Coarse-to-Fine Cross-Modality Generation for Enhancing Vehicle Re-Identification with High-Fidelity Synthetic Data

Leyang Jin[1], Wei Ji[2], Tat-Seng Chua[1], Zhedong Zheng[3†]

***Abstract*— Due to the critical issues of privacy and partial occlusion, license plate information is not always available in vehicle recognition systems. Consequently, researchers have increasingly turned towards vehicle re-identification (reID) techniques to bridge the gap between cross-view camera systems. Despite the growing interest, one major challenge persists: the scarcity of authentic, large-scale training datasets. To address this challenge, this paper introduces a coarse-to-fine generation pipeline designed to synthesize high-fidelity vehicle data, thereby facilitating subsequent vehicle representation learning. Specifically, the proposed approach consists of three stages: Prompt Processing, Diffusion Fine-tuning, and Semantic Filtering. First, we collect detailed prompts from vehicle websites and companies with fine-grained vehicle prototype attributes. Next, we leverage the prior knowledge of these automotive prototypes to fine-tune diffusion models. Finally, to ensure the quality of the synthesized data, we employ pre-trained vision-language models to filter out substandard images. Building upon the high-quality data generated by this pipeline, we validate the effectiveness using vanilla models. Extensive experimental evaluations demonstrate that our approach achieves competitive accuracy on public benchmarks such as VeRi-776, VehicleID and CityFlowV2, and is compatible with various model architectures.**

I. Introduction

Vehicle re-identification (reID) aims to match images of the same vehicle across multiple cameras, which is crucial for the deployment of autonomous vehicles [1] and intelligent traffic systems [2]. Given the minor intra-class differences between car models, vehicle reID is typically treated as a fine-grained representation learning task [3], [4]. However, privacy concerns [5] and annotation difficulties in multi-sensor systems [6], [7] result in a scarcity of realistic training data. To address this issue, recent research [8], [9], [10] has focused on generating synthetic data for vehicle reID. Despite these efforts, generating large-scale, high-fidelity training data that captures subtle inter-class discrepancies and intra-class consistencies remains challenging.

Existing efforts on vehicle reID data generation can be divided into two directions: 1) Graphics-engine-based methods, such as PAMTRI [8] and VehicleX [9]. They employ 3D

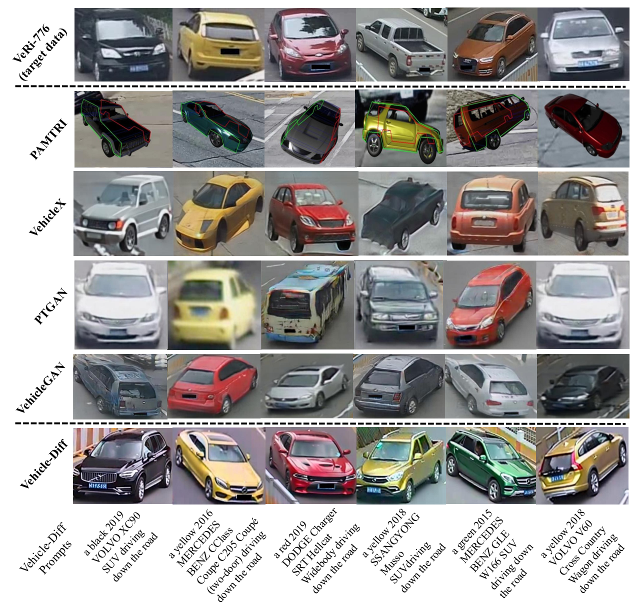


Fig. 1: We compare our Vehicle-Diff dataset to existing synthetic datasets. The second and third rows of datasets are based on 3D engines (PAMTRI [8] and VehicleX [9]), while PTGAN [11] and VehicleGAN [10] adopt the data-driven structure, *i.e.*, Generative Adversarial Networks [12]. We could observe that the proposed method is with a closer visual appearance compared to the real dataset, *i.e.*, VeRi-776. Besides, the generated images by the proposed method are associated with text captions, allowing for cross-modality knowledge to guide generation.

CAD models to generate vehicle images. While these methods have made significant strides, they still face challenges. There is a notable domain gap between rendered 3D CAD vehicle images and actual real-world images. Additionally, the process of generating the VehicleX dataset relies heavily on a large amount of labeled vehicle re-identification data, which is costly and raises privacy concerns. Similarly, synthetic data from PAMTRI needs to be combined with fully labeled re-identification datasets. 2) Data-driven methods, such as generative adversarial networks (GANs) [12]. For instance, PTGAN [11] and VehicleGAN [10] explore GANs to synthesize novel vehicle views. Although these methods generate vehicle images with relatively good visual quality, they under-explore the cross-modality guidance and thus the fine-grained attributes of the same vehicle are often inconsistent, compromising the training process of the vehicle reID.

To address the aforementioned challenges, we propose Vehicle-Diff, a new pipeline designed to synthesize large-

[1]Leyang Jin and Tat-seng Chua are with School of Computing, National University of Singapore, Singapore 117417 e0792447@u.nus.edu, dcscts@nus.edu.sg

[2]Wei Ji is with School of Intelligence Science and Technology, Nanjing University, China 215163 weiji@nju.edu.cn

[3]Zhedong Zheng is with the FST and ICI, University of Macau, China 999078 zhedongzheng@um.edu.mo † corresponding author

This work is supported by the University of Macau Start-up Research Grant SRG2024-00002-FST and Multi-Year Research Grant MYRG-GRG2024-00077-FST-UMDF

scale training data for vehicle re-identification, facilitating the representation learning. In particular, the pipeline consists of three primary stages: prompt processing, diffusion model tuning, and semantic filtering. We first collect and process the prompt for vehicles with a focus on the vehicle attribute. To harness the pre-trained inherent knowledge of car prototypes, we employ carefully crafted prompts. Then, we fine-tune the diffusion model using only 1% of unlabeled target data during the generation stage. It enables the diffusion model to adapt to the target vehicle domain at both the content and stylistic levels. In the subsequent filtering stage, we apply sophisticated post-processing techniques to enhance the semantic alignment of the generated data. Our pipeline is scalable and adaptable to multiple downstream scenarios, reducing labeling costs and privacy concerns. As shown in Fig. 1, the generated vehicle images are much closer to the real-world data. Finally, we construct a new labeled vehicle re-identification dataset, called Vehicle-Diff, comprising 149,472 images of 4,940 distinct vehicles. The efficacy of Vehicle-Diff is substantiated through comparative evaluations with synthetic datasets produced by existing approaches. In summary, our paper makes the following contributions:

- A new coarse-to-fine cross-modality generation pipeline by prompting the diffusion model to craft a synthetic vehicle re-identification dataset tailored to a downstream scene, with only about 1% unlabeled images in the original dataset. To the best of our knowledge, our work is among the early attempts for large-scale training data generation with attributes for vehicle re-identification.
- Extensive experiments have validated that our pipeline can minimize the gap between synthetic and real data, facilitating the subsequential reID model learning. The proposed method has achieved competitive performance, *e.g.*, 83.79% mAP on the VeRi-776 dataset.

II. Related work

Vehicle Re-Identification. Vehicle re-identification (reID) involves retrieving vehicles of interest from a database of images collected by traffic cameras. Previous studies [13], [14], [15], [16] have achieved significant success using supervised learning. However, this approach faces challenges such as high annotation costs and privacy concerns when collecting and labeling data. To mitigate these issues, some works [17], [18] have explored unsupervised learning to reduce annotation requirements. Despite these efforts, substantial real data is still needed for general vehicle reID tasks [16], and attribute annotations remain preferable [19], [20]. In contrast, we propose a multi-modality data synthesis approach that significantly reduces the need for both real data and annotations, addressing these limitations effectively.

Synthetic Datasets for Vehicle Re-Identification Task. Synthetic data are increasingly used to address privacy concerns and high annotation costs in creating re-identification datasets [21], [22]. Previous works [8], [9], [23], [24], [25] have employed 3D engines to generate characters and vehicles, but these assets suffer from the intrinsic domain gap between virtual and real scenes and are time-consuming to create. VehicleGAN [10] and PTGAN [11] deploy GANs for data augmentation, with VehicleGAN focusing on AutoReconstruction and pose consistency, and PTGAN generating novel vehicle views based on given poses. However, these methods still require large labeled datasets for effective training and are constrained by the quality and patterns of the original data. In contrast, our multi-modality data synthesis approach reduces the need for both real data and annotations, addressing these limitations effectively.

Text-to-image Diffusion Models. Diffusion models [26], [27] have recently emerged as promising generative models, particularly for text-to-image generation, where they can produce images based on textual descriptions. Recent advancements such as Stable Diffusion [28], Stable Diffusion XL [29], and Midjourney [30] have demonstrated remarkable results in this domain. Leveraging the power of these models, methods like [31], [32], [33] have utilized diffusion models, *e.g.*, GLIDE [34], to generate synthetic data for image classification. Despite their impressive visual outcomes and applications, the potential of text-to-image diffusion models for vehicle re-identification remains underexplored. In this paper, we evaluate multiple state-of-the-art text-to-image models and identify the optimal model for enhancing downstream vehicle re-identification performance.

III. Method

An overview of Vehicle-Diff is provided in Fig. 2. Vehicle-Diff generates high-fidelity data in a coarse-to-fine manner to enhance reID network training, comprising three stages: (1) prompt processing, (2) diffusion fine-tuning, and (3) semantic filtering. First, the prompt processing stage (§III-A) constructs a prompt library and specifies vehicle attributes such as models and colors for image generation. Next, during the diffusion fine-tuning stage (§III-B), Vehicle-Diff fine-tunes the diffusion model using unlabeled vehicle images, improving its adaptation to vehicle image generation. Finally, in the semantic filtering stage (§III-C), Vehicle-Diff generates vehicle images with different IDs using the prompt library and fine-tuned model, followed by filtering these images through off-the-shelf detection and cross-modality alignment.

A. Prompt Processing

The prompt processing stage aims to construct discriminative vehicle attribute prompts to guide image generation, thus enhancing inter-class consistency and intra-class diversity. We first filter the noisy online information to collect vehicle attributes, *i.e.*, brand, production year, and body style, for different car models from an online car information website [1]. It is worth noting that color is an important attribute, and we will use it again in the third stage for semantic filtering. Moreover, inspired by alternating optimization [38] and human-diffusion interaction [39], [40], [41], we also develop a prompt template to improve the quality of the generated images. Specifically, we adjusted one component of the prompt template based on feedback from the diffusion

[1] https://www.autoevolution.com/

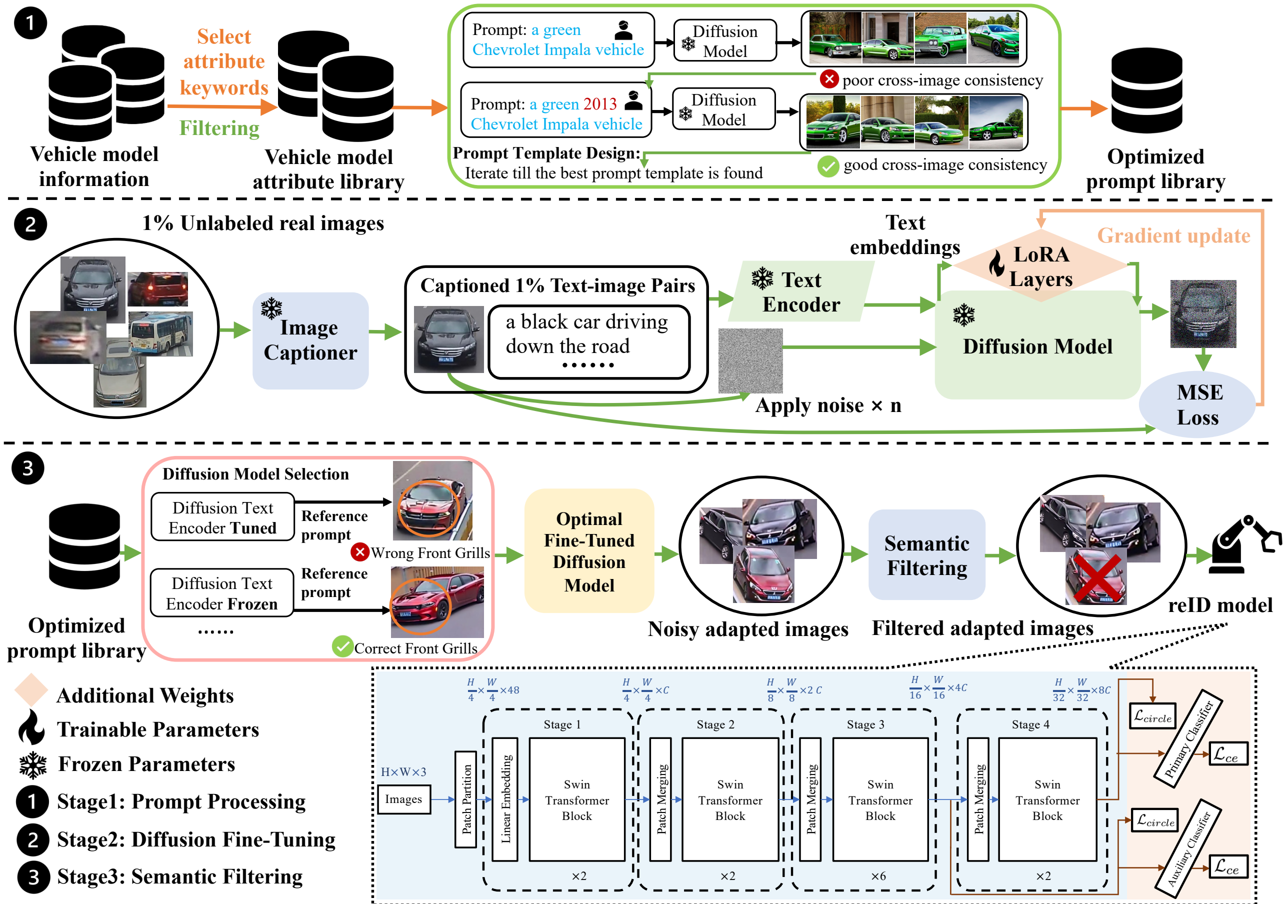


Fig. 2: An overview of our coarse-to-fine cross-modality pipeline Vehicle-Diff. It has three stages: Prompt Processing, Diffusion Fine-tuning, and Semantic Filtering. **(1)** We first scrape and filter vehicle model information from online vehicle websites. Given the diffusion model, we then select the prompt template according to the visual quality. **(2)** In the second stage, we leverage the off-the-shelf image captioner to generate the pseudo caption. It is worth noting that the proposed pipeline only requests a few unlabeled real images from the downstream dataset. After the data preparation, we fine-tune the diffusion model via Mean Squared Error (MSE) loss. **(3)** In the third stage, using the refined prompts, we choose the most effective diffusion model by comparing visual quality, such as consistency. Then, we create synthetic data for the vehicle re-identification task. We use the cross-modality model to filter out semantically misaligned data. Finally, we feed the high-fidelity data to train the reID model via cross-entropy loss [35], [36] and circle loss [37].

model. The final prompt template is designed as "a [color] [production year] [brand] [car model] [body style] driving down the road." In the bottom of Fig. 1, we show several examples of the prompt template and the resulting images.

B. Diffusion Fine-tuning

Vehicle-Diff leverages a text-to-image diffusion model to generate vehicle images according to prompts. However, a pre-trained diffusion model still struggles to adapt well to the real-world vehicle images, resulting in a domain gap between synthesized images and those in vehicle reID datasets. Therefore, we further fine-tune the diffusion model to mitigate the domain discrepancy while retaining its generation capability. As shown in Fig. 2 (Stage 2), we illustrate the step-by-step fine-tuning stage from the data preparation to the model optimization. To be specific, we first deploy an image captioner, *i.e.*, BLIP-2 [42], to predict text prompts for unlabeled vehicle images, and then employ the generated image-text pairs to fine-tune the text-to-image diffusion model. We incorporate additional weights [43] in the decoder part, while keeping the pre-trained weights unchanged. Therefore, the additional weights could adapt the final visual style, while maintaining the generative capability. The optimization objective is the mean squared error (MSE) loss. It is worth noting that, our Vehicle-Diff could be trained with **only a few (1%) unlabeled images of the vehicle dataset for fine-tuning,** *i.e.*, 378 images for VeRi-776 and 527 images for CityFlowV2, while previous methods either require large-scale datasets (GAN-based methods [10], [11]) or rely on labeled images (graphics-engine-based methods [8], [9]). Moreover, different from these methods, Vehicle-Diff harnesses the generative power of diffusion models, enabling to generate more realistic images, as shown in Fig. 1. Similarly, we fine-tune multiple candidate diffusion models in preparation for the next stage, which involves selecting the optimal diffusion model.

C. Semantic Filtering

We first sample approximately 10 prompts from the optimized prompt library to evaluate and select the optimal fine-tuned diffusion model. With a similar idea to our prompt template design, the selection of the fine-tuned model is informed by a qualitative assessment of the images generated by each candidate model. Fig. 2 (Stage 3) provides illustrative examples of fine-tuned models evaluated alongside the corresponding generated imagery. Through this evaluation, we opt for the fine-tuned diffusion model that maintains the text encoder in a frozen state. We then feed our designed prompts into the optimal fine-tuned diffusion model, which generates synthetic images automatically. Because of the limitations of text-to-image generation models in producing fine-grained and controllable outputs, directly using generated images is insufficient for training vehicle re-identification networks due to the following two major challenges, *i.e.*, multiple objects and semantic misalignment. We only need portions of the images that include the high-quality vehicle. Diffusion models can generate low-quality images, such as those with multiple vehicles, fragmented vehicles, or no vehicle at all. To tackle this issue, we utilize the YOLOv5x6 detection model [44], trained on high-resolution 1280×1280 images, for vehicle detection and cropping. The model is configured to detect only vehicle categories, with a single bounding box per image prioritizing the most prominent vehicle. We retain images with high-confidence detections and discard vehicles smaller than or equal to 250 pixels in height or width. After cropping, we have the vehicle in the center of the image, and we further screen out noisy images with semantic misalignment, such as vehicles with incorrect colors. In particular, we employ a cross-modal vision-language model, *i.e.*, CLIP [45], to extract the feature for both text and image modalities. We then remove semantic misaligned images that match wrong colors. Specifically, the test prompts are constructed as phrases, *e.g.*, "a red vehicle," where the color term is dynamically substituted from a predefined color list, such as "red," "yellow," "green," "white," and "black." The cosine similarity between image and test text in the feature level is:

$$\mathrm{sim}_k = \frac{\mathbf{f}_I \cdot \mathbf{f}_{T_k}}{\|\mathbf{f}_I\|\|\mathbf{f}_{T_k}\|}. \quad (1)$$

The predicted color $\hat{k}$ is identified as: $\hat{k} = \arg\max_k(\mathrm{sim}_k)$. We then compare the predicted color to the expected color, which is specified within the prompt used to generate the image. If the predicted color matches the expected color, the image is preserved; otherwise, it is discarded.

D. ReID Learning

In this paper, we do not pursue the network structure, but focus on the data aspect. Our generated data is compatible with different networks, and we are free to the reID model selection. Here, we take the typical transformer, Swin-V2 [46], as an example (please see the bottom of Fig. 2). We follow the existing works [47], [48] to add an auxiliary classifier to facilitate the backward gradients, especially for the large-scale dataset. To optimize the network, we adopt the classification loss [35], [36] and the circle loss [37] as $\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{circle}$, where $\mathcal{L}_{ce}$ is the cross-entropy loss to classify different vehicles, and the $\mathcal{L}_{circle}$ is to optimize the representation space by pulling closer positive images, while pushing away the negative samples. We apply the same loss terms to both the primary and auxiliary classifiers. It is worth noting that our synthetic data can be combined with real-world data to improve performance even further.

Dataset		#IDs	#Img	#Cam	Attr
Real	StanfordCars [50]	196	16,185	N/A	✓
	PKU-Vehicle [51]	N/A	10,000,000	N/A	✗
	CompCar [52]	4,701	136,726	N/A	✗
	PKU-VD1 [53]	1,232	1,097,649	1	✓
	PKU-VD2 [53]	1,112	807,260	1	✓
	VehicleID [54]	26,328	222,629	2	✗
	VehicleReID [55]	N/A	47,123	2	✗
	VeRi-776 [56]	776	49,357	20	✓
	CityFlow [57]	666	56,277	40	✗
	CityFlowV2 [58]	440	52,717	46	✗
	VRIC [59]	5,622	60,430	60	✗
Synthetic	PAMTRI [8]	402	41,000	Varied	✓
	VehicleX [9]	1,362	75,516†	Varied	✓
	Vehicle-Diff	4,896	149,472‡	Varied	✓

TABLE I: Statistic comparisons with public real-world and synthetic vehicle re-ID datasets in terms of the number of vehicle IDs, images, and viewpoints, and the availability of attributes. †: Number of images in their code. ‡: Given more text prompts, we could generate more images.

IV. Experiment

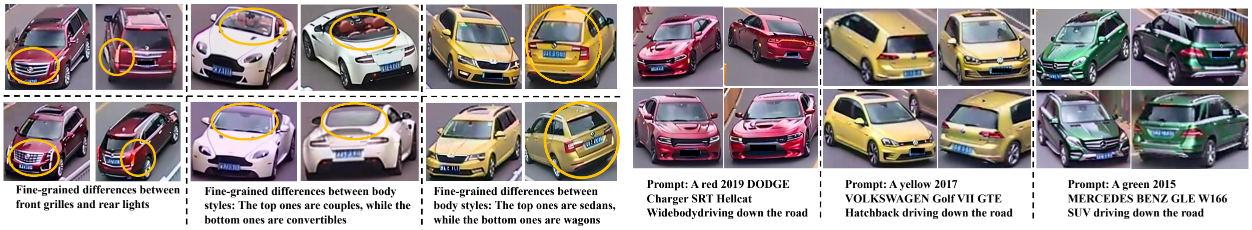
A. Implementation Details

Synthetic data generation. The Diffusion Fine-tuning process uses the Adam optimizer [49], with a learning rate of 0.0001 at the start and a polynomial scheduler for scheduling. We train the diffusion model for 100 epochs, with the first 20 serving as a warm-up. During inference, we set the guidance scale to 8, and the diffusion step to 50. The output size is set to 1024×1024. The vehicle detection threshold is set to 0.65. Our generation pipeline, Vehicle-Diff, yields 149,472 images of 4,940 vehicles on VeRi-776.

ReID baseline training. Following the setting of existing works [8], [9], we mainly study a CNN-based model, *i.e.*, Res50 [36], and a transformer model, *i.e.*, SwinV2 [46].

B. Comparison with the State-of-the-art

In Tab. I, we show the statistics of dataset generated by our Vehicle-Diff and other existing vehicle re-ID datasets. We observe that our pipeline could synthesize more high-fidelity images with more identities, *i.e.*, 4 times larger number of IDs compared with VehicleX [9]. It is worth noting that our proposed Vehicle-Diff could further generate more images, if more text prompts are provided. In Tab. II, Tab. III and Tab. IV, we compare our proposed Vehicle-Diff with existing vehicle re-ID methods on three real-world datasets, *i.e.*, VeRi-776 [56], VehicleID [54] and CityFlowV2 [58], respectively. For a fair comparison, we follow the setting in



(a) Inter-class Discrepancy. (b) Intra-class Variance.

Fig. 3: Our pipeline reflects the fine-grained discrepancy between two appearance-similar vehicles, *e.g.*, front grilles, rear lights, and body types, while we also depict reasonable intra-class variations of the same vehicle, such as vehicle pose.

Method	Backbone	Data	Mix	Rank-1	Rank-5	mAP
VehicleX [9]	Res50	S	-	51.25	67.70	21.29
Vehicle-Diff	Res50	S	-	**57.87**	**74.97**	**22.21**
VehicleX [9]	SwinV2-B	S	-	66.87	79.80	28.33
Vehicle-Diff	SwinV2-B	S	-	**74.14**	**84.45**	**34.73**
VANet [60]	Res50	R	-	89.78	95.99	66.34
AAVER [61]	Res101	R	-	90.17	94.34	66.35
baseline (IDE [36])	Res50	R	-	92.73	96.78	66.54
VehicleX [9]	Res50	R+S	D	93.44	97.26	70.62
Vehicle-Diff	Res50	R+S	D	94.52	**97.97**	71.50
PAMTRI [8]	Dense121	R+S	D	92.86	96.97	71.88
SAN [62]	Res50	R	-	93.30	-	72.50
VehicleGAN [10]	Res50	R+S	D	93.60	97.30	74.20
CAL [63]	Res50	R	-	**95.40**	97.90	74.30
MSDeep [15]	Res50	R	-	95.10	-	74.50
VehicleX (PCB) [9]	Res50	R+S	D	94.34	97.91	74.51
Vehicle-Diff (PCB)	Res50	R+S	D	94.40	97.56	**75.45**
baseline	SwinV2-B	R	-	96.72	98.57	77.99
Vanilla Diffusion [29]	SwinV2-B	R+S	B	96.31	98.39	78.92
CLIP-ReID [64]	ViT-B/16	R	-	95.70	-	79.30
DCAL [65]	ViT-B/16	R	-	96.90	-	80.20
GiT [66]	GiT	R	-	96.86	-	80.34
TransReID [14]	ViT-B/16	R	-	96.90	-	80.60
PCL-CLIP [67]	ViT-B/16	R	-	97.10	98.60	82.50
CLIP-ReID [64]	ViT-B/16	R	-	97.40	-	83.30
VehicleX [9]	SwinV2-B	R+S	D	97.32	98.69	80.36
Vehicle-Diff	SwinV2-B	R+S	D	97.38	98.51	80.98
VehicleX [9]	SwinV2-B	R+S	B	97.08	98.81	81.39
Vehicle-Diff	SwinV2-B	R+S	B	**97.68**	**98.93**	**83.79**

TABLE II: Comparisons with the state-of-the-art methods on VeRi-776 [56]. "S" and "R" denote synthetic and real data, respectively. "B" indicates that each training batch selects equal amounts of synthetic and real data (as introduced in § III-D), whereas "D" indicates that synthetic and real data are combined randomly. Results on two backbones, *i.e.*, Res50 and SwinV2-B, are both reported.

Method	Backbone	Data	Mix	Small			Medium			Large		
				Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
VehicleX [9]	Res50	S	-	30.54	48.39	39.20	26.93	42.76	34.64	23.46	38.07	30.73
Vehicle-Diff	Res50	S	-	**40.84**	**59.99**	**49.78**	**36.63**	**55.72**	**45.50**	**31.72**	**50.14**	**40.45**
VehicleX [9]	SwinV2-B	S	-	42.44	60.85	46.86	39.77	57.19	43.89	36.96	54.18	41.00
Vehicle-Diff	SwinV2-B	S	-	**50.41**	**64.90**	**53.87**	**46.83**	**63.46**	**50.64**	**42.91**	**59.98**	**46.89**
RAM [68]	VGGM [69]	R	-	75.20	91.50	-	72.30	87.00	-	67.70	84.50	-
AAVER [61]	Res101	R	-	74.69	93.82	-	68.62	89.95	-	63.54	85.64	-
GSTE [51]	VGGM	R	-	75.90	84.20	75.40	74.80	83.60	74.30	74.00	82.70	72.40
IDE [36]	Res50	R	-	77.35	90.28	83.10	75.24	87.45	80.73	72.78	85.56	78.51
PRN [70]	Res50	R	-	78.40	92.30	-	75.00	83.00	-	74.20	86.40	-
SAN [62]	Res50	R	-	79.70	94.30	-	78.40	91.30	-	75.60	88.30	-
SAVER [71]	Res50	R	-	79.90	95.20	-	77.60	91.10	-	75.30	88.30	-
MSDeep [15]	Res50	R	-	81.20	95.40	84.30	78.00	91.80	81.00	75.60	89.30	78.60
CFVMNet [72]	Res50	R	-	81.40	94.10	-	77.30	90.40	-	74.70	88.70	-
VehicleX [9]	Res50	R+S	D	81.50	94.85	87.33	77.62	92.20	83.88	74.87	89.90	81.35
CAL [63]	Res50	R	-	82.50	94.70	87.80	78.20	91.00	83.80	75.10	88.50	80.90
VehicleGAN [10]	Res50	R+S	D	83.50	**96.50**	-	78.20	93.20	-	75.70	90.60	-
Vehicle-Diff	Res50	R+S	D	**83.87**	96.11	**89.23**	**78.60**	**94.21**	**85.23**	**75.93**	**91.40**	**82.62**
baseline	SwinV2-B	R	-	81.80	96.01	84.79	79.12	91.70	81.84	77.68	90.34	80.43
TransReID [14]	ViT-B/16	R	-	83.60	97.10	-	-	-	-	-	-	-
VehicleX [9]	SwinV2-B	R+S	B	82.12	96.29	85.18	78.91	91.74	81.71	77.78	90.71	80.58
Vehicle-Diff	SwinV2-B	R+S	B	82.73	96.75	85.72	79.54	92.43	82.29	78.02	90.95	80.82
VehicleX [9]	SwinV2-B	R+S	D	83.75	97.10	86.64	80.16	93.40	83.03	79.27	91.68	82.00
Vehicle-Diff	SwinV2-B	R+S	D	**84.37**	**97.52**	**87.17**	**80.55**	**93.92**	**83.47**	**79.42**	**92.29**	**82.22**

TABLE III: Comparisons with the state-of-the-art methods on VehicleID [54].

Method	Data	CityFlowV2			CityFlowV2→VeRi-776		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
VehicleX [9]	S	22.21	28.83	35.09	62.04	76.16	81.59
Vehicle-Diff	S	**26.38**	**33.09**	**36.54**	**66.81**	**77.18**	**83.61**

TABLE IV: Comparisons with competitive VehicleX [9] on CityFlowV2 [58] only using synthetic data.

the existing work [9] and utilize the same number synthetic image during the reID model training. As shown in Tab. II, Vehicle-Diff enables to achieve competitive vehicle re-ID accuracy on VeRi-776. This indicates that our proposed coarse-to-fine generation pipeline adapts well to vehicle re-ID, and enables to generate high-fidelity training images, even through our generative diffusion model is fine-tuned only with 1% of the unlabeled training data. Specifically, when the reID model is trained solely on synthetic data, our approach improves mAP by 0.92% compared with VehicleX on VeRi-776. When the reID backbone is switched to SwinV2-Base, we observe a consistent mAP improvement, *i.e.*, +6.40%. Furthermore, combined with the original real-world training set, our generated dataset can further improve the reID performance. In particular, our approach achieves 0.94% and 3.57% improvements in mAP compared with VehicleX and PAMTRI, respectively, when jointly trained with the original VeRi-776 training set in Res50 backbone [73]. For SwinV2-Base reID backbone, our method shows a consistent improvement. In VeRi-776 dataset, Vehicle-Diff ourperforms VehicleX by 0.62% on mAP when using random combination strategy ("D" in Tab. II) and 2.4% on mAP when using balanced sampling combination strategy ("B" in Tab. II). Notably, Vehicle-Diff achieves a 0.6% increase in Rank-1 accuracy over VehicleX, whose Rank-1 is already high at 97.08%. This increase is non-trivial. Besides, compared with other state-of-the-art methods, Vehicle-Diff also shows competitive performances. Our Vehicle-Diff method achieves 97.68% Rank-1 and 83.79% mAP, which surpasses CLIP-ReID [64] of 97.40% Rank-1 and 83.30% mAP. Similarly, for the VehicleID dataset, Vehicle-Diff shows competitive performance in Tab. III. In CityFlowV2, Vehicle-Diff outperforms VehicleX by 4.17% on Rank-1 and 4.26% on Rank-5 (see the left section of Tab. IV). We further conduct experiments to evaluate the generalization capability of Vehicle-Diff. As shown in the right section of Tab. IV, we apply the reID model trained on synthesized source-domain data to assess performance on the target domain. Notably, only images from CityFlowV2 were used for fine-tuning the generative model, without any label information. Despite this, Vehicle-Diff consistently outperforms VehicleX.

We further evaluate the quality of the generated data through both quantitative and qualitative evaluation. For the quantitative assessment, we utilize the Frechet Inception Distance (FID) [74], a widely recognized evaluation metric. Unfortunately, since the PAMTRI dataset is not publicly

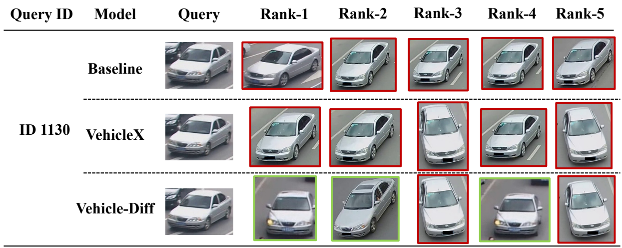


Fig. 4: Qualitative retrieval results. Here we compare our method with both our baseline and VehicleX. The ranking list is presented in descending order from left to right based on the similarity score. The images in red boxes are false-matched, whereas the green ones are true-matched.

Method	FID↓ VeRi-776	CityFlowV2
VehicleGAN [10]	233.0	-
PTGAN [11]	231.1	-
VehicleX	88.20	77.87
Vehicle-Diff	**44.84**	**54.84**

TABLE V: Quantitative comparisons on generated data quality. For a fair comparison, both Vehicle-Diff and VehicleX are trained on 1% images of VeRi-776.

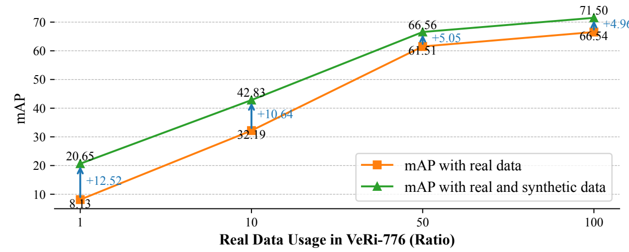


Fig. 5: Synthetic data provides notable mAP improvements, especially when the amount of real training data is small.

Components DFT	SF	#IDs	#Imgs	Rank-1	mAP	FID
		5,305	191,720	33.19	8.26	126.24
✓		4,940	160,758	58.34	22.00	**44.35**
✓	✓	4,896	149,472	**58.76**	**22.33**	44.78

TABLE VI: Ablation study on components, *i.e.*, diffusion fine-tuning (DFT) and semantic filtering (SF).

Baseline	#IDs	#imgs	Rank-1	Rank-5	mAP
IDE	4,894	45,338	57.87	**74.97**	22.21
	4,896	149,472	**58.76**	74.43	**22.33**

TABLE VII: Ablation study on the number of synthetic images for training the reID model on the IDE baseline.

available, we are unable to calculate its FID score. To ensure a fair comparison, we randomly selected 1% of the training datasets to train VehicleX and generate sample images. As shown in Tab. V, Vehicle-Diff achieves a lower FID score compared to all other generative methods. For qualitative comparison, we visualize the sample outputs of competitive generative methods in Fig. 1. The images in the first row are from the real-world dataset, while the images in the remaining five rows are from different synthetic data pipeline based on both 3D engines and GAN. We could observe that Vehicle-Diff produces images that are visually closer to the real-world dataset while keeping the fine-grained texture.

C. Ablation Studies and Further Discussion

Effectiveness of the coarse-to-fine strategy. Here, we evaluate the effectiveness of each component in our coarse-to-fine generation pipeline. Although the filtering process has minimal impact on visual quality and the Fréchet Inception Distance (FID) change after fine-tuning is negligible, the reID model performance shows consistent improvement (see Tab. VI). Tab. VI validates that quality matters more than quantity, and Tab. VII shows that more high-quality data leads to better results.

Effectiveness of the balanced sampling strategy. Previous methods, such as VehicleX and PAMTRI, typically conduct random sampling on mixed real and synthetic data to train the model. As a by-product of our pipeline, we introduce a balanced sampling strategy. We merge two mini-batch samples from real and synthetic datasets as a new mini-batch for training. We find that our balanced sampling strategy improves model learning on both VehicleX and Vehicle-Diff data. As shown in the last four rows of Tab. II, compared to the vanilla sampling strategy, our balanced sampling strategy yields a +1.06% boost in mAP for VehicleX and +2.81% boost in mAP for Vehicle-Diff.

Retrieval visualization. As shown in Fig. 4, we conduct the qualitative image retrieval comparison on VeRi-776. Our method has successfully recalled the target vehicle in the top-5 of the ranking list, surpassing the same model trained on real data or VehicleX. It is because that our Vehicle-Diff contains a large number of vehicle images with fine-grained attributes and intra-class variances such as camera angle, facilitating the discriminative feature learning (see Fig. 3). Therefore, the model trained on our Vehicle-Diff is able to handle challenging matches with fine-grained differences and significant camera angle variations.

Limited real data? To evaluate the effectiveness of Vehicle-Diff under limited real data conditions, we systematically reduce the amount of real data used. Specifically, we train the baseline using 1%, 10%, 50%, and 100% of the VeRi-776 dataset, each mixed with our synthetic data (see Fig. 5). The results show that synthetic data significantly enhance representation learning, particularly when real data is scarce.

V. CONCLUSION

In this paper, we study the state-of-the-art text-to-image synthetic data for vehicle re-identification (reID). We introduce Vehicle-Diff, a novel coarse-to-fine cross-modality generation pipeline that creates a synthetic reID dataset using only 1% of unlabeled images from the original dataset, tailored to specific downstream tasks. Extensive experiments show that our pipeline significantly reduces the gap between synthetic and real-world data, thereby enhancing reID performance. Specifically, our method achieves a competitive 83.79% mAP on VeRi-776. Furthermore, we analyze the strengths and limitations of synthetic data across various settings and identify optimal strategies for its use. In the future, we plan to integrate 3D-aware framework [75] into our pipeline to further improve the data quality.

REFERENCES

[1] F. Rollo, A. Zunino, N. Tsagarakis, E. M. Hoffman, and A. Ajoudani, "Continuous adaptation in person re-identification for robotic assistance," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 425–431.

[2] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[3] P. Shyam, K.-J. Yoon, and K.-S. Kim, "Adversarially-trained hierarchical feature extractor for vehicle re-identification," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 400–13 407.

[4] Z. Lu, R. Lin, Q. He, and H. Hu, "Mask-aware pseudo label denoising for unsupervised vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4333–4347, 2023.

[5] A. Khurshudov, "The smart city conundrum: technology, privacy, and the quest for convenience," *Smart and Sustainable Built Environment*, 2024.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[7] A. Wang, D. Sato, Y. Corzo, S. Simkin, A. Biswas, and A. Steinfeld, "Tbd pedestrian data collection: Towards rich, portable, and large-scale natural pedestrian data," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 637–644.

[8] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 211–220.

[9] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 775–791.

[10] B. Li, P. Liu, L. Fu, J. Li, J. Fang, Z. Xu, and H. Yu, "Veniclegan: Pair-flexible pose guided image synthesis for vehicle re-identification," *arXiv:2311.16278*, 2023.

[11] C.-S. Hu, S.-W. Tseng, X.-Y. Fan, and C.-K. Chiang, "Vehicle view synthesis by generative adversarial network," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, and J. Zhao, "Msinet: Twins contrastive search of multi-scale interaction for object reid," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 243–19 253.

[14] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 013–15 022.

[15] Y. Cheng, C. Zhang, K. Gu, L. Qi, Z. Gan, and W. Zhang, "Multi-scale deep feature fusion for vehicle re-identification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1928–1932.

[16] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 2683–2693, 2020.

[17] Y. Xu, N. Jiang, L. Zhang, Z. Zhou, and W. Wu, "Multi-scale vehicle re-identification using self-adapting label smoothing regularization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2117–2121.

[18] J. Yu, J. Kim, M. Kim, and H. Oh, "Camera-tracklet-aware contrastive learning for unsupervised vehicle re-identification," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 905–911.

[19] B. Jiao, L. Yang, L. Gao, P. Wang, S. Zhang, and Y. Zhang, "Vehicle re-identification in aerial images and videos: Dataset and approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1586–1603, 2024.

[20] R. Kishore, N. Aslam, and M. H. Kolekar, "Patreid: Pose apprise transformer network for vehicle re-identification," *IEEE Transactions on Emerging Topics in Computational Intelligence (Early Access)*, pp. 1–12, 2024.

[21] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[22] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4492–4501.

[23] Z. Zheng, X. Wang, N. Zheng, and Y. Yang, "Parameter-efficient person re-identification in the 3d space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7534–7547, 2024.

[24] K. Chen, W. Chen, T. He, R. Du, F. Wang, X. Sun, Y. Guo, and G. Ding, "Tagperson: A target-aware generation pipeline for person re-identification," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 560–571.

[25] S. Xiang, D. Qian, M. Guan, B. Yan, T. Liu, Y. Fu, and G. You, "Less is more: Learning from synthetic data with fine-grained attributes for person re-identification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 5s, pp. 1–20, 2023.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[29] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv:2307.01952*, 2023.

[30] Midjourney, "Midjourney - official website," https://www.midjourney.com/, accessed: 2024-09-15.

[31] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi, "Is synthetic data from generative models ready for image recognition?" in *The Eleventh International Conference on Learning Representations*, 2023.

[32] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic data from diffusion models improves imagenet classification," *arXiv:2304.08466*, 2023.

[33] H. Ju and Z. Zheng, "Video2bev: Transforming drone videos to bevs for video-based geo-localization," *arXiv preprint arXiv:2411.13610*, 2024.

[34] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv:2112.10741*, 2021.

[35] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 14, no. 1, pp. 1–20, 2017.

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 868–884.

[37] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.

[38] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Advances in Soft Computing—AFSS 2002: 2002 AFSS International Conference on Fuzzy Systems Calcutta, India, February 3–6, 2002 Proceedings*. Springer, 2002, pp. 288–300.

[39] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang, *et al.*, "Rich human feedback for text-to-

image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 401–19 411.

[40] X. Lin, Z. Dai, A. Verma, S.-K. Ng, P. Jaillet, and B. K. H. Low, "Prompt optimization with human feedback," *arXiv preprint arXiv:2405.17346*, 2024.

[41] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, "Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching," in *European Conference on Computer Vision*. Springer, 2024, pp. 213–231.

[42] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 19 730–19 742.

[43] 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," *arXiv:2106.09685*, 2021.

[44] R. Sapkota, R. Qureshi, M. F. Calero, M. Hussain, C. Badjugar, U. Nepal, A. Poulose, P. Zeno, U. B. P. Vaddevolu, H. Yan, *et al.*, "Yolov10 to its genesis: A decadal and comprehensive review of the you only look once series," *arXiv:2406.19407*, 2024.

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[46] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 009–12 019.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[48] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization in vivo," *IJCAI*, 2020.

[49] D. Kinga, J. B. Adam, *et al.*, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, vol. 5. San Diego, California;, 2015, p. 6.

[50] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[51] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2385–2399, 2018.

[52] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.

[53] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 562–570.

[54] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[55] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–31.

[56] X. Liu, W. Liu, T. Mei, and H. Ma, "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*. Springer, 2016, pp. 869–884.

[57] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8797–8806.

[58] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, C. E. Lopez, A. Sharma, Q. Feng, V. Ablavsky, and S. Sclaroff, "The 5th ai city challenge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 4263–4273.

[59] A. Kanaci, X. Zhu, and S. Gong, "Vehicle re-identification in context," in *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, September 10-12, 2018, Proceedings*, 2018.

[60] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8282–8291.

[61] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6132–6141.

[62] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 173–11 180.

[63] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1025–1034.

[64] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[65] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4692–4702.

[66] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, "Git: Graph interactive transformer for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 1039–1051, 2023.

[67] J. Li and X. Gong, "Prototypical contrastive learning-based clip fine-tuning for object re-identification," *arXiv preprint arXiv:2310.17218*, 2023.

[68] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: a region-aware deep model for vehicle re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[69] K. Chatfield, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[70] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3997–4005.

[71] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 369–386.

[72] Z. Sun, X. Nie, X. Xi, and Y. Yin, "Cfvmnet: A multi-branch network for vehicle re-identification based on common field of view," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3523–3531.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[74] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[75] W. Ma, Q. Liu, J. Wang, A. Wang, X. Yuan, Y. Zhang, Z. Xiao, G. Zhang, B. Lu, R. Duan, Y. Qi, A. Kortylewski, Y. Liu, and A. Yuille, "Generating images with 3d annotations using diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024.