Learning to Generate Synthetic Data via Compositing

Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, Visesh Chari

amazon Lab126

Goal

Efficient task-aware generation of synthetic data by compositing images. Target-in-the-loop training results in 2.7-3.5% improvement in classification and object detection accuracy.

TERSE Data Generation

Task-aware
Generate synthetic data which is tuned to the task

Efficient
Generate fewer and more useful composite images compared to random data generation and data augmentation

Realistic
Generated synthetic data should be context-aware and realistic in order to minimize domain gap

Task-aware Efficient Realistic Synthesis of Examples

Our approach

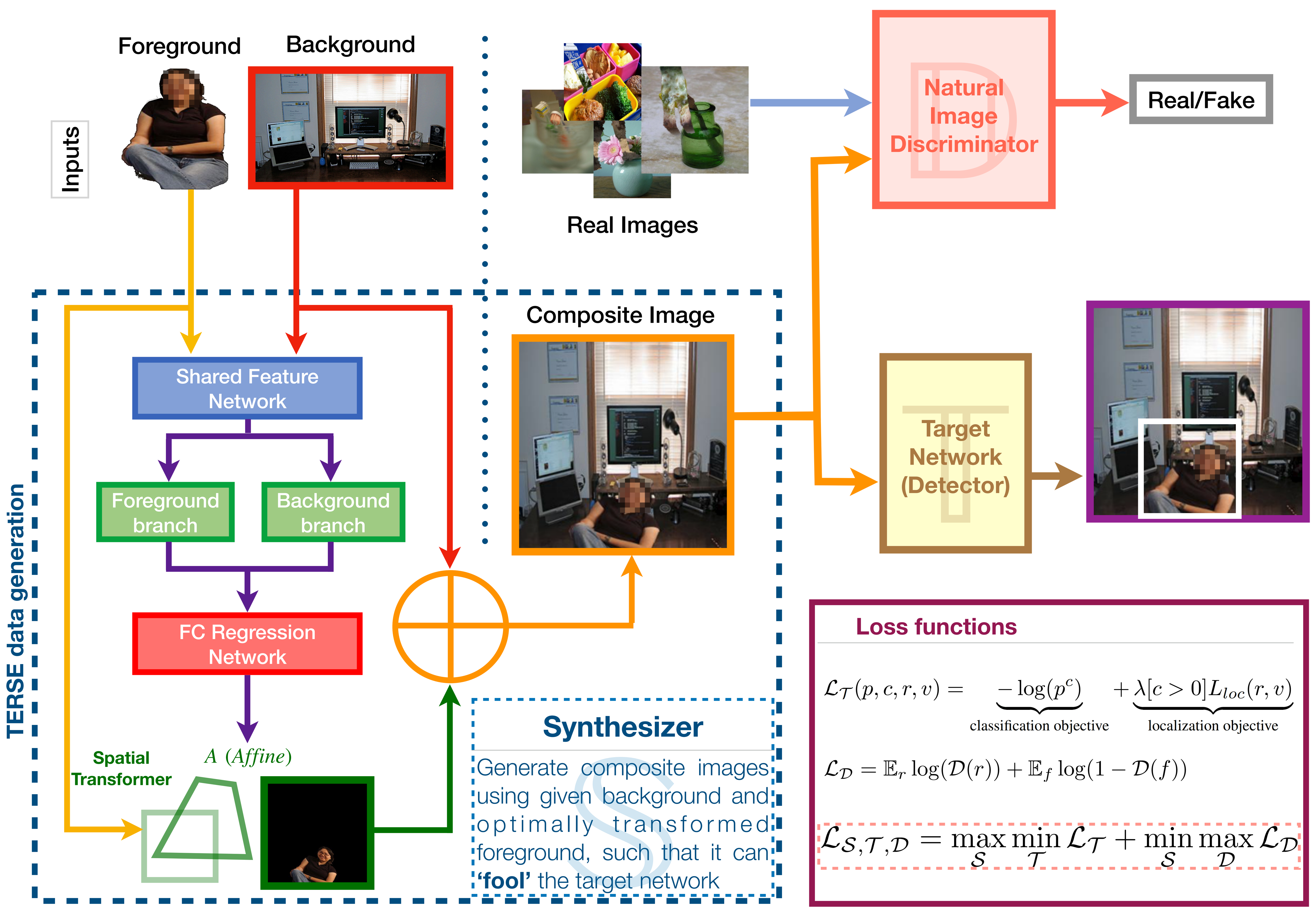
- Target-in-the-loop (TIL) training paradigm: Synthesizer and target network are trained iteratively, in lockstep
- Maximizing target network failure during compositing ensures efficient generation (EG) of synthetic data
- Implicit naturalness priors learnt by the discriminator result in context-aware (CA) generation
- Adding hallucinated artefacts in synthetic backgrounds achieves blending artefact robustness (BAR)
- Demonstrated applicability on image-classification, object/instance detection using various architectures while using <50% synthetic data compared to SOTA

Comparison with compositing works

Method	TIL	EG	CA	BAR
Cut-Paste-Learn [1]				⋆
Context-Data-Augmentation [2]			✓	⋆
Copy-pasting GAN [3]†				⋆
TERSE (Ours)	✓	✓	✓	✓

⋆ partial BAR using randomized blending/blurring
† does not generate synthetic data for training

[1] D Dwibedi, I Mishra, and Martial Herbert. Cut, paste and Learn: Surprisingly easy synthesis for instance detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
[2] N Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In IEEE European Conference on Computer Vision (ECCV), 2018
[3] R Arandjelovic, A Zisserman. Object discovery with a Copy-Pasting GAN. arXiv preprint arXiv:1905.11269 (2019)



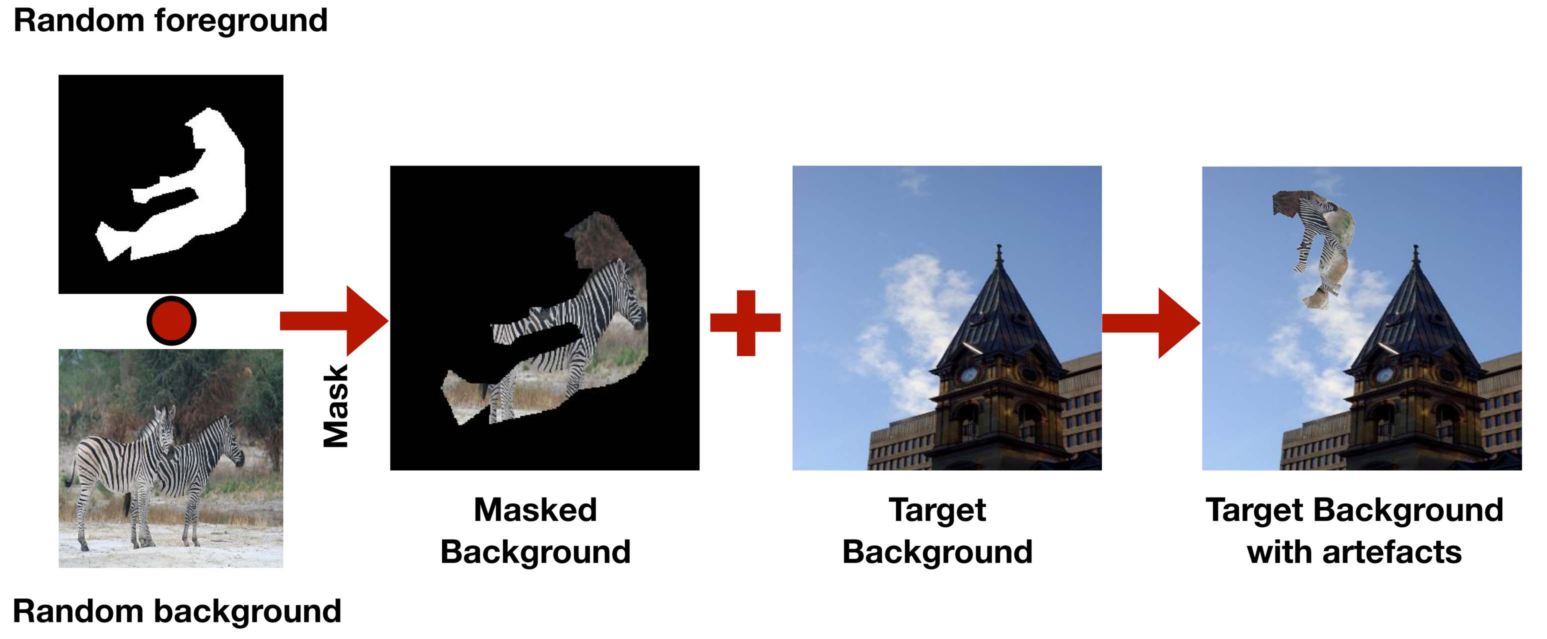
Loss functions

$$\mathcal{L}_{\mathcal{T}}(p,c,r,v) = \underbrace{-\log(p^c)}_{\text{classification objective}} + \underbrace{\lambda[c>0]L_{loc}(r,v)}_{\text{localization objective}}$$

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_r \log(\mathcal{D}(r)) + \mathbb{E}_f \log(1-\mathcal{D}(f))$$

$$\mathcal{L}_{\mathcal{S},\mathcal{T},\mathcal{D}} = \max_{\mathcal{S}} \min_{\mathcal{T}} \mathcal{L}_{\mathcal{T}} + \min_{\mathcal{S}} \max_{\mathcal{D}} \mathcal{L}_{\mathcal{D}}$$

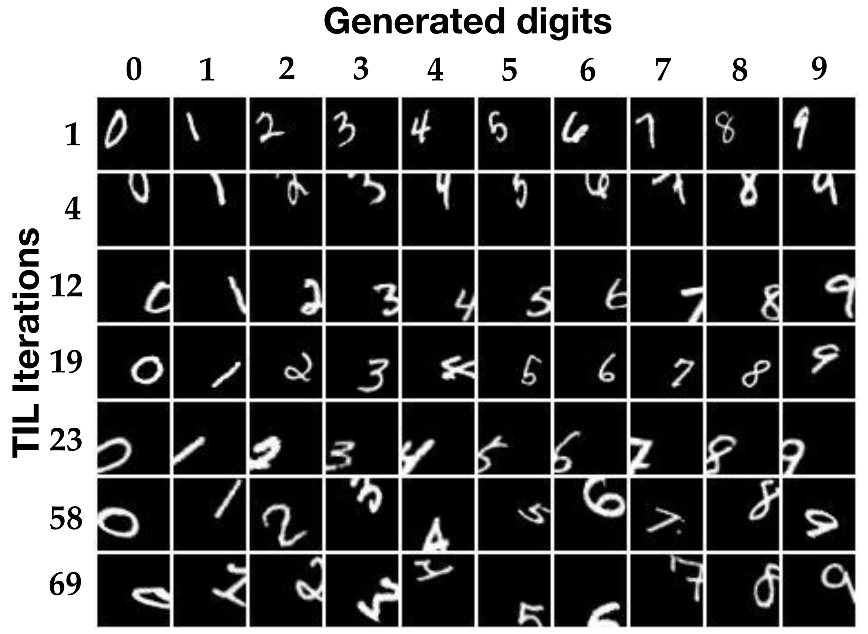
Blending Artefact Robustness



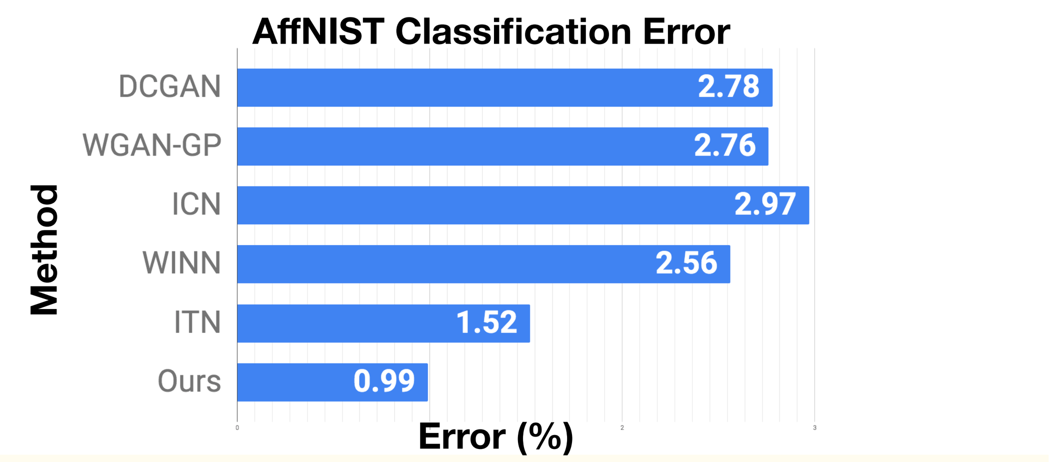
Experiments

Task-aware

- Generated increasingly diverse modes in successive TIL iterations



- Better classification error on AffNIST dataset



- Improved baseline SSD accuracy on Pascal VOC

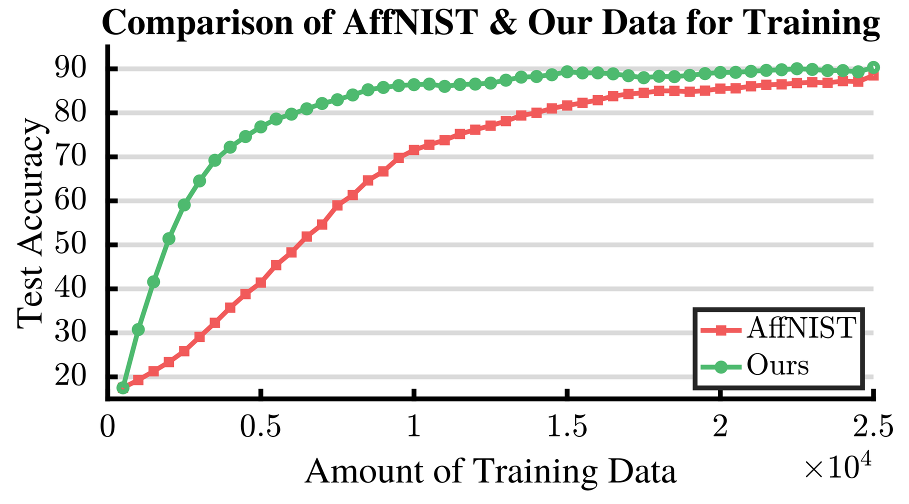
IoU	Baseline	[1]	[2]	Ours no-$\mathcal{D}$	Ours +$\mathcal{D}$
0.5	78.93	76.65	76.81	**79.61 (+0.68)**	79.53 (+0.60)
0.6	69.61	66.88	66.91	70.39 (+0.78)	**70.67 (+1.06)**
0.7	52.97	52.12	50.21	53.71 (+0.74)	**54.50 (+1.53)**
0.8	29.54	28.82	28.14	31.96 (+2.44)	**32.22 (+2.68)**

- Improved baseline Faster-RCNN accuracy on GMU Kitchen

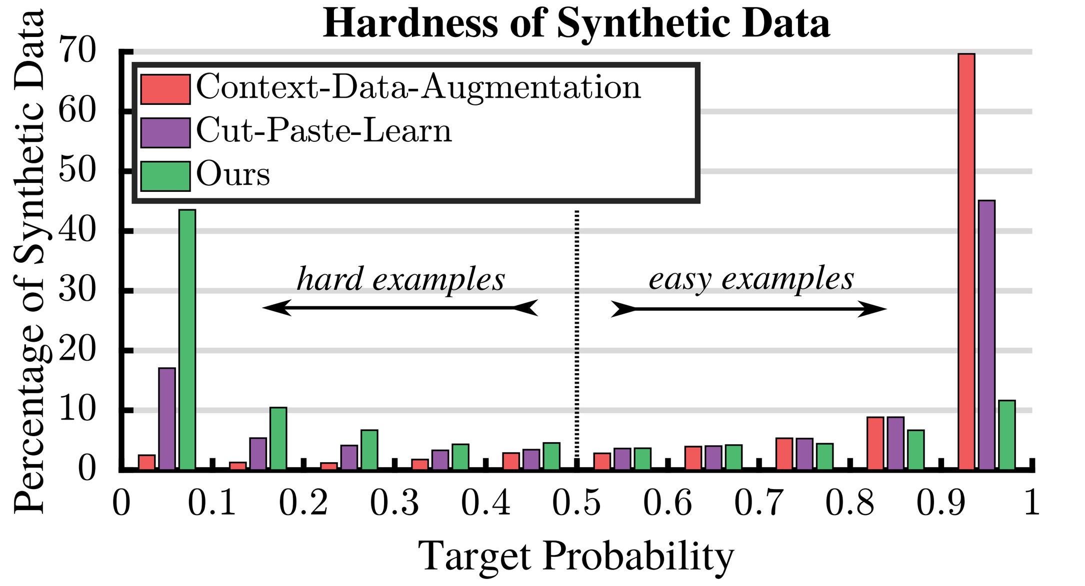
Dataset	coca cola	coffee mate	honey bunches	hunt's sauce	mahatma rice	nature v1	nature v2	palmolive orange	pop se-cret	pringles bbq	red bull	mAP
Baseline faster-RCNN	81.9	95.3	92.0	87.3	86.5	96.8	88.9	80.5	92.3	88.9	58.6	86.3
Cut-Paste-Learn	**88.5**	95.5	**94.1**	88.1	**90.3**	**97.2**	91.8	80.1	94.0	**92.2**	65.4	88.8
Ours	86.9	**95.9**	93.9	**90.2**	90.0	96.6	**92.0**	**87.6**	**94.9**	90.9	**69.2**	**89.8**

Efficient

- Achieved baseline accuracy with <50% synthetic data



- More hard positives generated than other image compositing approaches



Realistic

- Discriminator results in realistic synthetic data generation

Qualitative Results

- TIL SSD is more robust to occlusions, truncations, small scale, rare context etc.