Supplementary Material
H-NeXt: The next step towards roto-translation invariant networks

Tomáš Karella
karella@utia.cas.cz
Filip Šroubek
sroubekf@utia.cas.cz
Jan Flusser
flusser@utia.cas.cz
Jan Blažek
blazek@utia.cas.cz
Vašek Košík
kosik@utia.cas.cz

Institute of Information Theory and
Automation
Czech Academy of Sciences
Pod Vodárenskou věží 4
Prague 8, Czech Republic

A Invariance Measurement

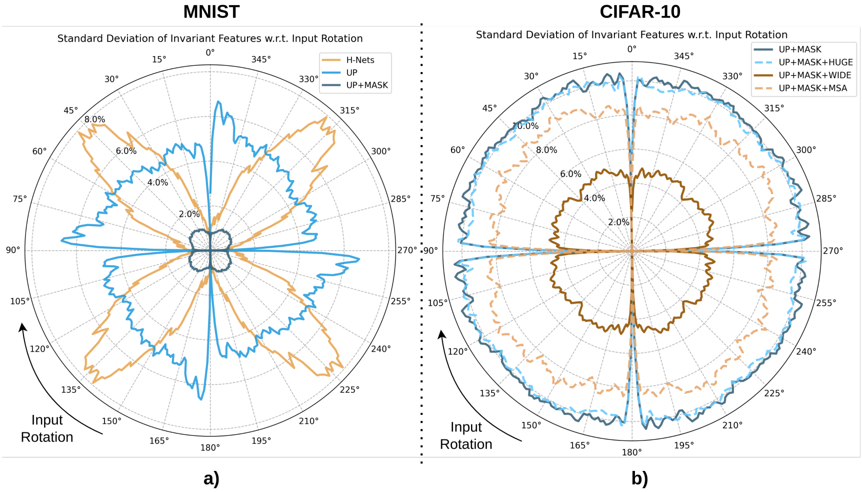


Figure 1: Standard deviation (radius) of the differences for inputs rotated by 0° and the given angles a) for the MNIST dataset and b) for the CIFAR-10 dataset.

As a measure of rotation robustness, we compute the standard deviation (STD) of the differences between the network outputs for images rotated by the given angle and images without rotation. The results are shown in Figure 1, where the angular axis represents the input rotation and the radius represents the STD.

As can be seen, the STD repeats periodically at 90° intervals with a minimum at the boundaries where there are no rotation artifacts. In the case of MNIST, since every digit of the training dataset is in an upright position, we can observe the angular effect on STD, the maximum being at an angle of 45°, where the effect of interpolation is the greatest. The model **UP** with large filters is strongly affected by the boundary effect, which exceeds the effect of interpolation. In the case of CIFAR, where object orientations vary across the dataset, therefore STD is uniformly affected except at 90° angles.

B Experiment Details

B.1 Setup

The models were trained for up to 100 epochs using the AdamW optimizer [4], batch size 64, and cross-entropy loss with an initial learning rate of 0.007, decreased whenever a validation loss plateau was reached. The regularization techniques used are L2 normalization, gradient clipping, label smoothing, and dropout (in the Classifier Network). Most importantly, no augmentation is applied to the training set. Each experiment was repeated 10 times and we report the mean and standard deviation.

B.2 SWN-GCN Comparison

To show the state-of-the-art results we follow the SWN-GCN evaluation setup [2], which consists of multiple test datasets rotated by fixed angles. The average accuracy and standard deviation are listed in Table 1 for MNIST and in Table 2 for CIFAR.

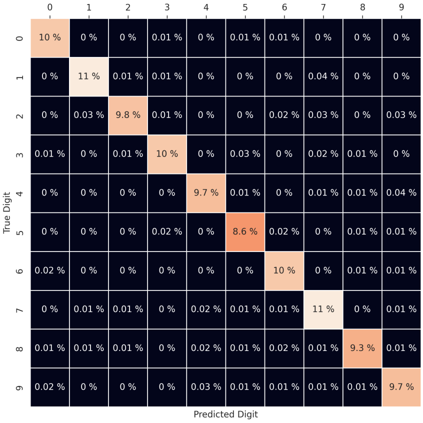
Table 1: Performance comparison on the MNIST invariance benchmark following the SWN-GCN [2] setup. OA is the overall accuracy of all fixed angles and σ stands for standard deviation.

MNIST Models	OA	0°	30°	60°	90°	120°	150°	180°	210°	240°	270°	300°	330°
E(2)-CNN [5]	87.50	99.30	98.10	95.90	96.30	86.20	74.90	70.70	71.10	81.80	95.10	92.90	97.00
TIGRANET [3]	85.10	89.10	82.70	79.80	89.10	82.70	79.80	89.10	82.70	79.80	89.10	82.70	79.80
SWN-GCN [2]	91.80	96.50	89.80	87.30	96.50	89.80	87.30	96.50	89.80	87.30	96.50	89.80	87.30
H-Nets [6]	92.89	98.70	89.41	90.55	98.70	89.41	90.55	98.70	89.41	90.55	98.70	89.40	90.55
UP+MASK	**98.68**	**98.94**	**98.55**	**98.55**	**98.94**	**98.55**	**98.55**	**98.94**	**98.55**	**98.55**	**98.94**	**98.55**	**98.54**
UP+MASK-σ	±.11	±.08	±.11	±.16	±.09	±.11	±.16	±.08	±.11	±.16	±.09	±.11	±.16

Table 2: Performance comparison on the CIFAR-10 invariance benchmark following the SWN-GCN [2] setup. OA is the overall accuracy of all fixed angles and σ stands for standard deviation.

CIFAR-10 Models	**OA**	0°	30°	60°	90°	120°	150°	180°	210°	240°	270°	300°	330°
RESNET-50 [2]	36.10	85.10	54.50	34.10	18.30	27.50	26.90	35.60	27.00	24.90	33.80	33.20	52.50
E(2)-CNN [5]	46.20	77.10	57.80	44.30	48.50	34.40	30.80	37.80	31.90	35.40	49.40	45.00	56.00
TIGRANET [3]	38.10	38.90	37.00	36.80	38.90	37.00	36.80	38.90	37.00	36.80	38.90	37.00	36.80
SWN-GCN [6]	50.50	51.30	49.60	50.10	51.30	49.60	50.10	51.30	49.60	50.10	51.30	49.60	50.10
UP+MASK	57.40	59.67	56.26	56.27	59.68	56.26	56.27	59.67	56.26	56.27	59.68	56.26	56.27
UP+MASK-σ	0.89	0.99	0.97	0.81	0.99	0.98	±.82	±1.0	±.98	±.82	±.99	±.97	±.82
UP+MASK+WIDE	61.16	62.80	**60.31**	**60.35**	62.80	**60.31**	**60.35**	62.80	**60.30**	**60.36**	62.81	**60.31**	**60.36**
UP+MASK+WIDE-σ	±.89	±.90	±.96	±.93	±.91	±.97	±.92	±.90	±.96	±.93	±.91	±.96	±.92
UP+MASK+MSA	**61.46**	**64.15**	60.09	60.13	**64.15**	60.08	60.13	**64.15**	60.09	60.13	**64.15**	60.09	60.13
UP+MASK+MSA-σ	±.60	±.59	±.78	±.61	±.59	±.77	±.62	±.59	±.77	±.61	±.58	±.77	±.62

B.3 Misclassification of digits 6 and 9



	0	1	2	3	4	5	6	7	8	9
0	10 %	0 %	0 %	0.01 %	0 %	0.01 %	0.01 %	0 %	0 %	0 %
1	0 %	11 %	0.01 %	0.01 %	0 %	0 %	0 %	0.04 %	0 %	0 %
2	0 %	0.03 %	9.8 %	0.01 %	0 %	0 %	0.02 %	0.03 %	0 %	0.03 %
3	0.01 %	0 %	0.01 %	10 %	0 %	0.03 %	0 %	0.02 %	0.01 %	0 %
4	0 %	0 %	0 %	0 %	9.7 %	0.01 %	0 %	0.01 %	0.01 %	0.04 %
5	0 %	0 %	0 %	0.02 %	0 %	8.6 %	0.02 %	0 %	0.01 %	0.01 %
6	0.02 %	0 %	0 %	0 %	0 %	0 %	10 %	0 %	0.01 %	0.01 %
7	0 %	0.01 %	0.01 %	0 %	0.02 %	0.01 %	0.01 %	11 %	0 %	0.01 %
8	0.01 %	0.01 %	0.01 %	0 %	0.02 %	0.01 %	0.02 %	0.01 %	9.3 %	0.01 %
9	0.02 %	0 %	0 %	0 %	0.03 %	0.01 %	0.01 %	0.01 %	0.01 %	9.7 %

Figure 2: Confusion matrix of the model **UP+MASK** on mnist-rot-test.

In TigraNet [3], the authors removed the digit 9 from MNIST, since its rotated version resembles digit 6, but SWN-GCN [2] doesn't explicitly mention the removal of digit 9. According to our experiments handwritten digits 6 and 9 are distinguishable from each other as illustrated in Figure 2.

C Roto-Translation Equivariance Proofs

Note that the H-Convs equivariance proof was first formulated by Worrall *et al.* [6], for completeness we have included our version. The goal is to show that stacks of H-Convs layers change the output predictably under roto-translation, i.e., the rotation orders of the streams and convolutional filters add up.

Definition 1 (Cross-Correlation)**.** *The cross-correlation of a convolution filter W with an input image or a feature map F is defined as follows*

$$[W \star F](\mathbf{x}) = \int_{\mathbb{R}^2} W(\mathbf{z}) F(\mathbf{x}-\mathbf{z}) d\mathbf{z}. \tag{1}$$

Lemma 1 (Rotation of a Harmonic Filter)**.** *The harmonic filter changes by $e^{im\theta}$ when the coordinates are rotated, where m is the rotation order of the harmonic filter and $\mathcal{R}$ is a 2D rotation matrix of angle θ.*

Proof.

$$\begin{aligned} W_m(\mathcal{R}^{-1}\mathbf{x}) \equiv \tilde{W}_m(r, \varphi - \theta) &= R(r) \cdot e^{-im(\varphi-\theta)} \\ &= e^{im\theta} \tilde{W}_m(r,\varphi) \equiv e^{im\theta} W_m(\mathbf{x}) \end{aligned} \tag{2}$$

□

Denote an input image that is rotated by the angle θ and translated by arbitrary t as

$$F'(\mathbf{x}) \triangleq F(\mathcal{R}\mathbf{x} + \mathbf{t}). \tag{3}$$

Theorem 2 (Rotation Order Additivity)**.** *When an input is rotated by θ and translated by* $\mathbf{t}$*, the output of the successive Harmonic Convolutions results in*

$$\left[W_{m_1} \star W_{m_2} \star \cdots \star F'\right](x) = e^{i(m_1+m_2+\cdots)\theta} \left[W_{m_1} \star W_{m_2} \star \cdots \star F\right](\mathcal{R}\mathbf{x}+\mathbf{t}) \tag{4}$$

Proof.

$$\begin{aligned} \left[W_{m_1} \star F'\right](\mathbf{x}) &= \int_{\mathbb{R}^2} W_{m_1}(\mathbf{z}) F'(\mathbf{x}-\mathbf{z}) d\mathbf{z} \\ &= \int_{\mathbb{R}^2} W_{m_1}(\mathbf{z}) F(\mathcal{R}\mathbf{x} - \mathcal{R}\mathbf{z} + \mathbf{t}) d\mathbf{z} \\ &= \int_{\mathbb{R}^2} W_{m_1}(\mathcal{R}^{-1}\mathbf{z}') F(\mathcal{R}\mathbf{x} - \mathbf{z}' + \mathbf{t}) d\mathbf{z}' \qquad \mathbf{z}' = \mathcal{R}\mathbf{z} \\ &= e^{im_1\theta} \int_{\mathbb{R}^2} W_{m_1}(\mathbf{z}') F((\mathcal{R}\mathbf{x}+\mathbf{t}) - \mathbf{z}') d\mathbf{z}' \\ &= e^{im_1\theta} \left[W_{m_1} \star F\right](\mathcal{R}\mathbf{x}+\mathbf{t}) \end{aligned} \tag{5}$$

Denote the output of the first convolution filter as

$$F_1'(\mathbf{x}) \equiv e^{im_1\theta} F_1(\mathcal{R}\mathbf{x}+\mathbf{t}) \equiv e^{im_1\theta} \left[W_{m_1} \star F\right](\mathcal{R}\mathbf{x}+\mathbf{t}). \tag{6}$$

The induction step:

$$\begin{aligned}
\left[W_{m_2} \star F_1'\right](\mathbf{x}) &= \int_{\mathbb{R}^2} W_{m_2}(\mathbf{z}) F_1'(\mathbf{x}-\mathbf{z}) d\mathbf{z} \\
&= e^{im_1\theta} \int_{\mathbb{R}^2} W_{m_2}(\mathbf{z}) F_1(\mathcal{R}\mathbf{x}-\mathcal{R}\mathbf{z}+\mathbf{t}) d\mathbf{z} \\
&= e^{im_1\theta} \cdot e^{im_2\theta} \int_{\mathbb{R}^2} W_{m_2}(\mathbf{z}') F_1((\mathcal{R}\mathbf{x}+\mathbf{t})-\mathbf{z}') d\mathbf{z}' \qquad \mathbf{z}' = \mathcal{R}\mathbf{z} \\
&= e^{i(m_1+m_2)\theta} \left[W_{m_2} \star F_1\right](\mathcal{R}\mathbf{x}+\mathbf{t})
\end{aligned} \tag{7}$$

□

D Computational Complexity

The main difference with CNN is the computational cost of cross-correlation, analogous to H-Nets [6]. The classical cross-correlation consists of $M = h \cdot w \cdot k^2 \cdot i \cdot o$ multiplications, where h, w is height/width of input, k is filter size, i, o is a number of input/output channels.

In the complex domain we need 4 multiplications. For H-Next with 2× upscale, input rotation order r_i and output rotation r_i multiplication results in $M_h = 4 \cdot 2h \cdot 2w \cdot k^2 \cdot i_h \cdot o_h \cdot r_i \cdot r_o$. By setting $i = o$, $i_h = o_h$ and $r_i = r_o$, the transformation rule is $i = 4 r_i i_h$. For example, H-NeXt with similar computational cost to a regular CNN with 64 channels per layer, with rotation order $r_i \in \{0, 1\}$ then the number of H-NeXt channels is $i_h = \frac{64}{4 \cdot 2} = 8$. Note that classical CNNs do not have sufficient accuracy in roto-translational invariance experiments, regardless of the number of parameters.

Table 3: Flop count estimates for individual models on Nvidia V100 using fvcore [1].

CIFAR	**UP+MASK**	**MSA**	**WIDE**	**HUGE**
GFLOPS	8.86	6.36	11.84	137.45
MNIST	**H-Net**	**UP**	**UP+MASK**	
GFLOPS	0.11	3.96	8.77	

References

[1] Facebookresearch. Facebookresearch/fvcore: Collection of common code that's shared among different research projects in fair computer vision team. URL https://github.com/facebookresearch/fvcore.

[2] Sungwon Hwang, Hyungtae Lim, and Hyun Myung. Equivariance-bridged SO (2)-Invariant representation learning using graph convolutional network. In *The 32nd British Machine Vision Conference (BMVC 2021)*. The British Machine Vision Association, 2021. doi: 10.48550/arXiv.2106.09996.

[3] Renata Khasanova and Pascal Frossard. Graph-based isometry invariant representation learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1847–1856. PMLR, 2017. doi: 10.48550/arXiv.1703.00356.

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1711.05101.

[5] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. *Advances in Neural Information Processing Systems*, 32, 2019. doi: 10.48550/arXiv.1911.08251.

[6] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5028–5037, 2017. doi: 10.48550/arXiv.1612.04642.