Using Robust Networks to Inform Lightweight Models in Semi-Supervised Learning for Object Detection: ATO-Action Results

Jonathan Worobey, Shawn Recker, and Christiaan Gribble
Applied Technology Operation
SURVICE Engineering
4695 Millennium Drive
Belcamp, MD 21017 USA
{jonathan.worobey,shawn.recker,christiaan.gribble}@survice.com

I. ATO-Action Dataset

ATO-Action is a video dataset capturing (usually) a single human subject near the center of the frame at various distances and from various angles. The subject performs aircraft handling signals including *forward*, *back*, *left*, *right*, *stop*, *wave off*, and *land*. For the purposes of this experiment (and for consistency with our experiments using the Okutama-Action dataset [1]), these ATO-Action classes are treated as a single class, *pedestrian*.

At 11,963 images and 12,957 *pedestrian* objects, ATO-Action has far fewer images (22%) and objects (4%) than does the Okutama-Action dataset. While the size of the datasets differs dramatically—especially in the number of objects—we use the same split ratios as in the Okutama-Action experiments and again divide the ground truth training set into validation, training, and unlabeled subsets. The size of each subset for each split ratio is shown in Figure 1, and Figure 2 depicts three ATO-Action exemplars.

II. Experiments

Section IV-B of the main text describes our full methodology; we summarize that methodology and highlight experimental differences here.

As with the Okutama-Action experiments, we perform five-fold cross-validation on each split ratio. For each cross-validation iteration, we train a robust model on a small training dataset. Following our proposed method, we then create a new inferenced dataset from the unlabeled data using the robust model. The inferenced datasets are then used to train both of our lightweight models, and we compare the accuracy of these models with models trained on both the full ground-truth dataset and the small hand-labeled dataset.

We train our robust models with most of the same hyperparameters as described in Section V of the main text. However, we reduce the number of training iterations from 100,000 to 50,000 for Faster R-CNN [2] with NASNet [3],

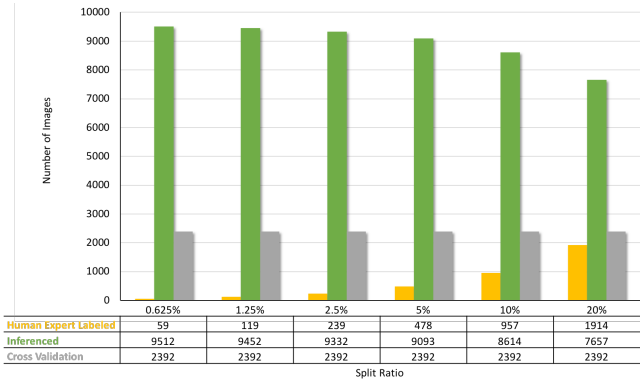


Fig. 1. *Split ratios used in our experiments.* We perform five-fold cross-validation on each dataset and split ratio. The relative sizes of ground-truth training dataset and inferenced dataset range from 0.625% to 20%, with absolute sizes as indicated.

as this model converges more quickly with the ATO-Action dataset. Likewise, we train our lightweight models as described in Section V of the main text, but we again reduce the number of training iterations for SSD [4] with MobileNetV2 [5] from 100,000 to 50,000, as this model also converges more quickly for ATO-Action.

III. Results

Figure 3 depicts the robust network training results. Here, the robust models train to acceptable levels of accuracy much more quickly and with less data than do the same models when trained on the Okutama-Action dataset. Recall that while a particular image may be considered *correctly inferenced*, this result guarantees only that each object in a frame is labeled with an intersection over union with the corresponding ground truth label of greater than 0.5 (as described in Section IV-B of the main text). Additional *human expert labeled* examples will likely lead to higher quality *correctly inferenced* examples and may further improve performance as the number of *correctly inferenced* examples starts to plateau—for example, at the larger split ratios of this experiment. (Though we do not

978-1-7281-4732-1/19/$31.00 ©2019 IEEE

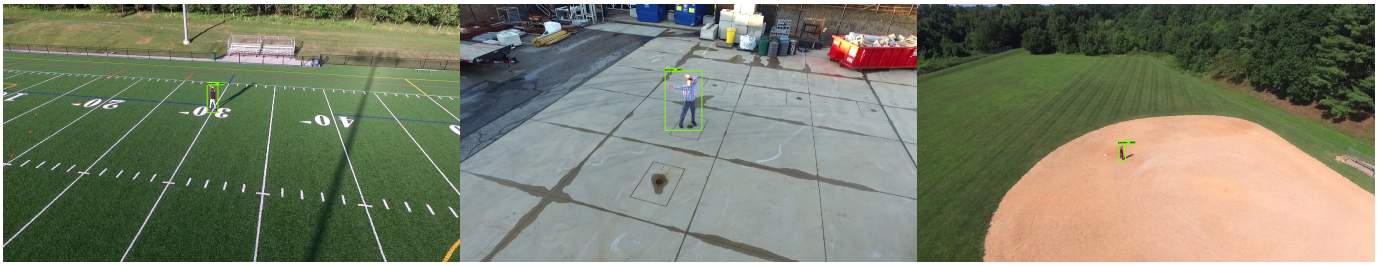
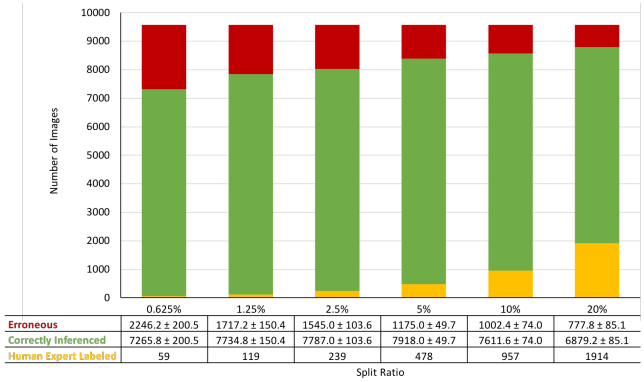


Fig. 2. *Dataset exemplars.* ATO-Action is an video dataset that tracks human subject (or *pedestrian* class) locations and actions from an aerial view. The training set consists of 11,963 images with (typically) a single person visible in each frame. In our context, we discard the action labels provided by this dataset and utilize only the objects' localizations instead.



	0.625%	1.25%	2.5%	5%	10%	20%
Erroneous	2246.2 ± 200.5	1717.2 ± 150.4	1545.0 ± 103.6	1175.0 ± 49.7	1002.4 ± 74.0	777.8 ± 85.1
Correctly Inferenced	7265.8 ± 200.5	7734.8 ± 150.4	7787.0 ± 103.6	7918.0 ± 49.7	7611.6 ± 74.0	6879.2 ± 85.1
Human Expert Labeled	59	119	239	478	957	1914

Fig. 3. *Inferenced dataset image distributions.* We train five robust models per split ratio. Each image of the inferenced dataset is considered correctly inferenced or erroneous. Erroneous images are handled in various ways, including ignoring, discarding, and replacing.

explicitly test this hypothesis, results with the lightweight models, which continue to improve after the combined *correctly inferenced* and *human expert labeled* example count plateaus, lend weight to its validity.)

As can be seen in Figure 4, both lightweight models show better performance improvement compared to the corresponding experiments with Okutama-Action—particularly at the smallest split ratios. Moreover, nearly every model using our proposed method sees improvement over the corresponding base model, which is in contrast to results with the Okutama-Action experiments. These results indicate that our proposed methods are better suited to the ATO-Action dataset across a wider range of split ratios.

We see nearly identical results compared to the Okutama-Action experiments in the *Human Expert Time Only*, *Ignore* (Table I(a), Row 1 and 2) curation speedup factors, with only slight differences in initial labeling and discarding time. In contrast, the results of *Human Expert Time Only*, *Replace* (Table I(a), Row 3) speedup factors are much better than those in the corresponding Okutama-Action experiments—this result is due to the significantly smaller number of erroneously labeled images that must be replaced by a human expert.

Finally, the speedup factors that include computer time (Table I(a) and (b)) are dramatically worse than the Okutama-Action experiment results; in this case, computer time dominates the total time necessary to execute our proposed method. Specifically, the time it takes to train a robust model, create the inferenced dataset, and finally train a lightweight model far outweighs the human time of creating the *human expert labeled* subset and fixing the inferenced dataset (via ignoring, discarding, or replacing). If computer time is valued equally to human expert time, dataset size becomes an important factor to consider when deciding whether or not to employ to our proposed method.

(a) Human Expert Time Only

	0.625%	1.25%	2.5%	5%	10%	20%
Ignore	166.13×	80.99×	39.75×	19.80×	10.19×	5.00×
Discard	29.83×	28.33×	21.39×	14.80×	8.80×	4.72×
Replace	3.89×	4.23×	4.14×	4.47×	3.77×	3.00×

(b) Faster R-CNN with NASNet and SSD with MobileNetV2

	0.625%	1.25%	2.5%	5%	10%	20%
Ignore	1.78×	1.76×	1.73×	1.66×	1.56×	1.37×
Discard	1.71×	1.70×	1.67×	1.62×	1.52×	1.35×
Replace	1.27×	1.30×	1.29×	1.32×	1.26×	1.17×

(c) Faster R-CNN with NASNet and SSD with Inception-v2

	0.625%	1.25%	2.5%	5%	10%	20%
Ignore	1.80×	1.78×	1.74×	1.68×	1.57×	1.37×
Discard	1.72×	1.72×	1.69×	1.64×	1.53×	1.35×
Replace	1.27×	1.31×	1.30×	1.33×	1.27×	1.18×

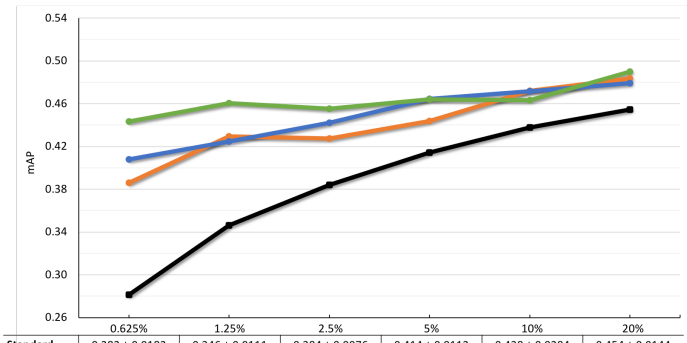
TABLE I
Dataset Curation Speedup Factors

REFERENCES

[1] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-Action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 28–35.

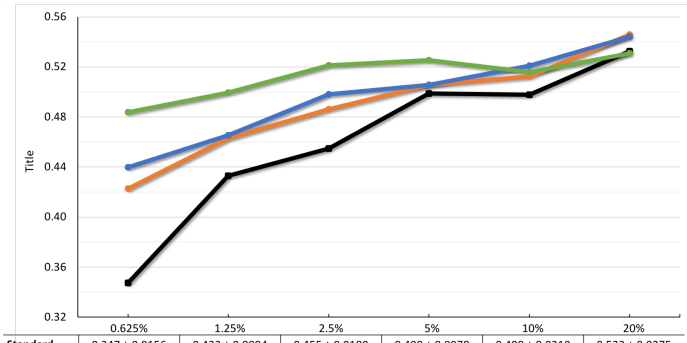
[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[3] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.



	0.625%	1.25%	2.5%	5%	10%	20%
Standard	0.282 ± 0.0182	0.346 ± 0.0111	0.384 ± 0.0076	0.414 ± 0.0112	0.438 ± 0.0284	0.454 ± 0.0144
Ignored	0.386 ± 0.0081	0.429 ± 0.0075	0.427 ± 0.0167	0.444 ± 0.0206	0.472 ± 0.0183	0.483 ± 0.0124
Discarded	0.408 ± 0.0126	0.425 ± 0.0116	0.442 ± 0.0183	0.464 ± 0.0123	0.472 ± 0.0206	0.479 ± 0.0288
Replaced	0.443 ± 0.0212	0.460 ± 0.0088	0.455 ± 0.0200	0.464 ± 0.0212	0.463 ± 0.0249	0.490 ± 0.0192

(a) *SSD with MobileNetV2*



	0.625%	1.25%	2.5%	5%	10%	20%
Standard	0.347 ± 0.0156	0.433 ± 0.0094	0.455 ± 0.0180	0.499 ± 0.0078	0.498 ± 0.0210	0.533 ± 0.0275
Ignored	0.423 ± 0.0224	0.463 ± 0.0048	0.486 ± 0.0165	0.506 ± 0.0128	0.512 ± 0.0171	0.546 ± 0.0101
Discarded	0.440 ± 0.0172	0.466 ± 0.0144	0.498 ± 0.0062	0.506 ± 0.0157	0.521 ± 0.0040	0.544 ± 0.0207
Replaced	0.484 ± 0.0118	0.499 ± 0.0376	0.521 ± 0.0114	0.525 ± 0.0094	0.516 ± 0.0319	0.531 ± 0.0239

(b) *SSD with Inception-v2*

Fig. 4. *Lightweight inference results.* Two different models, SSD with MobileNetV2 [5] (a) and SSD with Inception-v2 [6] (b), are evaluated against the three error-handling strategies, with the standard method shown for reference. Five-fold cross validation is used for each split level and each strategy. Both lightweight models show improvement across all three strategies at the smaller split ratios.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.