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
Václav Košík
kosik@utia.cas.cz

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

Abstract

The widespread popularity of equivariant networks underscores the significance of parameter efficient models and effective use of training data. At a time when robustness to unseen deformations is becoming increasingly important, we present H-NeXt, which bridges the gap between equivariance and invariance. H-NeXt is a parameter-efficient roto-translation invariant network that is trained without a single augmented image in the training set. Our network comprises three components: an equivariant backbone for learning roto-translation independent features, an invariant pooling layer for discarding roto-translation information, and a classification layer. H-NeXt outperforms the state of the art in classification on unaugmented training sets and augmented test sets of MNIST and CIFAR-10.

1 Introduction

Convolutional Neural Networks (CNNs) [30] are still one the most influential concepts in Computer Vision, and they are being actively researched [32, 46]. One of the key advantages that made CNNs dominant over fully connected networks is the weight sharing across the spatial dimension, i.e. the result of a translated input is an equally translated feature map. This property, called translation equivariance, makes CNNs implicitly robust and efficient with respect to translation.

This does not apply to other deformations, such as rotation, reflection, or affine transformation. Extensive data augmentation is a widely accepted practice to deal with these deformations [28]. As shown by Zeiler *et al*. [49], CNNs filters trained on augmented data consist of rotated, scaled, and translated copies of one another. A question arises whether we can use parameters more effectively, as in the case of translation. Moreover, if we wanted to

create a network robust to numerous deformations using data augmentation, the training set would grow exponentially and training then becomes prohibitively expensive.

Group Equivariant Convolutional Neural Networks (G-CNN) formulated by Cohen and Weiler [9] defined the notion of equivariance, which provides the basis for the effective use of weights. G-CNN led to the development of various methods that exploit symmetries by sharing weights, such as networks based on Steerable filters [2, 22, 27, 34, 45] or Scattering transform [4, 36, 38, 39, 40]. Primarily for rotation, several applications have emerged in domains where objects naturally occur with different orientations, such as aerial photography [5, 6, 15, 21], microscopy imaging [7, 8], or texture classification [33].

All of the aforementioned papers are based on benchmarks with small but augmented training sets. The achieved accuracy shows how effective the model is at using its weights on limited training data, but does not reflect the robustness to deformations that the network has not seen. The next natural step is to show model performance when it is trained only on the dataset without any augmentation. We call models trained without ever seeing an augmented image hard (mathematically) invariant networks, and they were first formulated by Khasonova and Frossard [25] and further improved by Hwang *et al*. [23].

This paper presents a roto-translation invariant model (H-NeXt). ① Within the ablation study, we outline the most common problems encountered when working with unaugmented datasets and show solutions adopted by our model. ② H-NeXt surpasses the state of the art with an order of magnitude smaller number of parameters compared to its predecessor. ③ Finally, our model is used to contextualize the invariant and equivariant roles within the rot-mnist dataset [29]. Source codes and datasets are available at https://github.com/karellat/h-next.

2 Related Work

Interest in transformation robust models has been growing in recent years, and these models were comprehensively summarized in a 2021 survey [35]. This article will focus on robustness to rotation and translation, as these deformations arise in a variety of natural tasks [5, 6, 7, 8, 15, 21, 33]. Articles addressing this topic can be neatly divided into three groups: soft (empirically) invariant, equivariant, and hard (mathematically) invariant.

Network (function) $\phi : X \to Y$ is equivariant with respect to the group G_X iff

$$\forall T \in G_X, \exists T' \in G_\phi \qquad \phi(T(x)) = T'(\phi(x)). \tag{1}$$

That is, transforming the input $x \in X$ by T and then processing it by network ϕ is the same as first passing the same input x through network ϕ and transforming the output by T'. Usually the equality of the groups G_X and G_ϕ is considered in most cases, but T and T' can be different, $T = T'$ implying that the transformations commute with ϕ. In the case of invariant networks, G_ϕ is an identity. [1]

Soft invariance/equivariance refers to networks that do not satisfy the mathematical definition of (1), but are forced to do so by common regularization techniques. This category includes the article by Lenc and Vivaldi [31], who minimize the distance between features of transformed and untransformed inputs in the loss function, Spatial Transformers [24],

[1] For clarity, we will consider the transformation group to be the 360°roto-translation group $G_X = SE(2)$ and $G_X = G_\phi$ for equivariant networks.

which estimate the deformation parameters of the input from input data and then normalize the input, or PDO-eCONvs [32] approximately equivariant convolution based on partial differential equations and many others [14, 20].

Equivariant networks have become a broad topic since the publication of G-CNNs [9], which are equivariant to 90 degree rotations. Using steerable filters [19] led to other discrete groups [11, 22, 27, 45], 360° rotations [47], sphere surfaces [10, 12, 16], or 3D volumetric space [13]. E(2)-CNN [44] then summarized the whole theory about steerable networks.

Khasanova and Frossard [25] were the first to formalize networks with hard invariance properties by introducing the roto-translation invariant TigraNet, which consists of spectral convolutions and dynamic pooling. They propose a new type of datasets focusing on invariance properties, i.e. the training sets do not contain any augmented images, but the performance is measured on augmented test sets. Their experiments showed that neither equivariant nor soft invariant networks can produce roto-translation robust features in the unaugmented setting. In particular, they compared TigraNet with the classical CNNs [8], Harmonic Networks (H-Nets) [47], Spatial Transformers [24] and Deep Scattering Networks [36]. Their work was followed by the SWN-GCN architecture [23] that uses an equivariant backbone corresponding to graph convolutional networks followed by Global Average Pooling (GAP) to obtain invariant descriptors. SWN-GCN outperformed TigraNet and became the state of the art on both MNIST and CIFAR-10. For both benchmarks, test sets were augmented by rotation, while only upright images were included in the training set.

3 Proposed Method

Our architecture is based on a general concept that includes H-NeXt, but also other models [23, 25, 41, 42, 47]. Suppose that objects within a single class can be described by a set of transformations decomposable into two subsets D_p and D_n. While D_p contains mathematically modelable transformations, D_n contains all the others. For example consider the class of kittens, where D_p could be the roto-translation group, and different fur, paw size, etc. would be transformations of D_n. The architecture, as illustrated in Figure 1, consists of three parts: Equivariant Backbone, Invariant Pooling, and Classification Network. The backbone is equivariant with respect to D_p $(= G_X = G_\phi)$, and the goal is to create features that use parameters optimally by paying no attention to geometrically modelable variations, and leave that to the next layer, which is the invariant pooling with respect to D_p. The object class is predicted by a classifier network, which is typically an MLP. For H-NeXt, we always consider D_p to be roto-translation and the backbone to be commutative with respect to D_p.

3.1 Equivariant Backbone

Traditional CNNs are translation equivariant, but not rotation equivariant, i.e. rotation changes not only the position of a feature, but also its value. To achieve rotation equivariance in our model, the values must be independent of input roto-translation.

The proposed backbone, shown in Figure 2, is based on H-Nets that operate on complex numbers[2]. Inside the backbone, the transformation $T_\theta \in D_p$, denoted by the rotation angle θ and an arbitrary translation, affects the channels (feature maps) $F_m^{T_\theta}$ with respect to the original channels F_m^0 as follows

[2]Note that complex polar form $z \in \mathbb{C}; z = re^{i\varphi}$ (r is the magnitude (distance) and φ the angle (phase)) is used.

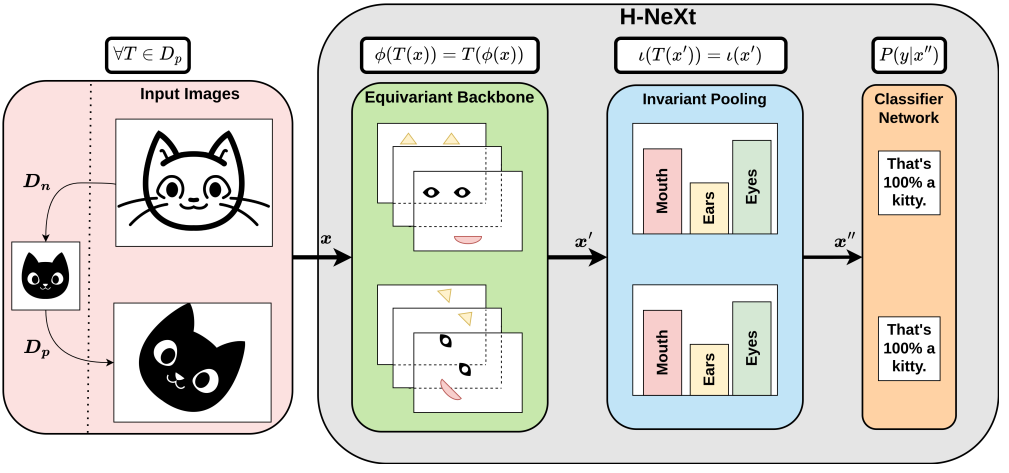

Figure 1: The proposed H-NeXt network invariant with respect to roto-translation (D_p) consisting of three parts: Roto-translation Equivariant Backbone, Roto-translation Invariant Pooling and Classifier Network.

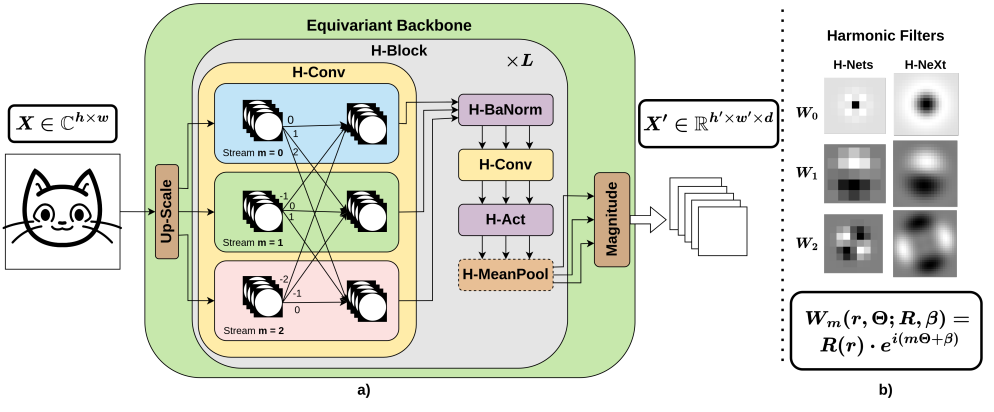

Figure 2: (a) Equivariant Backbone, the principal part of H-NeXt divided into its individual layers and (b), the comparison of H-Nets and H-NeXt Harmonic Convolution Filters

$$F_m^{T_\theta}(x,y) = e^{im\theta} F_m^0(T_\theta(x,y)). \tag{2}$$

Rotation changes only the phase, which is discarded at the end of the backbone, and the magnitude is used as the output. The channels are divided into streams according to $m \in \mathbb{N}$, which is called rotation order.

Harmonic convolution (H-Conv) H-Convs are convolution layers with filters limited to the circular harmonic family. Each filter is defined as

$$W_{m_1} = W_{m_1}(r, \Theta; R, \beta) = R(r) \cdot e^{i(m_1\Theta+\beta)}, \tag{3}$$

where (r,Θ) are polar coordinates, $R:\mathbb{R}^+\to\mathbb{R}$ is a radial function, $\beta\in[0,2\pi)$ is the phase offset and $m_1\in\mathbb{Z}$ is the rotation order, representing axial filter symmetry as illustrated in Figure 2b. Eq. (3) implies that convolution of channels $F^{T_\theta}_{m_2}$ with W_{m_1} follows

$$T_\theta(W_{m_1}(x,y)\star F^{T_\theta}_{m_2}(x,y)) = e^{i\theta(m_1+m_2)}\cdot(W_{m_1}(T_\theta(x,y))\star F^{0}_{m_2}(T_\theta(x,y)).^{3} \tag{4}$$

The result will then have the rotation order m_1+m_2. Even though streams are otherwise processed separately in the backbone, Eq. (4) allows H-Convs to mix the streams without breaking (2). Resulting channels are reassembled into corresponding streams; note diagonal and horizontal connections inside H-Convs layer in Fig 2a. To keep the number of streams fixed, H-Convs applies filters of rotation orders from $-m$ to m.

Each further block of the backbone must satisfy (2) to hold equivariance.

Operating on magnitude (H-Act, H-BaNorm) The condition defined by (2) is preserved if any arbitrary function, such as activation functions or batch normalization, is applied only to the magnitude, leaving the phase unchanged. *E.g.* Harmonic ReLU with bias b for input z is derived as follows

$$z=|z|\cdot e^{\alpha i}\quad HReLU_b(z)=ReLU(|z|+b)\cdot e^{\alpha i}. \tag{5}$$

From continuous to discrete space (Up-Scale, H-MeanPool) All backbone blocks in continuous space satisfy the condition given by (2), but the images and network channels are discrete square grids, which leads to discretization inaccuracies. For example, Mean Pooling clearly follows (2) in continuous space, but the rougher the feature discretization, the greater the difference for roto-translated input. To minimize this, we added an Up-Scale layer at the beginning of the backbone, which significantly reduces the inaccuracies of Mean Pooling and the other layers.

Not only the discrete features cause the equivariance violation, but also the filter discretization of the H-Convs has a non-negligible impact. As can be seen in Figure 2b, the H-Nets filters break the circular symmetry, which gets even worse with increasing rotation order. To avoid this, while keeping exactly the same number of weights, H-NeXt uses larger (15×15) filters. Each H-Conv filter has a learnable vector of size n representing the radial function R. But in contrast with H-Nets, where R could be seen as n rings with a distance of 1 px between each other, H-NeXt rings are evenly spread over the entire filter spatial size, as illustrated in Figures 2b and 3b. Using larger filters allows us to increase the maximum rotation order, contrary to Worall *et al.*'s [47] conclusions that streams with orders greater than 1 are not beneficial.

These changes are referred to as **UP** in the experiments that follow in Section 5.

Channel shape prior Another straightforward modification with a significant impact is to constrain the feature maps to a circular shape. The square channels used in most CNNs break the rotational equivariance due to the boundary effect, as shown in Figure 3a. In H-NeXt, the constraint is implemented by applying a circular mask to each of the channels. These changes are referred to as **MASK** in the experiments that follow in Section 5.

[3] Proof formulated by Worall *et al.* [47] in Supplementary Material.

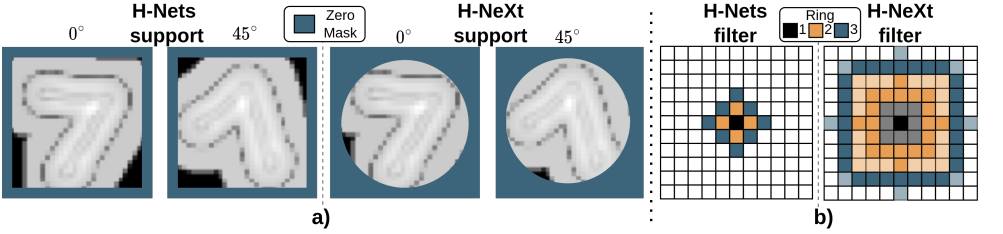

Figure 3: Comparison of H-Nets and H-NeXt (a) channel shapes (b) convolution filters discretization

Backbone output The last layer of the backbone discards the rotation-dependent phases, and returns only the magnitudes as roto-translation equivariant features. In H-Nets the classification follows directly, which means that the pooling yields logits equal to the number of classes. Therefore, either only the $m = 0$ stream is used, or the streams are merged by summing over the rotation orders, and in both cases aggregated by GAP.

Our experiments showed that for more complex tasks, we need to increase the complexity of the model, and simply adding more channels and blocks is not effective. Instead, we propose to use the magnitudes of each order separately, concatenate them, aggregate them by GAP, and feed them into a classification layer. The training loss decreases much faster due to a better gradient distribution across different rotation orders. These changes are referred to as **WIDE** in the experiments that follow in Section 5.

3.2 Invariant Pooling

We propose three approaches to obtain invariant features as illustrated in Figure 4. The simplest approach is to use GAP as in H-Nets, which is obviously roto-translation invariant.

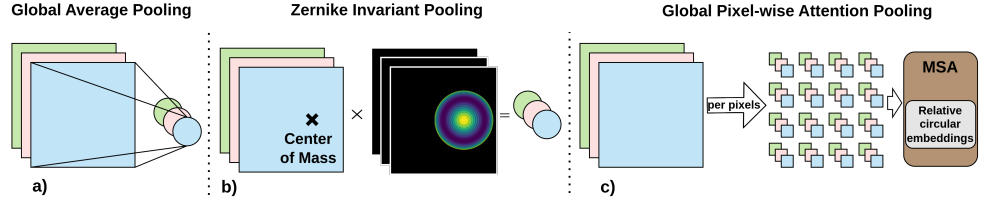

Figure 4: Three roto-translation invariant pooling options: (a) Global Average Pooling, (b) Zernike Invariant Pooling and (c) Invariant Multi-Head Self-Attention Pooling

Zernike moment invariants [43], as published by Singh *et al.* [41, 42], or other rotation moment-based invariants [1, 17, 18] can be used as pooling. Translation invariance is preserved by shifting the polynomials to the center of mass of each channel.

To further increase the complexity of the model, we propose to use Multi-Head Self-Attention (MSA) as an invariant pooling layer because it allows more intricate interactions. The input to MSA must be divided into patches, but roto-translation changes the pixel distribution into patches leading to invariance violation. To improve invariance, we work with a grid of the smallest possible windows $1 \times 1 \times d$, where d is the number of input channels. For the same reason, we cannot add classical positional embeddings to the patches. To pro-

vide spatial information, relative circular embeddings similar to iRPE have been used [18]. The rest of the MSA is roto-translation invariant because the ordering of the patches doesn't matter.

4 Datasets

The experiments were performed on three benchmarks summarized in Table 1: ① Two newly generated datasets with unaugmented training set and randomly rotated test sets were created to demonstrate the invariance improvements. ② For comparison with the state-of-the-art, we follow the evaluation setup of SWN-GCN, which includes multiple test sets rotated by fixed angles $(0°, 30°, \ldots, 330°)$. Unlike ①, the whole range of angles is not covered, *e.g.* 45°, where accuracy is affected the most. ③ To relate the equivariance and invariance tasks, H-NeXt was also tested on rot-mnist [29].

Table 1: Datasets used in experiments (Exp. id), including properties such as sizes, training (Rot. train) or test (Rot. test) set rotation, and whether the rotation performed (Rot. type) was at fixed (Fix) or at random angles (Rnd).

Dataset name	Exp. id	Original dataset	Train size	Test size	Valid size	Rot. train	Rot. test	Rot. type	Ref.
mnist-rot-test	①	MNIST	50k	10k	10k	✗	✓	Rnd	Ours
swn-gcn-mnist	②	MNIST	50k	10k	10k	✗	✓	Fix	[23]
rot-mnist	③	MNIST	10k	2k	50k	✓	✓	Rnd	[29]
cifar-rot-test	①	CIFAR10	42k	10k	8k	✗	✓	Rnd	Ours
swn-gcn-cifar	②	CIFAR10	42k	10k	8k	✗	✓	Fix	[23]

5 Experiments

Three experiment settings were proposed in accordance with the datasets: ① The impact of H-NeXt architecture changes is investigated on our benchmarks. ② We present the state-of-the-art results in comparison with the SWN-GCN [23] setup. ③ The equivariance is tested on rot-mnist [29], which is a standard benchmark for equivariance models.[4]

① **H-NeXt Ablation** Models are trained on images in the upright position and evaluated on the randomly rotated test set. To evaluate the invariance capability of the network we propose to measure the difference between the validation accuracy at fixed angles 0° (upright position) and at 45°.

The effects of each backbone enhancement on the mnist-rot-test are reported in Figure 5. The performance is compared to H-Nets, which has a significant gap of ∼13% between the classification of 0° and 45° rotated images. The **UP** model shrinks this gap to ∼7%. Decreasing the discretization effects by including an Up-Scale layer and larger filters improves the invariance, as expected. Adding circular channels (**UP+MASK**) further reduces the gap to ∼1%. Both of these changes lead to a significant improvement in test error and hence

[4] Comprehensive experimental settings are listed in the Supplementary Material.

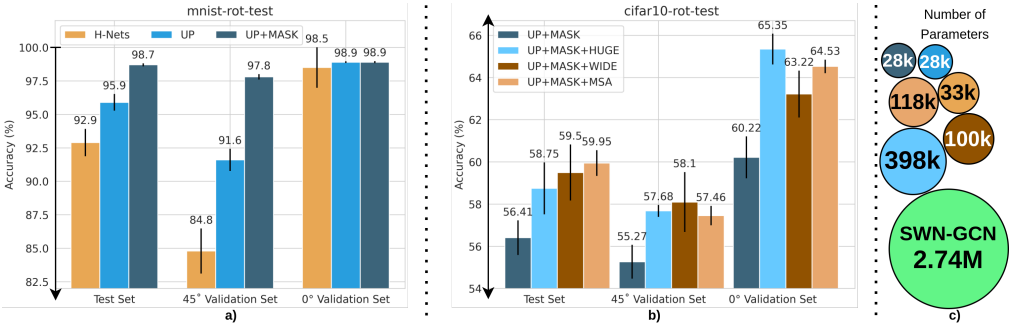

Figure 5: H-NeXt modifications impact on (a) mnist-rot-test accuracy (b) cifar10-rot-test accuracy and (c) model sizes. Test set contains randomly rotated images and validation sets are for specific angles.

higher invariance. Further improvements on the mnist-rot-test may not be possible because of interpolation artifacts, as the model did not see any rotated data during training.

Different types of pooling are tested on cifar-rot-mnist[5], the results are shown in Figure 5b. **HUGE** has the worst performance on the test set (58.75 ± 1.23), although it works best for $0°$, it does not perform well for the other rotations. It is outperformed by **WIDE** (59.50 ± 1.33) with four times fewer parameters, showing that increasing channels and blocks in **HUGE** is highly inefficient. By replacing the global average pooling with an invariant self-attention (**MSA**), the test accuracy is the highest (59.95 ± 0.61). **MSA** increases performance for $0°$, improves training stability and reduces variance, but slightly decreases $45°$ accuracy. This shows that **WIDE** is less affected by interpolation, but on average **MSA** performs better.

② **SWN-GCN comparison** We train H-Next without any augmentations and measure the error on the test sets rotated by fixed angles $0°, 30°, 60°, \ldots$, following SWN-GCN evaluation setup [23].[6] As can be seen in Table 2, H-NeXt for MNIST surpasses the current state-of-the-art using just 28k parameters, which is a fraction of the 2.7M SWN-GCN parameters as illustrated in Figure 5c. In the case of CIFAR (see Table 3), the accuracy is improved by 7% using models with up to 100k parameters, which is still significantly smaller than SWN-GCN. Advanced pooling adds another 2–4% using fewer parameters, where **MSA** works better for small rotations, but worse for large angles, and conversely the model **WIDE**, has more difficulties with the $0°$ but works better at other angles.

③ **Equivariant Context** To establish a connection between the equivariance and invariance tasks, we include the results on rot-mnist in Figure 6a. Equivariant models (E(2)-CNN [44], H-Nets [47]) perform differently when trained on augmented and unaugmented data. Unlike hard invariant models (H-NeXt), the equivariant models leverage the information about rotation, when trained on augmented data.

For a comparison of mnist-rot-test and rot-mnist, we show the performance of H-NeXt on a different size with respect to the training size in Figure 6b. The accuracy of mnist-rot-test is lower than that of rot-mnist when the training size is the same (10k). Achieving

[5] Note that CIFAR-10 images are circularly masked to remove rotation artifacts.

[6] A comprehensive accuracy list, including all angles, is included in the Supplementary Materials.

Table 2: Performance comparison on the MNIST invariance benchmark following the SWN-GCN [23] setup. OA is the overall accuracy of all fixed angles.

MNIST Models	0°	30°	120°	150°	210°	240°	300°	330°	**OA**
RESNET-50 [23]	**99.50**	91.90	29.70	48.80	51.10	33.70	35.20	90.00	42.40
E(2)-CNN [42]	99.30	98.10	86.20	74.90	71.10	81.80	92.90	97.00	87.50
TIGRANET [25]	89.10	82.70	82.70	79.80	82.70	79.80	82.70	79.80	85.10
SWN-GCN [23]	96.50	89.80	89.80	87.30	89.80	87.30	89.80	87.30	91.80
H-Nets [47]	98.70	89.41	89.41	90.55	89.41	90.55	89.40	90.55	92.89
UP+MASK	98.94	**98.55**	**98.55**	**98.55**	**98.55**	**98.55**	**98.55**	**98.54**	**98.68**

Table 3: Performance comparison on the CIFAR-10 invariance benchmark following the SWN-GCN [23] setup. OA is the overall accuracy of all fixed angles

CIFAR Models	0°	30°	120°	150°	210°	240°	300°	330°	**OA**
RESNET-50 [23]	**85.10**	54.50	27.50	26.90	27.00	24.90	33.20	52.50	36.10
E(2)-CNN [42]	77.10	57.80	34.40	30.80	31.90	35.40	45.00	56.00	46.20
TIGRANET [25]	38.90	37.00	37.00	36.80	37.00	36.80	37.00	36.80	38.10
SWN-GCN [23]	51.30	49.60	49.60	50.10	49.60	50.10	49.60	50.10	50.50
UP+MASK	59.67	56.26	56.26	56.27	56.26	56.27	56.26	56.27	57.40
UP+MASK+WIDE	62.80	**60.31**	**60.31**	**60.35**	**60.30**	**60.36**	**60.31**	**60.36**	61.16
UP+MASK+MSA	64.15	60.09	60.08	60.13	60.09	60.13	60.09	60.13	**61.46**

invariance is more difficult when no rotated data is seen during training.

Moment invariant pooling (Fig. 4) was tested in both settings ① and ③ with the same performance as our model **UP+MASK**. This leads us to the conclusion that the circular support, which is used by both the Zernike Moment Pooling and the **UP+MASK** model, affects the model performance the most.

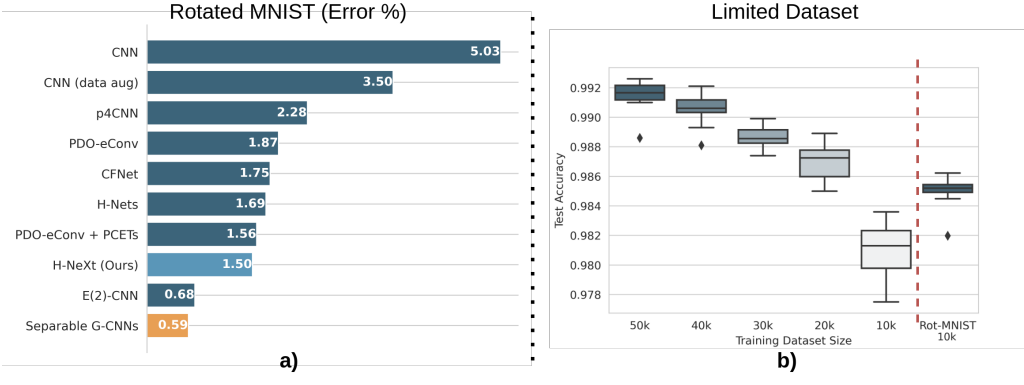

Figure 6: (a) Comparison of the models on the rot-mnist, namely: CNN [9], CNN(data aug) [9], G-CNN (p4CNN) [9], PDO-eConv [37], CFNet [8], H-Nets [47], PDO-eConv+PCETs [42], H-NeXt (Ours), E(2)-CNN[45] and G-CNNs[26]. (b) Comparison of H-NeXt performance between mnist-rot-test and rot-mnist (10k) on the limited training set.

6 Conclusion

We propose H-NeXt, a roto-translation invariant network that achieves state-of-the-art results with fewer parameters, as validated by the invariance benchmarks according to the SWN-GCN [23] setup. Our network targets tasks where a strong inductive bias is advantageous, for example, where roto-translated objects are naturally present. Restrictions on the architecture imposed by invariance constraints lower the maximum achievable model recognition capabilities. H-NeXt is thus not yet applicable in every context, but it is a step towards purely universal and invariant networks.

Acknowledgements: This work was supported by the Czech Science Foundation under the grant No. GA21-03921S and by the *Praemium Academiae*. Special thanks to Adam Harmanec for proofreading and valuable comments.

References

[1] Leonid Bedratyuk, Jan Flusser, Tomáš Suk, Jitka Kostková, and Jaroslav Kautsky. Non-separable rotation moment invariants. *Pattern Recognition*, 127:108607, 2022.

[2] Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 440–448. Springer International Publishing, 2018. ISBN 978-3-030-00928-1. doi: 10.1007/978-3-030-00928-1_50.

[3] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 111–118. Omnipress, 2010. ISBN 978-1-60558-907-7.

[4] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. doi: 10.1109/TPAMI.2012.230.

[5] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. doi: 10.1109/TGRS.2016.2601622.

[6] Gong Cheng, Junwei Han, Peicheng Zhou, and Dong Xu. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Transactions on Image Processing*, 28(1):265–278, 2018. doi: 10.1109/TIP.2018.2867198.

[7] Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, and Minh N Do. Enhanced rotation-equivariant u-net for nuclear segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1097–1104, 2019. doi: 10.1109/CVPRW.2019.00143.

[8] Benjamin Chidester, Tianming Zhou, Minh N Do, and Jian Ma. Rotation equivariant and invariant neural networks for microscopy image analysis. *Bioinformatics (Oxford, England)*, 35(14):i530–i537, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz353.

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016. doi: 10.48550/arXiv.1602.07576.

[10] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2019. doi: 10.48550/arXiv.1902.04615.

[11] Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1612.08498.

[12] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1801.10130.

[13] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12200–12209, 2021. doi: 10.48550/arXiv.2104.12229.

[14] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1889–1898. PMLR, 2016. doi: 10.48550/arXiv.1602.02660.

[15] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2849–2858, 2019. doi: 10.48550/arXiv.1812.00155.

[16] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1709.01889.

[17] Jan Flusser, Barbara Zitova, and Tomas Suk. *Moments and moment invariants in pattern recognition*. John Wiley & Sons, 2009.

[18] Jan Flusser, Tomas Suk, and Barbara Zitová. *2D and 3D Image Analysis by Moments*. John Wiley & Sons, 2016. ISBN 978-1-119-03935-8.

[19] William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991. ISSN 0162-8828. doi: 10.1109/34.93808.

[20] Robert Gens and Pedro M Domingos. Deep symmetry networks. *Advances in Neural Information Processing Systems*, 27, 2014.

[21] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. ReDet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2786–2795, 2021. doi: 10.48550/arXiv.2103.07733.

[22] Emiel Hoogeboom, Jorn W.T. Peters, Taco S. Cohen, and Max Welling. HexaConv. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1803.02108.

[23] 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.

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Kavukcuoglu, Koray. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015. doi: 10.48550/arXiv.1506.02025.

[25] 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.

[26] David M Knigge, David W Romero, and Erik J Bekkers. Exploiting redundancy: Separable group convolutional networks on lie groups. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11359–11386. PMLR, 2022. doi: 10.48550/arXiv.2110.13059.

[27] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2747–2755. PMLR, 2018. doi: 10.48550/arXiv.1802.03690.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. ISSN 0001-0782. doi: 10.1145/3065386.

[29] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 473–480. Association for Computing Machinery, 2007. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273556.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[31] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, 2015. doi: 10.48550/arXiv.1411.5908.

[32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. doi: 10.48550/arXiv.2201.03545.

[33] Diego Marcos, Michele Volpi, and Devis Tuia. Learning rotation invariant convolutional filters for texture classification. In *23rd International Conference on Pattern Recognition (ICPR)*, pages 2012–2017. IEEE, 2016. doi: 10.1109/ICPR.2016.7899932.

[34] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 10.1109/ICCV.2017.540, pages 5048–5057, 2017.

[35] Alhassan Mumuni and Fuseini Mumuni. CNN architectures for geometric transformation-invariant feature representation in computer vision: A review. *SN Computer Science*, 2:1–23, 2021. doi: 10.1007/s42979-021-00735-0.

[36] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2865–2873, 2015. doi: 10.48550/arXiv.1412.8659.

[37] Zhengyang Shen, Lingshen He, Zhouchen Lin, and Jinwen Ma. Pdo-econvs: Partial differential operator based equivariant convolutions. In *International Conference on Machine Learning*, pages 8697–8706. PMLR, 2020. doi: 10.48550/arXiv.2007.10408.

[38] Laurent Sifre and Stéphane Mallat. Combined scattering for rotation invariant texture analysis. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, volume 44, pages 68–81, 2012. ISBN 978-2-87419-049-0.

[39] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1233–1240, 2013. doi: 10.1109/CVPR.2013.163.

[40] Laurent Sifre and Stéphane Mallat. Rigid-motion scattering for texture classification. *arXiv preprint arXiv:1403.1687*, 2014. doi: 10.48550/arXiv.1403.1687.

[41] Jaspreet Singh and Chandan Singh. Learning invariant representations for equivariant neural networks using orthogonal moments. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9892412.

[42] Jaspreet Singh, Chandan Singh, and Ankur Rana. Orthogonal transforms for learning invariant representations in equivariant neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1523–1530, 2023. doi: 10.1109/WACV56688.2023.00157.

[43] Michael Reed Teague. Image analysis via the general theory of moments∗. *J. Opt. Soc. Am.*, 70(8):920–930, Aug 1980. doi: 10.1364/JOSA.70.000920. URL https://opg.optica.org/abstract.cfm?URI=josa-70-8-920.

[44] 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.

[45] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 849–858, 2018. doi: 10.48550/arXiv.1711.07289.

[46] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. doi: 10.48550/arXiv.2301.00808.

[47] 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.

[48] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10033–10041, 2021. doi: 10.48550/arXiv.2107.14222.

[49] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – European Conference on Computer Vision 2014*, pages 818–833. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10590-1_53.