ALICE BENCHMARKS: CONNECTING REAL WORLD OBJECT RE-IDENTIFICATION WITH THE SYNTHETIC

Anonymous authors
Paper under double-blind review

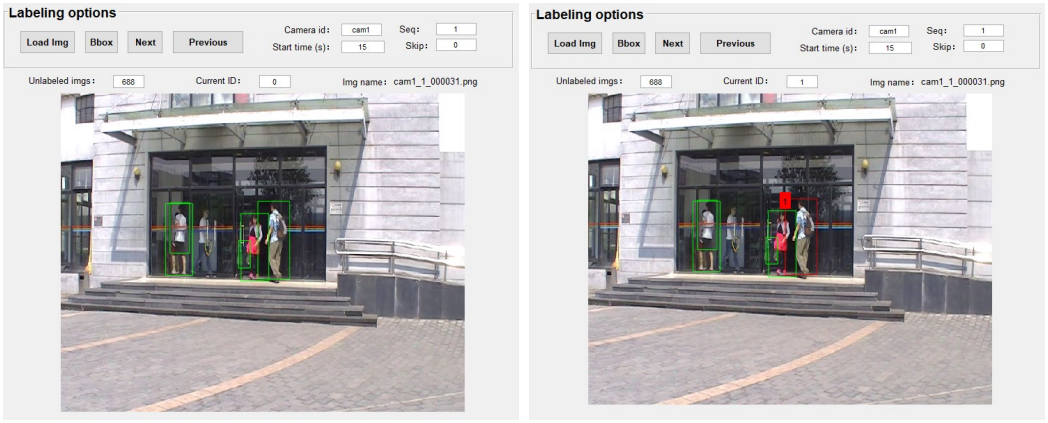
In the supplementary material, we present the summary of existing synthetic datasets (Section A), the dataset annotation procedure (Section B), sources of datasets and DA method (Section C and Section D), as well as discussion and future work (Section E).

A SUMMARY OF EXISTING SYNTHETIC DATASETS

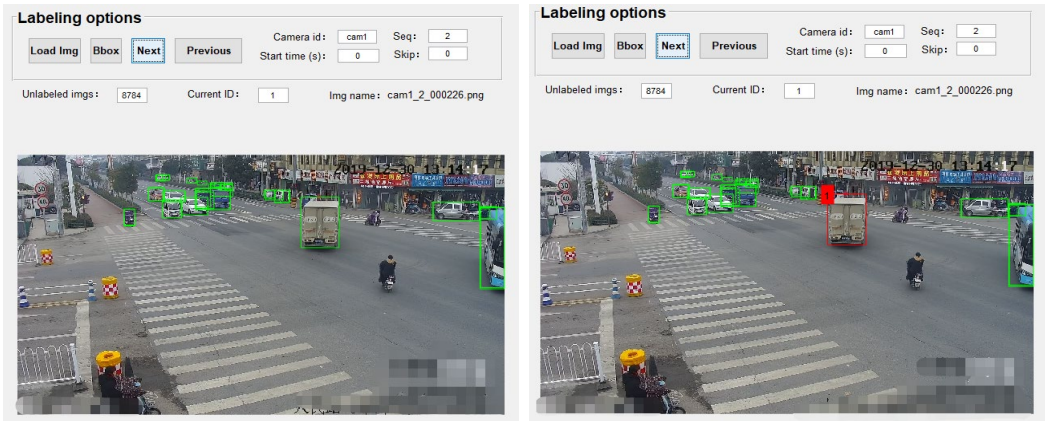
Table 1 presents some synthetic datasets published in recent years, designed for various purposes. This indicates synthetic data has emerged as a powerful tool, offering flexibility in data design and addressing challenges related to data scarcity and privacy concerns. In line with this, our research endeavors to push the boundaries of synthetic data in the field of object re-identification.

Table 1: Example synthetic datasets released in recent years.

name	task	data format	engine
SOMAset (Barbosa et al., 2018)	person re-ID	image	Unreal
SyRI (Bak et al., 2018)	person re-ID	image	Blender
PersonX (Sun & Zheng, 2019)	person re-ID	image/model	Unity
RandPerson (Wang et al., 2020)	person re-ID	image/video	Unity
UnrealPerson (Zhang et al., 2021)	person re-ID	image	Unreal
ClonedPerson (Wang et al., 2022)	person re-ID	image/model	Unity
WePerson (Li et al., 2021)	person re-ID	image/model	Script Hook V
GPR (Xiang et al., 2020)	person re-ID	image	GTA V
VehicleX (Yao et al., 2020)	vehicle re-ID	image/model	Unity
PAMTRI (Tang et al., 2019)	vehicle re-ID	image	Unreal
SYNTHIA (Ros et al., 2016)	segmentation	image	Unity
GTA5 (Richter et al., 2016)	segmentation	image	GTA V
SceneX (Xue et al., 2021)	segmentation	image/model	Unity
SAIL-VOS (Hu et al., 2019)	segmentation	image/video	GTA V
GCC (Wang et al., 2019)	crow counting	image	GTA V
G2D (Doan et al., 2018)	SfM	image/engine	GTA V
SDR (Prakash et al., 2019)	car detection	image/model	Unreal
CARLA (Dosovitskiy et al., 2017)	depth estimation & segmentation	image/model	Unreal
Sim4CV (Mueller et al., 2017)	navigation & tracking	image/model	Unreal
Virtual KITTI (Gaidon et al., 2016)	object tracking & detection & segmentation	image/video	Unity
synthetic2real (Peng et al., 2018)	classification & detection	image	CAD
VisDA (Peng et al., 2017)	classification & segmentation	image	CAD & GTA V
AI2-THOR (Kolve et al., 2017)	navigation &	image & model	Unity
RoboTHOR (Deitke et al., 2020)	navigation & tracking	image	Unity
Meta-Sim (Kar et al., 2019)	car detection	image	Unreal
SVIRO (Cruz et al., 2020)	interior vehicle & sensing	image	Blender
MultiviewX (Hou et al., 2020)	multi-view & detection	image	Unity



GUI for AlicePerson



GUI for AliceVehicle

Figure 1: **Graphical user interface (GUI) used in our annotation process.** AlicePerson (Top) and AliceVehicle (Bottom). Annotators work together according to the time of video recording. Each annotator is in charge of 1-2 cameras, looking for the same ID together.

B Data Annotation

Figure 1 displays the GUI, where **Top** and **Bottom** are examples for annotating different datasets. Annotators collaborate based on the video recording timestamps. Each annotator oversees 1-2 cameras, collectively identifying individuals with the same ID. After annotation, we conducted multiple rounds of checks on the data to ensure the accuracy of annotation. Our data is annotated by a data annotation company, which has a legitimate business license and meets the government's required wage levels (24 AUD / Hour).

Test set statistics are shown in Fig. 2. For AlicePerson, we can observe that the 4th camera captures a larger number of unique IDs compared to the other cameras from **A** of Fig. 2. Additionally, it is noticeable that most IDs appear in 4 or 3 cameras. For AliceVehicle, each camera includes about 300-800 images, and most IDs appear in 4 - 6 cameras. In the test set, we do not have ID only appears in one camera except for person on distractor images.

C Links to datasets

We make the real data we collected for Alice benchmarks publicly available on the links below:

Alice-train: https://drive.google.com/file/d/19sQdxFwF9LTmK8BjhINjWvp8o-UwjnRc/view?usp=sharing

Alice-validation-test: https://drive.google.com/file/d/1SAlSwfxUZ0QQCkja5BuBbVk6x0gOnkSh/view?usp=drive_link

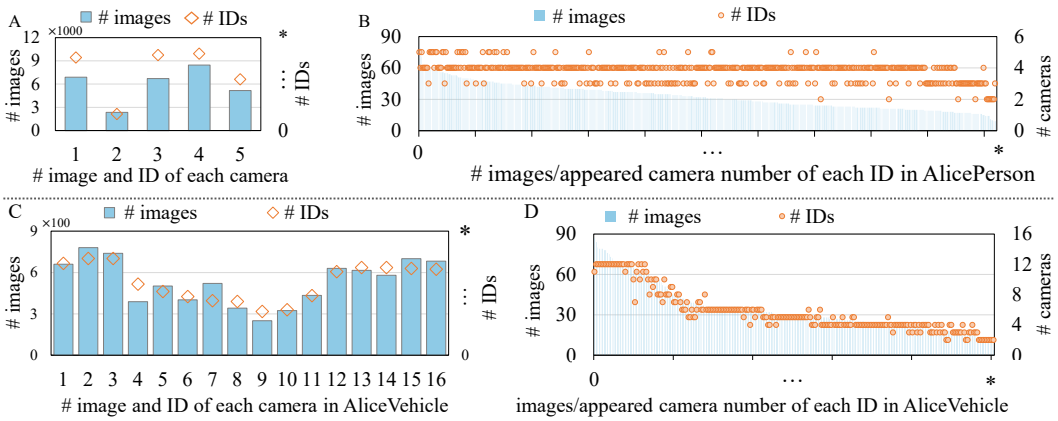


Figure 2: Statistics of the AlicePerson/AliceVehicle test sets. **A** and **C**: For each camera of AlicePerson and AliceVehicle, we display the number of images and object IDs that have appeared in that camera. **C** and **D**: For each ID of AlicePerson and AliceVehicle, we show the number of images and how many cameras the ID appeared. Note that the number of IDs in the test set is hidden for online evaluation. # is the notation of "the number of".

The synthetic data we reused is also available on these links:

PersonX: `https://github.com/sxzrt/Instructions-of-the-PersonX-dataset`

VehicleX: `https://github.com/yorkeyao/VehicleX`

D Source Code of DA methods

Attr. desc.: `https://github.com/yorkeyao/VehicleX`

IDE: `https://github.com/layumi/Person_reID_baseline_pytorch`

PCB: `https://github.com/layumi/Person_reID_baseline_pytorch`

CycleGAN and SPGAN: `https://github.com/Simon4Yan/eSPGAN`

PUL: `https://github.com/hehefan/Unsupervised-Person-Re-identification-Clustering-and-Fine-tuning`

ECN: `https://github.com/zhunzhong07/ECN`

UDA: `https://github.com/open-mmlab/OpenUnReID`

MMT: `https://github.com/open-mmlab/OpenUnReID`

E Discussion and Future Work

This section will further discuss open questions, challenges and interesting research problems for future investigations.

Understanding the domain gap between the synthetic and the real. The domain gap is usually studied with real-world datasets. For example, Torralba *et al.* (Torralba & Efros, 2011) analyze and summarize different types of domain bias, such as the selection bias (different collection sites, websites, *etc.*) and the caption bias (different viewpoints, resolutions ,*etc.*). In the context of "synthetic to real", the domain gap may be different from that in "real to real" DA. For example, the difference between 3D object models and real-world objects may create a domain gap that does not occur under the "real to real" setting. It is still largely unknown whether existing "real to real" DA methods can handle such new problems.

Also, simulated data has lower diversity than real-world data in complex environments. However, given that synthetic data can be simulated in large amounts, would it compensate for its lower data diversity? In this regard, an interesting question would be whether synthetic data makes an inferior

Table 2: FID values between target dataset and synthetic data with (Attr. desc) and without (Random) content-level domain adaptation.

Target datasets	Random	Attr. desc
Market-1501	104.78	83.24
AlicePerson	134.00	121.75
VeRi-776	80.22	57.45
AliceVehicle	78.05	56.07

source domain to real data. In other words, given the same target domain, which should we choose as a source: real data or editable synthetic data?

Designing "synthetic to synthetic" DA evaluation protocols for comprehensive evaluation of DA methods, dissecting the domain gap and providing references for "real to real" DA. The evaluation and understanding of DA methods are limited by the scale and variation of both the source and the target data. This problem can be overcome by controllable and customizable synthetic data. For example, we can generate source and target data with similar styles but very different content distributions for semantic segmentation. Conversely, we can synthesize data with similar content distributions but very different styles. With various data settings, we can conduct well-directed and comprehensive evaluations of domain adaptation methods, which will also be helpful for understanding the domain gap.

On the other hand, although quantitatively defining the domain gap caused by changes in various visual factors is rather infeasible, it is possible to analyze the domain gap in a higher capacity by using synthetic data. For example, the illumination of source data can be gradually changed for a fixed target set to analyze the influence of illumination differences on domain adaptation. Similarly, we can study the relative changes of various visual factors between the source and target domain, which may bring us a better and deeper understanding of the domain gap.

Need for new pseudo-label methods for object re-ID under more realistic settings. The object re-ID results in Tables 3 and 4 show that state-of-the-art pseudo-label methods, such as MMT, demonstrate high performance when using the Market-1501 and VeRi-776 as target sets, but do not work well on the Alice benchmarks (refer to Section 6.3). Given the decreased performance of pseudo-label methods, we should rethink the clustering strategies used for more realistic target sets. It would be interesting to study how to better leverage the camera style of the target set to assist with data clustering. Adopting this strategy, ECN yields the highest results for AlicePerson (Table 3).

New strategies to effectively use synthetic data to bridge the gap with the real world. This paper discusses several existing strategies such as pixel DA and the relatively new content DA. For the latter, there are still many open questions, *e.g.,* the design of new similarity measurements for distributions apart from the FID score used in (Yao et al., 2020) (Table 2 shows some results of the FID values). Moreover, the characteristics of a good training set are still largely unknown. In this regard, having a smaller domain gap between training and testing sets might not be the only objective. Other indicators such as appearance diversity or even noise are worth investigating. Furthermore, a variety of annotation types can be obtained from synthetic data, such as pixel-level semantics, bounding boxes and depth. These annotations may facilitate research in multi-task learning, which would also potentially improve real-world performance. In addition, it is an interesting task to explore how to combine simulation 3D models and real images in learning.

Can we finally get rid of real-world data when proposing new models? While state-of-the-art techniques in computer vision have been developed on real-world datasets, it would be interesting to study whether we can resort to pure synthetic data (training + testing) in model development, underpinned by the increasing ethics concerns in artificial intelligence (AI). Several challenges are yet to be resolved. Firstly, diverse, complex, and realistic synthetic data needs to be created. Secondly, it is necessary to confirm that training and testing on synthetic data are analogous to real-world performance. Thirdly, we need to explore the optimum composition of test data so as to comprehensively evaluate model performance. Apart from these challenges, there are many task-specific problems to be considered. At this point, it has not been determined how real-world test data can be replaced with synthetic data, but we will consider this to be an option with future higher-fidelity data generators and stronger theoretical support.

What is the relationship between the synthetic scene creator's control and its influence on the content gap with real-world data? Initially, human intervention in adjusting the content of simulated scenes can effectively minimize the content gap by exerting control over various factors, such as lighting conditions, object placement, and background variations. However, there are significant limitations, including the high cost associated with generating large-scale datasets. This constraint is a key driver behind the research community's exploration of machine-learning methods for automating content gap reduction.

Furthermore, content-level domain adaptation (DA) operates in a manner akin to manual control of visual attributes within scenes, although it remains far from perfect. This approach employs the Frechet Inception Distance (FID) to evaluate the quality of synthetic data and subsequently adjusts the parameters governing visual attributes in the simulation environment. This iterative process aims to progressively diminish the content gap and enhance alignment between synthetic and real data distributions.

We run experiments on a server that has 4 RTX-2080TI GPUs and a 16-core AMD Threadripper CPU @ 3.5Ghz.

REFERENCES

Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2018.

Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 167:50–62, 2018.

Steve Dias Da Cruz, Oliver Wasenmuller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3164–3174, 2020.

Anh-Dzung Doan, Abdul Mohsi Jawaid, Thanh-Toan Do, and Tat-Jun Chin. G2d: from gta to data. *arXiv preprint arXiv:1806.07381*, 2018.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Proceedings of the European Conference on Computer Vision*, 2020.

Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4551–4560, 2019.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3115–3123, 2021.

Matthias Mueller, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem. Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision*, 08 2017.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark for synthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *Proceedings of the International Conference on Robotics and Automation*, pp. 7249–7255, 2019.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision*, 2016.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2016.

Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 211–220, 2019.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the European Conference on Computer Vision*, 2011.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 8198–8207, 2019.

Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. *arXiv preprint arXiv:2006.12774*, 2020.

Yanan Wang, Xuezhi Liang, and Shengcai Liao. Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4900–4909, 2022.

Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. Unsupervised domain adaptation through synthesis for person re-identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2020.

Zhenfeng Xue, Weijie Mao, and Liang Zheng. Learning to simulate complex scenes for street scene segmentation. *IEEE Transactions on Multimedia*, 2021.

Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *Proceedings of the European Conference on Computer Vision*, 2020.

Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11506–11515, 2021.