A APPENDIX

A.1 DATA PRE-PROCESSING AND AUGMENTATION

In all our experiments, we consider the same data augmentation for all networks (i.e. CNNs and CapsNets with different routing mechanisms). For SVHN, CIFAR-10, and CIFAR-100 training, we first pad pixels with four zeros and randomly crop the image to 32×32. For Tiny-ImageNet, images are rotated randomly by up to 20 degrees. For ImageNet, all training images are resized to 256×256, then randomly cropped to yield an input training patch of 224×224. For all datasets except SVHN, training images are also horizontally flipped with probability 0.5. Finally, all images are normalized to zero mean and unit standard deviation. During evaluation, we do not perform data augmentation for all models.

For the SmallNORB dataset and following the training protocol of **?**, all images are resized to 48×48. During training, each image is randomly cropped to yield an input training patch of 32×32. Then, we horizontally flip the image with probability 0.5, and randomly jitter the brightness and contrast of the images. During inference, we crop the center of the test image to yield a 32×32 patch.

A.2 EFFECT OF THE NUMBER OF HEADS

Table 1: The mean test classification error rates ($\pm$std.) of the proposed Trans-Caps with SAR as a function of the number of attention heads.

# of heads	SVHN	CIFAR-10	CIFAR-100
1	**3.02±0.14**	6.68±0.21	25.32±0.24
2	3.13±0.13	**6.56±0.16**	**25.17±0.21**
4	3.23±0.17	6.69±0.19	25.42±0.25
8	3.12±0.14	6.76±0.22	25.38±0.25
16	3.15±0.13	6.76±0.20	25.50±0.31

A.3 MODEL CONFIGURATIONS

The following tables depict the architectures and parameter setups for the trained models discussed in the paper. Note that ConvCaps and FCCaps layers represent the convolutional capsule layers and fully-connected capsule layers, respectively.

Table 2: Architecture of the proposed CapsNet with SAR for SVHN and CIFAR-10.

Name	Operation	Output size
Backbone ($A = 64$)	ResNet-20 (input_dim=3, output_dim=64)	$8 \times 8 \times 64$
`PrimaryCaps` ($B = 32$)	3×3 Conv, input_dim=64, output_dim=512, stride=1, padding=1 + reshape to 32, 4×4-dim capsules	$8 \times 8 \times 32 \times 4 \times 4$
`ConvCaps` ($C = 32$)	3×3 ConvCaps SAR to 32, 4×4-dim capsules, stride=2, padding=1	$4 \times 4 \times 32 \times 4 \times 4$
`ClassCaps`	FCCaps SAR to 10, 4×4-dim. capsules	$10 \times 4 \times 4$
`Classifier`	input_dim=16, output_dim=1, linear	10×1

Table 3: Architecture of the proposed CapsNet with SAR for CIFAR-100.

Name	Operation	Output size
Backbone ($A = 128$)	ResNet-32 (input_dim=3, output_dim=128)	$8 \times 8 \times 128$
`PrimaryCaps` ($B = 32$)	3×3 Conv, input_dim=128, output_dim=512, stride=1, padding=1 + reshape to 32, 4×4-dim capsules	$8 \times 8 \times 32 \times 4 \times 4$
`ConvCaps1` ($C = 32$)	3×3 ConvCaps SAR to 32, 4×4-dim capsules, stride=2, padding=1	$4 \times 4 \times 32 \times 4 \times 4$
`ConvCaps2` ($D = 32$)	4×4 ConvCaps SAR to 32, 4×4-dim capsules, stride=2, padding=1	$32 \times 4 \times 4$
`ClassCaps`	FCCaps SAR to 100, 4×4-dim. capsules	$100 \times 4 \times 4$
`Classifier`	input_dim=16, output_dim=1, linear	100×1

Table 4: Architecture of the proposed CapsNet with SAR for Tiny-ImageNet.

Name	Operation	Output size
Backbone ($A = 128$)	ResNet-32 (input_dim=3, output_dim=128)	$16 \times 16 \times 128$
`PrimaryCaps` ($B = 32$)	3×3 Conv, input_dim=128, output_dim=512, stride=1, padding=1 + reshape to 32, 4×4-dim capsules	$16 \times 16 \times 32 \times 4 \times 4$
`ConvCaps1` ($C = 32$)	3×3 ConvCaps SAR to 32, 4×4-dim capsules, stride=2, padding=1	$8 \times 8 \times 32 \times 4 \times 4$
`ConvCaps2` ($D = 64$)	8×8 ConvCaps SAR to 64, 4×4-dim capsules, stride=2, padding=1	$64 \times 4 \times 4$
`ClassCaps`	FCCaps SAR to 200, 4×4-dim. capsules	$200 \times 4 \times 4$
`Classifier`	input_dim=16, output_dim=1, linear	200×1

Table 5: Architecture of the proposed CapsNet with SAR for ImageNet.

Name	Operation	Output size
Backbone ($A = 1024$)	ResNet-50 (input_dim=3, output_dim=1024)	$14 \times 14 \times 1024$
`PrimaryCaps` ($B = 64$)	1×1 Conv, input_dim=1024, output_dim=1024, stride=1, padding=0 + reshape to 64, 4×4-dim capsules	$14 \times 14 \times 64 \times 4 \times 4$
`ConvCaps1` ($C = 128$)	3×3 ConvCaps SAR to 128, 4×4-dim capsules, stride=2, padding=1	$7 \times 7 \times 128 \times 4 \times 4$
`ConvCaps2` ($D = 128$)	3×3 ConvCaps SAR to 128, 4×4-dim capsules, stride=1, padding=0	$5 \times 5 \times 128 \times 4 \times 4$
`ClassCaps`	FCCaps SAR to 1000, 4×4-dim. capsules	$1000 \times 4 \times 4$
`Classifier`	input_dim=16, output_dim=1, linear	1000×1

Table 6: Architecture of the proposed CapsNet with SAR for SmallNORB.

Name	Operation	Output size
Backbone ($A = 64$)	5×5 Conv, input_dim=1, output_dim=64, stride=2, padding=1	$15 \times 15 \times 64$
`PrimaryCaps` ($B = 8$)	1×1 Conv, input_dim=64, output_dim=128, stride=1, padding=0 + reshape to 8, 4×4-dim capsules	$15 \times 15 \times 8 \times 4 \times 4$
`ConvCaps1` ($C = 16$)	3×3 ConvCaps SAR to 16, 4×4-dim capsules, stride=2, padding=0	$7 \times 7 \times 16 \times 4 \times 4$
`ConvCaps2` ($D = 16$)	3×3 ConvCaps SAR to 16, 4×4-dim. capsules, stride=1, padding=0	$5 \times 5 \times 16 \times 4 \times 4$
`ClassCaps`	FCCaps SAR to 5, 4×4-dim. capsules	$5 \times 4 \times 4$
`Classifier`	input_dim=16, output_dim=1, linear	5×1

Table 7: Architecture of the baseline CNN used in SmallNORB experiments.

Name	Operation	Output size
`Conv1`	5×5 Conv, input_dim=1, output_dim=64, stride=2, padding=1 + BN + ReLU	$15 \times 15 \times 64$
`Conv2`	1×1 Conv, input_dim=64, output_dim=128, stride=1, padding=0 + BN + ReLU	$15 \times 15 \times 128$
`Conv3`	3×3 Conv, input_dim=128, output_dim=256, stride=2, padding=0 + BN + ReLU	$7 \times 7 \times 256$
`Conv4`	3×3 Conv, input_dim=256, output_dim=256, stride=1, padding=0 + BN + ReLU	$5 \times 5 \times 256$
`AvgPool`	5×5 global average pooling + flatten	256
`Classifier`	input_dim=256, output_dim=5, linear	5