DiagViB-6: A Diagnostic Benchmark Suite for Vision Models in the Presence of Shortcut and Generalization Opportunities

Elias Eulig Piyapat Saranrittichai Chaithanya Kumar Mummadi
Kilian Rambach William Beluch Xiahan Shi Volker Fischer
Bosch Center for Artificial Intelligence (BCAI)

A. Supplementary material

A.1. Image generation

We provide a comprehensive overview of all factor classes and respective factor space regions in Tab. A1. The five textures from which the texture crops $f_{\text{text.}}$ are sampled are shown in Fig. A1.

$\mathcal{F}_i$	$\mathcal{S}_i$	$\mathcal{N}_i$	$\mathcal{C}_{i,j}$	$\mathcal{S}_{i,j}$
position	$[0,1]^2$	9	top-left	$[1/7, 2/7] \times [1/7, 2/7]$
			top-center	$[1/7, 2/7] \times [3/7, 4/7]$
			top-right	$[1/7, 2/7] \times [5/7, 6/7]$
			center-left	$[3/7, 4/7] \times [1/7, 2/7]$
			center-center	$[3/7, 4/7] \times [3/7, 4/7]$
			center-right	$[3/7, 4/7] \times [5/7, 6/7]$
			bottom-left	$[5/7, 6/7] \times [1/7, 2/7]$
			bottom-center	$[5/7, 6/7] \times [3/7, 4/7]$
			bottom-right	$[5/7, 6/7] \times [5/7, 6/7]$
hue	$[0, 2\pi)$	6	red	$[345°, 15°]$
			yellow	$[45°, 75°]$
			green	$[105°, 135°]$
			cyan	$[165°, 195°]$
			blue	$[225°, 255°]$
			magenta	$[285°, 315°]$
lightness	$[0,1]^2$	4	dark	$[0, 1/11] \times [4/11, 5/11]$
			darker	$[2/11, 3/11] \times [6/11, 7/11]$
			brighter	$[4/11, 5/11] \times [8/11, 9/11]$
			bright	$[6/11, 7/11] \times [10/11, 1.]$
scale	$[1/1.45, 1.45]$	5	small	$[1/1.45, 1/1.35]$
			smaller	$[1/1.25, 1/1.15]$
			normal	$[1/1.05, 1.05]$
			larger	$[1.15, 1.25]$
			large	$[1.35, 1.45]$
shape	MNIST	10	‘0’	digits ‘0’
			‘1’	digits ‘1’
			$\vdots$	$\vdots$
texture	textures	5	tiles	tiles texture crops
			wood	wood texture crops
			carpet	carpet texture crops
			bricks	bricks texture crops
			lava	lava texture crops

Table A1: Overview of factors $\mathcal{F}_i$, respective factor spaces $\mathcal{S}_i$, number of classes $\mathcal{N}_i$, factor classes $\mathcal{C}_{i,j}$ and respective factor space regions $\mathcal{S}_{i,j}$ used in this work.

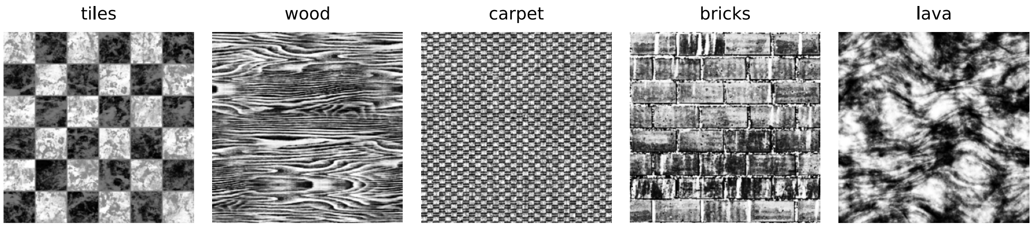


Figure A1: Overview of the textures used in this work. All textures are from `CC0Textures.com`, licensed under CC0 1.0 Universal.

A.2. Baselines

We train all baseline architectures using standard cross entropy loss and Adam optimizer [18] using early stopping. All test accuracies reported in this work correspond to the network with lowest validation loss. Batch size and learning rate were optimized on the ZSO study and weight initialization is fixed across all experiments. The output size of all the architectures are modified to predict the factor classes in our studies.

Following ASR, we use a variant of their U-Net architecture for the lens and ResNet18 as feature extractor. Instead of an auxiliary task, we train both the lens and feature extractor using the classification task directly. The lens is trained using least likely adversarial loss using the classification objective and also the reconstruction loss, both with equal weighting. Output of lens is used as inputs for training the feature extractor.

Both VAE methods use architectures similar to [25] with a latent size of 12. We train both VAE with batch size 64 and Adam optimizer [18] with learning rate 0.0001. For the Factor-VAE, we use $\gamma = 20$ and train the discriminator with learning rate 0.00001 throughout our experiments. Some examples of latent traversals from VAE and Factor-VAE are shown in Fig. A2. We experimented with β-VAEs but observed, that the reconstructions of are poor when we use higher β in our settings. One potential reason is the discretely distributed positions in the trained datasets. The `position` factor of DiagViB-6 has nine possible values with large gap among them violating the Gaussian distribution assumption in the KL-divergence term during training. To empirically demonstrate this claim, we train β-VAEs with different β on two settings: (1) with `position` factor freely assigned to three different values (2) with `position` factor fixed to the center of the images. According to the results shown in Fig. A3, the reconstructions from the training with fixed `position` are significantly better, supporting the aforementioned argument.

A.3. ZSO and ZGO

The mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO together with the mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ on the ZGO study for all baselines are shown in Fig. A5 and A6. Additionally, the aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the ZSO and ZGO studies are presented in Tab. A2 and A3.

From the exemplar VAE decoder reconstructions (Fig. A3) and ASR lens outputs (Fig. A4) we find that the high-frequency texture information is not preserved, leading to the observed low classification accuracies on the texture task.

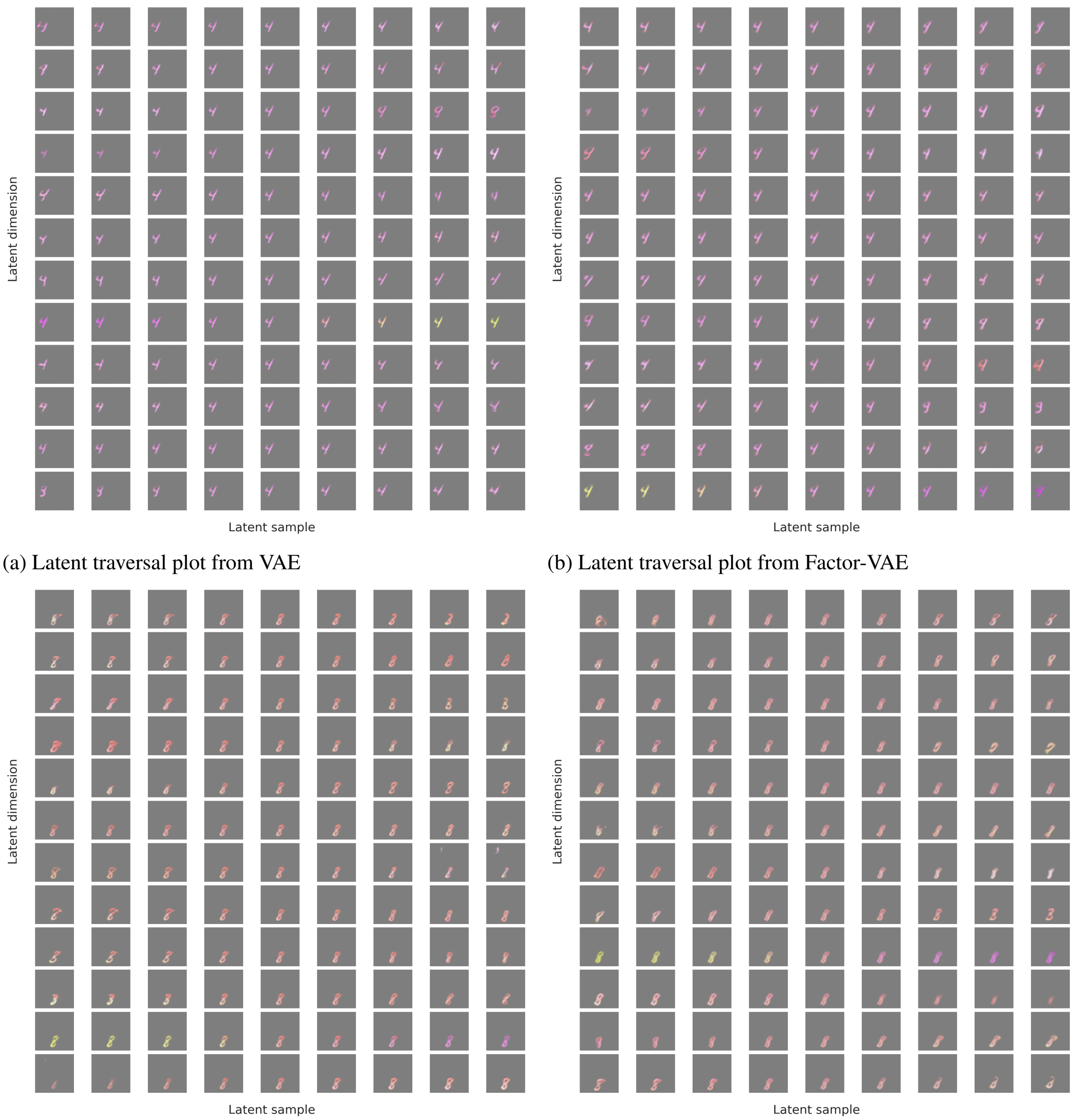
A.4. Compositional-based generalization opportunities

The mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO together with the mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ on the CGO and ZGO studies for all baselines are shown in Fig. A5 and A6. Additionally, the aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the CGO studies are presented in Tab. A4 to A6.

A.5. Frequency-based generalization opportunities

The mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO together with the mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ on the FGO and ZGO studies for all baselines are shown in Fig. A7 and A8. Additionally, the aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the FGO studies are presented in Tab. A7 to A9.

As discussed in Sec. 5, we find that ASR fails to exploit the presented GO for `hue` and `lightness`. This is likely attributed to the pixel-wise reconstruction loss that is used as a regularizer in the ASR objective. We present qualitative results of this behaviour in Fig. A9.



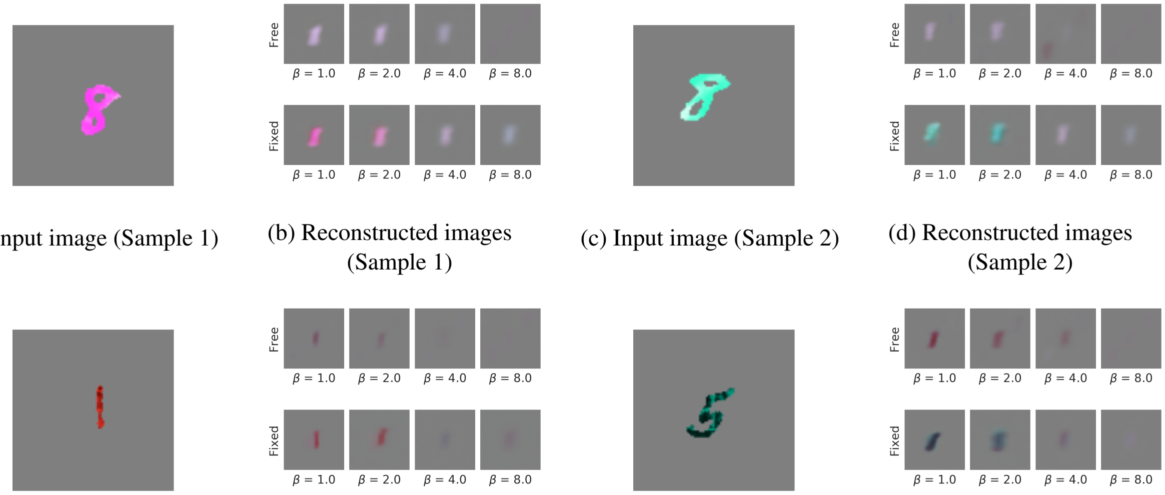
(a) Latent traversal plot from VAE

(b) Latent traversal plot from Factor-VAE

(c) Latent traversal plot from VAE

(d) Latent traversal plot from Factor-VAE

Figure A2: Comparison of latent traversals from VAE and Factor-VAE



(a) Input image (Sample 1)

(b) Reconstructed images (Sample 1)

(c) Input image (Sample 2)

(d) Reconstructed images (Sample 2)

(e) Input image (Sample 3)

(f) Reconstructed images (Sample 3)

(g) Input image (Sample 4)

(h) Reconstructed images (Sample 4)

Figure A3: Reconstruction images from the β-VAE networks trained with datasets of free and fixed positions from different β values

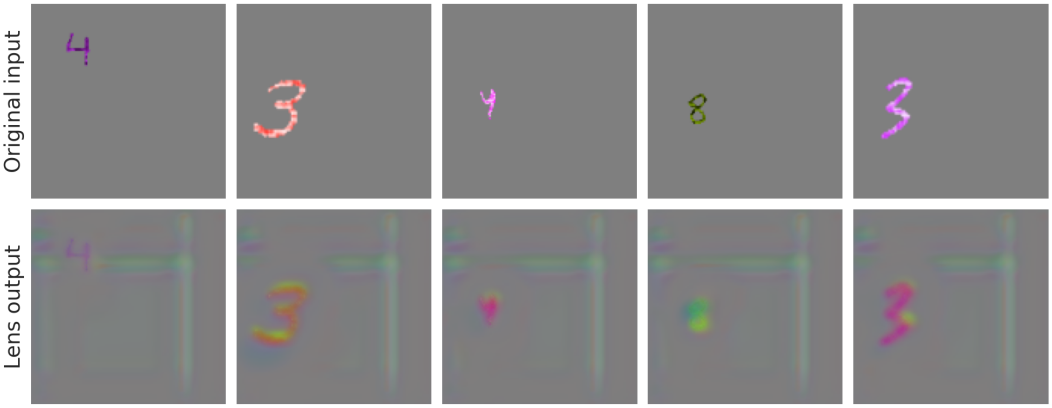


Figure A4: Examples of Automatic shortcut removal Lens output for ZSO study with factor `texture`.

	position	hue	lightness	scale	shape	texture
RN18	100 ± 0	100 ± 0	99 ± 0	99 ± 0	100 ± 0	62 ± 6
RN50	100 ± 0	100 ± 0	89 ± 9	99 ± 0	100 ± 0	77 ± 3
RN50-IN	99 ± 0	99 ± 0	94 ± 1	74 ± 2	83 ± 3	54 ± 1
AlexNet	100 ± 0	100 ± 0	98 ± 1	97 ± 0	99 ± 0	68 ± 3
DenseNet	100 ± 0	100 ± 0	99 ± 0	99 ± 0	100 ± 0	70 ± 3
WRN	100 ± 0	100 ± 0	99 ± 0	99 ± 0	100 ± 0	86 ± 2
ASR (RN18)	100 ± 0	77 ± 6	82 ± 2	92 ± 1	91 ± 1	45 ± 1
Factor-VAE	100 ± 0	97 ± 1	93 ± 2	84 ± 1	91 ± 2	34 ± 0
VAE	100 ± 0	98 ± 1	94 ± 1	86 ± 2	93 ± 2	34 ± 0

Table A2: FAAvg and respecive standard error for all baselines on the ZSO study.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	99 ± 1	74 ± 2	0 ± 0	57 ± 4	31 ± 5	33 ± 2	0 ± 0	41 ± 2	0 ± 0	2 ± 1	0 ± 0
RN50	100 ± 0	100 ± 0	72 ± 2	0 ± 0	58 ± 4	27 ± 1	32 ± 2	0 ± 0	42 ± 2	0 ± 0	3 ± 1	0 ± 0
RN50-IN	68 ± 3	18 ± 2	85 ± 2	64 ± 3	72 ± 4	50 ± 6	28 ± 2	4 ± 1	33 ± 5	8 ± 4	10 ± 1	1 ± 0
AlexNet	100 ± 0	100 ± 0	72 ± 2	0 ± 0	62 ± 6	31 ± 5	32 ± 2	0 ± 0	39 ± 2	0 ± 0	3 ± 1	0 ± 0
DenseNet	98 ± 1	88 ± 6	74 ± 3	7 ± 3	61 ± 6	34 ± 8	34 ± 1	0 ± 0	40 ± 2	0 ± 0	2 ± 0	0 ± 0
WRN	99 ± 1	96 ± 3	73 ± 3	4 ± 3	61 ± 6	37 ± 8	32 ± 2	0 ± 0	41 ± 2	0 ± 0	2 ± 1	0 ± 0
ASR (RN18)	99 ± 0	96 ± 1	27 ± 3	1 ± 0	43 ± 2	25 ± 0	51 ± 2	0 ± 0	60 ± 1	1 ± 0	8 ± 0	0 ± 0
Factor-VAE	100 ± 0	100 ± 0	65 ± 4	0 ± 0	64 ± 4	27 ± 0	32 ± 1	0 ± 0	35 ± 2	0 ± 0	4 ± 0	0 ± 0
VAE	100 ± 0	100 ± 0	65 ± 4	0 ± 0	64 ± 4	26 ± 0	33 ± 2	0 ± 0	37 ± 3	0 ± 0	4 ± 0	0 ± 0

Table A3: FAAvg and FAMin together with their respective standard errors for all baselines on the ZGO study.

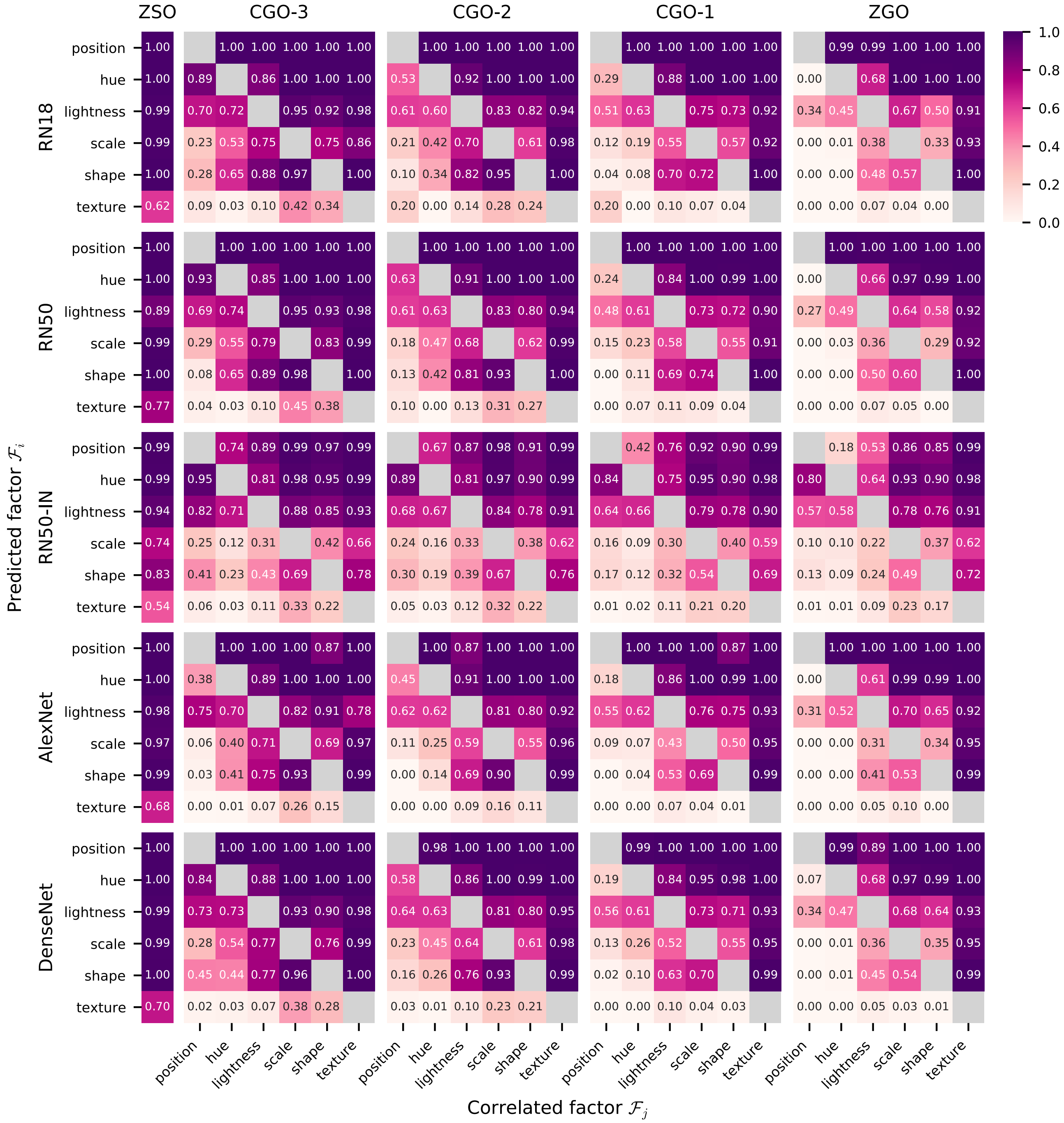


Figure A5: Mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO study and mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ for the RN18, RN50, RN50-IN, AlexNet and DenseNet on the CGO and ZGO studies.

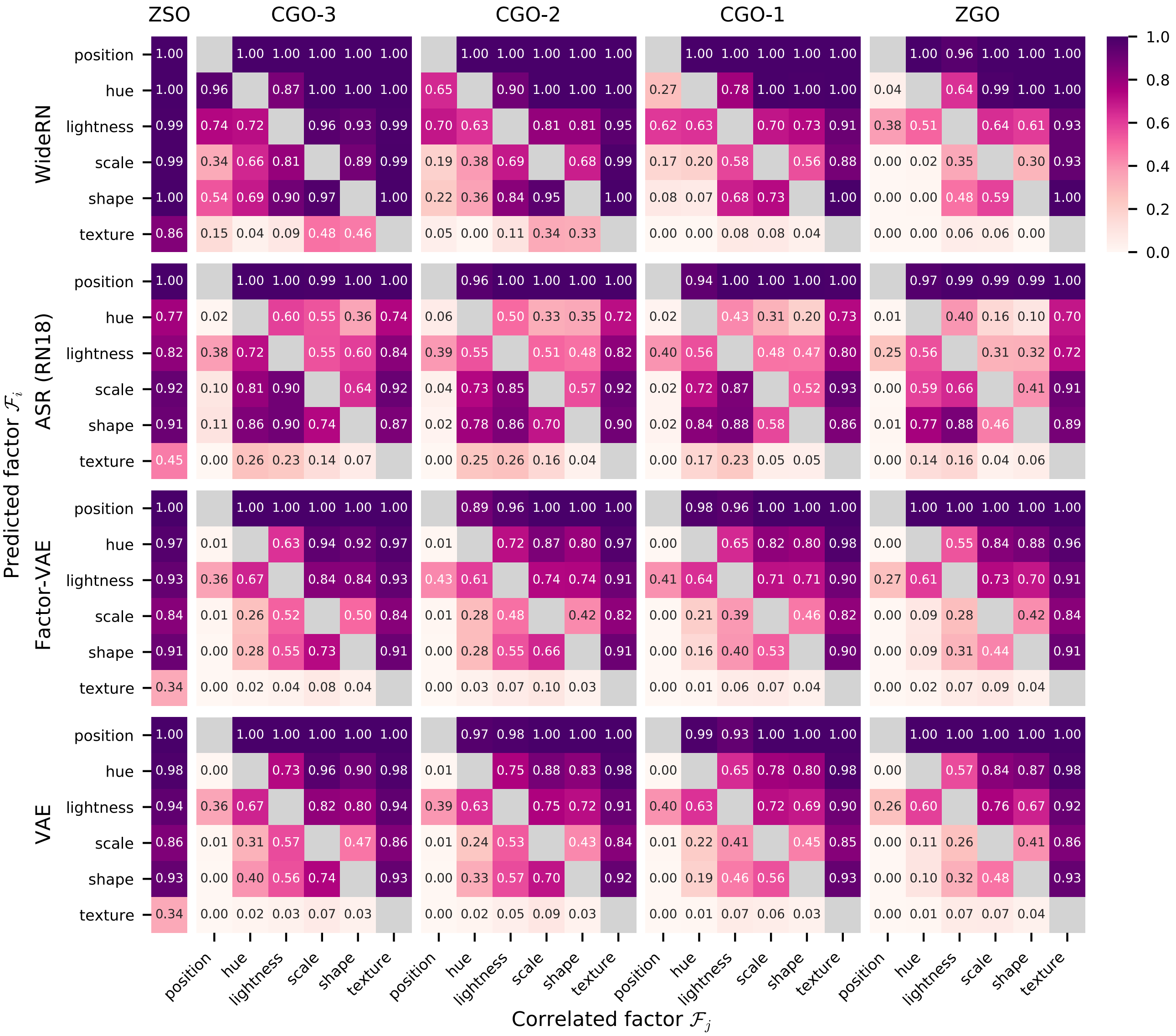


Figure A6: Mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO study and mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ for the WideRN, ASR (RN18), Factor-VAE and VAE on the CGO and ZGO studies.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	83 ± 1	29 ± 5	71 ± 8	49 ± 9	47 ± 3	6 ± 2	51 ± 3	2 ± 1	8 ± 2	0 ± 0
RN50	100 ± 0	100 ± 0	82 ± 1	24 ± 6	69 ± 8	46 ± 8	48 ± 4	9 ± 3	51 ± 3	0 ± 0	6 ± 2	0 ± 0
RN50-IN	80 ± 2	42 ± 4	88 ± 2	73 ± 3	75 ± 5	59 ± 8	31 ± 3	8 ± 2	37 ± 6	12 ± 5	11 ± 1	1 ± 0
AlexNet	97 ± 2	87 ± 12	81 ± 2	18 ± 7	72 ± 8	52 ± 10	41 ± 3	2 ± 1	45 ± 3	0 ± 0	3 ± 0	0 ± 0
DenseNet	100 ± 0	99 ± 1	79 ± 1	19 ± 1	71 ± 8	50 ± 10	48 ± 4	9 ± 3	49 ± 2	2 ± 1	3 ± 0	0 ± 0
WRN	100 ± 0	100 ± 0	81 ± 2	27 ± 4	72 ± 9	51 ± 10	48 ± 3	10 ± 3	51 ± 3	4 ± 2	4 ± 1	0 ± 0
ASR (RN18)	99 ± 1	94 ± 3	34 ± 3	2 ± 1	54 ± 4	40 ± 5	61 ± 2	2 ± 1	64 ± 3	2 ± 1	10 ± 1	0 ± 0
Factor-VAE	99 ± 0	93 ± 2	65 ± 3	0 ± 0	67 ± 6	41 ± 5	38 ± 3	0 ± 0	40 ± 3	0 ± 0	4 ± 0	0 ± 0
VAE	98 ± 1	92 ± 3	64 ± 3	0 ± 0	67 ± 6	40 ± 6	39 ± 3	1 ± 0	43 ± 4	0 ± 0	3 ± 0	0 ± 0

Table A4: Aggregated benchmark metrics FAAvg$_i$ and FAMin$_i$ for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the CGO-1 study.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	89 ± 4	53 ± 16	76 ± 11	51 ± 14	59 ± 5	21 ± 6	64 ± 2	4 ± 2	17 ± 4	0 ± 0
RN50	100 ± 0	100 ± 0	91 ± 4	59 ± 17	76 ± 10	56 ± 14	59 ± 4	18 ± 7	66 ± 3	13 ± 5	16 ± 4	0 ± 0
RN50-IN	88 ± 2	67 ± 6	91 ± 3	80 ± 4	77 ± 8	63 ± 12	35 ± 5	15 ± 4	46 ± 6	18 ± 5	15 ± 2	2 ± 1
AlexNet	97 ± 2	87 ± 12	87 ± 2	45 ± 7	75 ± 9	55 ± 14	49 ± 4	11 ± 5	55 ± 2	0 ± 0	7 ± 2	0 ± 0
DenseNet	100 ± 0	98 ± 2	89 ± 4	56 ± 14	77 ± 10	58 ± 15	58 ± 5	23 ± 4	62 ± 2	8 ± 3	12 ± 3	0 ± 0
WRN	100 ± 0	100 ± 0	91 ± 4	62 ± 17	78 ± 11	58 ± 14	59 ± 4	19 ± 3	67 ± 3	11 ± 6	17 ± 4	0 ± 0
ASR (RN18)	99 ± 1	96 ± 3	39 ± 6	6 ± 6	55 ± 7	39 ± 9	62 ± 2	4 ± 2	65 ± 3	2 ± 1	14 ± 2	0 ± 0
Factor-VAE	97 ± 2	89 ± 6	67 ± 4	1 ± 0	68 ± 8	43 ± 10	40 ± 4	1 ± 0	48 ± 3	0 ± 0	5 ± 1	0 ± 0
VAE	99 ± 1	97 ± 3	69 ± 4	1 ± 0	68 ± 7	39 ± 8	41 ± 4	1 ± 0	51 ± 4	0 ± 0	4 ± 0	0 ± 0

Table A5: Aggregated benchmark metrics FAAvg$_i$ and FAMin$_i$ for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the CGO-2 study.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	95 ± 3	81 ± 9	86 ± 6	65 ± 13	62 ± 7	22 ± 3	76 ± 3	28 ± 7	19 ± 4	1 ± 0
RN50	100 ± 0	100 ± 0	96 ± 2	85 ± 8	86 ± 7	66 ± 17	69 ± 7	29 ± 9	72 ± 2	8 ± 3	20 ± 4	1 ± 0
RN50-IN	92 ± 2	74 ± 7	94 ± 1	81 ± 5	84 ± 5	70 ± 10	35 ± 5	12 ± 3	51 ± 8	23 ± 10	15 ± 2	3 ± 1
AlexNet	97 ± 2	87 ± 12	86 ± 2	38 ± 7	79 ± 7	54 ± 11	56 ± 7	6 ± 2	62 ± 4	3 ± 2	10 ± 2	0 ± 0
DenseNet	100 ± 0	100 ± 0	94 ± 2	80 ± 6	86 ± 7	71 ± 13	66 ± 7	23 ± 4	72 ± 4	29 ± 11	16 ± 3	1 ± 0
WRN	100 ± 0	100 ± 0	97 ± 2	87 ± 6	87 ± 6	69 ± 13	74 ± 6	32 ± 7	82 ± 3	48 ± 9	24 ± 6	2 ± 0
ASR (RN18)	100 ± 0	99 ± 0	45 ± 1	2 ± 1	62 ± 8	34 ± 8	67 ± 3	10 ± 5	69 ± 3	11 ± 7	14 ± 1	0 ± 0
Factor-VAE	100 ± 0	100 ± 0	70 ± 3	1 ± 1	73 ± 6	36 ± 7	43 ± 4	1 ± 1	49 ± 3	0 ± 0	4 ± 1	0 ± 0
VAE	100 ± 0	100 ± 0	72 ± 3	0 ± 0	72 ± 7	36 ± 8	44 ± 5	1 ± 0	53 ± 4	0 ± 0	3 ± 1	0 ± 0

Table A6: Aggregated benchmark metrics FAAvg$_i$ and FAMin$_i$ for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the CGO-3 study.

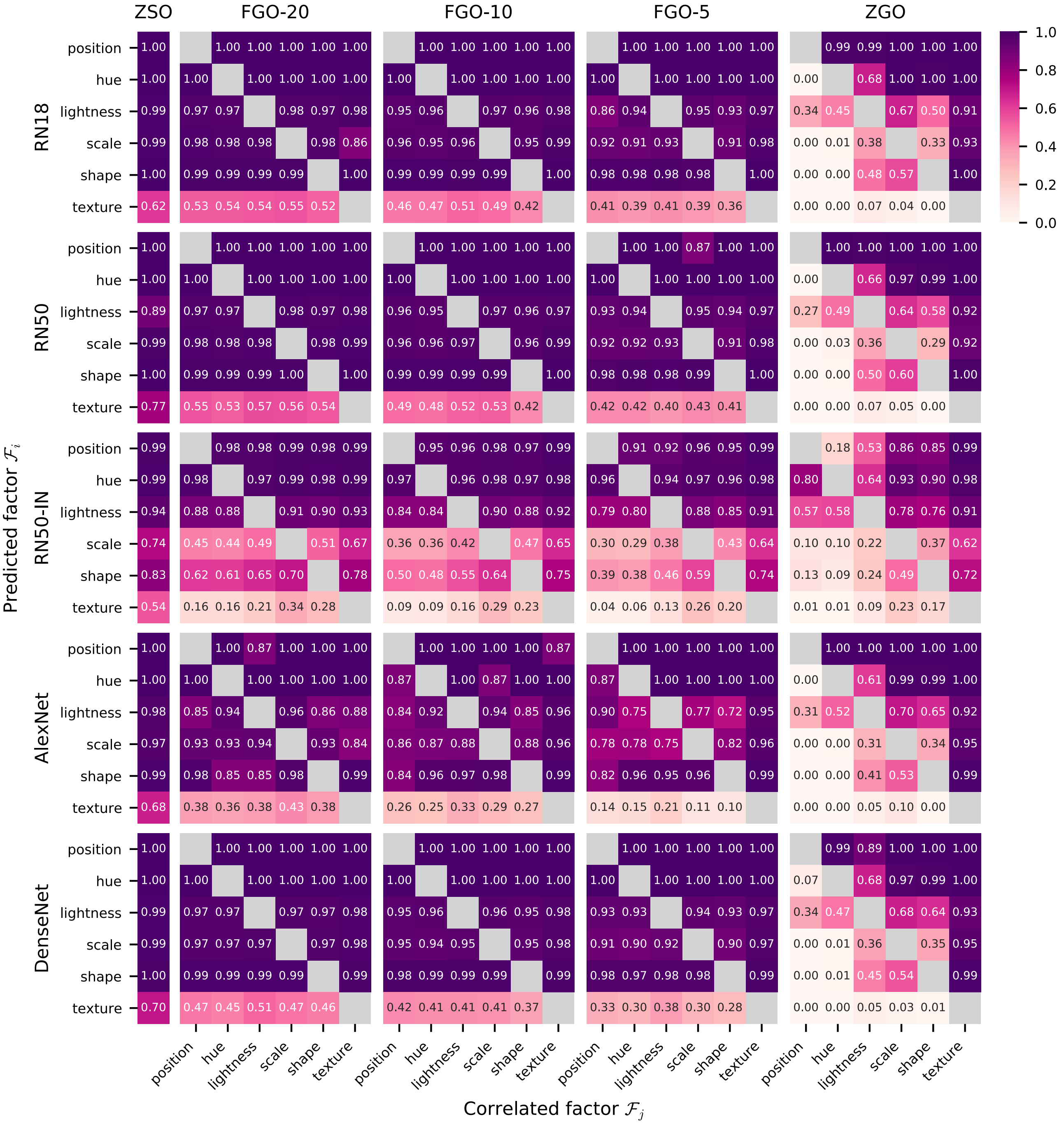


Figure A7: Mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO study and mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ for the RN18, RN50, RN50-IN, AlexNet and DenseNet on the FGO and ZGO studies.

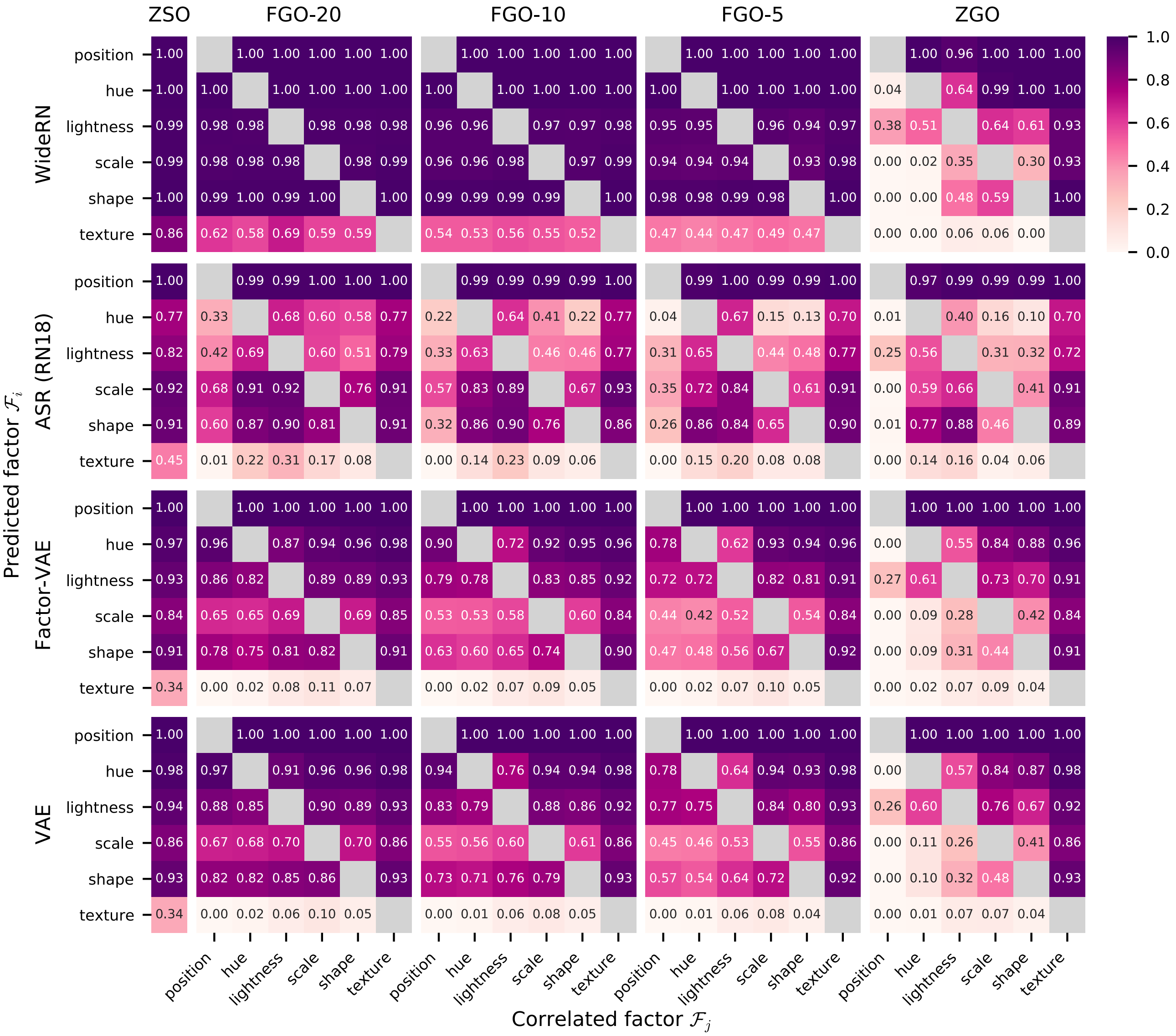


Figure A8: Mean accuracies P_i for all factors $\mathcal{F}_i$ on the ZSO study and mean accuracies $\mathrm{P}_{i,j}$ for all factor pairings $(\mathcal{F}_i, \mathcal{F}_j), i \neq j$ for the WideRN, ASR (RN18), Factor-VAE and VAE on the FGO and ZGO studies.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	100 ± 0	100 ± 0	93 ± 2	85 ± 8	93 ± 1	89 ± 1	99 ± 0	98 ± 0	39 ± 2	35 ± 3
RN50	97 ± 2	87 ± 12	100 ± 0	100 ± 0	95 ± 1	93 ± 2	93 ± 1	90 ± 1	99 ± 0	98 ± 0	41 ± 2	36 ± 1
RN50-IN	95 ± 1	91 ± 2	96 ± 1	94 ± 1	85 ± 3	78 ± 4	41 ± 4	28 ± 4	51 ± 8	37 ± 9	14 ± 1	4 ± 1
AlexNet	100 ± 0	100 ± 0	97 ± 2	87 ± 12	82 ± 4	64 ± 8	82 ± 3	69 ± 8	94 ± 3	82 ± 11	14 ± 2	6 ± 3
DenseNet	100 ± 0	100 ± 0	100 ± 0	100 ± 0	94 ± 1	92 ± 2	92 ± 1	88 ± 1	98 ± 0	97 ± 1	32 ± 2	26 ± 2
WRN	100 ± 0	100 ± 0	100 ± 0	100 ± 0	95 ± 1	94 ± 1	95 ± 1	92 ± 1	98 ± 0	98 ± 0	47 ± 3	42 ± 2
ASR (RN18)	99 ± 0	98 ± 0	34 ± 3	1 ± 1	53 ± 4	28 ± 1	69 ± 2	35 ± 6	70 ± 2	26 ± 7	10 ± 1	0 ± 0
Factor-VAE	100 ± 0	100 ± 0	84 ± 4	61 ± 9	80 ± 5	69 ± 7	55 ± 4	40 ± 4	62 ± 5	44 ± 7	5 ± 0	0 ± 0
VAE	100 ± 0	100 ± 0	86 ± 3	63 ± 8	82 ± 4	74 ± 5	57 ± 4	42 ± 4	68 ± 5	52 ± 6	4 ± 0	0 ± 0

Table A7: Aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the FGO-5 study.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	100 ± 0	100 ± 0	96 ± 1	95 ± 1	96 ± 1	95 ± 1	99 ± 0	99 ± 0	47 ± 3	41 ± 4
RN50	100 ± 0	100 ± 0	100 ± 0	100 ± 0	96 ± 1	95 ± 1	97 ± 0	95 ± 1	99 ± 0	99 ± 0	49 ± 3	41 ± 2
RN50-IN	97 ± 1	95 ± 1	97 ± 1	96 ± 1	87 ± 2	83 ± 3	45 ± 3	35 ± 4	59 ± 7	48 ± 9	17 ± 2	8 ± 1
AlexNet	97 ± 2	87 ± 12	95 ± 3	73 ± 15	90 ± 3	75 ± 9	89 ± 1	85 ± 2	95 ± 3	84 ± 11	28 ± 2	22 ± 2
DenseNet	100 ± 0	100 ± 0	100 ± 0	100 ± 0	96 ± 1	95 ± 1	95 ± 1	94 ± 1	99 ± 0	98 ± 1	40 ± 1	36 ± 1
WRN	100 ± 0	100 ± 0	100 ± 0	100 ± 0	97 ± 1	96 ± 1	97 ± 0	95 ± 0	99 ± 0	99 ± 0	54 ± 5	52 ± 5
ASR (RN18)	99 ± 0	98 ± 0	45 ± 4	11 ± 4	53 ± 3	31 ± 2	78 ± 1	50 ± 3	74 ± 4	32 ± 7	10 ± 0	0 ± 0
Factor-VAE	100 ± 0	100 ± 0	89 ± 3	72 ± 9	83 ± 4	76 ± 6	62 ± 4	51 ± 5	71 ± 5	60 ± 5	5 ± 0	0 ± 0
VAE	100 ± 0	100 ± 0	91 ± 2	76 ± 7	86 ± 3	79 ± 5	64 ± 4	53 ± 4	78 ± 5	70 ± 6	4 ± 0	0 ± 0

Table A8: Aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the FGO-10 study.

	position		hue		lightness		scale		shape		texture	
	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin	FAAvg	FAMin
RN18	100 ± 0	100 ± 0	100 ± 0	100 ± 0	98 ± 1	97 ± 1	95 ± 2	84 ± 11	99 ± 0	99 ± 0	54 ± 5	51 ± 5
RN50	100 ± 0	100 ± 0	100 ± 0	100 ± 0	98 ± 1	97 ± 1	98 ± 0	98 ± 0	100 ± 0	99 ± 0	55 ± 5	52 ± 5
RN50-IN	98 ± 0	98 ± 1	98 ± 0	97 ± 1	90 ± 2	88 ± 2	51 ± 3	44 ± 3	67 ± 6	60 ± 7	23 ± 2	15 ± 2
AlexNet	97 ± 2	87 ± 12	100 ± 0	100 ± 0	90 ± 4	77 ± 10	91 ± 3	80 ± 11	93 ± 5	85 ± 11	38 ± 2	34 ± 2
DenseNet	100 ± 0	100 ± 0	100 ± 0	100 ± 0	97 ± 1	96 ± 1	97 ± 0	97 ± 0	99 ± 0	99 ± 0	47 ± 3	43 ± 3
WRN	100 ± 0	100 ± 0	100 ± 0	100 ± 0	98 ± 0	97 ± 1	98 ± 0	97 ± 0	99 ± 0	99 ± 0	61 ± 5	57 ± 6
ASR (RN18)	100 ± 0	99 ± 0	59 ± 4	26 ± 8	60 ± 3	38 ± 2	84 ± 2	67 ± 5	82 ± 3	60 ± 7	16 ± 1	1 ± 1
Factor-VAE	100 ± 0	100 ± 0	94 ± 2	84 ± 5	88 ± 3	82 ± 5	71 ± 3	63 ± 3	81 ± 4	75 ± 5	5 ± 0	0 ± 0
VAE	100 ± 0	100 ± 0	96 ± 1	89 ± 4	89 ± 3	84 ± 4	72 ± 3	65 ± 3	85 ± 3	81 ± 4	5 ± 0	0 ± 0

Table A9: Aggregated benchmark metrics FAAvg_i and FAMin_i for all factors $\mathcal{F}_i, i \in \{0, 1, ..., 6\}$ together with their respective standard errors for all baselines on the FGO-20 study.

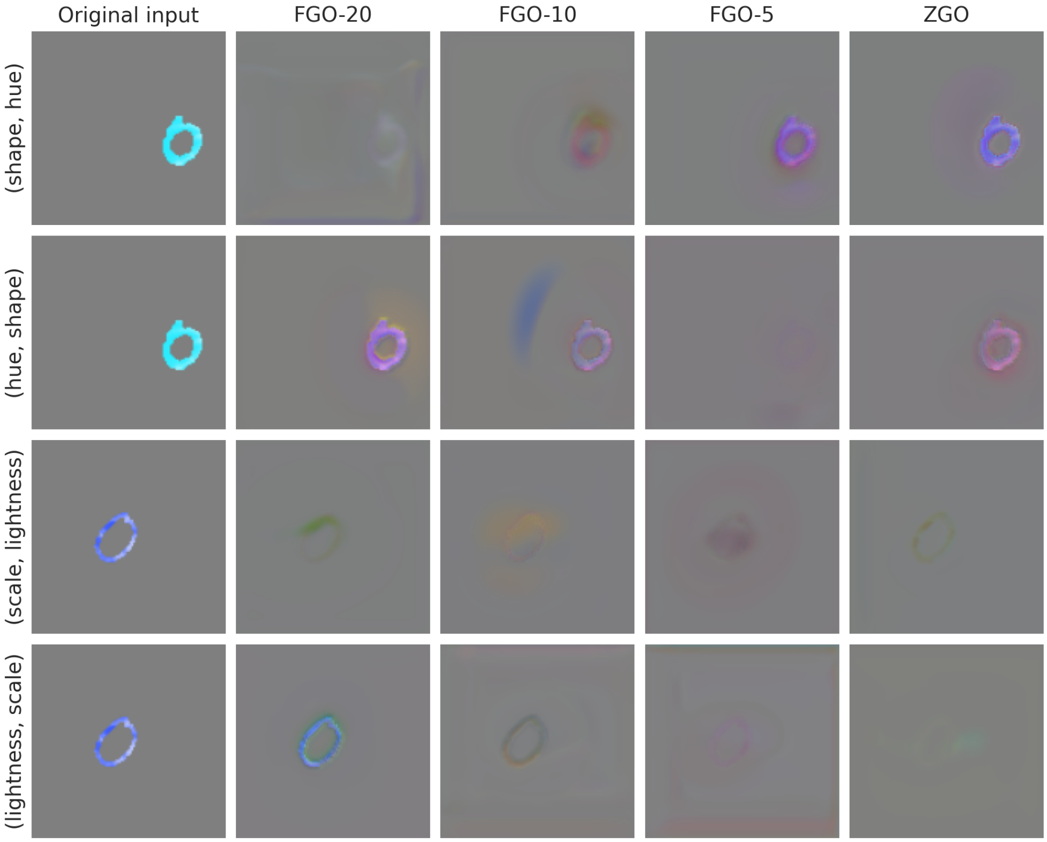


Figure A9: Examples of Automatic shortcut removal Lens output for different studies in our benchmark on datasets with following factor correlations: (`shape`, `hue`) (first row), (`hue`, `shape`) (second row), (`scale`, `lightness`) (third row), (`lightness`, `scale`) (fourth row).