Not Using the Car to See the Sidewalk: Quantifying and Controlling the Effects of Context in Classification and Segmentation

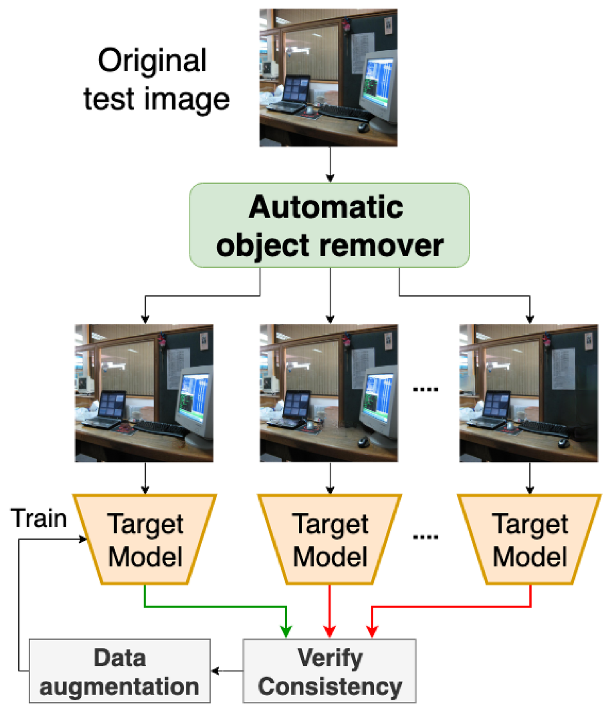
Rakshith Shetty[1] Bernt Schiele[1] Mario Fritz[2]

[1]Max Planck Institute for Informatics [2]CISPA Helmholtz Center for Information Security, Saarland Informatics Campus

Contact: rshetty@mpi-inf.mpg.de

Summary

- Context is useful but overuse can be harmful. Side effects include object hallucination and blindness to existing objects
- We present an **automatic test-case generation** system to quantify context dependency and identify failure modes
- Object removal is done using ground truth masks and an in-painter trained for adversarial scene editing [1]
- For example, removing *cars* causes segmenation models to fail to distinguish between *road* and *sidewalk* classes
- Data augmentation with generated samples **improves robustness** in both classification and segmentation networks **without sacrificing performance**.

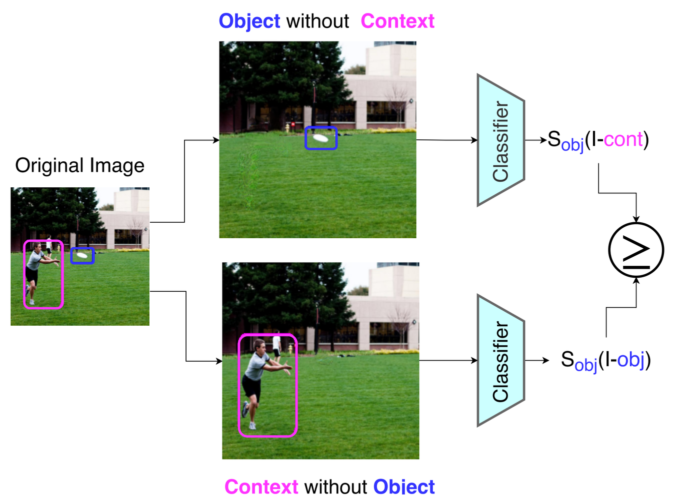


Automatically Testing Robustness to Context

Image Classification

- Use object removal to create a context without object image and a set of object without context images
- Count the number of violations to compute

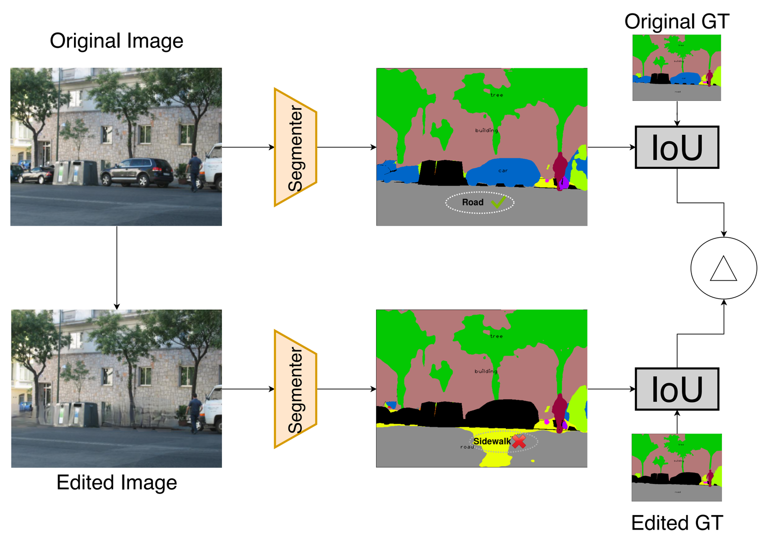
$$V^{\min}(c_i) = \frac{\Sigma_I \mathbb{1}\left[\left(\min_{\text{cont}} S_{c_i}(I - \text{cont})\right) < S_{c_i}(I - c_i)\right]}{\Sigma_I \mathbb{1}[c_i \in I]}$$



Semantic Segmentation

- Run segmentation on original and edited image with one object removed
- Measure the change in IoU for other objects

$$AR(c_i, c_j) = \frac{\Sigma_I \mathbb{1}\left[\left|\Delta \text{IoU}_{c_i c_j}\right| \geq \alpha\right]}{\Sigma_I \mathbb{1}[c_i, c_j \in I]}$$



- *AR* matrix captures inter-class dependency

Data augmentation

Image Classification

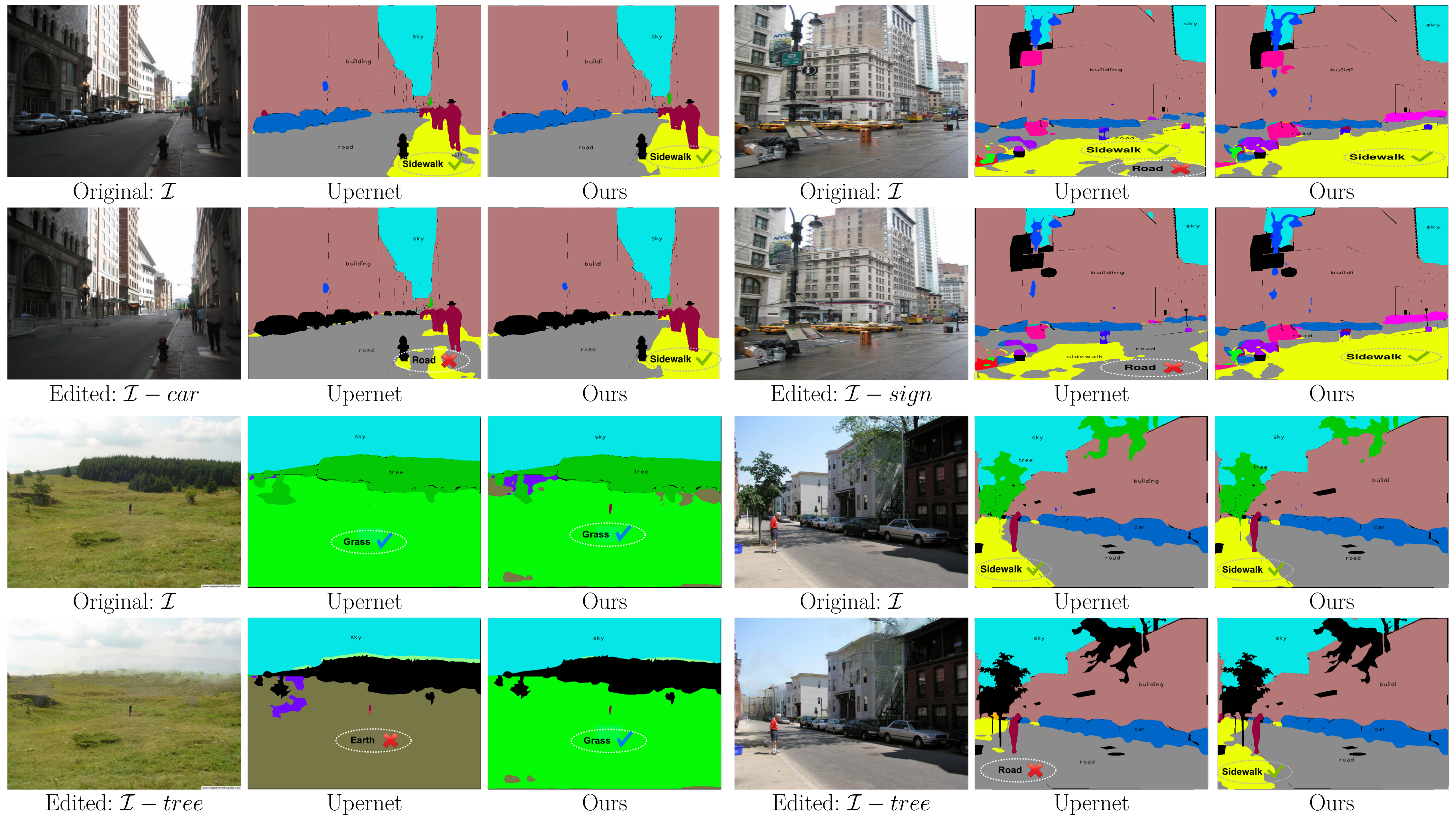
- *DA-Rand*: Randomly sample object to remove and use standard cross entropy loss
- *DA-Const*: Explicitly enforce constraints using hinge loss

$$\mathcal{L}_h(I) = \sum_{c_i \in I} \max\left[0, S_{c_i}(I - c_i) - \min_{c_j, j \neq i} S_{c_i}(I - c_j)\right]$$

Semantic Segmentation

- *DA-Size*: Sample the removed object inversely proportional to the area

Context in Semantic Segmentation



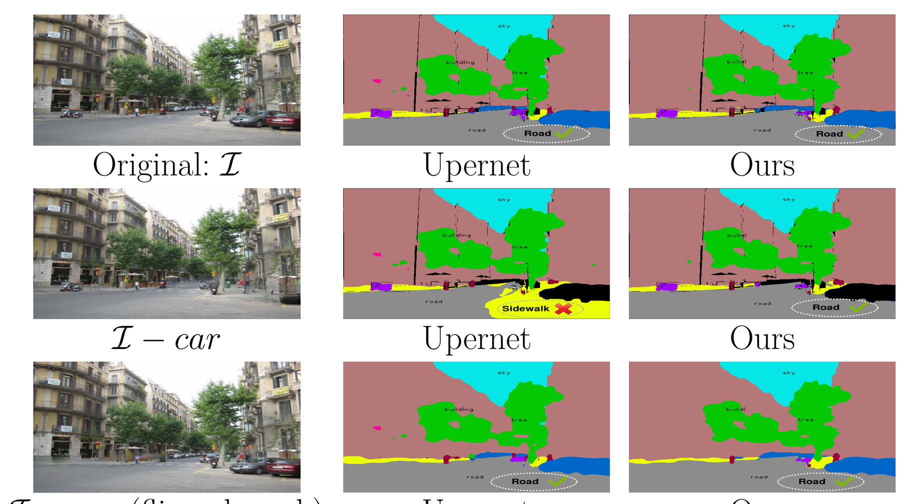
Quantitative results and ablations

Encoder	Decoder	mIoU	Sensitivity of sidewalk to car
mobilenet	conv [2]	0.324	18%
resnet-18	ppm [3]	0.380	18%
resnet-50	ppm [3]	0.408	20%
resnet-101	upernet [2]	0.420	22%
*resnet-50	upernet [2]	0.377	22%
*resnet-50 + DA-HardNeg	upernet [2]	0.385	**14%**

Better performing architectures are still sensitive to context changes. Our data augmentation increases the robustness to context changes. Models marked in * are trained with smaller batchsize.

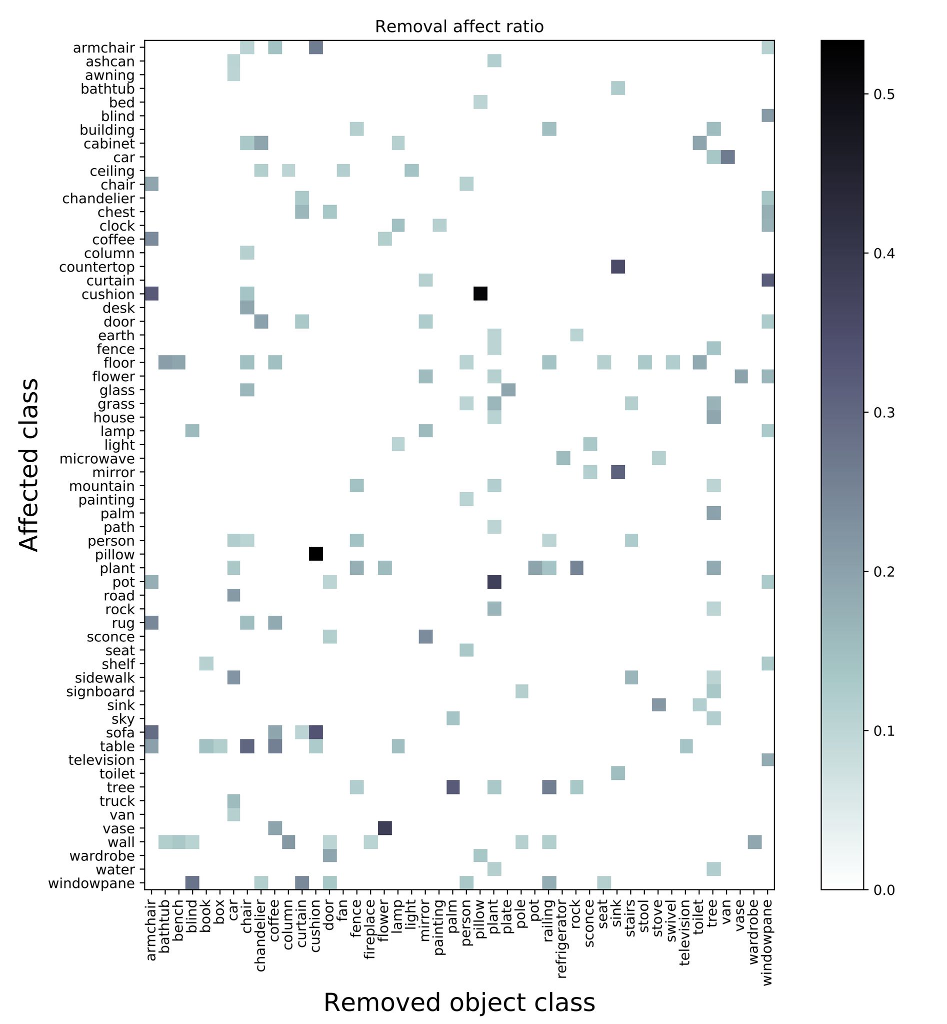
Model	all (407 images)		with car (258)		without car (149)	
	road	sidewalk	road	sidewalk	road	sidewalk
Upernet	0.81	0.59	0.86	0.67	0.68	0.40
DA-HardNeg	0.82	0.60	0.86	0.65	**0.72**	**0.46**

Context sensitivty is seen in real data as well Looking at subsets of real images with and without car, we see that the segmetation performance of road and sidewalk is significantly worse without car. Data augmentation improves this



Model	mIoU	Accuracy
Upernet [2]	0.377	78.31
DA-Size	0.377	78.25
DA-HardNeg	**0.385**	**78.47**

Performance comparison on ADE20k dataset. Hard negative data augmentation performs better than baseline and DA-size.



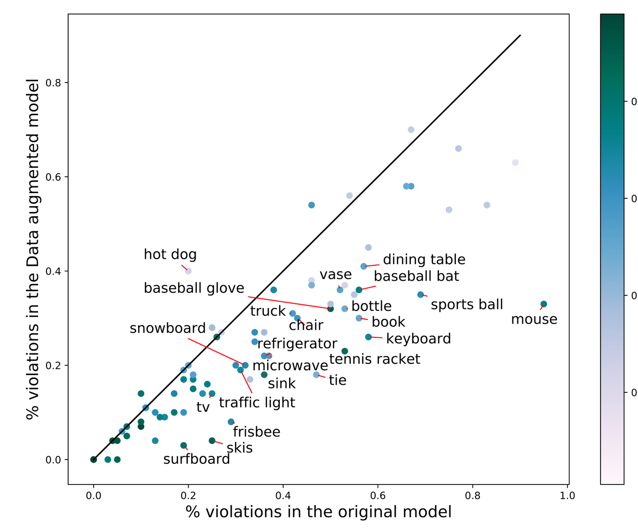
Visualizing frequency with which classes are affected by

Context in Classification

Examples of context sensitivity in classification. Baseline classifier weighs the contextual evidence more than the actual object. Data augmentation helps model learn to correctly rank these images

Model	Training Data	COCO test set		Robustness Metrics		UnRel
		Co-occur ↑	Single ↑	$V^{\min}$ ↓	V^{mean} ↓	dataset ↑
Baseline	Full (39k)	0.57	0.62	34%	24%	0.50
DA-Rand	Full (39k)	0.58	**0.65**	32%	22%	**0.54**
DA-Const	Full (39k)	0.58	0.63	**25%**	**14%**	0.52
Baseline	Co-occur (30k)	0.55	0.58	34%	24%	0.46
DA-Rand	Co-occur (30k)	**0.57**	**0.60**	31%	21%	0.49
DA-Const	Co-occur (30k)	**0.57**	**0.60**	**27%**	**15%**	**0.51**

Performance comparison on ADE20k dataset. Hard negative data augmentation performs better than baseline and DA-size.



Effect of data augmentation on robustness. Classes like 'mouse', 'keyboard', 'sports ball' get significantly more robust with data augmentation.

References

[1] R. Shetty, M. Fritz, and B. Schiele, "Adversarial scene editing: Automatic object removal from weak supervision," in *NeurIPS*, 2018.

[2] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for