Supplementary Materials for On-Off Center-Surround Receptive Fields for Accurate and Robust Image Classification

Zahra Babaiee [1] **Ramin Hasani** [2] **Mathias Lechner** [3] **Daniela Rus** [2] **Radu Grosu** [1]

S1. Theoretical Proofs and Calculations

In this section, we bring the mathematical calculations and theoretical proofs.

S1.1. Proof of Proposition 1

The DoG model used in the main document is defined as in Equation (S1), where γ with $\gamma < 1$, defines the ratio between the radius r of the center and that of the surround. This model allows us to analytically compute the variances, from the size of the receptive fields:

$$DoG_{\sigma,\gamma}(x,y) = \frac{A_c}{\gamma^2}\, e^{-\frac{x^2+y^2}{2\gamma^2\sigma^2}} - A_s\, e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{S1}$$

The coefficients A_c and A_s are determined, by requiring that the sum of all positive values in Equation (S1) are equal to those of the negative values. Here, we make them to sum up to 1 and to -1, respectively:

$$\iint [DoG_{\sigma,\gamma}(x,y)]^+ dxdy = 1, \tag{S2}$$

$$\iint [DoG_{\sigma,\gamma}(x,y)]^- dxdy = -1 \tag{S3}$$

By $[z]^+$ and $[z]^-$ we denote the positive and the negative half wave rectification functions, respectively:

$$[z]^+ = max(0,z), \quad [z]^- = min(0,z) \tag{S4}$$

Proposition 1 (DoG Coefficients)**.** *In the infinite continuous case, the coefficients A_c and A_s are equal.*

Proof. We have the following equalities:

$$\iint_{\mathbb{R}^2} [\frac{A_c}{\gamma^2}\, e^{-\frac{x^2+y^2}{2\gamma^2\sigma^2}} - A_s\, e^{-\frac{x^2+y^2}{2\sigma^2}}]^+ dxdy = 1, \tag{S5}$$

$$\iint_{\mathbb{R}^2} [\frac{A_c}{\gamma^2}\, e^{-\frac{x^2+y^2}{2\gamma^2\sigma^2}} - A_s\, e^{-\frac{x^2+y^2}{2\sigma^2}}]^- dxdy = -1 \tag{S6}$$

By transforming the integrals to the polar coordinates we have the equations below, where r_s is the radius of the surround, and $r_s \to \infty$.

[1]CPS, TU Wien [2]CSAIL, MIT [3]IST Austria. Correspondence to: Zahra Babaiee <zahra.babaiee@tuwien.ac.at>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$$\int_0^{2\pi}\int_0^{r_s}[\frac{A_c}{\gamma^2}\,r\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,r\,e^{-\frac{r^2}{2\sigma^2}}]^+drd\theta=1, \tag{S7}$$

$$\int_0^{2\pi}\int_0^{r_s}[\frac{A_c}{\gamma^2}\,r\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,r\,e^{-\frac{r^2}{2\sigma^2}}]^-drd\theta=-1 \tag{S8}$$

The positive values are in the center with radius of r_c and the negative values are in a ring between the center and surround. So we can remove the half wave rectifiers as follows:

$$2\pi\int_0^{r_c}\frac{A_c}{\gamma^2}\,r\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,r\,e^{-\frac{r^2}{2\sigma^2}}\,dr=1, \tag{S9}$$

$$2\pi\int_{r_c}^{r_s}\frac{A_c}{\gamma^2}\,r\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,r\,e^{-\frac{r^2}{2\sigma^2}}\,dr=-1 \tag{S10}$$

After calculating the integrals:

$$2\pi(A_c\,\sigma^2\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,\sigma^2\,e^{-\frac{r^2}{2\sigma^2}})\Big|_0^{r_c}=1,$$

$$2\pi(A_c\,\sigma^2\,e^{-\frac{r^2}{2\gamma^2\sigma^2}}-A_s\,\sigma^2\,e^{-\frac{r^2}{2\sigma^2}})\Big|_{r_c}^{r_s}=-1$$

$$2\pi\sigma^2(A_c\,e^{-\frac{r_c^2}{2\gamma^2\sigma^2}}-A_s\,e^{-\frac{r_c^2}{2\sigma^2}})-2\pi\sigma^2(A_c\,e^0-A_s\,e^0)=1,$$

$$2\pi\sigma^2(A_c\,e^{-\frac{r_s^2}{2\gamma^2\sigma^2}}-A_s\,e^{-\frac{r_s^2}{2\sigma^2}})-2\pi\sigma^2(A_c\,e^{-\frac{r_c^2}{2\gamma^2\sigma^2}}-A_s\,e^{-\frac{r_c^2}{2\sigma^2}})=-1 \tag{S11}$$

Adding the two equations together, we have:

$$\lim_{r_s\to\infty}2\pi\sigma^2(A_c\,e^{-\frac{r_s^2}{2\gamma^2\sigma^2}}-A_s\,e^{-\frac{r_s^2}{2\sigma^2}})-2\pi\sigma^2(A_c-A_s)=0 \tag{S12}$$

$$=2\pi\sigma^2(A_c\,e^{-\infty}-A_s\,e^{-\infty})-2\pi\sigma^2(A_c-A_s)=0$$

$$2\pi\sigma^2(A_c-A_s)=0\Rightarrow A_c=A_s \tag{S11}$$

□

S1.2. Computation of the Variance

The $DoG_{\sigma,\gamma}(x,y)$ is equal to zero on the border of the center and surround. The radius equals to r_c on this border, meaning that $x^2+y^2=r_c^2$ when $DoG_{\sigma,\gamma}(x,y)=0$. so by setting the $DoG_{\sigma,\gamma}(x,y)=0$ we have:

$$\frac{A_c}{\gamma^2}\,e^{-\frac{r_c^2}{2\gamma^2\sigma^2}}-A_s\,e^{-\frac{r_c^2}{2\sigma^2}}=0$$

$$\ln(A_c)-2\ln(\gamma)-\frac{r_c^2}{2\gamma^2\sigma^2}-\ln(A_s)+\frac{r_c^2}{2\sigma^2}=0$$

$$\ln(A_c)-\ln(A_s)-2\ln(\gamma)=\frac{r_c^2}{2\gamma^2\sigma^2}-\frac{r_c^2}{2\sigma^2}$$

$$\ln(\frac{A_c}{A_s})-2\ln(\gamma)=\frac{r_c^2(1-\gamma^2)}{2\gamma^2\sigma^2}$$

$$\sigma^2=\frac{r_c^2(1-\gamma^2)}{2\gamma^2(\ln(\frac{A_c}{A_s})-2\ln(\gamma))}$$

$$\sigma = \frac{r_c}{\gamma}\sqrt{\frac{1-\gamma^2}{2\ln(\frac{A_c}{A_s}) - 4\ln(\gamma)}} \tag{S12}$$

Based on Proposition 1, the values of A_c and A_s are equal in the infinite continuous case. Since in the finite discrete case those values are very close, we can approximate the value of σ:

$$\sigma \approx \frac{r_c}{2\gamma}\sqrt{\frac{1-\gamma^2}{-\ln\gamma}} \tag{S13}$$

S1.3. Proof of Theorem 1

We use Equation (S1) to compute the weights in the On-center kernel matrix DoG_{On}. For the Off-center kernel DoG_{Off}, we use the same equation with the signs inverted. For a given input χ, we calculate the On and Off responses by convolving χ with the computed fixed kernels separately:

$$\chi_{\mathrm{On}}[x,y] = (\chi * DoG_{\mathrm{On}})[x,y], \tag{S14}$$

$$\chi_{\mathrm{Off}}[x,y] = (\chi * DoG_{\mathrm{Off}})[x,y] \tag{S15}$$

Note that the On and Off convolutions cover the input image completely. These two convolutions result in the following equations, when the kernel is in the shape of a square:

$$\chi_{\mathrm{On}}[x,y] = \int_{-r_s}^{r_s}\int_{-r_s}^{r_s} \chi(x+\rho, y+\tau)\,(\frac{A_c}{\gamma^2}\,e^{-\frac{\rho^2+\tau^2}{2\gamma^2\sigma^2}} - A_s\,e^{-\frac{\rho^2+\tau^2}{2\sigma^2}})\,d\rho\,d\tau \tag{S16}$$

$$\chi_{\mathrm{Off}}[x,y] = \int_{-r_s}^{r_s}\int_{-r_s}^{r_s} \chi(x+\rho, y+\tau)\,(A_s\,e^{-\frac{\rho^2+\tau^2}{2\sigma^2}} - \frac{A_c}{\gamma^2}\,e^{-\frac{\rho^2+\tau^2}{2\gamma^2\sigma^2}})\,d\rho\,d\tau \tag{S17}$$

Theorem 1 (On-Off Complementarity)**.** *The on- and off-pathways learn unique and complementary features.*

Proof. We prove this theorem by contradiction. Assume that the features extracted by the On convolution are identical to the features extracted by the Off convolution. Now suppose the input image has a small spot of light (smaller than the center of our kernels) on a dark background. We first convolve this image with an On kernel. If the spot of light lies in the center of the kernel, the convolution will result in a response close to 1, according to the Equations (S5), (S6), and (S16). If the spot of light lies in the surround, the convolution will result in a negative response. As a result we obtain an activation map with values close to 1 where the light spot is located, negative values in the outer edges of the light spot, and zero values everywhere else.

When we convolve the same image with an Off kernel we obtain the following. If the spot of light lies in the center of the kernel, then we obtain a negative response. If it lies in the surround, then it will result in a positive response close to zero, according to the Equations (S5), (S6), and (S17). Hence, the Off convolution results in an activation map with small values in the outer edges of the light spot, negative values where the light spot is located, and zero values everywhere else. Comparing the two activation maps, one can see that: 1) The positive values are in different locations, and 2) These values are close to 1 for the On activation map, and close to zero for the Off activation map. This contradicts our initial assumption. Note that a similar but complementary argument can be made for an image with a small dark spot on a light background. □

S2. Experimental Setup

Here, we describe the experimental setup for the tasks discussed in Tables 1, 2, 3, 4 and 5.

S2.1. Dataset description

Image classification. We used a random subset of the Imagenet dataset (Deng et al., 2009) with 60000 images from 100 categories. We used 5000 of the images for each of the validation and test sets. All samples were cropped around the center if they were not originally in square shapes, and then resized to 192×192 pixels.

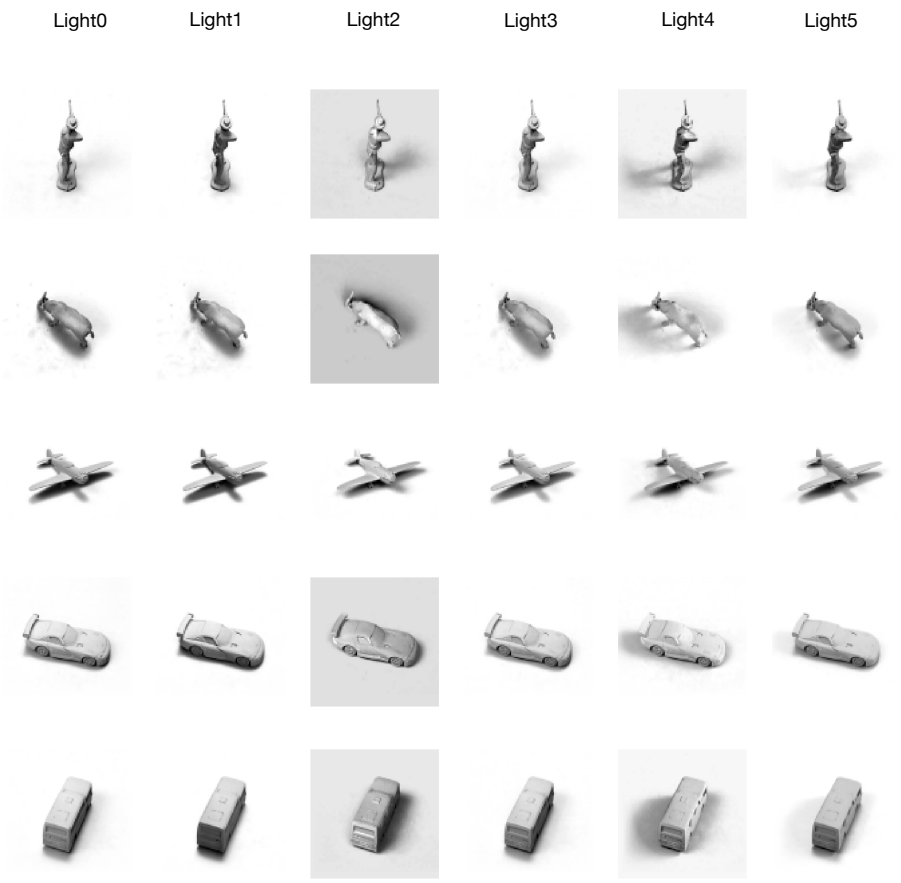

Figure S1. Example samples of the Norb dataset (LeCun et al., 2004) from each of the five categories and each of the lighting conditions.

Robustness to illumination change. We used small Norb dataset (LeCun et al., 2004) which contains images of toys from 5 generic categories: human figures, four-legged animals, airplanes, cars and trucks. The images of each category were taken from 10 toy instances in 6 different lighting conditions, 9 elevations, and 18 azimuths. The training set consists of the images from 5 of the instances of each category, and the rest 5 instances are in the test set. Figure S1 shows one sample from each category in each of the lighting conditions. We separated the dataset based on the lighting conditions, and used the images from the first light, Light0, as the training set and tested our networks on testsets from all 6 different lighting conditions. Each of the training and test sets contained 4050 images of size 96×96 pixels.

Robustness to distribution shifts. We used MNIST dataset (LeCun & Cortes, 2010) containing grayscaled images of handwritten digits. There are 60000 and 10000 samples in the training and test sets respectively. For testing, we inverted all the pixel values by subtracting them form 255.

S2.2. Network architectures and Hyper parameters

For each of the experiments, we used a CNN as the base network, with different numbers of layers depending the dataset image sizes. Figure S2 shows the architectures of the base networks. We construct the other models from the base networks as discussed in the main paper. For the On and Off Center convolutions in OOCS-CNNs, we used kernels of size 5×5 for Imagenet and Norb datasets. We used smaller kernels of size 3×3 for the MNIST dataset, since the images are of smaller size. We calculated the On and Off resposes from the inputs and directly fed their summation to the network.

We had batch sizes of 64 in all experiments. We used Adam optimiser (Diederik & Ba, 2015) for experiments on Imagenet and Norb, with a learning rate of 10^{-4}. In the experiment on Imagenet, we decreased the learning rate to half after 10 epochs which was mainly in favour of the baselines. In the Imagenet experiments with ResNet-34 we use SGD optimiser

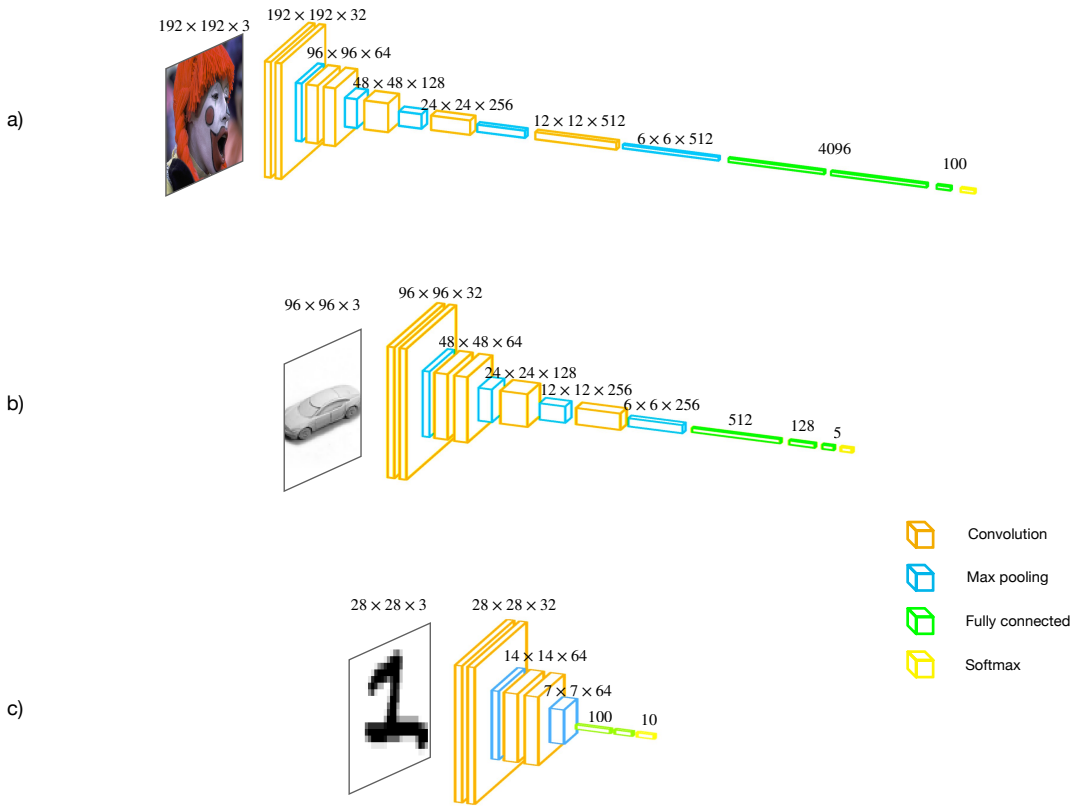

Figure S2. Base network architectures for a) Imagenet subset classification, b) robustness evaluation on Norb, and c) robustness evaluation on MNIST

Table S1. Test Accuracy and variance for test images with Gaussian noise. n=6.

	Gaussian Noise (σ)				
Models	0.01	0.02	0.03	0.04	0.05
ResNet-34	$61.2_{\pm 0.7}$	$58.5_{\pm 0.8}$	$53.14_{\pm 0.8}$	$46.4_{\pm 0.5}$	$40.23_{\pm 0.8}$
OOCS-ResNet-34	$\mathbf{62.7}_{\pm \mathbf{0.7}}$	$\mathbf{60.4}_{\pm \mathbf{0.9}}$	$\mathbf{56.0}_{\pm \mathbf{1.3}}$	$\mathbf{50.5}_{\pm \mathbf{2.1}}$	$\mathbf{44.8}_{\pm \mathbf{2.5}}$

and start with a learning rate of 0.1, which we decay by a factor of 0.1 every 20 epochs and we trained the networks for 60 epochs. For scaling the gradient descent steps, we use a Nesterov-momentum of 0.9.

S3. Experiments on Digital Distribution Shifts

In this section we describe the experiments to evaluate the robustness of a ResNet-34 on the Imagnet subset compared to the same network equipped with OOCS.

We altered the test set images with 5 different digital perturbations: adding Gaussian noise, decreasing and increasing the brightness (Gamma correction), and decreasing and increasing the contrast. Figure S3 shows one sample of Imagenet dataset in different brightness and contrast changes with different severities.

The results of this experiment are summarized in tables S1-5 and figure S4. As the results show, OOCS can enhance the robustness of a ResNet under digital distribution shifts.

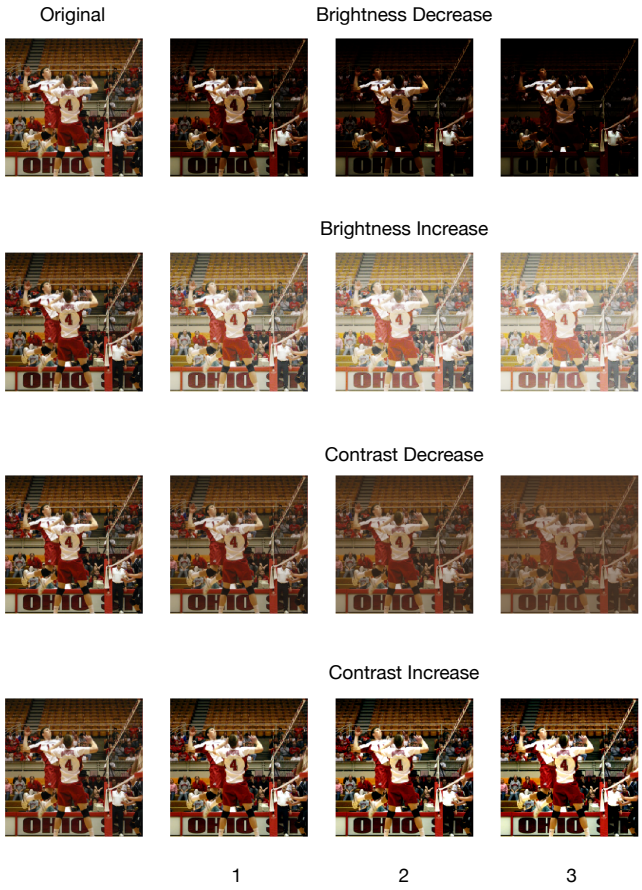

Figure S3. Sample Brightness and contrast variations we tested OOCS and Residual networks against.

S4. Code and Data Availability

All code and data are included in https://github.com/ranaa-b/OOCS.

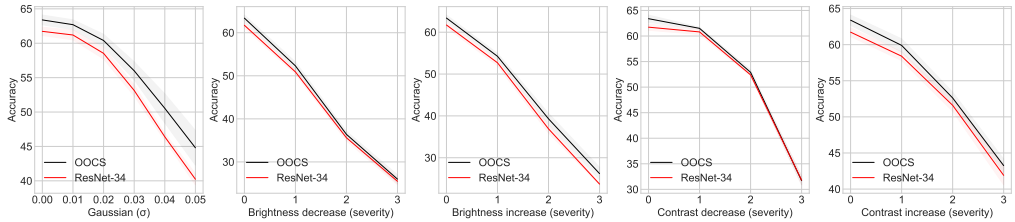

Figure S4. OOCS filters added to ResNet-34 consistently enhances the robustness of a network to perturbations such as Gaussian noise, Brightness and Contrast variations.

Table S2. Test Accuracy and variance for test images with decreasing brightness. n=6.

	Gamma Correction (γ)		
Models	2	3	4
ResNet-34	$50.9_{\pm 0.5}$	$35.6_{\pm 0.7}$	$25.5_{\pm 0.8}$
OOCS-ResNet-34	$\mathbf{52.3_{\pm 0.9}}$	$\mathbf{36.4_{\pm 0.7}}$	$\mathbf{26.0_{\pm 1.0}}$

Table S3. Test Accuracy and variance for test images with increasing brightness. n=6.

	Gamma Correction (γ)		
Models	1/2	1/3	1/4
ResNet-34	$52.7_{\pm 0.7}$	$36.9_{\pm 1.0}$	$23.7_{\pm 0.2}$
OOCS-ResNet-34	$\mathbf{54.2_{\pm 0.7}}$	$\mathbf{39.3_{\pm 0.7}}$	$\mathbf{26.2_{\pm 1.5}}$

Table S4. Test Accuracy and variance for test images with decreasing contrast. n=6.

	Contrast Factor		
Models	0.8	0.6	0.4
ResNet-34	$60.8_{\pm 0.5}$	$52.4_{\pm 0.2}$	$\mathbf{31.8_{\pm 0.9}}$
OOCS-ResNet-34	$\mathbf{61.5_{\pm 0.2}}$	$\mathbf{52.9_{\pm 0.5}}$	$31.7_{\pm 0.8}$

Table S5. Test Accuracy and variance for test images with increasing contrast. n=6.

	Contrast Factor		
Models	1.2	1.4	1.6
ResNet-34	$58.4_{\pm 0.6}$	$51.6_{\pm 0.8}$	$41.9_{\pm 1.3}$
OOCS-ResNet-34	$\mathbf{59.9_{\pm 0.9}}$	$\mathbf{52.62_{\pm 0.8}}$	$\mathbf{43.25_{\pm 0.9}}$

References

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, Florida, USA, June 2009. IEEE Computer Society.

Diederik, P. and Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.

LeCun, Y., Huang, F., and Bottou, L. Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 97–104, Washington DC, USA, July 2004. IEEE.