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

Qi Wang[1], Zeyu Zhang[1], Dong Wang[2*], Di Gai[1], Xin Xiong[3], Jiyang Xu[1], Ruihua Zhou[2]

[1]School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China
[2]School of Software, Nanchang University, Nanchang, China
[3]The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China
[*] corresponding author

{wangqi, gaidi, xiongxinxx}@ncu.edu.cn, {zeyuzhang1010, dongwang, xujiyang, zrh}@email.ncu.edu.cn

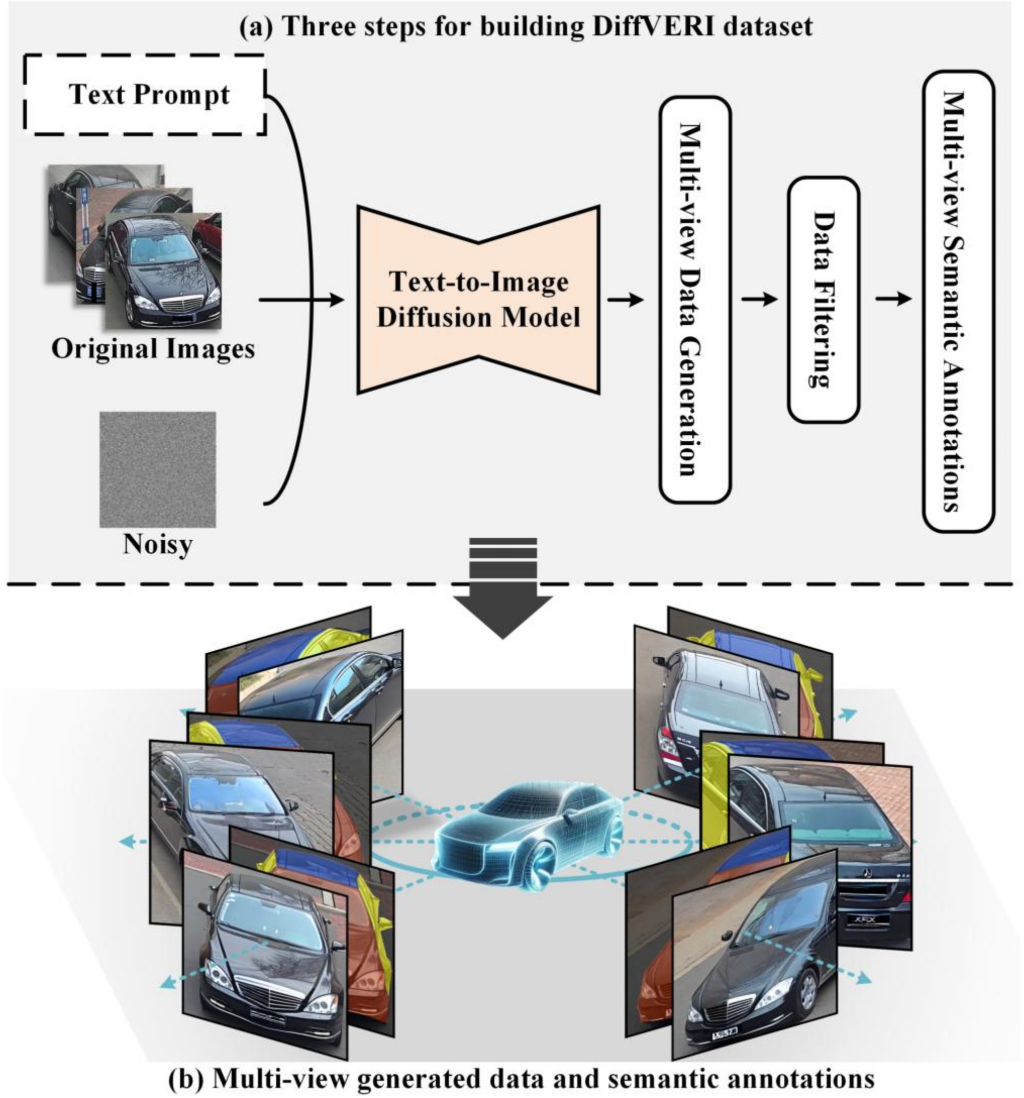
Background

- **Vehicle Re-ID:** Match same vehicle across different views
- **Challenge:** Large intra-class variation due to viewpoint changes
- **Problem:** Lack of large-scale multi-view vehicle datasets

Target

- **Solution:** Build DiffVERI dataset and propose VehicleMAE for pre-training

DiffVERI Dataset



Pipelines

- **Synthetic data generation** using DreamBooth (diffusion model).
- **1.7M+ images** with multi-view semantic annotations.
- **Data filtering** via YOLOv7 and manual annotation.
- **Multi-view segmentation** using fine-tuned SAM model.

DiffVERI Benchmark

Datasets	Images	Source	Multi-view Semantic Annotations	Resolution	Views
VeRi-776[25]	49,357	Real	✗	243 × 214	Constrained
VehicleID[26]	221,763	Real	✗	374 × 412	Constrained
VeRi-Wild[27]	416,314	Real	✗	415 × 354	Constrained
VehicleX[48]	192,150	Synthetic	✗	256 × 256	Constrained
VRAI[42]	137,613	Real	✗	349 × 234	Constrained
DiffVERI	1,712,703	Real-synthetic	✓	496 × 485	Diverse

DiffVERI

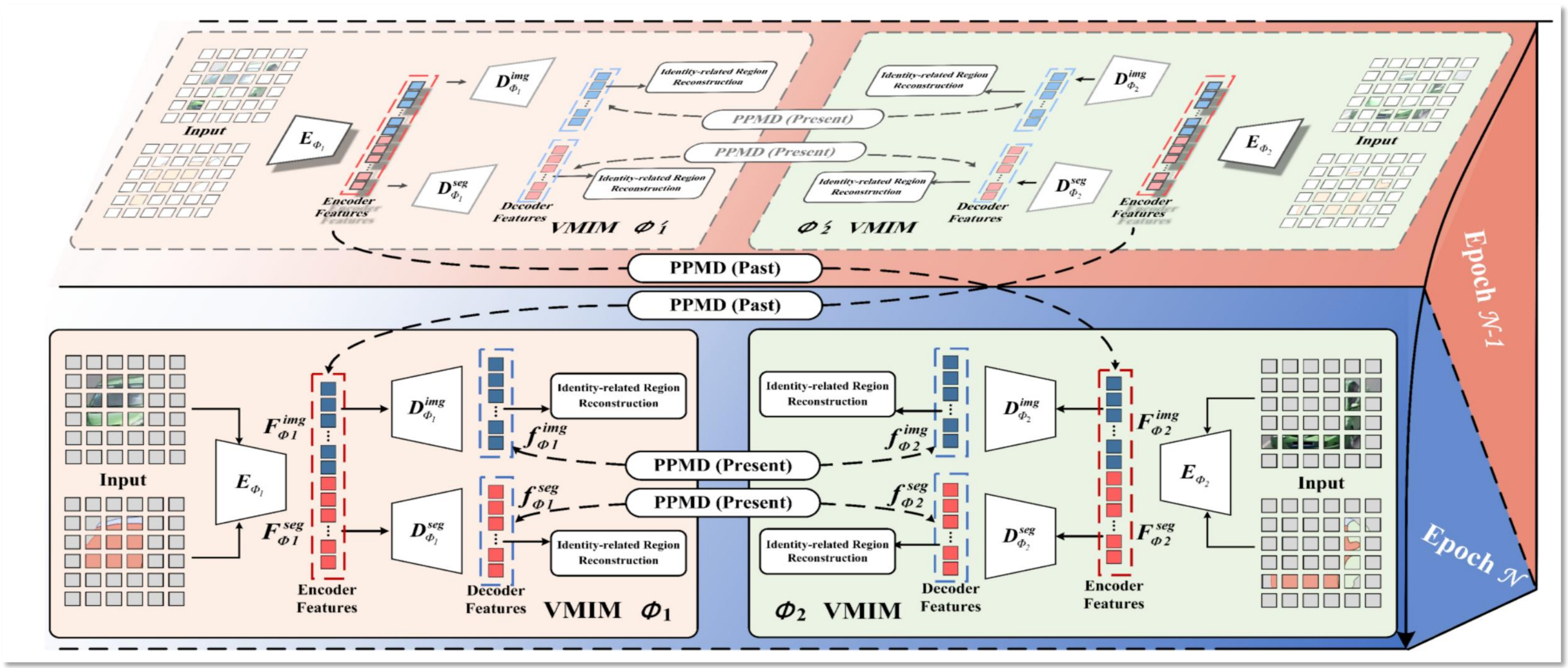
Comparison

- Comparison of the statistics between **DiffVERI** and other public vehicle Re-ID datasets. In contrast, **DiffVERI** is currently the largest multi-view vehicle Re-ID benchmark

Dataset Example

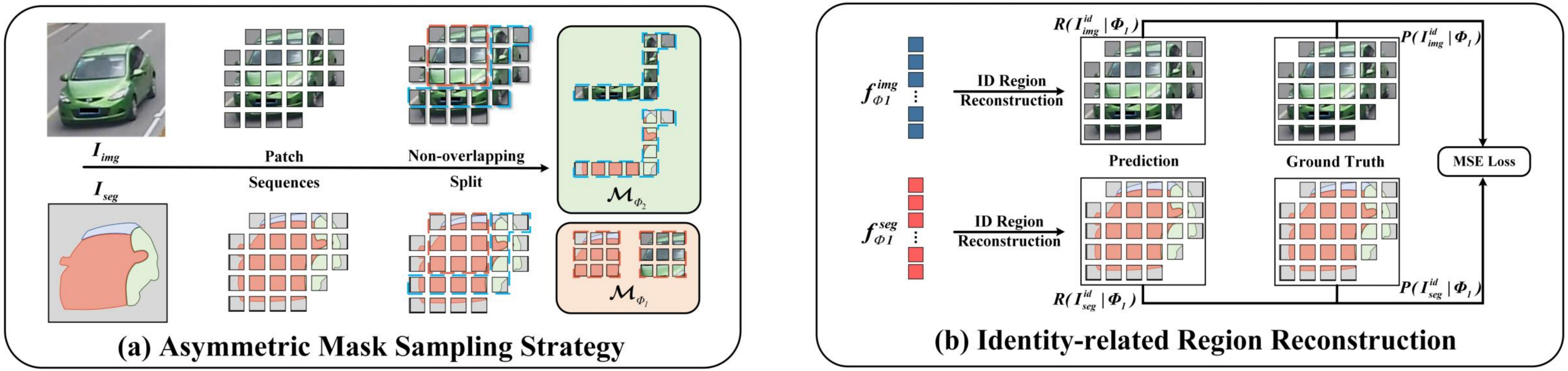
- Some synthesized instances and multi-view annotations. The two adjacent rows represent the synthesis images of multiple view ranges for two vehicle identities and the corresponding view masks.

VehicleMAE



Input
Encoder Features
Decoder Features
Identity-related Region Reconstruction
PPMD (Present)
PPMD (Past)
VMIM
Epoch N-1
Epoch N

View-asymmetry Masked Image Modeling



VMIM consist of two submodules:

(a) The asymmetric mask sampling strategy in VMIM module generates a pair of visible maps without overlapping patches to create diverse preservation clues for reconstruction tasks.

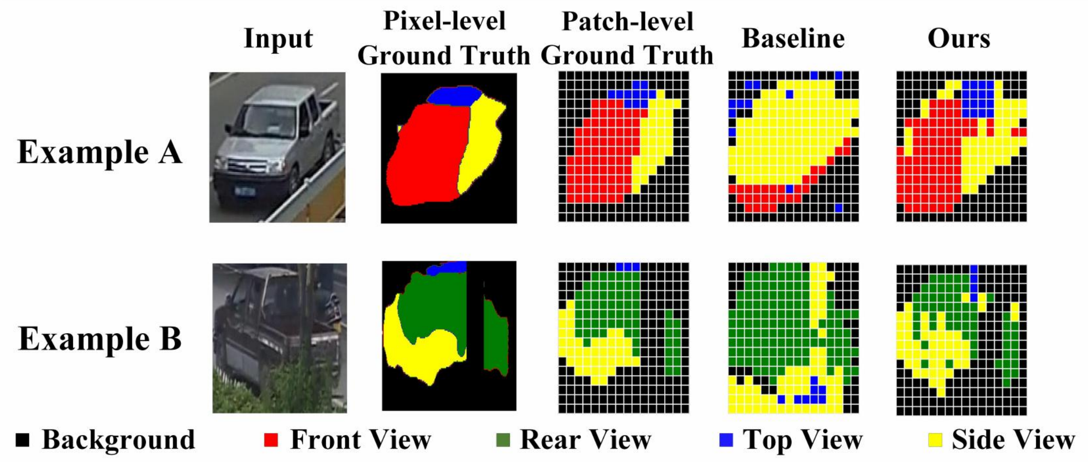
(b) Illustration of the identity-related region reconstruction in terms of ϕ1.

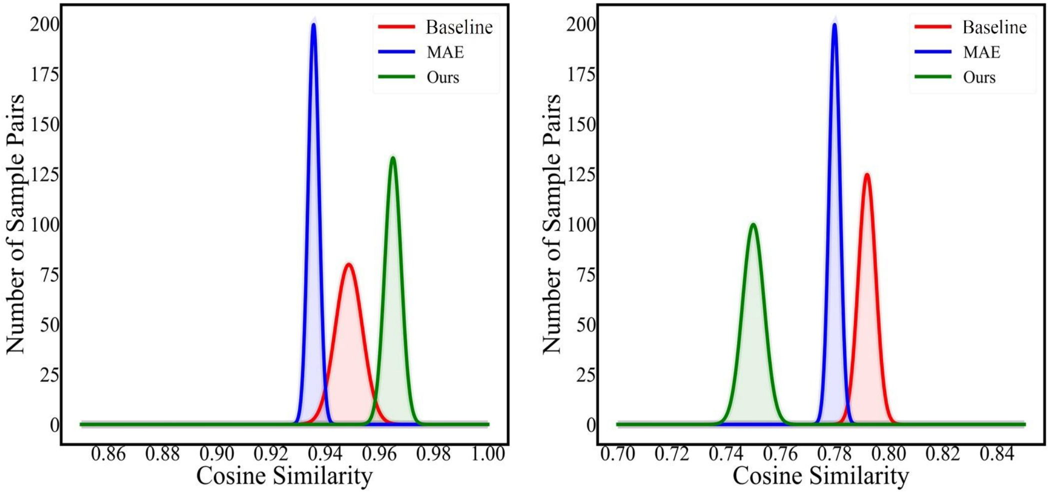
Methods	VeRi-776		VehicleID	
	mAP	Rank-1	Rank-1	Rank-5
Supervised Learning				
AAVER∗^{R}[19]	61.2	89.0	63.5	85.6
VehicleNet∗^{R}[54]	83.4	96.8	79.5	92.0
TransReID∗^{V}[14]	82.0	97.1	-	-
GiT∗^{V} [35]	80.3	96.9	77.9	-
LCNL∗^{V}[46]	81.8	97.4	-	-
TANet∗^{R}[17]	83.6	96.8	78.2	92.6
CFA-Net∗^{R}[41]	80.7	96.9	-	-
HCI-Net∗^{R}[36]	83.8	96.6	76.4	91.2
CLIP-ReID∗^{C}[24]	83.3	97.4	78.1	92.7
Ours†^{C}	**87.6**	**97.4**	**85.9**	**94.9**
Unsupervised Learning				
MMT∗^{R}[7]	25.4	60.9	31.0	42.4
SPCL∗^{R}[8]	36.9	79.9	53.0	66.4
RLCC∗^{R}[51]	39.6	83.4	-	-
PPLR∗^{R}[4]	41.6	85.6	-	-
VAPC TO∗^{R}[52]	30.4	76.2	-	-
ICL∗^{R}[37]	39.5	83.7	-	-
AdaMG∗^{R}[31]	41.0	86.2	-	-
NNNI∗^{R}[9]	42.3	86.3	-	-
STDA∗^{R} [13]	42.3	87.4	-	-
MAE†^{C}[12]	37.6	81.2	69.0	81.8
CL-MAE†^{C}[29]	41.0	84.0	70.8	82.6
CMAE†^{C}[18]	41.6	84.7	71.1	84.9
Ours†^{C}	**42.5**	**87.6**	**73.3**	**85.6**

Experiments

- **Datasets:** VeRi-776, VehicleID
- **Metrics:** mAP, Rank-1, Rank-5
- **Quantitative comparison:**

Our method delivers the best mAP and Rank accuracy for both supervised and unsupervised settings. The noticeable performance improvement implies that the masked image modeling based on VehicleMAE is better suited for downstream tasks of vehicle Re-ID.





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

Qualitative Comparison

- **Feature Distributions:** The first figure displays two visualization examples of the feature distribution at the patch-level extracted by Baseline and VehicleMAE.
- **Distance Distribution:** Next figure further explores the distance metirc performance of different pre-training models on positive and negative sample pairs.

Conclusion

This paper releases **DiffVERI**, a large-scale multi-view vehicle Re-ID dataset for learning view-invariant representations, and proposes a masked image modeling pre-training paradigm termed **VehicleMAE** specially for vehicle Re-ID downstream tasks. **VehicleMAE** first proposes a VMIM module that attempts to apply two homogeneous MAEs to predict the RGB pixels and multi-view semantic clues of vehicles in pairs, thereby gaining diverse multi-view inference capabilities. Subsequently, to facilitate learning collaboratively, a PPMD module is designed to progressively exchange knowledge with each other. Extensive experiments demonstrate that equipping our pre-training model can achieve competitive performance in generic vehicle ReID downstream tasks. Future work contributes to further expanding **VehicleMAE** into a unified multimodal pre-training paradigm.