Learning Invariant Representations of Images for Computational Pathology

Learning Invariant Representations of Images for Computational Pathology

This thesis was typeset by the author using LaTeX.
Cover design and layout by Maxime Lafarge.

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-5216-0

Learning Invariant Representations of Images for Computational Pathology

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op maandag 15 februari 2021 om 16:00 uur

door

Maxime Wallace Lafarge

geboren te Neuilly-sur-Seine, Frankrijk

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. M. Merkx
promotor:	prof.dr. J.P.W. Pluim
copromotor:	dr. M. Veta
leden:	prof.dr. P.J. van Diest (UMC Utrecht)
	prof.dr. A. Martel (University of Toronto)
	prof.dr. A. Vilanova
	dr. J. van der Laak (Radboud UMC)
	dr.ir. R. Duits

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

Chapter 1
Introduction

1.1 Clinical Background and Computational Pathology

At the microscopic scale, the structure of tissue specimens constitutes a primary source of information about the state of a disease. By observing tissue samples under a microscope, pathologists quantify disease-associated structures that support clinical decision making. In the case of cancerous tissues, this assessment is essential to determine the prognosis of patients. To measure the appearance abnormality of cancerous tissues, pathologists rely on well-established grading systems that define cancer-specific tissue characteristics to inspect [1]–[3]. For example, the Gleason Grading System [1] categorizes patterns in prostate cancer tissue samples based on the formation of glands and infiltrative patterns; the Nottingham Grading System [2] grades breast cancer tissue samples based on the assessment of tubule formation, the degree of nuclear pleomorphism and the count of mitotic figures.

The on-going development of whole-slide image scanners has enabled the digitization of glass slides of tissue specimens as an alternative to the use of conventional brightfield microscopes. In addition to transferring the examination and annotation procedures of tissue sections to a computer, the digitization of pathology labs equipped with this technology has induced a demand for automated image analysis systems (Figure 1.1) that can assist pathologists in their routine workflow [4], [5].

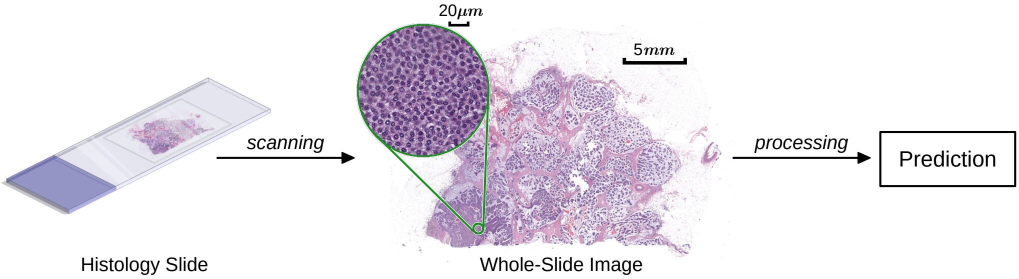


Figure 1.1: Illustration of the digital pathology acquisition and processing pipeline. A physical histology slide is digitized into a whole-slide image that is then processed by an algorithm to make a prediction or quantify a biomarker of interest.

Such systems present significant potential as they can perform tasks in a faster, systematic and reproducible way in comparison to human experts. With modern computational resources, it has become possible to process gigapixel whole-slide images (WSIs) at high resolution in a few minutes (whereas inspecting the same amount of data would be impracticable for a human). Yet, the most promising benefit of such automated systems is the reliable quantification of biomarkers, that can support pathologists to more accurately assess the prognosis of cancer patients, improving inter-observer agreement, reducing cancer interpretive errors and ultimately leading to better patient care. Beyond that, such computational power also enables prioritizing cases and refocusing the workload of pathologists on other tasks that require their expertise.

From a research viewpoint, this technology has fostered the creation and release of large datasets of whole-slide images, joined with molecular and clinical data. The field of computational pathology emerged from the access to these datasets, as they provide researchers with the material to develop and validate computational models and algorithms that can extract, quantify and discover disease-associated patterns [6], [7] and answer the demand for reliable automated image analysis systems. To leverage this large amount of image data, the computational pathology community has established machine learning and more specifically convolutional neural networks (CNNs) as the prominent methodology to achieve state-of-the-art performances across many histopathology image classification tasks [8]. By modeling associations between complex visual patterns in WSIs and corresponding target quantities (labels) that exist in the training data, CNNs have the potential to generalize and make accurate predictions of these quantities on new unseen images encountered in a clinical context.

1.2 Machine Learning for Computational Pathology

The key principle behind the success of this methodology is that CNN-based models have the ability to learn complex informative high-level features that are predictive of target quantities of interest, directly from data. As this feature representation is learned from training examples, the subsequent system does not have to rely on a restricted set of hand-crafted features that requires domain expertise to be defined, and whose implementation has typically a higher computational cost than the forward-pass of a CNN. Given sufficient training data and capacity, the universal approximation property of neural networks implies that hand-crafted features can be in principle reproduced by CNNs, and even surpassed if they are sub-optimal with respect to solving a given task. This explains why deep learning models have outperformed models based on feature engineering: we refer to [9] for an overview of recent machine learning advances in the field of medical image analysis, and to [8], [10]–[13] for an overview of machine learning techniques dedicated to histology image analysis.

As this data-driven paradigm has moved the manual feature extraction process of classical machine learning to the design of model architectures and training protocols, this thesis presents developments that concern these design decisions. Depending on the assumptions that can be made to train the models, the type of labels available, and the desired level of interpretability, we distinguish three families of CNN-based frameworks for WSI-based predictive/quantitative tasks (Figure 1.2).

1. Patch-level classification: a CNN is trained to classify known tissue biomarkers and patterns given a training set of image patch examples extracted from WSIs that were labeled by expert annotators. Here, CNNs make patch-level predictions (detection and classification of objects) that can then be used to derive slide-level quantities that have a clinical value (Figure 1.2(a)). Examples of such CNN-based applications include mitosis detection [14], nuclei segmentation [15], nuclei classification [16], tubule detection [17] or lesion detection [18].
2. Multiple-instance learning: in the scenario in which patch-level labels are not available, a CNN is trained to learn a feature representation of image patch instances extracted from WSIs (here WSIs are viewed as bags of instances). This feature representation is then aggregated from all the instances of a given WSI to produce a slide-level representation. This aggregated representation is then input to a classifier that is trained to predict slide-level target quantities (Figure 1.2(b)).
3. End-to-end predictions: a CNN-based model takes a full WSI as input and is directly trained to predict a slide-level target (Figure 1.2(c)). This approach relies on methods that enable overcoming memory and computational limitations related to processing gigapixel WSIs in a end-to-end fashion [19], [20].

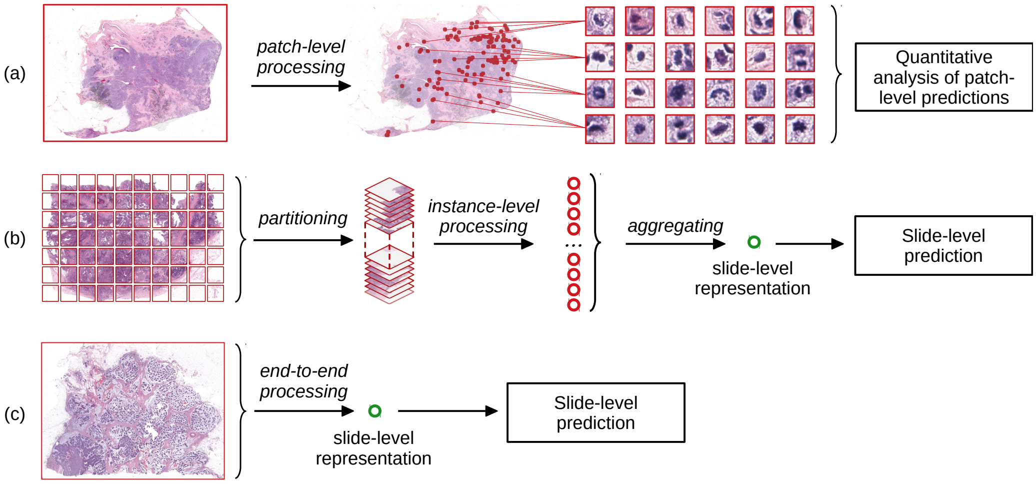

(a)
patch-level processing
Quantitative analysis of patch-level predictions
(b)
partitioning
instance-level processing
aggregating
slide-level representation
Slide-level prediction
(c)
end-to-end processing
slide-level representation
Slide-level prediction

Figure 1.2: Illustration of the main families of computational frameworks to process WSIs. (a) Patch-level classification (here, example of mitosis detection). (b) Multiple instance learning. (c) End-to-end WSI processing.

1.3 Invariant Representation Learning

The common point across this spectrum of WSI processing frameworks is that all use CNNs to learn abstract feature representations of the input histology images to compute an output. Although these representations are aimed to capture phenotypical features that are informative for the task at hand, in practice, this representation is highly sensitive to irrelevant factors of appearance that are present in the images. As these irrelevant factors are potentially captured in the learned representation, the output of the trained models is subject to vary in an unpredictable manner when these (confounding) factors change, making the developed systems unreliable to some extent.

Histology images (and more generally bioimages) are known to exhibit a high variability of appearance that is caused by factors that are independent of the inherent morphological characteristics of the imaged tissues [21], [22]. These inevitable factors of variation can thus be treated as irrelevant with respect to solving most tasks of interest. In the case of WSIs, we consider four categories of sources of variations that affect their appearance (Figure 1.3):

1. Histology slide preparation: as the preparation protocol varies from a pathology lab to another, this causes appearance variations associated to the thickness of the specimens, staining inconsistencies (concentration or fading), and any other slide-specific artifacts (tissue distortions, improper fixation or embedding).
2. The positioning of the tissue specimen on the slide is subject to variations, causing uninformative roto-translations of the acquired image.
3. Scanner characteristics (optical system, acquisition device/algorithm) and scanning parameters (illumination, camera focus) are another independent source of variation.
4. Other residual variations that are independent of the other categories.

In this thesis we consider histology images as samples of an unknown generative process that involves multiple latent random variables (Figure 1.3 (a)). With this probabilistic modeling perspective, solving a predictive task can be seen as an inference process in which we want to recover the variables that are informative of the morphology of tissues, that have a predictive power (Figure 1.3 (b)). Therefore, we address the inference problem of representing images by explanatory latent variables that are invariant to irrelevant factors of variations.

In the machine learning literature, the concept of learning invariant representations is covered in several topics including domain adaptation/generalization [23], representation disentanglement [24]–[26] and fair classification [27], [28]. These topics are of significant interest for computational pathology [7], [10], and thus motivates the development of the frameworks presented in this thesis.

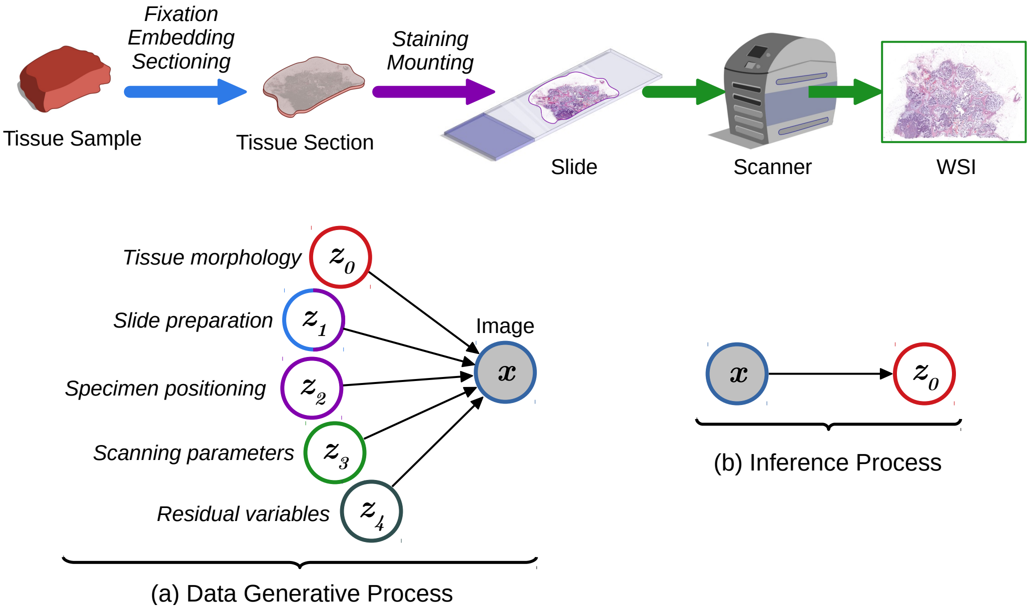


Figure 1.3: Illustration of the pipeline generating WSIs from tissue specimens. (a) Variations in the slide preparation steps and in the scanning procedures constitute sources of irrelevant variations in the image generative process. (b) In computational pathology, we are interested in inferring the morphological information of the imaged tissues.

1.4 Outline

Following this line of research, this thesis describes new frameworks to constrain deep learning models and provide them with invariance properties that improve their robustness to the irrelevant variability of histology data. Comparative analyses of the developed methods across multiple datasets and classification tasks are presented. The chapters are ordered in two parts. First, the Chapters 2, 3 focus on invariant representations in the context of supervised learning (when pre-assigned labels of a set of image examples are used to train the models) applied to patch-level classification tasks. Then, chapters 4, 5, 6 concern learning representations from unlabeled data which is a promising research direction with significant potential for computational pathology applications.

In Chapter 2, we investigate methods to make CNN-based models robust to slide-specific variations of appearance. Based on the assumption that all image patches extracted from a given WSI are drawn from a unique data distribution, we propose to consider every WSI of the training data as a domain, and address the problem of multi-domain generalization. Domain-adversarial training is investigated as an alternative solution to conventional standardization and augmentation approaches. A comparative analysis on two tasks is presented (inter-lab generalization of a mitosis classifier and multi-organ generalization of a nuclei segmentation system).

In Chapter 3, we address the robustness of CNN-based models to the arbitrary global orientation of tissue specimens. We propose a framework to encode the geometric structure of the special Euclidean motion group *SE(2)* in CNNs to yield translation and rotation equivariance via the introduction of *SE(2)*-group convolution layers. As a result, the robustness of the output of models equipped with this operation to rotations of the input is guaranteed by construction. Relative improvement of performances is shown on three different tasks (mitosis detection, nuclei segmentation and tumor detection) in comparison to baseline models trained with rotation augmentation.

In Chapter 4, we present a qualitative proof-of-concept study related to tri-dimensional rotational invariance. We describe an adversarial-driven method to infer a realistic tri-dimensional volume of stain concentrations that would produce realistic images under simulations of light transmission. This study gives insights into the possibility for deep learning models to represent the underlying tri-dimensional structure of imaged tissue slices from single two-dimensional views. This opens research directions towards granting additional rotational invariance into a learned representation.

In Chapter 5, we investigate an adversarial-driven extension of the variational autoencoder (VAE) framework to learn a useful latent representation of bioimages from unlabeled data. We show it is possible to train a latent variable model that is competitive with other popular models in a downstream classification task. As opposed to existing models, this method enables direct synthesis and reconstruction of realistic images from the latent variables, providing a tool for researchers to gain better insight into structure variations.

In Chapter 6, we further extend the VAE framework by leveraging the group structure of rotation-equivariant CNNs to learn orientation-wise disentangled generative factors of histology images. As a result, this learned representation can be directly used in downstream classification tasks (nuclear pleomorphism grading, mitotic activity assessment, cell type classification) and is competitive with respect to baseline VAEs.

Chapter 2
Learning domain-invariant representations of histological images

This chapter is based on:

M. W Lafarge, J. P. Pluim, K. A. Eppenhof, and M. Veta, "Learning domain-invariant representations of histological images", *Frontiers in Medicine*, vol. 6, p. 162, 2019.

M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images", *in Proceedings of the International Workshop on Deep Learning in Medical Image Analysis in Conjunction with MICCAI*, 2017, pp. 83–91.

2.1 Introduction

The traditional microscopy-based workflow of pathology labs is undergoing a rapid transformation since the introduction of whole-slide scanning. This new technology allows viewing of digitized histological slides on computer monitors and integration of advanced image analysis algorithms, which can enable pathologists to perform more accurate and objective analysis of tissue.

The process of producing a digital slide consists of several successive procedures: formalin fixation and paraffin embedding of the tissue, sectioning, staining and scanning. Each procedure has a multitude of parameters that vary between pathology labs and within the same lab over time. This results in significant tissue appearance variation in the digital slides, that adds to the underlying biological variability that can occur, for example, due to differences in tissue type or pathology.

In a real-world scenario, histological images are made available in pair with ground-truth annotations for the development of a predictive model to solve and automate a given task. Very often, these images were acquired in specific conditions (via the same scanner, following a lab-specific preparation process or from a small cohort for example) resulting in a narrower range of appearances than what could be observed in other conditions (different scanner, lab or cohort).

The discrepancy between the restricted data distribution available at training time and the higher variability of possible histological images on which a model is expected to perform, often limits the generalization of image analysis techniques, including deep learning-based methods.

This problem is typically addressed with ad-hoc methods based on known priors. For instance, one might correct for the known staining variability via a staining normalization approach. However, relying on such specifically chosen priors raises the risk to leave out or enhance domain-specific noise in the learned representation. For example, staining normalization methods will not handle other sources of variability such as specific tissue pleomorphism.

Deep learning methods learn abstract representations directly from the image data and have achieved state-of-the-art results in many computer vision and medical image analysis tasks including histopathology. Every histological slide results from a given set of latent parameters (corresponding to a specific case, hospital or tissue type for instance) and thus can be considered an individual domain. As such, all the image patches extracted from a given whole slide image (WSI) are samples of the same data distribution, and so, the same domain. We hypothesize that learning a representation that is explicitly invariant to the domains of the training data is likely to be also invariant, to some extent, to new unseen domains. This hypothesis is motivated by the fact that regular Convolutional Neural Networks (CNNs) preserve domain information in their representation that is not useful for the task at hand. This phenomenon is illus-

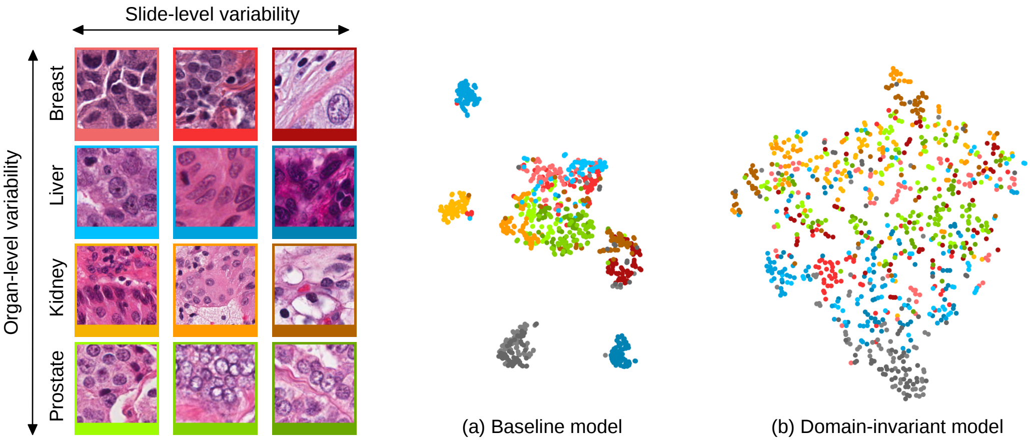


Figure 2.1: Illustration of the domain distribution of the internal representation of convolutional neural networks (CNN) trained for the task of nuclei segmentation. The scatter plots are t-SNE embeddings [29] of a random selection of 64 image patches for each of 12 digital slides, for which a representative patch is displayed on the left and framed with matching colors. Each image patch is represented by the concatenated means and standard deviations of its activations after the second convolutional layer of the CNNs. Two models are compared: (a) shows the representation learned by a baseline CNN model and (b) a model that uses stain-normalized inputs and domain-adversarial training. The models were trained with image patches from these 12 slides, and image patches from two hold-out slides (colon tissue type) were embedded the same way and are shown in gray. The baseline model induces domain clusters in the embedding space whereas the domain-invariant model produces a smoother distribution of the domains.

trated in Figure 2.1 (a): the appearance features present in some digital slides form separated clusters in the space of the learned representation, even if the slides share a known variability factor (patches from different liver tissue images, in blue, are distributed apart when represented by a baseline model). In the example of Figure 2.1 (a), image patches that originate from an unseen domain (colon tissue represented in gray), form a disjoint cluster, in a region that the model was not trained to process, and that is likely to lead to poor performances. However, this internal distribution can become smoother when strategies are employed to make the representation domain-invariant. The distribution of the embeddings shown in Figure 2.1 (b), illustrates how the representation of seen domains that was disjoint among the same organ now overlaps, and how unseen domains align with this smooth distribution: the gray cluster representing an unseen organ tissue type is now connected to the rest of the embeddings, and is more likely to lead to better generalization performances.

In this paper, we propose a domain-adversarial framework to constrain CNN models to learn domain-invariant representations (Section 2.3.2)), and compare it with staining normalization (Section 2.3.3), augmentation methods (Section 2.3.3, 2.3.3) and combinations of these methods. Domain-adversarial training differs from conventional methods in the sense it does not rely on defined hard priors: the proposed framework leverages the domain information that is available in most histopathology datasets in order to achieve domain invariance, whereas this information is usually left aside by conventional methods.

This work is an extension of the comparative analysis presented at the 2017 MICCAI-DLMIA workshop [30]. In addition to an extended set of experiments, we also make a novel technical contribution that enables the use of batch normalization when training a single network with different input data distributions, as is required with domain adversarial networks.

We show experiments for two different tasks: 1) mitosis detection with a testing set originating from pathology labs that were unseen during training and 2) nuclei segmentation with a testing set consisting of tissue types that were unseen during training.

2.2 Related Work

Machine learning models for histopathology image analysis that directly tackle the appearance variability can be grouped in two main categories: 1) methods that rely on pre-processing of the image data and 2) methods that directly modify the machine learning model and/or training procedure.

The first group of methods includes a variety of staining normalization techniques [31], [32]. Some image processing pipelines handle the variability problem via extensive data augmentation strategies, often involving color transformations [16], [30], [33]–[35]. Hybrid strategies that perturb the staining distributions on top of a staining normalization procedure have also been investigated [36]–[39].

The second group of methods is dominated by domain adaptation approaches. Domain adaptation assumes the model representation learned from a source domain can be adapted to a new target domain. Fine-tuning and domain-transfer solutions were proposed for deep learning models [40]–[43], and with applications to digital pathology [44]–[46]. Another approach consists in considering the convolutional filters of the CNN as domain-invariant parameters whereas the domain variability can be captured with the Batch Normalization (BN) parameters [47], [48]. Adaptation to new domains can be achieved by fine-tuning a new set of BN parameters dedicated to these new domains [48].

Adversarial training of CNNs was proposed to achieve domain adaptation from a source domain of annotated data to a single target domain from which unlabeled data is available [49].

Adversarial approaches aim at learning a shared representation that is invariant to the source and target domains via a discriminator CNN, that is used to penalize the model from learning domain-specific features [49]–[53]. This type of method has been successfully applied and adapted to the field of medical image analysis [54]. These methods, however, require that data from the target domains is available at training time, which is not a constraint of our approach and were not investigated on tasks involving histological images.

Finally, we proposed in [30] a similar approach that enforces the model to learn a domain-agnostic representation for a given extensive domain variability present within the training data and we investigated its ability to perform on new unseen domains.

2.3 Material and Methods

We evaluate the different approaches for achieving domain invariance on two relevant histopathology image analysis tasks: nuclei segmentation and mitosis detection. Automated nuclei segmentation is an important tool for many downstream analyses of histopathological images, such as assessment of nuclear pleomorphism. Mitosis detection is the first step towards assessment of the tumor proliferation activity, and is therefore an important biomarker for breast cancer prognostication and part of the widely used Bloom-Richardson-Elston grading system [55].

In this section, we first describe the datasets used for the two image analysis tasks, and specify the domain shift under which the generalization of trained models on new domains will be assessed. Then, we describe the baseline convolutional neural network model, the domain-adversarial framework, the staining normalization and the data augmentation approaches that will be used in the comparative analysis.

2.3.1 Datasets

The proposed comparative analysis was made on two datasets which expose two different types of domain variability. These datasets correspond to different tasks, enabling to study the framework viability in multiple analysis settings.

Inter-Lab Mitosis Dataset

We used the TUPAC16 dataset [56] that includes 73 breast cancer cases with histological slides stained with Hematoxylin-Eosin (H&E). The dataset consists of a selection of high power field images (HPF) that were annotated with mitotic figure locations, derived from the consensus of at least two pathologists.

The cases come from three different pathology labs ($\mathrm{PL_A}$, $\mathrm{PL_B}$ and $\mathrm{PL_C}$ with 23, 25 and 25 cases respectively) and were scanned with two different whole-slide image scanners (the slides from $\mathrm{PL_B}$ and $\mathrm{PL_C}$ were scanned with the same scanner). We split the dataset as follows:

- A training set of eight cases from $\mathrm{PL_A}$ (458 mitoses).

- A validation set with four other cases from $\mathrm{PL_A}$ (92 mitoses).
- A test set with the remaining 11 cases from $\mathrm{PL_A}$ (533 mitoses), in order to measure the intra-lab performance of the trained models in the same condition as the AMIDA13 challenge [57].
- A test set using the 50 cases from $\mathrm{PL_B}$ and $\mathrm{PL_C}$ (469 mitoses), in order to evaluate inter-lab generalization performance.

Multi-Organ Nuclei Dataset

We used the multi-organ dataset created in [32]: a subset of 30 HPF images, selected from single WSIs of H&E-stained tissue slices, prepared in 18 different hospitals, and provided by The Cancer Genome Atlas [58]. These 30 images consist of seven different tissue types with nuclei mask annotations publicly available [32].

To be in conditions similar to [32], we split the dataset in two groups of tissue types $\mathrm{T_A}$={*Breast, Liver, Kidney, Prostate*} and $\mathrm{T_B}$={*Bladder, Colon, Stomach*}. For experimental purpose, we split the dataset in the conditions of [32] as follows:

- A training set of 12 HPF images with three images for each tissue type of $\mathrm{T_A}$ (7337 nuclei).
- A validation set of four other HPF images with one images for each tissue type of $\mathrm{T_A}$ (1474 nuclei).
- A test set of eight other HPF images with two images for each tissue type of $\mathrm{T_A}$ (4130 nuclei).
- A test set using the six HPF images of $\mathrm{T_B}$ with two images of each type (4025 nuclei), in order to to evaluate cross-tissue-type generalization performance.

2.3.2 Domain-Adversarial Framework

The framework we propose is designed for classification tasks given images $\mathbf{x}$ that are associated with class labels $\mathbf{y}$.

The Underlying Convolutional Network

The proposed framework is applicable to any baseline CNN architecture that can be decomposed in two parts: a feature extractor CNN $\mathcal{F}$ and a classifier CNN $\mathcal{C}$, parameterized by $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_C$ respectively, as illustrated in Figure 2.2.

$\mathcal{F}$ takes images $\mathbf{x}$ as input and outputs an intermediate representation $\mathcal{F}(\mathbf{x};\boldsymbol{\theta}_F)$, whereas $\mathcal{C}$ takes $\mathcal{F}(\mathbf{x};\boldsymbol{\theta}_F)$ as input and outputs a classification probability $\mathcal{C}(\mathcal{F}(\mathbf{x};\boldsymbol{\theta}_F);\boldsymbol{\theta}_C)$. The $(\mathcal{F},\mathcal{C})$ pipeline can be trained by minimizing the cross-entropy loss $\mathcal{L}_C(\mathbf{x},\mathbf{y};\boldsymbol{\theta}_F,\boldsymbol{\theta}_C)$. $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_C$ are optimized by stochastic gradient descent using mini-batches of image-label pairs $(\mathbf{x},\mathbf{y})$.

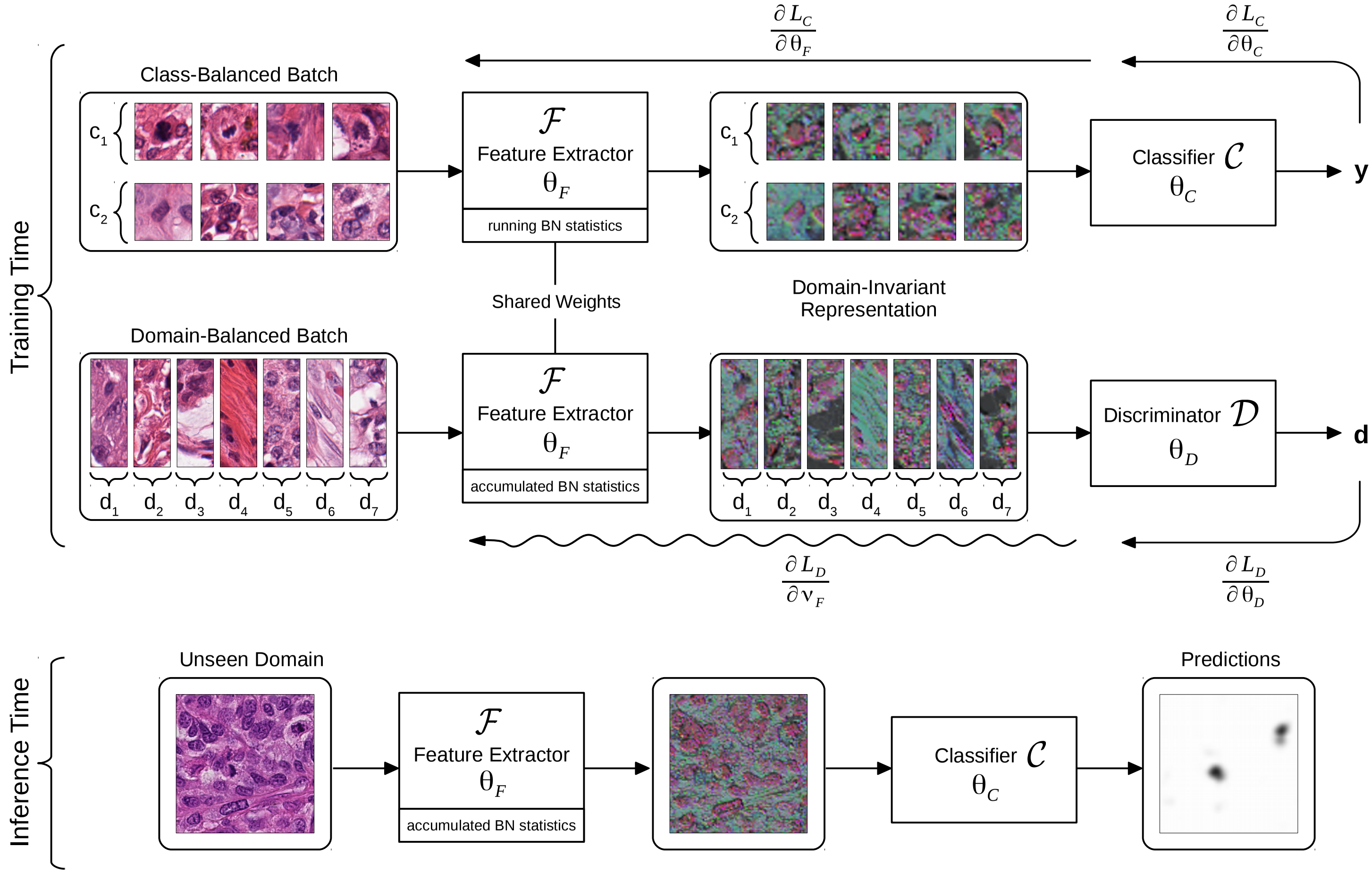


Figure 2.2: Flowchart of the domain-adversarial model. The model is trained using batches balanced across classes or domains. The intermediate representation is learned to optimize the classifier, while the discriminator is trained to identify domains of origin from this representation. Gradient back-propagations are shown with right-to-left arrows, adversarial back-propagation is shown with a sinuous arrow. The domain-invariant representation is illustrated as a selection of three activation maps that are output by the feature extractor, all assigned to RGB channels.

Domain-Adversarial Training

The goal of the framework is to make the intermediate representation $\mathcal{F}(\mathbf{x}, \boldsymbol{\theta}_F)$ invariant to the domains of the training data. We make the assumption that by making $\mathcal{F}(\mathbf{x}, \boldsymbol{\theta}_F)$ domain-agnostic, this will improve the cross-domain generalization of the classifier $\mathcal{C}$. By making the representation invariant to the known domain variability of the training data, we can expect, in some extent, it will also be invariant to unseen variability factors.

Towards this goal, we turned the baseline CNN to a domain-adversarial neural network (DANN) [49] by involving a discriminator CNN $\mathcal{D}$ with parameters $\boldsymbol{\theta}_D$. $\mathcal{D}$ takes the representation $\mathcal{F}(\mathbf{x}; \boldsymbol{\theta}_F)$ as input and predicts the domain probability $\mathcal{D}(\mathcal{F}(\mathbf{x}; \boldsymbol{\theta}_F), \boldsymbol{\theta}_D)$ of the input training images $\mathbf{x}$ via softmax activation. We define $\mathcal{L}_D(\mathbf{x}, d; \boldsymbol{\theta}_F, \boldsymbol{\theta}_D)$ as the cross-entropy loss of the domain discriminator given an input of image-domain pair ($\mathbf{x}$, d), with d a domain identifier, unique to each slide of the training dataset.

The minimization of $\mathcal{L}_D(\mathbf{x}, d; \boldsymbol{\theta}_F, \boldsymbol{\theta}_D)$ during training implies that domain-specific features get extracted from the shared representation that we want to make domain-invariant. Such domain identification is possible since regular models naturally distribute domains apart in the representation as shown in Figure 2.1 (a). In order to obtain domain-invariance, the weights $\boldsymbol{\theta}_F$ are jointly optimized by stochastic gradient ascent, to maximize $\mathcal{L}_D(\mathbf{x}, d; \boldsymbol{\theta}_F, \boldsymbol{\theta}_D)$. This process aims at removing domain-specific features from the representation that are useless for the task at hand, as it is illustrated in Figure 2.1 (b), while still being optimized to improve the performances of $\mathcal{C}$.

Handling Classification-Related and Domain-Related Input Distributions

Batch Normalization (BN) [59] is used throughout the networks $\mathcal{F}$, $\mathcal{C}$ and $\mathcal{D}$ as it is an efficient method that allows fast and stable training, in particular with adversarial components [60]. By normalizing every batch with computed mean and variance at every convolutional layer, BN implies that the distribution of the feature maps is a function of the distribution of the input batch. As a consequence, the distribution of the feature maps will vary with the balance of the batch associated with every pass (see Section 2.3.2).

It is necessary for the domain-adversarial update to be computed with a forward-pass in the same conditions as for the classification pass. Therefore, we propose to apply BN during the adversarial pass using the accumulated moments of $\mathcal{F}$, while keeping a regular BN computation and regular moment accumulation during the classification pass. To this end, we adjusted the adversarial update (2.4) to update only the convolutional weights $\boldsymbol{\vartheta}_F \subset \boldsymbol{\theta}_F$, so as not to interfere with the BN weights, updated according to (2.1) of the classification pass, with a similar motivation as in [48].

The domain-adversarial training procedure consists in alternating between four update rules:

Optimization of the feature extractor with learning rate λ_C:

$$\boldsymbol{\theta}_F \leftarrow \boldsymbol{\theta}_F - \lambda_C \frac{\partial \mathcal{L}_C}{\partial \boldsymbol{\theta}_F} \tag{2.1}$$

Optimization of the classifier :

$$\boldsymbol{\theta}_C \leftarrow \boldsymbol{\theta}_C - \lambda_C \frac{\partial \mathcal{L}_C}{\partial \boldsymbol{\theta}_C} \tag{2.2}$$

Optimization of the domain discriminator with learning rate λ_D:

$$\boldsymbol{\theta}_D \leftarrow \boldsymbol{\theta}_D - \lambda_D \frac{\partial \mathcal{L}_D}{\partial \boldsymbol{\theta}_D} \tag{2.3}$$

Adversarial update of the feature extractor:

$$\boldsymbol{\vartheta}_F \leftarrow \boldsymbol{\vartheta}_F + \alpha \lambda_D \frac{\partial \mathcal{L}_D}{\partial \boldsymbol{\vartheta}_F} \tag{2.4}$$

The update rules (2.1) and (2.4) work in an adversarial way: with (2.1), the parameters $\boldsymbol{\theta}_F$ are updated for the classification task (by minimizing $\mathcal{L}_C$), and with (2.4), a subset of the same parameters are updated to prevent the domains of origin to be recovered from the representation $\mathcal{F}(\cdot; \boldsymbol{\theta}_F)$ (by maximizing $\mathcal{L}_D$). The parameter $\alpha \in [0, 1]$ controls the influence of the adversarial component.

2.3.3 Comparison of Methods

For comparison purpose, we chose to study three different well-established standard methods that aim at improving the generalization of deep learning models in the context of histopathology image analysis and that do not require additional data. A visual overview of these methods is presented in Figure 2.3. We also analyzed combinations of these individual approaches together with the proposed domain-adversarial training framework.

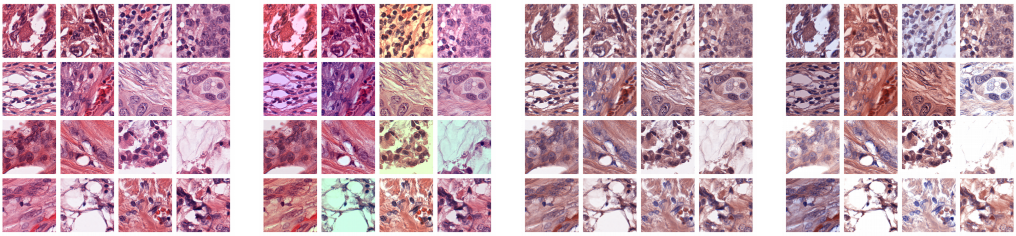


Figure 2.3: Illustration of different types of pre-processing augmentations: (a) Original images, (b) RGB Color Augmentation, (c) Staining Normalization, (d) Staining Augmentation.

Color Augmentation

Since the most prominent source of variability in histology images is the staining color appearance, one alternative to artificially produce new training samples consists in randomly perturbing the color distribution of sampled image patches. By increasing the amount of different color distributions in the training set, the model is expected to learn a representation that better generalize to this type of variability.

We performed color augmentation (CA) by transforming the contrast and shifting the intensities of every color channel $I_c \leftarrow a_c(\cdot I_c - \mu(I_c)) + \mu(I_c) + b_c$, where a_c and b_c are drawn from uniform distributions $a_c \sim U\,[0.9, 1.1]$ and $b_c \sim U\,[-13, +13]$ and where $\mu(I_c)$ is the mean intensity of I_c.

Staining Normalization

The opposite strategy is to reduce the appearance variability of all the images as a preprocessing step before training and evaluating a trained CNN model. For hematoxylin and eosin (H&E) stained slides, staining normalization (SN) methods can be used [61], [62].

The RBG pixel intensities of H&E-stained histopathology images can be modeled with the Beer-Lambert law of light absorption: $I_c = I_0 \exp\left(-\mathbf{A}_{c,*} \cdot \mathbf{C}\right)$. In this expression $c = 1, 2, 3$ is the color-channel index, $\mathbf{A} \in [0, +\infty]^{3\times 2}$ is the matrix of absorbance coefficients and $\mathbf{C} \in [0, +\infty]^2$ are the stain concentrations [61]. We perform staining normalization with the method described in [62]. This is an unsupervised method that decomposes any image with estimates of its underlying $\mathbf{A}$ and $\mathbf{C}$. The appearance variability over the dataset can then be reduced by recomposing all the images using some fixed reference absorbance coefficients $\mathbf{A}_{ref}$.

Staining Augmentation

An approach between CA and SN consists in artificially perturbing the distribution of the concentrations estimated in the unmixing step of SN before applying the recomposing step with constant $\mathbf{A}_{ref}$ [36]–[39].

We experimented with Staining Augmentation (SNA) for comparison, by randomly perturbing each estimated concentration map $\mathbf{C}_i$ linearly with $\mathbf{C}_i \leftarrow g_i \cdot \mathbf{C}_i + h_i$, where g_i and h_i are drawn from uniform distributions $g_i \sim U\,[0.9, 1.1]$ and $h_i \sim U\,[-0.1, +0.1]$.

2.4 Experiments

We implemented two DANN models [49], one for the mitosis detection task and one for the nuclei segmentation task. Both problems are approached with a patch-based classification setup. In the case of mitosis detection, $\mathcal{C}$ outputs the probability for the input patches to be centered on mitotic figures. In the case of nuclei segmentation, $\mathcal{C}$ outputs the 3-class probability vectors for the center of the image patches: nuclei foreground, nuclei edge, or background.

2.4.1 Architectures

For both problems, we chose straightforward convolutional networks, similar to the related literature [14], [30], [32], [56], [57]. We chose to investigate DANN models with a single bifurcation at the second max-pooling layer, corresponding to receptive fields of size 12×12 for the mitosis classifier and 16×16 for the nuclei classifier.

Every convolutional layer is activated by a leaky Rectified Linear Unit (with coefficient 0.01), except for the output layers that are activated by a softmax function. Architecture details are presented in Table 2.1 and Table 2.2.

Table 2.1: Architecture of the mitosis detection model. The feature extractor $\mathcal{F}$ and mitosis classifier $\mathcal{C}$ form a 10-layer CNN with a single class-probability output. The domain classifier $\mathcal{D}$ is a 3-layer network bifurcated at the second max-pooling layer of $\mathcal{F}$ and outputs a 8-domain probability vector.

	Feature Extractor and Mitosis Classifier				Domain Classifier				
	Layer	Size	Filter	Rec. F.	Layer	Output	Filter	Rec. F.	
$\mathcal{F}$	**Input**	$64 \times 64 \times 3$		1×1					
	Conv	$60 \times 60 \times 16$	5×5	5×5					
	Max Pool	$30 \times 30 \times 16$	2×2	6×6					
	Conv	$28 \times 28 \times 16$	3×3	10×10					
	Max Pool	$14 \times 14 \times 16$	2×2	12×12	**Bifurcation**	$14 \times 14 \times 16$		12×12	
$\mathcal{C}$	Conv	$12 \times 12 \times 16$	3×3	20×20	Conv	$12 \times 12 \times 32$	3×3	20×20	$\mathcal{D}$
	Max Pool	$6 \times 6 \times 16$	2×2	24×24	Conv	$10 \times 10 \times 64$	3×3	24×24	
	Conv	$4 \times 4 \times 16$	3×3	40×40	**Softmax**	$10 \times 10 \times 8$	1×1	24×24	
	Max Pool	$2 \times 2 \times 16$	2×2	48×48					
	Conv	$1 \times 1 \times 64$	2×2	64×64					
	Sigmoid	$1 \times 1 \times 1$	1×1	64×64					

2.4.2 Training Procedures

We used the same training procedure for the models of both the problems. For all experimental configurations, image patches were transformed by a baseline augmentation pipeline consisting of a random 90-degree rotation, random mirroring, $-10/+10\%$ spatial-scaling. Sampling of non-mitosis figures and nuclei background classes were adjusted by hard-negative mining using a first version of the baseline models to

Table 2.2: Architecture of the nuclei segmentation model. The feature extractor $\mathcal{F}$ and nuclei classifier $\mathcal{C}$ form a 9-layer CNN with a 3-class probability output. The domain classifier $\mathcal{D}$ is a 3-layer network bifurcated at the second max-pooling layer of $\mathcal{F}$ and outputs a 12-domain probability vector.

	Feature Extractor and Nuclei Classifier				Domain Classifier				
	Layer	Size	Filter	Rec. F.	Layer	Output	Filter	Rec. F.	
$\mathcal{F}$	**Input**	$52 \times 52 \times 3$		1×1					
	Conv	$48 \times 48 \times 24$	5×5	5×5					
	Max Pool	$24 \times 24 \times 24$	2×2	6×6					
	Conv	$20 \times 20 \times 24$	5×5	14×14					
	Max Pool	$10 \times 10 \times 24$	2×2	16×16	**Bifurcation**	$10 \times 10 \times 24$		16×16	
$\mathcal{C}$	Conv	$8 \times 8 \times 24$	3×3	24×24	Conv	$8 \times 8 \times 32$	3×3	24×24	$\mathcal{D}$
	Max Pool	$4 \times 4 \times 24$	2×2	28×28	Conv	$6 \times 6 \times 64$	3×3	28×28	
	Conv	$2 \times 2 \times 24$	3×3	44×44	**Softmax**	$6 \times 6 \times 12$	1×1	28×28	
	Conv	$1 \times 1 \times 96$	2×2	52×52					
	Softmax	$1 \times 1 \times 3$	1×1	52×52					

reject easy-to-classify image patches. The domain-balanced batches were built using patches of size 24×24 for the mitosis detection model and 28×28 for the nuclei segmentation model in order for the domain classifiers to output 1×1 predictions.

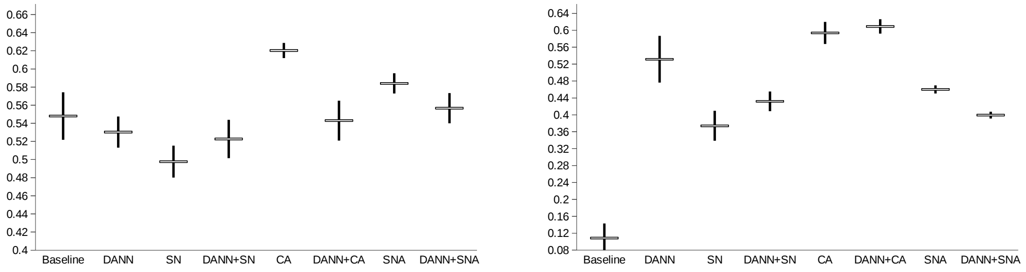
The model weights were optimized with Stochastic Gradient Descent with learning rates $\lambda_C = 0.01$ and $\lambda_D = 0.001$ and momentum $\mu = 0.9$. λ_C and λ_D were decayed by a factor of 0.9 every 5000 iterations. L_2-regularization was applied to all the convolutional weights. For stability purposes and as proposed in [54], we used a warm-up scheduling for the coefficient α, to control the influence of the adversarial component, by following a linear increase from 0.0 to 1.0 from the 5000^{th} to the 10000^{th} training iteration.

2.5 Results

This section presents quantitative and qualitative evaluations of the ability of the developed models to generalize to a known factor of variability of the test set that is absent from the training data.

2.5.1 Mitosis Detection

The performances of the mitosis detection models were evaluated with the F$_1$-score as described in [14], [56], [57]. We used the trained classifiers to produce dense mitosis probability maps for all test images. All local maxima above an operating point were considered detected mitotic figures. This operating point was determined as the threshold that maximizes the F_1-score over the validation set.



(a) F_1-scores – Seen Labs

(b) F_1-scores – Unseen Labs

Figure 2.4: Box-plot of the F_1-score of the mitosis classification models. Points represent the mean ± standard deviation of the F_1-score of each model across 3 repeats with random initialization and random patch sampling. DANN: Domain-Adversarial Neural Network, CA: Color Augmentation, SN: Staining Normalization, SNA: Staining Augmentation.

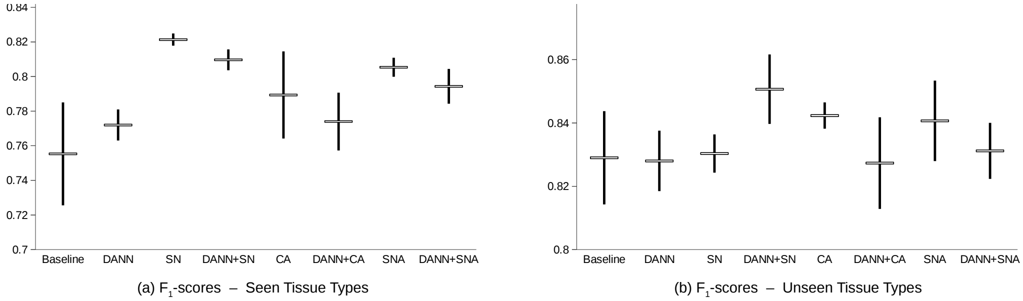
On the test made of images acquired in the same labs as the images of the training set, all methods and combinations have relatively good performances, in line with previously reported results [14], [33], [56], [57]. The best performing method is CA (F_1-score of 0.62 ± 0.008, see Figure 2.4). Adding domain-adversarial training does not improve performance of the conventional methods.

On the other hand, on the test set of images acquired in different labs than for the training set, the best performing method is the combination of CA and DANN (F_1-score of 0.609 ± 0.017). The baseline model does not generalize properly to unseen labs, and domain-adversarial training improves the performances except for the combination with SNA.

2.5.2 Nuclei Segmentation

We used the trained nuclei classifiers to produce segmented nuclei objects. First we generated a set of object seeds using the object foreground map prediction, thresholded by an operating point selected based on a validation set. A set of background seeds were generated using the background prediction, thresholded by a constant of 0.5. Finally a set of segmented nuclei objects was generated using the watershed algorithm given the computed background and foreground seeds and the predicted edges as the topographic relief.

All segmented objects with more than 50% overlap with ground-truth annotations were considered as hits. The performances of the nuclei segmentation models were evaluated with the F_1-score as described in [32], computed over a whole test set.



(a) F_1-scores – Seen Tissue Types

(b) F_1-scores – Unseen Tissue Types

Figure 2.5: Box-plot of the F_1-score of the nuclei segmentation models. Points represent the mean ± standard deviation of the F_1-score of each model across 3 repeats with random initialization and random patch sampling. DANN: Domain-Adversarial Neural Network, CA: Color Augmentation, SN: Staining Normalization, SNA: Staining Augmentation.

On the test set of images of seen tissue types, the best performing method is SN (F_1-score of 0.821 ± 0.004, see Figure 2.5). On the test set of unseen tissue types, the best performing method is the combination of SN and domain-adversarial training (F_1-score of 0.851 ± 0.011). On both test sets, domain-adversarial training produces a decrease in generalization performance when combined with augmentation methods (CA and SNA).

The baseline model generalizes properly due to the high variability already present in the training set, and therefore is in line with the results reported in [32]. We report a difference of the range of performances between the two test sets.

2.5.3 Qualitative Results

Qualitatively, we observe that the baseline models fail to generalize with images that have unseen low-contrast appearance (see *Bladder* and *Colon* examples in Figure 2.6). This limit is solved by methods involving staining normalization. The addition of domain-adversarial training tends to better separate touching nuclei, improving the F_1-score.

Likewise, low-contrast structures occurring in the images from the unseen labs entail false positive detection of mitotic figures (see Figure 2.7), whereas these do not occur for models trained using CA. The addition of domain-adversarial training tends to produce smoother distribution of the predictions, resulting in a higher rate of true positives and higher F_1-score.

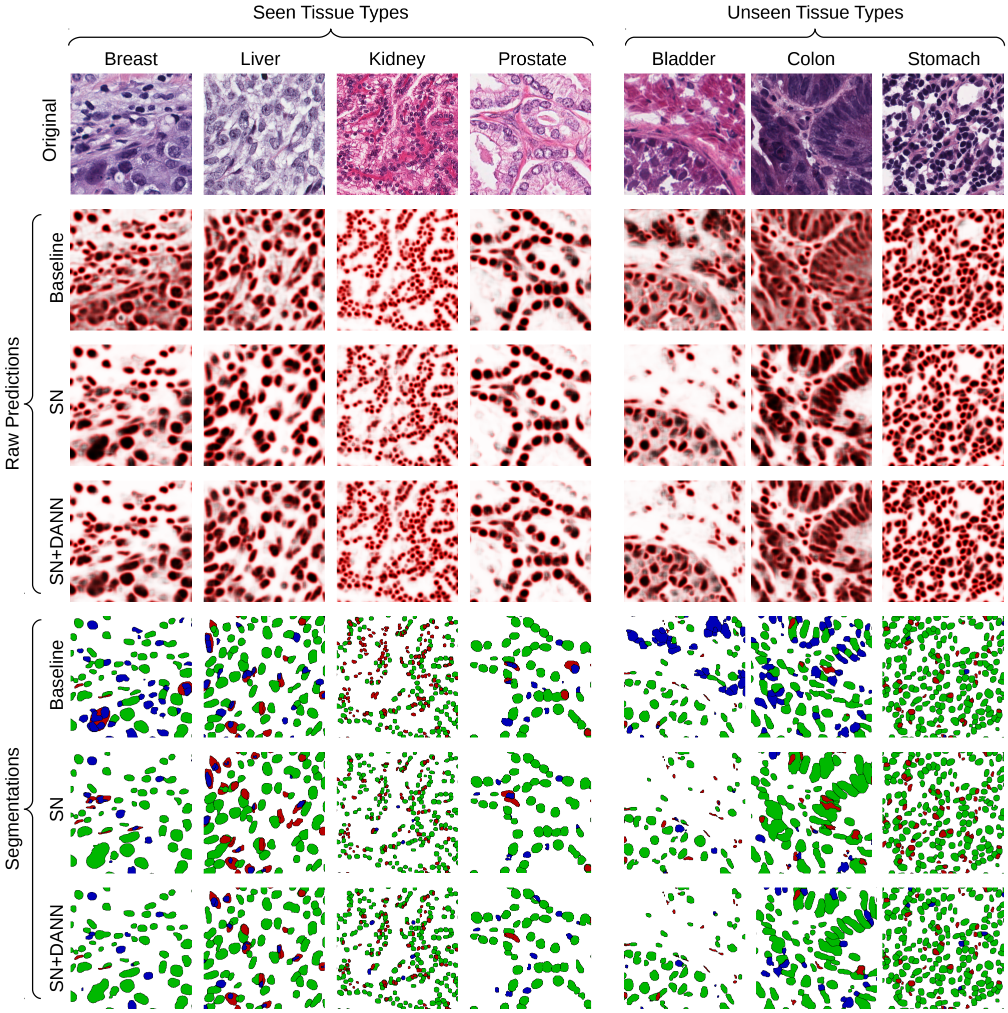


Figure 2.6: Visualization of the raw predictions (background: white, foreground: black, border: red) and resulting segmentation maps of the baseline and best-performing segmentation models. True positive, false positive and false negative are show in green, blue and red respectively.

2.6 Discussion and Conclusions

The relative improvement of performances brought by the analysed methods depends on the data and task at hand. In the case where the training data presents a high domain variability (images from multiple labs, multiple organ types), SN is the most effective method when testing on the test set that consists of the same tissue types, be-

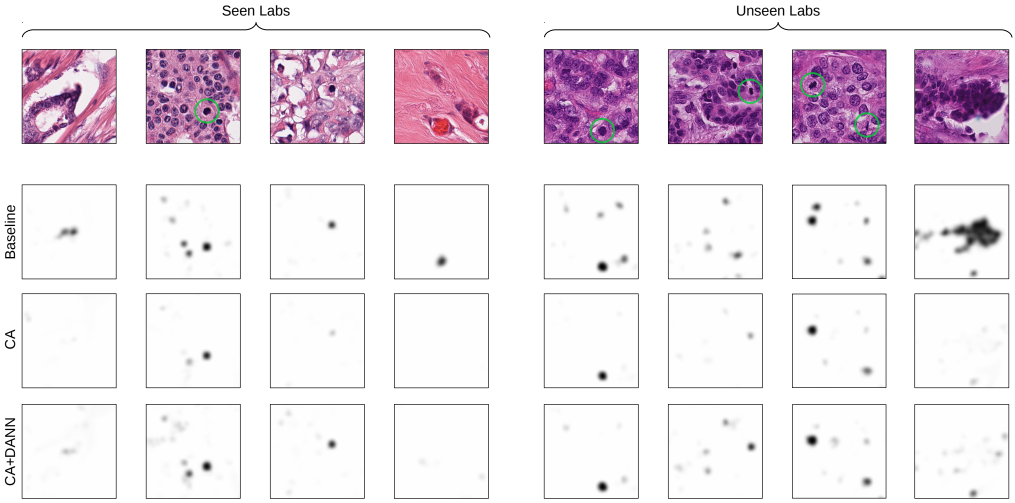


Figure 2.7: Visualization of the predictions of the baseline and best-performing models (CA for seen-lab test set, and CA+DANN for unseen-lab test set). Ground-truth mitotic figures are circled in green.

cause the model can learn to be efficient to the range of staining variability observed on these specific tissue types. However, CA and SNA generalize better than SN on the test set that consists of unseen tissue types as they allow to anticipate the new color and staining variability that can occur in these images. We assume this limit of SN is overcome when combined with domain-adversarial training, as it enables the model to improve the generalization of the learned representation beyond the range of staining distributions seen in the training set.

In the case where the training data presents a low domain variability (intra-lab variability only), CA and SNA, were the most effective methods when testing on unseen images whether they were obtained in the same lab as for the training data or in different labs. This implies that augmentation methods or domain-adversarial training can better anticipate unseen color/staining distributions than SN in this situation. The failure case of the baseline model indicates *overfitting* to the limited variability of domains of the training data and is avoided by CA, SNA or domain-adversarial training. The additional improvement of performances shown when domain-adversarial training is combined with CA indicates that this approach helps the model to generalize to factors other than colors.

Two design choices need to be considered in the proposed domain-adversarial framework as we assume they have an influence on the task performances. These parameters depend on the type of image, task at hand and type of domain variability, and thus need to be carefully tuned.

First, the depth level of $\mathcal{F}$ has to be chosen: with an early bifurcation, the low-level features can be made invariant (with the risk of over-fitting to the domains of the training data), whereas a late bifurcation can make the high-level features invariant with the risk that the early features do not get affected by the domain-adversarial update, thus failing to extract features in unseen domains. Fine-tuning this hyper-parameter is necessary to obtain optimal performances. An alternative solution could consist in using multiple bifurcations as it was proposed in [30], [54].

The receptive field of $\mathcal{D}$ on the input is another point to consider. Depending on the task at hand, the receptive field of $\mathcal{D}$ does not need to be necessarily the same as $\mathcal{C}$, especially if the source of domain variability can be captured in a field of view smaller than the objects that are being classified. Using too large a receptive field for $\mathcal{D}$ raises the risk of identifying, and removing from the representation, some features specific to a domain that might actually be relevant for the task at hand.

In conclusion, we proposed a domain-adversarial framework for training CNN models on histopathology images, and we made a comparative analysis against conventional pre-proccessing methods. We showed that exploiting slide-level domain information at training time, via an adversarial training process, is thus a suitable additional approach towards domain-invariant representation learning and to improve generalization performances. Still, the performances of a trained model vary with the type of normalization/augmentation method used and the type of variability present in the data at training and inference time. Analyzing these factors is therefore a critical decision step when designing machine-learning models for histology image analysis. Directions for further research include adapting the framework to other model architectures, other tasks, and exploiting known variability factors other than slide-level information. The relative top-performances that domain-adversarial training achieved, confirm it is a relevant research direction towards a general method for consistent generalization to any type of unseen variability of histological images.

Chapter 3
Roto-translation equivariant convolutional networks

This chapter is based on:

M. W. Lafarge, E. J. Bekkers, J. P. Pluim, R. Duits, and M. Veta, “Roto-translation equivariant convolutional networks: Application to histopathology image analysis”, *Medical Image Analysis,* In press, 2020.

E. J. Bekkers, M. W. Lafarge, M. Veta, K. A. Eppenhof, J. P. Pluim, and R. Duits, “Roto-translation covariant convolutional networks for medical image analysis”, *in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI),* 2018, pp. 440–448.

3.1 Introduction

Invariance to irrelevant factors of variability is a desirable property of machine learning models, in particular for medical image analysis problems for which models are expected to generalize to unseen shapes, appearances, or to arbitrary orientations. For example, histopathology image analysis problems require processing a digital slide of a stained specimen whose global orientation is strictly arbitrary. Indeed, in the preparation workflow of histology slides, resection of the tissue is done arbitrarily and local structures within the section can have any three-dimensional orientation. In this context, models whose output varies with the orientation of the input constitute a source of uncertainty. The output of such image analysis systems should be rotation invariant, meaning that the output of a model should not change when its input is rotated.

Convolutional Neural Networks (CNNs) are the method of choice to solve complex image analysis tasks, in part due to the translation co-variance induced by trainable $\mathbb{R}^2$ convolution operators. In theory, this structure allows CNNs to learn features in any orientation given sufficient capacity. For example, if a specific edge detector is a relevant filter for the task at hand, it is expected that the CNN learns this filter in all possible directions. Typical solutions to obtain rotation invariance consist in augmenting the dataset by generating additional randomly rotated samples, with the expectation that the model will learn the relevant features that are artificially observed under these additional orientations. Although data augmentation is a way to induce and encourage an invariance prior, such approaches do not guarantee conventional CNNs to be rotation-invariant. Furthermore, with such approaches it is common practice to average predictions of the trained model on a set of rotated inputs at test time: this can increase the robustness of the model, however it comes at the cost of a computational overhead.

We propose to replace convolutions in $\mathbb{R}^2$ by group convolutions using representations of the special Euclidean motion group *SE(2)* (roto-translation of a kernel) so as to explicitly encode the orientation of the learned features. This structure ensures that the learned representation is co-variant/equivariant with the orientation of the input for rotations that lay on the pixel grid and to some extent for rotations that are out of the pixel grid. This equivariance property implies that an oriented feature of interest will get extracted independently of the spatial orientation of the input. We achieve orientation encoding at resolution levels higher than 90-degree via bi-linear interpolation of the *SE(2)* convolution kernels. Finally rotation invariance can be achieved via a projection operation with respect to the encoded orientation of the learned representation.

Contributions This work builds upon our previous work presented at the MICCAI conference 2018 [63]. In addition to a more detailed description of the proposed framework, we now present a comparative analysis of models with different angular discretization levels of the *SE(2)*-image representations. Here we focus on three types of histopathology image analysis problems (mitosis detection, nuclei segmentation and tumor detection), for which we conduct experiments on popular and realistic benchmark datasets. With this we also show that the *SE(2)*-image representations can be integrated in other classical CNN architectures such as U-net [64]. Finally, in a new series of in-depth experimental analyses we show an increased robustness of the proposed group convolutional neural networks (G-CNNs) compared to standard CNNs with respect to rotational variations in the data. This includes a quantitative and qualitative assessment of rotational invariance of the trained networks, as well as a data regime analysis in which we investigate the effect of increased angular resolution when the data availability is reduced.

3.2 Rotation Invariance, Related Work, and Contributions

3.2.1 Rotation Invariance via G-CNNs

We distinguish between invariance and equivariance/covariance as follows. An artificial neural network (NN) is invariant with respect to certain transformations when the output of the network does not change under transformations on the input. We call a NN equivariant, or covariant[1], when the output transforms in a predictable way when the input is transformed (we formalize this statement in Subsection 3.3.2). The property of equivariance guarantees that no information is lost when the input is transformed. Standard CNNs are equivariant to translations: if the input is translated the output translates accordingly and we do not need to worry about learning how to deal with translated inputs. It turns out that *group convolution layers* are the only type of linear NN layers that are guaranteed to be equivariant (see e.g. [65, Thm. 1]) and that the standard convolution layer is a special case that is translation equivariant. In this paper, we construct *SE(2)* equivariant group convolution layers and with it build G-CNNs with which we solve problems in histopathology that require rotation invariance.

Nowadays, rotation invariance is often still dealt with via data augmentations. In such an approach the data is rotated during training time while keeping the target label fixed, thereby aiming for the network to learn how to classify input samples regardless of their orientation. Downsides of this approach are that 1) valuable network capacity is spend on learning geometric behavior at the cost of descriptive representa-

[1] Terminology changes between fields of study (mathematics, physics, machine learning) and often refer to the same. Following custom in machine learning research we will use the term equivariance.

tion learning, 2) rotation invariance is not guaranteed, and 3) augmentation only captures geometric invariance globally. G-CNNs solve these problems by hard-coding geometric structure into the network architecture such that 1) geometric behavior does not have to be learned, 2) rotation invariance is guaranteed by construction, and 3) each group convolution layer achieves local equivariance on its own, so that global equivariance is still obtained when the layers are stacked.

The local-to-global equivariance property means that G-CNNs recognize both low-level features (e.g. edges), mid-level features (e.g. individual cells), and high-level features (e.g. tissue structure) independent of their orientations. In this paper we experimentally show that $SE(2)$ equivariant G-CNNs indeed solve all three aforementioned problems and that in fact the added geometric structures leads to networks that significantly outperform classical CNNs trained with data-augmentation.

3.2.2 Related Work on G-CNNs

G-CNN Methods

In the seminal work by Cohen *et al.* [66] a framework is proposed for group equivariant CNNs. In G-CNNs, the convolution operator is redefined in terms of actions of a transformation group, and by consistent use of the group structure (rules for concatenating transformations) equivariance is ensured. They showed a significant performance gain of G-CNNs over classical CNNs, however, the practical applicability was limited to discrete transformation groups that leave the pixel grid intact (s.a. 90° rotations and reflections). Subsequent work in the field focused on expanding the class of transformation groups that are suitable for G-CNNs by:

1. Working with a grid that has more symmetries than the standard Cartesian grid [67].

2. Expanding convolution kernels in a special basis, tailored to the transformation group of interest, that enables to build steerable CNNs [68]

3. Relying on interpolation methods to transform kernels [63], or relying on analytic basis functions and sample the transformed kernels at arbitrary resolution [69], [70].

Extensions to 3D transformation groups are described in [71]–[74], generalization to equivariance beyond roto-translations are described in [65], [75], extension to spherical data are described in [76]–[79], and additional theoretical results and further generalizations of G-CNNs are described in [77], [80], [81]. Applications of G-CNN methods in medical image analysis are discussed below in Subsection 3.2.2.

Although the first of the above generalizations elegantly enables an exact implementation of G-CNNs of roto-translations with a finer resolution than the 90° rotation angles of [66], it is a very specific approach that does not generalize well to other groups. The second approach does not require to sample transformed kernels at all, but works exclusively by manipulations of basis coefficients in a similar way as standard 2D convolutions (and translations) can be described in the Fourier domain. This approach however requires careful bookkeeping of the coefficients, only optimizes over kernels expressible by the basis, and the choice for non-linear activation functions is limited. In this paper we rely on the third approach. We build upon our previous work [63] and use bi-linear interpolation to efficiently transform (unconstrained) convolution kernels. This allows us to build *SE(2)* equivariant G-CNNs at arbitrary angular resolutions.

Rotation Equivariant Machine Learning

Prior, and in parallel, to the above discussed G-CNN methods, group convolution methods for pattern recognition have been proposed that, at the time, were not regarded as G-CNNs or not treated in the full generality of (end-to-end) deep learning. E.g., Gens *et al.* [82] redefine the convolution operator and construct sparse (approximative) group convolution layers that are used to build what they called deep symmetry networks. Scattering convolution networks, as proposed by Mallat [83], involve a concatenation of separable group convolutions with well-designed hand-crafted filters followed by the modulus as activation function. Other examples are orientation score based template matching [84], cyclic symmetry networks [85], oriented response networks [86], and vector field networks [87], which can all be considered instances of roto-translation equivariant G-CNNs.

Other techniques that focus on equivariance properties of CNNs work via transformations on input feature maps, rather than transformations of convolution kernels as in G-CNNs, and are closely related to spatial transformer networks [88]. These methods include warped CNNs [89], polar transformer networks [90], and equivariant transformer networks [91]. Although these methods describe elegant and efficient ways for achieving (global) equivariance, they often break translation equivariance and local symmetries as the transformations act globally on the whole inputs.

Group Theory in Medical Image Analysis

Equivariance constraints and group theory take a prominent position in the mathematical foundations of “classical” image analysis, e.g., in scale space and wavelet theory. In medical image analysis, group theoretical algorithms enable to respect natural equivariance constraints and deal with context and the complex geometries that are abundant in medical images. Examples of group theoretical techniques, closely related to G-CNNs, are orientation score [92], [93] methods such as crossing preserving vessel enhancement based on gauge theory on Lie groups [94]–[96], vessel and nerve fiber enhancement (in diffusion imaging) via group convolutions with Gaussian (derivative) kernels [97]–[99], and anatomical landmark recognition via group convolutions[65]. In other, non-convolutional methods in medical image analysis, group theory provides a powerful tool to deal with symmetries and geometric structure, such as in statistical shape atlases [100], shape matching [101], registration [102], [103] and in general in statistics on non-Euclidean data structures [104]. Following this successful line of geometry driven methods in medical image analysis, we propose in this paper to rely on G-CNNs to solve tasks in histopathology in an end-to-end learning setting.

G-CNNs in Medical Image Analysis

For many medical image analysis tasks, the location, reflection or orientation of objects of interest should not affect the output of the developed models. Although typical solutions rely on data augmentation, several studies investigated G-CNNs in the context of medical image analysis to leverage this prior into building equivariant models that outperform classical CNNs.

In Winkels *et al.* [72], Andrearczyk *et al.* [74], and Winkels *et al.* [105], G-CNNs were used to detect pulmonary nodules in CT scans. G-CNNs were also investigated for segmentation tasks in dermoscopy images [106], retinal images [63] and microscopy images [63], [107], [108]. Chidester *et al.* [109] proposed a variation of G-CNNs for the classification of sub-cellular protein localization in microscopy images.

Rotation-equivariant models have shown to be particularly efficient for problems in histopathology images, at cell level for mitosis detection [63], nuclei segmentation [107], and at higher tissue levels for tumor detection in lymph node sections [110] and gland-lumen segmentation in colon histology images [108].

3.3 Material and Methods

We evaluate the proposed framework on three relevant histopathology image analysis tasks: mitosis detection, nuclei classification, and patch-based tumor detection. In this section, we first describe the benchmark datasets corresponding to the analysis tasks, that we used to train and evaluate the models. We then describe the relationship between the proposed framework and group theory, and our proposed implementation via bi-linear interpolation of rotated convolution kernels.

3.3.1 Datasets

We chose three popular benchmark datasets of hematoxylin-eosin stained histological slides, in order to assess the performances of the proposed framework and its variants in a controlled and reproducible setup. We chose datasets for which objects of interest are observed at different scales, thus covering a range of problems that are typically

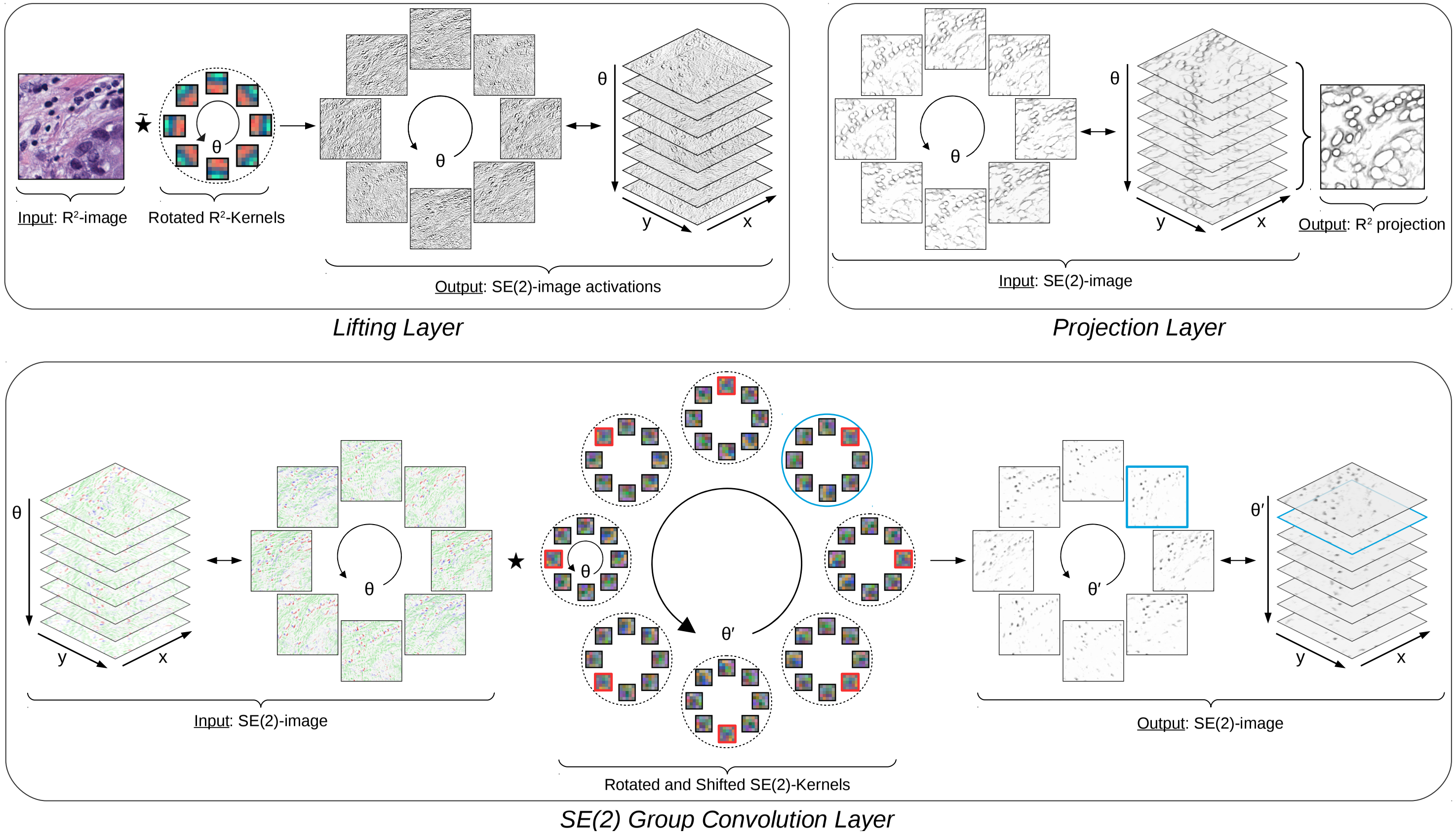


Figure 3.1: Illustration of the three types of layers investigated in our G-CNNs. The *lifting layer* uses a set of rotated kernels in $\mathbb{R}^2$ to output an activation map that is an image on *SE(2)*. The *SE(2) group convolution layer* applies a *shift-twist convolution* via a set of rotated-and-shifted kernels in *SE(2)* to output a *SE(2)*-image activation map (red border highlights the kernel transformation, cyan border highlights the output of a *SE(2)* kernel). The *projection layer* transforms an input *SE(2)*-image onto $\mathbb{R}^2$ via a rotation-invariant operation (pixel-wise maximum projection is used here). A 3-channel input is shown for the *SE(2)* group convolution layer and 1-channel outputs are shown for all the layers: this is done for illustrative purposes but more channels are used in practice. The example images used for the examples are extracted from a trained nuclei segmentation model with a 8-fold discretization of *SE(2)*.

addressed in histopathology image analysis. In these datasets, we assume that the orientation of the objects of interest is irrelevant for the classification task. Therefore we hypothesize that any bias in the orientation information captured by a non-rotation-invariant CNN could be reflected in its performance on the selected benchmarks. This hypothesis will be experimentally confirmed in Section 3.5.

Mitosis Detection We used the public dataset *AMIDA13* [57] that consists of high power-field (HPF) images (resolution $\sim 0.25 \mu\text{m/px}$) from 23 breast cancer cases. Eight cases (458 mitotic figures) were used to train the models and four cases (92 mitoses) for validation. Evaluation is performed on a test set of 11 independent cases (533 mitoses), following the evaluation procedure of the *AMIDA13* challenge, for details see [57].

Multi-Organ Nuclei Segmentation We used the subset of the public multi-organ dataset introduced by [32], that consists of 24 HPF images (resolution $\sim 0.25 \mu\text{m/px}$), selected from WSIs of four different tissue types (Breast, Liver, Kidney and Prostate), provided by *The Cancer Genome Atlas* [58], associated with mask annotations of nucleus instances. We used the balanced dataset split proposed in [111]: 4×3 HPF images for training (7337 nuclei), 4×1 HPF images for validation (1474 nuclei) and 4×2 HPF images for testing (4130 nuclei). Given the high staining variability of the dataset, all the images were stain normalized using the method described in [62].

Patch-Based tumor detection We used the public *PCam* dataset introduced by [110], that consists of $327,680$ image patches (resolution $\sim 1 \mu\text{m/px}$), selected from WSIs of lymph node sections derived from the *Camelyon16* Challenge [18]. The patches are balanced across the two classes (benign or malignant), based on the tumor area provided in [18], and we used the dataset split proposed by [110].

Data Regime Analysis In order to study the behavior of the compared models when data availability is reduced, we analyzed the performances under different data regimes, by using reduced versions of the training sets. We constructed:

- Three variations of the mitosis dataset by sequentially removing two cases out of the original eight.
- Two variations of the nuclei dataset by sequentially removing one HPF image per organ out of the original three HPF images per organ.
- Four variations of the patch-based tumor dataset by randomly removing 25%, 50%, 75% and 90% in each class-subset of the training data.

3.3.2 Group Representation in CNNs

The Roto-Translation group *SE(2)*

A group is a mathematical structure that consists of a set G, for example a collection of transformations, together with a binary operator $\cdot$ called the group product that satisfies four fundamental properties: *Closure*: For all $h, g \in G$ we have $h \cdot g \in G$; *Identiy*: There exists an identity element e; *Inverse*: for each $g \in G$ there exists an inverse element $g^{-1} \in G$ such that $g^{-1} \cdot g = g \cdot g^{-1} = e$; and *Associativity*: For each $g, h, i \in G$ we have $(g \cdot h) \cdot i = g \cdot (h \cdot i)$.

The group product essentially describes how two consecutive transformations, e.g. by $g, h \in G$, result in a single net transformation $(g \cdot h) \in G$. Here, we consider the group of roto-translations, denoted[2] by $SE(2) = \mathbb{R}^2 \rtimes SO(2)$, which consists of the set of all planar translations (in $\mathbb{R}^2$) and rotations (in $(SO(2))$, together with the group product given by

$$g \cdot g' = (\mathbf{x}, \mathbf{R}_\theta) \cdot (\mathbf{x}', \mathbf{R}_{\theta'}) = (\mathbf{R}_\theta \mathbf{x}' + \mathbf{x}, \mathbf{R}_{\theta+\theta'}), \tag{3.1}$$

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. Representations of a group G are linear transformations $\mathcal{R}_g : \mathbb{L}_2(X) \to \mathbb{L}_2(X)$, parameterized by group elements $g \in G$ that transform vectors, e.g. signals/images $f \in \mathbb{L}_2(X)$ on a space X, and which share the group structure via

$$(\mathcal{R}_g \circ \mathcal{R}_h)(f) = \mathcal{R}_{g \cdot h}(f), \qquad \text{with } g, h \in G.$$

We use different symbols for the representations of $SE(2)$ on different type of data structures. In particular, we write $\mathcal{R} = \mathcal{U}$ for the left-regular representation of $SE(2)$ on 2D images $f \in \mathbb{L}_2(\mathbb{R}^2)$, and it is given by

$$(\mathcal{U}_g f)(\mathbf{x}') = f(\mathbf{R}_\theta^{-1}(\mathbf{x}' - \mathbf{x})), \tag{3.2}$$

[2] It is the semi-direct product (denoted by $\rtimes$) of the group of planar translations $\mathbb{R}^2$ and rotations $SO(2)$, i.e., it is not the direct product since the rotation part acts on the translations in (3.1) in the group product of $SE(2)$.

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

$$(\mathcal{L}_g F)(g') = F(g^{-1} \cdot g') = F(\mathbf{R}_\theta^{-1}(\mathbf{x}' - \mathbf{x}), \theta' - \theta), \tag{3.3}$$

with $g = (\mathbf{x}, \theta), g' = (\mathbf{x}', \theta') \in SE(2)$. In Section 3.3.3 we define the G-CNN layers in terms of these representations.

Equivariance

Given the above definitions, we can formalize the notation of equivariance. An operator $\Phi : \mathbb{L}_2(X) \to \mathbb{L}_2(Y)$ is equivariant with respect to a group G if

$$\Phi(\mathcal{R}_g(f)) = \mathcal{R}'_g(\Phi(f)), \tag{3.4}$$

with $\mathcal{R}_g$ and $\mathcal{R}'_g$ representations of G on respectively functions the domains X and Y. I.e., if we transform the input by $\mathcal{R}_g$, then we know that the output transforms via $\mathcal{R}'_g$. To ensure that we maintain the equivariance property (3.4) of linear operators Φ it is required that we define such Φ in terms of representations of G, that is, via group convolutions (see e.g. [65, Thm. 1], [112, Thm. 21], or [80, Thm. 6.1]).

3.3.3 SE(2) Group Convolutional Network Layers

Notation and 2D Convolution Layers

In the following we denote the space of multi-channel feature maps on a domain X by $(\mathbb{L}_2(X))^N$, with N the number of channels. The feature maps themselves are denoted by $\underline{f} = (f_1, \ldots, f_N)$, with each channel $f_i \in \mathbb{L}_2(X)$. The inner product between such feature maps on X is denoted by

$$(\underline{k}, \underline{f})_{(\mathbb{L}_2(X))^N} := \sum_{c=1}^{N} (k_c, f_c)_{\mathbb{L}_2(X)}$$

with $(k, f)_{\mathbb{L}_2(X)} = \int_X k(\mathbf{x}') f(\mathbf{x}') \mathrm{d}\mathbf{x}'$ the standard inner product between real-valued functions on X. Then, with these notations we note that the classical 2D cross-correlation[3] operator can defined in terms of inner products of input feature map $\underline{f}$ with translated convolution kernels $\underline{k}$ via

$$\begin{aligned}(\underline{k} \star_{\mathbb{R}^2} \underline{f})(\mathbf{x}) :&= (\mathcal{T}_{\mathbf{x}} \underline{k}, \underline{f})_{(\mathbb{L}_2(\mathbb{R}^2))^N} \\ &= \sum_{c=1}^{N} \int_{\mathbb{R}^2} k_c(\mathbf{x}' - \mathbf{x}) f_c(\mathbf{x}') \mathrm{d}\mathbf{x}',\end{aligned} \tag{3.5}$$

[3] 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.

with $\mathcal{T}_{\mathbf{x}}$ the translation operator, the left-regular representation of the translation group $(\mathbb{R}^2, +)$. It is well known that convolution layers Φ, mapping between 2D feature maps (i.e. functions on $X = Y = \mathbb{R}^2$), are equivariant with respect to translations. I.e. in Eq. (3.4) we let $\mathcal{R}'_g = \mathcal{R}_g = \mathcal{T}_{\mathbf{x}}$ be the left-regular representation of the translation group with $g = (\mathbf{x}) \in \mathbb{R}^2$.

Roto-Translation Equivariant Convolution Layers

Next we define two types of convolution layers that are equivariant with respect to roto-translations. We do so simply by replacing the translation operator in Eq. (3.5) with a representation of $SE(2)$. When the input is a 2D feature map $\underline{f} \in (\mathbb{L}_2(\mathbb{R}^2))^N$ we need to rely on the representation $\mathcal{U}_g$ of $SE(2)$ on 2D images, and define the ***lifting correlation***:

$$\begin{aligned}(\underline{k} \tilde{\star} \underline{f})(g) :&= (\mathcal{U}_g \underline{k}, \underline{f})_{(\mathbb{L}_2(\mathbb{R}^2))^N} \\ &= \sum_{c=1}^{N} \int_{\mathbb{R}^2} k_c(\mathbf{R}_\theta^{-1}(\mathbf{x}' - \mathbf{x})) f_c(\mathbf{x}') \, \mathrm{d}\mathbf{x}'.\end{aligned} \tag{3.6}$$

These correlations *lift* 2D image data to data that lives on the 3D position orientation space $\mathbb{R}^2 \times S^1 \equiv SE(2)$ by matching convolution kernels under all possible translations and rotations.

We define the ***lifting layer***, recall Figure 3.1, as an operator $\tilde{\Phi}^{(l)} : (\mathbb{L}_2(\mathbb{R}^2))^{N_{l-1}} \rightarrow (\mathbb{L}_2(SE(2))^{N_l}$ that maps a 2D feature map $\underline{f}^{(l-1)} \in (\mathbb{L}_2(\mathbb{R}^2))^{N_{l-1}}$ with N_{l-1} channels to an $SE(2)$ feature map $\underline{F}^l \in (\mathbb{L}_2(SE(2))^{N_l}$ with N_l channels via lifting correlations with a collection of N_l kernels, denoted with $\mathbf{k}^{(l)} := (\underline{k}_1^{(l)}, \ldots, \underline{k}_{N_l}^{(l)})$, each kernel with N_{l-1} channels, via

$$\underline{F}^{(l)} = \tilde{\Phi}^{(l)}(\underline{f}^{(l-1)}) := \mathbf{k}^{(l)} \tilde{\star} \underline{f}^{(l-1)}, \tag{3.7}$$

where we overload the $\tilde{\star}$ symbol defined in Eq. (3.6) to also denote the lifting correlation between a set of convolution kernels and a vector valued feature map via $\mathbf{k}^{(l)} \tilde{\star} \underline{f}^{(l-1)} := \left(\underline{k}_1^{(l)} \tilde{\star} \underline{f}^{(l-1)} , \ldots , \underline{k}_{N_l}^{(l)} \tilde{\star} \underline{f}^{(l-1)} \right)$. Note that such operators are equivariant with respect to roto-translations when in (3.4) we let $\mathcal{T}_g = \mathcal{U}_g$ and $\mathcal{T}'_g = \mathcal{L}_g$ be the representations of $SE(2)$ given respectively in (3.2) and (3.3), indeed $\tilde{\Phi}^{(l)}(\mathcal{U}_g \underline{f}^{(l-1)}) = \mathcal{L}_g \tilde{\Phi}^{(l)}(\underline{f}^{(l-1)})$.

The lifting layer thus generates higher-dimensional feature maps on the space of roto-translations. An $SE(2)$ equivariant layer that takes such feature maps as input is then again obtained by taking inner products of the input feature map $\underline{F}$ with (3D) roto-translated convolution kernels $\underline{K}$, where the kernels are transformed by application of the representation $\mathcal{L}_g$ of $SE(2)$ on $\mathbb{L}_2(SE(2))$.

Group correlations are then defined as

$$(\underline{K} \star \underline{F})(g) := \sum_{c=1}^{N_c} (\mathcal{L}_g K_c, F_c)_{\mathbb{L}_2(SE(2))} \\ = \sum_{c=1}^{N_c} \int_{SE(2)} K_c(g^{-1} \cdot g') F_c(g') \mathrm{d}g'. \tag{3.8}$$

Note here, that a rotation of an $SE(2)$ convolution kernel is obtained via a shift-twist, a planar rotation and shift along the θ-axis, see Eq. (3.3) and Figure 3.1. The convolution kernels $\underline{K}$ are 3-dimensional and they assign weights to activations at positions and orientations relative to a central position and orientation (relative to $g \in SE(2)$). A set of $SE(2)$ kernels $\mathbf{K}^{(l)} := (\underline{K}_1^{(l)}, \ldots, \underline{K}_{N_l}^{(l)})$ then defines a ***group convolution layer***, which we denote with $\Phi^{(l)}$, and which maps from $SE(2)$ feature maps $\underline{F}^{(l-1)}$ at layer $l-1$, with N_{l-1} channels, to $SE(2)$-feature maps $\underline{F}^{(l)}$ at layer l, with N_l channels, via

$$\underline{F}^{(l)} = \Phi^{(l)}(\underline{F}^{(l-1)}) := \mathbf{K}^{(l)} \star \underline{F}^{(l-1)}, \tag{3.9}$$

where we overload the group correlation symbol $\star$, defined in (3.8), to also denote correlation between a set of convolution kernels and a vector valued feature map on $SE(2)$ via $\mathbf{K}^{(l)} \star \underline{F}^{(l-1)} := \left(\underline{K}_1^{(l)} \star \underline{F}^{(l-1)} \ , \ \ldots \ , \ \underline{K}_{N_l}^{(l)} \star \underline{F}^{(l-1)} \right)$.

Finally, we define the ***projection layer*** as the operator that projects a multi-channel $SE(2)$ feature map back to $\mathbb{R}^2$ via

$$\underline{f}^{(l)}(\mathbf{x}) = \mathcal{P}(F^{(l)})(\mathbf{x}) := \operatorname*{mean}_{\theta \in [0, 2\pi)} \underline{F}^{(l)}(\mathbf{x}, \theta). \tag{3.10}$$

Here we define the projection layer as taking the mean over the orientation axis, however, we note that any permutation invariant operator (on the θ-axis) could be used to ensure local rotation invariance, such as e.g. the commonly used $\max$ operator [63], [66].

3.3.4 Discretized *SE(2,N)* Group Convolutional Network

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 group $SE(2, N) := \mathbb{R}^2 \rtimes SO(2, N)$, with the space of 2D rotations in $SO(2)$ sampled with N rotation angles $\theta_i = \frac{2\pi}{N} i$, with $i = 0, \ldots, N-1$.

The discrete lifting kernels $\mathbf{k}^{(l)}$ at layer l, are used to map a 2D input image with N_{l-1} channels to an $SE(2, N)$-image with N_l channels, and thus have a shape of $n \times n \times N_{l-1} \times N_l$ (the discretization of $\mathbf{k}^{(l)}$ is illustrated in Figure3.1 as a set of n rotated $\mathbb{R}^2$ kernels, distributed on a circle). Likewise, the $SE(2, N)$ kernels $\mathbf{K}^{(l)}$ have a shape of $n \times n \times N \times N_{l-1} \times N_l$.

The lifting and group convolution layers require rotating the spatial part of the kernels and shift along the θ-axis for the $SE(2)$-kernels. We obtain the rotated spatial parts of each kernel via bi-linear interpolation. The discretization of a single lifting kernel $k_{i,j}^{(l)}$ and its N rotated versions is illustrated in the top-left part of Figure 3.1. The discretization of a single group correlation kernel $K_{i,j}^{(l)}$ and its N rotated and θ-shifted versions is illustrated in the bottom part of Figure 3.1.

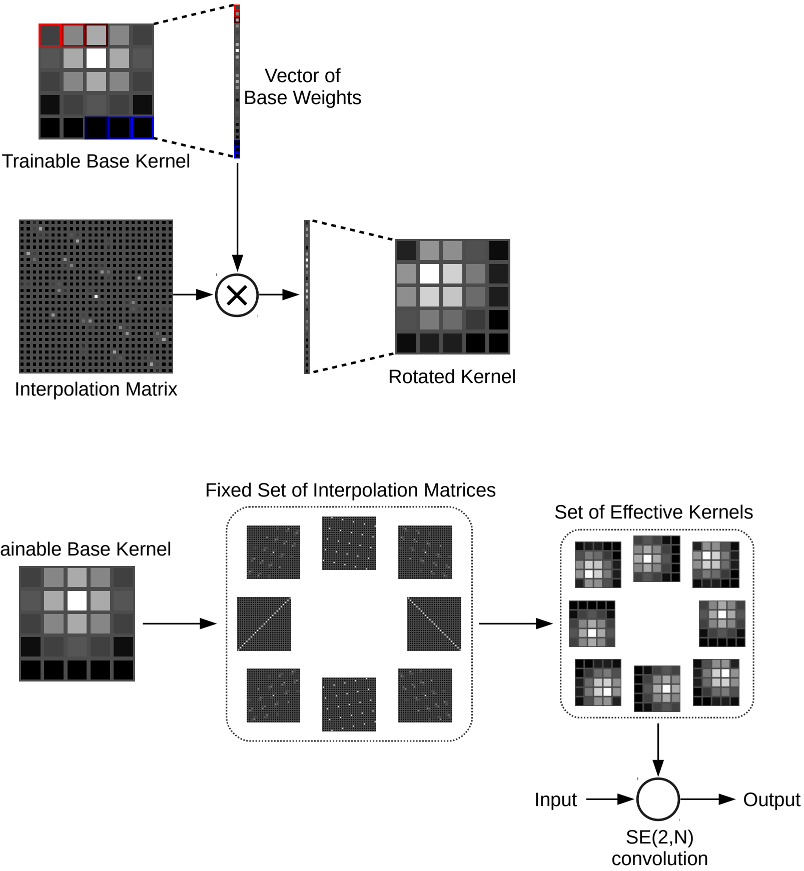


Figure 3.2: Illustration of the process generating a rotated set of effective kernels from a trainable vector of base weights via the introduction of fixed interpolation matrix in the computational pipeline.

In order to construct the rotated sets of effective kernels $\mathbf{k}^{(l)}$ or $\mathbf{K}^{(l)}$ we rely on bi-linear interpolation. We first define a set of trainable vectors containing base weights that are used to generate rotated versions of the same base 2D kernel via bi-linear interpolation. We implemented this rotation process in the computational pipeline via the definition of non-trainable interpolation matrices, each coding for a rotation step, and the introduction of respective matrix multiplication operations. This process is illustrated in Figure. 3.2.

Although these sets of rotated kernels are used in the computational pipeline, only the base weights are updated during the network optimization. By construction, the effective kernels are differentiable with respect to their base weight, enabling their update in back-propagation of gradients (since the matrix multiplication operation is differentiable).

3.4 Experiments

In this section, we present the G-CNN architectures that we build using the layers defined in Section 3.3.3 and we describe the experiments that we used to analyze and validate them. In the construction of the G-CNNs we adhere to the following principle of group equivariant architecture design.

G-CNN design principle A sequence of layers starting with a lifting layer (Eq. (3.7)) and followed by one or more group convolution layers (Eq. (3.9)), possibly intertwined with point-wise non-linearities, results in the encoding of roto-translation equivariant feature maps. If such a block is followed by a projection layer (Eq. (3.10)) then the entire block results in a encoding of features that is guaranteed to be rotationally invariant. Our implementation of the G-CNN layers is available at `https://github.com/tueimage/se2cnn`.

3.4.1 Applications and Model Architectures

For each task introduced in Section 3.3.1 we conducted two experiments: first, we trained a set of variations of a baseline CNN, by changing the orientation sampling level N of their *SE(2,N)* layers, while keeping the total number of weights of each model approximately the same. Second, we trained each model with the reduced data regime counterparts of the training sets introduced in Section 3.3.1. For each task we opted for versions of straight-forward architectures with a low number of parameters that were in-line with methods reported in the literature. This way, we propose new G-CNN baselines that facilitate comparative experiments and that can be extended to more sophisticated architectures.

Mitosis Detection We used the mitosis classification model originally described in [63] as a baseline: a 6-layer CNN with three down-sampling steps, such that the overall receptive field is of size 68×68.

We designed the G-CNN variants of this baseline described in Table 3.1, by replacing the first convolution layer by a lifting layer, replacing the following convolution layers by group convolution layers and inserting a projection layer before the last fully connected layer.

The models were trained with batches of size 64 balanced across classes. Non-mitosis class patches were sampled based on a hard negative mining procedure [14] using a first baseline model trained with random negative patches. The models were trained to minimize the cross-entropy of the binary-class predictions.

Nuclei Segmentation For the nuclei segmentation task, we opted for a 7-layer U-net that corresponds to two spatial down/up-sampling operations with an overall receptive field of size 44×44. The sequence of operations defining this G-CNN architecture is given in the first column of Table 3.2.

The label associated with each input image is a 3-class mask corresponding to the foreground, background and border of the nuclei it contains (these masks can then be used to retrieve an individual nucleus using a segmentation procedure such as described in Section 3.5).

The models were trained with batches of size 16 balanced across patients, to minimize the class-weighted cross-entropy of the *softmax* activated output maps corresponding to the three target masks.

Tumor detection The baseline architecture we used for the tumor detection model is a 6-layer CNN with three down-sampling steps, such that the overall receptive field is of size 88×88 (see Table 3.3 for the detailed architecture).

The models were trained with batches of size 64 balanced across classes. We refined both classes by running a hard negative mining procedure [14] using a first baseline model trained with the original dataset of the benchmark. The models were trained to minimize the cross-entropy of the binary-class predictions.

3.4.2 Implementation details

For all three baseline architectures, convolution kernels are of size 5×5 with circular masking and fully connected layers are implemented as convolutional layers with kernels of shape 1×1 to enable dense application (the resulting models can efficiently be applied on larger input sizes).

Batch Normalization [59] is used throughout the networks. Batch statistics are normally computed across batch and spatial dimensions of the activations, but we also included the orientation-axis of the *SE(2,N)*-image activation maps in the statistic computation to ensure their invariance with respect to the orientation of the input.

All models were trained with Stochastic Gradient Descent with momentum (learning rate 0.01, momentum 0.9) and a epoch-wise learning rate decay using a factor of 0.5 was applied. Training was stopped after convergence of the loss computed on the validation sets. All models were regularized with decoupled weight decay (coefficient

Table 3.1: Architecture of the investigated G-CNN models for mitosis detection. The left-most column indicates the operations applied in each layer. *Max. Proj.* indicates the projection operation on $\mathbb{R}^2$, achieved via maximum intensity projection along the orientations.

	SE(2,N) Groups			
Layers	N=1 ($\mathbb{R}^2$)	N=4 (p4)	N=8	N=16
Input	$68\times68\times3$			
Lifting Layer BN + ReLU MaxPool(2×2)	$1\times42\times42\times16$ (1040)	$4\times42\times42\times10$ (650)	$8\times42\times42\times8$ (520)	$16\times42\times42\times6$ (390)
Group Conv. BN + ReLU MaxPool(2×2)	$1\times14\times14\times16$ (5408)	$4\times14\times14\times10$ (8420)	$8\times14\times14\times8$ (10768)	$16\times14\times14\times6$ (12108)
Group Conv. BN + ReLU MaxPool(2×2)	$1\times5\times5\times16$ (5408)	$4\times5\times5\times10$ (8420)	$8\times5\times5\times8$ (10768)	$16\times5\times5\times6$ (12108)
Group Conv. BN + ReLU	$1\times1\times1\times64$ (21632)	$4\times1\times1\times16$ (13472)	$8\times1\times1\times8$ (10768)	$16\times1\times1\times4$ (8072)
Group Conv. BN + ReLU	$1\times1\times1\times16$ (1056)	$4\times1\times1\times16$ (1056)	$8\times1\times1\times16$ (1056)	$16\times1\times1\times16$ (1056)
Max. Proj.	$1\times1\times16$			
FC Layer + Sigmoid	$1\times1\times1$ (17)			
Total Weights	34561	32035	33897	33751

5×10^{-4}). Baseline augmentation transformations were applied to the training image patches (random spatial transposition, random 90-degree-wise rotation, random channel-wise brightness shifting).

3.4.3 Experiment: Orientation Sampling

In order to assess the effect of using the proposed *SE(2,N)* G-CNN structure on the benchmark performances, we trained every model with $N \in \{1, 4, 8, 16\}$. In order to allow fair comparison we adjusted the number of channels in every layer involving *SE(2,N)*-image representation such that the total number of weights in the models stay close to the count of the corresponding baselines. The detailed distributions of the weights are shown in Tables 3.1, 3.2 and 3.3: for each *SE(2,N)* group, the dimensions of the output of the layers are shown with the format $N\times Height\times Width\times C$, with C the number of output channels in the layer.

Each model was trained three times with random initialization seeds. We report

Table 3.2: Architecture and weight counting of the G-CNN models for patch-based tumor detection. The left-most column indicates the operations in each layer. *Concat(HL.x)* indicates the characteristic skip operation of the U-net architecture that consist in concatenating a centered crop of the output activation of the x^{th} layer of the network. *Max. Proj.* indicates the projection operation on $\mathbb{R}^2$, achieved via maximum intensity projection along the orientations.

	SE(2,N) Groups			
Layers	N=1 ($\mathbb{R}^2$)	N=4 (p4)	N=8	N=16
Input	60×60×3			
Lifting Layer BN + ReLU MaxPool(2×2)	1×28×28×16 (1040)	4×28×28×10 (650)	8×28×28×8 (520)	16×28×28×6 (390)
Group Conv. BN + ReLU MaxPool(2×2)	1×12×12×16 (5408)	4×12×12×10 (8420)	8×12×12×8 (10768)	16×12×12×6 (12108)
Group Conv. BN + ReLU	1×8×8×16 (5408)	4×8×8×10 (8420)	8×8×8×8 (10768)	16×8×8×6 (12108)
Up-sampling Concat(HL.2) Group Conv. BN + ReLU	1×12×12×16 (10784)	4×12×12×10 (16820)	8×12×12×8 (21520)	16×12×12×6 (24204)
Up-sampling Concat(HL.1) Group Conv. BN + ReLU	1×20×20×64 (43136)	4×20×20×16 (26912)	8×20×20×8 (21520)	16×20×20×4 (16136)
Group Conv. BN + ReLU	1×20×20×16 (1056)	4×20×20×16 (1056)	8×20×20×16 (1056)	16×20×20×16 (1056)
Max. Proj.	20×20×16			
FC Layer Softmax	20×20×3 (54)			
Total Weights	66886	62332	66206	66056

the mean and standard deviation of the performances across three random intializations.

3.4.4 Experiment: Data Regime Experiments

In order to assess the effect of using the proposed *SE(2,N)* with varying sampling factor N when data is availability is reduced, we trained each model on the data-regime subsets presented in Section 3.3.1. Likewise, each model was trained three times with random initialization seeds so as to report the variability of the performances.

Table 3.3: Architecture and weight counting of the G-CNN models for patch-based tumor detection. The left-most column indicates the operations in each layer. *Mean. Proj.* indicates the projection operation on $\mathbb{R}^2$, achieved via mean intensity projection along the orientations.

	SE(2,N) Groups			
Layers	N=1 ($\mathbb{R}^2$)	N=4 (p4)	N=8	N=16
Input	88×88×3			
Lifting Layer BN + ReLU MaxPool(2×2)	1×42×42×32 (2080)	4×42×42×19 (1235)	8×42×42×14 (910)	16×42×42×10 (650)
Group Conv. BN + ReLU MaxPool(2×2)	1×19×19×32 (21568)	4×19×19×19 (30362)	8×19×19×14 (32956)	16×19×19×10 (33620)
Group Conv. BN + ReLU MaxPool(3×3)	1×5×5×32 (21568)	4×5×5×19 (30362)	8×5×5×14 (32956)	16×5×5×10 (33620)
Group Conv. BN + ReLU	1×1×1×64 (43136)	4×1×1×16 (25568)	8×1×1×8 (18832)	16×1×1×4 (13448)
Group Conv. BN + ReLU	1×1×1×16 (1056)	4×1×1×16 (1056)	8×1×1×16 (1056)	16×1×1×16 (1056)
Mean Proj.	1×1×16			
FC Layer Sigmoid	1×1×1 (17)			
Total Weights	89425	88600	86727	82411

3.5 Results

This section summarizes the qualitative and quantitative results of the experiments we conducted. Each trained model was evaluated on the test set of its corresponding benchmark dataset based on standard performance metrics.

Mitosis Detection For the mitosis detection task, models were densely applied on test images, followed by a smoothing operation before extracting all local maxima to be considered candidate detections. We computed the F_1-score of the set of detections using an operating point that is optimized on the validation set, as described in the scoring protocol used in [57].

Nuclei Segmentation To quantify the performances of the nuclei segmentation model, generation of segmented candidate objects is obtained by following the protocol used in [32], [111]. First, marker seeds are derived from thresholded foreground and background predictions, border predictions are used as the watershed energy landscape. Then, candidate objects that overlap the nuclei ground-truth masks by at least 50% of their area are considered hits, enabling object-level detection quantification to be calculated using the F_1-score. Thresholds to generate marker seeds were selected such that the F_1-score is maximized on the validation set.

Patch-based tumor detection To evaluate the tumor detection model, we computed the class probability of every patch of the test dataset and calculated the accuracy of the model given the ground-truth labels as in Veeling *et al.* [110] after selection of the operating point that maximizes the accuracy on the validation set.

3.5.1 Qualitative Results

We qualitatively investigated the robustness of the prediction of different models to controlled rotations of the input. We see that the model predictions can be very inconsistent for our best baseline model, in comparison to G-CNN models (see Figure 3.3, Fig. 3.5 and 3.4) in particular for cell or tissue morphologies that are typically asymmetric. For example, the mitotic figures (h) and (i) shown in Fig. 3.3 are in telophase (directed separation of the pair of chromosomes) and the variance of the prediction of the baseline model is higher for these cases (green curve) compared to the G-CNN models (blue and red curves). We also observe that for the *SE(2,4)* model, predictions that are obtained for an input image rotated with an angle below $\pi/2$rad also produce some variance, but present a $\pi/2$rad-period cyclic pattern.

3.5.2 Quantitative Results

The performances of the trained models for both orientation sampling experiments and data regime experiments are summarized in the box plots of Fig. 3.6, 3.7 and 3.8.

Notes on absolute performances For the mitosis detection benchmark, the best result we obtained is in line with the results previously reported in [111] (best F_1-score of 0.62 ± 0.008). For the PCam benchmark, the best result we obtained is in line with the results previously reported in [110] (best accuracy of 0.898). For the nuclei segmentation task, we note that the performances we achieved are significantly lower than

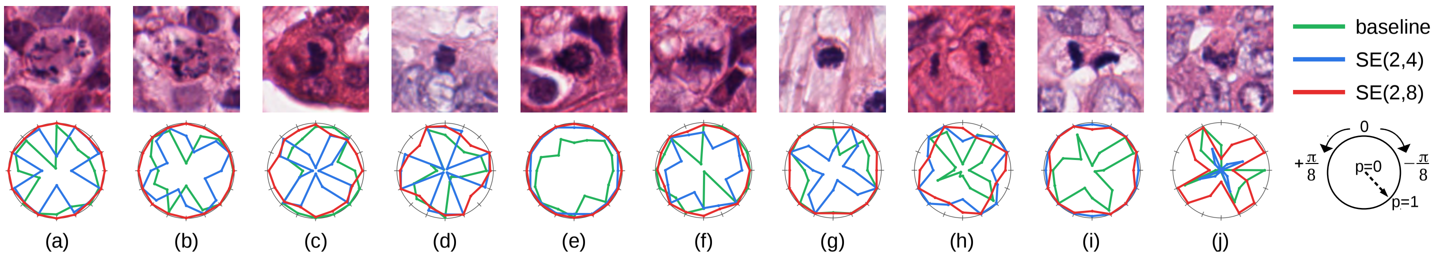


Figure 3.3: Example of mitosis-centered image patches selected from the test set. Below each, polar plots show model predictions (distance from origin) as a function of the orientation of the input (angle coordinate) using steps of $\pi/8$ rad. An ideal model would then produce a circle with maximum radius. Selected models are indicated with colors, and correspond to the best obtained models that were trained without reduced data regime over repeats (based on their F_1-score).

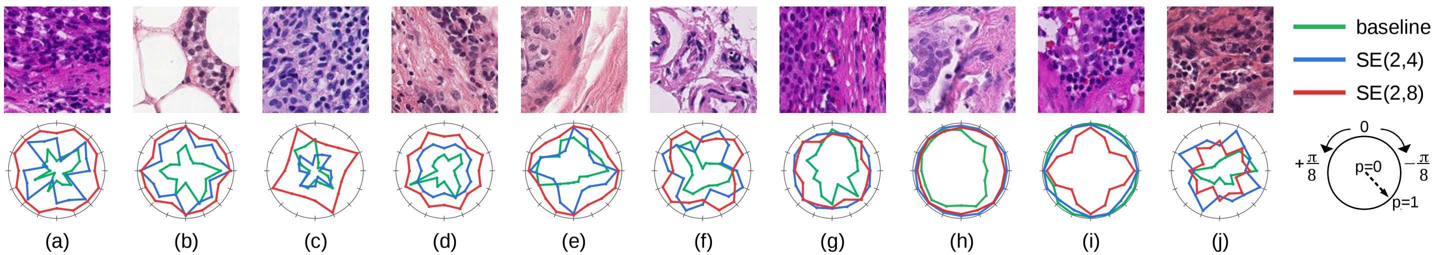


Figure 3.4: Example of image patches selected from the test set of the *PCam* benchmark, for which pixels in the center area were classified as *tumor tissue*. Below each, polar plots show model predictions (distance from origin) as a function of the orientation of the input (angle coordinate) using steps of $\pi/8$ rad. Selected models are indicated with colors, and correspond to the best obtained models that were trained without reduced data regime over repeats (based on their accuracy).

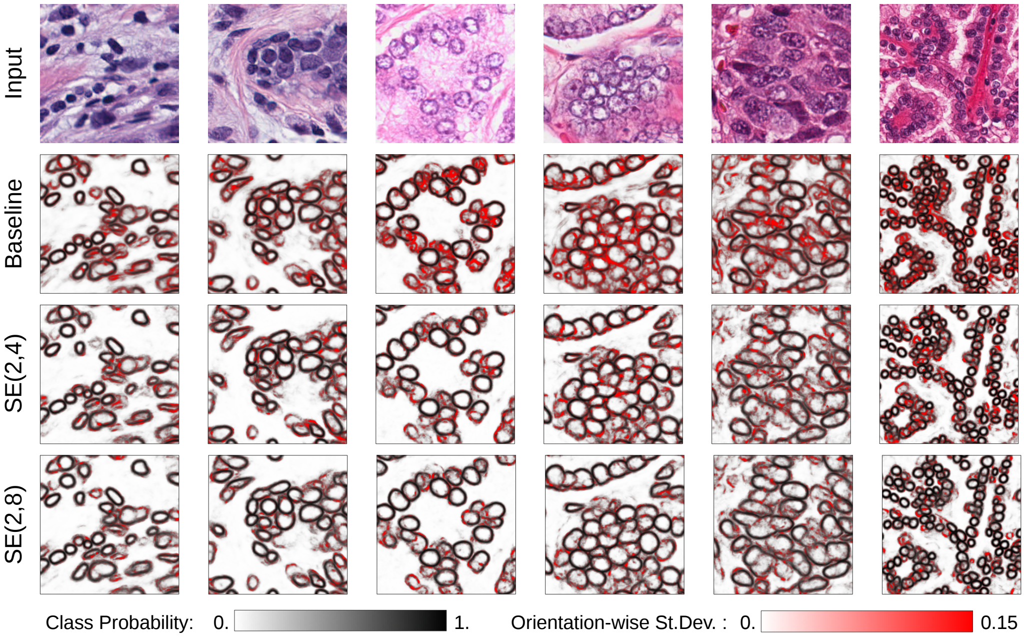


Figure 3.5: Example of image patches selected from the test set of the nuclei segmentation benchmark (column 1-2: breast tissue, column 3-4: prostate tissue, column 5: kidney tissue, column 6: liver). For each image, and a selection of models, the raw predictions of the nucleus boundary class were computed and stored for the set of rotated inputs using steps of $\pi/8$ rad. Predictions were re-aligned and their means were mapped to gray-scale and the standard deviations of the predictions were mapped to a white-to-red color scale. The overlap of these statistics is shown below each original image. Selected models are the best obtained models that were trained without reduced data regime over repeats (based on their F_1-score).

the performances previously reported in the literature (on the same test set, Lafarge *et al.* [111] reported a F_1-score of 0.821 ± 0.004). We explain these difference by the strict constraints we imposed in the design of the baseline segmentation model of this study (lower receptive field, shallower network, lower weight capacity).

Effect of orientation sampling For all three studied tasks, we observed an increase of performance with the number of sampled orientations from $N = 1$ to $N = 8$. For the full data regime of the mitosis detection experiments, the use of a *SE(2,8)* G-CNN improves the F_1-score to 0.626 ± 0.015 on average compared to 0.556 ± 0.016 for the baseline model without test-time rotation augmentation (see Fig. 3.6). A similar increase of performances is observed for the nuclei segmentation experiments with an improvement of the F_1-score from 0.754 ± 0.006 to 0.771 ± 0.06 (see Fig. 3.7), and for the tumor detection experiments with an improvement of the accuracy from 0.863 ± 0.003

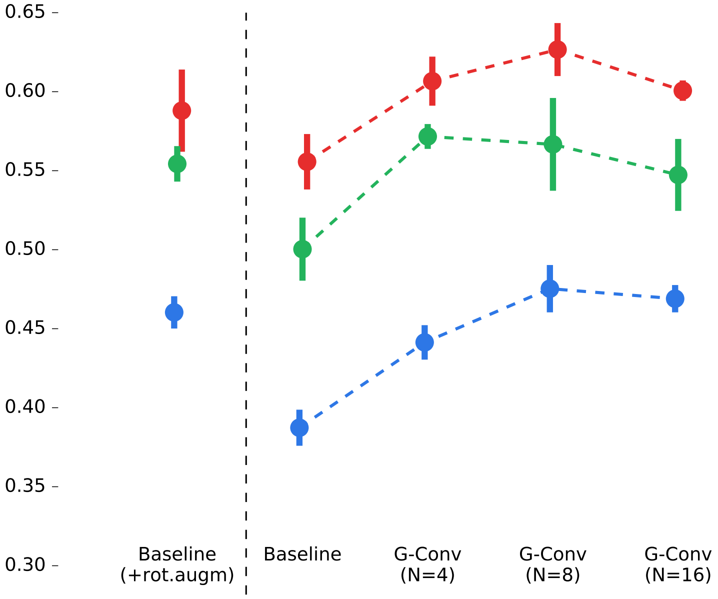


Figure 3.6: Mean and Standard Deviation plots summarizing the F_1-score of the mitosis detection models. Mean ± standard deviation is indicated. Color identifies the different data regime (red: 8 cases; green: 4 cases; blue: 2 cases).

to 0.892 ± 0.004 (see Fig. 3.8).

We remark that the performances of the *SE(2,4)* G-CNN models are better than the baseline with test-time rotation augmentation as was previously reported in literature for similar tasks [63], [110]. We also report that for all three tasks, *SE(2,16)* G-CNN models perform worse than the *SE(2,8)* G-CNN models.

Effect of reduced data regime with orientation sampling For all three tasks, we see a global consistent decrease of performances when less training data is available. In Fig. 3.8, the performances of the *SE(2,4)* and *SE(2,8)* G-CNN models trained with the 25%, 50% and 75% data regimes, are higher than for the baseline model at full data regime using test-time rotation augmentation. This reveals that under experimental conditions, data availability is not the only reason for limited performances since this experiment shows that the *SE(2,N)* G-CNN models enable achieving higher performances than the baseline models, even if less data is available.

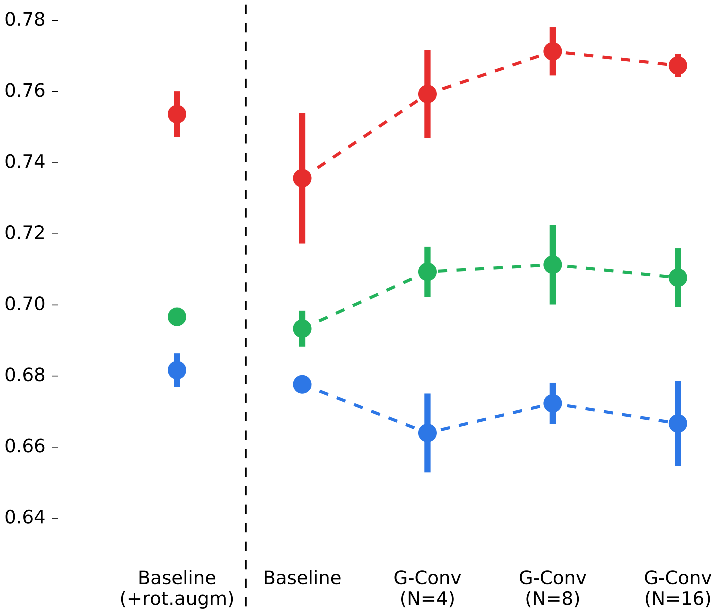


Figure 3.7: Mean and Standard Deviation plots summarizing the F_1-score of the nuclei segmentation models. Mean ± standard deviation is indicated. Color identifies the different data regime (red: 6 HPFs/organ; green: 4 HPFs/organ; blue: 2 HPFs/organ).

3.6 Discussion and Conclusions

The presented study investigated the effects of embedding the *SE(2)* group structure in CNNs, in the context of histopathology image analysis, across multiple controlled experimental setups.

The comparative analysis we conducted shows a consistent increase of performances for three different histopathology image analysis tasks when using the proposed *SE(2,N)* G-CNN architecture compared to conventional CNNs acting in $\mathbb{R}^2$ evaluated with test-time rotation augmentation. This is in line with previously reported results when using G-CNNs with groups that lay on the pixel grid (p4, p4m) [66], [110], but we also show that these performances can be surpassed when using groups with higher discretization levels of *SE(2)*.

This confirms that conventional $\mathbb{R}^2$ CNNs struggle to learn a rotation equivariant representation based on data solely and that enforcing equivariant representation learning enables reaching higher performances. G-CNNs with *SE(2,N)* structure have the advantage to guarantee higher robustness to input orientation without requiring training-time or test-time rotation augmentation. Furthermore, the slight computa-

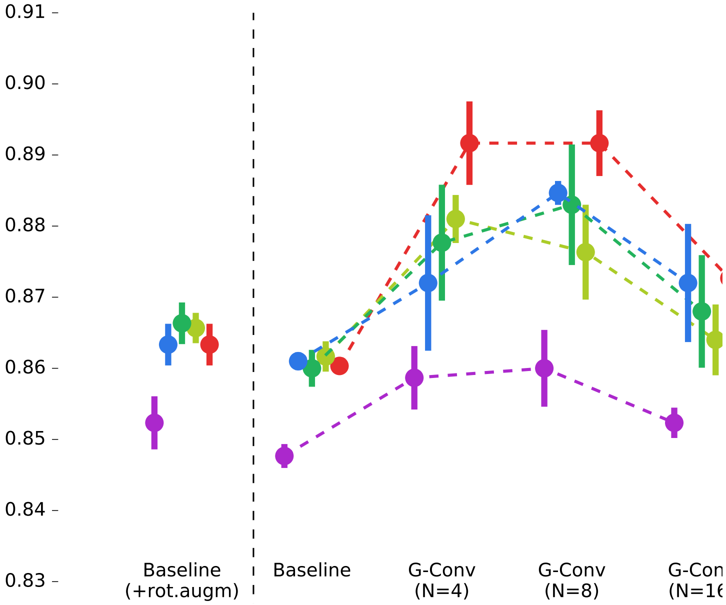


Figure 3.8: Mean and Standard Deviation plots summarizing the accuracy of the tumor detection models. Mean ± standard deviation is indicated. Color identifies the different data regime (red: 100%; lime: 75%; green: 50%; blue: 25%; purple: 10%).

tional overhead for computing rotated convolutional operators and their gradient, at training time, can be canceled at test-time by computing and fixing all final oriented *SE(2,N)* kernels, resulting in a model that is computationally equivalent to conventional $\mathbb{R}^2$ CNNs.

We show that these performances can be surpassed when using representations with higher angular resolution levels, as shown with experiments involving *SE(2,8)* G-CNNs and when the training data is of sufficient amount. This conclusion corroborates the results we reported on other medical image analysis tasks [63] and in studies that investigated models with rotated operators that lay outside of the pixel grid [67].

However, we also identified consistent lower performances for *SE(2,16)* G-CNNs compared to *SE(2,8)* G-CNNs at full data regime. We assume that this phenomenon is in part related to the model architectures we chose to enforce fixed model capacity, resulting in a number of channels in the representation of the *SE(2,N)* models being reduced when N increases. This reduced number of channels might affect the diversity of the features learned by the models, to the point that this limits their overall performances. Therefore, it appears there is a trade-off between performances and angular resolution at fixed capacity, further work would be necessary to confirm this hypothesis.

For the tumor detection task, we observed that the performances of the baseline models (with or without test-time rotation augmentation) reached a plateau, whatever the regime of available training data was among 25%, 50%, 75% or 100%. This indicates that in the conditions of the *PCam* dataset, the amount of available training data does not significantly influence the performances. However, the rotation-equivariant models were able to achieve better performances with increased data regime.

This behavior was not evidenced for the mitosis detection and nuclei segmentation experiments. We assume this result may be task-dependent or might be due to the fact that the plateau of performances observed for the tumor detection models was not reached yet for the two other tasks.

We qualitatively showed that in some cases, the predictions of conventional CNNs are inconsistent when inputs are rotated, whereas *SE(2)* G-CNNs show better stability in that sense. This suggests that the anisotropic learned features of conventional models only get activated when the input is observed in a specific orientation. On the shown examples (Section 3.5.1), the *SE(2)* models are more robust to the input orientation since their *SE(2)* structure guarantees the features to be expressed in multiple orientations. We also see that *SE(2)* models with a limited angular resolution can yet produce some variance for rotation angles lower than this resolution. This is also supported by the fact that higher performances were obtained for the experiments that compare *SE(2,4)* models to *SE(2,8)* models.

Still, variation of performances for these models was also observed when the input was rotated out of the pixel grid. We explain this limit from the approximation errors caused by two of the operators we used, and that have a weaker rotation equivariance property. First, the interpolation-based computation of the rotated kernels can cause small variations in the output when the input is rotated. Second, the pooling operators are not rotation equivariant by construction (since they lay on fixed down-sampled versions of the pixel grid), and so are another source of error.

In conclusion, we proposed a framework for *SE(2)* group-convolutional network and showed its advantages for histopathology image analysis tasks. This framework enables the learned models to be invariant to the natural roto-translational symmetry of histology images. We showed that G-CNNs models whose representation have a *SE(2)* structure yield better performances than conventional CNNs and our experiments suggest the ability of G-CNNs models to fully exploit the data amount of large datasets. Our results suggest the existence of a trade-off between network capacity and the chosen angular resolution of the *SE(2,N)* operators. We chose to experiment with light-weighted shallow model architectures in order to clearly show benefits of *SE(2,N)* equivariance: such light-weighted shallow model architectures allow for fair and transparent comparisons (where we control and fix the overall network capacity, see Table 3.1,3.2,3.3). The proposed framework can also be applied to more heavyweighted and deeper models via the replacement of conventional $\mathbb{R}^2$ convolutions by *SE(2,N)* convolutions, but this is beyond the scope of this article and is left for future work. Likewise, the use of more sophisticated data augmentation strategies that do not involve rotating the images can still be beneficial in practice. Other directions for future work include further analysis of the relationship between the newly introduced architecture-related hyper-parameters and their effect on model performances, as well as studying other prior structures that can improve model stability to other families of input transformations.

Chapter 4
Inferring a third spatial dimension from 2D histological images

This chapter is based on:

M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, “Inferring a third spatial dimension from 2D histological images”, *in Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 586-589.

4.1 Introduction

In clinical context, pathological diagnosis and prognosis commonly results from the analysis of bright-field microscopy images of histological slides. These 2D images are obtained by transmitting light through the histological specimens, stained beforehand, in order to attenuate light and produce contrast. To quantify biomarkers of interest in 2D images, pathologists rely on their experience and knowledge of the 3D context of the objects they observe, when 3D microscopy techniques are not considered.

Taking inspiration from the image formation process of bright-field microscopy, we propose a method to infer a realistic decomposition of hematoxylin and eosin (H&E) stained histological slides along the axis of their thickness (z-axis), resulting in 3D images. The decomposition of a given histological image is achieved by generating a volume of its underlying stain concentrations, such that new images obtained by simulating transmitted light along other directions are realistic according to a trained discriminative deep learning model.

This study is motivated by the recent developments in deep generative models [113], in particular for generating biological microscopy images [114]. In Gadelha *et al.* [115], the authors showed that it is possible to train a generative adversarial network to infer 3D volumes from 2D training images only, without having to rely on 3D training data. Likewise, our method trains a discriminator from 2D training images only, but can generate 3D volumes that correspond to the decomposition of 2D images, and therefore does not require a generator drawing samples from a latent space.

The proposed algorithm can be seen as generating realistic 3D scenarios for the 2D observed scenes. As an example of a possible application, the generated 3D volumes can be used for data augmentation as they allow to create new “views” of the same data. Generalization of deep learning models is a known problem in automated histopathology image analysis, and new augmentation methods can help improving generalization [30]. The 3D information inferred by our method can also be used for analysis by synthesis strategies [116], to improve histopathology image analysis models, as it is a way to include the prior that processed objects have a 3D structure.

4.2 Method

Histological images can be modeled as a set of stain concentrations at every pixel location [61] as illustrated in Figure 4.1. Thus, our method aims at solving the inverse problem of estimating the volume of stain concentrations that produced the original histological image, for a chosen model of light absorption. We hypothesize that decomposition in depth is possible since the thickness of the histological specimens is of the order of the image resolution. Such a volume is generated under two constraints:

(C1) the reconstruction of the original image must be possible from the estimated volume, and (C2) new images produced from the volume must be realistic.

4.2.1 Model of Stain Concentration Volume

The RGB pixel intensities can be modeled according to the Beer-Lambert law of light absorption [61], such that the image intensity at each pixel location (x, y) can be decomposed as $I_c(x, y) = I_0 \exp\left(-\mathbf{A}_{c,*}\mathbf{C}(x, y)\right)$ with $c = 1, 2, 3$ the color-channel index, $\mathbf{A} \in [0, +\infty]^{3\times2}$ the matrix of absorption coefficients specific to the current image, and $\mathbf{C}(x, y) \in [0, +\infty]^2$ the H&E stain concentrations. We used the method of Macenko *et al.* [62] to achieve unsupervised staining unmixing of the images.

Based on the same model, the stain concentrations can be discretized along the z-axis in N parts, such that $\mathbf{C}(x, y) = \sum_{z=0}^{N-1} \mathcal{C}(x, y, z)$.

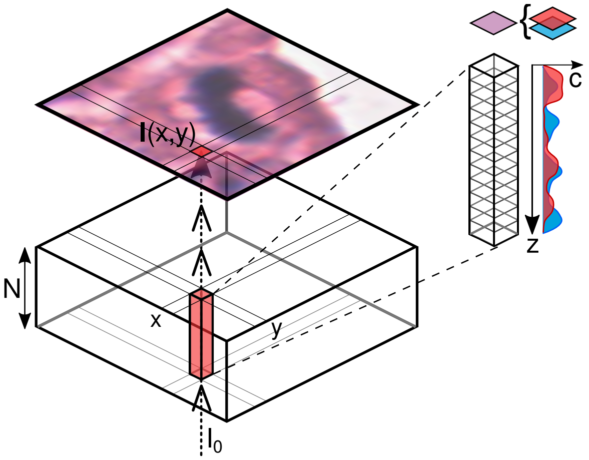


Figure 4.1: Decomposition of the estimated stain concentration values of a digital slide along the z-axis at (x, y).

The constraint (C1) can be enforced by reducing the problem to finding the vectors $\mathbf{V}(x, y, z) \in [0, 1]^2$, with $\mathcal{C}(x, y, z) = \mathbf{C}(x, y) \odot \mathbf{V}(x, y, z)$ and $\sum_{z=0}^{N-1} \mathbf{V}(x, y, z) = [1, 1]^\top$ describing how the concentrations $\mathbf{C}(x, y)$ are distributed along the z-axis (the operator $\odot$ is the element-wise multiplication).

4.2.2 Simulation of Transmitted Light

For a given volume of concentrations, new images can be generated by simulating transmitted light from different directions, using the same model of light absorption. In particular, new projection images $I^{proj}_{x=x_0,c}$ and $I^{proj}_{y=y_0,c}$ are generated by simulating light transmission along the x-axis and y-axis through the slices $x \in [x_0, x_0 + N - 1]$ and $y \in [y_0, y_0 + N - 1]$, as shown in Figure 4.2. The pixel intensities of these images are expressed in equation (4.1) as the sum of stain concentrations in the direction of projection.

$$\begin{aligned} I^{proj}_{x=x_0,c}(y,z) &= I_0 \exp\left(-\mathbf{A}_{c,*} \sum_{x=x_0}^{x_0+N-1} \mathcal{C}(x,y,z)\right) \\ I^{proj}_{y=y_0,c}(x,z) &= I_0 \exp\left(-\mathbf{A}_{c,*} \sum_{y=y_0}^{y_0+N-1} \mathcal{C}(x,y,z)\right) \end{aligned} \tag{4.1}$$

N is carefully chosen such that the pixel resolution in the xz-slices and yz-slices is the same as in the original xy-plane.

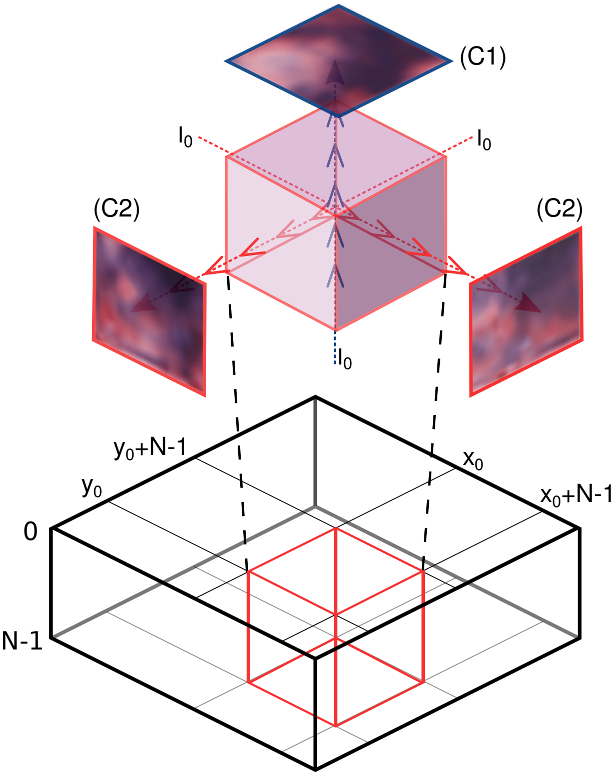


Figure 4.2: Illustration of an inferred concentration volume block of size $N \times N \times N$ pixels. (C1) is respected by enforcing the z-projection to reconstruct the original image patch. (C2) requires the x/y-projections, obtained by simulated transmitted light (red arrows), to be realistic.

4.2.3 Realism Constraint

A convolutional neural network can be trained to discriminate "fake" generated projection images that result from an underlying unrealistic concentration volume and "real" images that are assumed to be the result of realistic volume of concentration distributions.

For a given image patch, 3D volume inference starts from a 4D tensor $\mathbf{V}$ initialized with uniform concentration distributions. Then, the trained discriminative model (discriminator) can be used to update $\mathbf{V}$ by gradient descent, so that the generated projections of the updated volume appear slightly more realistic. The gradient of the loss of the discriminator with respect of the input is computed via back-propagation.

This update process (Figure 4.3) is iterated until convergence: when the discriminator classifies the generated projections as realistic with small error.

This image generation approach via optimization of the loss function of a neural network is similar to the methods developed in [117], [118], and plays a role comparable to the generator of standard generative adversarial networks [113] in the way how generated images are used as input to a discriminator.

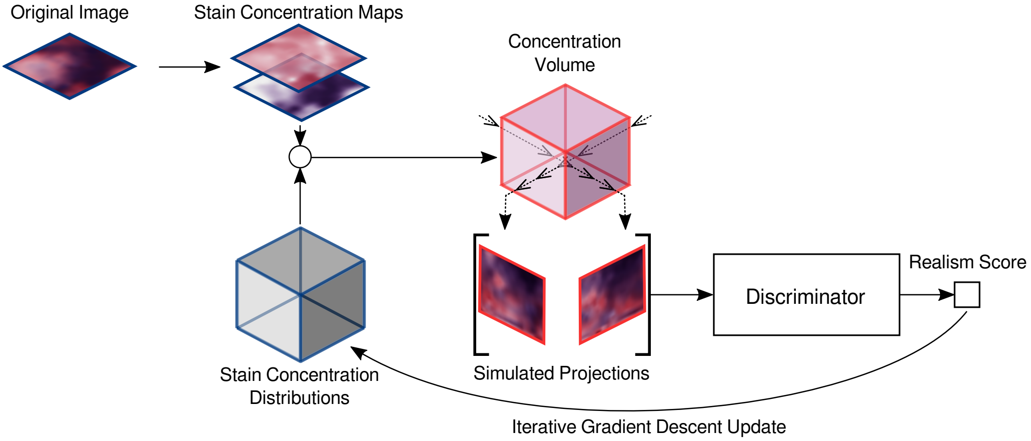


Figure 4.3: Iterative process of generating a volume of concentrations constrained by an original image. The stain concentration volume is updated by gradient descent in order to produce projection images that "fool" the fixed trained discriminator.

4.2.4 Discriminator Training

The discriminator is trained using two sets: a set of random real image patches S_{real}, and a set of adversarial examples S_{adv} that are generated during training, using the projections computed with (4.1), from previous states of the trained model.

The training procedure alternates between two steps. First, the current state of the model is used to infer volumes from real images via gradient update using the process presented in Section 4.2.2, and x/y-axis projections produced from this volume are added to S_{adv}. Secondly, a batch of image patches balanced between samples of S_{real} and S_{adv} is used to train the discriminator. Images in S_{adv} are sampled according to their misclassification probability such that the model learns from the “fake” generated images that are the most realistic and that are more challenging to classify.

4.3 Experiments and Results

4.3.1 Dataset

We used the high power field images of H&E stained slides of the public AMIDA13 dataset [57] for the experiments. 232 images of size 2000×2000 pixels from 8 different breast cancer cases were used for training and the remaining images were used to generate test examples.

4.3.2 Discriminator Architecture and Training Procedure

We implemented the discriminator that can classify input patches as “fake” or realistic as a 6-layer convolutional neural network. The network takes 24×24 image patches transformed to H&E concentration maps as input. Kernels of size 3×3, batch normalization, average-pooling and leaky ReLU non-linearities were used throughout. The network was trained by minimizing the cross-entropy loss using the Adam optimizer.

4.3.3 Generative Process and Extension to Large Images

We set the z-axis discretization to 24 pixels as we considered 6 micrometers as the maximum thickness of a tissue slice, in which case the z-axis pixel resolution of the inferred volumes can be the same as in the xy-plane (0.25 micrometers).

The discriminator, as such, can only infer volumes from images of size 24×24. To overcome this limitation, volumes of larger images can be inferred by optimizing overlapping $24 \times 24 \times 24$ sub-volumes in parallel. This solution was used to produce stain concentration volumes from 64×64 images.

The generated projections presented in Figure 4.4 indicate that the optimization process is able to distribute the stain concentrations of unseen images across the z-axis, and is able to create new tissue structures that are realistic for the trained discriminator.

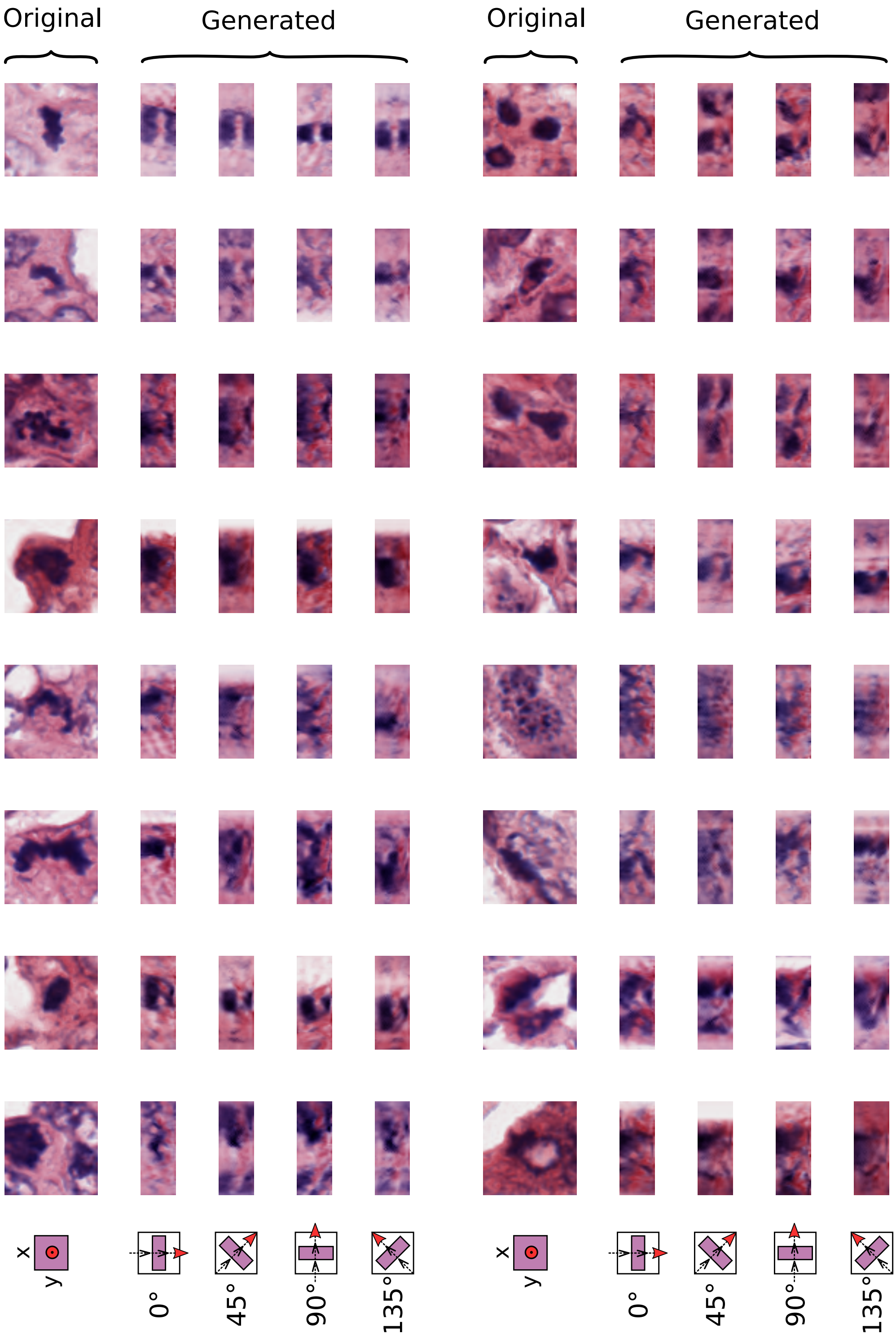


Figure 4.4: Examples of projection images from generated volumes of stain concentration. The first row of each block shows the real image patches the volumes were inferred from. The other rows show the projections obtained by simulating light transmission in different oriented slices as indicated in the left column. The top block shows results on mitotic figures that were annotated by expert pathologists, and the bottom block includes non-mitotic figures only.

4.4 Discussion and Conclusions

We proposed a method for inferring the 3D structure of 2D histological images. The method showed good qualitative performance when applied to an image dataset of mitoses and non-mitosis objects extracted from breast cancer histology slides. Although the volumes generated by our method cannot be considered as representing the actual tissue structure, the generated projections can still be considered as a likely scenario and thus used as a data augmentation tool.

In addition to being driven by the image formation process of bright-field microscopy, our method has the property that the generated images are directly produced from the available data, the same way transformation-based augmentation methods work. In contrast, generators drawing inputs from a latent space, such as generative adversarial networks, do not have this property.

Directions of future work include, further research to assess the realism quality of the generated images, and application of the generated 3D representation for data augmentation.

Chapter 5
Capturing Phenotypic Variations via Unsupervised Representation Learning

This chapter is based on:

M. W. Lafarge, J. C. Caicedo, A. E. Carpenter, J. P. Pluim, S. Singh, and M. Veta, "Capturing single-cell phenotypic variation via unsupervised representation learning", *in Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, vol. 102, 2018, pp. 315–325.

5.1 Introduction

Microscopy images provide rich information about cell state. Image-based profiling—an approach where images of cells are used as a data source—is a powerful tool with several applications in drug discovery and biomedicine [119].

Cell samples are treated using chemical or genetic perturbations, then stained using fluorescent markers, and imaged under a microscope. Image-based profiles of these genes or compounds are created by summarizing the single-cell level information extracted from these images. When executed using high-throughput technologies, this framework can be used to generate profiles of tens to hundreds of thousands of perturbations.

Creating profiles that accurately capture variations in cellular structure is an open problem [119]. Central to this problem is the task of generating representations of single cells, which can then be appropriately summarized into a profile representing the population (e.g. as the mean of the individual cell representations). In recent years, several methods have been proposed for generating single cell representations, spanning both feature engineering approaches [120], as well as feature learning using deep neural networks [121]–[125]. While the resulting profiles perform well in downstream analysis, none are able to provide much biological insight into what cellular structure variations are important for discerning phenotypes (i.e. visible appearance). This lack of insight hinders a better understanding of what drives similarities or differences between perturbations.

Recently, generative adversarial networks (GANs) were shown to learn feature representations [126], while also to synthesize cell images to help biologists visualize salient phenotypic variations. However, while the images generated were highly realistic, the accuracy of resulting profiles was relatively poor, and a direct reconstruction from the learned representations was not possible.

Here, we propose using an adversarial-driven similarity constraint applied to the standard variational autoencoder (VAE) framework [127] that addresses these limitations: (1) VAEs enable direct reconstruction given a feature representation, (2) our proposed model is demonstrably better in learning representations for profiling applications, and (3) our proposed training procedure allows higher quality reconstructions than standard VAEs, making the visualizations comparable with previous GAN models.

By proposing a novel training procedure for learning representations of single cells, we provide researchers a new tool to match cellular phenotypes effectively, and also to gain greater insight into cellular structure variations that are driving differences between populations, offering insights into gene and drug function.

5.2 Related Work

Image-based profiling measures multiple phenotypic features of single cells to characterize the effect of drugs or the function of genes. The phenotypic features can be obtained by engineering representations that capture known relevant properties of cell state, such as cell size. Previous studies using feature engineering approaches demonstrate that profiles generated using standard feature sets in bioimaging software (e.g. CellProfiler [128]) are successful in grouping compounds based on mechanism-of-action [120], [129]–[132], grouping genes into pathways [133]–[135], predicting genetic interactions [136]–[139], and several other applications [119].

Deep convolutional neural networks (CNN) have been evaluated for computing cellular features using models pretrained on natural images. A deep metric network trained on a large collection of consumer images was evaluated [121] for predicting mechanism-of-action in the *BBBC021* benchmark dataset (used in this paper), as were CNNs trained on the *ImageNet* dataset [123]. Both gave competitive results without requiring cell segmentation or image preprocessing.

Representations can be learned directly from biological images. Multiple instance learning and supervised learning using mechanism-of-action labels directly have been used to train neural networks that process full images without segmentation [124], [140]. Given that ground truth labels are rarely available for training in high-throughput projects, other strategies that require less supervision have also been explored. Weakly supervised learning using treatment replicates has been proposed to learn single-cell feature embeddings for profiling [122], and a similar technique has been developed for full fields of view [125].

However, these approaches encode cellular features without an explicit mechanism for interpreting phenotypic variations, a major limitation for many applications in biology. Goldsborough *et al.* [126] proposed to tackle this problem using the *CytoGAN* model to generate explanatory visualizations of cell variations between two treatments, but the models do not allow direct reconstructions, and have relatively poor classification accuracy on at least one benchmark dataset (*BBBC021*).

Unlike GAN models, autoencoder (AE) models are optimized to learn embeddings that can directly produce good reconstructions, and were successfully applied on cell images [141], [142]. In particular, the variational autoencoder (VAE) framework [127] implies a constraint on the embedding that produces some desired properties: smooth embedding interpolation [142] and disentanglement of generative factors [143]. To improve the limited reconstruction quality of standard VAE models, some methods involving adversarial training were proposed [144]–[147]. Here we propose to follow the concept proposed by Larsen *et al.* [144] to address the requirements of the cell profiling pipeline [148], while allowing high-quality reconstructions from the embeddings.

5.3 Material and Methods

5.3.1 Datasets

We use the *BBBC021* dataset, a popular benchmark for image-based profiling that has been adopted in several studies, mostly for evaluating assignment of chemicals to mechanisms-of-action using the leave-one-compound-out evaluation protocol [120]. The dataset is from a high-throughput experiment performed in multi-well plates; each plate has 96 wells, and in each well, a sample of cells has been treated with a compound at a specific concentration.

The subset used in all profiling experiments, including ours, has 103 unique treatment conditions (i.e. compounds at a specific concentration) representing 12 mechanisms-of-action [149]. After treatment with a given compound, the cells were stained using fluorescent markers for DNA, F-Actin and β-Tubulin and imaged under a microscope, capturing four 3-channel images for each well, and approximately one million cells across the entire dataset. These channels are stacked and treated as RGB images by mapping DNA $\mapsto$ R, β-Tubulin $\mapsto$ G, F-Actin $\mapsto$ B.

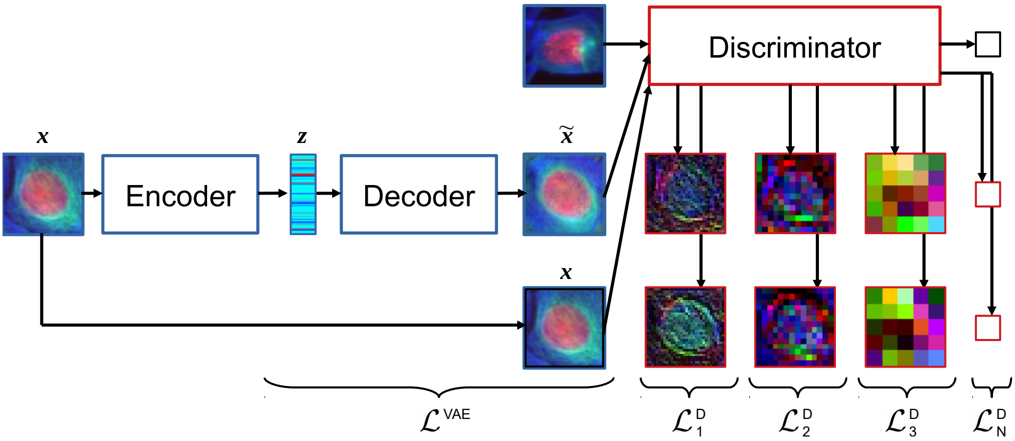


Figure 5.1: Flowchart of CNN models. The auto-encoder (blue-framed components) describes the original VAE formulation. The adversarial-driven reconstruction losses are illustrated by the representation learned by the discriminator (red-framed images).

5.3.2 The VAE framework

In this study, we are interested in methods that can directly generate low-dimensional embeddings $\mathbf{z}$ and reconstructions $\tilde{\mathbf{x}}$ of given input images $\mathbf{x}$. Therefore we chose the VAE framework as a baseline [127]. VAE models consist of an encoder convolutional neural network (CNN) that models an approximation of the posterior $q_\phi(\mathbf{z}|\mathbf{x})$ on the latent $\mathbf{z}$, parameterized by ϕ, and a decoder CNN that models the likelihood of the data $p_\theta(\mathbf{x}|\mathbf{z})$, parameterized by θ. The model is then optimized by maximizing a lower bound on the marginal log likelihood of the data $\mathcal{L}^{\text{VAE}}(\mathbf{x}, \mathbf{z}; \theta, \phi) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta \cdot D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})\right)$, with $p(\mathbf{z})$ a defined prior distribution to constrain the embeddings, D_{KL} the Kullback-Leibler divergence, and β an hyper-parameter controlling the strength of this constraint [127], [143].

5.3.3 Transition from Pixel-Wise to Adversarial-Driven Reconstructions

The limited reconstruction quality of standard VAE models can be explained by the pixel-wise reconstruction objective related to the Gaussian observation process modeled by $p_\theta(\mathbf{x}|\mathbf{z})$ [144], [150].

Learned similarity As proposed by Larsen *et al.* [144], we define a discriminator CNN $\mathcal{D}$ with parameters χ that is trained to classify real images $\mathbf{x}$ from independent reconstructions $\widetilde{\mathbf{x}}$. The discriminator outputs the probability for the input to originate from the distribution of real images and is optimized via minimization of the binary cross-entropy.

The activations resulting from the hidden layers of the discriminator $D_i\left(\mathbf{x}\right)$ and $D_i\left(\widetilde{\mathbf{x}}\right)$ are used as additional, synthetic Gaussian observations, with i the layer indices. These observations are drawn from $p_\chi\left(D_i(\mathbf{x})|\mathbf{z}\right)$, modeled as normal distributions with means $D_i(\widetilde{\mathbf{x}})$ and identity covariances. We thus define additional reconstruction losses $\mathcal{L}_i^D(\mathbf{x}, \mathbf{z}; \theta, \phi, \chi) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\chi(D_i(\mathbf{x})|\mathbf{z})\right]$ for every hidden layer i of the discriminator. Figure 5.1 illustrates how the different losses of the framework arise in the full model pipeline.

Progressive Training We conjecture that the reconstruction term in $\mathcal{L}^{\text{VAE}}$ should not be discarded and that the additional losses $\mathcal{L}_i^D$ can be all used to compensate the limited reconstruction ability induced by $\mathcal{L}^{\text{VAE}}$, as opposed to the formulation of Larsen *et al.* [144]. Therefore, we propose to use $\mathcal{L}^{\text{VAE}} + \mathcal{L}^{\text{D}}$ as the full objective for the encoder and decoder with $\mathcal{L}^{\text{D}} = \sum_i \gamma_i \cdot \mathcal{L}_i^D$, and (γ_i) a set of parameters to control the contribution of each reconstruction loss.

For stability purposes when dealing with adversarial training Karras *et al.* [151], we chose to define the $\gamma_i(t)$ as a function of the iteration step t. By defining $\gamma_i(t) = \min(1, \max(0, t/T - i))$, we induce a progressive training procedure, such that the abstraction levels of the discriminator contribute sequentially to $\mathcal{L}^{\text{D}}$. T is thus the hyper-parameter defining the period between two losses $\mathcal{L}_i^D$ and $\mathcal{L}_{i+1}^D$ to contribute to the final objective.

5.3.4 Model Architectures

The encoder takes image patches of size 68×68 as input, and estimates the mean and standard deviation of the Gaussian posterior, that allow sampling an embedding of size 256 using the reparameterization trick Kingma *et al.* [127].

The encoder, decoder and discriminator have four convolution layers with filters of size 5×5, with an additional 1×1 layer for the last layer of the decoder and an additional fully connected layer for the discriminator. Leaky Rectified Linear Units (coefficient 0.01) and max-pooling/up-sampling layers were used throughout the CNNs, except for the last layer of the discriminator, which is activated by a sigmoid.

The decoder is a mirrored version of the encoder, by using transposed convolutions followed by 2×2 up-sampling layers. Batch normalization (BN) layers were used throughout the CNNs, and the BN moments for the discriminator were computed only using batches balanced with input reconstructions and independent real images. The implementation of the model is available at https://github.com/tueimage/cytoVAE.

5.4 Experiments and Results

5.4.1 Experiments

We investigated three variations of the proposed model for comparison purposes. We trained standard AE and VAE models by setting the parameter β to 0.0 and 1.0 respectively while excluding $\mathcal{L}^D$ from the full objective. The proposed model (VAE+) was trained using the full objective (see Section 5.3.3), β was set to 2.0 to compensate the additional reconstruction losses, and T was set to 2500 iterations.

Mini-batches were built by sampling a random image patch from each treatment of the dataset. Every channel of the image patches was normalized by its maximum intensity. Two independent mini-batches B_1 and B_2 were used at every iteration: B_1 was used to compute $\mathcal{L}^{\text{VAE}}$ through the encoder-decoder, B_2 was paired with the reconstructions of B_1 to train the discriminator. Finally, B_1 and its reconstructions were used to compute $\mathcal{L}^D$.

We used the *Adam* optimizer to train the encoder and decoder (learning rate 0.001; momentum 0.9), and Stochastic Gradient Descent with momentum (learning rate 0.01; momentum 0.9) to train the discriminator. All the convolutional weights were regularized with weight decay (coefficient 0.0001). Training was stopped after $40,000$ iterations.

5.4.2 Creating Profiles and Classifying Compounds

Given images of cells treated with a compound (the input), the challenge in the *BBBC021* dataset is to predict the mechanism-of-action (the label) of the compound. Centers of each cell were precomputed using CellProfiler [128] and were used to extract patches. Representations of these patches were generated using the trained models. Given a representation per cell, a profile for each well was computed as the average of all the cells in that well. Next, the profile for each unique treatment (a compound at a specific concentration) was computed by computing the median of all wells with that treatment. Treatments were classified using 1-nearest-neighbors, using one of two hold-out procedures as proposed by Ando *et al.* [121]: (1) Not-Same-Compound (*NSC*), where all profiles of the same compound (regardless of concentration) were held out, and (2) Not-Same-Compound-and-Batch (*NSCB*), where in addition to *NSC* constraints, profiles from the same experimental batch where held out.

NSCB indicates how sensitive the profiling method is to variations across experimental batches; better *NSCB* performance indicates better resilience to batch variations. Ando *et al.* [121] transformed the profiles on a given plate using a whitening transform learned from the control wells on that plate, which improved *NSCB* performance; we tested this procedure (indicated by “Whitened” in Table 5.1). Further, Rohban *et al.* [135] created profiles by summarizing using standard deviations as well as means; we tested this approach (indicated by “Mean+S.D.” in Table 5.1).

5.4.3 Results

Classification Performances

The proposed VAE model (VAE+ in Table 5.1) significantly outperforms the best GAN-based models (68% *NSC*; *NSCB* unavailable), which is the only model to our knowledge that can provide reconstructions. Further, whitening consistently improves accuracy across all configurations where mean is the summary statistic, and for some where both mean and S.D. are used as summary statistics. The VAE+ model, with mean + S.D. summaries followed by whitening (last column) performs similarly to the best performing classical approach (90% *NSC*; 85% *NSCB* Singh *et al.* [152]). While none of these models, including VAE+, achieve classification performance as high as the best performing deep-learning-based model (96% NSC; 95% *NSCB* Ando *et al.* [121]), they nonetheless provide valuable insight (discussed below) into the variations in cellular morphologies that underlie the similarities and differences between the treatment conditions. Finally, we observe that the AE model implemented performed very similarly to VAE+. The VAE+ reconstructions are however superior to AE, making the former overall better suited for profiling applications.

Visualizing Structural Variations in Cell Phenotypes

The proposed VAE+ model produces the most realistic images (Figure 5.2); both AE and VAE images are consistently blurrier than VAE+ images. Similar to Goldsborough *et al.* [126], we assessed the quality of reconstructed images by presenting three ex-

Table 5.1: Classification Accuracy of the compared models. Mean result ± standard deviation across 3 repeated experiments with random initialization and random input sampling. The numbers in bold indicate the method-summarization combination that was best performing for each hold-out procedure (NSC and NSCB).

	Method	Mean	Mean +Whitened	Mean+S.D.	Mean+S.D. +Whitened
NSC	*VAE+*	**90.6** ± 1.5	90.3 ± 1.0	**92.2** ± 1.7	**92.9** ± 2.4
	VAE	83.5 ± 1.0	80.6 ± 4.4	90.9 ± 1.1	87.1 ± 0.6
	AE	87.6 ± 2.0	**92.2** ± 1.0	90.3 ± 0.0	92.5 ± 0.6
	Ando et al.	N.A	96.0	N.A	N.A
	Singh et al.	90.0	N.A	N.A	N.A
NSCB	*VAE+*	71.0 ± 1.2	76.1 ± 1.1	72.5 ± 2.3	**82.2** ± 2.6
	VAE	68.8 ± 0.6	69.9 ± 5.1	74.6 ± 0.6	71.0 ± 0.6
	AE	**75.0 ± 2.0**	**79.0** ± 0.6	**76.8** ± 0.7	80.8 ± 1.7
	Ando et al.	N.A	95.0	N.A	N.A
	Singh et al.	85.0	N.A	N.A	N.A

pert biologists with 50 real cell images and 50 cells reconstructed using VAE+. The cells were balanced across the available treatments, including controls and the biologists were blinded with respect to this treatment information. Images were randomly shuffled and presented to experts to assess whether each cell was real or synthetic. On average, 40.7% of the time the synthetic cells were realistic enough to deceive the experts into labeling them as real, compared to 30% previously reported with GANs [126].

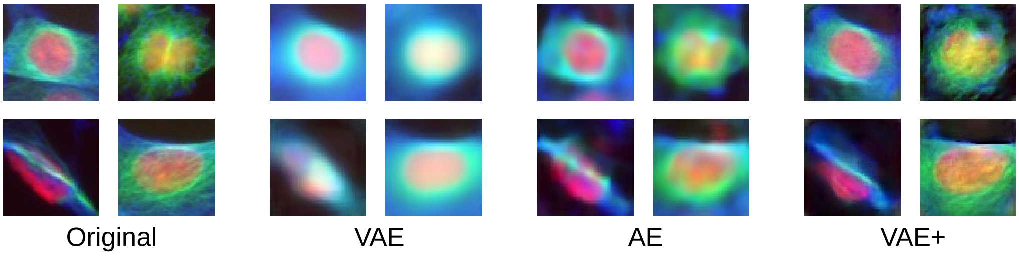


Figure 5.2: Comparison between original images of four randomly sampled single cells, and their reconstructions produced via different auto-encoders.

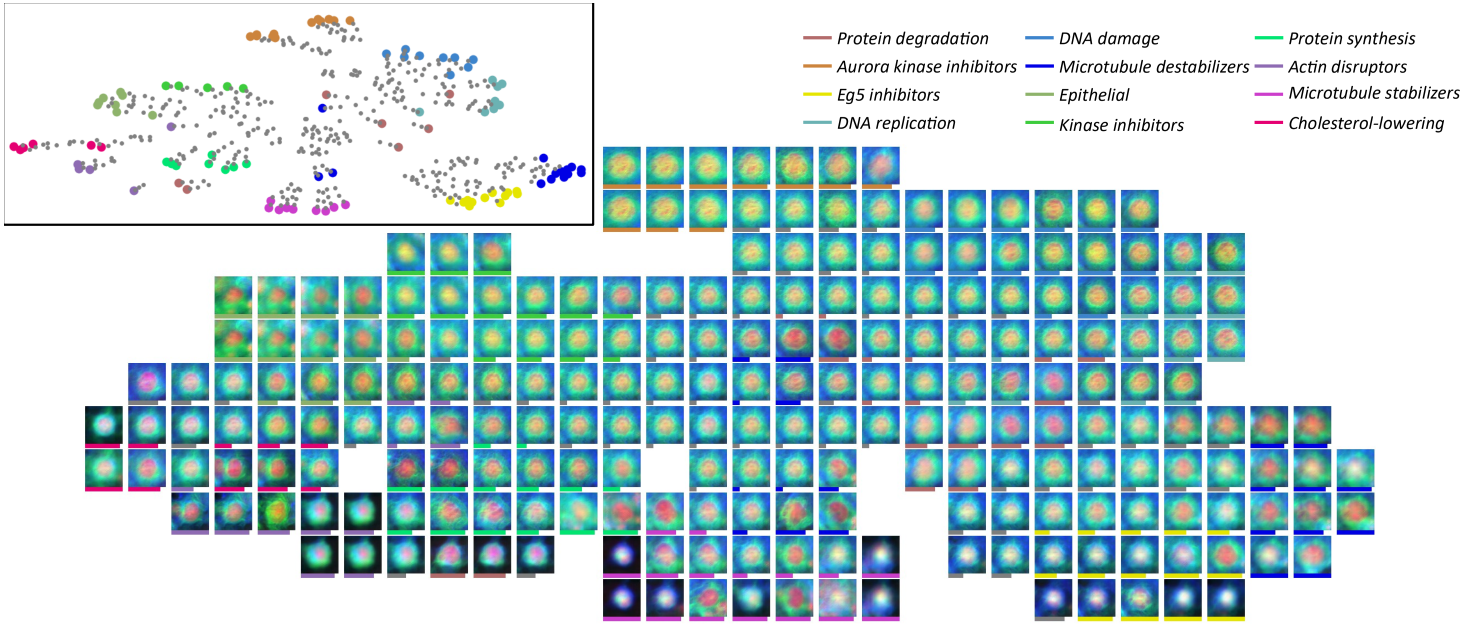


Figure 5.3: t-SNE visualization [29] of the learned representation. Feature representations of single-cells from each individual treatment (in color) and their interpolations with the control cell population (gray) were embedded. The displayed images were generated using interpolated embeddings based on a grid sampling procedure in the space spanned by the t-SNE embedding. The learned representation forms clusters that correspond to the ground-truth mechanisms-of-action.

The ability to interpolate between real cells from different treatment conditions and produce realistic images is powerful tool to visualize how a compound affects cellular structure (Figure 5.4 and Figure 5.3). Compounds from different mechanisms induce visually distinct phenotypes. Interpolating between a control cell and a treated cell presents a hypothetical path in phenotypic space that the cell may have taken to arrive at the observed (target) state. Verifying these hypotheses would require further followup experiments. Regardless, these visualizations give valuable insight into how each compound is affecting cellular structure. For instance, an actin disrupting chemical (*cytochalasin D*) appears to make the cells smaller, with both actin and tubulin condensing more tightly and symmetrically around the nucleus. A cholesterol lowering chemical (*simvastatin*) has a similar effect but makes the tubulin more asymmetric. Both results match expectations and inspection of real images.

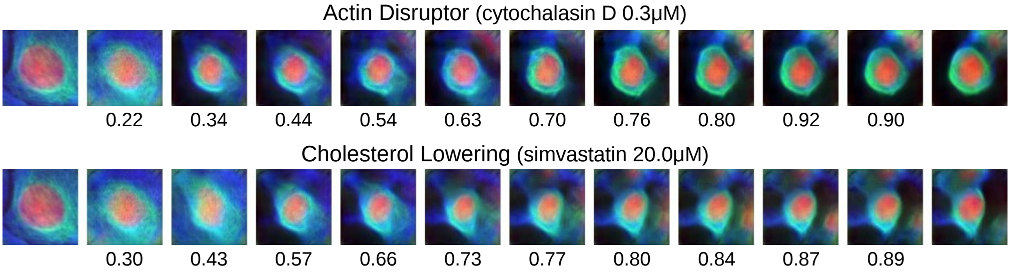


Figure 5.4: Translation in VAE+ latent space of a control cell (left) to target cells (right) corresponding to compounds with different mechanisms-of-action. The target cell is the one closest to the mean of the compound. Each interpolation step is a shift of features with highest absolute difference w.r.t. the target features. Cosine similarity between the embedding of an image and its target is shown below each.

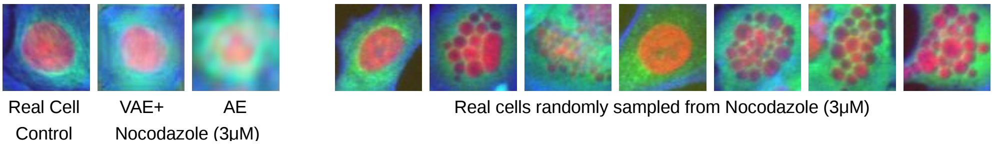


Figure 5.5: VAE+ captures β-tubulin structure better (less blurry) and correctly identifies *Nocodazole* as a microtubule destabilizer. AE incorrectly classifies it as an actin disruptor. However, neither captures the fragmented nucleus phenotype seen in a fraction of cells' real images (right).

However, we noticed one interesting anomaly when exploring a case where VAE+ correctly classified a drug and AE did not (recall their overall classification accuracies across all classes are similar (Table 5.1)). For the drug *Nocodazole*, a known microtubule destabilizer, AE yields a blurry reconstruction of tubulin while VAE+ yields a more accurate texture (Figure 5.5). Upon inspection of randomly sampled cell images, however, it becomes clear that neither representation is able to capture the distinctive fragmented nucleus phenotype caused in some cells by *Nocodazole*. We suspect that the selection of the target cell is thus a crucial choice in the proposed strategy, particularly when a population of cells shows two very distinct types of appearances.

5.5 Discussion and Conclusions

We proposed an auto-encoding approach competitive with other unsupervised learning approaches while overcoming the challenge of high quality reconstructions.

We introduced adversarial-driven representation learning for the problem of image-based profiling using a straightforward extension of the VAE framework, by proposing a generic method inline with the work of Larsen *et al.* [144]. Some methods are other plausible solutions for this task, such as Adversarially Learned Inference Dumoulin *et al.* [146] and are worth investigating for future work.

The unsupervised training context explains the limited classification performances reported here, but could be improved when combined with more effective approaches (weakly/fully supervised training).

This model offers researchers a powerful tool to probe the structural changes in a cell induced by genetic and chemical perturbations, or even disease states. This is a step towards filling the gap of interpretability in image-based profiling approaches: to reveal not just which perturbations are similar or different, but also to provide clues about the underlying biology that makes them so. We identified room for improvement in capturing phenotypes for very heterogeneous cell populations. The proposed strategy may be applied to other domains in biomedical imaging that require capturing phenotypic variations, particularly detecting, understanding, and reversing disease.

Chapter 6
Orientation-Disentangled Unsupervised Representation Learning

This chapter is based on:

M. W. Lafarge, J. P. Pluim, and M. Veta. "Orientation-Disentangled Unsupervised Representation Learning for Computational Pathology", (2020) [In review, arXiv preprint arXiv:2008.11673].

6.1 Introduction

Dimensionality reduction is an efficient strategy to facilitate the analysis of large image datasets by representing individual images by a small set of informative variables, which can be used in place of the original images. Unsupervised learning methods can be used to obtain such an informative latent representation from a given dataset without the need for expert annotations. For this purpose, popular unsupervised learning frameworks such as the Variational Auto-Encoder [127] or flow-based approaches [153] can be used to model a joint distribution between an image dataset and a set of latent generative factors. As these frameworks provide a posterior distribution over a space of latent variables, they enable the estimation of the latent factors of new previously unseen images, that can then be used for any subsequent task.

However, irrelevant factors that affect the appearance of images but are independent of the factors of interest can get *entangled* in the learned representation [25], [28], [143]. These irrelevant factors can be treated as *nuisance variables* that affect the learned representation in an unpredictable way [154], consequently perturbing any *downstream analysis*[1] performed on a distribution of generative factors. Therefore, there is a need for *disentangling* such nuisance variables from the informative generative factors of interest.

In computational pathology, these nuisance variables are known to affect the generalization power of machine learning models. They affect the appearance of the images across slides, scanners and hospitals and can be associated with the inevitable variations in tissue slide preparation and scanner-dependent digitization protocols.

In a supervised learning context, strategies were developed to filter-out such irrelevant factors from the learned representation; popular methods applied in computational pathology include: staining normalization [31], staining/style transfer [45], [155], [156], data augmentation [36], domain-adversarial training [30], [111] and rotation-equivariant modeling [63], [157].

In this paper, we focus on a specific generative factor that can be considered as a nuisance variable in some specific tasks: the orientation of individual image patches. In digital pathology, the orientation of localized image patches in a dataset of WSIs is arbitrary in the sense that tissue structures are likely to be observed in any orientation, as opposed to natural images or organ-level medical images for which the orientation of the imaged objects is typically not uniformly distributed.

We propose an unsupervised learning framework to model a partitioned latent space of generative factors in which specific independent latent variables either code for oriented or non-oriented (isotropic) morphological components of histopathology images.

[1] We refer to *downstream task/analysis* to express any task/analysis performed on an image dataset for which a learned representation is used in place of original images.

Motivation We identified several points that motivate the development of methods to handle nuisance variables for computational pathology in an unsupervised learning context:

- Using an informative representation in place of large and complex images can reduce the computational cost and facilitate training of subsequent task-specific models. In particular, such a representation can be used to directly process Whole Slide Images (WSIs) via patch-based compression [19] or to represent bags that consist of a high number of image patches in two-stage multiple-instance-learning frameworks [158]. By removing irrelevant factors from the representation, such existing frameworks can be further improved.
- A representation learned without supervision can better conserve the extent of the morphological information of tissue images making it suitable for a wide range of potential downstream tasks. This is opposed to using the representation of a supervised model that potentially discards information that is irrelevant for the task for which it was trained, but that might be relevant for other downstream tasks.
- Latent variable models equipped with a generative component enable visual inspection of the individual learned factors. This can support the interpretation of a model, as a tool to gain insights into the morphological factors that are predictive for a given downstream task.

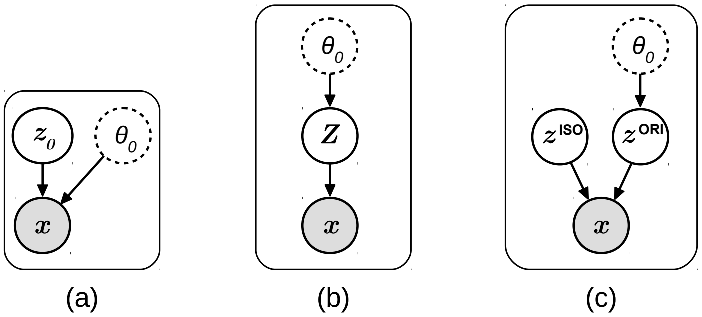


Figure 6.1: Bayesian networks of three latent variable models in which a hidden nuisance variable θ_0 is involved. (a) Classical generative model; $\boldsymbol{z}_0$ and θ_0 are independent sources of the observed images $\boldsymbol{x}$. (b) Chain-structured model; the images $\boldsymbol{x}$ are generated from intermediate latent variables $\boldsymbol{Z}$ that depend on θ_0. (c) Proposed disentangled model; the images $\boldsymbol{x}$ are generated by two independent variables: $\boldsymbol{z}^{\mathsf{ISO}}$ that is independent of θ_0 and $\boldsymbol{z}^{\mathsf{ORI}}$ that encodes θ_0.

The proposed method To enable such a partitioning of the latent space, we leveraged the structure of *SE(2)*-group convolutional networks [63], [157] to build the com-

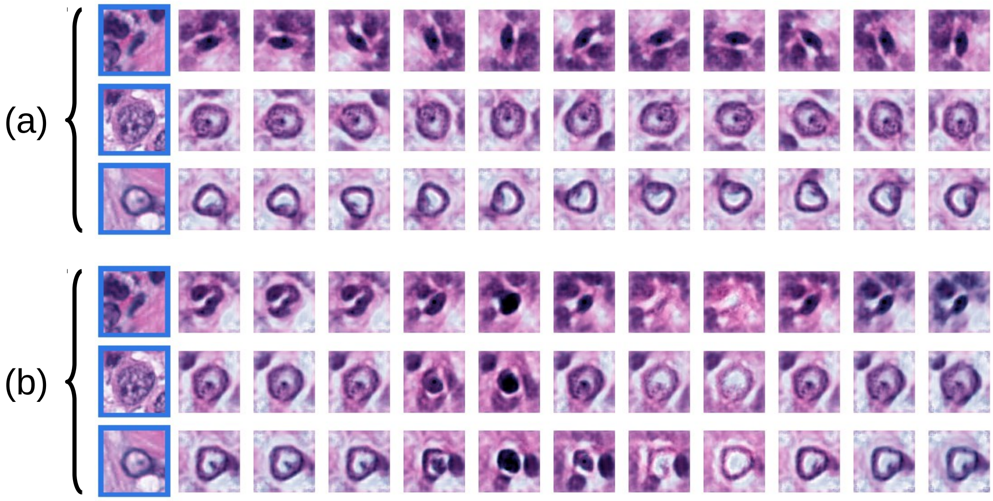


Figure 6.2: Generated images using the proposed latent variable model in which images are represented by a set of two types of variables: real-values z^{ISO} that code for isotropic components and angle variables z^{ORI} that code for oriented components. Left-most images are original and were used to estimate initial values of z^{ISO} and z^{ORI}. In (a), z^{ISO} is kept fixed and the values of z^{ORI} are sequentially incremented by a fixed angle measurement (cycle-shifted), causing a spatial rotation of the generated images. In (b), z^{ORI} is kept fixed while the values of z^{ISO} are sequentially varied causing *isotropic* morphological changes in the generated images.

ponents of a new VAE that produces a pair of rotation-equivariant and rotation-invariant embeddings of image patches. This means that instead of representing an image by a vector of scalar latent variables (z_0 in Figure 6.1 (a)), the proposed framework learns in parallel a vector of real-valued isotropic variables (z^{ISO}) and a vector of angular orientation variables (z^{ORI}, see Figure 6.1 (c)) that enable the disentanglement of the orientation information in images.

In Figure 6.2, we illustrated the orientation disentanglement property obtained with the proposed framework, in which the effect of varying each type of generative factor can be observed in the generated examples. The resulting structure of the learned representation is also illustrated in Figure 6.3, in which we show that the translation between existing datapoints in the latent space along the oriented or isotropic dimenisons causes distinctive generative effects.

The independence between the two types of variables is guaranteed by the structure of the *SE(2)*-CNN-based auto-encoder. In the spirit of the original VAE framework, we propose an extension of the objective function so as to encourage the mutual independence between all the introduced latent variables.

We made a comparative analysis of the proposed framework with a baseline VAE

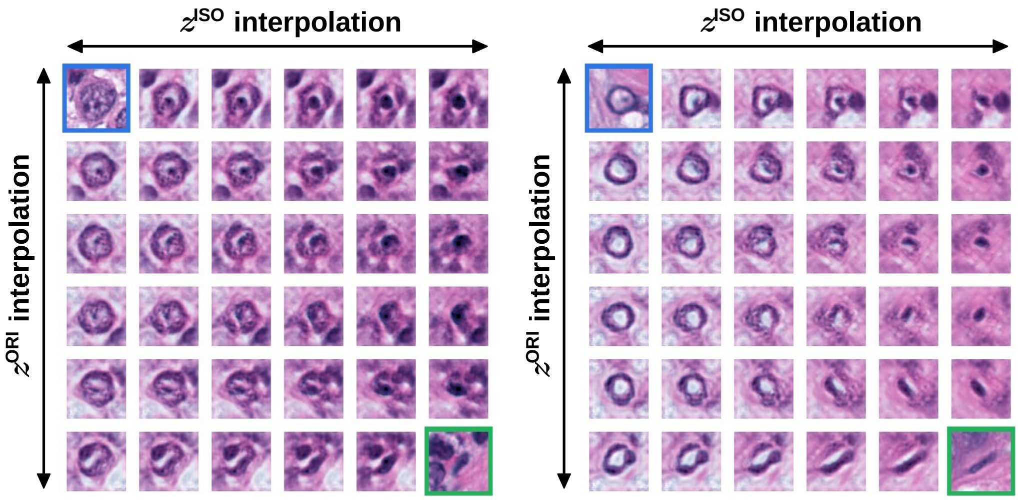

z^{ISO} interpolation
z^{ISO} interpolation
z^{ORI} interpolation
z^{ORI} interpolation

Figure 6.3: Generated image from linearly interpolated latent variables: along the horizontal axis the isotropic variables z^{ISO} are interpolated, along the vertical axis the angular components z^{ORI} are interpolated. Original images used to estimated the original values of the latent variables are framed in blue (top-left images), and green (bottom-right images).

and well-established hand-crafted measurements in the context of histopathology image analysis. We trained and evaluated the models using a dataset of nucleus-centered images extracted from histological cases of 148 breast cancer patients, exposing a large variability of nuclear morphology to be learned. We evaluated the quality of the unsupervised learned representation by training simple logistic regression models on downstream classification tasks. We compared the ability of the learned embedding to predict the pleomorphism grade and tumor proliferation grade associated to each case of a hold-out test set using multi-class ROC-AUC metrics.

Contributions

- To our knowledge, this is the first time that an auto-encoder is proposed to explicitly disentangle the orientation information in images by learning a 2-part structured representation consisting of rotation-invariant real-valued variables and rotation-equivariant angle variables.
- We propose to use *SE(2)*-structured CNNs to generate latent variables with guaranteed equivariance/invariance properties.
- We show that such an unsupervised model can quantify nuclear phenotypical variation in histopathology images and that the learned representation can be used in dowstream analysis to predict slide-level target values.

6.2 Related Work

Representation Learning for Microscopy Image Analysis Automated quantification of morphological features of single-cell images has been a paradigm for comparing populations of cells in high-throughput studies across microscopy modalities and applications [148], [159], [160].

In particular, machine learning methods were proposed to learn representation directly from image data. Transfer learning methods use the internal representation of deep learning models that were trained with an independent dataset and task [161]–[165]. Weakly-supervised and self-supervised models use the representation of deep learning models that are trained with the data at hand but optimized to solve an auxiliary pretext task [122], [125], [126], [166].

We argue that methods that rely on training a generative adversarial network and that exploit the inner feature maps of the discriminator network as a representation fall into this category as these feature maps do not necessarily correspond to generative factors. These methods rely on the hypothesis that deep learning models can learn generic features that will generalize to an external task without further assumptions. However, such generalization is not guaranteed for a difficult medical-related task whose domain is too different from the task and dataset that were used for the original training of the models. Also, these methods often rely on subsequent fine-tuning using the data at hand, and such a transferred learned representation cannot give any direct visual insight into the nature of the individual features.

Generative models and in particular latent variable models, were developed to learn the generative factors of cell images as a representation to be used in downstream tasks. These methods are based on (sparse) dictionary learning [167], sparse auto-encoding [35], [168], variational auto-encoding [169]–[171], conditional auto-encoding [142] or other auto-encoding frameworks [172]–[175].

Other Applications of Representation Learning for Computational Pathology We consider unsupervised representation learning of random patches of digital slides related work: such representations facilitated the achievement of downstream tasks [161], [176]–[195]. However, when supervision is possible, patch-based representation can be achieved in a end-to-end fashion in a multiple-instance-learning framework for a given task [158], [182], [190], [194], [196]–[202]. In [19], the authors compared different latent variable models in order to compress WSIs and enable their processing in a single run, and investigated their potential on downstream tasks. Studies on realistic generation of histopathology images [203]–[206] showed that decoder-based generative models can embed the fine-grained morphological structures of tissues in a low-dimensional latent space, which is in line with the motivation of our work.

Structured Latent Variable Models Although unsupervised latent variable modeling is appealing, Locatello *et al.* [207] and Dai *et al.* [208], supported by the work of Ilse *et al.* [209] showed that learning disentangled generative factors is not possible without constraining the learned representation. This argument justifies the limited performances of the learned representation of baseline VAE models in downstream tasks. As a solution, methods were proposed to structure the VAE latent space as a form of inductive prior: hyperspherical latent structure [210], supervised nuisance variable separation [28] or domain-wise latent space factorization [209] for example.

In the context of single-cell representation learning, Johnson *et al.* [142] proposed a structured latent space via a conditional VAE model that encourages separation of cell/nuclear shape information from sub-cellular component localization.

The framework proposed in this paper is in the direction of research of these prior works but specifically address the spatial orientation of the generative factors of cell images.

Rotation-Equivariance in Convolutional Networks Deep Learning methods were proposed to learn representations that are equivariant to the orientation of images. These methods enable learning a representation that changes in a deterministic way when the input image is rotated. In particular, group convolutional networks [63], [66], [67], [71], [73], [76], [81] extend standard CNNs by replacing the convolution operation. Advantages of using group-structured convolutional networks were shown on computational pathology tasks in a supervised training context [63], [107], [108], [110], [157], [211].

Here we leverage the structure of *SE(2)*-CNNs in an unsupervised context as a new way to structure the latent space of VAE-based models.

6.3 Datasets

To train and compare the models investigated in this study, we used one dataset for training purposes and to assess slide-level classification performances and another patch-based benchmark dataset to assess cell-level classification performances.

TUPAC-ROI We used a dataset of 148 WSIs of Hematoxylin-Eosin-stained (H&E) tissue slices of breast cancer patients. These WSIs are part of the training set of the *TUPAC16* challenge [56] and are originally provided by The Cancer Genome Atlas Network [58]. Three Regions of Interest (ROIs) were annotated by a pathologist to indicate tumor regions with high cellularity, that pathologists would typically select for cancer grading. Note that we used this subset of WSIs as it was the only subset for which ROIs were provided.

Heng *et al.* [212] provided several patient-level metrics on these cases, including molecular and genetic information as well as expert-based visual morphological assessment (mitosis grading, tubular formation, pleomorphism grading). We used the pleomorphism grade and tumor proliferation grade (discrete grades in $\{1, 2, 3\}$) associated to each WSI as a target value to evaluate the quality of the learned representations investigated in this study. In the case of tumor proliferation grading, we make the assumption that this value can be associated to cell-level patterns and their local distributions.

To reduce the inter-case staining variability, we pre-processed all images by applying the well-established staining normalization method described in [62]. We applied our internal nuclei segmentation deep learning model [157] within each ROI, and used the center of mass of the segmented instances as an estimate of the nuclei center locations. Image patches of size $68{\times}68\text{px}^2$ at a resolution of $\sim 0.25\mu\text{m/px}$ centered on these locations were extracted and constitute the effective dataset of cell-centered images we used to train and test our models. We made a training-validation-test split of this dataset (including respectively 104, 22 and 22 cases). See the supplementary material for details about the class distributions across the splits. We will refer to this refined dataset as *TUPAC-ROI*.

CRCHistoPhenotypes In order to assess the single cell-level quality of the trained models and investigate the transfer ability of the learned representation to other tissue types, we used the classification subset of the *CRCHistoPhenotypes - Labeled Cell Nuclei Data* (CRCHP) dataset provided by Sirinukunwattana *et al.* [16]. This dataset consists of $22{,}444$ localized nuclei extracted from 100 ROIs, themselves originating from WSIs of H&E stained histology images of colorectal adenocarcinomas. Cell-type labels (*epithelial, inflammatory, fibroblast, miscellaneous*) were provided for each nucleus. We made a training-validation-test split of this dataset free of any ROI-overlap (including respectively $11{,}090$, $3{,}133$ and $8{,}221$ nuclei). We resampled and cropped image patches centered at the nuclei locations so that the resolution and dimensions matched the ones of the *TUPAC-ROI* dataset. Finally, we applied the same staining normalization protocol as for the *TUPAC-ROI* dataset.

6.4 Methods

This section describes the proposed framework, first summarizing the baseline VAE framework, then presenting its expansion using *SE(2,N)*-group convolutions and finally developing how we enable disentanglement of the orientation information.

We formalize the representation learning problem in this generative modeling setting as the problem of learning the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ of the observed images $\boldsymbol{x}$ with their latent generative factors $\boldsymbol{z}$. Typically, we want to estimate the distribution that maximizes the marginal likelihood $p(\boldsymbol{x})$ of this model for a given dataset.

6.4.1 Variational Auto-Encoder

In the VAE framework [127], the likelihood of the observed images given a latent embedding $p_\psi(\boldsymbol{x}|\boldsymbol{z})$ is modeled by a decoder CNN with parameters ψ. It is assumed that the latent $\boldsymbol{z}$ are drawn from a given prior distribution $p(\boldsymbol{z})$, typically a multivariate normal distribution. By introducing an approximation of the posterior on the latent $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ modeled by a CNN encoder (parameterized by ϕ), Kingma *et al.* [127] propose to optimize ψ and ϕ by maximizing a tractable lower bound on the marginal log likelihood, as written in Eq.6.1.

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z}; \psi, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})] - \beta \cdot \mathrm{D}_{\text{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})\right)] \tag{6.1}$$

The Kullback-Leibler divergence term D_{KL}, encourages the distribution of the sampled latents to be close to the prior distribution. The β hyper-pararameter controls the strength of this constraint as introduced by Higgins *et al.* [143].

Orientation Encoding The encoder and decoder CNNs of conventional VAE models are built as a series of alternating trainable 2D convolution operations, non-linearity activation functions and down/up-pooling operations. For a given image $\boldsymbol{x} \in \mathbb{L}_2[\mathbb{R}^2]$, the encoder CNN q_ϕ generates low-dimensional embedding samples $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (of M elements) with $\boldsymbol{z} = [z_i]_{i=1}^M$ and $z_i \in \mathbb{R}$.

In the context of tissue imaging, we make the hypothesis that every image can be decomposed as a pair $\boldsymbol{x} = (\boldsymbol{x}_0, \theta_0)$ of independent variables such that $p(\boldsymbol{x}) = p(\boldsymbol{x}_0) \cdot p(\theta_0)$. With this formulation, we assume the existence of a reference distribution of images $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$ that get rotated by an angle drawn from a uniform distribution $\theta_0 \sim U(0, 2\pi)$. This uniformity assumption is logical for histopathology since tissue slices are prepared independently of their orientation and thus tissue images can be acquired in any possible orientation.

We write $\boldsymbol{x} = \mathcal{L}_{\theta_0}[\boldsymbol{x}_0]$ as the relationship between these variables, with $\mathcal{L}_{\theta_0} : \mathbb{L}_2[\mathbb{R}^2] \to \mathbb{L}_2[\mathbb{R}^2]$ the *left-regular representation* on 2D images of the rotation group *SO(2)*, parameterized by θ_0. In this notation θ_0 indicates the action of a planar rotation $R_{\theta_0} \in$ *SO(2)*, such that $\mathcal{L}_{\theta_0}[\boldsymbol{x}_0](\boldsymbol{u}) = \boldsymbol{x}_0(R_{\theta_0}^{-1} \cdot \boldsymbol{u})$, given a vector location $\boldsymbol{u} \in \mathbb{R}^2$.

From this point of view, θ_0 is a generative factor of the observed images, and so it would be expected that the model learns to isolate this factor in the learned latent $\boldsymbol{z}$. Without loss of generality, we can assume that an optimal model would learn to decompose the latent variables such that $\boldsymbol{z} = [\boldsymbol{z}_0, \theta_0]$ with $\boldsymbol{z}_0$ the subset of latent variables that are independent of θ_0.

This decomposition would imply that the posterior distribution of θ_0 is *equivariant* under the deterministic action of the *rotation group* on the image domain, and via the concatenation of rotations on the domain of orientations (written as the addition of angles modulo 2π in Eq. 6.2a). The desired (conditional) independence relationship between z_0 and θ_0 is equivalent to the posterior being *invariant* under the same group actions (see relationship of Eq. 6.2b).

$$p(\theta_0 + \theta \,|\, \mathbf{x} = \mathcal{L}_\theta[\boldsymbol{x}_0]) = p(\theta_0 \,|\, \mathbf{x} = \boldsymbol{x}_0) \ ; \forall\theta \in [0, 2\pi) \tag{6.2a}$$

$$p(\boldsymbol{z}_0 \,|\, \mathbf{x} = \mathcal{L}_\theta[\boldsymbol{x}_0]) = p(\boldsymbol{z}_0 \,|\, \mathbf{x} = \boldsymbol{x}_0) \ ; \forall\theta \in [0, 2\pi) \tag{6.2b}$$

However empirical experiments have shown that perfect independence of the latent variables is hard to achieve in a generative model and that such disentanglement of the generative factors is typically not obtained without further constraints on the models [143]. At that, the assumed uniform distribution of θ_0 is not encouraged by the Gaussian prior distribution of the formulation of the standard VAE. Therefore, we propose to consider a network architecture that explicitly encodes the orientation information of the images, guarantees the relationships of Eq. 6.2a-b and enables the modeling of the Bayesian network illustrated in Figure 6.1-c.

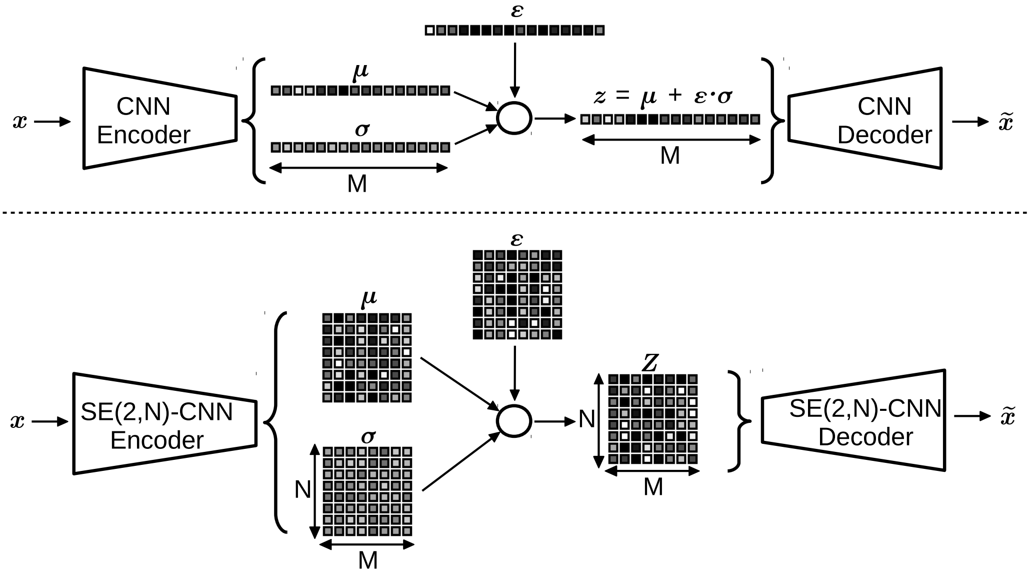


Figure 6.4: Comparison of the auto-encoding pipeline of an image $\boldsymbol{x}$ in a conventional VAE (top) and an *SE(2,N)*-CNN-based VAE (bottom). Here, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ parameterize $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $q_\phi(\boldsymbol{Z}|\boldsymbol{x})$ such that samples $\boldsymbol{z} = [z_i]_{i=1\ldots M}$ and $\boldsymbol{Z} = [Z_{i,j}]_{i,j=1\ldots M}$ are drawn using the reparametrization trick with $\boldsymbol{\epsilon} = [\epsilon_i \sim \mathcal{N}(0,1)]_{i=1\ldots M}$ or $\boldsymbol{\epsilon} = [\epsilon_{i,j} \sim \mathcal{N}(0,1)]_{i,j=1\ldots M}$.

6.4.2 *SE(2)* Variational Auto-Encoder

Although conventional CNNs are equivariant to translations (since 2D convolutions are equivariant to planar translations), they are not guaranteed to be equivariant with respect to rotations of the input images.

Group Structured CNNs Group convolution operations were proposed to give CNNs the structure of the roto-translation group *SE(2)* $:= \mathbb{R}^2 \rtimes$ *SO(2)* [66]. The internal feature maps of CNNs with such group structure can be treated as *SE(2)*-images $F \in \mathbb{L}_2[$*SE(2)*$]$ and the application of convolutional operations with *SE(2)*-image kernels are equivariant under the action of the elements of *SO(2)*. This architecture provides an end-to-end roto-translation equivariance property to CNNs.

The architecture of a *SE(2)-CNN* can be implemented by discretizing the sub-group *SE(2,N)* $:= \mathbb{R}^2 \rtimes$ *SO(2,N)* by sampling *SO(2)* with the elements that correspond to the N rotation angles of $\{2\pi n/N \mid n{=}0...N-1\}$ [63], [157]. This way, the internal feature maps of the network can be implemented as tensors of shape $H \times W \times N \times M$ with H and W the size of the spatial dimensions, N the size of the discretized orientation-axis and M the number of channels in the layer.

Application to the VAE framework We propose to replace the 2D-convolution operations of the conventional CNNs in the VAE framework by *SE(2)*-group convolutions to yield rotation-equivariant mappings from the input images to the sampled latent variables and from the latent variables to the reconstructed images. This change of architecture also relies on the replacement of the first layer of the encoder by a *lifting layer* to produce *SE(2)*-image representation maps; and the replacement of the penultimate layer of the decoder by a *projection layer* to output 2D images [63], [157].

The bottleneck of a conventional CNN-based encoder corresponds to feature vectors of size M. Likewise, in a *SE(2)*-CNN-based VAE, the feature vectors at the bottleneck of the encoder can be defined in terms of the rotation elements of the circle group *SO(2)* solely. We treat these feature vectors as $\mathcal{U}_g[f]$ where $\mathcal{U}_g$ is the *left-regular group-representation* of *SO(2)* on functions $f \in \mathbb{L}_2[$*SO(2)*$]$ with $g = R_\theta \in$ *SO(2)* (we simplify this notation to $\mathcal{U}_\theta f$).

In practice, we consider the sub-group *SO(2,N)*, so that the *SO(2,N)*-vectors $\mathcal{U}_\theta \boldsymbol{f}$ can be implemented as tensors of shape $N \times M$ with N the size of the discretized o-rientation-axis and M the number of latent variables. The difference in embedding structure between conventional VAEs and *SE(2)*-CNN-based VAEs is illustrated in Figure 6.4.

SE(2,N)-Structured Latent Variables Instead of considering real-valued latent variables, we propose to model the latent variables as *SO(2,N)*-vector-valued random variables $\boldsymbol{Z}$. The group structure of the *SE(2)*-modified encoder enables to model the posterior distribution $q_\phi(\boldsymbol{Z} \mid \boldsymbol{x})$ with the property of being equivariant under the action of *SO(2,N)*. Thus, the modeled distribution verifies the relationship of Eq. **??**.

$$q_\phi(\mathcal{U}_\theta[\boldsymbol{Z}] \mid \boldsymbol{x} = \mathcal{L}_\theta[\boldsymbol{x}_0]) = q_\phi(\boldsymbol{Z} \mid \boldsymbol{x} = \boldsymbol{x}_0)\,; \forall\theta \in \textit{SO(2,N)} \tag{6.3}$$

The *SE(2)*-CNN decoder takes the samples $\boldsymbol{Z}$ as input and models the likelihood $p_\psi(\boldsymbol{x}|\boldsymbol{Z})$ as a multivariate Gaussian with identity covariance. Note that this *SE(2)*-CNN-based VAE can be trained with the same objective as conventional VAEs (see Eq.6.1) after adjusting the prior on the latent to a multivariate normal distribution that matches the dimensions of $\boldsymbol{Z}$.

Consequences for Downstream Analysis By construction, the equivariance property of the variational posterior q_ϕ guarantees that rotating the input images by an angle measure θ will cause a cycle-shift on the posterior distribution as expressed by the relationship of Eq. 6.3. Likewise, the rotational equivariance of p_ψ implies that cycle-shifting the values of a latent sample will cause a rotation of the reconstructed images.

As a result, the generative process of the images does not depend directly on θ_0 anymore: this variable becomes encoded in the *SO(2,N)* latent variables as a *hidden shift* on the orientation-axis. Still, the dependence of each variable $Z_{i,j}$ to θ_0 makes downstream analysis of these generative factors subject to the variability of this arbitrary orientation within a dataset.

6.4.3 Separation of Isotropic and Oriented Latents

In order to achieve disentanglement of the orientation information in the latent variables, we make the hypothesis that the set of generative factors can be split in two sets of independent variables: a set of real-valued variables $\boldsymbol{z}^{\mathsf{ISO}} = [z_i^{\mathsf{ISO}}]_{i=1}^{M}$ that codes for non-oriented/isotropic features in the images, and a set of angle variables $\boldsymbol{z}^{\mathsf{ORI}} = [z_i^{\mathsf{ORI}}]_{i=1}^{M}$ with values in $[0, 2\pi]$ that code for oriented structures in the images.

To achieve such partitioning of the latent space, we design the *SE(2,N)*-CNN encoder to approximate two posterior distributions $q_\phi(\boldsymbol{z}^{\mathsf{ISO}}|\boldsymbol{x})$ and $q_\phi(\boldsymbol{z}^{\mathsf{ORI}}|\boldsymbol{x})$ by producing three output components that parameterize these distributions (as illustrated in Figure 6.5):

- Two sets of *SO(2,N)*-vectors that are projected via the *mean* operator along the orientation-axis resulting in a *mean vector* $\boldsymbol{\mu}^{\mathsf{ISO}} \in \mathbb{R}^M$ and a *variance vector* $(\boldsymbol{\sigma}^{\mathsf{ISO}})^2 \in (\mathbb{R}^{+})^M$.
- One set of *softmax-activated SO(2,N)*-vectors Q_i^{ORI} that correspond to discretized approximations of $q_\phi(z_i^{\mathsf{ORI}}|\boldsymbol{x})$ as defined in Eq. 6.4 with $i = 1 \ldots M$. Here the softmax function is used to ensure that each vector Q_i^{ORI} represents a probability mass function.

$$Q^{\mathsf{ORI}}_{i,j} = \int_{\frac{(j-1)2\pi}{N}}^{\frac{j2\pi}{N}} q_\phi(z^{\mathsf{ORI}}_i|\boldsymbol{x})\,\mathrm{d}z^{\mathsf{ORI}}_i \; ; \; j = 1...N \tag{6.4}$$

Finally, the z^{ISO}_i can be drawn from $\mathcal{N}(\mu^{\mathsf{ISO}}_i, (\sigma^{\mathsf{ISO}}_i)^2)$ and the z^{ORI}_i can be directly drawn from $q_\phi(z^{\mathsf{ORI}}_i|\boldsymbol{x})$ using the approximations Q^{ORI}_i (the implementation of these sampling procedures are detailed in the next paragraph).

By construction, the μ^{ISO}_i and σ^{ISO}_i are rotation-invariant, and thus ensure that $\boldsymbol{z}^{\mathsf{ISO}}$ verifies the *invariance* relationship of Eq. 6.2b. The modeled posteriors $q_\phi(z^{\mathsf{ORI}}_i|\boldsymbol{x})$ follow the equivariance relationship of Eq. 6.2a, as θ_0 becomes encoded as a shared hidden shift (modulo 2π) across the variables z^{ORI}_i.

Implementation Details on Latent Sampling During training, the stochastic sampling process of the $\boldsymbol{z}^{\mathsf{ISO}}$ and $\boldsymbol{z}^{\mathsf{ORI}}$ variables requires implementation via differentiable operations in the computational graph, to enable gradient back-propagation through the encoder.

As proposed by Kingma *et al.* [127], we implemented the sampling process of $\boldsymbol{z}^{\mathsf{ISO}}$ via the reparameterization trick to obtain sampled latents $\boldsymbol{z}^{\mathsf{ISO}} = \mu^{\mathsf{ISO}} + \epsilon^{\mathsf{ISO}} \cdot \sigma^{\mathsf{ISO}}$ (as illustrated in Figure 6.4). Here, $\epsilon^{\mathsf{ISO}} \sim \mathcal{N}(0, 1)$ is an auxiliary noise variable to simulate sampling from a Gaussian distribution.

To approximate the sampling process of z^{ORI} we implemented an *Inverse Transform Sampling* layer that returns a set of angle measures drawn from the distributions coded by Q^{ORI}. This layer computes a *continuous inverse cumulative distribution* Q^{-1}_i for each vector Q^{ORI}_i and calculates each angle measure as $\boldsymbol{z}^{\mathsf{ORI}} = Q^{-1}_i(\epsilon^{\mathsf{ORI}})$ with $\epsilon^{\mathsf{ORI}} \sim \mathcal{U}(0, 1)$ a uniformly distributed auxiliary noise variable. An example of this two-step procedure is shown in Figure 6.5.

So as to conserve the end-to-end equivariance property of the framework, the latent variables are reshaped as *SE(2,N)*-vectors, such that the decoding procedure can occur in the same conditions as described in Section 6.4.2. This is done by expanding and repeating the values of $\boldsymbol{z}^{\mathsf{ISO}}$ along the orientation-axis, and via *label smoothing* to encode the sampled angles $\boldsymbol{z}^{\mathsf{ORI}}_i$ into *SE(2,N)*-vectors.

Extended Objective Since the images are generated from two independent sources of generative factors we simply further developed the lower bound in the VAE formulation such that the proposed model can be trained by maximizing the loss written in Eq. 6.5. The constraints are computed by fixing the prior $p(\boldsymbol{z}^{\mathsf{ISO}})$ as a multivariate normal distribution and the prior $p(\boldsymbol{z}^{\mathsf{ORI}})$ as a multivariate uniform distribution on $[0, 2\pi]$.

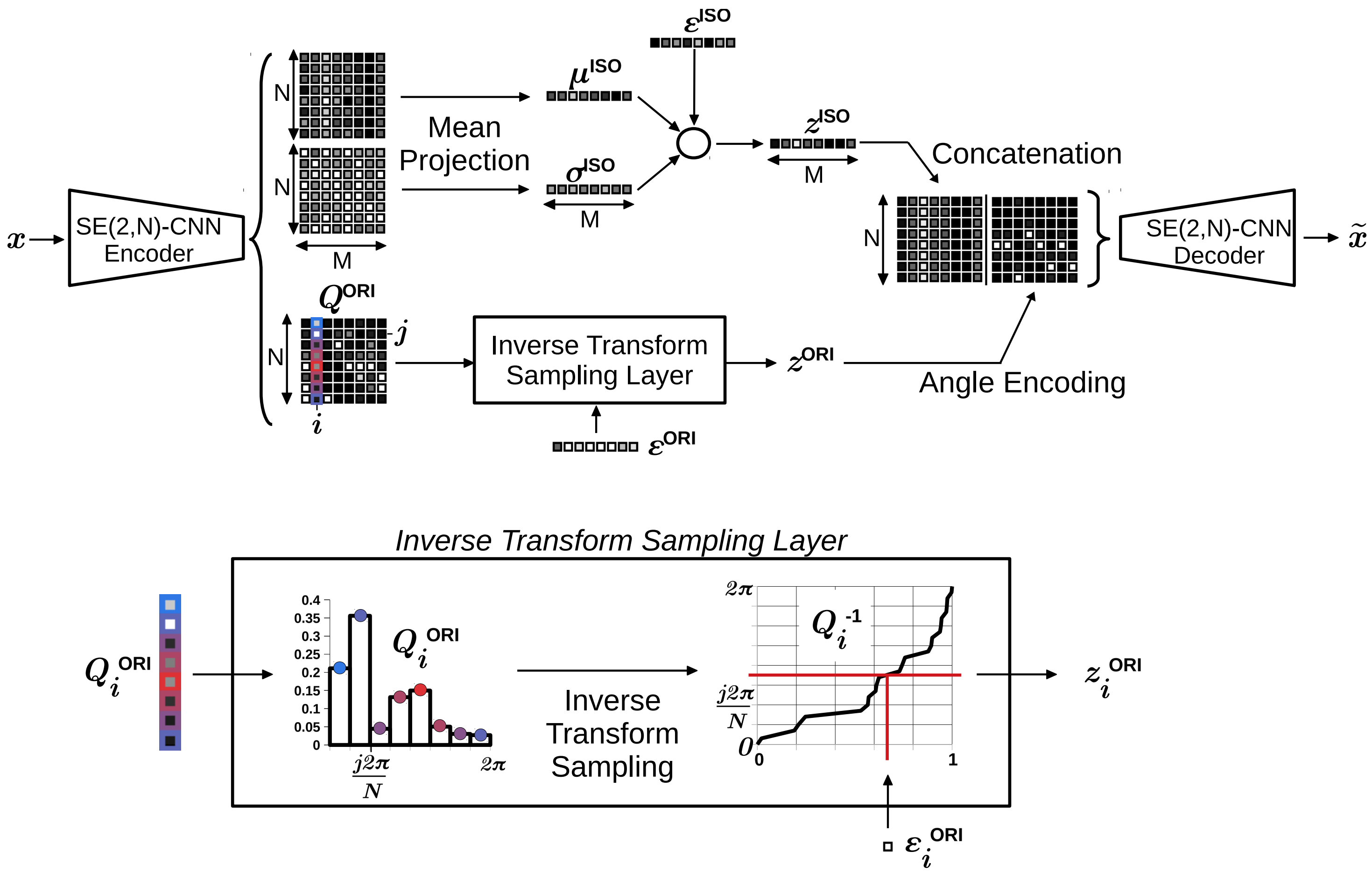


Figure 6.5: Illustration of the auto-encoding pipeline of an image $\boldsymbol{x}$ in the proposed VAE. Here, the embedding consists of two components $\boldsymbol{z}^{\text{ISO}}$ and $\boldsymbol{z}^{\text{ORI}}$. $\boldsymbol{z}^{\text{ISO}}$ is drawn from $q_\phi(\boldsymbol{z}^{\text{ISO}}|\boldsymbol{x})$ parameterized by $\boldsymbol{\mu}^{\text{ISO}}$ and $\boldsymbol{\sigma}^{\text{ISO}}$ such that samples $\boldsymbol{z}^{\text{ISO}} = [z_i^{\text{ISO}}]_{i=1\ldots M}$ are drawn using the reparametrization trick with $\boldsymbol{\epsilon}^{\text{ISO}} = [\epsilon_i^{\text{ISO}} \sim \mathcal{N}(0,1)]_{i=1\ldots M}$. $\boldsymbol{z}^{\text{ORI}}$ is drawn from $q_\phi(\boldsymbol{z}^{\text{ORI}}|\boldsymbol{x})$ via Inverse Transform Sampling applied on the cumulative density functions approximated by $\boldsymbol{Q}^{\text{ORI}}$ using a noise variable $\boldsymbol{\epsilon}^{\text{ISO}} = [\epsilon_i^{\text{ORI}} \sim \mathcal{U}(0,1)]_{i=1\ldots M'}$ (an example of this sampling process is shown for a variable z_i^{ORI}).

$$\begin{aligned}\mathcal{L}_{\theta\text{-VAE}}(\boldsymbol{x}, \boldsymbol{z}^{\mathsf{ISO}}, \boldsymbol{z}^{\mathsf{ORI}}; \psi, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}^{\mathsf{ISO}}, \boldsymbol{z}^{\mathsf{ORI}}|\boldsymbol{x})} &\left[\log p_\psi(\boldsymbol{x} \mid \boldsymbol{z}^{\mathsf{ISO}}, \boldsymbol{z}^{\mathsf{ORI}})\right] \\ &-\beta^{\mathsf{ISO}} \cdot \mathrm{D}_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}^{\mathsf{ISO}}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}^{\mathsf{ISO}})\right] \\ &-\beta^{\mathsf{ORI}} \cdot \mathrm{D}_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}^{\mathsf{ORI}}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}^{\mathsf{ORI}})\right]\end{aligned} \tag{6.5}$$

Consequences for Downstream Analysis The isotropic generative factors $\boldsymbol{z}^{\mathsf{ISO}}$ are guaranteed to be independent to θ_0 by construction, so they can be compared and aggregated within/across populations of tissue image patches independently of their spatial orientation. Likewise, the distribution of angles z^{ORI} in a given population characterizes the variability of oriented features independently of isotropic factors.

Complementary Reconstruction Loss

Conventional VAE models are known for generating/reconstructing *blurry* images of a lower quality than original images. Poor reconstructions might imply that the high-frequency details in the images do not get encoded in the latent representation and might entail poor performances in downstream tasks. This limited quality of generated images is often associated to the pixel-wise reconstruction term of the VAE objective (see Eq. 6.1).

To ensure the reconstruction term enables generation of realistic images, we used the extension of the VAE objective that was described in [169]. This method is inspired by the initial work of Larsen *et al.* [144] that introduces a discriminator CNN D with parameters χ in the VAE framework.

This discriminator is trained to classify batches balanced between real images $\boldsymbol{x}{\sim}p(\boldsymbol{x})$ and reconstructed images $\widetilde{\boldsymbol{x}}{\sim}p_\psi(\boldsymbol{x}|\boldsymbol{z})$. The internal feature maps of D for a given input image $\boldsymbol{x}$ are defined as $D_i(\boldsymbol{x})$ with i a layer index. The $D_i(\widetilde{\boldsymbol{x}})$ can be considered as additional Gaussian observations with identity covariance drawn from $p_\chi(D_i(\boldsymbol{x})|\boldsymbol{z})$. We thus define extra reconstruction losses $\mathcal{L}_i^D$ that we use to complement the model objective as defined in Eq. 6.6 where γ is a weighting hyper-parameter.

$$\begin{aligned}\mathcal{L}_{\text{VAE+}}(\boldsymbol{x}, \boldsymbol{z}; \psi, \phi, \chi) &= \mathcal{L}_{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z}; \psi, \phi) + \gamma \sum_i \mathcal{L}_i^D(\boldsymbol{x}, \boldsymbol{z}; \psi, \phi, \chi) \\ \mathcal{L}_i^D(\boldsymbol{x}, \boldsymbol{z}; \psi, \phi, \chi) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\chi(D_i(\boldsymbol{x}) \mid \boldsymbol{z})]\end{aligned} \tag{6.6}$$

The training procedure of the full model pipeline consists of alternating training steps between updating ψ and ϕ by maximizing $\mathcal{L}_{\text{VAE+}}$ and updating χ by minimizing the cross-entropy loss on the predictions of D.

This extension of the VAE framework has the benefit to keep the rest of the model formulation intact and was shown to be beneficial in autoencoder-based applications involving cell images [142], [169]. The discriminator is used at training time only and is discarded at inference time, which thus does not cause any computational overhead for downstream analysis. All the models investigated in this paper were extended by this method.

6.5 Experiments

In this section we detail the network architectures we designed to implement the proposed orientation-disentangled VAE (Section 6.4.3) and a comparable baseline VAE model (Section 6.4.1). We also describe their training procedures and the evaluation protocols we applied to compare the quality of the learned representation and gain insights into the effect of the disentanglement and newly introduced hyper-parameters.

6.5.1 Model Architectures

We designed the conventional CNN encoders and SE(2,N)-CNNs encoders of the VAEs as straight-forward sequences of four blocks that each consists of a convolutional layer, a batch normalization layer (BN), a leaky reLU non-linearity and a max-pooling layer.

In the case of the *SE(2,N)*-CNNs, conventional $\mathbb{R}^2$-convolutional layers were replaced by *SE(2,N)* convolutional layers and BN layers were adjusted to include the orientation-axis in the computation of the batch statistics. We also introduced intermediate *projection layers* similar to the locally rotation-invariant CNNs proposed by Andrearczyk *et al.* [74] within all the hidden layers in order to reduce the computational cost of the network.

In order to generate embedding samples at the bottleneck of the models during training, we implemented the computational sampling procedures presented in Section 6.4.1 for the conventional VAEs and in Section6.4.3 for the orientation-disentangled VAEs.

The decoder networks correspond to mirrored versions of their encoder counterparts via the use of up-sampling layers and transposed convolutions. We included a mean-projection layer at the end of the *SE(2,N)*-CNN decoder to project the features maps on $\mathbb{R}^2$. Finally we added an extra $1{\times}1$-convolutional layer to the decoders to output images whose dimensions match the input dimensions.

For all the networks, we used kernels with spatial size $5{\times}5$ so as to enable proper rotation of the *SE(2,N)*-kernels as described in [63], [157]. We fixed the angular resolution of the *SE(2,N)* layers to $N{=}8$ as it was previously shown to give optimal performances [157]. To ensure fair comparison of the models, we balanced the number of channels in each layer such that the total number of weights between the two types of VAEs is approximately the same. In order to have a fair comparison,, we also fixed the total number of variables in the bottleneck of all the models (size of $\boldsymbol{z}$ is 64 in the conventional VAEs and $\boldsymbol{z}^{\mathsf{ISO}}$ and $\boldsymbol{z}^{\mathsf{ORI}}$ are both of size 32 in the orientation-disentangled VAEs).

6.5.2 Training Procedures

All the models were trained using the training set of the *TUPAC-ROI dataset* described in Section 6.3. We used mini-batches that consist of 35 image patches, such that the distribution of the WSIs of origin within each batch was approximately uniform. We used the *Adam* optimizer to update the weights of the encoders and decoders (learning rate 0.001, β_1=0.9, $\beta_2 = 0.999$), and *Stochastic Gradient Descent* with momentum to update the weights of the discriminator (learning rate 0.001, momentum 0.9). All convolutional kernels were regularized via decoupled weight decay with coefficient $1{\times}10^{-4}$. We stopped the training process after convergence of the loss on the validation set (approximately $20{\times}10^{3}$ iterations). The weighting coefficient of the disciminator-based reconstruction loss was fixed to γ=0.01 across all experiments.

In order to assess the effect of the weighting coefficient of the prior constraint as it was evidenced by Higgins *et al.* [143] we trained the models with varying values of β, β^{ISO} and β^{ORI} in $\{0.1, 0.5, 1.0, 2.0, 4.0\}$.

6.5.3 Downstream Analysis

In order to compare the quality of the learned representation at the bottleneck of the different trained VAE models, we trained subsequent logistic regression models that take the learned representation as input to solve downstream tasks.

We investigated two types of downstream tasks to evaluate the value of the learned representation in different contexts: patient-level classification tasks (predicting the pleomorphism grade and tumor proliferation grade of a given WSI) using the *TUPAC-ROI* dataset and single-cell-level classification tasks (predicting the cell-type of a given nucleus) using the *CRCHP* dataset.

Each downstream task was investigated with different types of representation of the data points based on the different latent variables estimated in the VAEs (see $\boldsymbol{z}$ in Section 6.4.1, $\boldsymbol{z}^{\mathsf{ISO}}$ and $\boldsymbol{z}^{\mathsf{ORI}}$ in Section 6.4.3. For comparison, we also considered well-established nuclear morphometric measurements for comparison (mean nuclear area and standard deviation of the nuclear area) as well as combinations of representations.

Patient-level classification tasks: we first aggregated the representation of all nucleus-centered image patches within every ROI of the dataset (see Section 6.3). For the representations corresponding to $\boldsymbol{z}$ and $\boldsymbol{z}^{\mathsf{ISO}}$, we considered the mean $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{\mathsf{ISO}}$ of the estimated posterior distributions as cell-level embeddings, and we further computed ROI-level embeddings by computing the mean of all the embeddings obtained within a ROI. We then trained a single-layered multi-class logistic regression model to minimize the cross-entropy loss given the WSI-level ground-truth labels.

For the representation corresponding to $\boldsymbol{z}^{\mathsf{ORI}}$, we used their discrete distribution at the ROI-level as an aggregated representation (see Q^{ORI} in Section 6.4.3). Likewise, we trained a two-layer multi-class logistic regression model with an intermediate maximum projection layer to ensure rotational-invariance of the predictions.

Single-cell-level classification task: we directly used the mean $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{\mathsf{ISO}}$ of the posterior distributions on $\boldsymbol{z}$ and $\boldsymbol{z}^{\mathsf{ISO}}$ as representation of nucleus-centered image patches and trained a single-layered multi-class logistic regression model to minimize the cross-entropy loss given cell-type ground-truth labels.

All logistic regression models were regularized with L_2-weight decay and the associated coefficient was fine-tuned on the validation set.

6.6 Results

This section details the evaluation protocols and metrics we used to assess the quality of the learned representation we obtained with the methods presented in Section 6.4. We present the downstream performances for each type of representation on the tasks described in Section 6.5.

6.6.1 Slide-Level Downstream Performances

We evaluated the trained multi-class logistic regression models on the hold-out test set described in Section 6.3. For each model, we considered a set of binary classifiers that correspond to pairwise comparisons of each class against the other classes. For each set of such binary classifiers, we computed a set of *Receiving Operating Curves* (ROCs) based on the model predictions. Within each set of ROCs, we computed the corresponding *Areas Under the Curve* (AUC) and the *mean of AUCs* (mAUC) across that set to summarize the AUC metric given the multi-class setting at hand.

To assess the robustness of the learned representations given perturbations of the training data, we resampled the training set ten times, re-trained the models and reported the mean and standard deviation of the mAUCs across these repeats.

Pleomorphism Grading The results for the pleomorphism grade prediction task are summarized in Figure 6.6. We report an improvement of the mAUC of 0.038 from the isotropic learned representation in comparison to the learned representation of the baseline VAE. The isotropic learned representation performed better than the oriented learned representation for all the tested values of β^{ISO}. We obtained a consistent additional improvement of performance when combining the isotropic learned representation with segmentation-based features.

Tumor Proliferation Grade Prediction The results for the tumor proliferation grade prediction task are summarized in Figure 6.7. We do not report any significant im-

Representation	multi-class AUC
Mean Nuclear Area (MNA)	0.519 ± 0.016
MNA + St.Dev Nuclear Area (SDNA)	0.681 ± 0.016
🟦 Baseline VAE	0.757 ± 0.045
🟩 SE(2,8)-VAE - ISO	0.795 ± 0.038
🟥 SE(2,8)-VAE - ORI	0.706 ± 0.039
🟦 Baseline-VAE + MNA + SDNA	0.776 ± 0.034
🟩 SE(2,8)-VAE - ISO + MNA + SDNA	**0.821 ± 0.031**
🟥 SE(2,8)-VAE - ORI + MNA + SDNA	0.768 ± 0.020

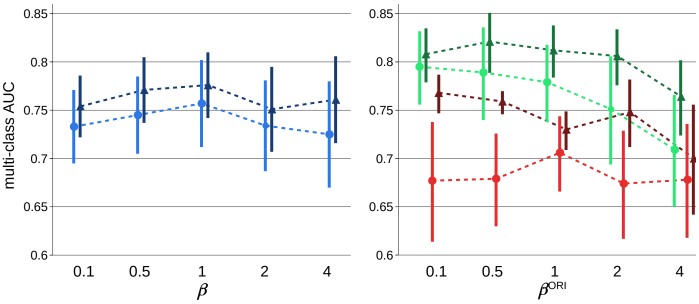


Figure 6.6: Performances in downstream analysis for pleomorphism grading. The table shows best obtained scores for each type of investigated representation. The plots shows the effect of different hyper-parameters: β for the baseline VAE, and β^{ORI} with fixed $\beta^{\text{ISO}} = 1$ for the proposed orientation-disentangled VAE. Mean ± standard deviation of the multi-class AUC are indicated in the table and shown with a bar in the plots.

provement of the mAUC from using isotropic or oriented learned representation in comparison to the learned representation of the baseline VAE and to segmentation-based features. We report a consistent additional improvement of performance when combining the isotropic learned representation with segmentation-based features.

Representation	multi-class AUC
Mean Nuclear Area (MNA)	0.572 ± 0.007
MNA + St.Dev Nuclear Area (SDNA)	0.722 ± 0.016
🟦 Baseline VAE	0.699 ± 0.044
🟩 SE(2,8)-VAE - ISO	0.701 ± 0.057
🟥 SE(2,8)-VAE - ORI	0.723 ± 0.033
🟦 Baseline-VAE + MNA + SDNA	0.738 ± 0.048
🟩 SE(2,8)-VAE - ISO + MNA + SDNA	**0.806 ± 0.050**
🟥 SE(2,8)-VAE - ORI + MNA + SDNA	0.725 ± 0.030

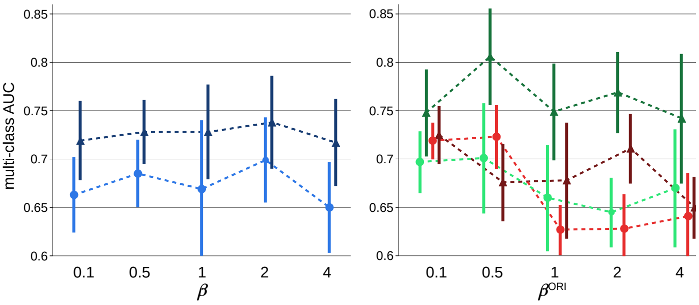


Figure 6.7: Performances in downstream analysis for tumor proliferation grade prediction. The table shows best obtained scores for each type of investigated representation. Curves shows the effect of different hyper-parameters: β for the baseline VAE, and β^{ORI} with fixed $\beta^{\text{ISO}} = 1$ for the proposed orientation-disentangled VAE. Mean ± standard deviation of the multi-class AUC are indicated in the table and shown with a bar in the plots.

6.6.2 Cell-Level Downstream Performances

We used the same protocol to evaluate the cell-type classification models. The results for the cell-type classification task are summarized in Figure 6.8. We report a consistent improvement of the mAUC from using the isotropic representation in comparison to the learned representation of the baseline VAE. The oriented learned representation performed worse than the representation of the baseline VAE or the isotropic learned representation for all the tested values of β and β^{ISO}.

Representation	multi-class AUC
🟦 Baseline VAE	0.802 ± 0.002
🟩 SE(2,8)-VAE - ISO	**0.823 ± 0.003**
🟥 SE(2,8)-VAE - ORI	0.739 ± 0.004

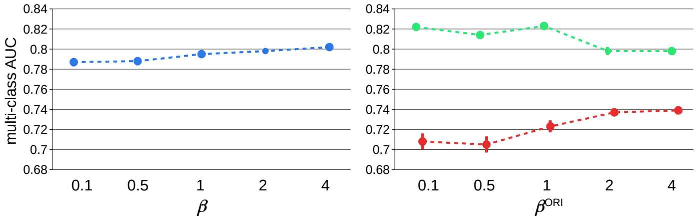


Figure 6.8: Performances in downstream analysis for single-cell type classification. The table shows best obtained scores for each type of investigated representation. Curves shows the effect of different hyper-parameters: β for the baseline VAE, and β^{ORI} with fixed $\beta^{\text{ISO}} = 1$ for the proposed orientation-disentangled VAE. Mean $\pm$ standard deviation of the multi-class AUC are indicated in the table and shown with a bar in the plots.

6.7 Discussion and Conclusions

In this study, we proposed a novel rotation-equivariant variational auto-encoder framework that learns two types of generative factors: isotropic real-valued components and oriented angular components. We showed that this two-fold low-dimensional structure can efficiently represent histopathology images. We investigated in a controlled experimental setup, the predictive power of the learned representation using the proposed frameworks and show its advantage in comparison to unsupervised baseline counterparts. The difference of generative action of each type of variable was qualitatively demonstrated via smooth transitions of generated examples given interpolated embeddings (see Figures. 6.2-6.3).

Qualitatively, we observed that the isotropic learned representation captures morphological factors such as stain ratios, nuclear sparsity, thickness of the nuclear boundary (see Figure 6.2(b)) whereas the oriented learned representation codes for the radial location of the surrounding objects (non-centered neighboring nuclei) and asymmetric structures (see Figure 6.2(a)).

Quantitatively, using isotropic representation was always better or as good as the representation learned by conventional VAE or segmentation-based nuclear area features (see Section 6.6). This is in agreement with our hypothesis that the orientation information that is entangled in the representation learned by the baseline VAE affects the quality of the aggregated representation, and subsequent downstream performances. This was both observed for patient-level grading tasks based on aggregated representation and for cell-level classification tasks. This is also in line with previous results reported in studies on supervised rotation-invariant CNN models trained to solve computational pathology tasks [63], [107], [108], [110], [157], [211].

Although our results suggest that unsupervised learned representation is an efficient alternative to hand-crafted feature-based representations, the fact that the combination of hand-crafted nuclear feature representation with unsupervised learned representation gave a consistent improvement of performances also reveals the limitations of the proposed framework. Indeed, complex knowledge-based quantities such as the *mean nuclear area* that were relevant for the task at hand could not be extracted from the learned representation as they do not necessarily correspond to independent generative factors that the models learned. However we conjecture that this limitation might be due to the restricted architectural design of the model that we chose for the comparative analysis. We thus believe that this limitation can be potentially overcome by using more complex and sophisticated architectures for the latent variable models and downstream classification models.

The performances achieved on the pleomorphism grading task using the isotropic learned representation indicate that these features were more predictive than the oriented features to solve this specific task. This is expected as by definition the pleomorphism grade was labeled based on rotation-invariant nuclear morphological factors. However on the tumor proliferation grading task, oriented features were slightly more predictive than isotropic features. This suggests that these two types of features were equally informative of this specific patient-level grade.

Besides these applications, the proposed framework can be used as a generic tool to quickly gain insights, and with minimal training effort, into the slide-level predictive value of fixed-scale image patches. Indeed, we showed that logistic regression models could be trained using aggregated representation of cell populations to predict patient-level target values such as the pleomorphism grade and tumor proliferation grade. But the same approach could potentially be applied to estimate any other slide-level value. Also, extensions to other convolutional latent variable models are possible including more complex architectures (such as flow-based generative models), other families of variational distributions (such as structured posterior distributions of the oriented variables) and other training paradigms (such as in a semi-supervised framework). For computational convenience, we proposed implementing the sampling of the angle variables by means of a straight-forward *inverse sampling layer*, however, other end-to-end sampling strategies could be investigated for further improvement. This framework is also transferable to other problems in which one wants to model a posterior distribution on the rotation group. Other interesting applications for future work include pre-training for patch-based classification tasks, and compression of WSIs.

Chapter 7
Discussion

Deep convolutional neural networks are powerful computational systems to learn abstract representation of images: they have become a methodology of choice to achieve state-of-the-art performances across many computational pathology tasks. However, the high variability of histopathology data limits the robustness of deep learning systems to inconsistent and unforeseen variations.

With the goal of improving the quality of this learned representation, the main contributions of this thesis consist of the development of new frameworks that improve the robustness of deep learning models by making their learned representation invariant to irrelevant factors of variations. This thesis focused on two known categories of nuisance variables: slide-specific variations (via domain-adversarial methods, in Chapter 2) and variations in terms of spatial orientation (via *SE(2,N)* convolutional layers, in Chapter 3). We demonstrated the applicability of these frameworks to several patch-level computational pathology tasks in both supervised (Chapters 2 and 3) and unsupervised (Chapter 4-5-6) learning contexts.

The main goal for the frameworks presented in this thesis, was to enable the development of deep learning models that can be safely integrated in a clinical environment, as opposed to current state-of-the-art deep learning models that are potentially subject to making mispredictions due to the unforeseen dataset shifts (distributional changes between the training data and the data encountered in practice), and for which no design decisions were made to improve or guarantee their robustness.

Even when trained models are properly validated on a hold-out set, they can still be subject to making mispredictions caused by dataset shifts, whereas the decisions made by a human expert in the same circumstances would not be affected (see examples of error made by baseline models in Chapters 2 and 3). This lack of guaranteed robustness to distributional variations goes beyond the scope of computational pathology, and constitutes a well-known artificial intelligence safety problem [213], [214]. This problem is also linked to another legitimate trust issue towards deep learning systems that stems from the fact these models can be viewed as black boxes whose feature representation cannot be directly interpreted and verified by humans. Along with this lack of direct interpretability, we argue that this black-box status is an other reason why the robustness of the models cannot be fully guaranteed.

In this chapter, we summarize how the studies presented in this thesis contribute to solving this open problem in the context of computational pathology, we discuss the understanding of the limitations of the proposed solutions and present recent progress and directions for future work.

In their review, Bengio *et al.* [24] present building invariant features as a generic desired property for learning good representations. In this thesis, we followed this general principle by explaining away irrelevant factors of variations and build up more robust representations. One way to achieve this goal consisted in a loss-based domain-adversarial approach to improve robustness to slide-specific factors of variations (Chapter 2). The proposed framework exploits the available known inter-slide variability of a training set to encourage learning of a domain-invariant representation. In comparison to existing ad-hoc methods, the generic aspect of the proposed framework can in principle make a model robust to a wider range of variations than the ones addressed by normalization and augmentation approaches. As shown in a controlled set of experiments, generalization to images from external datasets was improved by combining this approach with conventional ones. Since generalization performances can be improved by such a framework, the same trained model can be integrated across various labs and can potentially produce a more consistent output than models that do not specifically address inter-slide variability. Yet, we acknowledge that this method only acts as a generic regularizer to encourage domain invariance and does not bring any generalization guarantee. Nonetheless, domain-adversarial approaches still are a promising research direction as demonstrated by recent studies in computational pathology [215], [216]. It is important for future work to combine the use of this type of method with a large number of domains as a way to provide a soft guarantee regarding the range appearance variability the model was exposed to. Recent developments in domain-adversarial training and alternative approaches have been proposed since our original publication and these constitute relevant research directions [209], [217]–[220].

As opposed to constraining a model by means of a penalty in the training loss function, we investigated solutions to build-in invariance properties into the architecture of CNNs. Indeed, we showed in Chapter 3 how we can achieve robustness to the global orientation of WSIs by construction, which is not a goal specifically addressed by conventional CNNs. Without an approach such as the proposed equivariant convolutional layers, the end users of CNN-based models might get a different output when rotating input images, which brings an undesired uncertainty for making a clinical decision. With the proposed framework, end users now have a guarantee that the internal features learned by a CNN to produce outputs are rotation-equivariant. In consequence, this framework removes the typical extra computational cost of performing rotation augmentation at test time, and removes the uncertainty of the output related to the orientation of the input images. Furthermore, we showed in a set of comparative analyzes that this framework significantly improves performances across multiple classification tasks. Limitations of this approach include its specificity to a single type of transformation and the relative increased computational cost related to the increased usage of single weights in the computational pipeline. Recent studies present promising directions to overcome these limitations by providing models with invariance to other transformations [65], [75], and in a more efficient way via attention mechanisms [221].

These two investigated approaches aimed at removing irrelevant information from the learned representation, however we can potentially achieve the same goal with a generative perspective, with the concept of disentanglement[24], in which one wants to extract and set apart distinct informative factors of the data from each other. We followed this concept in Chapters 5 and 6, by proposing an auto-encoder that can disentangle the orientation information in image patches. The developed models learn low-dimensional representations of images that enable both high quality reconstructions and the ability to achieve reasonable performances in downstream tasks. Previously proposed unsupervised latent variables models for the analysis of large scale microscopy image datasets were not competitive with feature engineering methods, or supervised approaches, and did not enable the high quality of reconstruction we achieved. With the orientation disentanglement framework we proposed, this is the first time to our knowledge that a CNN-based representation can partition isotropic and oriented variables. This framework enables end users to analyze the learned generative factors that correspond to oriented or isotropic components independently, and offers the possibility to visually inspect the structural changes associated with each factor via latent manipulation. This is opposed to conventional latent variable models, for which the arbitrary orientation of the input images is likely to affect the latent variables in an unknown manner, which in consequence, unreliably affects any subsequent classification model. With tasks for which the orientation of images is assumed to be irrelevant, we showed this information can now be discarded from the representation, and that this process can improve performances in downstream slide-level classification tasks. For future work, we consider extensions to disentangle domain information in the learned representation as proposed in [209], [222], [223]. As such frameworks are modular, other extensions to control the disentanglement of multiple known irrelevant factors should be investigated [224], [225] and performances in downstream tasks could potentially benefit from other recently proposed latent variable models [226]–[229].

Towards the end goal of providing robust image analysis systems, the quantification of the predictive uncertainty of machine learning models is an important research topic. By providing ways to detect unexpected data variations that are not part of the training data, the safety of the systems integrated in the clinic can be improved. We did not investigate methods to quantify the predictive uncertainty of the developed models, but we assert it is an important topic that should be further investigated, in particular in the context of the dataset shift problem addressed in this thesis [230].

We argue that the development of robust models is not possible without proper construction and use of datasets. To properly validate trained models given the high variability of histopathology data, it is important to evaluate them on independent cohorts ideally acquired in different hospitals/institutes, as it has become standard according to recent large-scale studies [216], [231], [232]. Since the developed models are data-driven, we argue it is also critical to design training sets that expose models to variability that cannot be simulated via augmentation techniques solely. This concept can be extended to strategies that aim at injecting prior knowledge about data variations (as is the motivation of Chapter 4) into the model, such that robustness to these variations can be improved.

Furthermore, as new whole-slide high-throughput imaging techniques are emerging [233]–[237], these are reshaping the field of computational pathology by providing novel co-localized molecular data. This scaled up source of information can support the development of powerful machine learning models, but also comes with the robustness-related challenges addressed in this thesis. In the line of the research we presented, we believe that future models will have to be carefully designed to ensure robustness to various dataset shifts associated with these new techniques. For an efficient progress, methodologies combining this new data with conventional WSI of H&E-stained specimens while exploiting knowledge learned from existing models [238] are promising research directions.

As shown in this thesis, proper design of deep learning models with respect to their architectures and training procedures are key elements to produce robust automated image analysis systems for computational pathology. As the proposed frameworks are modular, this work present significant potential to contribute to advanced CNN-based WSI processing pipelines and forthcoming challenging slide-level classification tasks.

References

[1] J. I. Epstein, W. C. Allsbrook Jr, M. B. Amin, L. L. Egevad, I. G. Committee, *et al.*, "The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma," *The American Journal of Surgical Pathology*, vol. 29, pp. 1228–1242, 2005.

[2] E. A. Rakha, M. E. El-Sayed, A. H. Lee, C. W. Elston, M. J. Grainge, Z. Hodi, R. W. Blamey, and I. O. Ellis, "Prognostic significance of nottingham histologic grade in invasive breast carcinoma," *Journal of Clinical Oncology*, vol. 26, no. 19, 3153–3158, 2008.

[3] B. Delahunt, J. C. Cheville, G. Martignoni, P. A. Humphrey, C. Magi-Galluzzi, J. McKenney, L. Egevad, F. Algaba, H. Moch, D. J. Grignon, *et al.*, "The international society of urological pathology (isup) grading system for renal cell carcinoma and other prognostic parameters," *The American Journal of Surgical Pathology*, vol. 37, no. 10, pp. 1490–1504, 2013.

[4] N. Stathonikos, T. Q. Nguyen, C. P. Spoto, M. A. Verdaasdonk, and P. J. van Diest, "Being fully digital: perspective of a dutch academic pathology laboratory," *Histopathology*, vol. 75, pp. 621–635, 2019.

[5] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, "Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives," *Journal of Pathology Informatics*, vol. 9, 2018.

[6] D. N. Louis, M. Feldman, A. B. Carter, A. S. Dighe, J. D. Pfeifer, L. Bry, J. S. Almeida, J. Saltz, J. Braun, J. E. Tomaszewski, *et al.*, "Computational pathology: A path ahead," *Archives of Pathology & Laboratory Medicine*, vol. 140, no. 1, pp. 41–50, 2016.

[7] E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, *et al.*, "Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the digital pathology association," *The Journal of Pathology*, vol. 249, no. 3, pp. 286–294, 2019.

[8] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101 813, 2021.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[10] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, 170–175, 2016.

[11] O. Jimenez-del Toro, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, M. Rousson, H. Müller, and M. Atzori, "Analysis of histopathology images: From traditional machine learning to deep learning," in *Biomedical Texture Analysis*, 2017, pp. 281–314.

[12] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018.

[13] E. Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2312–2325, 2020.

[14] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013, pp. 411–418.

[15] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, *et al.*, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[16] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[17] D. Romo-Bucheli, A. Janowczyk, H. Gilmore, E. Romero, and A. Madabhushi, "Automated tubule nuclei quantification and correlation with oncotype dx risk categories in er+ breast cancer whole slide images," *Scientific Reports*, vol. 6, p. 32 706, 2016.

[18] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Journal of the American Medical Association (JAMA)*, vol. 318, no. 22, pp. 2199–2210, 2017.

[19] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. (In Press), 2019.

[20] H. Pinckaers, B. van Ginneken, and G. Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. (In Press), 2020.

[21] Y. Yagi, "Color standardization and optimization in whole slide imaging," in *Diagnostic Pathology*, vol. 6, 2011, S15.

[22] S. A. Taqi, S. A. Sami, L. B. Sami, and S. A. Zaki, "A review of artifacts in histopathology," *Journal of Oral and Maxillofacial Pathology: JOMFP*, vol. 22, no. 2, p. 279, 2018.

[23] G. E. Csurka, *Domain adaptation in computer vision applications*. Springer, 2017, vol. 8.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, “Towards a definition of disentangled representations,” *arXiv preprint arXiv:1812.02230*, 2018.

[26] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 4114–4124.

[27] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013, pp. 325–333.

[28] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, “The variational fair autoencoder,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[29] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[30] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, “Domain-adversarial neural networks to address the appearance variability of histopathology images,” in *Proceedings of the International Workshop on Deep Learning in Medical Image Analysis in Conjunction with MICCAI*, vol. 10553, 2017, pp. 83–91.

[31] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, and J. van der Laak, “The importance of stain normalization in colorectal tissue classification with convolutional networks,” in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2017, pp. 160–163.

[32] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, “A dataset and a technique for generalized nuclear segmentation for computational pathology,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[33] M. Veta, P. J. van Diest, M. Jiwa, S. Al-Janabi, and J. P. Pluim, “Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method,” *PloS One*, vol. 11, no. 8, e0161286, 2016.

[34] M. Veta, P. J. Van Diest, and J. P. Pluim, “Cutting out the middleman: Measuring nuclear area in histopathology slides without segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 632–639.

[35] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, *et al.*, “Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images,” *Pattern Recognition*, vol. 86, pp. 188–200, 2019.

[36] D. Tellez, M. Balkenhol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi, “H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection,” in *Proceedings of SPIE Medical Imaging*, 2018, 105810Z.

[37] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, *et al.*, "Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 9, pp. 2126–2136, 2018.

[38] Y.-R. Van Eycke, C. Balsat, L. Verset, O. Debeir, I. Salmon, and C. Decaestecker, "Segmentation of glandular epithelium in colorectal tumours to automatically compartmentalise ihc biomarker quantification: A deep learning approach," *Medical Image Analysis*, vol. 49, pp. 35–45, 2018.

[39] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Proceedings of the International Conference on Image Analysis and Recognition*, 2018, pp. 737–744.

[40] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[41] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, *et al.*, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017, pp. 516–524.

[42] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 801–814, 2018.

[43] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4500–4509.

[44] M. Gadermayr, M. Strauch, B. M. Klinkhammer, S. Djudjaj, P. Boor, and D. Merhof, "Domain adaptive classification for compensating variability in histopathological whole slide images," in *Proceedings of the International Conference on Image Analysis and Recognition*, 2016, pp. 616–622.

[45] M. Gadermayr, V. Appel, B. M. Klinkhammer, P. Boor, and D. Merhof, "Which way round? A study on the performance of stain-translation for segmenting arbitrarily dyed histological images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 165–173.

[46] Y. Huang, H. Zheng, C. Liu, X. Ding, and G. K. Rohde, "Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1625–1632, 2017.

[47] H. Bilen and A. Vedaldi, "Universal representations: The missing link between faces, text, planktons, and cat breeds," *arXiv preprint arXiv:1701.07275*, 2017.

[48] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain mr segmentation across scanners and protocols," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[49] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[50] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 443–450.

[51] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Domain separation networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[52] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7167–7176.

[53] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3722–3731.

[54] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, *et al.*, “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” in *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*, 2017, pp. 597–609.

[55] C. W. Elston and I. O. Ellis, “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*, vol. 19, no. 5, pp. 403–410, 1991.

[56] M. Veta, Y. J. Heng, N. Stathonikos, B. E. Bejnordi, F. Beca, T. Wollmann, K. Rohr, M. A. Shah, D. Wang, M. Rousson, *et al.*, “Predicting breast tumor proliferation from whole-slide images: The tupac16 challenge,” *Medical Image Analysis*, vol. 54, pp. 111–121, 2019.

[57] M. Veta, P. J. Van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. Larsen, J. S. Vestergaard, A. B. Dahl, *et al.*, “Assessment of algorithms for mitosis detection in breast cancer histopathology images,” *Medical Image Analysis*, vol. 20, no. 1, pp. 237–248, 2015.

[58] Cancer Genome Atlas Network *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, p. 61, 2012.

[59] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[60] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[61] A. C. Ruifrok, D. A. Johnston, *et al.*, “Quantification of histochemical staining by color deconvolution,” *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[62] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2009, pp. 1107–1110.

[63] E. J. Bekkers, M. W. Lafarge, M. Veta, K. A. Eppenhof, J. P. Pluim, and R. Duits, "Roto-translation covariant convolutional networks for medical image analysis," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 11070, 2018, pp. 440–448.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[65] E. J. Bekkers, "B-spline CNNs on lie groups," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[66] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 2990–2999.

[67] E. Hoogeboom, J. W. T. Peters, T. Cohen, and M. Welling, "Hexaconv," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[68] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5028–5037.

[69] M. Weiler, F. A. Hamprecht, and M. Storath, "Learning steerable filters for rotation equivariant cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 849–858.

[70] E. J. Bekkers, M. Loog, B. M. ter Haar Romeny, and R. Duits, "Template matching via densities on the roto-translation group," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 452–466, 2018.

[71] D. Worrall and G. Brostow, "Cubenet: Equivariance to 3d rotation and translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 567–584.

[72] M. Winkels and T. S. Cohen, "Pulmonary nodule detection in ct scans with equivariant cnns," *Medical Image Analysis*, vol. 55, pp. 15–26, 2019.

[73] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen, "3d steerable cnns: Learning rotationally equivariant features in volumetric data," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 381–10 392.

[74] V. Andrearczyk, J. Fageot, V. Oreiller, X. Montet, and A. Depeursinge, "Exploring local rotation invariance in 3d cnns with steerable filters," in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, vol. 102, 2019, pp. 15–26.

[75] D. Worrall and M. Welling, "Deep scale-spaces: Equivariance over scale," in *Advances in Neural Information Processing Systems*, 2019, pp. 7366–7378.

[76] T. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[77] R. Kondor and S. Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," 2018.

[78] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds," *arXiv preprint arXiv:1802.08219*, 2018.

[79] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning so(3) equivariant representations with spherical cnns," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–68.

[80] T. S. Cohen, M. Geiger, and M. Weiler, "A general theory of equivariant cnns on homogeneous spaces," in *Advances in Neural Information Processing Systems*, 2019, pp. 9145–9156.

[81] T. S. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling, "Gauge equivariant convolutional networks and the icosahedral cnn," 2019.

[82] R. Gens and P. M. Domingos, "Deep symmetry networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 2537–2545.

[83] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[84] E. Bekkers, R. Duits, and M. Loog, "Training of templates for object recognition in invertible orientation scores: Application to optic nerve head detection in retinal images," in *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2015, pp. 464–477.

[85] S. Dieleman, J. De Fauw, and K. Kavukcuoglu, "Exploiting cyclic symmetry in convolutional neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[86] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 519–528.

[87] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Rotation equivariant vector field networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5048–5057.

[88] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[89] J. F. Henriques and A. Vedaldi, "Warped convolutions: Efficient invariance to spatial transformations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1461–1469.

[90] C. Esteves, C. Allen-Blanchette, X. Zhou, and K. Daniilidis, "Polar transformer networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[91] K. S. Tai, P. Bailis, and G. Valiant, "Equivariant transformer networks," *arXiv preprint arXiv:1901.11399*, 2019.

[92] R. Duits, M. Felsberg, G. Granlund, and B. ter Haar Romeny, "Image analysis and reconstruction using a wavelet transform constructed from a reducible representation of the euclidean motion group," *International Journal of Computer Vision*, vol. 72, no. 1, pp. 79–102, 2007.

[93] M. H. Janssen, A. J. Janssen, E. J. Bekkers, J. O. Bescós, and R. Duits, "Design and processing of invertible orientation scores of 3d images," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 9, pp. 1427–1458, 2018.

[94] E. Franken and R. Duits, "Crossing-preserving coherence-enhancing diffusion on invertible orientation scores," *International Journal of Computer Vision*, vol. 85, no. 3, p. 253, 2009.

[95] J. Hannink, R. Duits, and E. Bekkers, "Crossing-preserving multi-scale vesselness," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2014, pp. 603–610.

[96] R. Duits, M. H. Janssen, J. Hannink, and G. R. Sanguinetti, "Locally adaptive frames in the roto-translation group and their applications in medical imaging," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 3, pp. 367–402, 2016.

[97] R. Duits and E. Franken, "Left-invariant diffusions on the space of positions and orientations and their application to crossing-preserving smoothing of hardi images," *International Journal of Computer Vision*, vol. 92, no. 3, pp. 231–264, 2011.

[98] J. Zhang, E. Bekkers, S. Abbasi, B. Dashtbozorg, and B. ter Haar Romeny, "Robust and fast vessel segmentation via gaussian derivatives in orientation scores," in *Proceedings of the International Conference on Image Analysis and Processing*, 2015, pp. 537–547.

[99] J. M. Portegies, R. H. J. Fick, G. R. Sanguinetti, S. P. Meesters, G. Girard, and R. Duits, "Improving fiber alignment in hardi by combining contextual pde flow with constrained spherical deconvolution," *PloS One*, vol. 10, no. 10, e0138122, 2015.

[100] M. S. Hefny, T. Okada, M. Hori, Y. Sato, and R. E. Ellis, "A liver atlas using the special euclidean group," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 238–245.

[101] B. Hou, N. Miolane, B. Khanal, M. C. Lee, A. Alansary, S. McDonagh, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz, "Computing cnn loss and gradients for pose estimation with riemannian geometry," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 756–764.

[102] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-euclidean framework for statistics on diffeomorphisms," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2006, pp. 924–931.

[103] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[104] X. Pennec, S. Sommer, and T. Fletcher, *Riemannian Geometric Statistics in Medical Image Analysis*. 2019.

[105] M. Winkels and T. S. Cohen, "3d g-cnns for pulmonary nodule detection," in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[106] X. Li, L. Yu, C.-W. Fu, and P.-A. Heng, "Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images," in *Proceedings of the International Skin Imaging Collaboration Workshop*, 2018, pp. 235–243.

[107] B. Chidester, T.-V. Ton, M.-T. Tran, J. Ma, and M. N. Do, "Enhanced rotation-equivariant u-net for nuclear segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 1097–1104.

[108] S. Graham, D. Epstein, and N. Rajpoot, "Rota-net: Rotation equivariant network for simultaneous gland and lumen segmentation in colon histology images," in *Proceedings of the European Congress on Digital Pathology (ECDP)*, 2019, pp. 109–116.

[109] B. Chidester, T. Zhou, M. N. Do, and J. Ma, "Rotation equivariant and invariant neural networks for microscopy image analysis," *Bioinformatics*, vol. 35, no. 14, pp. i530–i537, 2019.

[110] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 210–218.

[111] M. Lafarge, J. Pluim, K. Eppenhof, and M. Veta, "Learning domain-invariant representations of histological images," *Frontiers in Medicine*, vol. 6, p. 162, 2019.

[112] R. Duits, "Perceptual organization in image analysis," Eindhoven University of Technology, the Netherlands, 2005.

[113] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[114] A. Osokin, A. Chessel, R. E. Carazo Salas, and F. Vaggi, "Gans for biological image synthesis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2233–2242.

[115] M. Gadelha, S. Maji, and R. Wang, "3D shape induction from 2D views of multiple objects," in *Proceedings of the International Conference on 3D Vision*, 2017, pp. 402–411.

[116] M. Hejrati and D. Ramanan, "Analysis by synthesis: 3D object recognition by object reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2449–2456.

[117] L Gatys, A. Ecker, and M Bethge, "Texture synthesis using convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 262–270.

[118] L. Gatys, A. Ecker, and M Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423.

[119] J. C. Caicedo, S. Singh, and A. E. Carpenter, "Applications in image-based profiling of perturbations," *Current Opinion in Biotechnology*, vol. 39, pp. 134–142, 2016.

[120] V. Ljosa, P. D. Caie, R. Ter Horst, K. L. Sokolnicki, E. L. Jenkins, S. Daya, M. E. Roberts, T. R. Jones, S. Singh, A. Genovesio, *et al.*, "Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment," *Journal of Biomolecular Screening*, vol. 18, no. 10, pp. 1321–1329, 2013.

[121] D. M. Ando, C. McLean, and M. Berndl, "Improving phenotypic measurements in high-content imaging screens," *bioRxiv*, p. 161 422, 2017.

[122] J. C. Caicedo, C. McQuin, A. Goodman, S. Singh, and A. E. Carpenter, "Weakly supervised learning of single-cell feature embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9309–9318.

[123] N. Pawlowski, J. C. Caicedo, S. Singh, A. E. Carpenter, and A. Storkey, "Automating morphological profiling with generic deep convolutional networks," *BioRxiv*, p. 085 118, 2016.

[124] W. J. Godinez, I. Hossain, S. E. Lazic, J. W. Davies, and X. Zhang, "A multi-scale convolutional neural network for phenotyping high-content cellular images," *Bioinformatics*, vol. 33, no. 13, pp. 2010–2019, 2017.

[125] W. J. Godinez, I. Hossain, and X. Zhang, "Unsupervised phenotypic analysis of cellular images with multi-scale convolutional neural networks," *BioRxiv*, p. 361 410, 2018.

[126] P. Goldsborough, N. Pawlowski, J. C. Caicedo, S. Singh, and A. Carpenter, "CytoGAN: Generative modeling of cell images," *bioRxiv*, p. 227 645, 2017.

[127] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[128] C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, *et al.*, "CellProfiler 3.0: Next-generation image processing for biology," *PLoS Biology*, vol. 16, no. 7, e2005970, 2018.

[129] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, "Multidimensional drug profiling by automated microscopy," *Science*, vol. 306, no. 5699, pp. 1194–1198, 2004.

[130] D. W. Young, A. Bender, J. Hoyt, E. McWhinnie, G.-W. Chirn, C. Y. Tao, J. A. Tallarico, M. Labow, J. L. Jenkins, T. J. Mitchison, *et al.*, "Integrating high-content screening and ligand-target prediction to identify mechanism of action," *Nature Chemical Biology*, vol. 4, no. 1, pp. 59–68, 2008.

[131] F. Reisen, A. Sauty De Chalon, M. Pfeifer, X. Zhang, D. Gabriel, and P. Selzer, "Linking phenotypes and modes of action through high-content screen fingerprints," *Assay and Drug Development Technologies*, vol. 13, no. 7, pp. 415–427, 2015.

[132] J. L. Ochoa, W. M. Bray, R. S. Lokey, and R. G. Linington, "Phenotype-guided natural products discovery using cytological profiling," *Journal of Natural Products*, vol. 78, no. 9, pp. 2242–2248, 2015.

[133] Y. Ohya, J. Sese, M. Yukawa, F. Sano, Y. Nakatani, T. L. Saito, A. Saka, T. Fukuda, S. Ishihara, S. Oka, *et al.*, "High-dimensional and large-scale phenotyping of yeast mutants," *Proceedings of the National Academy of Sciences*, vol. 102, no. 52, pp. 19 015–19 020, 2005.

[134] F. Fuchs, G. Pau, D. Kranz, O. Sklyar, C. Budjan, S. Steinbrink, T. Horn, A. Pedal, W. Huber, and M. Boutros, “Clustering phenotype populations by genome-wide RNAi and multiparametric imaging,” *Molecular Systems Biology*, vol. 6, no. 1, p. 370, 2010.

[135] M. H. Rohban, S. Singh, X. Wu, J. B. Berthet, M.-A. Bray, Y. Shrestha, X. Varelas, J. S. Boehm, and A. E. Carpenter, “Systematic morphological profiling of human gene and allele function via cell painting,” *Elife*, vol. 6, e24060, 2017.

[136] T. Horn, T. Sandmann, B. Fischer, E. Axelsson, W. Huber, and M. Boutros, “Mapping of signaling networks through synthetic genetic interaction analysis by RNAi,” *Nature Methods*, vol. 8, no. 4, pp. 341–346, 2011.

[137] C. Laufer, B. Fischer, M. Billmann, W. Huber, and M. Boutros, “Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping,” *Nature Methods*, vol. 10, no. 5, pp. 427–431, 2013.

[138] C. Laufer, B. Fischer, W. Huber, and M. Boutros, “Measuring genetic interactions in human cells by RNAi and imaging,” *Nature Protocols*, vol. 9, no. 10, pp. 2341–2353, 2014.

[139] B. Rauscher, F. Heigwer, L. Henkel, T. Hielscher, O. Voloshanenko, and M. Boutros, “Toward an integrated map of genetic interactions in cancer cells,” *Molecular Systems Biology*, vol. 14, no. 2, e7656, 2018.

[140] O. Z. Kraus, J. L. Ba, and B. J. Frey, “Classifying and segmenting microscopy images with deep multiple instance learning,” *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.

[141] X. Ruan and R. F. Murphy, “Evaluation of methods for generative modeling of cell and nuclear shape,” *Bioinformatics*, vol. 35, no. 14, pp. 2475–2485, 2019.

[142] G. R. Johnson, R. M. Donovan-Maiye, and M. M. Maleckar, “Generative modeling with conditional autoencoders: Building an integrated cell,” *arXiv preprint arXiv:1705.00092*, 2017.

[143] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “Beta-vae: Learning basic visual concepts with a constrained variational framework,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[144] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 1558–1566.

[145] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[146] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, “Adversarially learned inference,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[147] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, “Variational approaches for auto-encoding generative adversarial networks,” *arXiv preprint arXiv:1706.04987*, 2017.

[148] J. C. Caicedo, S. Cooper, F. Heigwer, S. Warchal, P. Qiu, C. Molnar, A. S. Vasilevich, J. D. Barry, H. S. Bansal, O. Kraus, *et al.*, “Data-analysis strategies for image-based cell profiling,” *Nature Methods*, vol. 14, no. 9, pp. 849–863, 2017.

[149] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, “Annotated high-throughput microscopy image sets for validation.,” *Nature Methods*, vol. 9, no. 7, pp. 637–637, 2012.

[150] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[151] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[152] S. Singh, M.-A. BRAY, T. Jones, and A. Carpenter, “Pipeline for illumination correction of images for high-throughput microscopy,” *Journal of Microscopy*, vol. 256, no. 3, pp. 231–236, 2014.

[153] D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 1530–1538.

[154] C. Belthangady and L. A. Royer, “Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction,” *Nature Methods*, vol. 16, 1215—1225, 2019.

[155] A. BenTaieb and G. Hamarneh, “Adversarial stain transfer for histopathology image analysis,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 792–802, 2017.

[156] T. de Bel, M. Hermsen, J. Kers, J. van der Laak, G. Litjens, *et al.*, “Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology,” in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019, pp. 151–163.

[157] M. W. Lafarge, E. J. Bekkers, J. P. Pluim, R. Duits, and M. Veta, “Roto-translation equivariant convolutional networks: Application to histopathology image analysis,” *Medical Image Analysis*, In press, 2020.

[158] M. Ilse, J. M. Tomczak, and M. Welling, “Deep multiple instance learning for digital histopathology,” in *Handbook of Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 521–546.

[159] N. M. Carleton, G. Lee, A. Madabhushi, and R. W. Veltri, “Advances in the computational and molecular understanding of the prostate cancer cell nucleus,” *Journal of Cellular Biochemistry*, vol. 119, no. 9, pp. 7127–7142, 2018.

[160] A. Gupta, P. J. Harrison, H. Wieslander, N. Pielawski, K. Kartasalo, G. Partel, L. Solorzano, A. Suveer, A. H. Klemm, O. Spjuth, *et al.*, “Deep learning in image cytometry: A review,” *Cytometry Part A*, vol. 95, no. 4, pp. 366–380, 2019.

[161] A. Cruz-Roa, J. Arévalo, A. Judkins, A. Madabhushi, and F. González, “A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning,” in *Proceedings of SPIE Medical Imaging*, vol. 9681, 2015, p. 968 103.

[162] N. Pawlowski, J. C. Caicedo, S. Singh, A. E. Carpenter, and A. Storkey, “Automating morphological profiling with generic deep convolutional networks,” *BioRxiv*, p. 085 118, 2016.

[163] D. M. Ando, C. McLean, and M. Berndl, "Improving phenotypic measurements in high-content imaging screens," *bioRxiv*, p. 161 422, 2017.

[164] N. Bayramoglu and J. Heikkilä, "Transfer learning for cell nuclei classification in histopathology images," in *ECCV*, 2016, pp. 532–539.

[165] L. Hou, K. Singh, D. Samaras, T. M. Kurc, Y. Gao, R. J. Seidman, and J. H. Saltz, "Automatic histopathology image analysis with cnns," in *Proceedings of the New York Scientific Data Summit (NYSDS)*, 2016, pp. 1–6.

[166] B. Hu, Y. Tang, I Eric, C. Chang, Y. Fan, M. Lai, and Y. Xu, "Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1316–1328, 2018.

[167] A. Das, M. S. Nair, and S. D. Peter, "Sparse representation over learned dictionaries on the riemannian manifold for automated grading of nuclear pleomorphism in breast cancer," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1248–1260, 2018.

[168] J. Xu, L. Xiang, R. Hang, and J. Wu, "Stacked sparse autoencoder (ssae) based framework for nuclei patch classification on breast cancer histopathology," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2014, pp. 999–1002.

[169] M. W. Lafarge, J. C. Caicedo, A. E. Carpenter, J. P. Pluim, S. Singh, and M. Veta, "Capturing single-cell phenotypic variation via unsupervised representation learning," in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, vol. 102, 2018, pp. 315–325.

[170] K. D. Yang, K. Damodaran, S. Venkatchalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler, "Autoencoder and optimal transport to infer single-cell trajectories of biological processes," *bioRxiv*, p. 455 469, 2018.

[171] G. F. Schau, G. Thibault, M. A. Dane, J. W. Gray, L. M. Heiser, and Y. H. Chang, "Variational autoencoding tissue response to microenvironment perturbation," in *Proceedings of SPIE Medical Imaging*, 2019, p. 109491M.

[172] V. Murthy, L. Hou, D. Samaras, T. M. Kurc, and J. H. Saltz, "Center-focusing multi-task cnn with injected features for classification of glioma nuclear images," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 834–841.

[173] T.-H. Song, V. Sanchez, H. EIDaly, and N. M. Rajpoot, "Hybrid deep autoencoder with curvature gaussian for detection of various types of cells in bone marrow trephine biopsy images," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2017, pp. 1040–1043.

[174] Y. Feng, L. Zhang, and Z. Yi, "Breast cancer cell nuclei classification in histopathology images using deep neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 2, pp. 179–191, 2018.

[175] C.-H. Huang and D. Racoceanu, "Exclusive autoencoder (XAE) for nucleus detection and classification on hematoxylin and eosin (H&E) stained histopathological images," *arXiv preprint arXiv:1811.11243*, 2018.

[176] C.-H. Huang, A. Veillard, L. Roux, N. Loménie, and D. Racoceanu, “Time-efficient sparse analysis of histopathological whole slide images,” *Computerized Medical Imaging and Graphics*, vol. 35, pp. 579–591, 2011.

[177] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, “Biological interpretation of morphological patterns in histopathological whole-slide images,” in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2012, 218–225.

[178] J. Arevalo, A. Cruz-Roa, and F. A. González, “Hybrid image representation learning model with invariant features for basal cell carcinoma detection,” in *Proceedings of SPIE Medical Information Processing and Analysis*, vol. 8922, 2013, p. 89220M.

[179] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, “Classification of tumor histopathology via sparse feature learning,” in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 410–413.

[180] H. Chang, N. Nayak, P. T. Spellman, and B. Parvin, “Characterization of tissue histopathology via predictive sparse decomposition and spatial pyramid matching,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013, pp. 91–98.

[181] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, “A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013, pp. 403–410.

[182] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I Eric, and C. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1626–1630.

[183] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, “Classification of histology sections via multispectral convolutional sparse coding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3081–3088.

[184] A. Cruz-Roa, J. Arévalo, A. Basavanhally, A. Madabhushi, and F. González, “A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation,” in *Proceedings of SPIE Medical Imaging*, vol. 9287, 2015, 92870G.

[185] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, “Stacked predictive sparse decomposition for classification of histology sections,” *International Journal of Computer Vision*, vol. 113, no. 1, pp. 3–18, 2015.

[186] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, “Histopathological image classification using discriminative feature-oriented dictionary learning,” *IEEE Transactions on Medical Imaging*, vol. 35, pp. 738–751, 2015.

[187] S. Otálora, A. Cruz-Roa, J. Arevalo, M. Atzori, A. Madabhushi, A. R. Judkins, F. González, H. Müller, and A. Depeursinge, "Combining unsupervised feature learning and riesz wavelets for histopathology image representation: Application to identifying anaplastic medulloblastoma," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 581–588.

[188] J. Arevalo, A. Cruz-Roa, V. Arias, E. Romero, and F. A. González, "An unsupervised feature learning framework for basal cell carcinoma image analysis," *Artificial Intelligence in Medicine*, vol. 64, no. 2, pp. 131–145, 2015.

[189] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, I Eric, and C. Chang, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 947–951.

[190] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2424–2433.

[191] J. T. Kwak and S. M. Hewitt, "Multiview boosting digital pathology analysis of prostate cancer," *Computer Methods and Programs in Biomedicine*, vol. 142, pp. 91–99, 2017.

[192] C. T. Sari and C. Gunduz-Demir, "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1139–1149, 2018.

[193] W. Bulten and G. Litjens, "Unsupervised prostate cancer detection on h&e using convolutional adversarial autoencoders," *arXiv preprint arXiv:1804.07098*, 2018.

[194] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," vol. 50, no. 9, pp. 3950–3962, 2020.

[195] H. Muhammad, C. S. Sigel, G. Campanella, T. Boerner, L. M. Pak, S. Büttner, J. N. IJzermans, B. G. Koerkamp, M. Doukas, W. R. Jarnagin, *et al.*, "Unsupervised subtyping of cholangiocarcinoma using a deep clustering convolutional autoencoder," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019, pp. 604–612.

[196] C. Liu, Y. Huang, L. Han, J. A. Ozolek, and G. K. Rohde, "Hierarchical feature extraction for nuclear morphometry-based cancer diagnosis," in *Proceedings of the International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis in Conjunction with MICCAI*, 2016, pp. 219–227.

[197] B. Chidester, M. N. Do, and J. Ma, "Discriminative bag-of-cells for imaging-genomics," in *Proceedings of the Pacific Symposium on Biocomputing*, 2018, pp. 319–330.

[198] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 2127–2136.

[199] A. Momeni, M. Thibault, and O. Gevaert, "Deep recurrent attention models for histopathological image analysis," *BioRxiv*, p. 438 341, 2018.

[200] M. Combalia and V. Vilaplana, "Monte-carlo sampling applied to multiple instance learning for whole slide image classification," in *Proceedings of the International Workshop on Deep Learning in Medical Image Analysis in Conjunction with MICCAI*, 2018, pp. 274–281.

[201] J. M. Tomczak, M. Ilse, M. Welling, M. Jansen, H. G. Coleman, M. Lucas, K. de Laat, M. de Bruin, H. Marquering, M. J. van der Wel, *et al.*, "Histopathological classification of precursor lesions of esophageal adenocarcinoma: A deep multiple instance learning approach," in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[202] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[203] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz, "Unsupervised histopathology image synthesis," *arXiv preprint arXiv:1712.05021*, 2017.

[204] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz, "Robust histopathology image analysis: To label or to synthesize?" In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8533–8542.

[205] A. C. Quiros, R. Murray-Smith, and K. Yuan, "Pathology GAN: learning deep representations of cancer tissue," *arXiv preprint arXiv:1907.02644*, 2019.

[206] N. Dey, A. Chen, and S. Ghafurian, "Group equivariant generative adversarial networks," *arXiv preprint arXiv:2005.01683*, 2020.

[207] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 4114–4124.

[208] B. Dai and D. Wipf, "Diagnosing and enhancing VAE models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[209] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "Diva: Domain invariant variational autoencoders," in *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, 2020, pp. 322–348.

[210] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2018.

[211] S. Graham, D. Epstein, and N. Rajpoot, "Dense steerable filter cnns for exploiting rotational symmetry in histology images," *IEEE Transactions on Medical Imaging*, vol. (In Press), 2020.

[212] Y. J. Heng, S. C. Lester, G. M. Tse, R. E. Factor, K. H. Allison, L. C. Collins, Y.-Y. Chen, K. C. Jensen, N. B. Johnson, J. C. Jeong, *et al.*, "The molecular basis of breast cancer pathological phenotypes," *The Journal of Pathology*, vol. 241, pp. 375–391, 2017.

[213] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[214] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[215] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3852–3861.

[216] K. Sirinukunwattana, E. Domingo, S. D. Richman, K. L. Redmond, A. Blake, C. Verrill, S. J. Leedham, A. Chatzipli, C. Hardy, C. M. Whalley, *et al.*, "Image-based consensus molecular subtype (imcms) classification of colorectal cancer using deep learning," *Gut*, vol. (In Press), 2020.

[217] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5334–5344.

[218] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: training deep neural networks with biased data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9012–9020.

[219] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, 2019, pp. 6450–6461.

[220] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[221] D. W. Romero, E. J. Bekkers, J. M. Tomczak, and M. Hoogendoorn, "Attentive group equivariant convolutional networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[222] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101 535, 2019.

[223] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *Procceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*, 2019, pp. 249–261.

[224] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos, "Deforming autoencoders: Unsupervised disentangling of shape and appearance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 650–665.

[225] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 955–10 964.

[226] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[227] K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, “Multi-object representation learning with iterative variational inference,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, 2424—2433.

[228] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with VQ-VAE-2,” in *Advances in Neural Information Processing Systems*, 2019, pp. 14 866–14 876.

[229] D. Nielsen, P. Jaini, E. Hoogeboom, O. Winther, and M. Welling, “Survae flows: Surjections to bridge the gap between vaes and flows,” *arXiv preprint arXiv:2007.02731*, 2020.

[230] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift,” in *Advances in Neural Information Processing Systems*, 2019, pp. 13 991–14 002.

[231] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung, “Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis,” *Nature Cancer*, vol. 1, no. 8, pp. 800–810, 2020.

[232] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, *et al.*, “Pan-cancer image-based detection of clinically actionable genetic alterations,” *Nature Cancer*, vol. 1, no. 8, pp. 789–799, 2020.

[233] S. Blom, L. Paavolainen, D. Bychkov, R. Turkki, P. Mäki-Teeri, A. Hemmes, K. Välimäki, J. Lundin, O. Kallioniemi, and T. Pellinen, “Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis,” *Scientific Reports*, vol. 7, no. 1, pp. 1–13, 2017.

[234] Y. Goltsev, N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan, “Deep profiling of mouse splenic architecture with codex multiplexed imaging,” *Cell*, vol. 174, no. 4, pp. 968–981, 2018.

[235] C. C. Anyaegbu, T. F. Lee-Pullen, T. J. Miller, T. N. Abel, C. F. Platell, and M. J. McCoy, “Optimisation of multiplex immunofluorescence for a non-spectral fluorescence scanning system,” *Journal of Immunological Methods*, vol. 472, pp. 25–34, 2019.

[236] D. Jaishankar, C. Cosgrove, R. J. Deaton, and I. C. Le Poole, “A rapid method for multispectral fluorescence imaging of frozen tissue sections,” *Journal of Visualized Experiments*, no. 157, e60806, 2020.

[237] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai, H. H. M. Ng, D. Wu, Y. T. F. Wee, J. C. T. Lim, J. Yeong, and T. K. H. Lim, “Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy,” *Cancer Communications*, vol. 40, no. 4, pp. 135–153, 2020.

[238] A. Achille, T. Eccles, L. Matthey, C. Burgess, N. Watters, A. Lerchner, and I. Higgins, “Life-long disentangled representation learning with cross-domain latent homologies,” in *Advances in Neural Information Processing Systems*, 2018, pp. 9873–9883.

Summary

Learning Invariant Representations of Images for Computational Pathology

The on-going development of whole-slide imaging techniques with their integration in pathology labs have enabled the digitization of glass slides of tissue specimens as an alternative to the use of conventional bright field microscopes. This digitization of pathology labs has led to an increasing demand for automated, reliable, robust and fast histopathology image analysis systems that can quantify disease-associated patterns in digital slides to support pathologists in their routine workflow.

The development of such systems has become possible thanks to the creation and release of large histopathology image datasets that give opportunities for researchers to develop and validate new algorithms. To leverage this large amount of image data made available, the computational pathology community has established machine learning and more specifically convolutional neural networks (CNNs) as the prominent methodology to achieve state-of-the-art performances across many histopathology image classification tasks.

The key principle behind the success of this methodology is that CNN-based models have the ability to learn complex informative high-level features of histopathology images that are predictive of target disease-associated quantities of interest. As this feature representation is learned directly from training examples, the subsequent system does not rely on possibly biased hand-crafted features and can therefore effectively generalize to unseen images encountered in a practical clinical context.

Yet, learning a representation from histopathology image data comes with some methodological challenges that are addressed in this thesis. One main characteristic of histopathology images is the fact that they exhibit a high variability of appearances that are irrelevant to solve a task of interest. An example of such a factor of variability is the arbitrary orientation in which tissue specimens are digitally viewed. Another example is any change of appearance related to the preparation or the acquisition procedure of a digital slide. These examples of factors are independent of the inherent morphological characteristics of the imaged tissues and can thus be treated as irrelevant. Without taking these data biases into account, these irrelevant factors can still be captured in the learned representation, thus affecting the trained models in an unpredictable manner, making the developed systems unreliable to some extent. In this thesis, new frameworks are proposed to constrain machine learning models and provide them with invariance properties that improve their robustness to the irrelevant variability of the data. Comparative analyses of the developed methods across multiple datasets and classification tasks are presented.

The first chapters of the thesis focus on yielding invariances in the context of supervised learning. To make CNN-based models robust to slide-specific variations of appearance, domain-adversarial training is investigated as a solution to directly constrain a task-driven learned representation to be invariant to any slide-related variability present in the training data. A comparative analysis against conventional methods on two tasks is presented (inter-lab generalization of a mitosis classifier and multi-organ generalization of a nuclei segmentation system). To make deep learning models robust to the arbitrary global orientations of tissue specimens, a new type of convolutional layer is introduced to encode the orientation information of images into the learned representation. As a result, CNN-based models equipped with this operation are shown to achieve better performances on three different tasks in comparison to baseline counterparts. Regarding tri-dimensional rotational invariance, a qualitative proof-of-concept study is presented, giving insights on the possibility for deep learning models to represent the underlying tri-dimensional structure of the imaged tissue slices from single two-dimensional views. This opens research directions towards granting additional rotational invariances into a learned representation.

As opposed to supervised learning, learning a representation from data solely is a promising research direction with significant potential for computational pathology applications. This motivates the second part of the thesis which concerns yielding invariances into unsupervised learning frameworks. An adversarial-driven extension of the variational auto-encoder framework that enables learning a useful latent representation of cell images that can reconstruct images with a high quality is presented. This framework is then further developed to disentangle the orientation information of tissue image patches in a partitioned learned representation. This structure is shown to produce higher performances than baseline counterparts for slide-level classification tasks.

About the Author

Maxime Lafarge was born in Neuilly-sur-Seine, France, in 1992. After obtaining his baccalauréat at lycée Dumont D'Urville in Toulon, France, he was trained in the classes préparatoires aux grandes écoles at lycée Jeanne D'Albret in Saint-Germain-en-Laye, France. He then enrolled for the joint ISBS biomedical engineering program of Paris-Est Créteil University and of the École Supérieure d'Ingénieurs en Électrotechnique et Électronique. He gained experience in biomedical image analysis during the research internships he performed at the Fondation Rothschild Hospital, France, at Chiba University, Japan and at Philips Research, The Netherlands. He obtained his Diplôme d'Ingénieur and double Master's degree in Computer Science in 2015. From 2016 to 2020, he was a PhD candidate in the Medical Image Analysis group at Eindhoven University of Technology and focused on designing machine learning methods tailored for histopathology image analysis. The progress of his PhD project was further stimulated over the course of a 3-month visit at the Imaging Platform of the Broad Institute in Cambridge, MA, USA and by active collaborations with researchers from the Pathology department of UMC Utrecht and from the department of Mathematics and Computer Science of Eindhoven University of Technology. The results of the research conducted during this period are presented in this thesis.

List of Publications

M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, “Domain-adversarial neural networks to address the appearance variability of histopathology images”, *in Proceedings of the International Workshop on Deep Learning in Medical Image Analysis in Conjunction with MICCAI*, 2017, pp. 83–91.

P. Moeskops, M. Veta, **M. W. Lafarge**, K. A. Eppenhof, and J. P. Pluim, “Adversarial training and dilated convolutions for brain MRI segmentation”, *in Proceedings of the International Workshop on Deep Learning in Medical Image Analysis in Conjunction with MICCAI*, 2017, pp. 56-64

M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, “Inferring a third spatial dimension from 2D histological images”, *in Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 586-589.

E. J. Bekkers, **M. W. Lafarge**, M. Veta, K. A. Eppenhof, J. P. Pluim, and R. Duits, “Roto-translation covariant convolutional networks for medical image analysis”, *in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 440–448.

M. W. Lafarge, J. C. Caicedo, A. E. Carpenter, J. P. Pluim, S. Singh, and M. Veta, “Capturing single-cell phenotypic variation via unsupervised representation learning”, *in Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, vol. 102, 2018, pp. 315–325.

K. A. Eppenhof, **M. W. Lafarge**, P. Moeskops, M. Veta, and J. P. Pluim, “Deformable image registration using convolutional neural networks”, *in Proceedings of SPIE Medical Imaging*, vol. 10574, 2018, p. 105740S.

M. W Lafarge, J. P. Pluim, K. A. Eppenhof, and M. Veta, “Learning domain-invariant representations of histological images”, *Frontiers in Medicine*, vol. 6, p. 162, 2019.

K. A. Eppenhof, **M. W. Lafarge**, and J. P. Pluim, “Progressively growing convolutional networks for end-to-end deformable image registration”, *in Proceedings of SPIE Medical Imaging*, vol. 10949, 2019, p. 109491C.

K. A. Eppenhof, **M. W. Lafarge**, M. Veta, and J. P. Pluim, “Progressively trained convolutional neural networks for deformable image registration”, *IEEE Transactions on Medical Imaging (TMI)*, vol. 39(5), pp. 1594-1604, 2019.

M. W. Lafarge, E. J. Bekkers, J. P. Pluim, R. Duits, and M. Veta, “Roto-translation equivariant convolutional networks: Application to histopathology image analysis”, *Medical Image Analysis*, In press, 2020.

M. W. Lafarge, J. P. Pluim, and M. Veta. “Orientation-Disentangled Unsupervised Representation Learning for Computational Pathology”, 2020 [In review, arXiv preprint arXiv:2008.11673].

Acknowledgments

I would like to acknowledge numerous people who have impacted this work by the support they provided.

I would like to express my sincere gratitude to my promotor Josien Pluim for offering me an exceptional PhD opportunity: the research challenges in computational pathology were the perfect ingredients that have fed my motivation and creativity over the past years. Thank you for the freedom you have given me, letting me explore the wackiest ideas I could think of to process histology images. I am grateful for the support and encouragement you have given me to always diversify my academic activities through teaching opportunities, writing my first grant proposal, allowing me to expand my work during a research visit, being a reviewer for several journals and giving me multiple opportunities to give talks about my research.

To my co-promotor and daily supervisor Mitko Veta, thank you for trusting me and introducing me to the exciting field of computational pathology. Thank you for sharing your experience: you have given me the right resources and guidance to quickly comprehend the fundamental of deep learning and how to apply this knowledge for computational pathology. I greatly appreciate that you always took the time to listen and challenge my ideas, but also for continuously encouraging me to exchange know-hows with colleagues and collaborators. I keep great memories of the travels we have done together to attend conferences.

I would like to thank Erik Bekkers and Remco Duits for exchanging know-hows about their group-theoretical expertise, which has led to our successful joint development of the rotation-equivariant deep learning framework presented in this thesis. The constructive feedback you have given me throughout the project was formative and our collaboration has contributed to expanding my research interests.

I would like to thank Anne Carpenter, Shantanu Singh and Juan Caicedo for hosting me at the Imaging Platform of the Broad Institute of MIT and Harvard. I am grateful for the collaboration we have initiated to further develop advanced generative models for cell profiling applications. I have greatly appreciated the hospitality of your group: thank you for making these three months in the Boston area so enjoyable. Also, I would like to credit the Neubias COST Action for supporting the work conducted during this research visit with a travel grant.

I would like to thank prof.dr. van Diest, prof.dr. Martel, prof.dr. Vilanova, dr. van der Laak and dr.ir. Duits for being part of my doctorate committee and for assessing my work.

Koen Eppenhof, it was a privilege to share an office with you for several years. Thank you for being such a supportive officemate, always ready to take some time to discuss, brainstorm and develop constructive criticism about any fresh idea we have had come up with.

Suzanne Wetstein and Friso Heslinga, thank you for your support and friendship over the past years, especially for keeping our regular stand-up meetings productive in this extended time of remote working.

I would like to thank all my colleagues of the Medical Image Analysis group (Alexander, Andrey, Aymen, Bart, Behdad, Fan, Ishaan, Jiong, Marcel, Maureen, Nastaran, Pim, Ralf, Rina, Samaneh, Seb, Sina, Sylvia, Veronika and Yasmina) for the insightful and inspiring discussions we have had and for all the advice you have given me.

Merci à ma famille et à mes amis pour leur soutien et pour toutes les distractions tant bienvenues que vous m’avez apportées au cours de ces années si spéciales.