Appendix

A. Supplementary figures

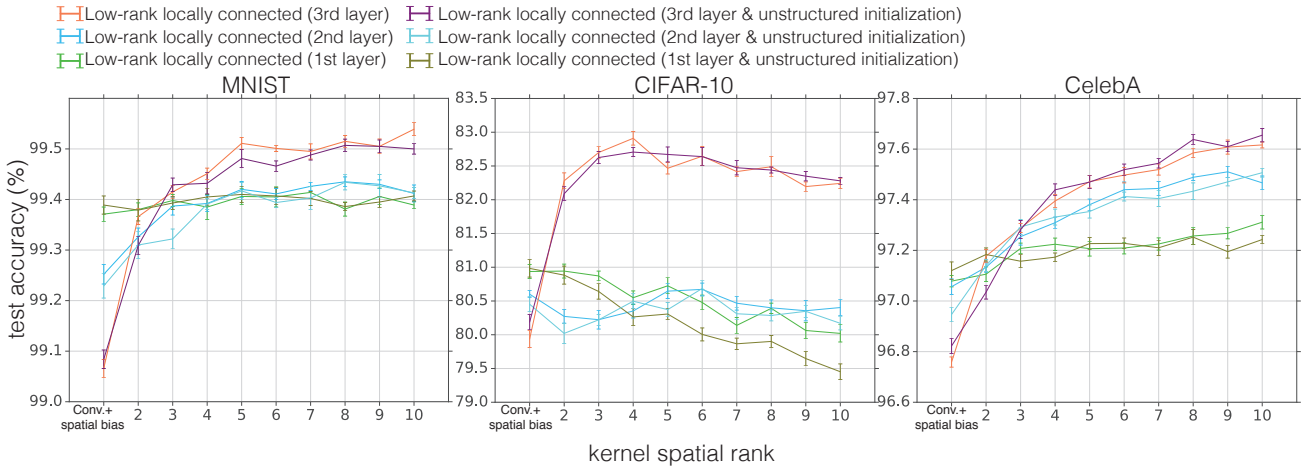

Figure Supp.1: **Structured vs unstructured initialization.** Top 1 accuracy similar to Figure 3. We study the effect of the structured initialization we used in our experiments for the LRLC layers (i.e., initialization to a convolution layer with a random kernel). In the structured initialization, we initialized the layer combining weights to constant equal to $1/\sqrt{\text{spatial rank}}$. We compared this initialization to a random initialization of the combining weights. Our results show that the structured initialization is generally quite similar to the unstructured initialization. Error bars indicate $\pm$ standard errors computed from training models from 10 different random initialization.

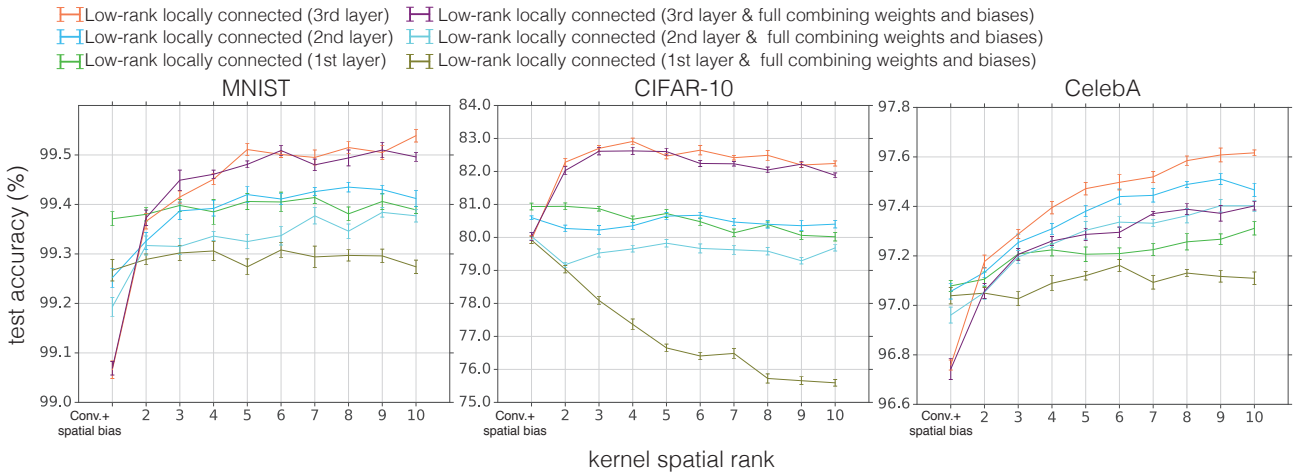

Figure Supp.2: **Factorized vs full combining weights and biases.** Top 1 accuracy similar to Figure 3. We study the effect of factorizing the combining weights and biases in Equations 3 and 5. We compare the performance of a LRLC layer with factorized weights and bias to a LRLC without this factorization. The layer with the factorization seems to perform better.

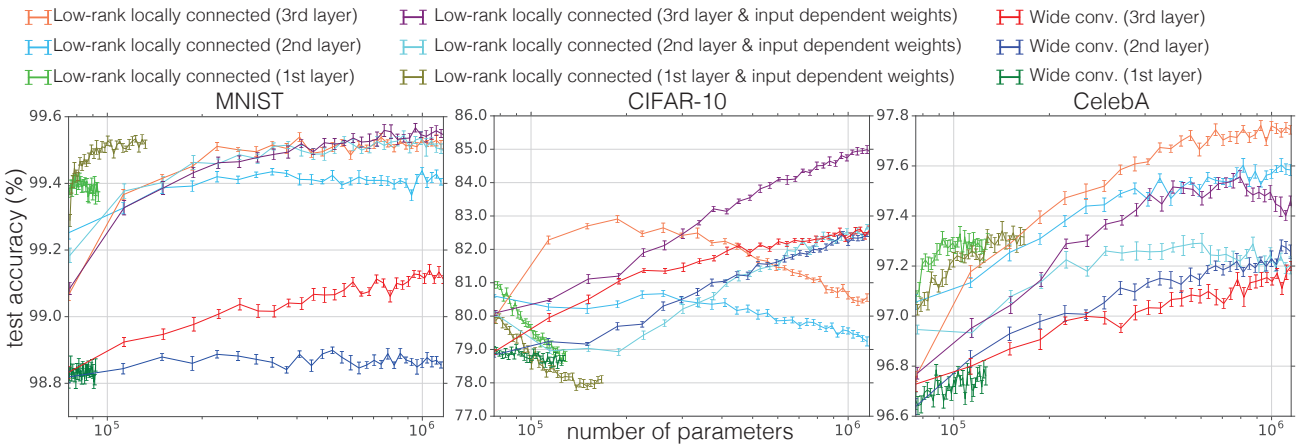

Figure Supp.3: **Accuracy as a function of model parameters.** Classification accuracy as a function of network parameters. Error bars indicate $\pm$ standard errors computed from training models from 10 different random initialization.

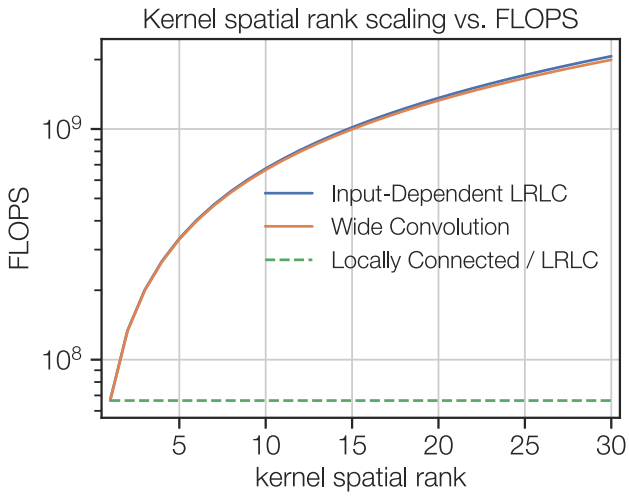

Figure Supp.4: **Computational cost as function of the spatial rank of the locally connected kernel.** As the spatial rank of the locally connected kernel increases, the computational cost, as measured by the number of floating point operations (FLOPS), of the input-dependent LRLC layer and the convolution layer with similar trainable parameter (wide convolution) grows at a similar rate, while the computational cost of the LRLC layer stays constant because it can be converted into a locally connected layer at inference time.

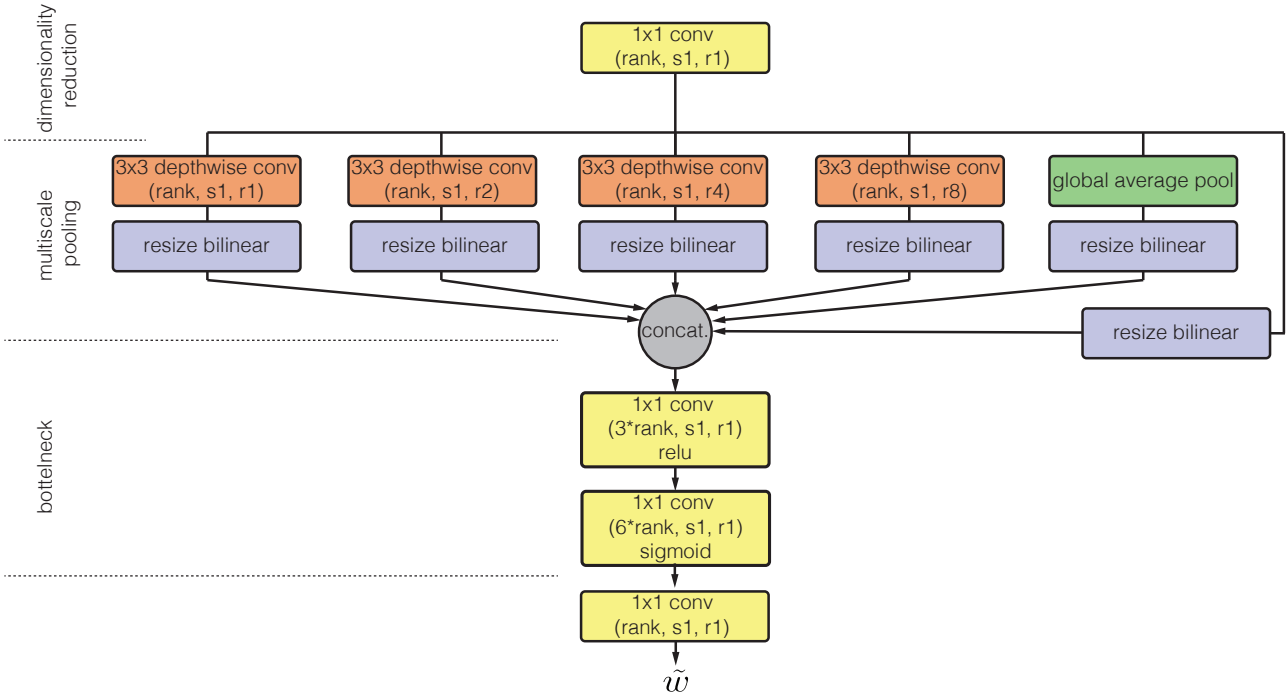

Figure Supp.5: **Input-dependent combining weights network architecture.**

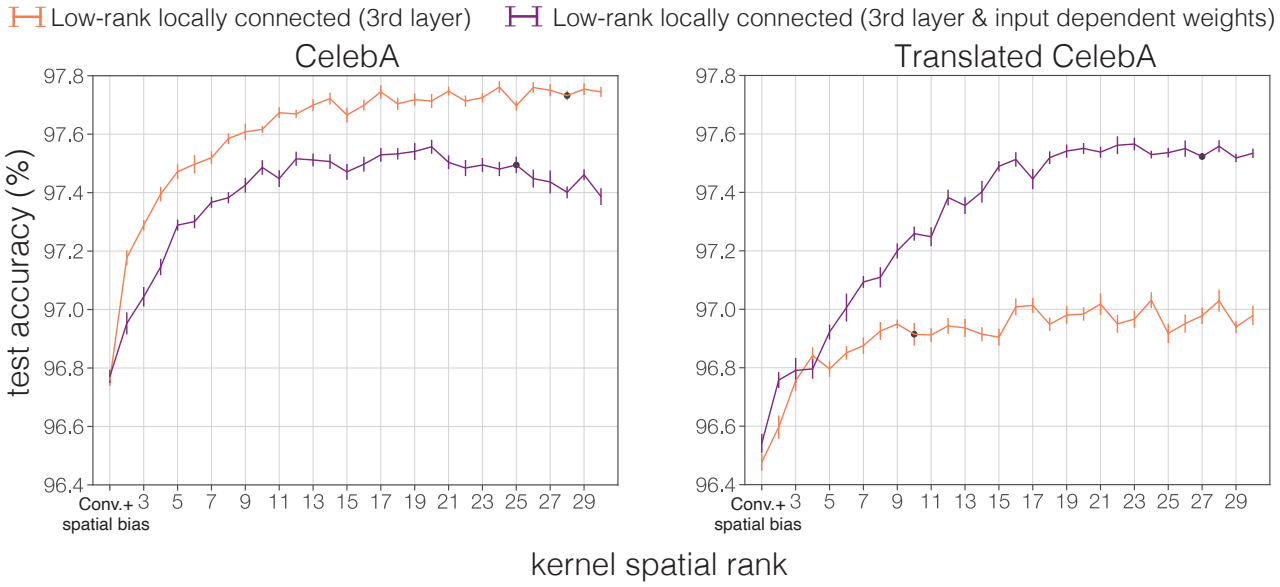

Figure Supp.6: **Input-dependent LRLC is invariant to translation.** Comparing the performance of LRLC and input dependent LRLC networks in CelebA and Translated CelebA datasets.

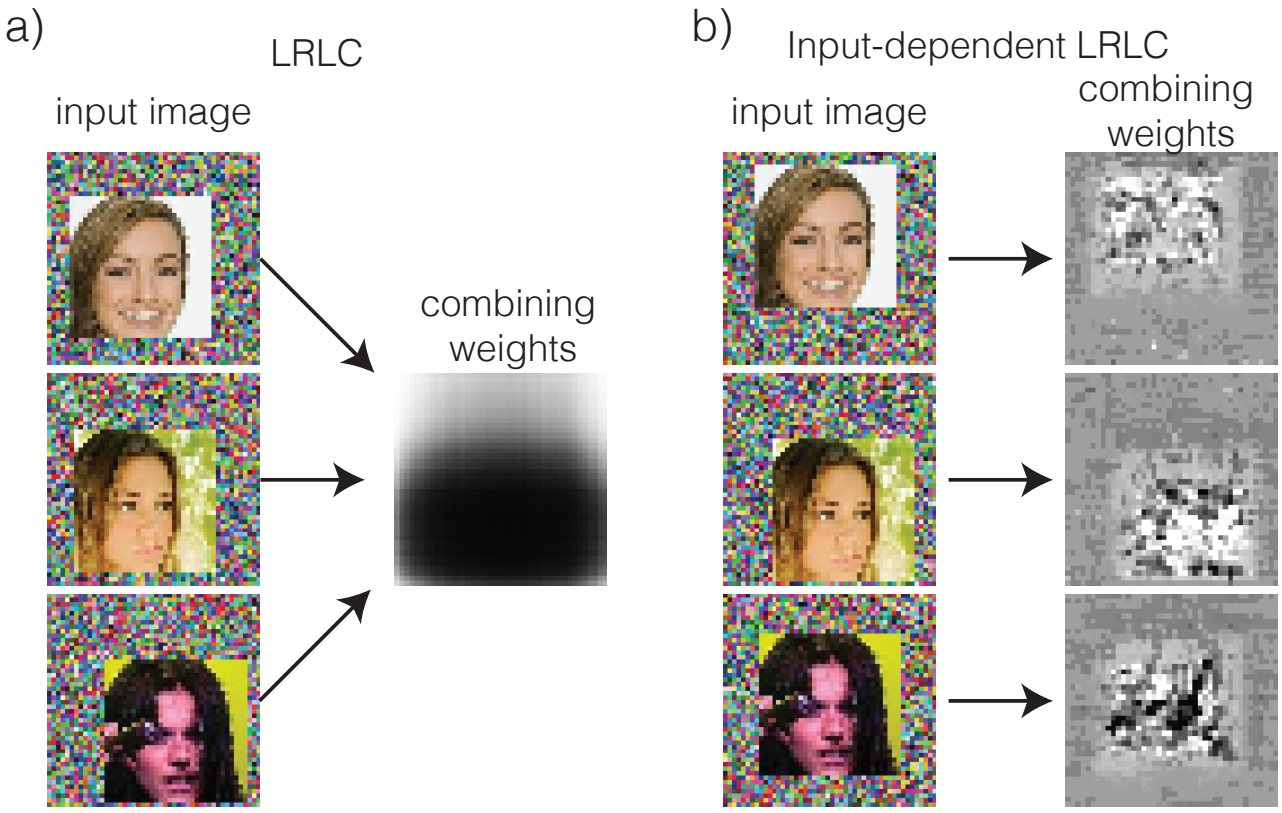

Figure Supp.7: **Visualization of combining weights.** Combining weights for an LRLC network a) and input-dependent LRLC network in b) with rank 2 trained on Translated CelebA dataset.

B. Input-dependent combining weights network

The architecture of the input-dependent combining weights network (g) is illustrated in Figure Supp.5. The initial operation of g is to project the input channels to a low-dimensional space using a 1×1 convolution. This projection is used to allow g to have small number of parameters, and also because selection of filter banks in the basis set is potentially a simpler task than the classification task the network is performing. Motivated by work on segmentation (Chen et al., 2017a; Yu & Koltun, 2015; Chen et al., 2017b), the second operation collects statistics across different scales of the input using parallel pooling and dilated depth-wise 3×3 convolution layers followed by bilinear resizing. Note the increase in parameters here is small due to the initial projection step and the use of depth-wise convolution. The next stage is a nonlinear low-dimensional bottleneck followed by nonlinear dimensionality expansion with 1×1 convolution. This operation has similar flavor to the Squeeze-and-Excitation operation (Hu et al., 2018), and is included to give g the power to learn useful embedding of the input. The last layer is a linear 1×1 convolution that reduce the channels size to the spatial rank.

C. ImageNet training

We divide the standard ImageNet ILSVRC 2012 training set into training and development subsets. We trained our models on the training subset and chose best rank based on the development subset. We follow common practice and report results on the separate ILSVRC 2012 validation set, which we do not use for training or hyperparameter selection.

We trained the network by optimizing the cross entropy loss plus ℓ^2-regularization on the model weights. We optimized all models using Stochastic Gradient Descent with Nesterov momentum of 0.9. We preprocessed images by subtracting the mean and dividing by the standard deviation of training examples. During optimization, we augmented the training data by taking a random crop within the image and then performing bicubic resizing to model's resolution. We used a batch size of 2048 and ℓ^2-regularization scale of $8e-5$. We trained our models for 150 epochs starting with a linear warmup period of 10 epochs and used a cosine decay schedule afterwards. We used Tensor Processing Unit (TPU) accelerators in all our training. We computed our results by computing the top-1 accuracy $\pm$ standatd error based on models trained from 3 different random initializations.

D. Supplementary tables

Table Supp.1: **Number of examples in each dataset.**

SUBSET	MNIST	CIFAR-10	CELEBA
TRAIN	55000	45000	162770
VALIDATION	5000	5000	19867
TEST	10000	10000	19962

Table Supp.2: **Summary of results (train subset).** Top-1 train accuracy of different models (mean $\pm$ SE).

LAYER	MNIST	CIFAR-10	CELEBA
CONVOLUTION	100.00 ± 0.00	97.47 ± 0.07	100.00 ± 0.00
CONVOLUTION (FC)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
CONVOLUTION + SPATIAL BIAS (1ST LAYER)	100.00 ± 0.00	97.99 ± 0.04	100.00 ± 0.00
CONVOLUTION + SPATIAL BIAS (2ND LAYER)	100.00 ± 0.00	97.99 ± 0.06	100.00 ± 0.00
CONVOLUTION + SPATIAL BIAS (3RD LAYER)	100.00 ± 0.00	97.60 ± 0.05	100.00 ± 0.00
CONVOLUTION + SPATIAL BIAS (ALL LAYERS)	100.00 ± 0.00	98.36 ± 0.06	100.00 ± 0.00
COORDCONV	100.00 ± 0.00	97.30 ± 0.08	100.00 ± 0.00
LRLC (1ST LAYER & INPUT DEPENDENT WEIGHTS)	100.00 ± 0.00	99.04 ± 0.05	100.00 ± 0.00
LRLC (1ST LAYER)	100.00 ± 0.00	98.53 ± 0.04	100.00 ± 0.00
LRLC (2ND LAYER & INPUT DEPENDENT WEIGHTS)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LRLC (2ND LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LRLC (3RD LAYER & INPUT DEPENDENT WEIGHTS)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LRLC (3RD LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LRLC (ALL LAYERS)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LOCALLY CONNECTED (1ST LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LOCALLY CONNECTED (2ND LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
LOCALLY CONNECTED (3RD LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
WIDE CONVOLUTION (1ST LAYER)	100.00 ± 0.00	98.25 ± 0.05	100.00 ± 0.00
WIDE CONVOLUTION (2ND LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
WIDE CONVOLUTION (3RD LAYER)	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

CONVOLUTION (FC) is a convolution network with a fully connected last layer and without global average pooling.

Table Supp.3: **Summary of results (validation subset).** Top-1 test accuracy of different models (mean ± SE).

LAYER	MNIST	CIFAR-10	CELEBA
CONVOLUTION	98.89 ± 0.03	79.68 ± 0.13	97.27 ± 0.02
CONVOLUTION (FC)	98.79 ± 0.03	68.36 ± 0.17	97.48 ± 0.02
CONVOLUTION + SPATIAL BIAS (1ST LAYER)	99.28 ± 0.02	81.61 ± 0.14	97.61 ± 0.02
CONVOLUTION + SPATIAL BIAS (2ND LAYER)	99.16 ± 0.02	81.08 ± 0.09	97.56 ± 0.02
CONVOLUTION + SPATIAL BIAS (3RD LAYER)	98.99 ± 0.02	80.83 ± 0.11	97.28 ± 0.05
CONVOLUTION + SPATIAL BIAS (ALL LAYERS)	99.25 ± 0.02	81.29 ± 0.12	97.59 ± 0.02
COORDCONV	99.33 ± 0.02	81.97 ± 0.15	97.83 ± 0.02
LRLC (1ST LAYER & INPUT DEPENDENT WEIGHTS)	99.36 ± 0.02	80.50 ± 0.11	97.82 ± 0.03
LRLC (1ST LAYER)	99.30 ± 0.02	81.76 ± 0.11	97.79 ± 0.02
LRLC (2ND LAYER & INPUT DEPENDENT WEIGHTS)	99.35 ± 0.03	83.55 ± 0.14	97.82 ± 0.02
LRLC (2ND LAYER)	99.31 ± 0.02	81.18 ± 0.10	98.03 ± 0.03
LRLC (3RD LAYER & INPUT DEPENDENT WEIGHTS)	99.39 ± 0.02	85.61 ± 0.06	98.00 ± 0.03
LRLC (3RD LAYER)	99.41 ± 0.01	82.87 ± 0.13	98.16 ± 0.02
LRLC (ALL LAYERS)	99.31 ± 0.02	81.75 ± 0.10	97.98 ± 0.02
LOCALLY CONNECTED (1ST LAYER)	98.77 ± 0.02	63.72 ± 0.20	96.95 ± 0.01
LOCALLY CONNECTED (2ND LAYER)	98.69 ± 0.02	70.24 ± 0.15	97.63 ± 0.02
LOCALLY CONNECTED (3RD LAYER)	99.00 ± 0.01	72.49 ± 0.14	97.79 ± 0.02
WIDE CONVOLUTION (1ST LAYER)	98.91 ± 0.02	79.94 ± 0.14	97.37 ± 0.02
WIDE CONVOLUTION (2ND LAYER)	98.91 ± 0.02	82.98 ± 0.06	97.90 ± 0.02
WIDE CONVOLUTION (3RD LAYER)	99.13 ± 0.02	83.11 ± 0.07	97.72 ± 0.02

CONVOLUTION (FC) is a convolution network with a fully connected last layer and without global average pooling.

Table Supp.4: **Summary of results (test subset).** Top-1 test accuracy of different models (mean ± SE). The optimal rank in LRLC and the optimal width in wide convolution models are obtained by evaluating models on a separate validation subset.

LAYER	MNIST	CIFAR-10	CELEBA
CONVOLUTION	98.84 ± 0.01	78.80 ± 0.12	96.66 ± 0.04
CONVOLUTION (FC)	99.05 ± 0.02	67.59 ± 0.19	96.98 ± 0.02
CONVOLUTION + SPATIAL BIAS (1ST LAYER)	99.37 ± 0.01	80.94 ± 0.10	97.08 ± 0.02
CONVOLUTION + SPATIAL BIAS (2ND LAYER)	99.25 ± 0.02	80.60 ± 0.05	97.06 ± 0.03
CONVOLUTION + SPATIAL BIAS (3RD LAYER)	99.07 ± 0.02	79.94 ± 0.12	96.76 ± 0.02
CONVOLUTION + SPATIAL BIAS (ALL LAYERS)	99.36 ± 0.01	80.89 ± 0.08	97.12 ± 0.03
COORDCONV	99.46 ± 0.01	81.29 ± 0.13	97.40 ± 0.02
LRLC (1ST LAYER & INPUT DEPENDENT WEIGHTS)	99.52 ± 0.02	79.99 ± 0.07	97.32 ± 0.03
LRLC (1ST LAYER)	99.39 ± 0.02	80.94 ± 0.10	97.25 ± 0.02
LRLC (2ND LAYER & INPUT DEPENDENT WEIGHTS)	99.53 ± 0.01	82.62 ± 0.07	97.25 ± 0.02
LRLC (2ND LAYER)	99.42 ± 0.02	80.67 ± 0.09	97.53 ± 0.01
LRLC (3RD LAYER & INPUT DEPENDENT WEIGHTS)	$\mathbf{99.57 \pm 0.01}$	$\mathbf{84.91 \pm 0.07}$	97.49 ± 0.03
LRLC (3RD LAYER)	99.51 ± 0.01	82.70 ± 0.08	$\mathbf{97.73 \pm 0.02}$
LRLC (ALL LAYERS)	99.45 ± 0.01	81.03 ± 0.11	97.51 ± 0.03
LOCALLY CONNECTED (1ST LAYER)	98.74 ± 0.01	62.54 ± 0.14	96.32 ± 0.02
LOCALLY CONNECTED (2ND LAYER)	98.72 ± 0.02	69.29 ± 0.09	97.19 ± 0.02
LOCALLY CONNECTED (3RD LAYER)	99.10 ± 0.01	71.86 ± 0.10	97.31 ± 0.01
WIDE CONVOLUTION (1ST LAYER)	98.84 ± 0.02	78.96 ± 0.11	96.77 ± 0.04
WIDE CONVOLUTION (2ND LAYER)	98.89 ± 0.02	82.36 ± 0.09	97.28 ± 0.02
WIDE CONVOLUTION (3RD LAYER)	99.10 ± 0.02	82.40 ± 0.10	97.20 ± 0.02

CONVOLUTION (FC) is a convolution network with a fully connected last layer and without global average pooling.