Supplementary Material: Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias

Krishna Kumar Singh[1], Dhruv Mahajan[2], Kristen Grauman[2,3], Yong Jae Lee[1], Matt Feiszli[2], Deepti Ghadiyaram[2]

[1]University of California, Davis, [2]Facebook AI, [3]University of Texas at Austin

1. Additional implementation Details

Choosing the biased category pairs: As mentioned in Sec. 3.1, our method is built on the following intuition: a given category b is most biased by c if (1) the prediction probability of b drops significantly in the *absence* of c and (2) b co-occurs frequently with c. Regarding (2), the co-occurring class for the biased categories appeared at least 20% of the times with the biased categories on COCO-Stuff and Animals with Attributes dataset, and 10% of the times for the DeepFashion dataset.

For the COCO-Stuff, we partition the training data into non-overlapping $80-20$ split. We train a standard multi-label classifier with BCE loss on the 80% split and compute bias (Eq. 1) on the 20% split. For the DeepFashion, we train the classifier on the entire training data and determine the bias on the validation data. For the Animals with Attributes dataset, we need to use the test data to determine the biased classes as the test set has different distribution than the training data (test set consists of animal classes unseen during the training).

Choice of $\alpha_{\min}$: $\alpha_{\min}$ is set to 3 for COCO-Stuff and Animals with Attributes, whereas it is set to 5 for DeepFashion dataset. We found these values through cross-validation. During inference, a single forward pass of an image takes 0.2 ms on a single Titan X GPU.

2. More results

Another baseline *split biased*: In addition to all baselines we describe in the main text, we also designed another baseline: *split biased*. For this, we split each b into two categories: (1) $\mathrm{b} \setminus \mathrm{c}$ and (2) $\mathrm{b} \cap \mathrm{c}$. This setup adds K additional categories to each dataset and explicitly separates the two scenarios (exclusive and co-occur) for biased categories. This baseline is similar to [4], where a separate classifier is learned for a visual phrase consisting of objects associated with a relation (e.g. "person riding horse"). Here, instead of visual phrases, we learn a separate classifier for each co-occurring biased class pair.

Methods	Exclusive	Co-occur
split biased	19.1	64.3
ours-CAM	26.4	64.9
ours-feature-split	28.8	66.0

Table 1. **Performance on COCO-Stuff** for the 20 most biased categories. *ours-CAM* and *ours-feature-split* outperform *split biased* with significant margin on both exclusive and co-occurring images.

Methods	60 non biased categories	171 object + stuff
standard	75.4	57.2
ours-CAM	75.2	57.0
ours-feature-split	75.2	57.1

Table 2. **mAP of the non-biased object classes** and entire object+ stuff classes. Our approach loses only negligible mAP compared to *standard* classifier in these cases.

2.1. Object Classification

Comparison with *split biased*: Results in Table 1 shows that *ours-feature-split* outperforms *split biased* with a significant margin on COCO-Stuff (28.8 vs. 19.1). Also, *ours-CAM* gives much better performance than *split biased* (26.4 vs. 19.1). Given that *split biased* cannot take full advantage of the co-occurring images (and vice-versa), it has inferior performance compared to both our methods.

Performance on non-biased classes: In Table 2, we show the mAP of our approach and *standard* classifier on the non-biased object classes (60 classes) and on the entire COCO-Stuff dataset (*object* + *stuff*, 171 classes). We can see that our approach very marginally (0.02%) reduces the performance on non-biased object and stuff classes, while improving performance when biased categories occur away from their context.

Measuring cosine-similarity between $\mathrm{W_o}$ and $\mathrm{W_s}$: We verify that $\mathrm{W_o}$ and $\mathrm{W_s}$ capture distinct information by computing a cosine similarity metric between them. From Table 3, we observe that both our approaches yield a lower similarity score compared to *standard*.

Methods	Cosine-similarity
standard	0.21
ours-CAM	0.19
ours-feature-split	0.17

Table 3. **Cosine similarity** between classifier weights of the biased class pairs (b,c). Our approach reduces the similarity between them indicating the biased class b is less dependent on c for prediction.

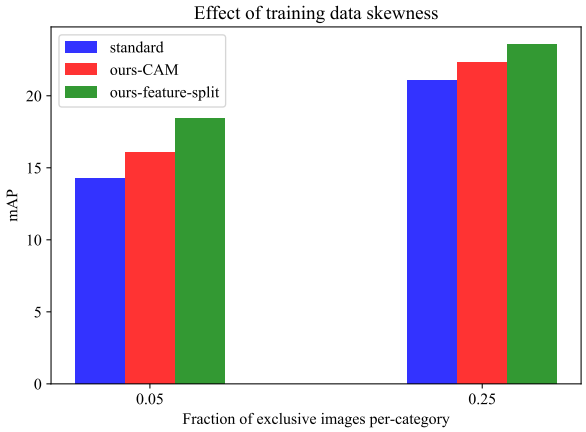

Figure 1. mAP of *standard*, *ours-CAM*, and *ours-feature-split* classifier by varying fraction of exclusive images during training. If ratio is more skewed then we get a bigger boost for the exclusive cases.

Per class mAP and co-occurrence bias for 20 biased classes: In the Table 6, we show per class results for the COCO-Stuff for the top 20 biased classes. We also show the co-occurrence bias value for each class computed according to Eq. 1 in the main paper. From these results, we may observe that when a category occurs out of its context *ours-feature-split* gives better performance compared to *standard* classifier while maintaining the performance when a category co-occurs with context. *ours-CAM* performs better than *standard* when a category occurs away from its context, but struggles when categories co-occur.

Ablation study of *ours-feature-split* by varying fraction of biased category images: Here, we study the performance of our method as we vary the fraction of training images with biased categories occurring away from their typical context for COCO-Stuff. Specifically, for each of the 20 biased categories in COCO-Stuff, we fix the total number of training images and vary the fraction of exclusive images. From Fig. 1, we note that *standard* performs rather poorly at lower fractions compared to both approaches (*ours-CAM* and *ours-feature-split*). Thus, both proposed methods achieve higher boosts at a fraction of 0.05 compared to 0.25. We also observe that a higher fraction of exclusive images benefits all the approaches, yet, our methods consistently outperform *standard*. This indicates that our approaches are more robust than the baseline especially on heavily skewed training data.

Methods	Exclusive	Co-occur
standard	4.9	17.8
split biased	3.5	14.3
remove co-occur labels	6.0	**20.4**
remove co-occur images	4.2	5.4
negative penalty	5.5	18.9
class balancing loss [1]	5.2	19.4
ours-feature-split	**9.2**	20.1

Table 4. **Top-3 recall on DeepFashion** for the 20 most biased attributes. *ours-feature-split* yields a significant boost over all approaches for the exclusive test split, without hurting performance on the co-occurring split. *ours-CAM* is not extensible to attributes hence not reported here. The above baseline methods are described in our main paper.

Methods	Exclusive	Co-occur
standard	19.4	72.2
split biased	19.7	66.8
remove co-occur labels	19.1	62.9
remove co-occur images	**22.7**	58.3
negative penalty	19.2	68.4
class balancing loss [1]	20.4	68.4
attribute decorrelation [2]	18.4	70.2
ours-feature-split	20.8	**72.8**

Table 5. **Performance on Animals with Attributes** for the 20 most biased attributes. Our proposed method *ours-feature-split* outperforms other methods. *ours-CAM* is not extensible to attributes hence not reported here.

2.2. Comparison with other baselines for attribute classification

Table 4 reports performance on DeepFashion [3]. We outperform all baselines by a significant margin on the exclusive test set. Although *remove co-occur labels* has slightly higher performance when attributes co-occur (20.4 vs. 20.1), *ours-feature-split* performs significantly better when attributes occur exclusively (6.0 vs. 9.2).

From Table 5, we observe that *ours-feature-split* offers gains on the exclusive test split compared to most methods for Animals with Attributes dataset. Though *remove co-occur images* yields higher gains on the exclusive test split, unlike *ours-feature-split*, it severely hurts the performance of co-occurring cases. Meanwhile *ours-feature-split* achieves good gains in exclusive cases without hurting co-occurring cases.

Finally, in Table 7 and 8, we show per category performance for the top 20 biased categories for two datasets: DeepFashion and Animals with Attributes. These results show that *ours-feature-split* gives better performance than the *standard* classifier when attributes occur exclusively without their co-occurring context. At the same time, *ours-feature-split* maintains performance when biased attribute categories appear with co-occurring context.

References

[1] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

Classes			Exclusive			Co-occur		
Biased class	Co-occur class	Bias	*standard*	*ours-CAM*	*ours-feature-split*	*standard*	*ours-CAM*	*ours-feature-split*
cup	dining table	1.76	33.0	35.4	27.4	68.1	63.0	70.2
wine glass	person	1.8	35.0	36.3	35.1	57.9	57.4	57.3
handbag	person	1.81	3.8	5.1	4.0	42.8	41.4	42.7
apple	fruit	1.91	29.2	29.8	30.7	64.7	64.4	64.1
car	road	1.94	36.7	38.2	36.6	79.7	78.5	79.2
bus	road	1.94	40.7	41.6	43.9	86.0	85.3	85.4
potted plant	vase	1.99	37.2	37.8	36.5	50.0	46.8	46.0
spoon	bowl	2.04	14.7	16.3	14.3	42.7	35.9	42.6
microwave	oven	2.08	35.3	36.6	39.1	60.9	60.1	59.6
keyboard	mouse	2.25	44.6	42.9	47.1	85.0	83.3	85.1
skis	person	2.28	2.8	7.0	27.0	91.5	91.3	91.2
clock	building	2.39	49.6	50.5	45.5	84.5	84.7	86.4
sports ball	person	2.45	12.1	14.7	22.5	75.5	75.3	74.2
remote	person	2.45	23.7	26.9	21.2	70.5	67.4	72.7
snowboard	person	2.86	2.1	2.4	6.5	73.0	72.7	72.6
toaster	ceiling	3.7	7.6	7.7	6.4	5.0	5.0	4.4
hair drier	towel	4	1.5	1.3	1.7	6.2	6.2	6.9
tennis racket	person	4.15	53.5	59.7	61.7	97.6	97.5	97.5
skateboard	person	7.36	14.8	22.6	34.4	91.3	91.1	90.8
baseball glove	person	339.15	12.3	14.4	34.0	91.0	91.3	91.1
Mean	-	-	24.5	26.4	**28.8**	**66.2**	64.9	66.0

Table 6. COCO-Stuff dataset. Per class mAP and bias for 20 most biased classes. *ours-feature-split* outperforms *standard* on the exclusive set while maintaining the performance on the co-occurring cases.

Classes			Exclusive		Co-occur	
Biased class	Co-occur class	Bias	*standard*	*ours-feature-split*	*standard*	*ours-feature-split*
bell	lace	3.15	5.4	22.8	3.1	9.4
cut	bodycon	3.3	8.6	12.5	29.3	36.2
animal	print	3.31	0.0	1.9	1.9	2.8
flare	fit	3.31	18.4	32.0	56.0	62.0
embroidery	crochet	3.44	4.1	1.8	4.8	0.0
suede	fringe	3.48	12.0	19.6	65.2	73.9
jacquard	flare	3.68	0.0	0.9	0.0	9.1
trapeze	striped	3.7	8.7	29.9	42.9	50.0
neckline	sweetheart	3.98	0.0	0.0	0.0	0.0
retro	chiffon	4.08	0.0	0.4	0.0	0.0
sweet	crochet	4.32	0.0	0.5	0.0	0.0
batwing	loose	4.36	11.0	12.0	27.5	15.0
tassel	chiffon	4.48	13.0	16.8	25.0	25.0
boyfriend	distressed	4.5	11.6	11.6	49.2	38.1
light	skinny	4.53	2.0	1.3	14.9	8.5
ankle	skinny	4.56	1.0	14.6	13.2	27.9
french	terry	5.09	0.0	0.8	9.6	7.9
dark	wash	5.13	2.6	2.1	8.7	13.0
medium	wash	7.45	0.0	0.0	0.0	0.0
studded	denim	7.8	0.0	3.2	4.0	24.0
Mean	-	-	4.9	**9.2**	17.8	**20.1**

Table 7. DeepFashion dataset. Per class top-3 recall and bias for 20 most biased classes. *ours-feature-split* outperforms *standard* on the exclusive set while maintaining the performance on the co-occurring cases.

[2] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share.

Classes			Exclusive		Co-occur	
Biased class	Co-occur class	Bias	*standard*	*ours-feature-split*	*standard*	*ours-feature-split*
white	ground	3.67	24.8	24.6	85.8	86.2
longleg	domestic	3.71	18.5	29.1	89.4	89.3
forager	nestspot	4.02	33.6	33.4	96.6	96.5
lean	stalker	4.46	11.5	12.0	54.5	55.8
fish	timid	5.14	60.2	57.4	98.3	98.3
hunter	big	5.34	4.1	3.6	32.9	30.0
plains	stalker	5.4	6.4	6.0	44.7	59.9
nocturnal	white	5.84	13.3	13.1	71.2	60.5
nestspot	meatteeth	5.92	13.4	14.9	62.8	67.6
jungle	muscle	6.26	33.3	31.3	88.6	86.6
muscle	black	6.39	9.3	9.3	76.6	73.6
meat	fish	7.12	4.5	3.8	76.1	73.6
mountains	paws	9.24	10.9	10.0	49.9	39.9
tree	tail	10.98	36.5	55.0	93.2	92.7
domestic	inactive	11.77	11.9	13.1	73.7	76.6
spots	longleg	20.15	43.8	45.2	61.8	59.1
bush	meat	29.47	19.8	22.1	70.2	75.1
buckteeth	smelly	34.01	7.8	8.9	27.1	45.3
slow	strong	76.59	15.5	14.6	95.8	93.3
blue	coastal	319.98	8.4	8.2	94.2	95.8
Mean	-	-	19.4	**20.8**	72.2	**72.8**

Table 8. Animals with Attributes dataset. Per class mAP and bias for 20 most biased classes. *ours-feature-split* outperforms *standard* on the exclusive set while maintaining the performance on the co-occurring cases.

In *CVPR*, 2014. 2

[3] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2

[4] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1