



Roto-Translation Covariant Convolutional Networks for Medical Image Analysis

Erik J. Bekkers¹(✉), Maxime W. Lafarge²(✉), Mitko Veta²,
Koen A. J. Eppenhof², Josien P. W. Pluim², and Remco Duits¹

¹ Department of Mathematics and Computer Science,
Eindhoven University of Technology, Eindhoven, The Netherlands
e.j.bekkers@tue.nl

² Department of Biomedical Engineering, Eindhoven University of Technology,
Eindhoven, The Netherlands
m.w.lafarge@tue.nl

Abstract. We propose a framework for rotation and translation covariant deep learning using $SE(2)$ group convolutions. The group product of the special Euclidean motion group $SE(2)$ describes how a concatenation of two roto-translations results in a net roto-translation. We encode this geometric structure into convolutional neural networks (CNNs) via $SE(2)$ group convolutional layers, which fit into the standard 2D CNN framework, and which allow to generically deal with rotated input samples without the need for data augmentation.

We introduce three layers: a *lifting layer* which lifts a 2D (vector valued) image to an $SE(2)$ -image, i.e., 3D (vector valued) data whose domain is $SE(2)$; a *group convolution layer* from and to an $SE(2)$ -image; and a *projection layer* from an $SE(2)$ -image to a 2D image. The lifting and group convolution layers are $SE(2)$ *covariant* (the output roto-translates with the input). The final projection layer, a maximum intensity projection over rotations, makes the full CNN rotation *invariant*.

We show with three different problems in histopathology, retinal imaging, and electron microscopy that with the proposed group CNNs, state-of-the-art performance can be achieved, without the need for data augmentation by rotation and with increased performance compared to standard CNNs that do rely on augmentation.

Keywords: Group convolutional network · Roto-translation group · Mitosis detection · Vessel segmentation · Cell boundary segmentation

1 Introduction

In this work we generalize \mathbb{R}^2 convolutional neural networks (CNNs) to $SE(2)$ group CNNs (G-CNNs) in which the data lives on position orientation space,

E.J. Bekkers and M.W. Lafarge—Joint main authors.

and in which the convolution layers are defined in terms of representations of the special Euclidean motion group $SE(2)$. In essence this means that we replace the convolutions (with translations of a kernel) by $SE(2)$ group convolutions (with roto-translations of a kernel). The advantage of the proposed approach compared to standard \mathbb{R}^2 CNNs is that rotation covariance is encoded in the network design and does not have to be learned by the convolution kernels. E.g., a feature that may appear in the data under several orientations does not have to be learned for each orientation, but only once. As a result, there is no need for data augmentation by rotation and the kernel weights (that no longer need to learn rotation covariance) become available to increase the CNNs expressive capacity. Moreover, the proposed group convolution layers are compatible with standard CNN modules, allowing for easy integration in popular CNN designs.

A main objective of medical image analysis is to develop models that are invariant to the shape and appearance variability of the structures of interest, including their arbitrary orientations. Rotation-invariance is a desired property, which our G-CNN framework generically deals with. We show state-of-the-art results with improvement over standard 2D CNNs on three different medical imaging tasks: mitosis detection in histopathology images, vessel segmentation in retinal images and cell boundary segmentation in electron microscopy (EM).

1.1 Related Work

In relation to other approaches that incorporate rotation invariance/covariance in the network design, such as harmonic networks [1], local transformation invariance learning [2], deep symmetry nets [3], scattering CNNs [4,5], and warped convolutions [6], the group convolution approaches [4,5,7–10] most naturally extend the standard CNNs by simply replacing the convolution operators.

In the work by Cohen and Welling [7] a comprehensive theoretical framework for G-CNNs is developed for discrete groups whose transformations stay on the pixel grid. In particular their focus was on the wall-paper groups $p4$ (group of translations + 90° rotations), for which a G-CNN approach was also developed by Dieleman et al. [8], and $p4m$ ($p4$ + reflections). In their work it was convincingly demonstrated that including such symmetries, by replacing standard convolutions by group convolutions, substantially increases the network’s performance without increasing the number of network variables. Although their theoretical G-CNN framework [7] holds for more general groups, their actual application scope was limited to discrete groups that stay on the pixel grid. In this paper, we are not restricted to such groups, but include efficient bi-linear interpolation that allows us to employ the full structure of the continuous roto-translation group $SE(2)$, which we can discretize to the sub-group $SE(2, N)$, with N rotations. Special cases of our framework are standard 2D CNNs when $N = 1$ and the $p4$ G-CNNs as proposed in [7,8] when $N = 4$.

In very recent work, Weiler et al. [9] describe a different approach to $SE(2)$ G-CNNs. Instead of relying on interpolation they used 2D complex-valued steerable kernels, which has the advantage that kernel rotations are exact. A disadvantage

is, however, that these kernels are constrained to a specific combination of complex valued basis functions. With our interpolation approach, kernel rotation simply appears in the CNN architecture as a (sparse) matrix-vector multiplication, that maps a set of base weights to a full set of rotated kernels.

In work by Mallat, Ouyallon, and Sifre [4,5] roto-translation invariant deep networks are formulated in the context of scattering theory. Their design involves a concatenation of separable group convolutions with hand-crafted (but well underpinned) filters, followed by the modulus as activation function. Learning takes place via support vector machines on the generated $SE(2)$ invariant descriptors. In our approach, the filters are learned without restrictions, the convolutions do not have to be separable, and we here use the common ReLU activation function.

In work by Bekkers et al. [10], an effective template matching method was proposed using group correlations in orientation scores, which are $SE(2)$ images obtained from a 2D image via lifting convolutions with a specific choice of kernel [11]. The $SE(2)$ templates were put in a B-spline basis (allowing for exact kernel rotations) and optimized via logistic regression. Their architecture fits within our framework as a single channel G-CNN of depth 2 with a fixed lifting kernel.

2 $SE(2)$ Convolutional Neural Networks

2.1 Group Theoretical Preliminaries

The Lie Group $SE(2)$: The group $SE(2) = \mathbb{R}^2 \rtimes SO(2)$ is the semi-direct product of the group of planar translations \mathbb{R}^2 and rotations $SO(2)$, and its group product is given by

$$g \cdot g' = (\mathbf{x}, \mathbf{R}_\theta) \cdot (\mathbf{x}', \mathbf{R}_{\theta'}) = (\mathbf{R}_\theta \mathbf{x}' + \mathbf{x}, \mathbf{R}_{\theta + \theta'}), \quad (1)$$

with group elements $g = (\mathbf{x}, \theta)$, $g' = (\mathbf{x}', \theta') \in SE(2)$, with translations \mathbf{x}, \mathbf{x}' and planar rotations by θ, θ' . The group acts on the space of positions and orientations $\mathbb{R}^2 \times S^1$ via $g \cdot (\mathbf{x}', \theta') = (\mathbf{R}_\theta \mathbf{x}' + \mathbf{x}, \theta + \theta')$. Since $(\mathbf{x}, \mathbf{R}_\theta) \cdot (\mathbf{0}, 0) = (\mathbf{x}, \theta)$, we can identify the group $SE(2)$ with the space of positions and orientations $\mathbb{R}^2 \times S^1$. As such we will often write $g = (\mathbf{x}, \theta)$, instead of $(\mathbf{x}, \mathbf{R}_\theta)$. Note that $g^{-1} = (-\mathbf{R}_\theta^{-1} \mathbf{x}, -\theta)$ since $g \cdot g^{-1} = g^{-1} \cdot g = (\mathbf{0}, 0)$.

Group Representations: The structure of the group can be mapped to other mathematical objects (such as 2D images) via representations. The left-regular $SE(2)$ representation on 2D images $f \in \mathbb{L}_2(\mathbb{R}^2)$ is given by

$$(\mathcal{U}_g f)(\mathbf{x}') = f(\mathbf{R}_\theta^{-1}(\mathbf{x}' - \mathbf{x})), \quad (2)$$

with $g = (\mathbf{x}, \theta) \in SE(2)$, $\mathbf{x}' \in \mathbb{R}^2$. It corresponds to a roto-translation of the image. The left-regular representation on functions $F \in \mathbb{L}_2(SE(2))$ on $SE(2)$, which we refer to as $SE(2)$ -images, is given by

$$(\mathcal{L}_g F)(g') = F(g^{-1} \cdot g') = F(\mathbf{R}_\theta^{-1}(\mathbf{x}' - \mathbf{x}), \theta' - \theta), \quad (3)$$

with $g = (\mathbf{x}, \theta)$, $g' = (\mathbf{x}', \theta') \in SE(2)$. It is a shift-twist (rotation + θ -shift) of F , see e.g. Figure 1. Next we define the G-CNN layers in terms of these representations.

2.2 The $SE(2)$ Group Convolution Layers

In CNNs one can take a convolution or a cross-correlation viewpoint and since these operators simply relate via a kernel reflection, the terminology is often used interchangeably. We take the second viewpoint, our G-CNNs are implemented using cross-correlations. On \mathbb{R}^2 we define cross-correlation via inner products of translated kernels:

$$(k \star_{\mathbb{R}^2} f)(\mathbf{x}) := (\mathcal{T}_{\mathbf{x}}k, f)_{\mathbb{L}_2(\mathbb{R}^2)} := \int_{\mathbb{R}^2} k(\mathbf{x}' - \mathbf{x})f(\mathbf{x}')d\mathbf{x}', \quad (4)$$

with $\mathcal{T}_{\mathbf{x}}$ the translation operator, the left-regular representation of the translation group $(\mathbb{R}^2, +)$. In the $SE(2)$ lifting layer we now simply replace translations of k by roto-translations via the $SE(2)$ representation \mathcal{U}_g defined in Eq. (2).

The $SE(2)$ Lifting Layer: Let $\underline{f}, \underline{k} : \mathbb{R}^2 \rightarrow \mathbb{R}^{N_c}$ be a vector valued 2D image and kernel (with N_c channels), with $\underline{f} = (f_1, \dots, f_{N_c})$ and $\underline{k} = (k_1, \dots, k_{N_c})$, then the group lifting correlations for vector valued images are defined by

$$(\underline{k} \tilde{\star} \underline{f})(g) := \sum_{c=1}^{N_c} (\mathcal{U}_g k_c, f_c)_{\mathbb{L}_2(\mathbb{R}^2)} = \sum_{c=1}^{N_c} \int_{\mathbb{R}^2} k_c(\mathbf{R}_\theta^{-1}(\mathbf{y} - \mathbf{x}))f_c(\mathbf{y})d\mathbf{y}. \quad (5)$$

These correlations *lift* 2D image data to data that lives on the 3D position orientation space $\mathbb{R}^2 \times S^1 \equiv SE(2)$. The *lifting layer* that maps from a vector image $\underline{f}^{(l-1)} : \mathbb{R}^2 \rightarrow \mathbb{R}^{N_{l-1}}$, with N_{l-1} channels at layer $l-1$, to an $SE(2)$ vector image $\underline{F}^{(l)} : SE(2) \rightarrow \mathbb{R}^{N_l}$ using a set of N_l kernels $\mathbf{k}^{(l)} := (\underline{k}_1^{(l)}, \dots, \underline{k}_{N_l}^{(l)})$, each with N_{l-1} channels, is then defined by

$$\underline{F}^{(l)} = \mathbf{k}^{(l)} \tilde{\star} \underline{f}^{(l-1)} := \left(\underline{k}_1^{(l)} \tilde{\star} \underline{f}^{(l-1)}, \dots, \underline{k}_{N_l}^{(l)} \tilde{\star} \underline{f}^{(l-1)} \right). \quad (6)$$

The $SE(2)$ Group Convolution Layer: Let $\underline{F}, \underline{K} : SE(2) \rightarrow \mathbb{R}^{N_c}$ be a vector valued $SE(2)$ image and kernel, with $\underline{F} = (F_1, \dots, F_{N_c})$ and $\underline{K} = (K_1, \dots, K_{N_c})$, then the group correlations are defined as

$$(\underline{K} \star \underline{F})(g) := \sum_{c=1}^{N_c} (\mathcal{L}_g K_c, F_c)_{\mathbb{L}_2(SE(2))} = \sum_{c=1}^{N_c} \int_{SE(2)} K_c(g^{-1} \cdot h)F_c(h)dh, \quad (7)$$

with $(K, F)_{\mathbb{L}_2(SE(2))} := \int_{SE(2)} K(h)F(h)dh$, the inner product on $\mathbb{L}_2(SE(2))$. A set of $SE(2)$ kernels $\mathbf{K}^{(l)} := (\underline{K}_1^{(l)}, \dots, \underline{K}_{N_l}^{(l)})$ defines a group convolution layer, mapping from $\underline{F}^{(l-1)}$ with $N_{(l-1)}$ channels to $\underline{F}^{(l)}$ with $N_{(l)}$ channels, via

$$\underline{F}^{(l)} = \mathbf{K}^{(l)} \star \underline{F}^{(l-1)} := \left(\underline{K}_1^{(l)} \star \underline{F}^{(l-1)}, \dots, \underline{K}_{N_l}^{(l)} \star \underline{F}^{(l-1)} \right). \quad (8)$$

The Projection Layer: Projects a multi-channel $SE(2)$ image back to \mathbb{R}^2 via

$$\underline{f}^{(l)}(\mathbf{x}) = \max_{\theta \in [0, 2\pi]} \underline{F}^{(l)}(\mathbf{x}, \theta). \quad (9)$$

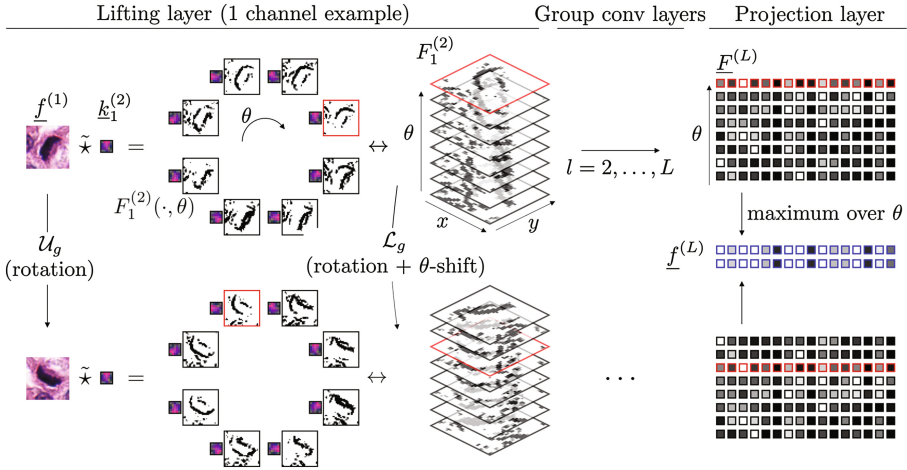


Fig. 1. Rotation co- and invariance. Top row: the activations after the lifting convolutions with a single kernel $k_1^{(2)}$, stacked together it yields an $SE(2)$ image $F_1^{(2)}$ (cf. Eq. (6)). The projection layer at the end of the pipeline gives a rotation invariant feature vector. Bottom row: the same figures with a rotated input.

2.3 Discretization and Network Architecture

Discretization, Kernel Sizes and Rotation: Discretized 2D images are supported on a bounded subset of $\mathbb{Z}^2 \subset \mathbb{R}^2$ and the kernels live on a spatially rectangular grid of size $n \times n$ in \mathbb{Z}^2 , with n the kernel size. We discretize the Lie group $SE(2, N) := \mathbb{Z}^2 \rtimes SO(2, N)$, with the space of 2D rotations in $SO(2)$ sampled with N rotation angles $\theta_i = \frac{2\pi}{N}i$, with $i = 0, \dots, N - 1$. The discrete lifting kernels $\mathbf{k}^{(l)}$ at layer l , mapping from a 2D image with N_{l-1} input channels to an $SE(2, N)$ image with N_l channels, thus have a shape of $n \times n \times N_{l-1} \times N_l$. The $SE(2, N)$ kernels $\mathbf{K}^{(l)}$ have a shape of $n \times n \times N \times N_{l-1} \times N_l$. A complete set of rotations of kernels $\mathbf{k}^{(l)}$ or $\mathbf{K}^{(l)}$ can be constructed with a single matrix multiplication from a vector that contains the shared kernel weights. This matrix is sparse and encodes bi-linear interpolation and kernel rotation.

3 Experiments and Results

We consider three different tasks in three different modalities. In each we consider the $SE(2, N)$ samplings with $N \in \{1, 2, 4, 8, 16\}$ to study the effect of the choice of N in the $SE(2, N)$ discretization. See Table 1 for the network settings. In each experiment the data is augmented at train and test time with transposed versions of the 2D input. For reference we also include transpose plus 90° rotation augmentation for the $N = 1$ experiment (as in [12, 13]) in order to be able to show that these are not necessary in our $SE(2, N)$ networks for $N \geq 4$. Each experiment is repeated 3 times with random initialization and

Table 1. $SE(2, N)$ chain settings for different orientation samplings N .

N (Group)	1 (\mathbb{Z}^2)	2 ($\mathbb{Z}^2 \times p2$)	4 ($\mathbb{Z}^2 \times p4$)	8 ($SE(2, 8)$)	16 ($SE(2, 16)$)
<i>Layer 1 - lifting with Eq. (6), $n = 5$</i>					
N_1 ($\#_w$)	16 (1040)	13 (845)	10 (650)	8 (520)	6 (390)
<i>Layer 2, 3, 4 - group conv. with Eq. (8), $n = 5$</i>					
N_2 ($\#_w$)	16 (5408)	13 (7124)	10 (8420)	8 (10768)	6 (12108)
N_3 ($\#_w$)	16 (5408)	13 (7124)	10 (8420)	8 (10768)	6 (12108)
N_4 ($\#_w$)	64 (21632)	32 (17536)	16 (13472)	8 (10768)	4 (8072)
<i>Layer 5 - group conv. with Eq. (8) + projection with Eq. (9), $n = 1$</i>					
N_5 ($\#_w$)	16 (1056)	16 (1056)	16 (1056)	16 (1056)	16 (1056)
<i>Layer 6 - standard conv. (output) layer, $n = 1$</i>					
N_6 ($\#_w$)	1 (17)	1 (17)	1 (17)	1 (17)	1 (17)
Total $\#_w$	34561	33702	32035	33897	33751

sampling to get a rough estimate of the mean and variance on the performance. For a fair comparison for different N the overall number of weights is matched. For a fair comparison with the \mathbb{R}^2 approach, the number of “2D” activations ($N_l N$) in the last three layers is also matched. Each network optimizes a logistic loss using stochastic gradient descent with momentum using the same settings as in [12]. Our G-CNN implementations are available at <https://github.com/tueimage/se2cnn>. The results are given in Fig. 2, the tasks and metrics are summarized as follows.

Histopathology - Mitosis Detection: The task aims at detecting mitotic figures in hematoxylin-eosin stained slides. We used the public dataset AMIDA13 [14] that consists of high power field images from 23 breast cancer cases. Eight cases (458 mitoses) were used to train the networks with random batches of 68×68 image patches, balanced between mitotic and hard negative figures. This receptive field was obtained by means of max-pooling operations in the first three layers. Sets of candidate detections were generated as in [13] after selection of an operating point on four validation cases (92 mitoses). We assessed an F_1 -score for each model based on the 11 test cases (533 mitoses) in the conditions of [14].

Retina - Vessel Segmentation: In this task the blood vessels in the retina are segmented. For validation we use the public DRIVE database [15], which consists of 40 retinal images with manual segmentations. The set is split in a training set (of which we use 16 for training, and 4 for validation) and a test set of also 20 images. The G-CNNs produce a probability for the vessel and background class. Training is done with 10000 patches (17×17) per class per image. The output probabilities can be thresholded to create a binary segmentation, which can be used to quantify performance in terms of sensitivity and specificity. The area under the receiver operator characteristic (ROC) curve, in short AUC, summarizes these performances into a single value.

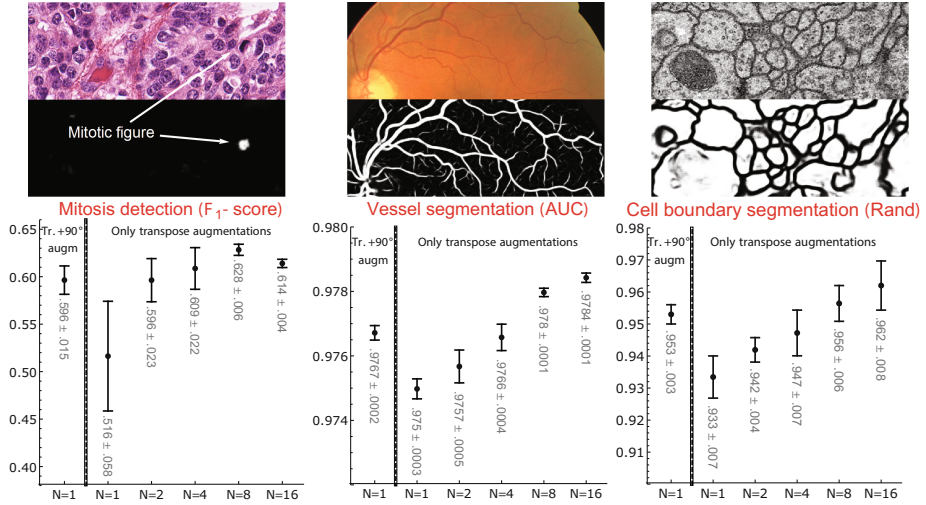


Fig. 2. Top row: Crop outs of images of the three tasks with the class probabilities generated by our method. Bottom row: Mean results (± 1 std. dev.).

Electron Microscopy - Cell Boundary Segmentation: This task consists of segmenting the boundaries of cells that are imaged with EM. We use the data and evaluation system of the ISBI EM segmentation challenge [16]. The data consists of 2 volumes (1 train, 1 test), each containing 30 consecutive images from a serial section transmission EM. Both the segmentation and the evaluation is done by treating the volumes as sequences of 2D slices. To increase receptive field size we include max pooling in the first 2 layers. Training is done with 10000 patches (48×48) per class per image. The main evaluation criterion for the challenge is the Rand score, which measures the similarity between clusterings/connected components [17]. The reported Rand score is the maximum score (for several thresholds) computed for the connected components obtained after thinning of the binary cell boundary segmentation, see [16] for more details.

Results: In each experiment we see that the performance of the baseline with extra rotation augmentations is reached by the non-augmented G-CNNs for $N \geq 4$, and even is surpassed for $N \geq 8$. In the first two experiments we also observe that the variance on the output is reduced with increasing N . Our results on the public datasets match or improve upon the state of the art with the following scores: F_1 -score = 0.628 ± 0.006 for mitosis detection, AUC = 0.9784 ± 0.0001 for vessel segmentation, Rand = 0.962 ± 0.008 for cell boundary segmentation.

4 Discussion and Conclusions

We showed a consistent improvement of performances across three medical image analysis tasks when using G-CNNs compared to their corresponding CNN baselines. The reported results are in line with the benchmark of each dataset and the

best performances were obtained for an orientation capacity $N \geq 4$, indicating the advantage of learning such rotation-invariant representations. We observed improved stability over the repeated experiments in mitosis detection and vessel segmentation for $N = 8$ and $N = 16$, suggesting a regularization effect due to the increased weight sharing with increasing N .

We conclude that it is beneficiary to include $SE(2)$ group convolution layers in CNN network design, as this avoids the need for rotation augmentation and it improves overall performance. In all three medical imaging problems we achieved state-of-the-art results with the same (basic) network design for each task. Based on these results we expect that our $SE(2)$ layers may lead to a further performance increase when embedded in more complex network designs, such as the popular UNets and ResNets.

Acknowledgements. The research leading to these results has received funding from the ERC council under the EC's 7th Framework Programme (FP7/2007–2013)/ERC grant agr. No. 335555.

References

1. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Harmonic networks: deep translation and rotation equivariance. In: CVPR, pp. 5028–5037 (2017)
2. Sohn, K., Lee, H.: Learning invariant representations with local transformations. In: CVPR, pp. 1339–1346. Omnipress (2012)
3. Gens, R., Domingos, P.M.: Deep symmetry networks. In: Advances in Neural Information Processing Systems, pp. 2537–2545 (2014)
4. Sifre, L., Mallat, S.: Rotation, scaling and deformation invariant scattering for texture discrimination. In: CVPR, pp. 1233–1240. IEEE (2013)
5. Oyallon, E., Mallat, S., Sifre, L.: Generic deep networks with wavelet scattering. arXiv preprint [arXiv:1312.5940](https://arxiv.org/abs/1312.5940) (2013)
6. Henriques, J.F., Vedaldi, A.: Warped convolutions: efficient invariance to spatial transformations. In: International Conference on Machine Learning, pp. 1461–1469 (2017)
7. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International Conference on Machine Learning, pp. 2990–2999 (2016)
8. Dieleman, S., De Fauw, J., Kavukcuoglu, K.: Exploiting cyclic symmetry in convolutional neural networks. arXiv preprint [arXiv:1602.02660](https://arxiv.org/abs/1602.02660) (2016)
9. Weiler, M., Hamprecht, F.A., Storath, M.: Learning steerable filters for rotation equivariant CNNs. arXiv preprint [arXiv:1711.07289](https://arxiv.org/abs/1711.07289) (2017)
10. Bekkers, E.J., Loog, M., ter Haar Romeny, B.M., Duits, R.: Template matching via densities on the roto-translation group. IEEE tPAMI **40**(2), 452–466 (2018)
11. Duits, R., Felsberg, M., Granlund, G.H., ter Haar Romeny, B.M.: Image analysis and reconstruction using a wavelet transform constructed from a reducible representation of the Euclidean motion group. IJCV **72**(1), 79–102 (2007)
12. Lafarge, M.W., Pluim, J.P.W., Eppenhof, K.A.J., Moeskops, P., Veta, M.: Domain-adversarial neural networks to address the appearance variability of histopathology images. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 83–91. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_10

13. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 411–418. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_51
14. Veta, M., et al.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. *MEDIA* **20**(1), 237–248 (2015)
15. Staal, J., et al.: Ridge-based vessel segmentation in color images of the retina. *IEEE TMI* **23**(4), 501–509 (2004)
16. Arganda-Carreras, I., et al.: Crowdsourcing the creation of image segmentation algorithms for connectomics. *Front. Neuroanat.* **9**, 142 (2015)
17. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)