Towards True Explainable Models: Rotation Invariant Prototypes for Enhanced Interpretability

Huzeyfe Ayaz[1,2][0000−0002−0585−2400], Andreas Urlberger[1,2][1111−2222−3333−4444], Elvin Abdinli[1,2][2222−−3333−4444−5555], Johannes Kaiser[1,2][2222−−3333−4444−5555], and Sarah Lockfisch[1,2][2222−−3333−4444−5555]

[1] Technical University of Munich, Munich, Germany
[2] AI in Medicine Lab., MRI-TUM, Ismaninger Str. 22, 81675 Munich, Germany
{name.surname}@tum.de

Abstract. In Explainable AI, the prototypical part network (ProtoPNet) is a widely used approach due to its expert-level reasoning, finding prototypical parts, and making final classification decisions based on them. Despite ProtoPNet's high classification accuracy, its unstable behavior and incorrect reasoning, especially on images with applied rotation transformations, raise a crucial question: How many rotation augmentations are necessary to ensure that the prediction result and the model's reasoning remain consistent? A new rotation equivariant solution called ReProtoPNet is proposed to address these issues. ReProtoPNet is a Group-equivariant Convolutional Neural Networks (G-CNN) powered version of ProtoPNet, which enhances the network's capacity to handle rotations without increasing the number of parameters. Our experiments show that ReProtoPNet can achieve consistent reasoning across multiple rotations of the same image, with improved stability in class probabilities compared to ProtoPNet. This paper details the architecture of ReProtoPNet and the modifications made to incorporate G-CNNs. The performance of ReProtoPNet is evaluated on various benchmark datasets, and its superior stability and accuracy are demonstrated in classification tasks involving rotated images. Results indicate that the proposed method retains the interpretability of ProtoPNet and significantly enhances its robustness to rotation transformations. This advancement has substantial implications for deploying explainable AI models in real-world applications where image orientations vary, especially on medical datasets.

Keywords: Explainable-AI · G-CNN · Interpretable Models.

1 Introduction

Artificial intelligence (AI) has appeared as a transformative force in the field of medicine, revolutionizing various applications from diagnostics to treatment planning [1]. Its ability to process vast amounts of data rapidly and accurately

has made it a valuable tool for healthcare professionals, offering significant enhancements in decision-making processes [9]. In medical practice, AI's importance is particularly evident in its ability to aid in complex decisions by analyzing patterns in medical images, predicting disease progression, and personalizing treatment plans. These capabilities have improved patient outcomes and alleviated the burden on healthcare providers by enabling more efficient and informed decision-making.

AI's integration into medical applications is broad, including radiology, which helps detect abnormalities in imaging, and pathology, which improves cancer detection. Additionally, AI-driven tools assist doctors in cross-referencing symptoms, suggesting diagnoses, and recommending treatments based on vast datasets that humans cannot process quickly. However, while AI offers these powerful tools, its "black-box" nature —where the inner workings of the algorithms remain unclear— poses significant challenges in medical decision-making. Clinicians must understand how and why AI arrives at certain conclusions to trust and effectively use these technologies in critical scenarios.

To bridge this gap, the development of explainable AI (XAI) has become a crucial focus. XAI aims to create models that perform well and provide clear, interpretable insights into their decision-making processes. In medicine, interpretable models are precious as they allow doctors to understand the reasoning behind AI-driven diagnoses and treatment suggestions, fostering greater confidence in these tools [3]. By making AI's decisions transparent, healthcare professionals can validate the model's conclusions, ensuring they align with clinical expertise and patient-specific factors.

One notable example of an interpretable model is ProtoPNet [4], a prototype-based neural network designed to make AI decision-making more transparent. ProtoPNet operates by classifying images by comparing them with learned prototypes—specific parts of images from known classes. For instance, in medical imaging, ProtoPNet can identify X-ray regions that resemble a particular disease's prototypical features. The model then bases its predictions on the degree of similarity between the observed image regions and these prototypes. This approach, which mirrors how radiologists compare suspected abnormalities with known pathological patterns, provides a level of interpretability that allows doctors to understand the AI's decision-making process.

Despite its advantages, ProtoPNet has its limitations. A significant challenge arises in the model's stability and consistency, particularly when images are rotated during inference. When an input image is rotated, the corresponding activation maps—used by the model to make predictions—do not remain stable, leading to inconsistent predictions. This lack of rotation equivariance [5] can undermine the reliability of AI-driven diagnostics, as even slight variations in image orientation can result in different outputs, which is unacceptable in critical medical contexts.

To address this issue, our research proposes a solution that integrates Group Equivariant Convolutional Neural Networks (G-CNNs) [5], specifically Steerable-CNNs [6], into ProtoPNet. G-CNNs are designed to maintain consistent feature

extraction across different image transformations, such as rotations. By incorporating Steerable-CNNs, a specific type of G-CNN, our approach ensures that the activation maps remain stable even when the input image is rotated. This enhancement not only preserves the interpretability of ProtoPNet but also significantly improves its reliability and robustness in clinical applications.

While ProtoPNet offers a promising approach to interpretable AI in medicine, its limitations in handling image rotations necessitate further refinement. By combining ProtoPNet with Steerable-CNNs, our research aims to overcome these challenges, ensuring that AI tools in medicine are both interpretable and reliable, ultimately supporting better decision-making and patient care.

2 Methods

2.1 ProtoPNet Architecture

The ProtoPNet's architecture consists of the four following major parts: a backbone model (convolutional layers), prototype comparison, conversion to prototype similarity scores, and finally, an affine layer. These individual steps are visualized in Fig. 1.

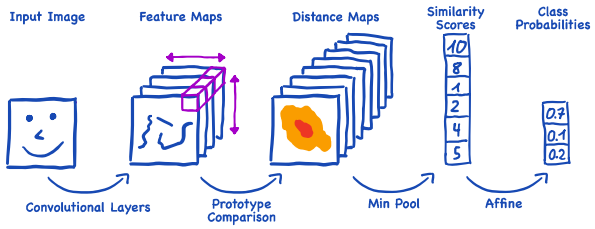

Fig. 1. Simplified architecture overview of ProtoPNet.

In the first part, convolutional layers are separated into two distinct sections: an easily replaceable backbone model and the dynamic add-on layers. The backbone might be any model as long as the spatial dimension of the latent space remains sufficiently large. For example, a spatial dimension of 1x1 would significantly deteriorate the interpretability, as no specific location of the highest prototype similarity could be determined. The add-on layers are a group of 1x1 convolutional layers that serve as a connector between the backbone and the ProtoPNet, allowing us a flexible configuration of the depth of the latent space, i.e., the depth of the prototypes.

The second part, the prototype comparison, is computing the L2 distance between the latent space and each prototype for every spatial location. This operation operates similarly to a convolutional layer with the prototypes being the kernels, just that it is not computing the scalar product of the kernel and

the latent patch but the L2 norm of the difference between the two. This means that we get a distance map for each prototype.

The next part is performing a min pool over the height and width of the distance maps, computing the min distance for each prototype. This distance is then converted into a similarity score, e.g., by taking the negative distance.

The final part is an affine layer, which is responsible for producing the final class probabilities. The weights of this layer are initialized in such a way that the reasoning of the network is rather "It looks like a duck; therefore, it is a duck" than "It does not look like a pig; therefore, it is a duck" [4]. Also, to preserve the weights of a similar structure during training, a particular loss term is added to the overall loss function.

2.2 ReProtoPNet Architecture

Our architecture is closely related to that of ProtoPNet. The key differences are the replacement of the convolutional layers by rotation equivariant convolutional layers followed by a group pool operation, as can be seen in Fig. 2, and a further restriction on the allowed prototype shapes. In the following, when we talk about rotation equivariant convolutional layers, we specifically mean E(n)-equivariant Steerable CNNs implemented in the escnn library [2, 11].

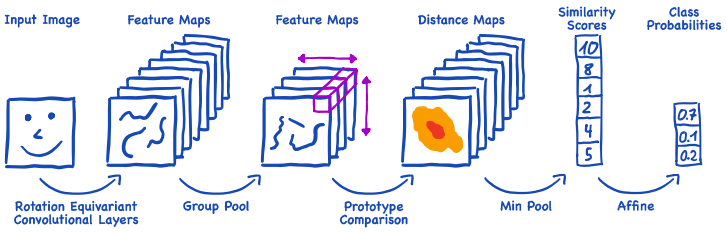

Fig. 2. Simplified architecture overview of ReProtoPNet.

Convolutional and add-on layers in the first and second parts of the architecture, respectively, are replaced with their corresponding rotation equivariant versions. Lastly, we add a group pool layer, which can be considered a max pool over the finite symmetry group, to obtain the same latent space for the rotated versions of an image, e.g. p16 [2]. In theory, the output of these rotation equivariant convolutional layers is rotation equivariant such that Eq. 1 holds, where x is the input image, r is any rotation, and f is the convolutional layers.

$$f(r(x)) = r(f(x)) \tag{1}$$

When initializing ProtoPNet, the prototype shape (spatial height, width, depth) must be defined. The prototype comparison takes place directly after the group pooling, as seen in Fig. 2. To preserve the rotational equivariance,

the spatial height and width of the prototype must be 1, such that a rotation does not affect the resulting L2 distance. This does not pose a fundamental limitation on the expressiveness of ProtoPNet since it aligns very well with the parameter choice made in [4]. Therefore, we apply this constraint to the shape of the prototype.

As outlined above, the rotational equivariance is preserved until the distance maps. The next step is a min pool over the spatial dimensions to achieve rotational and translational invariance. This step outputs the smallest overall distance of each prototype to the latent input image. After passing the min distances through the similarity function, we have a rotation and translation invariant similarity score for each prototype. The final class probabilities are, therefore, also invariant to these operations.

2.3 Weight Down Similarities in Corners

Rotational equivariance does not hold outside of a particular circular image patch. For this reason, we use escnn's masking module to apply a smooth circular mask to the input image, setting all values outside of the mask to zero. This ensures that no corner information, which would be cut off when the image is rotated, is used for the network's decision-making. Although this helps a lot, there will still be artifacts, i.e., parts that are not rotation equivariant, in the output of the convolutional layers. These will negatively affect the rotational invariance of the class probabilities. Therefore, we do not want prototypes to focus on these regions. Hence, we suggest applying a similar mask to the distance maps. By increasing the distances outside the mask, we weigh down the similarity scores of these patches. Fig. 3 visualizes the final architecture.

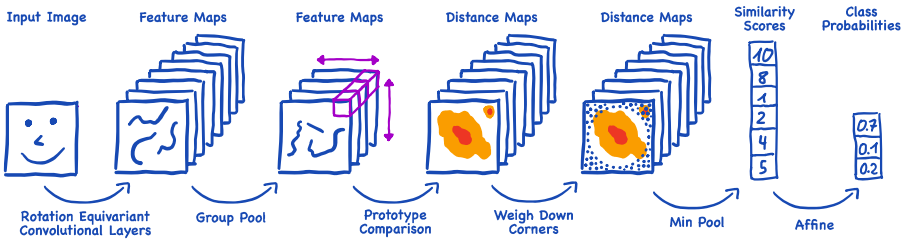

Fig. 3. Simplified architecture overview of ReProtoPNet with weighted down corners.

2.4 Datasets, Model and Implementation Details

MedMNIST [12] is a collection of biomedical image classification datasets comprising 18 different datasets. All datasets are standardized with multiple size options, including 28, 64, 128, and 224x224 images. The BloodMNIST dataset is used in all experiments with the size of 224x224 images unless stated otherwise. The regular ProtoPNet is trained with the augmented dataset, which

includes images rotated at 0, 90, 180, and 270 degrees. However, we do not use any augmentation for the ReProtoPNet.

We used the C8SteerableCNN from the escnn [2] example model, designed for 45-degree rotational equivariance by leveraging the cyclic group C_8 to enhance robustness in tasks requiring rotational invariance. The input image is treated as a scalar field corresponding to the trivial representation of C_8. The model consists of six convolutional blocks, each featuring equivariant convolutions that output feature maps transforming under the regular representation of C_8. After each block, batch normalization [7] and ReLU [8] activations are applied, with anti-aliased average pooling performed after certain blocks to downsample the feature maps. The final convolutional layers reduce the feature maps to 64 channels, after which group pooling aggregates information across all rotations, effectively removing the group dimension while retaining spatial information. The resulting features are passed through fully connected layers for classification, producing the model's output. This architecture ensures that the network's features are robust to 45-degree rotations.

For the original ProtoPNet, all Steerable-CNN blocks are simply replaced with their equivalent regular CNN blocks.

3 Results

In this section, we will present our key findings, which are organized according to the following primary research questions: How do the models perform in terms of accuracy? How stable is the new method compared to the original ProtoPNet? How does the Weight Down Corner (WDC) method affect stability?

For the two models, we pre-defined 10 prototypes per class like in the original ProtoPNet paper, but the accuracy results of various numbers of prototypes per class are also shown in Fig. S1.

3.1 Performance Evaluation

Accuracy is a simple yet effective method to show the overall performance of the models/methods for various tasks. In this work, we also shared the performance of the ProtoPNet and the ReProtoPNet across different datasets, as shown in Fig. 4. The baseline model shares the same architecture with the C8SteerableCNN model but is converted to the equivalent CNN model, as we mentioned previously. PP represents the ProtoPNet since it is an extension of the backbone model. In three datasets, ReProtoPNet (Steerable CNN + PP) does not compromise the accuracy much by gaining prediction stability across any rotation, which we mentioned in section 4.3.

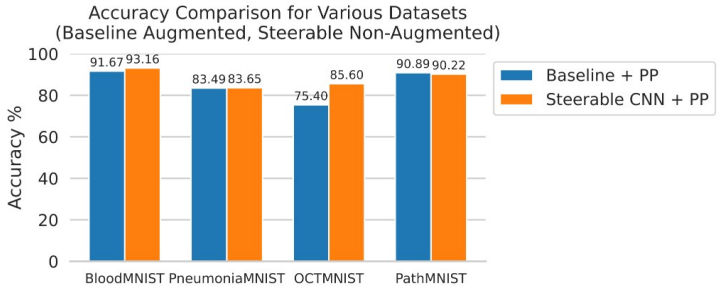

Fig. 4. Comparison of Accuracies across different MedMNIST datasets between ProtoPNet, and ReProtoPNet

3.2 Stability of Diagnosis With Respect to Rotation

It is essential to have stable and consistent results in machine learning, especially in medical applications. A doctor must not misdiagnose a patient due to the rotation of the input image, which is why they need explainable and interpretable models. To estimate how stable and reliable our models are, we performed various tests, two of which we present here.

Average Prediction per Class In the analysis presented in Fig. 5, we compared ProtoPNet, ReProtoPNet, and ReProtoPNet-WDC regarding their average probability across all samples for each class as the input images were rotated. As can be seen, ProtoPNet is inherently unstable, while ReProtoPNet and especially ReProtoPNet-WDC, by design, achieve much more consistent predictions.

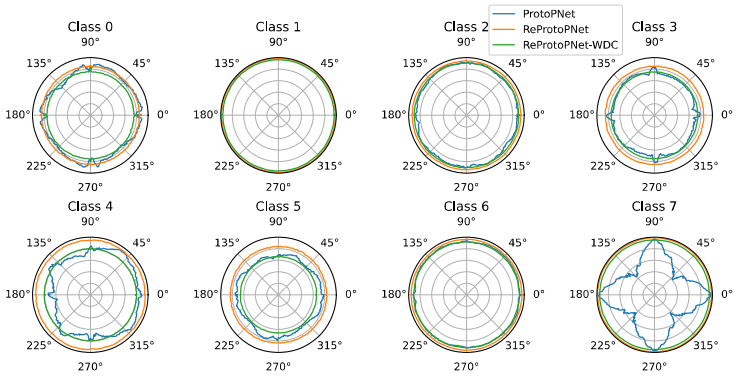

Fig. 5. Comparison of average probability over all samples for each ground truth class and rotation between ProtoPNet, ReProtoPNet, and ReProtoPNet-WDC.

Average Variance of Prototype Distances We compute the average variance of each prototype's distance the following way. The normalized prototype distance for prototype p, sample s, and rotation r is computed by $\frac{d_{spr}}{\bar{d}_{sp}+\epsilon}$, where $\bar{d}_{sp}$ is the mean distance over all rotations, and ϵ is a small value such that very small distances are not scaled too drastically. The average variance of prototype distances over all samples for each prototype is then given by Eq. 2

$$var_p = \frac{1}{S}\sum_{s=1}^{S} Var(\frac{d_{sp}}{\bar{d}_{sp}+\epsilon}) \tag{2}$$

By comparing these variances between ProtoPNet, ReProtoPNet, and ReProtoPNet-WDC, we can see that ReProtoPNet-WDC achieves roughly 500 times smaller variance in the prototype distances compared to the original ProtoPNet (see Fig. S2 and Fig. S3).

We also measured the average variance of class probabilities for each class for correctly classified samples (prediction is classified based on majority vote over all rotations). We computed this variance in a manner similar to the one described above. Our proposed architecture achieved a roughly 40 times smaller variance (Fig. S4).

4 Discussion

In the original ProtoPNet, rotation augmentation is used to generalize the model across different orientations. However, applying 16 rotations during training increases computational costs by 16 times, as noted by Veeling et al. [10]. Despite this, it does not ensure generalization during inference, as the model treats each rotated image as a distinct sample, resulting in different activation maps for each rotation, as shown in Fig. S5.

For the mask operation, the optimal size and smoothness (sigma) of the weight-down mask are unclear, which may affect stability. Using large filters in multiple convolutions can blur the latent space's border region, shifting rotation artifacts toward the center. For circular datasets, where key information is centered, a weight-down mask can help eliminate edge artifacts, as seen in BloodMNIST. However, while this approach shows promise, it may only be effective for specific datasets and could reduce interpretability for important features outside the circular mask.

5 Conclusion

To sum up, stability and explainability are significant parts of medical applications for robust diagnosis. In this study, we present rotation-invariant prototypes for interpretable models, specifically on ProtoPNet, to reduce prediction and reasoning variance caused by rotation. As a result, the proposed architecture achieved rotational invariance of diagnosis while preserving the explainability (97.5% reduced variance). It makes the variance of the prototype distances more than 500 times smaller (99.8% reduction).

References

1. Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., et al.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education **23**(1), 689 (2023)
2. Cesa, G., Lang, L., Weiler, M.: A program to build e (n)-equivariant steerable cnns. In: International conference on learning representations (2022)
3. Chaddad, A., Peng, J., Xu, J., Bouridane, A.: Survey of explainable ai techniques in healthcare. Sensors **23**(2), 634 (2023)
4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
5. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International conference on machine learning. pp. 2990–2999. PMLR (2016)
6. Cohen, T.S., Welling, M.: Steerable cnns. arXiv preprint arXiv:1612.08498 (2016)
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
8. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
9. Nazar, M., Alam, M.M., Yafi, E., Su'ud, M.M.: A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access **9**, 153316–153348 (2021)
10. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 210–218. Springer (2018)
11. Weiler, M., Cesa, G.: General e (2)-equivariant steerable cnns. Advances in neural information processing systems **32** (2019)
12. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)

6 Supplementary Materials

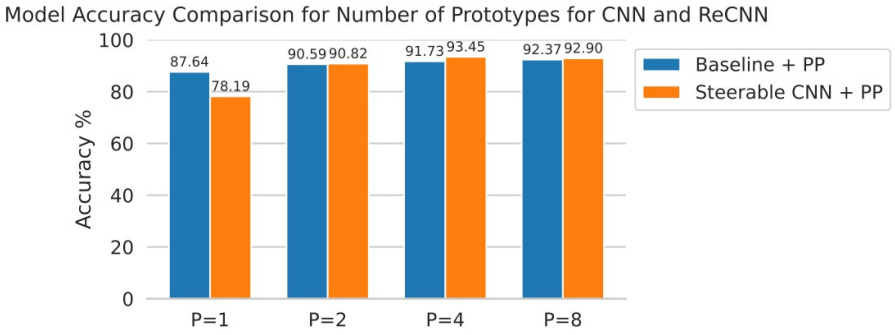

Fig. S1. Accuracy comparison of ProtoPNet (Baseline + PP) and ReProtoPNet (Steerable CNN + PP) for different number of prototypes per class. P stands for the number of Prototypes per class.

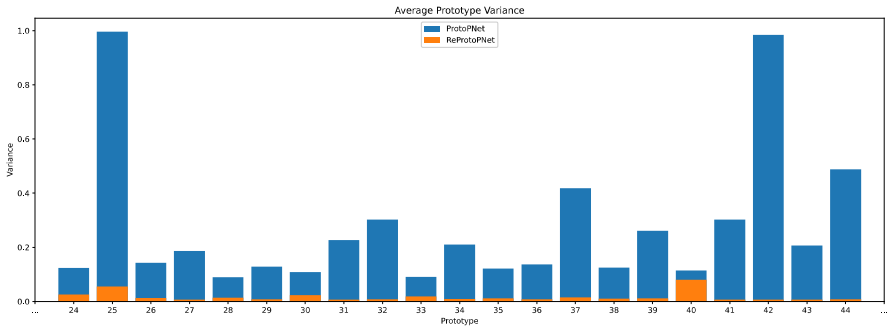

Fig. S2. Average variance of each prototype over all samples. Comparison between ProtoPNet and ReProtoPNet.

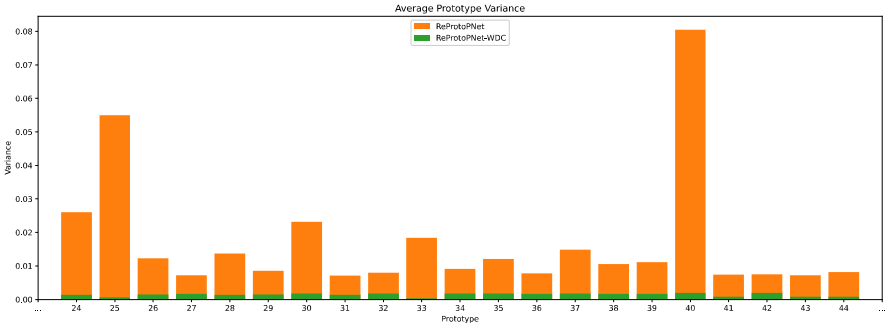

Fig. S3. Average variance of each prototype over all samples. Comparison between ReProtoPNet and ReProtoPNet-WDC.

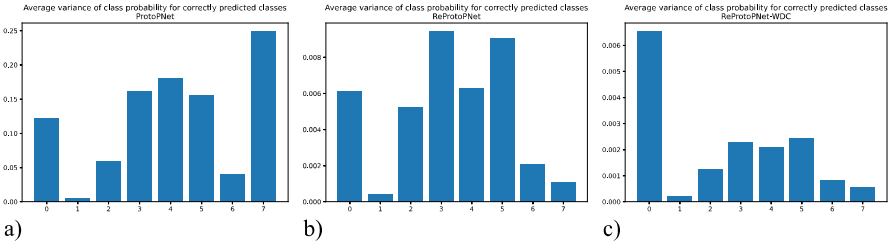

Fig. S4. Average variance of class probabilities for each class for correctly classified samples. Shown for ProtoPNet (a) ReProtoPNet (b) and ReProtoPNet-WDC (c).

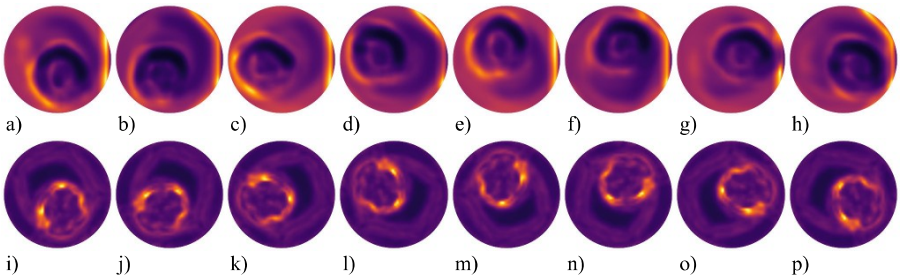

Fig. S5. Prototype distances of ProtoPNet (a – h) and ReProtoPNet (i – p) for various rotations of the input image (45° steps from 0° to 315°).