

打地鼠困境

A Whac-A-Mole Dilemma : Shortcuts Come in Multiples Where Mitigating One Amplifies Others

[†]Zhiheng Li² [‡]Ivan Evtimov¹ Albert Gordo¹ Caner Hazirbas¹ Tal Hassner¹

Cristian Canton Ferrer¹ Chenliang Xu² [‡]Mark Ibrahim¹

¹Meta AI ²University of Rochester

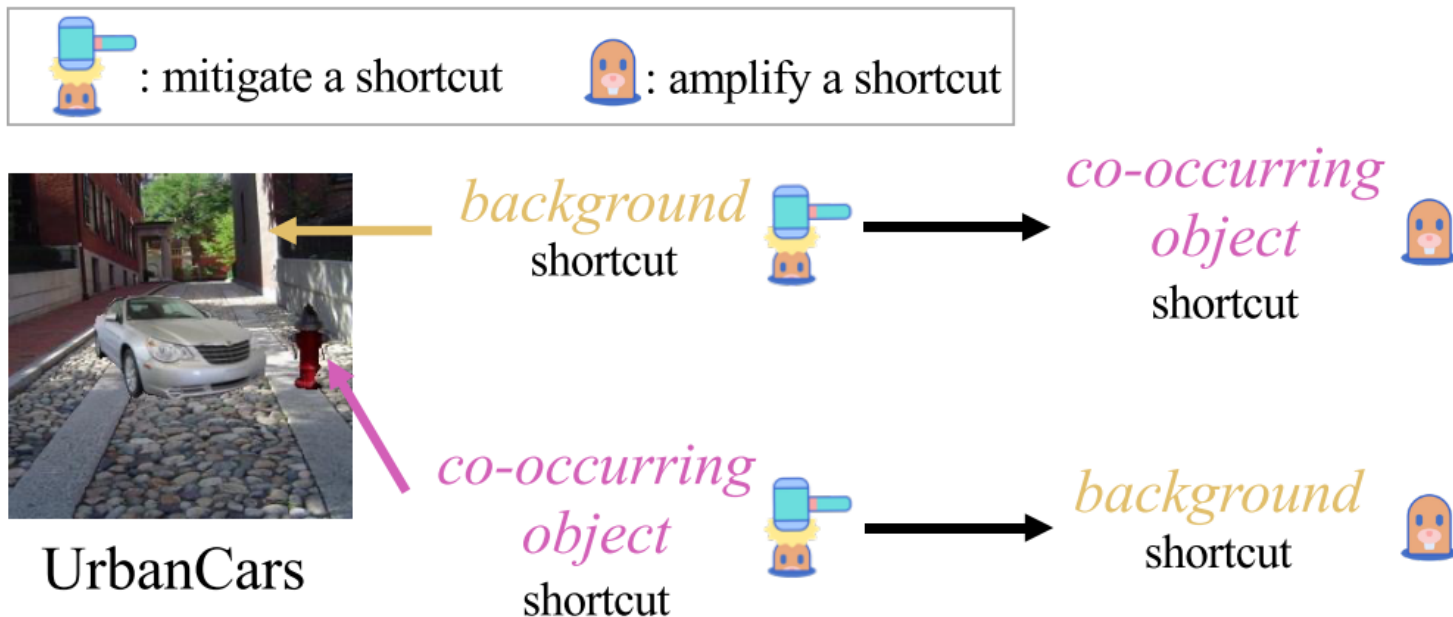
{ivanevtimov, agordo, hazirbas, thassner, ccanton, marksibrahim}@meta.com

{zhiheng.li, chenliang.xu}@rochester.edu

Rui Hu
2023.1.7

Motivation









- Most existing works design and evaluate methods under the tenuous assumption that a **single shortcut** is present in the data.
- Existing methods struggle in a Whac-A-Mole game, i.e., where mitigating one shortcut amplifies reliance on others.



New Datasets for Multi-Shortcut Mitigation

- UrbanCars Dataset

- Target: the car's body type
- Shortcut: background & co-occurring object

	Common BG Common CoObj	Uncommon BG Common CoObj	Common BG Uncommon CoObj	Uncommon BG Uncommon CoObj
Frequency	90.25%	4.75%	4.75%	0.25%
urban car				
country car				

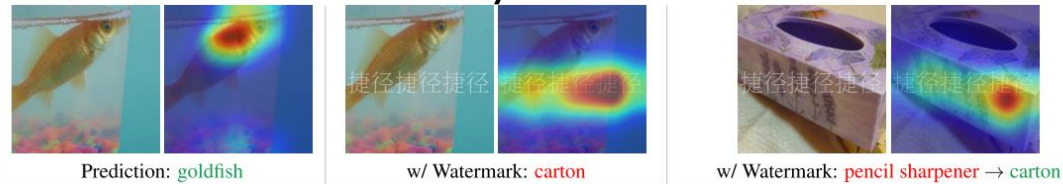
- ImageNet-Watermark

- Target: object
- Shortcut: texture & background & **watermark**



Ubiquitous reliance on the watermark shortcut

- Models of different **architectures**, augmentations, regularizations and **pretraining supervision** rely on the watermark as a shortcut.
 - models with larger architectures or extra training data can decrease reliance on the watermark shortcut.
 - CLIP with zero-shot transfer is least affected by watermark shortcuts.



method	architecture	(pre)training data	IN-1k Acc \uparrow	$P(\hat{y} = \text{carton}) (\%)$	IN-W Gap \uparrow	$\Delta P(\hat{y} = \text{carton}) (\%) \downarrow$	Carton Gap \downarrow	$\Delta P(\hat{y} = \text{carton} y = \text{carton}) (\%) \downarrow$
Supervised	ResNet-50 [30]	IN-1k [17]	76.1	0.07	-26.7	+7.56	+40	+42.46
MoCov3 [13] (LP)	ResNet-50	IN-1k	74.6	0.08	-20.7	+2.94	+44	+44.37
Style Transfer [26]	ResNet-50	SIN [26]	60.1	0.10	-17.3	+4.91	+52	+50.06
Mixup [93]	ResNet-50	IN-1k	76.1	0.07	-18.6	+3.43	+38	+39.78
CutMix [92]	ResNet-50	IN-1k	78.5	0.09	-14.8	+1.92	+22	+29.61
Cutout [19,96]	ResNet-50	IN-1k	77.0	0.08	-18.0	+2.93	+32	+38.06
AugMix [35]	ResNet-50	IN-1k	77.5	0.09	-16.8	+2.61	+36	+34.44
Supervised	RG-32gf	IN-1k	80.8	0.09	-14.1	+3.74	+32	+33.43
SEER [27] (FT)	RG-32gf [67]	IG-1B [27]	83.3	0.09	-6.5	+0.56	+18	+24.26
Supervised	ViT-B/32 [21]	IN-1k	75.9	0.09	-8.7	+1.20	+34	+34.31
Uniform Soup [89] (FT)	ViT-B/32	WIT [66]	79.9	0.09	-7.9	+0.32	+24	+23.87
Greedy Soup [89] (FT)	ViT-B/32	WIT	81.0	0.09	-6.5	+0.35	+16	+23.87
Supervised	ViT-L/16	IN-1k	79.6	0.08	-6.2	+0.82	+34	+32.57
CLIP [66] (zero-shot)	ViT-L/14	WIT	76.5	0.06	-4.4	+0.01	+12	+1.75
CLIP (zero-shot)	ViT-L/14	LAION-400M [74]	72.7	0.05	-4.9	+0.03	+12	+13.76
MAE [29] (FT)	ViT-H/14	IN-1k	86.9	0.08	-3.5	+0.43	+30	+29.59
SWAG [79] (LP)	ViT-H/14	IG-3.6B [79]	85.7	0.09	-4.9	+0.19	+8	+12.80
SWAG (FT)	ViT-H/14	IG-3.6B	88.5	0.09	-3.1	+0.35	+18	+20.25
CLIP (zero-shot)	ViT-H/14	LAION-2B [73]	77.9	0.06	-3.6	+0.03	+16	+12.01
average			78.6	0.08	-10.7	+1.74	+26.7	+27.96

Benchmark Methods and Settings

- We comprehensively evaluate shortcut mitigation methods across four categories based on the level of shortcut information required.

Category	Summary	Shortcut Information	Methods
1	Standard Augmentation and Regularization	None	Mixup [93], Cutout [19,96], CutMix [92], AugMix [35], SD [63]
2	Targeted Augmentation for Mitigating Shortcuts	Types of shortcuts (w/o shortcut labels)	CF+F Aug [11], Style Transfer (TXT Aug) [26], BG Aug [71,91], WMK Aug
3	Using Shortcut Labels	Image-level ground-truth shortcut label	gDRO [72], DI [87], SUBG [38], DFR [45]
4	Inferring Pseudo Shortcut Labels	Image-level pseudo shortcut label	LfF [60], JTT [58], EIIL [15], DebiAN [53]

Analytic Experiment on New Dataset

- Standard training relies on multiple shortcuts.
- Standard augmentation and regularization methods can mitigate some shortcuts (e.g., texture) but amplify others .
- Augmentations tackling a specific type of shortcut (e.g., style transfer for texture shortcut) can amplify other shortcuts (e.g., watermark).

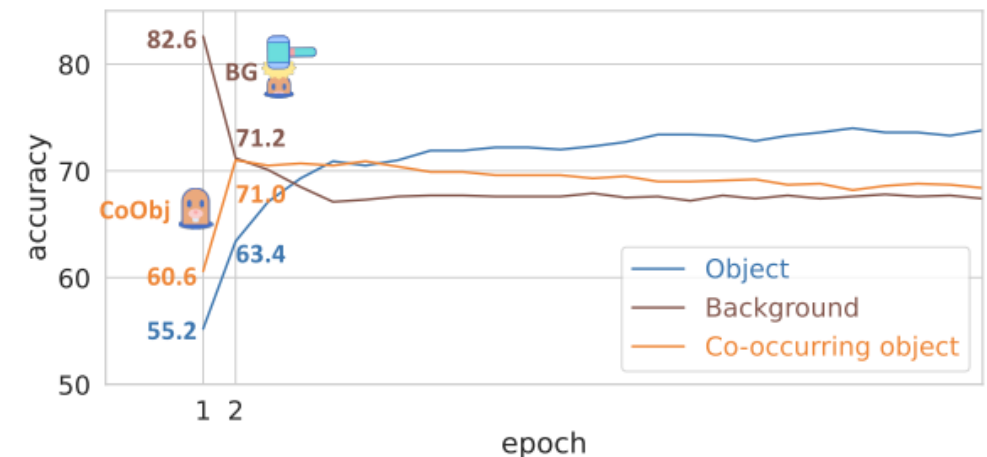
	I.D. Acc	shortcut reliance		
		BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	97.6	-15.3	-11.2	-69.2
Mixup	98.3	-12.6	-9.3	-61.8
CutMix	96.6	-45.0 ($\times 2.94$ 🤖)	-4.8	-86.5
Cutout	97.8	-15.8 ($\times 1.03$ 🤖)	-10.4	-71.4
AugMix	98.2	-10.3	-12.1 ($\times 1.08$ 🤖)	-70.2
SD	97.3	-15.0	-3.6	-36.1
CF+F Aug	96.8	-16.0 ($\times 1.04$ 🤖)	+0.4	-19.4
LfF	97.2	-11.6	-18.4 ($\times 1.64$ 🤖)	-63.2
JTT (E=1)	95.9	-8.1	-13.3 ($\times 1.18$ 🤖)	-40.1
EiIL (E=1)	95.5	-4.2	-24.7 ($\times 2.21$ 🤖)	-44.9
JTT (E=2)	94.6	-23.3 ($\times 1.52$ 🤖)	-5.3	-52.1
EiIL (E=2)	95.5	-21.5 ($\times 1.40$ 🤖)	-6.8	-49.6
DebiAN	98.0	-14.9	-10.5	-69.0
LLE (ours)	96.7	-2.1	-2.7	-5.9

	IN-1k	shortcut reliance				
		Watermark (WTM)		Texture (TXT)		Background (BG)
		IN-W Gap ↑	Carton Gap ↓	SIN Gap ↑	IN-R Gap ↑	IN-9 Gap ↑
ERM	76.39	-25.40	+30	-69.43	-56.22	-5.19
Mixup	76.17	-24.87	+34 ($\times 1.13$ 🤖)	-68.18	-55.79	-5.60 ($\times 1.08$ 🤖)
CutMix	75.90	-25.78 ($\times 1.01$ 🤖)	+32 ($\times 1.06$ 🤖)	-69.31	-56.36	-5.65 ($\times 1.09$ 🤖)
Cutout	76.40	-25.11	+32 ($\times 1.06$ 🤖)	-69.39	-55.93	-5.35 ($\times 1.03$ 🤖)
AugMix	76.23	-23.41	+38 ($\times 1.26$ 🤖)	-68.51	-54.91	-5.85 ($\times 1.13$ 🤖)
SD	76.39	-26.03 ($\times 1.02$ 🤖)	+30	-69.42	-56.36	-5.33 ($\times 1.03$ 🤖)
WTM Aug	76.32	-5.78	+14	-69.31	-56.22	-5.34 ($\times 1.03$ 🤖)
TXT Aug	75.94	-25.93 ($\times 1.02$ 🤖)	+36 ($\times 1.20$ 🤖)	-63.99	-53.24	-5.66 ($\times 1.09$ 🤖)
BG Aug	76.03	-25.01	+36 ($\times 1.20$ 🤖)	-68.41	-54.51	-4.67
LfF	76.35	-26.19 ($\times 1.03$ 🤖)	+36 ($\times 1.20$ 🤖)	-69.34	-56.02	-5.61 ($\times 1.08$ 🤖)
JTT	76.33	-26.40 ($\times 1.04$ 🤖)	+32 ($\times 1.06$ 🤖)	-69.48	-56.30	-5.55 ($\times 1.07$ 🤖)
EiIL	71.55	-33.48 ($\times 1.31$ 🤖)	+24	-66.04	-61.35 ($\times 1.09$ 🤖)	-6.42 ($\times 1.24$ 🤖)
DebiAN	76.33	-26.40 ($\times 1.04$ 🤖)	+36 ($\times 1.20$ 🤖)	-69.37	-56.29	-5.53 ($\times 1.07$ 🤖)
LLE (ours)	76.25	-6.18	+10	-61.00	-54.89	-3.82

Analytic Experiment on New Dataset

- **Methods using shortcut labels** mitigate the labeled shortcut but amplifies the unlabeled one.
- **Methods inferring pseudo shortcut labels** still amplify shortcuts. (previous page)
 - Because ERM learns different shortcuts asynchronously during training, making it hard to infer labels of all shortcuts for mitigation.

	shortcut label		I.D. Acc	shortcut reliance		
	Train	Val		BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	✗	BG+CoObj	97.6	-15.3	-11.2	-69.2
gDRO	BG+CoObj	BG+CoObj	91.6	-10.9	-3.6	-16.4
DI	BG+CoObj	BG+CoObj	89.0	-2.2	-1.0	+0.4
SUBG	BG+CoObj	BG+CoObj	71.1	-4.7	-0.3	-6.3
DFR	BG+CoObj	BG+CoObj	89.7	-10.7	-6.9	-45.2
ERM	✗	BG	97.8	-14.6	-11.3	-68.5
gDRO	BG 🚗	BG	96.0	-4.2	-26.9 (×2.39 🚗)	-56.5
DI	BG 🚗	BG	94.7	+2.2	-27.0 (×2.40 🚗)	-25.2
SUBG	BG 🚗	BG	92.6	+1.3	-36.4 (×3.24 🚗)	-35.8
DFR	BG 🚗	BG	97.4	-9.8	-13.6 (×1.21 🚗)	-58.9
ERM	✗	CoObj	97.6	-15.4	-11.0	-68.8
gDRO	CoObj 🚗	CoObj	95.7	-31.4 (×2.03 🚗)	-0.5	-54.9
DI	CoObj 🚗	CoObj	94.2	-36.1 (×2.34 🚗)	+2.8	-35.8
SUBG	CoObj 🚗	CoObj	93.1	-60.2 (×3.90 🚗)	+2.5	-62.4
DFR	CoObj 🚗	CoObj	97.4	-19.1 (×1.24 🚗)	-8.6	-64.9



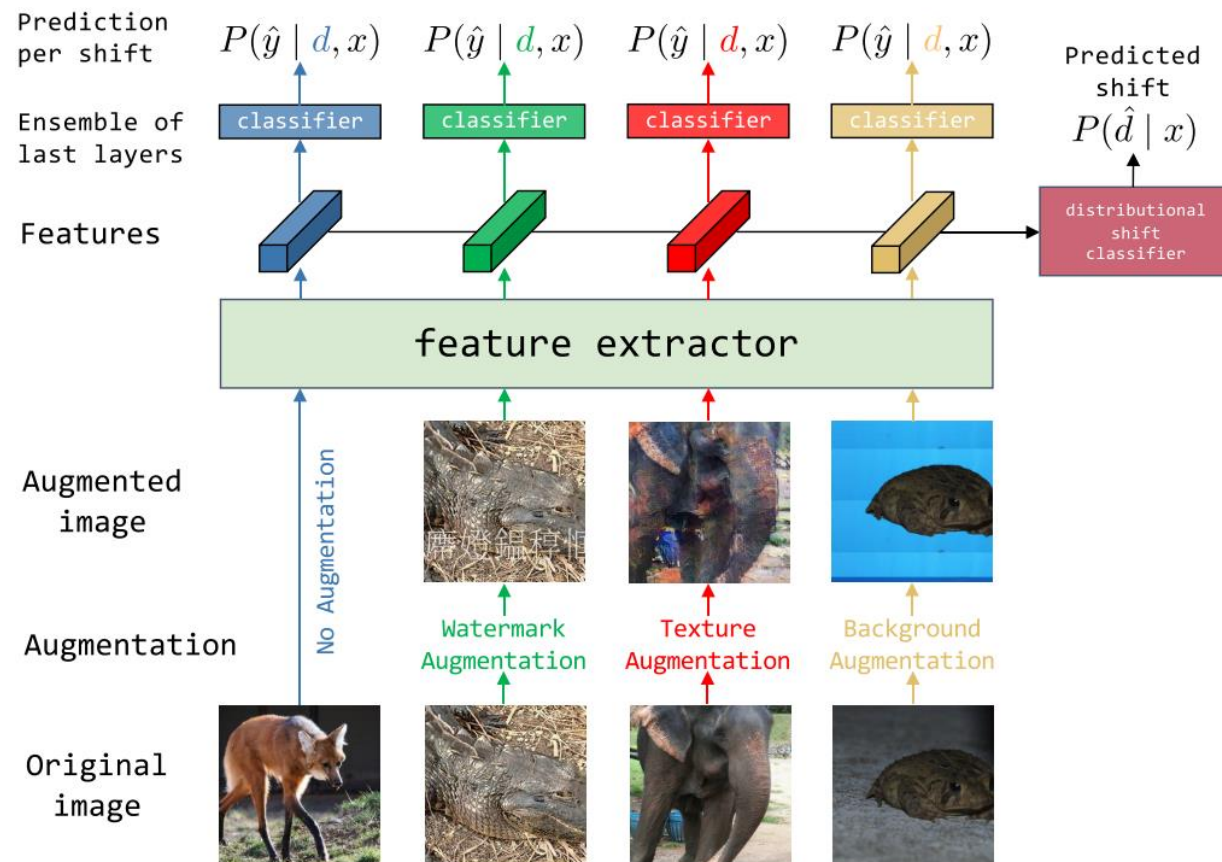
Analytic Experiment on New Dataset

- Self-supervised and foundation models can mitigate some shortcuts but amplify others.

	IN-1k	shortcut reliance				
		Watermark		Texture		Background
		IN-W Gap \uparrow	Carton Gap \downarrow	SIN Gap \uparrow	IN-R Gap \uparrow	IN-9 Gap \uparrow
<i>arch: RG-32gf</i>						
ERM	80.88	-14.15	+32	-69.27	-52.43	-6.40
SEER (FT,IG-1B)	83.35	-6.50	+18	-73.04 ($\times 1.05$ 🤖)	-50.42	-7.14 ($\times 1.11$ 🤖)
<i>arch: ViT-B/32</i>						
ERM	75.92	-8.71	+34	-57.16	-49.45	-6.86
Uniform Soup (FT,WIT)	79.96	-7.90	+24	-59.67 ($\times 1.04$ 🤖)	-27.51	-7.78 ($\times 1.13$ 🤖)
Greedy Soup (FT,WIT)	81.01	-6.47	+16	-59.61 ($\times 1.04$ 🤖)	-30.01	-7.21 ($\times 1.05$ 🤖)
<i>arch: ViT-B/16</i>						
ERM	81.07	-6.69	+26	-62.60	-50.36	-5.36
SWAG (LP,IG-3.6B)	81.89	-7.76 ($\times 1.16$ 🤖)	+18	-67.33 ($\times 1.08$ 🤖)	-19.79	-10.39 ($\times 1.94$ 🤖)
SWAG (FT,IG-3.6B)	85.29	-5.43	+24	-66.99 ($\times 1.07$ 🤖)	-29.55	-4.44
MoCov3 (LP)	76.65	-16.0 ($\times 2.39$ 🤖)	+22	-63.36 ($\times 1.01$ 🤖)	-56.86 ($\times 1.12$ 🤖)	-7.80 ($\times 1.45$ 🤖)
MAE (FT)	83.72	-4.60	+24	-65.20 ($\times 1.04$ 🤖)	-47.10	-4.45
MAE+LLE (ours)	83.68	-2.48	+6	-58.78	-44.96	-3.70
<i>arch: ViT-L/16 or 14</i>						
ERM	79.65	-6.14	+34	-61.43	-53.17	-6.50
SWAG (LP,IG-3.6B)	85.13	-5.73	+6	-60.26	-10.17	-7.26 ($\times 1.12$ 🤖)
SWAG (FT,IG-3.6B)	88.07	-3.16	+20	-63.45 ($\times 1.03$ 🤖)	-12.29	-2.92
CLIP (zero-shot,WIT)	76.57	-4.47	+12	-61.27	-6.26	-3.68
CLIP (zero-shot,LAION)	72.77	-4.94	+12	-56.85	-8.43	-4.54
MAE (FT)	85.95	-4.36	+22	-62.48 ($\times 1.02$ 🤖)	-36.46	-3.53
MAE+LLE (ours)	85.84	-1.74	+12	-56.32	-34.64	-2.77

Method: Last Layer Ensemble

- We focus on mitigating **multiple known shortcuts**, i.e., the number and types of shortcuts are given, but shortcut labels are not.



Limitation

- Not all shortcuts can be data augmented.
- Need to know the type of shortcut.

Conclusion

- Studying the Whac-A-Mole dilemma of **Large Pretrained Models** in shortcut mitigation is meaningful.