Appendix: Related Competitions

Members of our team have organized a similar Visual Domain Adaptation (VisDA) challenge at several computer vision conferences in recent years:

- The 1st VISDA (2017) challenge aimed to test domain adaptation methods' ability to transfer source knowledge and adapt it to novel target domains, focusing on Sim2Real transfer from a synthetic source domain to a real target domain. It featured an object classification and a semantic image segmentation track.

- The 2nd VISDA (2018) also tackled the Sim2Real problem, but featured object detection and open-set classification tracks.

- The 3rd VISDA (2019) promoted the multi-source and the semi-supervised domain adaptation settings on a newly collected 6-domain DomainNet dataset (Peng et al., 2019) (real, clipart, painting, drawing, infograph and sketch domains).

- The 4th VISDA (2020) focused on domain adaptive instance retrieval, where the source and target domains have completely different classes (instance IDs), for example, pedestrian IDs.

There have been several related competitions at NeurIPS. AutoML for Lifelong Machine Learning (NeurIPS'18 competition) addressed concept drift in lifelong learning, which is different from unsupervised domain adaptation. Inclusive Images evaluated classification on images drawn from geographic regions underrepresented in the training data. However they provided a labeled validation set from the target distribution, while our competition is focused on truly 'novel' distributions for which no labels were provided. Also, their distributional shift was more narrowly defined as a change in the geographic location where the image was collected by a user (via a crowdsourcing app).

Predicting Generalization in Deep Learning (NeurIPS'20 competition) invited competitors to design metrics that accurately predict the generalization performance of deep neural networks without using a test set. The competitors were asked to implement a function that takes a trained model and its training data, and returns a single scalar that correlates with the generalization ability of the model. Our competition was somewhat related in its desire to measure generalization, but addressed a very different task. We did not ask competitors to predict a given models' performance, but evaluate it's actual performance. We also considered novel domains rather than the same distribution of data.

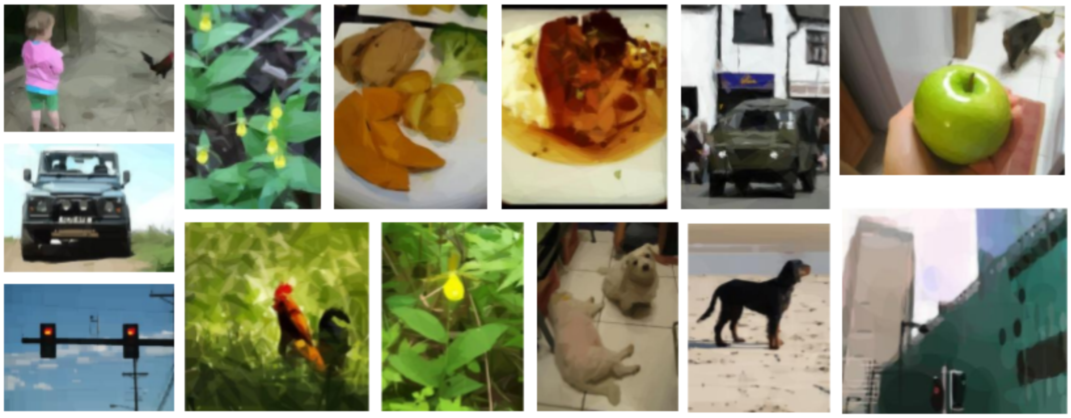

Figure 5: Examples from ImageNet-G we generated for the test phase of the challenge.

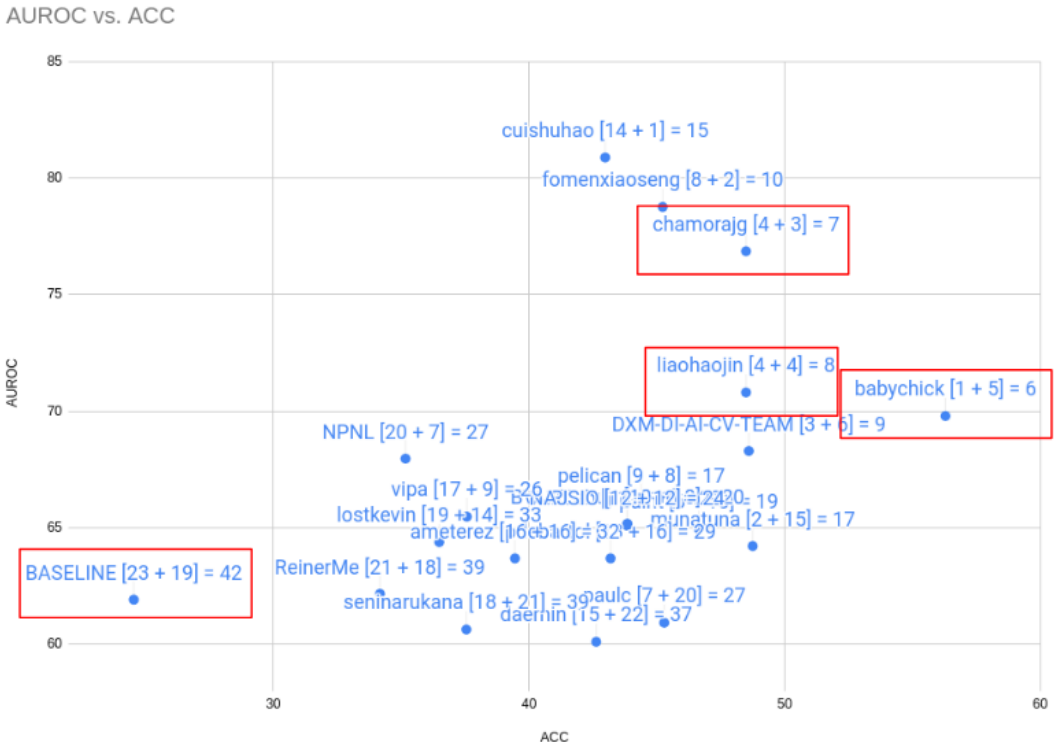

Figure 6: An overview of the final test leaderboard submissions made public by participants.