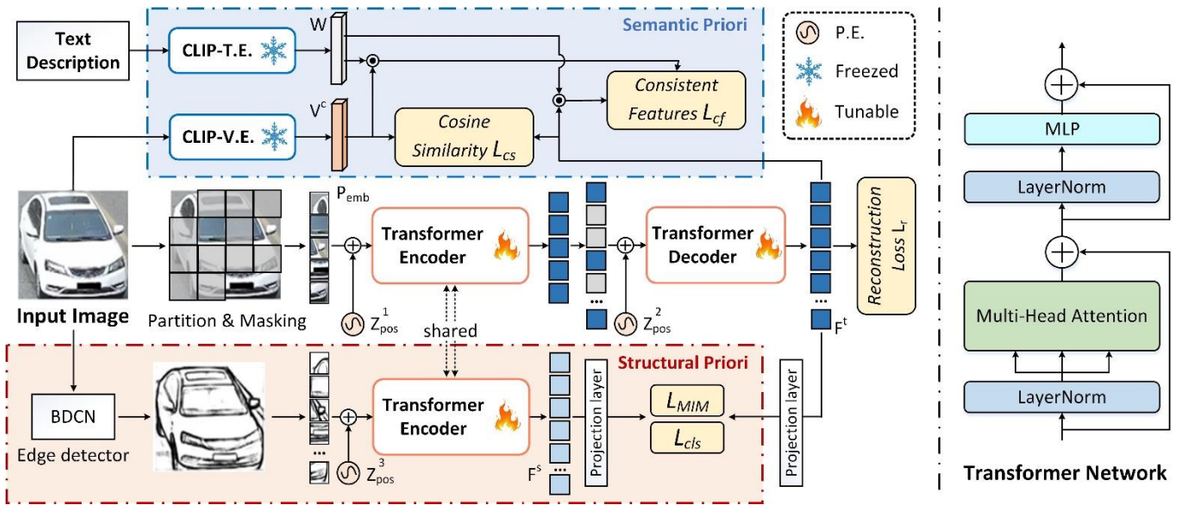

Figure 2: An overview of our proposed Structural and Semantic Prior Guided Masked Auto-Encoder Framework for General Vehicle-centric Perception, termed VehicleMAE.

Some details about the network architecture are listing below.

Masked Auto-Encoder. We partition the input vehicle image into non-overlapping patches. Most of these tokens are randomly masked, and the remaining ones are used as input to the network. In this paper, we adopt the Transformer encoder and Transformer decoder to achieve the masked token reconstruction.

Structural Prior Module. We adopt the BDCN edge detector to get the contour map and partition it into non-overlapping regions. The contour representation can be obtained using the shared Transformer encoder, which is an important clue for efficient vehicle reconstruction.

Semantic Prior Module. We adopt the off-the-shelf pre-trained vision-language models CLIP to encode the natural language descriptions and build contrastive learning schemes for high-level semantic information guided reconstruction. More in detail, the cosine similarity between the CLIP visual embedding and Transformer encoder, and the KL-Distance between the similarity of language embeddings and visual features predicted by the CLIP model and learned Transformer encoder are considered.

Experiment

In this work, the proposed VehicleMAE big model is pretrained on our newly proposed Autobot1M dataset. Then, we validate its effectiveness and generalization on three datasets corresponding to four downstream tasks.

Method	Dataset	VAR mA	VAR Accuracy	VAR Precision	VAR Recall	VAR F1	V-Reid mAP	V-Reid R1	VFR Accuracy	VPS mIou	VPS mAcc
Scratch	-	84.67	80.86	84.66	85.77	84.90	35.3	57.3	24.8	49.36	59.22
MoCov3	ImagNet-1K	90.38	93.88	95.57	95.48	95.33	75.5	94.4	91.3	73.17	78.60
DINO	ImagNet-1K	89.92	91.09	92.84	93.60	93.11	64.3	91.5	-	68.43	73.37
IBOT	ImagNet-1K	89.51	90.17	91.95	93.03	92.37	68.9	92.6	81.1	66.03	71.06
MAE	ImagNet-1K	89.69	93.60	94.81	95.54	95.08	76.7	95.8	91.2	69.54	75.36
MAE	Autobot1M	90.19	94.06	95.45	95.68	95.43	75.5	95.4	91.3	69.00	75.36
VehicleMAE	Autobot1M	92.21	94.91	96.00	96.50	96.17	85.6	97.9	94.5	73.29	80.22

Table 1: Experimental results of ours and other pre-trained models on vehicle attribute recognition (VAR), re-identification (V-Reid), fine-grained recognition (VFR), and partial segmentation (VPS).

Training Data	Method	Dataset	VAR mA	VAR Acc	VAR Prec	VAR Rec	VAR F1	V-ReID mAP	V-ReID R1	VFR Acc	VPS mIoU	VPS mAcc
20%	Scratch	-	80.94	71.33	76.18	79.64	77.27	25.2	34.9	7.1	39.87	49.50
	MAE	ImageNet-1K	89.32	92.65	94.35	94.87	94.41	64.8	89.7	42.5	64.86	72.20
	MAE	Autobot1M	89.58	92.36	94.09	95.06	94.29	60.0	85.5	66.5	65.04	70.81
	VehicleMAE	Autobot1M	91.50	94.53	95.74	96.33	95.91	80.9	95.2	83.58	68.72	76.02
10%	Scratch	-	78.47	66.48	72.35	75.47	73.25	-	-	4.5	35.44	46.22
	MAE	ImageNet-1K	88.61	90.78	92.74	93.64	92.95	-	-	17.1	52.31	62.30
	MAE	Autobot1M	86.49	89.59	91.61	93.33	92.13	-	-	21.4	62.35	69.56
	VehicleMAE	Autobot1M	89.29	93.76	94.94	95.86	95.25	-	-	71.4	65.09	71.19

Table 2: Results of different scales of training data used in downstream tasks. - denotes no corresponding results.

As shown in Fig. 3, we visualize the feature maps of MAE and our VehicleMAE big model, the masked tokens, and reconstructed images.

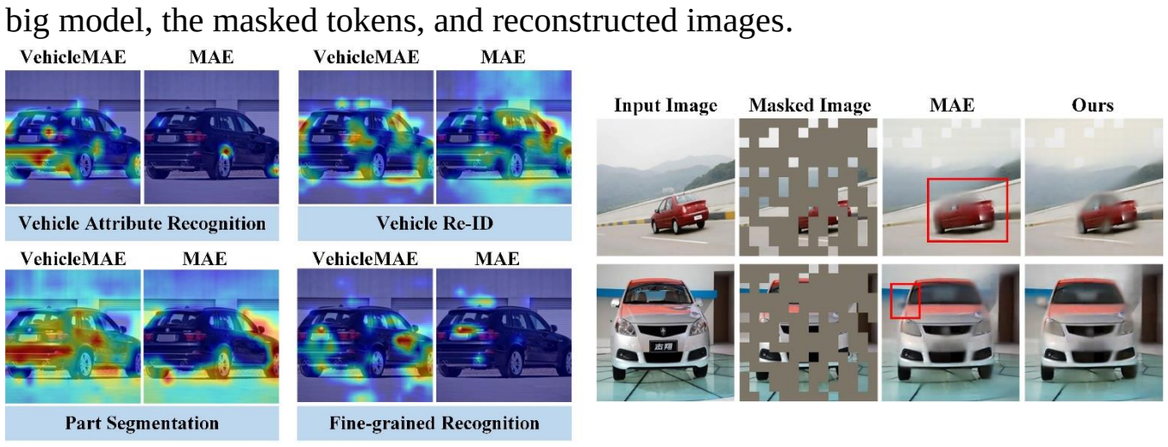

Figure 3: Visualization of attentions and reconstructed vehicle images.

To help researchers better understand the method we proposed, we conduct comprehensive experiments of component analysis on three datasets corresponding to four downstream tasks to check their influence on the overall model.

MAE Loss L_r	Structural Prior L_{mim}	Structural Prior L_{cls}	Semantic Prior L_{cs}	Semantic Prior L_{cf}	VAR mA	VAR Acc	VAR Prec	VAR Rec	VAR F1	V-ReID mAP	V-ReID R1	VFR Acc	VPS mIou	VPS mAcc
✓					90.19	94.06	95.45	95.68	95.43	75.5	95.4	91.3	69.00	75.36
✓	✓				91.27	94.11	95.29	95.82	95.50	79.7	96.1	93.2	70.34	75.70
✓	✓	✓			91.71	94.54	95.65	96.28	95.88	83.4	96.6	93.7	70.65	76.04
✓			✓		92.12	94.28	95.42	96.23	95.71	84.1	97.1	94.1	71.90	76.47
✓			✓	✓	92.15	94.58	95.69	96.36	95.92	85.2	97.1	94.3	71.87	77.93
✓	✓	✓	✓	✓	92.21	94.91	96.00	96.50	96.17	85.6	97.9	94.5	73.29	80.22

Table 3: Ablation study on loss functions in Structural Prior and Semantic Prior.

Conclusion

In this paper, we propose the first large-scale pre-trained big model for vehicle-centric perceptron, termed VehicleMAE. Given the vehicle image, we first divide and partition it into non-overlapping patches. Then, we randomly mask these patches with a high ratio and project the rest tokens into feature embeddings. The ViT network is adopted as the backbone to process these embeddings, then, masked tokens are padded for reconstruction using a Transformer decoder network. More importantly, the vehicle profile information and high-level natural language descriptions are taken into consideration for effective masked vehicle reconstruction. To bridge the data gap, we propose a large-scale dataset to pre-train our model termed Autobot1M, which contains about 1M vehicle images and 12693 text information. Extensive experiments fully validated the effectiveness and benefits of our VehicleMAE and Autobot1M dataset.

Reference

1. Bao, H. Dong, L. Piao, S. and Wei, F." **BEiT: BERT Pre-Training of Image Transformers.**" NAACL-HLT-2019.
2. K. He, X. Zhang, S. Ren, and J. Sun, R. "**Masked autoencoders are scalable vision learners.** " CVPR 2022.
3. Ramesh, A. Pavlov, M. Goh, G. Gray, S. Voss, C. Radford, A. Chen, M. and Sutskever, I. "**Zero-shot text-to-image generation.** " ICML-2021.