VisDA-2021 Competition: Universal Domain Adaptation to Improve Performance on Out-of-Distribution Data

Dina Bashkirova[1*] DBASH@BU.EDU
Dan Hendrycks[2*] HENDRYCKS@BERKELEY.EDU
Donghyun Kim[1*] DONHK@BU.EDU
Haojin Liao[5*] LIAOHAOJIN@163.COM
Samarth Mishra[1*] SAMARTHM@BU.EDU
Chandramouli Rajagopalan[*] CHANDUIYER.RAJA@GMAIL.COM
Kate Saenko[1,3*] SAENKO@BU.EDU
Kuniaki Saito[1*] KEISAITO@BU.EDU
Burhan Ul Tayyab[4*] BURHAN@SHIRLEYROBOTICS.COM
Piotr Teterwak[1*] PIOTRT@BU.EDU
Ben Usman[1*] USMAN@BU.EDU
[1]*Boston University,* [2]*UC Berkeley,* [3]*MIT-IBM Watson AI Lab,* [4]*Shirley Robotics,* [5]*BUPT*

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

Progress in machine learning is typically measured by training and testing a model on samples drawn from the same distribution, i.e. the same domain. This over-estimates future accuracy on out-of-distribution data. The Visual Domain Adaptation (VisDA) 2021 competition tests models' ability to adapt to novel test distributions and handle distributional shift. We set up unsupervised domain adaptation challenges for image classifiers and evaluate adaptation to novel viewpoints, backgrounds, styles and degradation in quality. Our challenge draws on large-scale publicly available datasets but constructs the evaluation across domains, rather than the traditional in-domain benchmarking. Furthermore, we focus on the difficult "universal" setting where, in addition to input distribution drift, methods encounter missing and/or novel classes in the test set. In this paper, we describe the datasets and evaluation metrics and highlight similarities across top-performing methods that might point to promising future directions in universal domain adaptation research. We hope that the competition will encourage further improvement in machine learning methods' ability to handle realistic data in many deployment scenarios. See http://ai.bu.edu/visda-2021/

Keywords: domain adaptation; out-of-distribution detection; dataset bias

1. Introduction

In machine learning, "dataset bias" happens when the training data is not representative of future test data. Finite datasets cannot include all variations possible in the real world, so every machine learning dataset is biased in some way. Yet, machine learning progress is traditionally measured by testing on in-distribution data: almost every new approach is trained and evaluated on i.i.d samples from the same original distribution. This traditional evaluation obscures the real danger that models will fail on new data distributions. For example, a pedestrian detector trained on pictures of people in a sidewalk could fail on

[*] Equal Contribution.

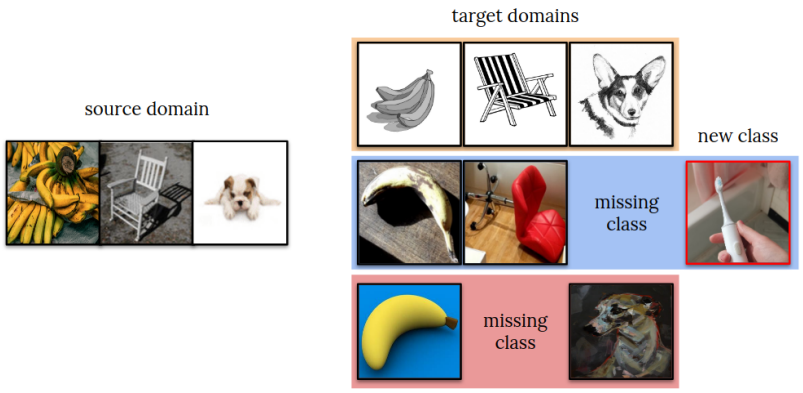

Figure 1: The **Universal Domain Adaptation** task addressed in our competition. Given a labeled dataset (source) and an unlabeled dataset (target), the task is to achieve the best possible performance on the target dataset. "Universal" refers to fact that, in addition to the input distribution shift, there will be a category shift between the source and target domains unknown a priori. The unknown category shift may include either all classes being shared, or missing classes in source, or missing classes in target, or both (Saito et al., 2020).

people outside of one, if the original data collection omitted them. A medical classifier could fail on data from a new hospital or a slightly different patient population. While deep neural networks have significantly improved performance on recognition tasks (Deng et al., 2009; Simonyan and Zisserman, 2014; Krizhevsky et al., 2012; Ren et al., 2015; He et al., 2017), they still suffer from poor generalization to out-of-domain data (Tzeng et al., 2014).

Domain adaptation techniques aim to adapt models and improve their performance on out-of-domain data. The **Unsupervised Domain Adaptation (UDA)** setting transfers models from a label-rich source domain to an unlabeled target domain without additional supervision. Recent UDA methods achieve this through unsupervised learning on the target domain, e.g., by minimizing the feature distribution shift between source and target domains (Ganin and Lempitsky, 2014; Long et al., 2015; Sun et al., 2016), classifier confusion (Jin et al., 2020), clustering (Saito et al., 2020), and pseudo-label based methods (Zou et al., 2018). Promising UDA results have been demonstrated on image classification (Long et al., 2018; Zou et al., 2019; Chen et al., 2019; Xu et al., 2019; Tang et al., 2020), semantic segmentation (Hoffman et al., 2016) and object detection (Chen et al., 2018) tasks.

Traditional unsupervised domain adaptation (UDA) methods assume that all source categories are present in the target domain, however, in practice, little might be known about the category overlap between the two domains. Ours is the first competition to address the more generally applicable **Universal Domain Adaptation (UniDA)** task. The task is as follows: given a labeled dataset (source) and an unlabeled dataset with unknown shift (target), achieve the best possible performance on the target dataset. "Universal" refers to fact that, in addition to the input distribution shift, there is an unspecified category overlap between the source and target domains. Specifically, the exact category overlap between source and target is unknown a priori and may include: full overlap, missing classes in source, missing classes in target, or both (Saito et al., 2020), as illustrated in Fig. 1.

Robustness or **Domain Generalization** is a closely related problem that involves performing well on OOD data without training on it. Common methods for this problem include clever augmentations of the training data (Hendrycks et al., 2020; Hendrycks* et al., 2020) and image stylization (Geirhos et al., 2018). Our UniDA task is similar, but allows model *adaptation* via training on the unlabeled target data, and also requires models to handle missing and novel classes in the target.

This is the fifth edition of the VisDA challenge on domain adaptation (see appendix for past competitions). Compared to previous editions, it focuses on a new task: universal domain adaptation (UniDA) where we have unknown overlap between the source and target classes. In past challenges, all target data was assumed to come from known categories. We also put emphasis on handling *multiple* types of realistic distribution shifts at once.

Why is the UniDA problem important? Distribution shift occurs in many, if not all, real-life application scenarios. When models are trained on offline datasets and deployed in the real world (e.g., a self-driving car, a hospital, etc.) the kinds of data they see inevitably shifts and changes relative to the static training distribution. Collecting more labelled training data in each new domain is not always feasible and interferes with the operation of the system. On the other hand, even very large labeled datasets like ImageNet have been found to have a distribution shift from similar datasets collected in the same way (Recht et al., 2019), so dataset bias is an inevitable problem with finite datasets.

Considering the ubiquity of the problem, there is a strong need for machine learning algorithms that generalize beyond their training data, or at least have the ability to adapt to novel distributions without requiring human supervision. Such algorithms would significantly impact many applications of machine learning. Our competition focuses on visual data (although the problem also occurs in other modalities), where applications could include: autonomous vehicles navigating in a new environment, a robot encountering objects in the real world and dealing with changing pose, lighting and other factors, or a medical imaging application receiving data from a novel facility, sensor or population.

2. VisDA-21 Competition

In this section we describe how we built the challenge dataset and measured the performance of participating methods. We also discuss rules and evaluation protocol we used to ensure fair competition.

2.1. Datasets

We used ImageNet (Russakovsky et al., 2015) as our **source domain**. This is a large-scale annotated dataset containing 1.4M images from 1,000 categories collected from the Web. While deep learning models work well on the test set in ImageNet, they can learn representations biased to incorrect texture cues (Geirhos et al., 2018) and often do not perform well on data which contains domain shift, including changes in artistic visual style, viewpoints, illumination, or even just re-collecting the test set (Recht et al., 2019).

The **target domains** in this competition contained images from the following sources:

- ObjectNet (Barbu et al., 2019) contains 50,000 images containing 313 object classes. Only 113 classes out of the 313 classes overlap with ImageNet. The dataset is both

easier than ImageNet – objects are largely centered and unoccluded – and harder, due to controlled variations in pose, background and viewpoints;

- ImageNet-R (Hendrycks et al., 2020) contains 30,000 images of the 200 classes in ImageNet (partial DA). The images contains different visual styles and textures;
- ImageNet-C (Hendrycks and Dietterich, 2019) contains the same validation images with 1,000 classes in ImageNet (closed set DA) but consists of 15 diverse corruption types, such as blur or noise, with different level of severity;
- ImageNet-O (Hendrycks et al., 2021) is a dataset built for out-of-distribution detection using imagenet models. It contains a set of 2000 images from Imagenet-22K which do not appear in Imagenet-1K, and are typically classified with high confidence by Imagenet-1K classifiers;
- ImageNet-G is a subset of Imagenet-val images distorted using a geometrize operation[1] illustrated in Figure 5 in the Appendix;
- Imagenet-val: We also included images from the original Imagenet-1K validation set.

2.2. UniDA Task

The objective of the competition was to leverage labeled data in the source domain and unlabeled data from the target domain to classify each target example as either a member of one of the source classes, or as an unknown class. Since our setting involves unsupervised domain adaptation, we allowed access to each unlabeled target domain during training. The domain label (i.e. which target sub-domain the image comes from) was not provided during the competition, but was used later to analyze results.

In accordance with the universal domain adaptation problem definition (Saito et al., 2020), both source and target domains had classes not present in the other domain. More concretely, we constructed the data to have all label overlap scenarios, including full class overlap (closed set), missing classes in the target (partial set), novel classes in the target (open set), and a mix of these (open-partial domain adaptation). For the images from classes only in the target domain, the task was to label them as a "novel" class not found in the source domain, without any further classification into categories.

2.3. Metrics

For each target example participating models predicted the most likely source *class label* and a *novelty score*, representing how likely the input is to be from an "unknown" target class not present in the source. We used two metrics to evaluate predictions and ranked models according to the average rank across the two metrics, as described below.

Area Under the ROC Curve (AUC). One desirable characteristic of a UniDA model is to be able to identify which target samples in the evaluation data belong to *novel* classes *not present in the source data*. To do this, we computed Area Under the ROC Curve (AUC) for the binary novelty detection task using the *novelty score* generated by participating methods.

1. https://github.com/fogleman/primitive

Split	Original Dataset	Classes		Images	
		ID	OOD	ID	OOD
Dev	Imagenet-C	1000	-	9080	-
	Imagenet-R	53	-	8224	-
	Imagenet-O	-	100	-	1034
	Objectnet	30	30	5026	5029
Test	Imagenet-C	1000	-	9094	-
	Imagenet-R	147	-	21776	-
	Imagenet-G	1000	-	9019	-
	Objectnet	83	170	26860	12069
	Imagenet-O	-	100	-	1036
	Imagenet-val	100	-	1000	-

Table 1: Development (dev) and Test Set Compositions. ID ("in-domain") are classes in source distribution, OOD ("out-of-domain") are classes out of source distribution.

Accuracy (ACC). We also computed the class label prediction accuracy, averaged over all samples belonging to classes present in the source data.

Final ranking. To obtain the final ranking for the leaderboard, we computed the ranking for each metric (ACC and AUC), took the average over these two rankings, and broke ties using ACC. For example, if Alice is 1st in terms of ACC and 3rd in terms of AUC (2.0 average rank), and Bob is 2nd in both (2.0 average rank), and Carol is 2nd in ACC and 3rd in AUC ranking (2.5 average rank), then Alice and Bob have a tie, but Alice wins because her ACC raking is higher, so the final ranking is: Alice (1st), Bob (2nd), Carol (3rd). This is consistent with the way prior challenges (Liu et al., 2020) ranked participants across multiple metrics.

2.4. Phases

The competition was run in two phases. During phase I, besides source images (Imagenet-1K) participants had access to a **development set** of images with class labels. This dataset is a combination of subset of Imagenet-C, -R and -O and ObjectNet introduced in Section 2.1. The number of images drawn from each datasets is summarized in the top half of Table 1. This set is meant to be used for model development and hyperparameter tuning. Participants could upload their model's predictions on the validation set to our evaluation server, and this information was used to populate entries on the validation leaderboard.

Phase II was a brief two week long period during which participants were provided the images from our test set without the ground truth labels. The content of the **test dataset** is summarized in the bottom half of Table 1. While the four out of five datasets we used to build the test set were also used in the validation set, we ensured that a significant domain shift between validation and test was present. First, the information about the content of the test split was **not** shared with participants. Second, different proportions datasets were used: while dev set contained predominantly images from ImageNet-C, the test set was for

Method	Position	Imagenet C+R+G	Objectnet		Imagenet val+O		Overall	
		ACC	ACC	AUC	ACC	AUC	ACC	AUC
Source-only*	Baseline	26.77	13.18	53.52	72.70	43.20	24.54	61.91
Ovanet*	Baseline	26.63	16.17	54.22	70.60	41.22	25.07	52.51
Tayyab *et al.*	1st	**60.22**	**40.99**	**64.24**	**84.00**	**90.02**	**56.29**	69.79
Rajagopalan	2nd	52.94	30.85	57.00	83.80	70.90	48.49	**76.86**
Liao *et al.*	3rd	52.66	31.89	55.15	82.20	59.21	48.49	70.80

Table 2: Test set performance of baselines and top-3 methods from the competition. We do not report AUC on C+R+G because it has no outliers. *Resnet-50 backbone.

the most part drawn from ObjectNet. Finally, we used non-overlapping subsets of outlier classes from ImageNet-O and ObjectNet in dev and test.

Rules. To ensure equal comparison, we did not allow training on any other external data or any form of manual data labeling. To encourage improvements in universal domain adaptation, rather than optimization of underlying model architectures, model size was limited to a total size of 100M parameters. During the test phase, participants were limited to a maximum of 5 uploads to the evaluation server to ensured that the participants could not exploit the test set to tune hyperparameters. Other rules are provided on our website.

Development Kit. The competition provided a development kit[2], containing links to the training, development and test data. It provided baseline code and instructions on how to submit results to the evaluation server, along with the evaluation code itself.

2.5. Baselines

We provide the two baseline methods: Source Only (SO) and OVANet (Saito and Saenko, 2021). SO is trained on the source data (*i.e.*, ImageNet); we directly use the pre-trained model provided in PyTorch (Paszke et al., 2017). OVANet is the state-of-the-art open-set domain adaptation method, which uses one-vs-all classifiers to detect outliers in the target domain. We tuned the hyper-parameters of OVANet on the validation set.

3. Competition Results

Over 140 teams registered on the evaluation server throughout the challenge, submitting over 170 predictions to the test leaderboard, summarized in the appendix Figure 6. Table 2 contains the test performance of the top-3 methods along with two baselines. Besides the overall accuracy and AUC on the test set, it also includes the performance of the methods on different partitions of the test set. Compared to the standard in-domain ImageNet task the accuracies are not very high, speaking to the difficulty of the problem of accurately determining image class under distribution shift while detecting outliers. However, the top ranked teams made impressive improvements to both ACC (56%, compared to 25% for baselines) and AUC (76% compared to 61% for baselines.) Implementations of the top

2. https://github.com/VisionLearningGroup/visda21-dev

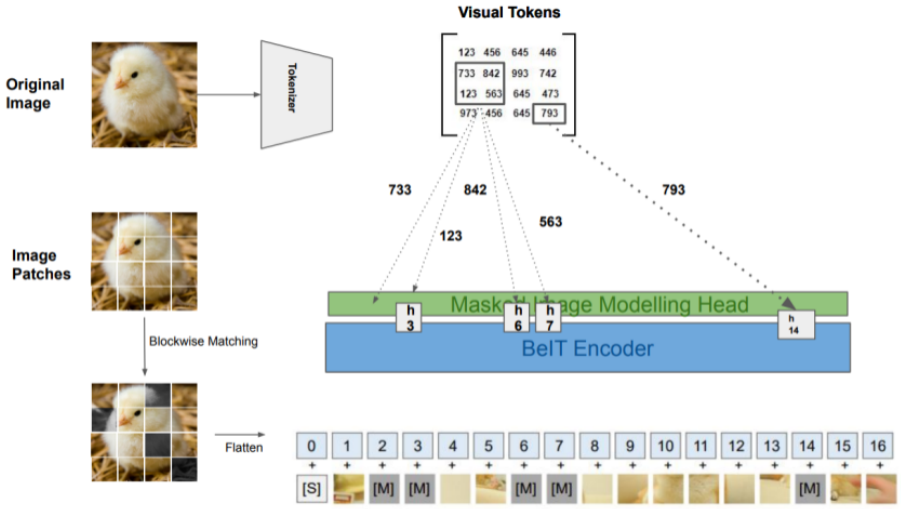

Figure 2: The first place solution from Tayyab and Chua (2021) used BEiT (Bao et al., 2021) self-supervised pre-training: a linear sequence of patches, with a random subset of patches masked out, is fed into a transformer along with their positional embeddings. The transformer encoder and a masking head are trained to predict labels of masked out patches generated using a DALL-E tokenizer.

three performing methods will be made available at our github solutions repo[3]. A brief overview of each approach is provided below.

3.1. First Place Solution

The first place solution from Tayyab and Chua (2021) used a transformer backbone pre-trained on a BERT-like (Devlin et al., 2018) self-supervised task. It was then fine-tuned for classification on the challenge source dataset. To generate anomaly scores, authors introduced an additional outlier class consisting of inlier images corrupted with random augmentations, and trained a classifier on the resulting dataset with 1000 inlier classes and one outlier class.

More specifically, the model was first pre-trained on the Imagenet-1K dataset using a self-supervised masked image modeling task. Authors used the backbone and pre-training strategy of BeIT-B (Bao et al., 2021) briefly described below. The backbone is a 12-layer transformer with 12 attention heads and a hidden dimension of size 786. In this stage, each input image was resized to 224x224 and then divided into a grid of 14x14 patches 16x16 pixels each. Each patch was tokenized using an off-the-shelf DALL-E tokenizer[4] (Ramesh et al., 2021) with a codebook of 8192 visual tokens. Similar to Bao et al. (2021), a linear sequence of patches with 40% of patches randomly masked out, along with their corre-

3. https://github.com/VisionLearningGroup/visda-21-solutions
4. https://github.com/openai/DALL-E

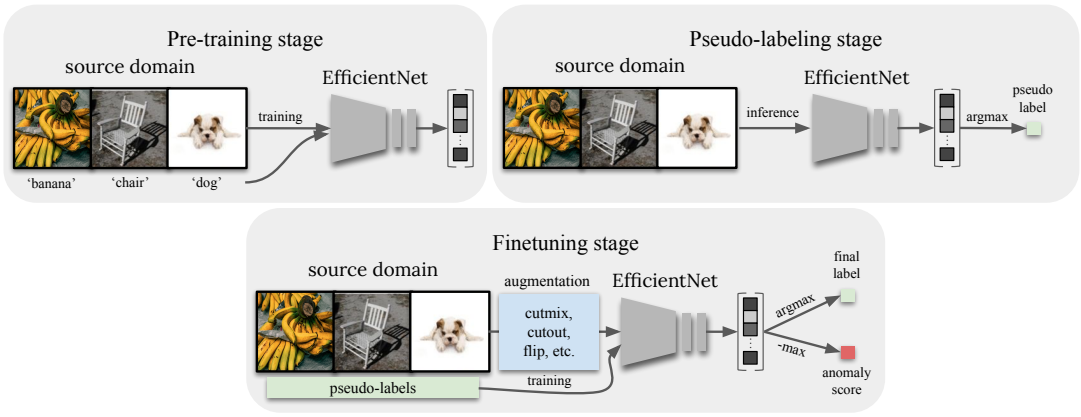

Figure 3: An overview of the second place solution from Rajagopalan. An EfficientNet-B7 convolutional backbone based classifier is trained in three stages. It is first trained on the source domain data without augmentations. This trained network is then used to generate soft labels for all source data which are used as the pseudo-labels in the final finetuning stage.

sponding positional embeddings, were passed to a transformer-based encoder and a masked modeling head, trained to predict masked-out visual tokens (see Figure 2).

Next, the masked modeling head was replaced with a classification head that was trained to classify images across 1000 ImageNet classes and a single outlier classes. The additional outlier class was generated by corrupting the same inlier images with random augmentations: random changes in hue and saturation and random cropping. The logit of the outlier class was used as the anomaly score. It is worth noting that the method did not use any unlabeled images from the target domain for training. For the first 400 epochs, an image size of 224x224 pixels was used. For the remaining 100 epochs, an image size of 384x384 pixels was used with the same patch size: the length of the patch sequence fed to the encoder increased, but the embedding size remained same. The final pipeline took 12 days to train on four Tesla V100s (32 GB each).

We acknowledge that, since a large proprietary dataset with 250 million samples (Ramesh et al., 2021) was used for unsupervised pre-training of DALL-E tokenizer, the first place solution used data outside of Imagenet-1K, and therefore technically violated the rules of the challenge. Nevertheless, we note that El-Nouby et al. (2021) showed (see Table 2 in that paper) that BeIT performs equally well with a random codebook instead of a DALL-E tokenizer, suggesting that pre-training on a large scale dataset is, in fact, not necessary to achieve high performance on ImageNet. We refer readers to our VisDA github solutions repo[3] with solutions for updated experiments with a random codebook.

3.2. Second Place Solution

The second place solution from Rajagopalan (2021), used an EfficientNet-B7 convolutional backbone (Tan and Le, 2019). The model was pre-trained and finetuned in two stages shown in Fig. 3. During pre-training, the model was trained on all source domain images (i.e. Imagenet-1K) without augmentations. Outputs of this trained model were then collected

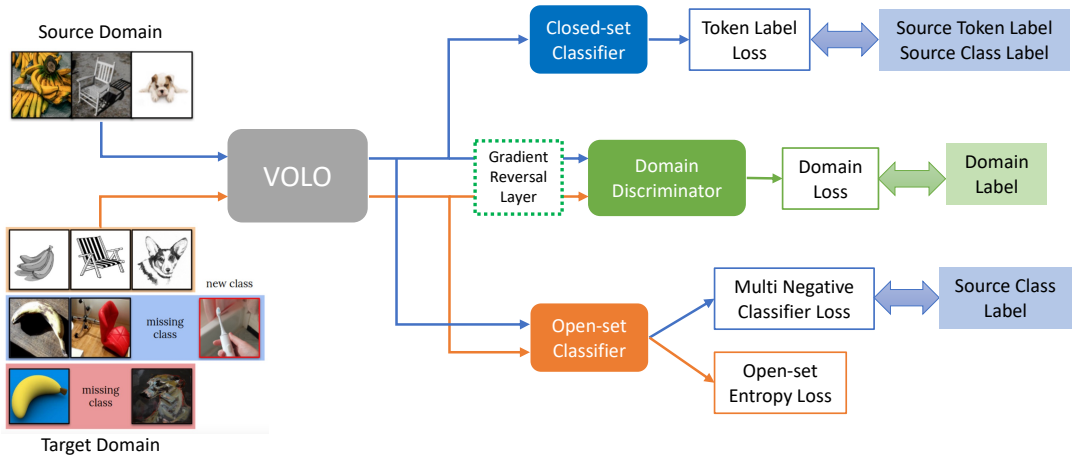

Figure 4: An overview of the third place solution from Liao *et al*. A VOLO model pre-trained on Imagenet-1K was used to initialize the network and then used in conjunction with a closed set classifier and an OVANet open-set classifier (Saito and Saenko, 2021).

to use as pseudo-labels in the final finetuning stage, as a form of teacher-student training to smooth labels.

Both training stages used a cosine learning rate schedule and model weights were updated using an exponential moving average with coefficient 0.9998. Pre-training used a cross-entropy criterion and fine-tuning used binary cross-entropy for each class output, both stages using label smoothing with a parameter 0.1. The finetuning stage involved cutmix (Yun et al., 2019) and mixup (Zhang et al., 2017). It also used dropconnect (Wan et al., 2013) for additional regularization.

During inference, this method predicted the inlier classes with the highest probability, and used the negative maximal probability across these inlier classes, $1 - \max_y p_y$, as the anomaly score, where p_y is the output probability for the class y. Like the first place solution, this method also did not make use of any target unlabeled data during training.

3.3. Third Place Solution

The third place solution from Liao et al. (2021) also employed a transformer backbone in the form of VOLO (Yuan et al., 2021). The participants used a VOLO model pre-trained on Imagenet-1K to initialize their backbone. They added one-vs-all classifiers to the backbone as in OVANet (Saito and Saenko, 2021) and trained the model in two stages.

The first stage optimized following objectives: (1) A cross-entropy loss from all the source labeled examples computed using the output of a closed-set classifier. (2) A domain discriminator loss (cross-entropy based on predicting the domain of the input), which is computed based on a linear domain discriminator. The linear discriminator was trained to minimize the loss, while the backbone was trained to maximize it. (3) Multi-negative classifier loss computed using the appropriate positive and a collection of "nearest negative class" one-vs-all classifiers. This loss is similar to the hard negative classifier sampling of

OVANet described by Saito and Saenko (2021) in Sec 3.1, with the difference being in the use of multiple hard negatives instead of just the one. (4) Open-set entropy minimization, which minimizes the average one-vs-all classification entropy over the different output classifiers, is also one of the criteria used in OVANet.

The second stage removes the domain discriminator and its accompanying loss term, and trains only to optimize the remaining three objectives. Finally, inference is done using the trained model by taking 5 crops of an input image and averaging the model's output on them. Similar to OVANet, the anomaly score for a given image is the negative class probability output by the one-vs-all classifier corresponding to the closest class according to the output of the closed-set classifier. A modular outline of the approach is in Fig. 4.

4. Discussion

Methods proposed throughout this competition made great advances in performance compared to the prior state of the art method OVAnet (Saito and Saenko, 2021) (See Fig 6 in appendix). These results, while impressive compared to the baseline, are still far from perfect, showing there is quite some room for improvement. This also suggests a new benchmark might be necessary to foster future research on universal domain adaptation. The challenging benchmark proposed in this competition serves as a good potential candidate.

The winners all used stronger backbones (transfomers or EfficientNet) than our baseline backbone (ResNet-50), which is evident in the in-distribution ImageNet val ACC results in Table 2 and very likely contributed to the improved out-of-domain peformance. The 1^{st} place solution far outperformed the runner-ups. This might point to the efficacy of the large transformer model and BEiT pre-training in overcoming dataset shift, but would need to be confirmed in ablations.

All three of our winning solutions made effective use of data augmentation techniques to improve both domain generalization and out-of-distribution detection. Additionally, the fact that two of our three winners had solutions based on transformer backbones, is an indication that they are effective tools for robustness to distribution shifts. This is supported by recent work from Bai et al. (2021), which finds that transformers are fundamentally better at generalization on target domain samples. The original report by Liao et al. (2021) provides careful ablations for their 3^{rd} place method and identifies components important for their method; they find that the domain discriminator, 5-crop inference, and token labelling are all are very useful. It's worth investigating whether these techniques could make the 1^{st} and 2^{nd} place solutions even more effective.

5. Conclusion

Generalization and adaptation to images that contain distribution shift and outliers is a hard problem for image classifiers. However our challenge revealed that new benchmarks can push the field to quickly improve on this task. While prior state-of-the-art performance was unimpressive on our benchmark, participant entries more than doubled accuracy and also significantly improved outlier detection. Methods used by VisDA21 participants indicate tools that are effective for this problem. Further experimentation with careful ablations should be used to probe the validity of these indications in future work, informing subsequent research about best approaches to similar problems.

References

Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? In *NeurIPS*, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32:9453–9463, 2019.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. 2019.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. 2014.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012.

Haojin Liao, Xiaolin Song, Sicheng Zhao, Shanghang Zhang, Xiangyu Yue, Xingxu Yao, Yueming Zhang, Tengfei Xing, Pengfei Xu, and Qiang Wang. 2nd place solution for visda 2021 challenge–universally domain adaptive image recognition. *arXiv preprint arXiv:2110.14240*, 2021.

Zhengying Liu, Adrien Pavao, Zhen Xu, Sergio Escalera, Fabio Ferreira, Isabelle Guyon, Sirui Hong, Frank Hutter, Rongrong Ji, Julio C. S. Jacques Junior, Ge Li, Marius Lindauer, Zhipeng Luo, Meysam Madadi, Thomas Nierhoff, Kangning Niu, Chunguang Pan, Danny Stoll, Sebastien Treguer, Jin Wang, Peng Wang, Chenglin Wu, Youcheng Xiong, Arbër Zela, and Yang Zhang. Winning solutions and post-challenge analyses of the ChaLearn AutoDL challenge 2019. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 17, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Chandramouli Rajagopalan. Final report. NeurIPS'21 VisDA Challenge Report, 2021. URL https://ai.bu.edu/visda-2021/assets/pdf/Chandramouli_Report.pdf.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811, 2019. URL http://arxiv.org/abs/1902.10811.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. *arXiv preprint arXiv:2104.03344*, 2021.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. 2020.

Burhan Tayyab and Nicholas Chua. Pre-training transformers for domain adaptation. NeurIPS'21 VisDA Challenge Report, 2021. URL https://ai.bu.edu/visda-2021/assets/pdf/Burhan_Report.pdf.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv*, 2014.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. 2019.

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv preprint arXiv:2106.13112*, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. 2018.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. 2019.