# Supplementary Material for VehicleMAE: View-asymmetry Mutual Learning for Vehicle Re-identification Pre-training via Masked AutoEncoders

In this supplementary material, we will provide more detailed visualization examples of DiffVERI and qualitative experimental results of our pre-training model.

## A. Dataset Visualization

To visually demonstrate the synthesis quality of DiffVERI, we provide some multi-view synthetic examples and view-mask annotations in Figure 1. It can be observed that each vehicle identity extremely retains its original appearance. In brief, the advantages of DiffVERI are summarized as follows:

**(1) Large-scale number of images:** DiffVERI includes over 1700k images and is currently the largest benchmark in terms of data size.

**(2) High quality Data:** Utilizing diffusion models for data generation and filtering operations on VERI-Wild overcomes the problems of motion blur and low light in the raw data collected by real-world surveillance systems. In contrast, the images in the DiffVERI dataset have higher resolution than other ones.

**(3) Diverse views and annotations:** DiffVERI covers a diverse range of views for each identity vehicle. Owing to the fine-tuning of the SAM model, we also equipped each image with a multi-view semantic mask to provide rich view clues for a series of subsequent studies.

Although DiffVERI can provide comprehensive view data for each vehicle instance, the generated images inevitably contain some detail errors, such as stickers on the windshield or surrounding background. This implies that there are more or less a small number of identity-related pixel-level errors in DiffVERI. Thus, VehicleMAE can effectively decouple the dependence on identity-level annotations and pay more attention to the appearance discrepancy at the view-level through a self-supervised pre-training manner on the DiffVERI. Additionally, the VMIM module incorporates multi-view semantic reconstruction tasks alongside pixel-level reconstruction tasks, effectively compensating for pixel-level representation errors through view semantic learning.

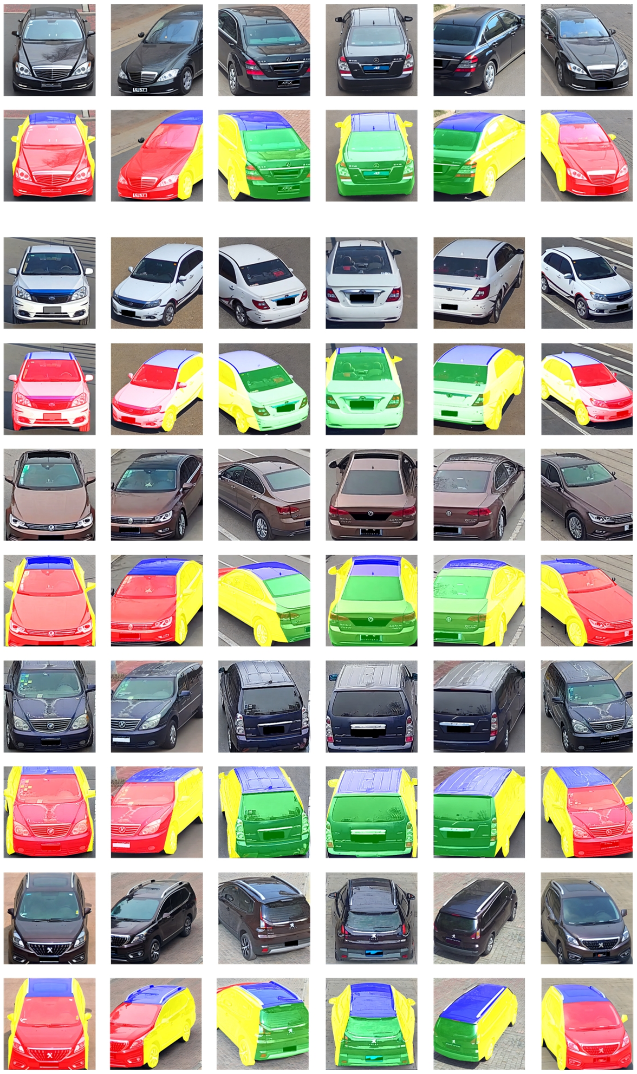
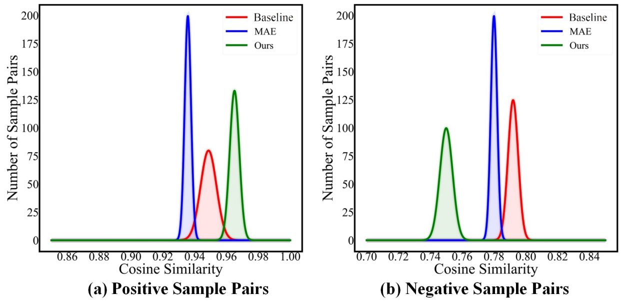


Figure 1. Some synthesized instances and multi-view annotations. Each two adjacent rows represent the synthesis images of multiple view ranges for two vehicle identities and the corresponding view-masks.



**(a) Positive Sample Pairs** **(b) Negative Sample Pairs**

Figure 2. Comparison of distance distribution for positive and negative sample pairs using different pre-training models. The horizontal and vertical axes of figure 2 (a) and (b) represent the cosine distance and number of sample pairs, respectively.

# B. Additional Qualitative Results

## B.1. Visualization of Feature Distance Distribution

Figure 2 further explores the distance metirc performance of different pre-training models on positive and negative sample pairs. Specifically, we randomly select 1600 positive and negative sample pairs from VeRi-776. Then, under a challenging unsupervised learning manner in VeRi-776, we adopt the fine-tuned Baseline, MAE, and our pre-training model for feature extraction and cosine similarity calculation. From Figure 2 (a) and (b), it can be observed that compared to Baseline and MAE, our pre-training model curve reaches its peak in the interval where the cosine distance between positive sample pairs is more similar, while the distribution of negative sample pairs is exactly the opposite. This observation confirms that our pre-training model can robustly reduce the intra-class differences and increase the inter-class discriminative ability.

## B.2. Visualization of Attention Map

To more intuitively demonstrate the ability of our pre-training model to capture discriminative clues, we provide different vehicle instances generated attention maps by pre-training on Baseline, MAE, and VehicleMAE, as shown in Figure 3. We can clearly observe that Baseline and MAE inevitably focus on some unrelated background regions. In contrast, our model focuses more on the informative region of each vehicle, such as the headlights on the roof and the text at the rear. These findings are in line with our expectations and confirm that our pre-training model can accurately capture discriminative regions and filter out identity-irrelevant clues from different views.

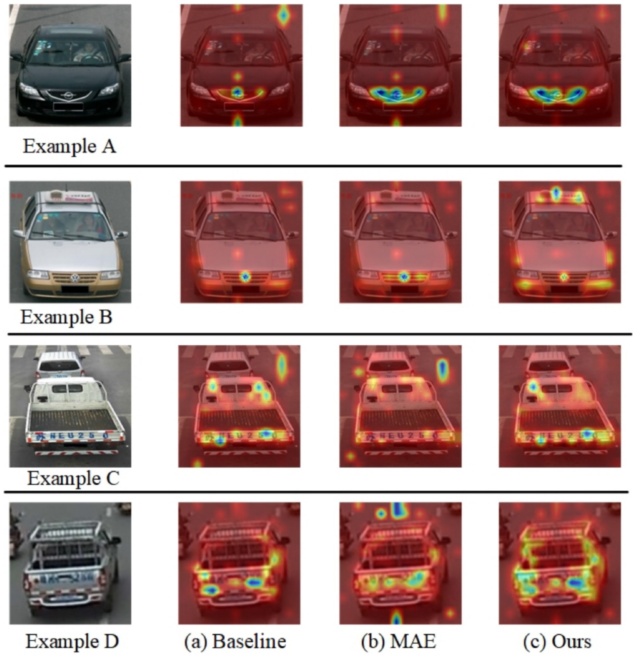


Figure 3. The visualization of attention maps extracted by Baseline, MAE and Ours Pre-training model.