# Interpretable CNN Pruning for Preserving Scale-Covariant Features in Medical Imaging

Mara Graziani[1,3], Thomas Lompech[2], Henning Müller[1,3], Adrien Depeursinge[1,4], and Vincent Andrearczyk[1]

[1]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[2]INP-ENSEEIHT, Toulouse, France
[3]University of Geneva (UNIGE), Geneva, Switzerland
[4]Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

**Abstract.** Image scale carries crucial information in medical imaging, e.g. the size and spatial frequency of local structures, lesions, tumors and cell nuclei. With feature transfer being a common practice, scale-invariant features implicitly learned from pretraining on ImageNet tend to be preferred over scale-covariant features. The pruning strategy in this paper proposes a way to maintain scale covariance in the transferred features. Deep learning interpretability is used to analyze the layer-wise encoding of scale information for popular architectures such as InceptionV3 and ResNet50. Interestingly, the covariance of scale peaks at central layers and decreases close to softmax. Motivated by these results, our pruning strategy removes the layers where invariance to scale is learned. The pruning operation leads to marked improvements in the regression of both nuclei areas and magnification levels of histopathology images. These are relevant applications to enlarge the existing medical datasets with open-access images as those of PubMed Central. All experiments are performed on publicly available data and the code is shared on GitHub.

**Keywords:** interpretability · scale · histopathology · transfer

## 1 Introduction

Transfer learning has become a standard approach in tasks with a limited amount of training data [35]. In medical imaging, it has led to significant improvements in various applications in terms of accuracy and speed of convergence [21,26,17,25]. Scale invariance is required and learned implicitly by Convolutional Neural Networks (CNNs) in the object recognition task on ImageNet, as they normally appear at different distances from the observation point. Despite the controlled viewpoint and the considerable domain shift (i.e. reduced number of classes, less color, texture and object variety [25]), medical imaging applications often reuse basic features from pretraining on natural images, i.e. color, edges and textures [17,13].

The implicitly learned invariances could have different impacts in medical imaging. Global and local rotation invariance, for instance, were shown to be relevant [31,1]. The scale invariance, however, could be detrimental. The viewpoint in medical images is controlled and the pixel (or voxel) size has a corresponding physical dimension. The size of an object of interest within an image carries relevant information [7,6]. Nuclei size histopathology applications is a clear example of a discriminant factor of tumor regions [8,14]. Approaches introducing scale analysis in the sense of either scale covariant networks [33] or multi-scale learning [3,32,24,16] showed that analyzing tissue at various magnifications benefits from the combination of fine-grained details and global tissue information. Histopathology is not the only application that benefits from information about scale. Nodule detection and classification in computed tomography is another example in the medical domain [16]. From a larger perspective, other applications can be remote sensing, defect detection, material recognition and biometrics (e.g. iris recognition) [28]. It is thus relevant to analyze the role of information about scale in state-of-the-art CNNs that are often used for transfer learning such as inception-based [27] and residual-learning networks [15].

A key question is how to quantify the degree of scale invariance at each layer in the network. Taking inspiration from previous research in concept-based interpretability of CNNs [20,12], we define the layer-wise quantification of scale-covariance as an interpretability task. Image scale is seen as a concept that is learned during training. This is analyzed with Regression Concept Vectors (RCVs) [12]. RCVs extend previous research on binary-expressed concept interpretability (where the concept is either present or not present) [20,5] and were already used to analyze the effects of transfer in [13]. They are particularly suited for our task since they allow us to measure scale with continuous values obtained from the bounding box annotations in the publicly available PASCAL-VOC dataset. The degree of invariance at each layer is evaluated as a regression task of the scale measures. Besides, the layer-wise quantification of scale covariance is used to implement a pruning strategy that preserves the scale-covariance of the features. Differently from the scale-covariant designs that explicitly model the requirements of specific applications [19,23,33,10,30,4], this pruning can be applied to state-of-the-art CNNs. In this way, ImageNet pre-trained weights[1] can be used without the need of retraining from scratch. Being based on the interpretability analysis, the pruning has an interpretation that promotes its algorithmic transparency. It removes, in fact, the layers that introduce scale-invariance to the features. The experimental results on Estrogen Receptor-positive Breast Cancer (ERBCA+) images show a marked benefit in the magnification regression of open-access histopathology images [24]. This can help predicting the magnification range of images where the physical dimension of voxels is unknown, e.g. the large open-access biomedical data repository PubMed Central[2], to extend existing medical datasets.

---

[1] downloadable at https://keras.io/api/applications/

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

## 2    Methods

### 2.1    Notations

We consider an input image $X \in \mathbb{R}^{w \times h}$, where $w$ is the image width and $h$ is the height. The function $\phi(\cdot)$, defined as $\phi : \mathbb{R}^{h \times w} \to \mathbb{R}^d$ maps the input image to a vector of arbitrary dimension $d$. For instance, it transforms X into a collection of $d$ scalars obtained from averaged feature maps at a given intermediate layer. At the final fully-connected layer, $\phi(\cdot)$ transforms X into a set of predictions. When analyzing scale information, we are interested in covariance[3], thus whether we can find a transformation $g' : \mathbb{R}^d \to \mathbb{R}^d$ that predicts the transformation $g : \mathbb{R}^{h \times w} \to \mathbb{R}^{h \times w}$ of the input image X in the feature space obtained by $\phi(g(X))$. The scaling transformations are expressed as $g_\sigma(\cdot)$, being parameterized by a scale factor $\sigma$.

### 2.2    Representation of Scale Information

Our interest is in finding a linear transformation $g'_\sigma(\cdot)$ that is a predictable transformation of the scaling operation $g_\sigma(\cdot)$. To this end, a regression vector $\mathbf{v}$ can be searched in the feature space to predict the scaling factor $\sigma$ as[4]:

$$\sigma = \sum_i v_i \phi_i(g_\sigma(X)) = \mathbf{v} \cdot \phi(g_\sigma(X)). \tag{1}$$

Therefore, $g'_\sigma(\cdot)$ can be represented as a translation matrix (in $\mathbb{R}^d$) by $\sigma$ along $\mathbf{v}$, so that $g'_\sigma(\phi(X)) = \phi(X) + \mathbf{v} \cdot \sigma$.

### 2.3    Bounding-Box Size vs. Image Size

This section clarifies our definition and measurement of the image scale. Indications of scale are commonly used to relate the dimensions of two objects. In design modeling and cartography, the scale is the ratio comparing the length of the represented segment to the one in the real world (i.e. 1 cm:1000 Km). Computer vision and image processing mostly refer to the act of scaling, namely the transformation that generates a new image with a larger or smaller number of pixels. If the input size is changed with the scaling, however, the transformation causes the "train-test" resolution discrepancy in [29] during network inference. For this reason, it is recommendable to fix the input size to the default model input size $S_i = 299 \times 299$ when measuring scale information as shown in [22]. By focusing only on ImageNet-like images that only contain a single object, image

---

[3] Following the same terminology, the equivariance, as opposed to covariance, implies that the function $\phi(\cdot)$ maps an input image to a point in the same domain, i.e. $\phi : \mathbb{R}^{h \times w} \to \mathbb{R}^{h \times w}$.

[4] For simplicity, we omit the intercept. In Eq. (1), the intercept would be $v_0$ with $\phi_0(g_\sigma(X)) = 1$

scale can be pragmatically defined as the solid angle of the object in the image, namely the proportion of the field of view occupied by an object [34]. More directly, we measure the bounding-box area $S_b$ occupied by the object in the image. The image has area $S_o = h_o \times w_o$, where $h_o$ and $w_o$ are respectively the original image width and height. A small bounding box corresponds to a smaller space in the field of view of the camera, and thus a smaller solid angle. Scale measures are thus defined as the ratio $r = \frac{S_b}{S_o} = \frac{h_b \times w_b}{h_o \times w_o}$, where $h_b$ and $w_b$ are the bounding box height and width. Fig. 1 shows an example of scale measures on input images from the same class appearing at different scales.

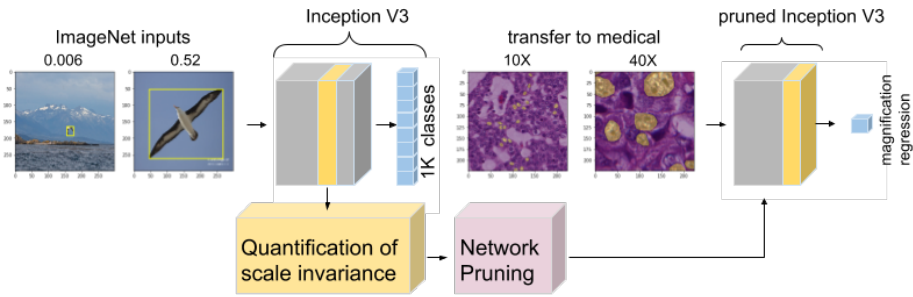## 2.4 Network Architectures and Tools



Fig. 1: Scale quantification and network pruning for better transfer in the medical domain. The bounding boxes for the ImageNet class *albatross* and the segmentation masks for the ERBCa+ inputs are overlaid in yellow on the images. The bounding box ratios $r$ are reported on top of the inputs. Images are shown at magnifications 10X and 40X. The layer evidenced in yellow encodes the most of information about scale. The pruned network drops the layers after this for solving the medical task. Best seen on screen.

*ImageNet-weights Initialization* InceptionV3 [27] and ResNet50 [15] are used for the analysis with pretrained ImageNet weights.

*Regression of Scale* The regression of the scale of multiple objects of the same class that appear naturally at various scales is sought to approximate $g'_\sigma(\cdot)$ as in Eq. 1. This corresponds to computing the RCV representing "scale" [22]. The regression is sought at several layers in the network to compare different depths. Aggregation is performed on the feature maps in the form of Global Average Pooling (GAP) as in [11] to obtain the feature vector $\phi(X)$ (except for the prediction layer which is already pooled). The determination coefficient $R^2$ is used to evaluate the prediction of the scale ratio $r$ on unseen test data

of the same class[5]. This evaluation is informative about the scale-covariance of the features. The $R^2$ is a measure between zero and one when the regression is evaluated on the training data. The $R^2$, however, could take negative values when evaluated on the test data (test $R^2$). Differently from what one may think, this is not due to a bad choice in the evaluation technique but it rather shows that the prediction on the test samples is far worse than predicting their mean. To address this issue, a normalization of the test $R^2$ is performed by evaluating $\frac{e^{R^2}}{e}$. In this way, the performance of the RCV on test data is kept in a [0,1] range, with values below $\frac{1}{e}$ evidencing bad performance.

*Pruning strategy* Network pruning is performed by comparing the test $R^2$ to identify the layer where the scale covariance is the highest. This evaluation is averaged across different object categories to remove the dependence on the class of the inputs. The layer with the highest test $R^2$ (the yellow layer in Fig. 1) is where the scale covariance is the highest. Layers deeper than this one are pruned off the architecture and a Global Average Pooling operation (GAP) is added to obtain a vector of the aggregated features.

*Transfer and network pruning* Transfer is performed from both the original and pruned architectures. To predict the average nuclei area, a single-unit dense layer is trained with the mean squared error loss between the true areas and the predicted ones. The nuclei area is expressed for each image as the average number of pixels within the segmentation of the nuclei. The regression is evaluated by the Mean Average Error (MAE). The magnification category (i.e. 5X, 8X, 10X, 15X, 20X, 30X, 40X) is also obtained from the average nuclei areas. The predicted areas are mapped to the magnification category that has the closest mean average value of the nuclei areas in the training set. This approach outperformed the direct classification of the magnification in [24]. Cohen's kappa coefficient is used to measure the inter-rater reliability of the magnification prediction. The networks are implemented in Keras and trained for five epochs with standard hyperparameters ($lr = 1e - 4$). The full pipeline is reported in Fig. 1 and the source code is available on github for reproducibility[6].

### 2.5   Datasets

The experiments in this paper involve two different datasets since the scale analysis is performed on inputs of natural images and the proposed final architecture is evaluated on a medical imaging task. For the scale quantification part, images with manual annotations of bounding boxes are selected from the publicly available PASCAL-VOC dataset [9]. We restrict our analysis to three object categories and to images containing a single bounding box, chosen among

---

[5] We compute $R^2 = \frac{\sum_{i=1}^{N}(\hat{r}_i - \bar{r})}{\sum_{i=1}^{N} r_i - \bar{r}}$, were $N$ is the number of test data samples, $\hat{r}$ is the ratio predicted by the regression model, $\bar{r}$ is the mean of the true ratios $\{r_i\}_{i=1}^{N}$.

[6] `https://bit.ly/2N6teMA`

the available annotated classes. These are *albatross* (ID: n02058221, 441 images), *kite* (ID: n01608432, 406 images) and *racing car* (ID: n04037443, 365 images).

The data for the histopathology application consist of 141 whole slide images of $2K \times 2K$ pixels taken at a maximum magnification of 40x ERBCa+ images. In these, $12,000$ nuclei boundaries were manually annotated [18]. Patch sampling was performed at 5, 8, 10, 15, 20, 30, and 40x magnification. A total of $69,019$ patches with nuclei segmentation masks were split into training, validation and test partitions (approximately 60%, 20%, 20% respectively) as shown in Table 1. The imbalance in the different magnification categories is due to the area covered by each magnification level, with the least number of patches being extracted at 5x and 8x. The average nuclei area is extracted for each input image by computing the average number of pixels in the relative nuclei segmentation mask. Example images with overlaid segmentation masks are displayed in Fig. 1.

| Split/# patches | 5X | 8X | 10X | 15X | 20X | 30X | 40X | Total |
|---|---|---|---|---|---|---|---|---|
| Train | 94 | 2,174 | 4,141 | 7,293 | 9,002 | 10,736 | 11,638 | 45,078 |
| Validation | 8 | 588 | 1,197 | 2,132 | 2,604 | 3,504 | 3,150 | 12,733 |
| Test | 36 | 428 | 900 | 1,728 | 2,198 | 2,802 | 3,166 | 11,208 |
| *Total* | 138 | 3,190 | 6,238 | 11,153 | 13,804 | 16,592 | 17,904 | 69,019 |

Table 1: Number of ERBCa+ patches extracted per magnification and partition.

## 3    Experiments and Results

### 3.1    Layer-wise Quantification of Scale invariance

The layerwise analysis of scale representation in InceptionV3 and ResNet50 is shown in Fig. 2[7]. The object categories *racing-car*, *albatross* and *kite* are used for the analysis. For each class, 70% of the available images are used for learning the regression and the rest for the evaluation of the $R^2$. The evaluation was performed for ten splits of images. To remove the dependency of the evaluation on the image selection (by multiple split) and category (by analyzing multiple classes) we average the 10 repetitions for all classes (a total of 30 evaluations). The regression of scale in a randomly initialized network (orange line) is compared to a pretrained model (blue line) in Fig. 2. An additional baseline (green line) shows the performance of regressing random scale measures, i.e. the scale ratios were shuffled to break the true image-label correspondence. Values of $R^2$ close to one reflect the linear covariance of the intermediate layers to object scale as defined in Section 2.1. Individual results for each class were discussed in [22], while the generalization on different test classes is further analyzed in [2].

---

[7] Layer names refer to the Keras implementation names.

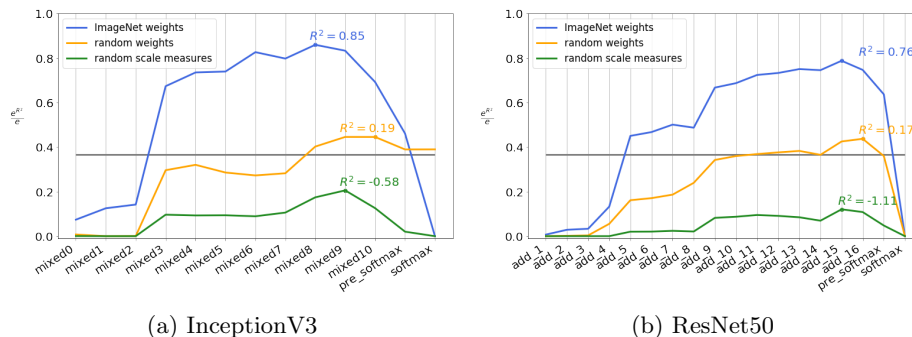(a) InceptionV3                                    (b) ResNet50

Fig. 2: Regression of scale measures on test data (performance of the RCV) at different layers.

### 3.2   Improvement of Transfer

The performance of the networks predicting average nuclei areas is compared between the original and pruned versions in Table 2. We report the MAE of ten repetitions[8] and the relative standard deviation. In the same table, we also report the kappa for the prediction of the magnification category.

Table 2: Mean Average Error (MAE) of the nuclei area regression (in pixels) and Cohen's kappa coefficient between the true and predicted magnification categories. Results are averaged over ten repetitions, the standard deviation is reported in brackets.

| model | layer | MAE (std) | kappa (std) |
|---|---|---|---|
| InceptionV3 | mixed10 | 81.85 (11.08) | 0.435 (0.02) |
| pruned InceptionV3 | mixed8 | **54.93** (4.32) | **0.571** (0.05) |
| ResNet50 | add16 | 70.08 (12.49) | 0.610 (0.03) |
| pruned ResNet50 | add15 | **54.76** (3.10) | **0.623** (0.04) |

## 4   Discussion

The experiments were designed for analyzing the presence of scale-informative features in state-of-the-art CNNs pretrained on ImageNet. Our results in Sec. 3.1, particularly highlight the linear covariance of medium-deep layers, with invariance being learned before the classification layer. The scale of unseen objects (test data) is regressed with the highest determination coefficient $R^2 = 0.85$ in InceptionV3 (blue line in Fig.2a), independently from the object class being

---

[8] Different seeds were used to initialize the dense connections to the last dense layer.

tested. In comparison, the RCV learned from random scale ratios cannot predict test data. This sanity check provides a lower bound $R^2 = -0.58$. For randomly initialized weights (orange line in Fig. 2a), the performance is around zero in almost all layers. This suggests that architectures with random weights do not contain linearly covariant representations, that hence must be learned during training. Similar observations apply to the ResNet50 models in Fig. 2b.

One important comment is about the low $R^2$ values at early layers of both architectures. We attribute this result to the limited size of the effective receptive field, that being at early layers only contains information from a very small fraction of the image. This affects the prediction of the scale ratios. The drop in the layers before the class prediction, namely in *mixed10* (for InceptionV3), *add16* (for ResNet50) and in *pre_softmax* (for both architectures), shows that deep network features learn scale invariance to classify image categories. Invariance to scale is thus achieved gradually in these layers preceding the last layer.

The quantification of scale invariance is applied to the image magnification regression as described in Sec. 3.2. The pruning strategy drops the layers with scale-invariant features. For InceptionV3, the pruned features are a result of a GAP on top of the *mixed8* features. As shown in Table 2, the MAE=54.93 of the nuclei area regression in *mixed8* is markedly lower than the MAE=81.85 in *mixed10*. This corresponds to a better prediction of the magnification range, hence to a higher kappa coefficient.

## 5   Conclusions and Future work

This paper proposed the analysis of scale covariance in state-of-the-art CNNs pretrained on ImageNet and a pruning strategy to mantain such covariance for better transfer. Feature extraction and finetuning are very diffused techniques, and the pruned features can lead to improved performances on imaging tasks where scale carries crucial information, as for example the medical task of nuclei area regression and scale magnification prediction shown in our application. This work shows, in addition, that research in deep learning interpretability can be actively used to improve model development. Other transformations could also be analyzed, e.g. rotation, to improve the feature extraction process without the need for explicit equivariant designs. Such analysis could be relevant not only in other medical imaging tasks, but also in remote sensing, defect detection, material analysis and biometrics.

A limitation of this work is that the regression only captures linear correlations in the data, whereas nonlinear relationships could be necessary to model other transformations. In future work, we will investigate non-linear regression and manifold learning of the feature space to formally address this point.

# References

1. Andrearczyk, V., Fageot, J., Oreiller, V., Montet, X., Depeursinge, A.: Exploring local rotation invariance in 3D CNNs with steerable filters. In: International Conference on Medical Imaging with Deep Learning (2019)
2. Andrearczyk, V., Graziani, M., Müller, H., Depeursinge, A.: Consistency of scale equivariance in internal representations of CNNs. In: Irish Machine Vision and Image Processing (2020)
3. Bejnordi, B.E., Zuidhof, G., Balkenhol, M., Hermsen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.: Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. Journal of Medical Imaging **4**(4), 044504 (2017)
4. Bruna, J., Mallat, S.: Invariant scattering convolution networks. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1872–1886 (2013)
5. Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., et al.: Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2019)
6. Depeursinge, A., Foncubierta-Rodriguez, A., Van De Ville, D., Müller, H.: Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities. Medical Image Analysis **18**(1), 176–196 (2014)
7. Depeursinge, A.: Multi-Scale and Multi-Directional Biomedical Texture Analysis: Finding the Needle in the Haystack. In: Biomedical Texture Analysis: Fundamentals, Applications and Tools, pp. 29–53. Elsevier-MICCAI Society Book series, Elsevier (2017)
8. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **19**(5), 403–410 (1991)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
10. Ghosh, R., Gupta, A.K.: Scale steerable filters for locally scale-invariant convolutional neural networks. In: Workshop on Theoretical Physics for Deep Learning at the International Conference on Machine Learning (2019)
11. Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H.: Concept attribution: Explaining CNN decisions to physicians. Computers in Biology and Medicine p. 103865 (2020)
12. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 124–132. Springer (2018)
13. Graziani, M., Andrearczyk, V., Müller, H.: Visualizing and interpreting feature reuse of pretrained CNNs for histopathology. In: Irish Machine Vision and Image Processing (IMVIP) (2019)
14. Graziani, M., Müller, H., Andrearczyk, V.: Interpreting intentionally flawed models with linear probes. In: SDL-CV workshop at the IEEE International International Conference on Computer Vision (2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

16. Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q.: Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition **83**, 134–149 (2018)
17. Huh, M., Agrawal, P., Efros, A.A.: What makes ImageNet good for transfer learning? In: Workshop on Large Scale Computer Vision Systems at NeurIPS 2016 (2016)
18. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics **7** (2016)
19. Kanazawa, A., Sharma, A., Jacobs, D.W.: Locally scale-invariant convolutional neural networks. In: Advances in Neural Information Processing Systems (2014)
20. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning. pp. 2673–2682 (2018)
21. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
22. Lompech, T., Graziani, M., Otálora, S., Depeursinge, A., Andrearczyk, V.: On the scale invariance in state of the art CNNs trained on ImageNet. In: (submitted) (2020)
23. Marcos, D., Kellenberger, B., Lobry, S., Tuia, D.: Scale equivariance in CNNs with vector fields. In: FAIM workshop at the International Conference on Machine Learning (2018)
24. Otálora, S., Atzori, M., Andrearczyk, V., Müller, H.: Image magnification regression using densenet for exploiting histopathology open access content. In: Computational pathology and ophthalmic medical image analysis, pp. 148–155. Springer (2018)
25. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning with applications to medical imaging. In: arXiv preprint arXiv:1902.07208 (2019)
26. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on Medical Imaging **35**(5), 1285–1298 (2016)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
28. Szeliski, R.: Computer vision: Algorithms and applications. Springer Science & Business Media (2010)
29. Touvron, H., Vedaldi, A., Douze, M., Jégou, H.: Fixing the train-test resolution discrepancy. In: Advances in Neural Information Processing Systems (2019)
30. Van Noord, N., Postma, E.: Learning scale-variant and scale-invariant features for deep image classification. Pattern Recognition **61**, 583–592 (2017)
31. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. pp. 210–218. Springer (2018)
32. Wan, T., Cao, J., Chen, J., Qin, Z.: Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. Neurocomputing **229**, 34–44 (2017)

33. Worrall, D.E., Welling, M.: Deep scale-spaces: Equivariance over scale. In: arXiv preprint arXiv:1905.11697 (2019)
34. Yan, E., Huan, Y.: Do CNNs Encode Data Augmentations? arxiv.org/2003.08773 (2020)
35. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. pp. 3320–3328 (2014)