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

Xiao Wang[1,2,3], **Wentao Wu**[1,2,4], **Chenglong Li**[1,2,4 *], **Zhicheng Zhao**[1,2,4], **Zhe Chen**[5], **Yukai Shi**[6], **Jin Tang**[1,2,3]

[1] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei 230601, China
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China
[3] School of Computer Science and Technology, Anhui University, Hefei 230601, China
[4] School of Artificial Intelligence, Anhui University, Hefei 230601, China
[5] School of Computing, Engineering and Mathematical Sciences, La Trobe University
[6] School of Information Engineering, Guangdong University of Technology, Guangzhou, China
{wangxiaocvpr, lcl1314}@foxmail.com, wuwentao0708@163.com, {zhaozhicheng, tangjin}@ahu.edu.cn, zhechen91@gmail.com, ykshi@gdut.edu.cn

Abstract

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 fine-tune 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 pre-training framework called VehicleMAE, which incorporates the structural information including the spatial structure from vehicle profile information and the semantic structure from informative high-level natural language descriptions 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. The source code and pre-trained models will be released at https://github.com/Event-AHU/VehicleMAE.

Introduction

Vehicles play a very important role in modern real life, such as the private car, public transport bus, trucks, etc. With the development of artificial intelligence, the problem of vehicle-centered perception has attracted more and more attention, especially for autonomous driving, security monitoring, and smart city. Many computer vision problems are proposed for the vehicles, including vehicle detection and tracking (Chadwick, Maddern, and Newman 2019; Wang et al. 2022a), segmentation (He et al. 2022a), attribute recognition (Liu et al. 2016; Wang et al. 2022b), re-identification (Wang et al. 2020), fine-grained classification (Krause et al. 2013), and text-based retrieval (Scribano et al. 2021). In the early stages of deep learning, these research topics are usually studied in a relatively independent form. To be more specific, they first collect and annotate a small subset, then, train a neural network from scratch or based on a pre-trained backbone network. Then, they conduct the evaluation on the testing subset. Although good performance can be achieved compared with previous ones, however, these problems are far from being solved due to the complicity of the real world.

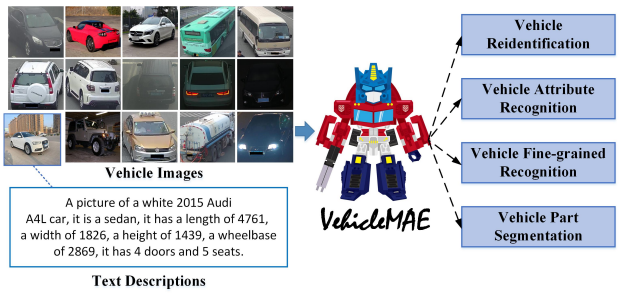

Figure 1: Our proposed pre-trained big model VehicleMAE takes the large-scale vehicle images and corresponding natural language descriptions as input and supports multiple downstream vehicle-based tasks.

Recently, the self-attention-based Transformer (Vaswani et al. 2017) networks illustrate their powerful ability on capturing long-range relations than the widely used Convolutional Neural Networks (CNN). Also, the big models obtained by stacking the Transformer blocks and pre-training on large-scale datasets demonstrate their amazing performance (Wang et al. 2023). Compared with the traditional small-scale deep neural network, these large Transformer models have achieved obvious advantages in accuracy and

*Corresponding author: Chenglong Li

generalization. To be specific, Large Language Models (short for LLMs, like the GPT series (Radford et al. 2018, 2019; Brown et al. 2020; OpenAI 2023), LLaMA (Touvron et al. 2023), T5 (Raffel et al. 2020)) have demonstrated their powerful performance on various text tasks. They can recognize, summarize, translate, predict, and generate text according to the knowledge learned from massive datasets. In addition to the large language models, the pre-training technique also showed superior performance in computer vision. Some pioneer models like the ViT (Dosovitskiy et al. 2020) and Swin-Transformer (Liu et al. 2021) achieved compelling performance in the image classification task. These pre-trained models are also widely used in other computer vision tasks, such as object detection, tracking, segmentation, etc. The deep generative models are also significantly boosted with the large model and training data, for example, DALLE (Ramesh et al. 2021), GPT-4 (OpenAI 2023), SparkDesk, and ERNIE-Bot. In order to obtain a more comprehensive and accurate model, the researchers pre-train their model on multimodal data (Wang et al. 2023), such as RGB, language, audio, depth, thermal, event stream, etc. Many multimodal large models demonstrate that different modalities can achieve the effect of information complementary in the training stage. In other words, it performs better than the unimodal-based model even under modality-missing scenarios in the inference phase.

Inspired by the success of the aforementioned large models, in this work, we begin to investigate the vehicle-based pre-training algorithms based on MAE (Masked Auto-Encoder) (He et al. 2022b) for high-performance and generalized vehicle perception. The vanilla MAE takes the high-ratio masked tokens as input and learns to reconstruct them under an auto-encoder framework. Directly adapting the MAE for our vehicle perception is a natural and intuitive approach, however, we believe that it could be difficult for the original MAE to focus on the vehicle representation without specific designs, and this framework can be further improved from the following two aspects: **Firstly**, vehicles typically possess certain distinctive visual features, such as their outlines which are often formed by lines and curves, and their consistent coloration. These features are distinct from those of regular objects. We believe that these features can be leveraged during the pre-training phase of the model to enhance its ability to perceive the visual structure and form of vehicles more effectively. **Secondly**, the natural language descriptions of the vehicles can easily be obtained from various vehicle websites. Therefore, from a multimodal pre-training perspective, it is possible to better explore the high-level semantic guidance to improve vehicle reconstruction. Also, existing multimodal big models, such as CLIP (Radford et al. 2021), can be fully utilized to further enhance the reconstruction results.

Based on the above observations and reflections, in this paper, 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. To be specific, given the input vehicle image, we first partition it into non-overlapping regions. Following the MAE (He et al. 2022b), we mask most of the input tokens and feed them into a Transformer encoder for feature representation learning. Then, a Transformer decoder is used for masked token prediction. Meanwhile, we also adopt the pre-trained CLIP visual encoder to get the visual representations and the edge detector BDCN network (He et al. 2020) to get the contour image. For the text input, we adopt the CLIP text encoder to obtain the text embeddings and tune the parameters of our neural network to make the predicted similarity between image-text input consistent with the CLIP model. Note that, the parameters of CLIP are fixed and only the Transformer encoder and decoder are adjustable. For the structural prior module, we first partition the contour visual image into non-overlapping patches and fed it into the shared Transformer encoder for contour information extraction. This information can be seen as a guide for the MAE branch for better image reconstruction.

To train our proposed VehicleMAE model, we collect a large-scale dataset that contains about 1M vehicle images, termed **Autobot1M**. These data are mainly obtained from existing datasets, public visual surveillance systems, and vehicle websites. These data fully reflect the key challenges in vehicle-centric perception, such as illumination, motion blur, viewpoints, and occlusion. Note that, part of these images are crawled from the Internet and the corresponding natural language descriptions (12693 sentences) are also available for pre-training. More details can be found in Section and some representative samples of our data can be found in our supplementary materials.

To sum up, the contributions of this paper can be concluded as the following three aspects:

- We propose the first multimodal pre-training framework for vehicle-centric perception, termed **VehicleMAE**. The structural contour information and high-level semantic prior are proposed for a more accurate masked token reconstruction.
- We propose a large-scale dataset to boost the research of pre-training on vehicle images, termed **Autobot1M**. It contains a total of 1M images and part of them with a corresponding language description.
- We conduct extensive experiments on four downstream tasks to validate the effectiveness of our proposed VehicleMAE, including Vehicle Attribute Recognition, Vehicle-based Re-identification, Fine-grained Vehicle Classification, and Vehicle Detection.

Methodology

Overview As shown in Fig. 2, our proposed VehicleMAE follows the masked auto-encoder framework which takes the high-ratio masked tokens as the input. Specifically speaking, we first 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. More importantly, we introduce two kinds of prior information to guide the token reconstruction, i.e., the structural and the semantic prior. For the structural prior,

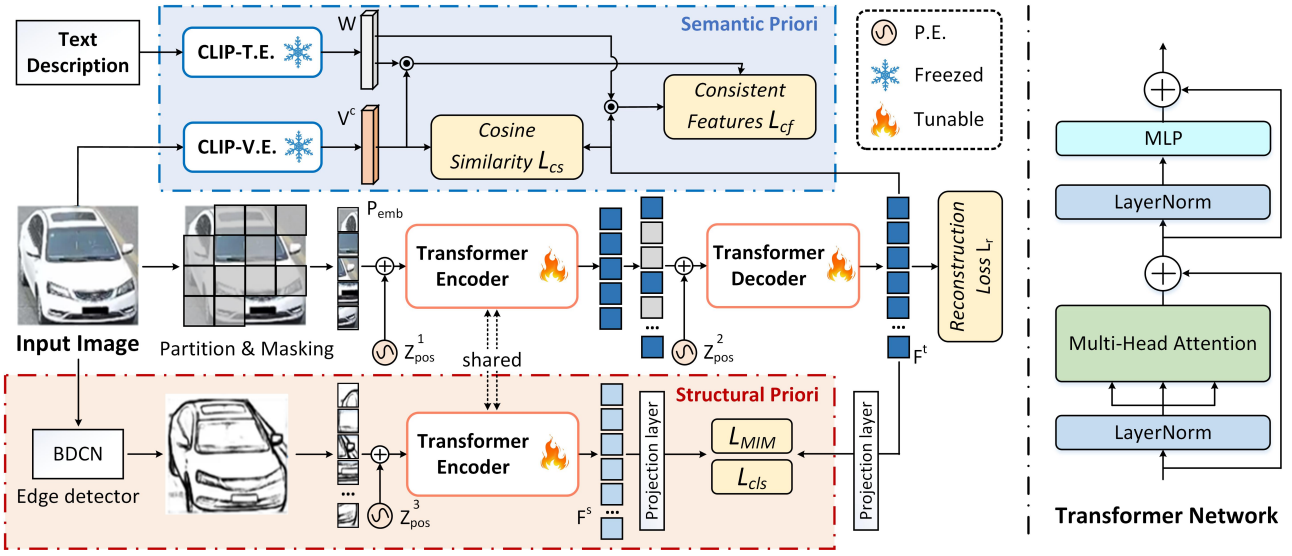

Figure 2: An overview of our proposed Structural and Semantic Prior Guided Masked Auto-Encoder Framework for General Vehicle-centric Perception, termed VehicleMAE.

we adopt the BDCN (He et al. 2020) 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. For the semantic prior, we adopt the off-the-shelf pre-trained vision-language models (CLIP (Radford et al. 2021) is adopted in this work) 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 (Radford et al. 2021) visual embedding and Transformer encoder, and the KL-Distance between the similarity of language embeddings and visual features predicted by the CLIP (Radford et al. 2021) model and learned Transformer encoder are considered. Extensive experiments on multiple downstream tasks demonstrate that the semantic and structural prior improves the pre-training significantly.

Network Architecture

Our proposed VehicleMAE contains three main modules, i.e., the masked auto-encoder, structural prior, and semantic prior module.

Masked Auto-Encoder Given the vehicle image $I \in \mathbb{R}^{224\times224\times3}$, we first partition it into 196 non-overlapping patches $P_i \in \mathbb{R}^{16\times16\times3}, i \in \{1, 2, ..., 196\}$. Following the Masked Auto-Encoder (MAE) (He et al. 2022b), we randomly mask 75% of these tokens and only feed the resting 25% into the following networks. A convolution layer with three kernels 16×16 is used to project the image patches P_i into the token embeddings $P^j_{emb} \in \mathbb{R}^{1\times768}, j \in \{1, 2, ..., 49\}$. The CLS-token is also integrated, therefore, we have the input token embeddings $P_{emb} \in \mathbb{R}^{50\times768}$. Meanwhile, we introduce the position encoding $Z^1_{pos} \in \mathbb{R}^{50\times768}$ to encode the spatial coordinates of input tokens. Similar to existing works, we random initialize the position encoding and add it with token embeddings, therefore, we have $\tilde{P}_{emb} = Z^1_{pos} + P_{emb}$.

After the input embedding $\tilde{P}_{emb} \in \mathbb{R}^{50\times768}$ is obtained, we feed them into ViT-B/16 (Dosovitskiy et al. 2020) encoder which contains 12 Transformer blocks. Each Transformer block consists of layer normalization, multi-head self-attention (MSA), and Multi-Layer Perceptron (MLP), as shown in the right part of Fig. 2. The output $\bar{P} \in \mathbb{R}^{50\times768}$ from the Transformer encoder is the same as the input tokens. After a 512-dimensional linear projection layer is used to project the output $\bar{P} \in \mathbb{R}^{50\times768}$ from Transformer encoder into the decode embedding $\bar{P}^k_{emb} \in \mathbb{R}^{1\times512}, k \in \{1, 2, ..., 50\}$. Two kinds of input tokens are fed into the Transformer decoder, i.e., the mask tokens $P_{mask} \in \mathbb{R}^{147\times512}$ and the encoded visible tokens $\bar{P}_{emb} \in \mathbb{R}^{50\times512}$. Note that, the mask tokens (Kenton and Toutanova 2019) are shared and learnable vectors. Position encoding $Z^2_{pos} \in \mathbb{R}^{197\times512}$ is also introduced and combined with input tokens. The decoder network contains 8 Transformer blocks and is only used in the pre-training phase for image reconstruction.

The mean square error (MSE) over pixel space from the masked token in the original image and the reconstructed token is adopted as the reconstruction loss to optimize the MAE module, which can be written as:

$$L_r = \frac{1}{N_m} \sum_{t \in P_m} ||V_t - V^r_t||_2 \quad (1)$$

where V is the RGB pixel value of the input image, and V^r is the predicted pixel value. N_m is the number of masked pixels, P_m is the index of pixel of the mask, $|| * ||_2$ refers to the L_2 loss.

Structural Prior Module The aforementioned MAE can already achieve good performance on vehicle-based perception, however, we think this model can be further improved

based on the design of auxiliary tasks. The structural prior guided image reconstruction is the first auxiliary task we proposed in this work, as shown in Fig. 2. The core motivation of this module is that we observed vehicles to be distinct from typical objects, possessing significant contour information such as horizontal and vertical lines, curves, and so on. This structural information about vehicles will play a crucial role in effectively enhancing vehicle image reconstruction.

In our practical implementation, we first adopt the edge detector BDCN (Bi-Directional Cascade Network) (He et al. 2020) to obtain the outline of the vehicle. Then, similar to the operations conducted on the input vehicle image, the skeleton map is also partitioned into non-overlapping patches, and each patch is projected into a token whose scale is 196×768 using a convolution layer. Differently, we don't conduct masking operations on the skeleton map.The CLS-token is also integrated, accordingly, we initialize the position encoding $Z^3_{pos} \in \mathbb{R}^{197\times768}$ and add it with the tokens as the input of the Transformer encoder. Note that this encoder is shared with the Transformer encoder used for vehicle image encoding. The skeleton feature vectors are treated as a guide for vehicle reconstruction.

To achieve structural priori-guided masked token reconstruction, we adopt the widely used knowledge distill (Bao et al. 2021) scheme. We project the skeleton tokens and reconstructed invisible tokens into probability distributions with K dimensions (i.e., $P^{patch}_{\theta'}(F^s_i)$ and $P^{patch}_{\theta}(F^t_i)$ in Eq. 2, respectively) using two separate projection layers whose parameters are θ' and θ. F^s is the skeleton feature output by the structural prior model. F^t is the feature corresponding to the reconstructed invisible token. The distill loss function can be formally written as:

$$L_{mim} = -\sum_{i=1}^{N} P^{patch}_{\theta'}(F^s_i)^T log P^{patch}_{\theta}(F^t_i) \tag{2}$$

where N is the number of masked patches in the encoding phase. In order to obtain better visual semantic information, we also project the skeleton CLS token and reconstructed CLS token to obtain their respective classification distributions. The distill loss function can be formally written as:

$$L_{cls} = -P^{cls}_{\theta'}(F^s)^T log P^{cls}_{\theta}(F^t) \tag{3}$$

Semantic Prior Module The MAE focuses on reconstructing the vehicle image in an auto-regression manner, and our newly proposed structural prior helps the reconstruction from the spatial contour layout. Although better performance can be obtained, however, the high-level semantic information about vehicles is still ignored. For example, on automotive manufacturer websites, you can find pervasive descriptive introductions, attribute information, specific parameters, and hardware configurations about vehicles. Multi-modal pre-training can leverage more clues and this allows the big model to better understand the vehicle.

In this work, we adopt pre-trained vision-language model CLIP to process the vehicle image I and corresponding natural language descriptions $T = [w_1, w_2, ..., w_m]$, where w_i denotes the i^{th} English word. Specifically, the language encoder of CLIP (Radford et al. 2021) embeds the sentence into a set of word tokens $W \in \mathbb{R}^{12693\times512}$. Meanwhile, the vision encoder of CLIP (Radford et al. 2021) transforms the vehicle image into visual tokens $V^c \in \mathbb{R}^{1\times512}$. Note that, the parameters of both CLIP (Radford et al. 2021) visual and language encoder are fixed. The cosine similarity between the CLIP (Radford et al. 2021) visual features and MAE Transformer decoded features are considered in the reconstruction process. In addition, we also introduce cross-modality contrastive learning to achieve semantic-aware feature learning in the decoding phase.

We normalize the CLIP visual features and MAE transform decode features using L2 normalization and calculate the similarity loss between the two features as follows:

$$L_{cf} = (\frac{F^t}{||F^t||_2} - \frac{V^c}{||V^c||_2})^2 = (\widetilde{F^t} - \widetilde{V^c})^2 \tag{4}$$

where F^t is decoded features from MAE, and V^c is CLIP visual features.

On the other hand, we also consider the consistency constraint between the similarity of CLIP text-visual features and CLIP text-MAE decoded visual features. To be specific, the similarity between the text $W = w_j, j = \{1, 2, ..., m\}$ and the MAE decoded features is firstly calculated by:

$$s_j(\widetilde{F^t}, W) = \frac{exp((\widetilde{F^t} * w_j)/\tau)}{\sum_{m=1}^{M} exp((\widetilde{F^t} * w_m)/\tau)} \tag{5}$$

where τ is a temperature hyper-parameter and we set it as 1 in our experiments. The similarity distribution between the text and MAE decoded features can be obtained via:

$$S(\widetilde{F^t}, W) = [s_1(\widetilde{F^t}, W), s_2(\widetilde{F^t}, W), ..., s_M(\widetilde{F^t}, W)]. \tag{6}$$

Similarly, we can get the similarity distribution between the text and CLIP visual features via:

$$S(\widetilde{V^c}, W) = [s_1(\widetilde{V^c}, W), s_2(\widetilde{V^c}, W), ..., s_M(\widetilde{V^c}, W)] \tag{7}$$

The consistency constraint can be achieved by minimizing the two similarity distributions, i.e., $KL(S(\widetilde{V^c}, W), S(\widetilde{F^t}, W))$, here, KL is short for relative entropy loss function. The regularization term of the similarity distribution between the MAE decoded features and CLIP text features is also introduced to enhance our model, which can be written as $H(S(F^t, W))$, here, H is short for entropy. Therefore, the regularized similarity distribution consistency loss function can be formulated as:

$$L_{cs} = \sum_{i=1}^{N} KL(S(\widetilde{V^c}, W), S(\widetilde{F^t}, W)) + \sum_{i=1}^{N} H(S(\widetilde{F^t}, W)) \tag{8}$$

Finally, the overall loss function of our proposed VehicleMAE can be expressed as:

$$L = L_r + L_{mim} + L_{cls} + L_{cf} + L_{cs}. \tag{9}$$

Downstream Tasks In this work, four downstream tasks are adopted to validate the effectiveness and generalization of our proposed VehicleMAE big model, including vehicle re-identification, attribute recognition, fine-grained recognition, and part segmentation. A brief introduction to these tasks can be found in the supplementary materials.

			VAR				V-Reid		VFR	VPS	
Method	**Dataset**	**mA**	**Accuracy**	**Precision**	**Recall**	**F1**	**mAP**	**R1**	**Accuracy**	**mIou**	**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).

Experiments

Dataset, Metric, and Implementation Details In this work, the proposed VehicleMAE big model is pre-trained on our newly proposed **Autobot1M** dataset. Then, we validate its effectiveness and generalization on three datasets corresponding to four downstream tasks. A brief introduction to these datasets is given below.

Pre-training Dataset. In this paper, we propose a large-scale, high-quality vehicle-centric dataset, termed Autobot1M. It contains 1026394 vehicle images from diverse scenarios and sources, including existing vehicle dataset CompCars (Yang et al. 2015) and VERI-Wild (Lou et al. 2019). There are 732112 surveillance images and 294282 network images. These images fully reflect the key features of vehicles, such as illumination, motion blur, viewpoints, and occlusion. Our dataset also considers multiple challenging factors such as illumination, motion blur, viewpoints, occlusion, etc. It is worth noting that part of these images are crawled from the Internet and the corresponding natural language descriptions are available for the pre-training.

Downstream Datasets. In our four different downstream tasks, three datasets are adopted for the downstream validation, including the **VeRi** dataset (Liu et al. 2016), **Stanford Cars** dataset (Krause et al. 2013), and **PartImageNet** dataset (He et al. 2022a). More details about these datasets can be found in our supplementary materials.

Evaluation Metric. In this work, multiple evaluation metrics are used for different downstream tasks, including mA, Accuracy (Acc), Precision, Recall, F1-score, map, R1, mIoU, and mAcc.

Implementation Details. In our pre-training phase, the learning rate is set as 0.00025, and the weight decay is 0.04. The AdamW (Loshchilov and Hutter 2018) is selected as the optimizer to train our model. The batch size is 512 and training for a total of 100 epochs on our Autobot1M dataset. The tradeoff parameters between various loss functions are set as 4, 0.02,0.02, 2, and 0.1, respectively. All the experiments are implemented using Python based on the deep learning toolkit PyTorch (Paszke et al. 2019). A server with four RTX3090 GPUs is used for the pre-training. About 58 hours are needed for our pre-training phase.

Compare with Other SOTA Algorithms In this experiment, four downstream tasks are used for the validation of our VehicleMAE pre-trained big model. Three different settings of the training data in the pre-training phase are evaluated, i.e., full data, 20%, and 10% of the training data. We compare with our baselines and also other state-of-the-art models on each downstream task, as shown in Table 1. More in detail, the models without the pre-training (i.e., learning from scratch), the pre-trained MAE model on the ImageNet dataset, and the pre-trained MAE model on our newly proposed Autobot1M dataset.

For vehicle attribute recognition, we report and compare the attribute recognition results on the Veri-776 dataset. The baseline approach VTB (Cheng et al. 2022) proposed by Cheng et al. in the year 2022 is selected and equipped with the pre-trained big models to validate the effectiveness. We can find that the VTB achieves 84.67%, 80.86%, 84.66%, 85.77%, and 84.90% on mA, Accuracy, Precision, Recall, and F1 metrics, respectively, based on ViT-base when learning from scratch. When initializing the ViT-base backbone network using parameters learned by MAE on the ImageNet dataset, the recognition results can be improved to 89.69%, 93.60%, 94.81%, 95.54%, 95.08%. This experiment demonstrates that the visual features learned by self-supervised pre-training contribute significantly. When replacing the ImageNet using our proposed Autobot1M dataset, the recognition performance can be further improved which demonstrates that the pre-training on vehicle images performs better than the generalized data in natural scenarios. Note that, our proposed pre-trained framework performs the best when compared with the aforementioned models, i.e., 92.21%, 94.91%, 96.00%, 96.50%, 96.17%. In contrast, existing pre-trained big models like the MoCov3 (Chen, Xie, and He 2021), DINO (Caron et al. 2021), and IBOT (Zhou et al. 2021), are all inferior to our model. These experimental results and comparisons fully validated the effectiveness of our proposed pre-training strategy for vehicle-based pre-training. Similar conclusions can also be drawn from the other three downstream tasks.

Ablation Study

Effects of Structural Prior In this paper, we introduce the structural prior to guide the reconstruction of given vehicle images. Two loss functions are involved here, i.e., L_{mim} and L_{cls}. As shown in Table 2, we introduce the two loss functions into the pre-training, and the performance is all improved on four downstream tasks. For example, the results are boosted to 91.27%, 94.11%, 95.29%, 95.82%, 95.50%

MAE Loss	Structural Prior		Semantic Prior		VAR					V-ReID		VFR	VPS	
L_r	L_{mim}	L_{cls}	L_{cs}	L_{cf}	mA	Acc	Prec	Rec	F1	mAP	R1	Acc	mIou	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 2: Ablation study on loss functions in Structural Prior and Semantic Prior.

Ratio of Masked Token	VAR					V-ReID		VFR	VPS	
	mA	Acc	Prec	Rec	F1	map	R1	Acc	mIoU	mAcc
0.25	90.48	94.34	95.63	96.02	95.72	84.9	97.0	94.0	72.31	78.20
0.50	91.88	94.35	95.55	96.11	95.72	85.2	97.3	94.3	71.90	77.66
0.75	92.21	94.91	96.00	96.50	96.17	85.6	97.9	94.5	73.29	80.22
0.85	90.73	94.18	95.32	95.97	95.55	82.1	96.3	93.5	70.91	77.12

Table 3: Ablation study on the ratio of masked tokens.

when the L_{mim} is adopted, and to 91.71%, 94.54%, 95.65%, 96.28%, 95.88% when both L_{mim} and L_{cls} are used. These experimental results and comparison fully validated the effectiveness of our proposed structural prior. Similar conclusions can also be drawn from other tasks.

Effects of Semantic Prior We introduce two loss functions for the semantic prior, i.e., L_{cs} and L_{cf}. As shown in Table 2, 75.5%, 95.4% are obtained on mAP and R1 for the vehicle re-identification task when only the MAE loss function is used. When L_{cs} is utilized, the results can be improved to 84.1%, 97.1%, which are significant improvements. Note that, these results can be further improved to 85.2%, 97.1% when both L_{cs} and L_{cf} are used. We can achieve the best performance on four downstream tasks when the four loss functions are all used based on MAE, which fully validates the effectiveness of our proposed structural and semantic prior information for the MAE based vehicle reconstruction.

Analysis on Ratio of Masked Tokens As shown in Table 3, we set different ratios of masked tokens to check their influence. Specifically, 0.25, 0.50, 0.75, 0.85 are tested and the experimental results on the four downstream tasks demonstrate that better performance can be obtained when 75% of the input tokens are randomly masked.

Analysis on Different Sizes of Training Data in Downstream Tasks To validate the effectiveness of our pre-trained VehicleMAE on tasks with few training samples, in this part, we utilize 10% and 20% of the training data in the downstream tasks. As shown in Table 4, we can find that when 10% of training data is used for learning from scratch, the results mIoU and mAcc on vehicle part segmentation are 35.44%, 46.22%, respectively. When utilizing the pre-trained model using MAE on ImageNet-1K dataset, the results are improved to 52.31%, 62.30%. This comparison demonstrates that the self-supervised pre-training contributes significantly to the feature representation. If our proposed Autobot1M dataset is used, the results can be further boosted to 62.35%, 69.56%. We can find that the overall results are best on the part segmentation task when both our dataset and VehicleMAE framework are all adopted, i.e., 65.09%, 71.19%. Similar conclusions can also be drawn from other tasks and the settings of 20% of the training dataset. These experiments fully validated the key contributions of our model and dataset for the vehicle-based perceptron.

Efficiency Analysis

As shown in Table 5, we report the FLOPs[1], MACs[2], and Parameters of our proposed VehicleMAE and other pre-trained big models to help the readers better understand the efficiency metrics. It is easy to find that the FLOPs and MACs of MAE are all 9.43G, while ours are 10.98G. The parameters of MAE is 111.65M and ours is 121.62M, which demonstrates that the scale of MAE and our proposed VechileMAE are similar. Meanwhile, our model achieves better results on multiple downstream tasks than the MAE framework.

Visualization

In this section, we visualize the feature maps of MAE and our VehicleMAE big model, the masked tokens, and reconstructed images. As shown in Fig. 3, we give the feature maps of four vehicle images on four downstream tasks. The GradCAM[3] is adopted to visualize the attention maps of 11^{th} Transformer block. Compared with our baseline MAE, we can find that the heat maps are higher in the key regions. Therefore, our model performs better on the tested vehicle-based perceptron tasks.

We also visualize the masked tokens in the third and fourth row of Fig. 3, the first and second columns are input images and images with masked regions. The third and

[1]https://pypi.org/project/ptflops/

[2]https://pypi.org/project/thop/

[3]https://mmpretrain.readthedocs.io/en/latest/useful_tools/cam_visualization.html

Training Data	Method	Dataset	VAR					V-ReID		VFR	VPS	
			mA	Acc	Prec	Rec	F1	mAP	R1	Acc	mIoU	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 4: Results of different scales of training data used in downstream tasks. - denotes no corresponding results.

MoCov3	DINO	IBOT	MAE	VehicleMAE
18.00	16.89	18.52	9.43	10.98
17.58	16.88	18.51	9.43	10.98
86.57	108.87	96.29	111.65	121.62

Table 5: Comparison of FLOPs (G, 1^{st} row), MACs (OPs, 2^{sec} row), and Parameters (M, 3^{rd} row) of ours and other pre-trained big models.

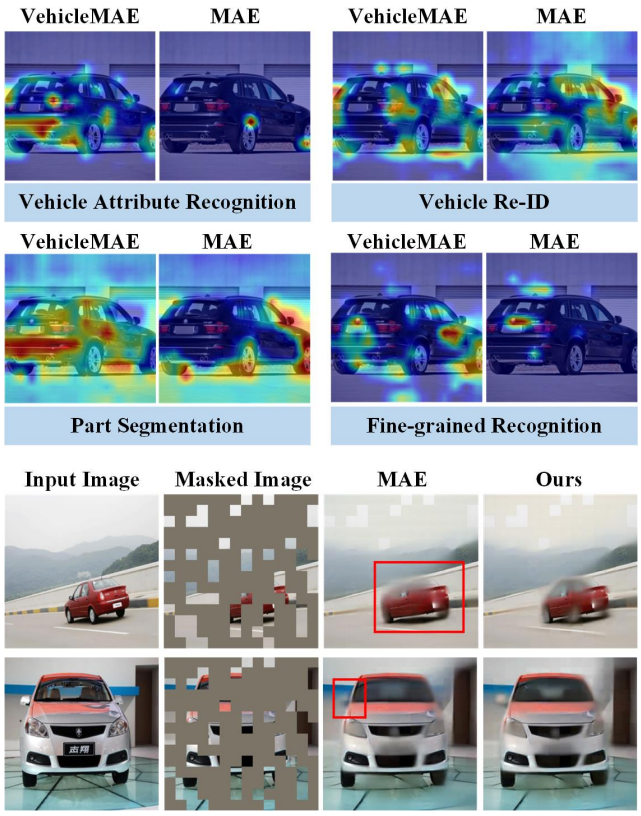

Figure 3: Visualization of attentions and reconstructed vehicle images.

fourth columns are images with reconstructed patches using MAE and our proposed VehicleMAE. Thanks to the structural and semantic prior information, the predicted regions using our model are significantly better than the MAE model, as highlighted in red bounding boxes.

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 (about 75%) 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. The sketch lines of vehicles are extracted as a form of structured information to guide vehicle reconstruction. Knowledge distill from the big multimodal model CLIP on the similarity between the paired/unpaired vehicle image-text sample is also considered. 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. Four downstream tasks including vehicle attribute recognition, fine-grained recognition, re-identification, and part segmentation, are adopted for the evaluation and comparison. Extensive experiments fully validated the effectiveness and benefits of our VehicleMAE and Autobot1M dataset.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62102205, 62002069, 62376004, 62306005), Australian Research Council Projects IH-180100002, Natural Science Foundation of Anhui Province (No. 2208085J18), Natural Science Foundation of Anhui Higher Education Institution (No. 2022AH040014), Anhui Provincial Key Research and Development Program (202104d07020008). The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Chadwick, S.; Maddern, W.; and Newman, P. 2019. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, 8311–8317. IEEE.

Chen, X.; Xie, S.; and He, K. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 9620–9629. IEEE.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6994–7004.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

He, J.; Yang, S.; Yang, S.; Kortylewski, A.; Yuan, X.; Chen, J.-N.; Liu, S.; Yang, C.; Yu, Q.; and Yuille, A. 2022a. Partimagenet: A large, high-quality dataset of parts. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, 128–145. Springer.

He, J.; Zhang, S.; Yang, M.; Shan, Y.; and Huang, T. 2020. BDCN: Bi-directional cascade network for perceptual edge detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 100–113.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022b. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 869–884. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; and Duan, L. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3235–3243.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Scribano, C.; Sapienza, D.; Franchini, G.; Verucchi, M.; and Bertogna, M. 2021. All you can embed: Natural language based vehicle retrieval with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4253–4262.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wang, H.; Peng, J.; Chen, D.; Jiang, G.; Zhao, T.; and Fu, X. 2020. Attribute-guided feature learning network for vehicle reidentification. *IEEE MultiMedia*, 27(4): 112–121.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 1–36.

Wang, X.; Chen, Z.; Jiang, B.; Tang, J.; Luo, B.; and Tao, D. 2022a. Beyond Greedy Search: Tracking by Multi-Agent Reinforcement Learning-Based Beam Search. *IEEE Transactions on Image Processing*, 31: 6239–6254.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022b. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121: 108220.

Yang, L.; Luo, P.; Change Loy, C.; and Tang, X. 2015. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3973–3981.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*.