A Detailed Classification Results

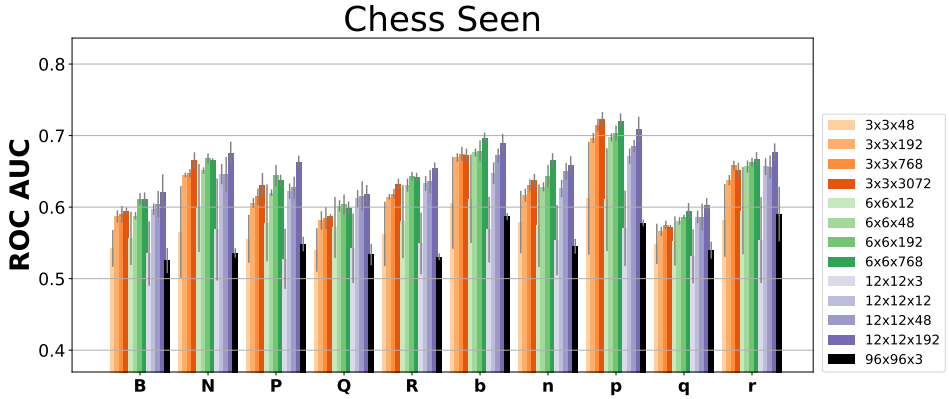

Figure A.1: Classification performance by class for the Chess dataset. The classifiers were trained on representations from samples that the CAE has seen during training. Configurations that have a common feature map size share the same color in the bar plot. Color intensity represents the amount of channels in the bottleneck (darker = more channels). Classification based on the raw inputs (baseline) is shown in black. Error bars indicate standard deviation over 9 runs (3 CAE seeds $\times$ 3 classifier seeds), except for the baseline where only the 3 classifier seeds are considered.

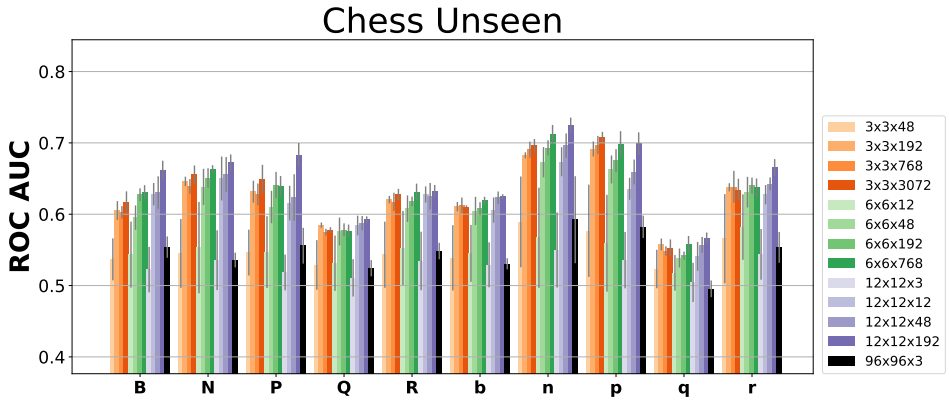

Figure A.2: Classification performance by class for the Pokemon dataset. The classifiers were trained on representations from samples that the CAE has not seen during training. Configurations that have a common feature map size share the same color in the bar plot. Color intensity represents the amount of channels in the bottleneck (darker = more channels). Classification based on the raw inputs (baseline) is shown in black. Error bars indicate standard deviation over 9 runs (3 CAE seeds $\times$ 3 classifier seeds), except for the baseline where only the 3 classifier seeds are considered. Missing bars indicate that the class was not present in the data.

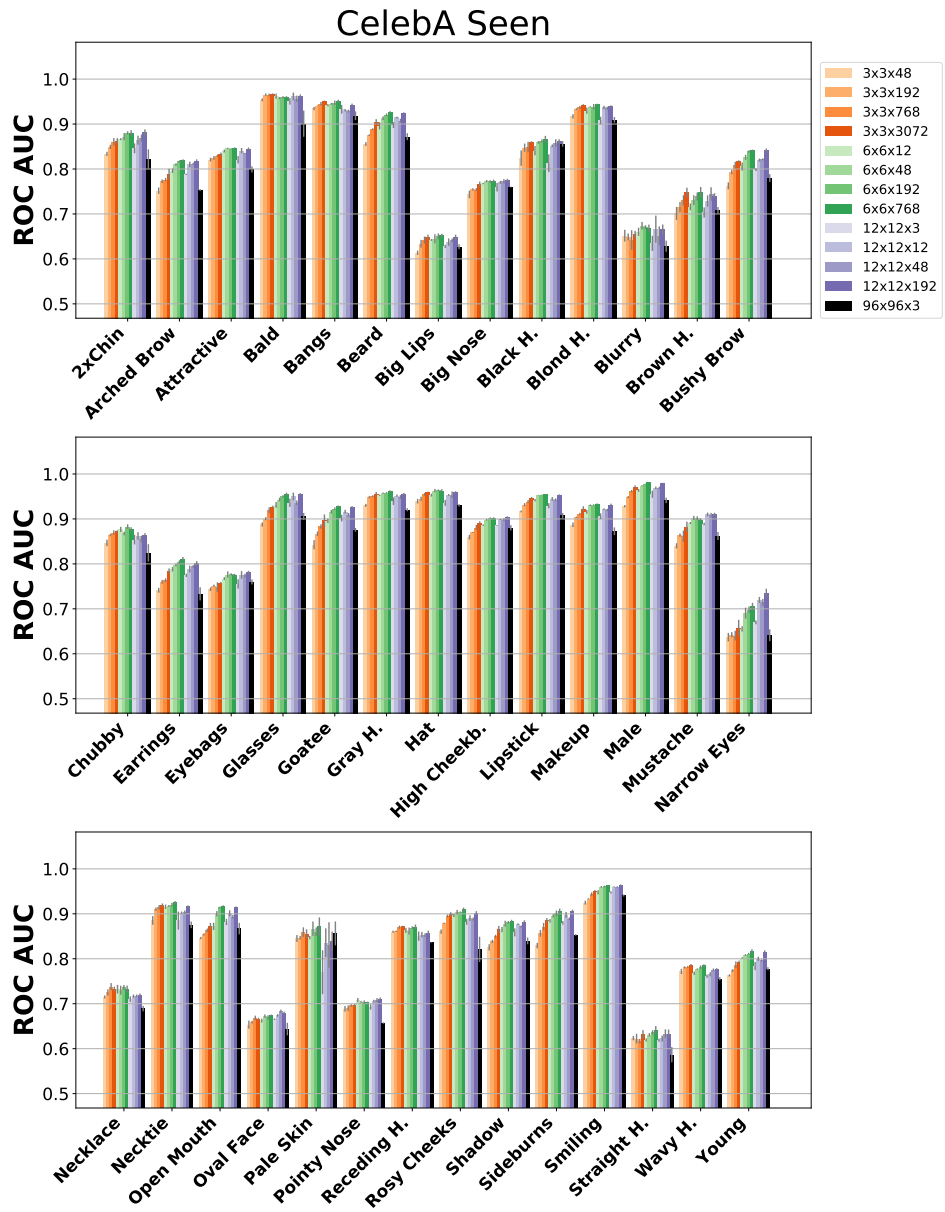

Figure A.3: Classification performance by class for the CelebA dataset. The classifiers were trained on representations from samples that the CAE has seen during training. Configurations that have a common feature map size share the same color in the bar plot. Color intensity represents the amount of channels in the bottleneck (darker = more channels). Classification based on the raw inputs (baseline) is shown in black. Error bars indicate standard deviation over 9 runs (3 CAE seeds $\times$ 3 classifier seeds), except for the baseline where only the 3 classifier seeds are considered.

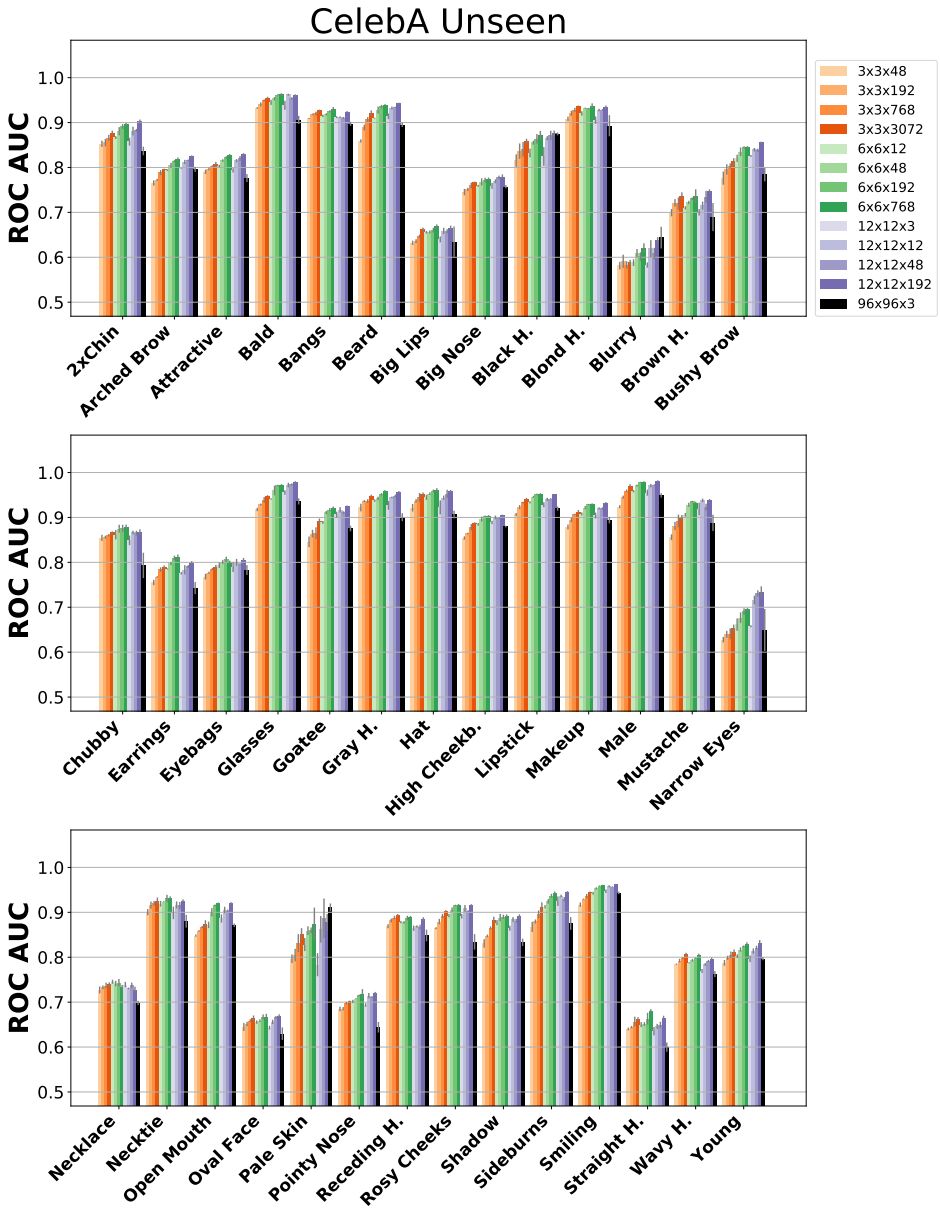

Figure A.4: Classification performance by class for the CelebA dataset. The classifiers were trained on representations from samples that the CAE has not seen during training. Configurations that have a common feature map size share the same color in the bar plot. Color intensity represents the amount of channels in the bottleneck (darker = more channels). Classification based on the raw inputs (baseline) is shown in black. Error bars indicate standard deviation over 9 runs (3 CAE seeds $\times$ 3 classifier seeds), except for the baseline where only the 3 classifier seeds are considered.

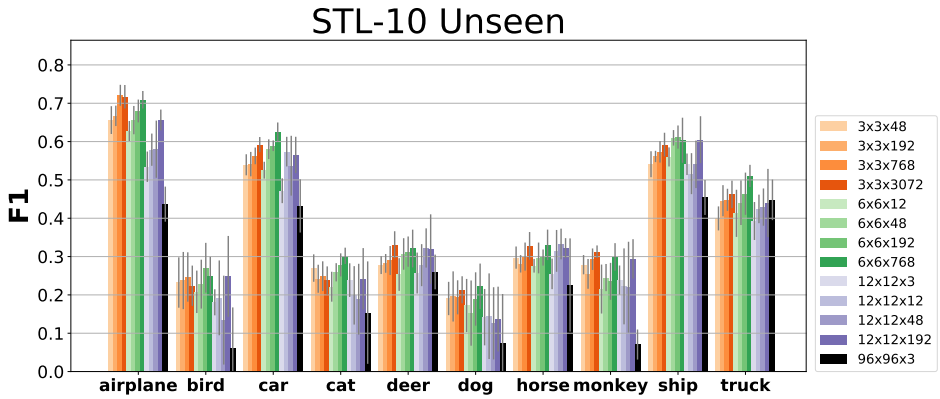

Figure A.5: Classification performance by class for the STL-10 dataset. The classifiers were trained on representations from samples that the CAE has not seen during training. Configurations that have a common feature map size share the same color in the bar plot. Color intensity represents the amount of channels in the bottleneck (darker = more channels). Classification based on the raw inputs (baseline) is shown in black. Error bars indicate standard deviation over 9 runs (3 CAE seeds $\times$ 3 classifier seeds), except for the baseline where only the 3 classifier seeds are considered.

B Reconstruction Samples

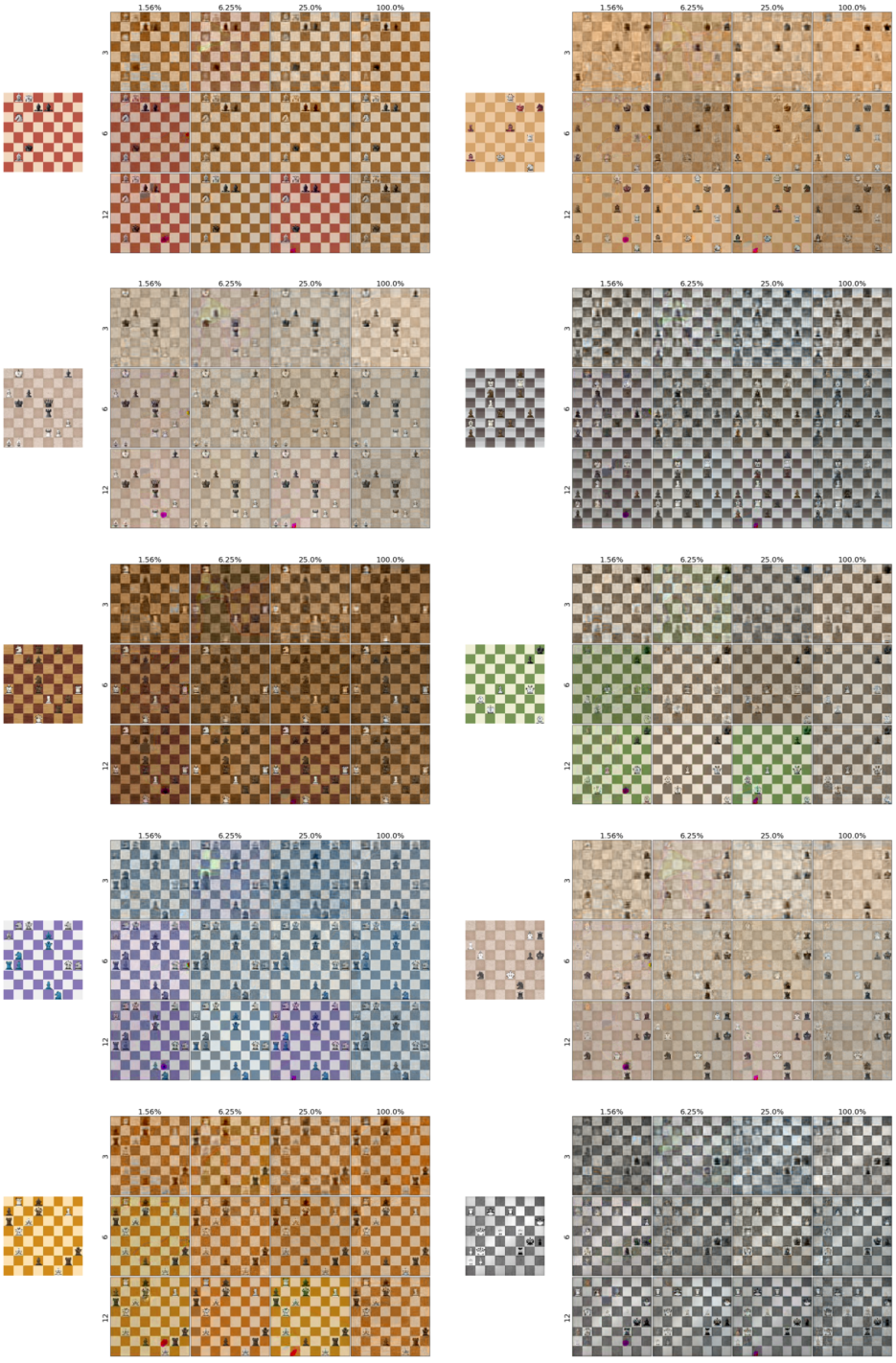

Figure B.1: Reconstructions of randomly picked samples from the Chess dataset. The left column contains samples from the training data, while on the right, we show samples from the test data. In each subfigure, the rows correspond to CAEs with the same bottleneck size (height, width), increasing from top to bottom. The columns group CAEs by the number of channels in the bottleneck, expressed as percentage relative to input given bottleneck size. The image to the left of each grid is the input image.

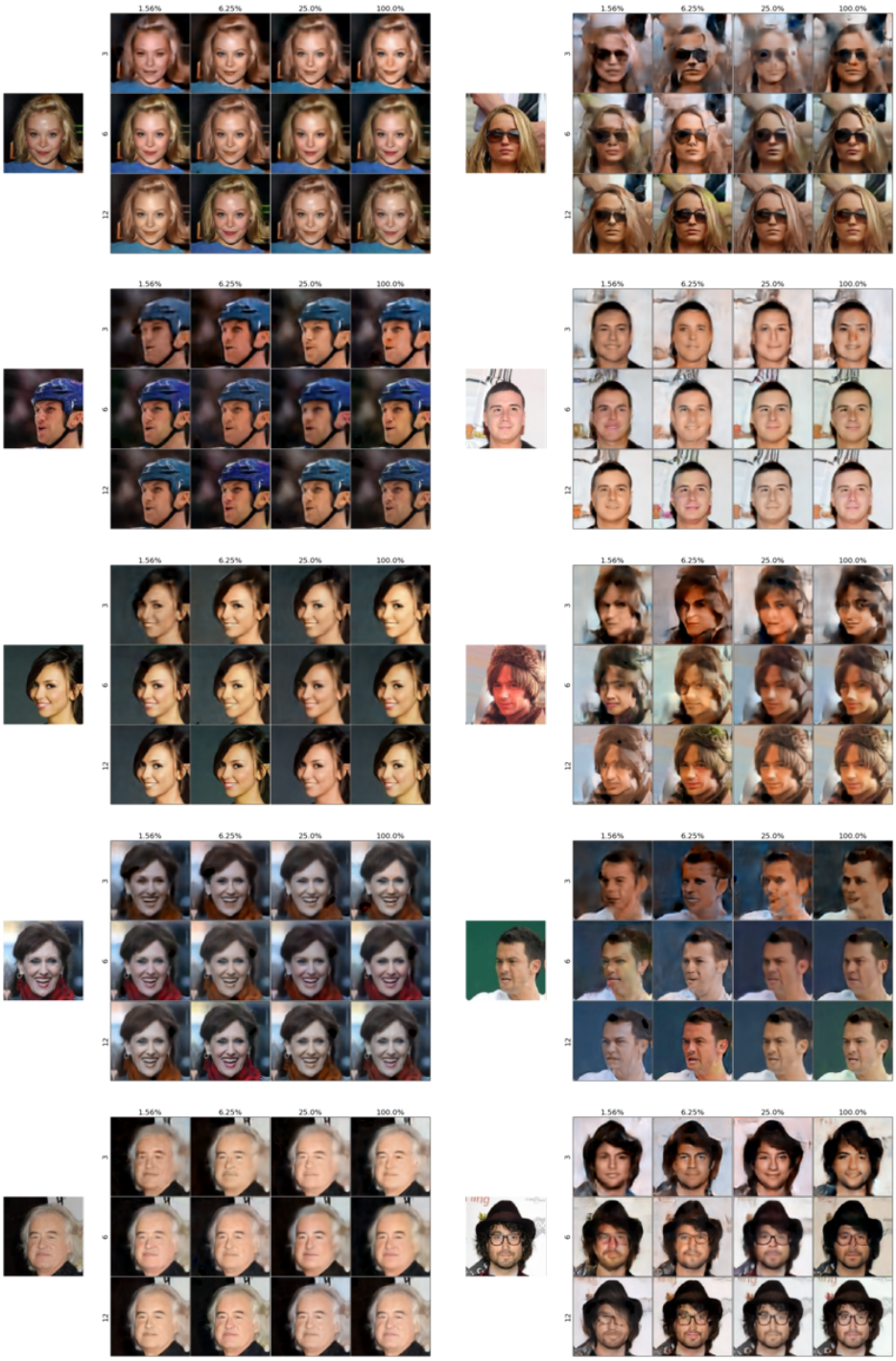

Figure B.2: Reconstructions of randomly picked samples from the CelebA dataset. The left column contains samples from the training data, while on the right, we show samples from the test data. In each subfigure, the rows correspond to CAEs with the same bottleneck size (height, width), increasing from top to bottom. The columns group CAEs by the number of channels in the bottleneck, expressed as percentage relative to input given bottleneck size. The image to the left of each grid is the input image.

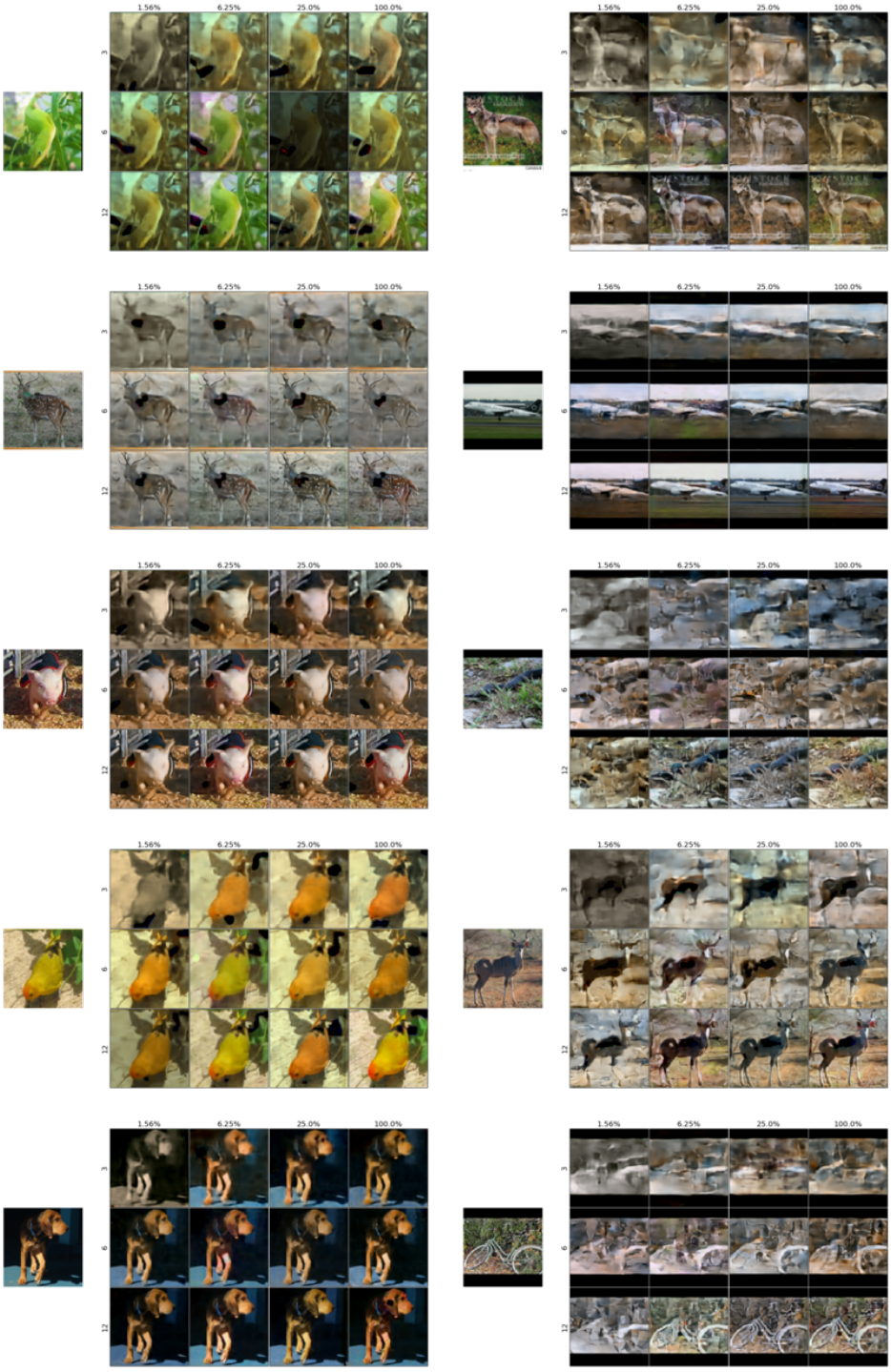

Figure B.3: Reconstructions of randomly picked samples from the STL-10 dataset. The left column contains samples from the training data, while on the right, we show samples from the test data. In each subfigure, the rows correspond to CAEs with the same bottleneck size (height, width), increasing from top to bottom. The columns group CAEs by the number of channels in the bottleneck, expressed as percentage relative to input given bottleneck size. The image to the left of each grid is the input image.

C Utility for Outlier Detection

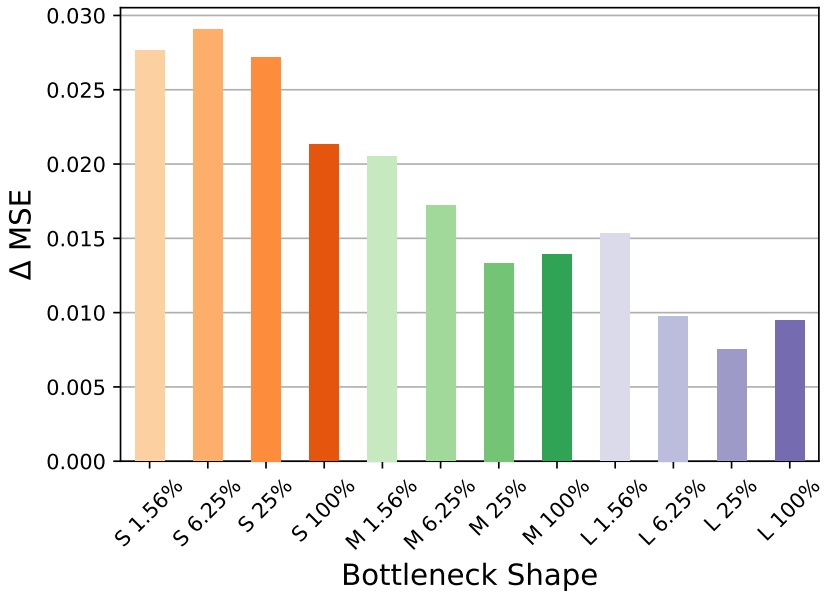

Figure C.1: Plot showing the difference in reconstruction MSE when passing unseen samples from CelebA and STL-10 to a CAE trained only on CelebA. Color of the bar distinguishes different bottleneck areas, while intensity represents the bottleneck depth.

D Wilcoxon Rank-Sum Test

dataset	grouped by	setting 1	setting 2	minimal test error	epoch @ minimal test error	epoch @ 0.001 training error
celeba	area	L	M	**0.00053**	0.08326	**0.00034**
		L	S	**0.00003**	**0.00048**	**0.00003**
		M	S	**0.00003**	**0.00426**	**0.00003**
	depth	$^1/_{64}$	$^1/_{16}$	0.20041	0.26969	0.62721
		$^1/_{64}$	$^1/_4$	0.08509	0.09340	0.69110
		$^1/_{64}$	1	0.12228	0.17110	0.45291
		$^1/_{16}$	$^1/_4$	0.56599	0.26969	0.89463
		$^1/_{16}$	1	0.35384	0.45291	0.72393
		$^1/_4$	1	0.96478	0.69110	0.62721
stl-10	area	L	M	**0.00009**	**0.00937**	0.07825
		L	S	**0.00003**	**0.00295**	**0.00005**
		M	S	**0.00003**	0.48842	**0.00043**
	depth	$^1/_{64}$	$^1/_{16}$	0.30988	0.17110	0.17110
		$^1/_{64}$	$^1/_4$	0.26969	0.06369	0.40154
		$^1/_{64}$	1	0.20041	**0.00172**	0.20041
		$^1/_{16}$	$^1/_4$	0.89463	0.42678	0.69110
		$^1/_{16}$	1	0.62721	0.20041	0.96478
		$^1/_4$	1	0.75728	0.82528	0.65884
chess	area	L	M	**0.00009**	**0.00014**	**0.02092**
		L	S	**0.00003**	**0.00053**	**0.00003**
		M	S	**0.00003**	**0.01531**	**0.00004**
	depth	$^1/_{64}$	$^1/_{16}$	0.89463	0.45291	0.89463
		$^1/_{64}$	$^1/_4$	0.56599	0.50780	0.96478
		$^1/_{64}$	1	0.45291	0.75728	0.82528
		$^1/_{16}$	$^1/_4$	0.62721	0.89463	0.96478
		$^1/_{16}$	1	0.62721	0.50780	0.82528
		$^1/_4$	1	0.96478	0.75728	0.82528

Table 1: P-values resulting from pairwise Wilcoxon rank-sum tests. This table is meant to accompany Figure 2 in the main paper. The test was applied to see whether the values of each variable (minimal test error, epoch at which the minimal test error occurred and first epoch to drop beneath 0.001 training error) are significantly different for different bottleneck areas and depths. We chose the Wilcoxon rank-sum test over the more common unpaired Student's t-test because we do not believe that its requirements are met (i.e. Gaussian distribution and identical standard deviation). Values below 0.05 are highlighted with bold font.