VisDA 2022 Challenge: Domain Adaptation for Industrial Waste Sorting

Dina Bashkirova[*†], **Samarth Mishra**[*†], **Diala Lteif**[*†], **Piotr Teterwak**[*†], **Donghyun Kim**[*†], **Fadi Alladkani**[‡†], **James Akl**[‡†], **Berk Calli**[‡†], **Sarah Adel Bargal**[§†], **Kate Saenko**[*¶†]
Daehan Kim[‖], **Minseok Seo**[**], **YoungJin Jeon**[**], **Dong-Geol Choi**[‖]
Shahaf Ettedgui[††], **Raja Giryes**[††], **Shady Abu-Hussein**[††]
Binhui Xie[‡‡], **Shuang Li**[‡‡]

Editors: Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

Abstract

Label-efficient and reliable semantic segmentation is essential for many real-life applications, especially for industrial settings with high visual diversity, such as waste sorting. In industrial waste sorting, one of the biggest challenges is the extreme diversity of the input stream depending on factors like the location of the sorting facility, the equipment available in the facility, and the time of year, all of which significantly impact the composition and visual appearance of the waste stream. These changes in the data are called "visual domains", and label-efficient adaptation of models to such domains is needed for successful semantic segmentation of industrial waste. To test the abilities of computer vision models on this task, we present the *VisDA 2022 Challenge on Domain Adaptation for Industrial Waste Sorting*. Our challenge incorporates a fully-annotated waste sorting dataset, ZeroWaste, collected from two real material recovery facilities in different locations and seasons, as well as a novel procedurally generated synthetic waste sorting dataset, SynthWaste. In this competition, we aim to answer two questions: 1) can we leverage domain adaptation techniques to minimize the domain gap? and 2) can synthetic data augmentation improve performance on this task and help adapt to changing data distributions? The results of the competition show that industrial waste detection poses a real domain adaptation problem, that domain generalization techniques such as augmentations, ensembling, etc., improve the overall performance on the unlabeled target domain examples, and that leveraging synthetic data effectively remains an open problem. See https://ai.bu.edu/visda-2022/

Keywords: domain adaptation, semantic segmentation, AI for environment

1. Introduction

Efficient post-consumer waste recycling is one of the key challenges of modern society as countries struggle to find sustainable solutions to rapidly rising waste levels. World waste production is estimated to reach 2.6 billion tonnes a year in 2030, an increase from its current level of around 2.1 billion tonnes (Kaza et al., 2018). In the US, one of the leading countries in waste generation by volume, less than 35% of recyclable waste is being actually recycled (EPA, 2017), which leads to increased soil and sea pollution and is one of the major concerns of environmental researchers as well as the common public. One of the

[*] Boston University [†] Organizer [‡] Worcester Polytechnic Institute [§] Georgetown University [¶] MIT-IBM Watson Lab [‖] Hanbat National University [**] SI Analytics [††] Tel-Aviv University [‡‡] Beijing Institute of Technology

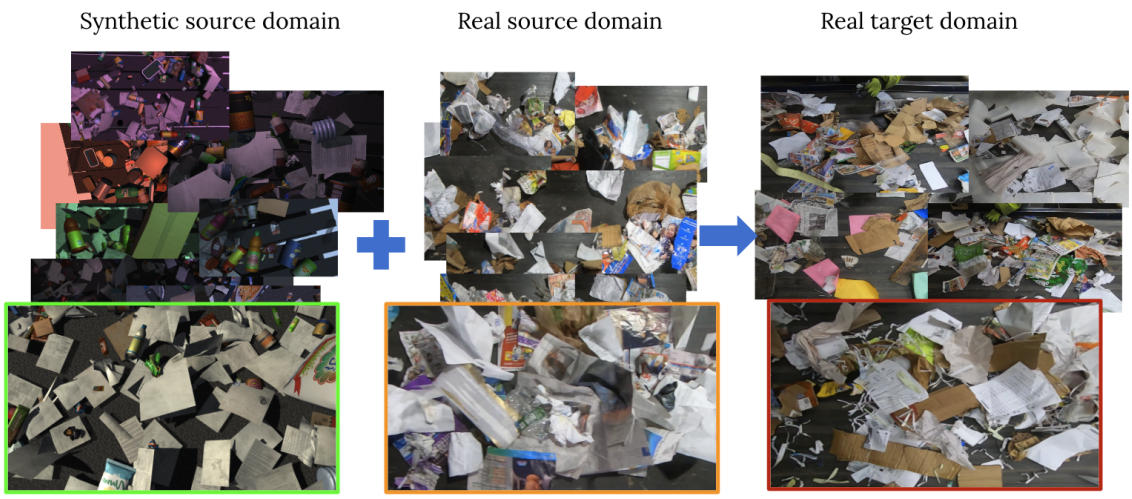

Figure 1: **Domain Adaptation for Semantic Segmentation of Recyclables:** Given a large and diverse labeled synthetic dataset **(left)** and a relatively small labeled real dataset **(center)** as source domains, the challenge task is to adapt the segmentation model trained on source data to the new unlabeled target domain **(right)** which introduces a natural domain shift as it was collected at a different location and season than the real source.

major challenges in recycling is waste composition analysis and sorting. In the US and many other countries, recyclable waste is sorted in material recovery facilities (MRFs). MRFs usually use special machinery to automatically sort recyclable waste on a conveyor belt according to the material type, however, they still heavily rely on manual sorting. As such, manual sorting is a mundane, physically demanding, and often dangerous task, as workers are exposed to sharp or contaminated objects on a daily basis. Therefore, an automated solution to aid waste sorting is necessary to make it both safe and profitable, and to ultimately solve the pollution problem.

Computer vision is instrumental for automating waste sorting since it enables segmentation of objects of various material types such as soft plastic, rigid plastic, metal and cardboard. Unfortunately, modern image segmentation models rely on large labeled datasets. It is extremely challenging to collect real in-the-wild waste stream images due to the disruption it causes to the facility's operation. Furthermore, the data annotation for this task involves pixel-level annotation and is prohibitively expensive. At the same time, the waste stream varies significantly by object appearance, season, location of the facility, as well as the sorting machinery used at a particular MRF, all of which introduce a significant natural domain shift during deployment and reduce segmentation accuracy. Therefore, *domain adaptation* methods that can adapt a model trained on a labeled *source* dataset to a novel *target* data stream without any additional labels are a promising approach for this problem.

While real data annotation is disruptive and expensive, unlimited amounts of data can be easily generated from 3D models using game engines like Unity, see Figure 1. Simulation promises to solve the limited and long tailed data problem, but models need to be adapted to an additional visual domain gap, that between non-photorealistic simulations and real images from the MRF. Inspired by the success of simulated training in self-driving applications (Richter et al., 2016; Geiger et al., 2012; Cordts et al., 2016), we propose a Sim2Real

challenge for industrial recycling. The task is to train a segmentation model on a source dataset consisting of a small amount of real data and a large amount of simulated data *and* achieve good results on a held-out real target domain. Algorithms can use the unlabeled target images to improve adaptation. We utilize two fully-labeled datasets for semantic segmentation of recyclable waste: the existing ZeroWaste dataset (Bashkirova et al., 2022), a novel ZeroWaste target dataset (from a different facility and season than the source), and a novel synthetic SynthWaste dataset designed according to the collection protocol of ZeroWaste. We also propose SynthWaste-aug, an augmented version of SynthWaste with instance-level texture augmentations for increased diversity.

Relationship to Previous Challenges. This challenge is the 6th iteration of the annual VisDA competition. This year, the organizing team consists of researchers from Boston University and Worcester Polytechnic Institute and the challenge was part of NeurIPS 2022 Competitions. A subset of our team also co-organized past VisDA competitions:

1. The 1st VISDA (ICCV 2017) proposed a single-source synthetic-to-real domain adaptation challenge for object classification and semantic segmentation, focusing on street-view data.
2. The 2nd VISDA (ECCV 2018) tackled synthetic-to-real open-set domain adaptation for object detection, where the target dataset contained examples of classes that were absent in the source domain.
3. The 3rd VISDA (ICCV 2019) introduced multi-source and semi-supervised domain adaptation settings of the DomainNet dataset (Peng et al., 2019) consisting of object classification in six domains (real, clipart, painting, drawing, infograph and sketch).
4. The 4th VISDA (ECCV 2020) for domain adaptive instance retrieval, where the target domain had a set of classes (instance IDs) novel with respect to the source domain.
5. The 5th VISDA (NeurIPS 2021) challenge studied a universal domain adaptation setup for object classification, in which the sets of classes in the source and target domains have a significant overlap but both source and target domains have classes that were not present in the other domain.

Our challenge is different from the previous iterations of VisDA, as it 1) includes the novel synthetic and real datasets for semantic segmentation of recyclable waste (see Figure 1), 2) proposes a setup in which synthetic data is used to improve the adaptation from one real dataset to another via supervised Sim2Real domain adaptation, as opposed to VisDA 2017 and 2018 that used only synthetic data as a labeled source domain, and 3) focuses on a challenging application of recyclable waste sorting, a problem that contains different types of distributional shift compared to prior domain adaptation setups, and thus is significantly different from the standard benchmarks.

Another line of work relevant to the proposed challenge is the simulation-to-real domain adaptation benchmarks, such as GTA (Richter et al., 2016) or SYNTHIA (Ros et al., 2016)-to-KITTI (Geiger et al., 2012) or Cityscapes (Cordts et al., 2016), that focus on the autonomous driving applications. (Bousmalis et al., 2018) also proposed to use Sim2Real domain adaptation to improve the quality of robotic grasping. These benchmarks also propose a challenging Sim2Real setup, but on tasks that are different from industrial waste sorting, and therefore, solutions developed for these datasets are not tailored to our task. In addition to that, they propose an unsupervised Sim2Real domain adaptation setup, whereas

we aim to leverage limited real supervision to minimize the domain gap and improve generalization to the unseen data in the real domain.

Another relevant challenge at NeurIPS is the AutoML for Lifelong Machine Learning (NeurIPS'18 competition) for continuous learning. Although this challenge also addresses the continuously changing data distributions, it is a lifelong AutoML setup that assumes a large-scale labeled dataset similar to the test sets (and in particular, during evaluation, test-set labels were *revealed* to the algorithm being evaluated after it made predictions on the most recent batch), whereas our challenge tackles the problem of unsupervised domain adaptation with a large randomized synthetic dataset and a smaller-scale real dataset as source domains.

2. Challenge Overview

2.1. Task

In this challenge, we propose a Sim+Real domain adaptation task, in which we provide fully-labeled data from two source domains: the novel large-scale synthetic **SynthWaste dataset** and a relatively small real **ZeroWaste dataset** for waste detection. The task at hand is to use these two datasets to adapt the segmentation model to the unlabeled real target domain (**ZeroWaste-v2**) that introduces a domain shift naturally occurring in the waste sorting application. Models have access to the target data during training. ZeroWaste-v2 is a novel dataset collected according to the ZeroWaste protocol at an MRF at a different location and season. An overview of the proposed task and datasets can be found in Figure 1.

2.2. Datasets

Our challenge was based on the following four datasets:

1. Real-world **ZeroWaste** dataset (http://ai.bu.edu/zerowaste/) (Bashkirova et al., 2022) is an open-access dataset for industrial waste sorting distributed under the Creative Commons Attribution 4.0 License. This dataset consists of $4,503$ fully annotated frames shot at a USA MRF during two hours of its operation. The frames are annotated with polygon semantic segmentation of 4 classes: cardboard, metal, soft plastic and rigid plastic. All other objects, including paper, as well as the conveyor belt, are labeled as background.

2. Synthetic **SynthWaste** dataset is designed specifically for this challenge to improve generalization and robustness to domain shifts. This dataset consists of 20990 procedurally generated frames of various recyclable objects randomly spawned onto a conveyor belt using Unity Development Platform that allows free usage for non-commercial purposes. The following simulation parameters are randomized: lighting type, intensity, direction and color, camera angle and position, level of clutter and overall distribution of object classes.

3. As there are style differences between synthetic and real data, we additionally provide a **texture-randomized** version of the synthetic **SynthWaste** dataset (SynthWaste-aug, see Figure 1). SynthWaste frames are augmented on the instance level using

the style transfer-based augmentation with Domain Aware Universal Style Transfer (DAUST) (Hong et al., 2021), which further increases visual diversity of waste objects. Objects in a frame are augmented using DAUST using random textures from the Flickr Material Database (Sharan et al., 2009) according to their material type.

4. We also collected a real-world **ZeroWaste-v2** dataset as target domain. This dataset consists of 7,720 frames collected according to the protocol of ZeroWaste, but in a different season (fall vs spring) and state in the USA (MA vs VT). We annotated 250 and 1004 frames for validation and final testing, respectively, and we provide 6,466 unlabeled frames for training. This novel dataset introduces a real-life domain shift typically occurring in industrial waste sorting.

2.3. Organization, Metrics and Baselines

Phases The competition consisted of two stages:

1. **Development (June 24 – September 30):** the labeled training and test sets of ZeroWaste, the SynthWaste and SynthWaste-aug datasets were released to the competitors along with the unlabeled ZeroWaste-v2 training set.
2. **Evaluation (September 30th – October 10th):** the test examples from ZeroWaste-v2 are released, and the teams were asked to submit the prediction results on the unlabeled ZeroWaste-v2 test set to our server, where the solutions are automatically evaluated.

Metrics and evaluation To evaluate the effectiveness of the competing solutions, mean intersection over union (mIoU), the standard semantic segmentation metric, was used to evaluate the performance on the test examples from the *target domain*. We used EvalAI (Yadav et al., 2019) for hosting our competition. The mean accuracy of per-pixel predictions (mAcc) was also reported.

Source-Only Baselines We evaluate baseline segmentation models trained only on source data (ZeroWaste-v1 or ZeroWaste-v1+SynthWaste datasets as stated in the second column of Table 1) and evaluated on the target data, without any domain adaptation. One such source-only baseline is the transformer-based **SegFormer** (Xie et al., 2021). Another is a convolutional network, **DeepLabv2** (Chen et al., 2017). In Table 1, we include the test results on the annotated test set frames from ZeroWaste-v2. Our results indicate that there is a significant domain gap between ZeroWaste-v1 and -v2 when the convnet-based DeepLabv2 is used as a backbone. Notably, the state-of-the-art transformer-based SegFormer is a stronger and more robust to this domain shift, with 10.41% source-only mIoU gap, in contrast to 16.32% gap with DeepLabv2. We note that fine-tuning the SegFormer model on SynthWaste slightly improves the overall mean accuracy and obtains similar mIoU as the original SegFormer model pretrained on ImageNet-1K (Deng et al., 2009). We observe that synthetic data improves performance on frequently occurring classes, such as cardboard and soft plastic.

Domain-Adaptive Baselines We used the state-of-the-art **DAFormer** domain adaptation method by (Hoyer et al., 2022a) trained either on ZeroWaste-v1 or on the combined data consisting of ZeroWaste-v1 and SynthWaste samples, as source domain, and the unlabeled examples from ZeroWaste-v2 as target domain. DAFormer uses the same visual

	train on	eval on	mIoU	mAcc
Source-only Baselines				
DeepLabv2	v1	v1	47.83	60.65
DeepLabv2	v1	v2	30.54	41.72
SegFormer	v1	v1	56.00	95.45
SegFormer	v1	v2	45.49	91.64
SegFormer	Synth+v1	v2	42.61	91.22
Adaptation Baselines				
DAFormer	v1+v2	v2	**52.26**	**91.20**
DAFormer	Synth+v1+v2	v2	48.31	90.63
Winning Solutions				
SI-Analytics (#1)	v1+v2	v2	**59.66**	**92.81**
Pros (HRDA) (#2)	v1+v2	v2	55.46	92.59
BIT-DA(PICO++) (#3)	v1+v2	v2	54.38	91.80

Table 1: Source-only, baseline domain adaptation results, and the results of the top-3 solutions with ZeroWaste-v1 (**v1**), ZeroWaste-v2 (**v2**), and SynthWaste (**Synth**) datasets. The source-only results of DeepLabv2 (Chen et al., 2017) and SegFormer (Xie et al., 2021) backbones show that while ZeroWaste-v2 introduces a domain shift that is significant for convnet-based DeepLabv2 architecture, features learned by SegFormer are more robust to this shift. The top submitted solutions are able to improve results significantly above our baselines.

transformer backbone as SegFormer. It is evident that the domain adaptation technique introduced in DAFormer improves the mIoU target (v2) domain w.r.t. the SegFormer source-only performance.

We also see that a naive baseline of training DAFormer on the combined SynthWaste and ZeroWaste-v1 reduces segmentation quality, likely due to a significant domain shift between the real and synthetic datasets. Therefore, the given baselines leave room for improvement, which is what we had hoped to achieve in the proposed challenge. As we will see below, none of the top solutions used the synthetic data, so this remains an open problem.

Materials and code We provide a starting kit that includes the data, data loaders and code to reproduce our baseline results at the start of the first phase of the competition. We also provide the executable used to generate SynthWaste to allow the participants to explore meta-learning approaches to further improve the synthetic data. All the code and materials provided can be found on our github page: https://github.com/dbash/visda2022-org

3. Results

In this section, we provide the main results of our challenge, including the participation statistics, the overview of the winning solutions, as well as the main takeaways. The top three solutions' results are summarized in Table 1.

Participation statistics 14 teams actively participated in the development phase of our challenge, with 314 submissions made total. In the evaluation phase, we limited the total

Method	background	rigid plastic	cardboard	metal	soft plastic	avg. mIoU
Baseline (DAFormer)	90.78	38.4	59.73	23.84	48.57	52.26
SIA_Adapt	**92.81**	**48.38**	**65.87**	**36.46**	**54.80**	**59.66**
HRDA	92.20	41.80	63.90	28.30	51.20	55.50
PICO++	91.36	44.35	61.71	31.24	43.25	54.38

Table 2: Per-category and average mIoU for models trained on ZeroWaste v1 and unlabeled frames from ZeroWaste v2 and evaluated on the ZeroWaste v2 **test** set.

number of submissions per team to avoid overfitting to the test set, and we received 40 submissions from 8 competing teams. Based on the results of the evaluation phase, we selected three winning solutions from teams SI-Analytics, Pros, and BIT-DA.

Reproducibility The results of the top-3 solutions according to the evaluation phase were tested and reproduced by the organizing team. The links to the code for reproducing the baseline and top-3 solutions can be found on our website.

3.1. First place solution: `SIA_Adapt`

The first place solution, `SIA_Adapt`, uses DAFormer (the baseline method) as the first step. There were two notable variations to DAFormer used by SIA_Adapt. First, the team found rare class sampling, which was used in training DAFormer, to be unhelpful for performance, and so they dropped it. Second, and more importantly, instead of an Imagenet-1K pre-trained transformer backbone, an Imagenet-22K pretrained ConvNeXt-L (Liu et al., 2022) backbone was used. This allowed the method to use a strong feature representation to build on top of and even without any target data at training time, perform at an impressive 56.4% mIoU on the target (ZeroWaste v2) test set. As another comparison, when the method was initialized with an Imagenet-1K pre-trained ConvNeXt-L backbone, it achieved 57.29% mIoU on the target test set, compared to the 59.66% mIoU (See Table 1) achieved with an Imagenet-22K initialization, thus isolating the effect that pre-training had on SIA_Adapt's performance. To better decouple the effects of the backbone architecture and pre-training, we conducted a study with the DAFormer baseline (See Appendix B).

With a trained DAFormer (including the modifications as described above), the method proceeds by pseudo-labeling the target and further self-training three different copies of this initial trained DAFormer, each using a different data-augmentation method (See Figure 2). Finally, these three networks are combined in a model soup (Wortsman et al., 2022), i.e., their weights are averaged, to obtain the final model.

3.2. Second place solution: `HRDA`

Overview: HRDA is a context-aware high resolution domain adaptation method for semantic segmentation (Hoyer et al., 2022b). The method comprises a multi-resolution training approach for UDA that combines small high-resolution crops and large low-resolution crops to preserve fine segmentation details as well as capture long-range context information. Predictions from both resolution crops are fused using a learned scale attention, which can enable adapting objects at the better-suited scale. As a backbone of their framework, this solution uses the DAFormer (Hoyer et al., 2022a) architecture that is based on a Trans-

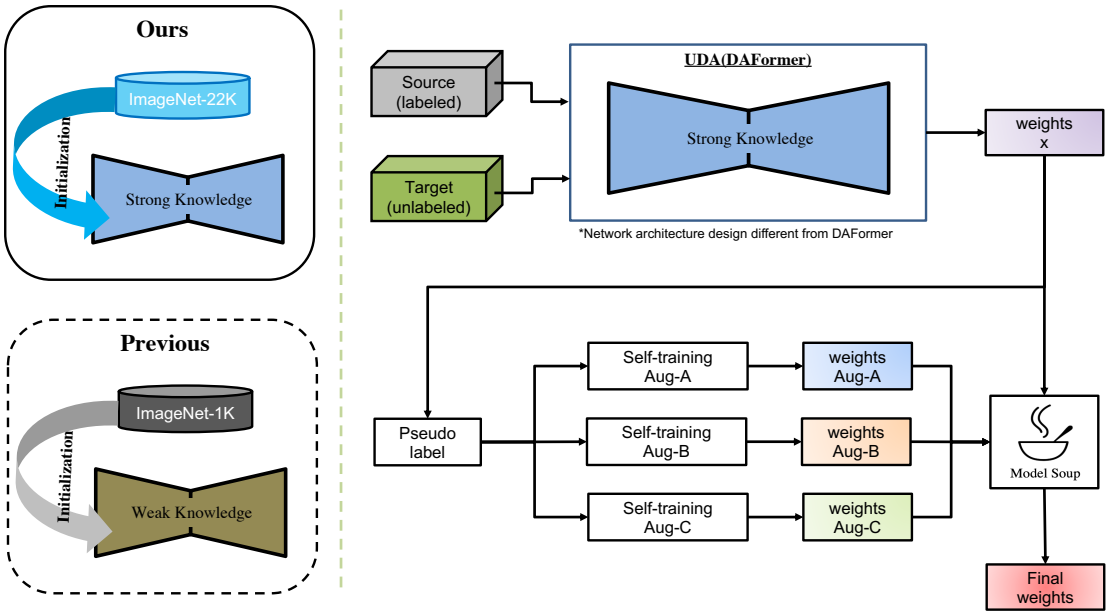

Figure 2: Overview of `SIA_Adapt`, the first place solution for VisDA-2022. The method uses a strong backbone initialization in the form of an Imagenet-22K pre-trained ConvNeXt-L and pseudolabeling. Also key to the method are self-training using different augmentations and using the resulting models together in a model soup.

former network which utilizes self-training. The latter uses pseudo labels generated by a teacher network to iteratively adapt the model to the target domain. Similar to the first place solution, the team concludes after an ablation study in Table 3 that rare class sampling (RCS) used to train DAFormer is ineffective for performance.

Results: Results of this solution are reported in Table 1, showing that HRDA yields a remarkable improvement in mIoU and mean accuracy compared to the source-only method. In addition, a detailed breakdown of the method's performance is reported in Table 2. The participating team also provides an ablation study with different source datasets and RCS configurations. In Table 3, the ablation study shows that training on the Zerowaste real-world dataset alone is enough to yield the best performance and that rare class sampling actually deteriorates it.

Source Dataset	RCS	Validation mIoU
Synthwaste	×	20.4
Synthwaste + Zerowaste	×	45.5
Synthwaste + Zerowaste, Equal Size	×	51.1
Zerowaste	✓	47.1
Zerowaste	×	**56.6**

Table 3: Ablation study by `HRDA`, the second place solution, examining different source data and the RCS (rare class sampling) configuration of the DAFormer backbone.

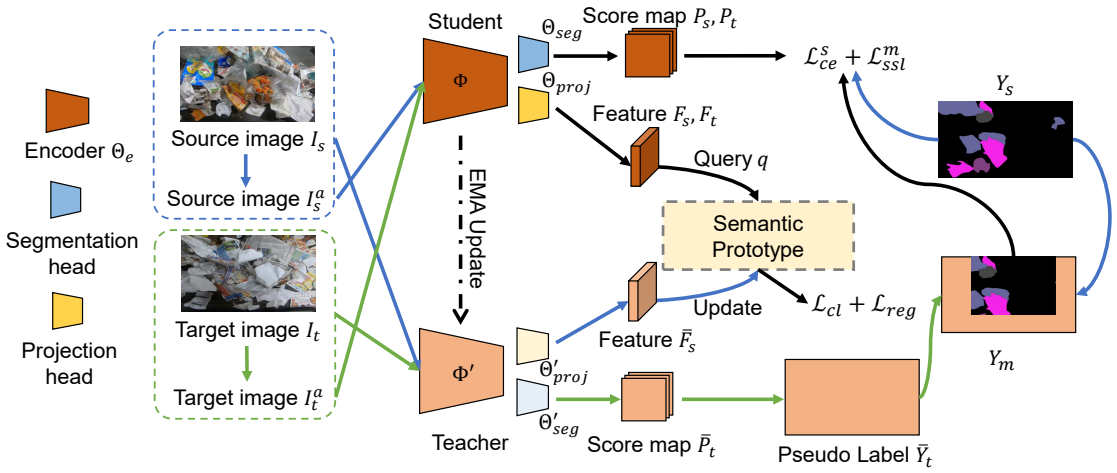

Figure 3: PICO++. A student-teacher framework, where the teacher is updated using EMA. The teacher then is used as a target data pseudo-labeler for both supervised contrastive and cross-entropy losses. Details are in Appendix A.2.

3.3. Third place solution: PICO++

Overview: PICO++ is a variant of SePiCo (Xie et al., 2023). The method composes a student-teacher architecture with learning signals from semi-supervised contrastive and semi-supervised cross-entropy losses. The student updates the teacher with EMA, while the teacher provides pseudo-labels for the student network to learn from target data. The samples are contrasted against class prototypes, which are computed from teacher representations. Different from other solutions, the contrastive loss provides explicit source-target domain alignment. Similar to other winning solutions, PICO++ is built on top of the very strong DAFormer (Hoyer et al., 2022a) architecture. The EMA update provides an implicit ensembling through model parameter averaging, which has been shown to improve domain generalization performance (Wortsman et al., 2021). For details, please see Appendix A.

Results: Results of PICO++ are reported in Table 1. PICO++ achieves substantial performance gains compared to the baseline method. As seen in Table 2, the gain is much greater for classes like rigid plastic and metal, implying that PICO++ has more potential in recognizing classes featuring a relatively regular shape. Nonetheless, a consistent improvement can be observed from all foreground classes, showing the effectiveness of PICO++.

3.4. Lessons learned

In this challenge, our goal was to investigate which domain adaptation and generalization techniques are particularly efficient for the real-world scenario afforded by the waste detection application. Below, we summarize our key observations:

1. **A small real dataset is better than a large synthetic one.** One of the unique features of our challenge compared to the previous Sim2Real competitions was that we propose both the synthetic source domain data and a small-scale real labeled dataset. The main assumption is that SynthWaste has higher visual diversity but is

less realistic, and the small-scale ZeroWaste can be used to bridge the gap between the synthetic and real domains. Interestingly, the results from the winning solutions indicate that omitting the sim2real adaptation step and performing domain adaptation from ZeroWaste-v1 to ZeroWaste-v2 results in higher performance.

2. **Pretraining and backbone architecture matter.** The results of the first place solution from SI-Analytics in Sec. 3.1 show a significant boost in segmentation quality when using ImageNet22K-pretrained ConvNeXt-L model compared to an ImageNet1K transformer model used in DAFormer. Our analysis in Appendix B with these changes made in the DAFormer baseline indicate the extent to which each affects target segmentation performance.
3. **Ensembling-based techniques and image augmentations are efficient.** The winning solutions commonly use some form of ensembling and / or augmentation, which is shown to improve model generalization. For example, SI-Analytics used model soup (Wortsman et al., 2022) of models trained on data augmented with different kinds of augmentations; Pros used HRDA (Hoyer et al., 2022b) that fuses predictions at various resolutions with attention; BIT-DA used a student-teacher paradigm and update the student network with the exponential moving average of the teacher weight update which is an ensembling / regularization technique, and a new variant of a contrastive loss for the source-target domain alignment.
4. **DAFormer is a strong baseline.** Even though ZeroWaste-v2 introduces a significant domain shift w.r.t. ZeroWaste with a 17.29% mIoU (47.83 versus 30.54) performance gap with DeepLabv2, DAFormer proved to be a strong baseline, with only 4 out of 14 teams beating the baseline in the development phase, and only 3 out of 8 in the final evaluation phase.

4. Conclusion

In this paper, we introduced the VisDA 2022 Visual Domain Adaptation Challenge that focuses on domain adaptation for industrial waste sorting. We show that domain shift occurs naturally in the industrial waste detection, and propose a new domain adaptation setup in which a large-scale and diverse synthetic dataset is used alongside the small real dataset to adapt the segmentation model to the real target domain. We propose a novel synthetic dataset, SynthWaste, as well as ZeroWaste-v2 collected according to the protocol of ZeroWaste at a different location and time. Our goal in this challenge was to reach out to the computer vision community to investigate efficient solutions for pressing and socially important applications and to popularize one of the applications of AI for environment. The results of our challenge suggest that state-of-the-art generalization methods significantly improve the overall performance on the target domain. We believe that our challenge opens new avenues of research in the fields of domain adaptation, and increases awareness and popularity of environment-centered applications of computer vision.

Acknowledgements This work was partially funded by the NSF FW-HTF #1928477 grant. We thank Vitaly Ablavsky for his guidance and support with this project. We are grateful to Nataliia Pyvovar, Guile Domingo and Ahmed Mudawar for their tireless work on data annotation and generation.

References

Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. 2022.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4243–4250. IEEE, 2018.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

US EPA. National overview: Facts and figures on materials, wastes and recycling, 2017.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognitio*, 2012.

Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14609–14617, 2021.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022b.

Silpa Kaza, Lisa Yao, Perinaz Bhada-Tata, and Frank Van Woerden. *What a waste 2.0: a global snapshot of solid waste management to 2050*. World Bank Publications, 2018.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 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 conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: domain adaptation via cross-domain mixed sampling. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1378–1388. IEEE, 2021. doi: 10.1109/WACV48630.2021.00142. URL https://doi.org/10.1109/WACV48630.2021.00142.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. https://arxiv.org/abs/2109.01903.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.

Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. Evalai: Towards better evaluation systems for ai agents. *arXiv preprint arXiv:1902.03570*, 2019.

Appendix A. Solution Details

A.1. SIA_Adapt

Datasets SIA_Adapt uses ZeroWastev1 as the labeled source domain and ZeroWastev2 as the unlabeled target domain according to the VisDA 2022 challenge rule[1]. In training, SynthWaste and SynthWaste-aug is not used. As per the challenge instructions, ZeroWastev2 test set is used for final evaluation.

Training The model was implemented using the DAFormer official code[2][3], except for rare class sampling which was not used in SIA_Adapt. IN-22K pre-trained weights for ConvNeXt-L are publicly available[4]. An NVIDIA RTX8000 GPU was used for training and all hyperparameter tuning experiments. 40,000 iterations of training was done for the initial adaptive model and 10,000 iterations for the fine-tuned model.

Fine-tuninig Model soup recipe was used to combine model weights after self-training. However, no EMA (exponential moving average) was used for training the individual self-trained models. The 3 different augmentations used, each for a different self-trained model, were: `PhotoMetricDistortion` implemented by mmseg[5], `GaussNoise` and `RandomGridShuffle` implemented by albumentations[6].

A.2. PICO++

Student and Teacher Network Architectures: The student and teacher networks are identical in architecture; they are built on top of the very strong DAFormer architecture, which is also used as a baseline for the challenge.Additionally, an extra projection head is added to reduce dimensionality (512→256) for both the student and teacher. During training, the student is updated with loss gradients while the teacher is update with an exponential moving average (EMA) of student iterates.

Cross-Entropy Losses: There are two cross-entropy losses used to update the student network. The first is a standard cross entropy loss applied on (augmented) source samples, denoted as $\mathcal{L}_{ce}^{s}$. Then, augmented target samples are pseudo-labeled using the teacher network, and then mixed with a source sample, creating mixed image I_m^a. The target pseudo-label is also mixed with the source label, creating mixed label Y_m. Another cross-entropy loss is applied to the student with the resulting mixed image-label pair $\mathcal{L}_{ce}^{m}(I_m^a, Y_m)$. The ratio of mixed image predictions whose confidence exceed β, which is called λ_β, reweights the $\mathcal{L}_{ce}^{m}$. The final loss on mixed images is $\mathcal{L}_{ssl}^{m} = \lambda_\beta \mathcal{L}_{ce}^{m}$. This method follows the DACS (Tranheden et al., 2021) methodology and allows for self-training using unlabelled target data.

Contrastive Losses: In addition to the cross-entropy losses, a semi-supervised contrastive loss is used following SePiCo (Xie et al., 2023). First, the training is warm-started using

[1] https://ai.bu.edu/visda-2022/
[2] https://github.com/lhoyer/DAFormer
[3] https://github.com/dbash/visda2022-org
[4] https://github.com/facebookresearch/ConvNeXt
[5] https://github.com/open-mmlab/mmsegmentation
[6] https://github.com/albumentations-team/albumentations

cross-entropy losses for $T_w = 3000$ iterations. Then, using the source data, per-label gaussians are fit to teacher projection-head features. These per-label gaussians are used to create proto-types to contrast the student features against. For a single sample of class i, the contrastive loss is formulated as

$$\mathcal{L}_{cl}^{q_i} = -\log\left[\frac{exp(\frac{q_i^\top \mu_i}{\tau} + \frac{q_i^\top \Sigma_i q_i}{2\tau^2})}{exp(\frac{q_i^\top \mu_i}{\tau} + \frac{q_i^\top \Sigma_i q_i}{2\tau^2}) + \sum_{k=1,k\neq i}^{C} exp(\frac{q_i^\top \mu_k}{\tau} + \frac{q_i^\top \Sigma_k q_i}{2\tau^2})}\right] + \frac{q_i^\top \Sigma_i q_i}{2\tau^2}, \quad (1)$$

where q_i represents a query feature q of i^{th} class, and C is the class number, τ denotes the smoothing factor common in contrastive learning. μ_i and Σ_i are the mean and covariance of the distribution within the prototype from i^{th} class. For source samples, ground truth class labels. For target samples, pseudo-labels are used. This contrastive loss ensures cross-domain feature alignment. For the derivation of this loss, please see the SePiCo paper (Xie et al., 2023), section 3.3.2.. Finally the regularization term is formulated as:

$$\mathcal{L}_{reg}^{\bar{q}} = \frac{1}{C\log C}\sum_{k=1}^{C}\log\frac{e^{\bar{q}^\top \mu_k/\tau}}{\sum_{l=1}^{K} e^{\bar{q}^\top \mu_l/\tau}}, \quad (2)$$

where $\bar{q}$ is the average of all features for either source or target domain. This prevents collapse of unsupervised samples.

Overall loss: Overall, PICO++ trained with the following objective:

$$\mathcal{L} = \mathcal{L}_{ce}^{s} + \mathcal{L}_{ssl}^{m} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{reg}\mathcal{L}_{reg} \quad (3)$$

where λ_{cl} and λ_{reg} are constant weights.

Hyperparameters and training details: All models are trained on a single NVIDIA A100-SXM4-40GB. AdamW (Loshchilov and Hutter, 2019) with betas (0.9, 0.999) and a weight decay of 0.01 is used. The initial learning rate is set to 6e-5 for encoder and 6e-4 for decoder. Note that only the student model is optimized, and the teacher model is momentum updated by the student. DAFormer (Hoyer et al., 2022a) is followed to employ a learning rate warmup policy in the first 1,500 iterations, and set pseudo confidence threshold α to 0.968, momentum coefficient β to 0.999, respectively. The model is trained with a batch of two 640×640-cropped images for 40,000 iterations. To get better initialization of the distributions, contrastive learning is started from the 3,000th iteration, and merely update the class prototypes before this. For the weights before loss terms, they are simply fixed to $\lambda_{cl} = \lambda_{reg} = 1$. Similar to the multi-view scheme proposed in DACS (Tranheden et al., 2021), ColorJitter and GaussianBlur is applied only to the samples entering the student model, with a uniform possibility.

Appendix B. Effect of Backbone architecture and Pre-training

Since the first place solution SIA_Adapt used a strong initialization and architecture in the form of an Imagenet-22K pre-trained ConvNeXt-L backbone, in this section we decouple and isolate the effects of these via experiments using the DAFormer baseline. Table 4 shows

ConvNeXt-L backbone	**Imagenet-22K pre-training**	mIoU
x	x	52.26
✓	x	56.41
✓	✓	58.72

Table 4: Decoupling the effect of backbone architecture and pre-training on Zerowaste-v2 test performance for the DAFormer method.

the mIoU of DAFormer using different backbone architectures and initializations on the Zerowaste-v2 test set (a 'x' in a column means using the default setting in DAFormer—which corresponds to the MiT transformer backbone and Imagenet-1K pre-training respectively).