Spot the Difference: Topological Anomaly Detection via Geometric Alignment

Anonymous Author(s)
Affiliation
Address
email

Abstract

Geometric alignment appears in a variety of applications, ranging from domain adaptation, optimal transport, and normalizing flows in machine learning; optical flow and learned augmentation in computer vision and deformable registration within biomedical imaging. A recurring challenge is the alignment of domains whose topology is not the same; a problem that is routinely ignored, potentially introducing bias in downstream analysis. As a first step towards solving such alignment problems, we propose an unsupervised topological difference detection algorithm. The model is based on a conditional variational auto-encoder and detects topological anomalies with regards to a reference alongside the registration step. We consider both a) topological changes in the image under spatial variation and b) unexpected transformations. Our approach is validated on a proxy task of unsupervised anomaly detection in images.

1 Introduction

Geometric alignment is a fundamental component of widely different algorithms, ranging from domain adaptation [6], optimal transport [34] and normalizing flows [29, 35] in machine learning; optical flow [18, 43] and learned augmentation [17] in computer vision, and deformable registration [5, 15, 16, 33, 44] within biomedical imaging. A recurring challenge is the alignment of domains whose topology is not the same. When the objects to be aligned are probability distributions [29], this appears when distributions have different numbers of modes whose support is separated into separate connected components. When the objects to be aligned are scenes or natural images, the problem occurs with occlusion or temporal changes [43]. In biomedical image registration, the problem is very common and happens when the studied anatomy differs from "standard" anatomy [30]. Despite being extremely common, this problem is routinely ignored or accepted as inevitable, potentially introducing bias in downstream analysis.

We study the special case of medical image registration of brain MRI scans, where tumors give common examples of anatomies that are topologically different from healthy brains. In deformable image registration, a "moving image" is mapped via a nonlinear transformation to make it as similar as possible to a "target" image, enabling matching local features or transferring information from one image to another. It is common to numerically stabilize the estimation of the transformation by constraining the predicted transformation to be diffeomorphic, that is, bijective and continuously differentiable. In particular, this implies that a common topology is assumed across all images [12, 15]. This topology is often provided by a common template image $\mathbf{I}_{\text{template}}$, from which all other images are obtained via the transformation Φ from the group of diffeomorphisms $\mathcal{G}$. Under this

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

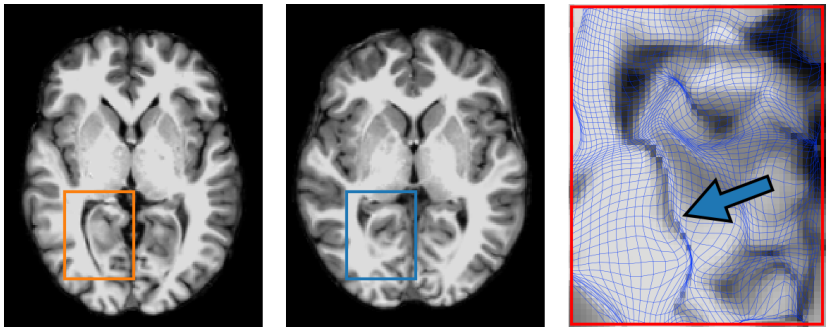


Figure 1: Example of the adverse effects of topological differences on image registration. We register a brain with an extended ventricle (orange square) to a closed ventricle (blue square). Due to the topological mismatch, the resulting registration (enlarged red square) shrinks the ventricle to a thin line. The transformation, visualized by the blue lines morphed from a uniform grid, distorts the surrounding tissue.

common topology assumption, the set of all images is given by

$$\mathcal{I} = \{\mathbf{I}_{\text{template}} \circ \Phi | \Phi \in \mathcal{G}\} \ .$$

Topological differences in biological anatomy can be caused by a variety of processes. For instance, tumor growth or surgical tissue removal alters the topology of an image and can also lead to replacement or deformation of nearby tissue through swellings or fluids, which cannot be mapped to the original image. Similarly, tumor cells invade surrounding tissue, changing the features of tissue that *does* diffeomorphically correspond to tissue in a normal brain. As most registration algorithms align images based on intensity, e.g. minimizing mean squared error (MSE), these tissue changes make it difficult to map images correctly. The strong local deformations required to deal with the non-diffeomorphic part of the image inevitably also deform the surrounding area, leading to distorted transformation fields in topologically matching parts of the image [30]. These transformation fields adversely affect downstream tasks, for example indicating false size changes in adjacent regions.

Previous work on aligning topologically inconsistent domains. Attempting to relax the diffeomorphism assumptions of image registration is not new. In the context of organs sliding against each other, several approaches exist, most of which rely on pre-annotating the sliding boundary using organ segmentation [9, 19, 31, 36, 39], with a few extensions to un-annotated images [32, 38].

For the case where topological holes are created or removed in the domain, several methods exist, ranging from masking or weighting of the loss function [23, 25, 26], to growing an artificial insection to correct anatomies [30]. These approaches rely on annotation of the topological differences, which have to be provided manually or by segmentation. An exception is given by Li and Wyatt [26], which detects topological anomalies from the difference between the aligned images. This depends crucially on the ability to find a good diffeomorphic registration *outside* the anomaly, which is difficult all the while the applied transformation is still diffeomorphic.

Our contribution. We propose an unsupervised topological difference detection algorithm. To this end, we train a conditional variational autoencoder for predicting image-to-image alignment, obtaining a per-target-pixel probability of being obtained from the moving image via diffeomorphic transformation. We combine a semantic loss function trained to segment healthy brain regions [7], with a learnable prior of transformations [8], allowing us to incorporate both the reconstruction error, as well as knowledge about the expected transformation strength.

We test the validity of our approach on a proxy task, detecting brain tumors using an image alignment model that was only trained to align tumor-free brains. We also validate our approach by investigating a spatial "topological inconsistency likelihood", and showing that this likelihood is higher in regions where topological inconsistencies are known to be common. Our model is able to detect topological inconsistencies with a purely registration-driven framework, and thus provides the first step towards an end-to-end registration model for images with topological discrepancies.

2 Background

2.1 Notation of images and transformations

We view an image $\mathbf{I}$ interchangeably as two different structures. First, it is a continuous function $\mathbf{I} : \Omega_{\mathbf{I}} \to \mathbb{R}^C$, where $\Omega_{\mathbf{I}} = [0,1]^D$ is the domain of the image, and C the number of channels. This function can be approximated by a grid of n pixels with positions $x_k \in \Omega_{\mathbf{I}}$ leading to the image representation $\mathbf{I}_k^{(c)}$, where c is an index over the channels and $\mathbf{I}_k = (\mathbf{I}_k^{(1)}, \ldots, \mathbf{I}_k^{(C)})^T = \mathbf{I}(x_k)$. Second, this pixel grid is accompanied by a graph structure that encodes the neighbourhood of each pixel. In this view, the set of neighbours of a pixel with index k (for example the 4-neighbourhood of a pixel on the image grid) is referred to as $N(k)$ and $|N(k)|$ is the number of neighbours. The neighborhoods of a pixel gives rise to a graph which can be described via the graph laplacian $\Lambda \in \mathbb{R}^{n \times n}$ with $\Lambda_{k,k} = |N(k)|$ and $\Lambda_{k,k'} = -1$ when pixel $k' \in N(k)$, and zero otherwise.

Applying a spatial transformation $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ to an image is written as $\mathbf{J} = \mathbf{I} \circ \Phi$, which can be seen as its own image with domain $\Omega_{\mathbf{J}} = [0,1]^D$ with pixel coordinates $y_k \in \Omega_{\mathbf{J}}$ and $\mathbf{J}_k = \mathbf{I}(\Phi(y_k))$. The transformation Φ can be seen as a vector field on the image domain which assigns each pixel in $\mathbf{J}$ a position on $\mathbf{I}$ and thus it can be parameterized as a pixel grid $\Phi_k^{(d)}$, $d = 1, \ldots, D$ at the pixel coordinates of $\mathbf{J}$ using $\Phi(y_k) = y_k + \Phi_k$. To make this choice of coordinate system clear, we will refer to a transformation that moves a pixel position from the domain $\Omega_{\mathbf{J}}$ to the corresponding pixel in domain $\Omega_{\mathbf{I}}$ as $\Phi_{\mathbf{J} \to \mathbf{I}}$, whenever it is not clear from the context. If Φ is a diffeomorphism, it can alternatively be parameterized by a vector field V on the tangent space around the identity, where the mapping between the tangent space and the transformation is given by $\Phi = \exp(V)$, which amounts to integration over the vector field [2].

2.2 Variational registration framework

It is possible to phrase the problem of fitting a registration model in terms of variational inference, using an approach similar to conditional variational autoencoders [40]. Here, we summarize the approach taken by [8, 27]. For a D-dimensional image pair $(\mathbf{I}, \mathbf{J})$, we assume that $\mathbf{J}$ is generated from $\mathbf{I}$ by drawing a transformation Φ from a prior distribution $p(\Phi|\mathbf{I})$, apply it to $\mathbf{I}$ and then add some pixel-wise Gaussian noise:

$$p(\mathbf{J}|\mathbf{I}) = \int p_{\text{noise}}(\mathbf{J}|\mathbf{I} \circ \Phi) p(\Phi|\mathbf{I}) \, d\Phi$$

This includes the common topology assumption implicitly via $p(\Phi|\mathbf{I})$, which is typically chosen to produce invertible transformations depending only on the topology of $\mathbf{I}$, as well as the noise model which does not assume systematic changes between $\mathbf{J}$ and $\mathbf{I}$. This model can be learned using variational inference using a proposal distribution $q(\Phi|\mathbf{I}, \mathbf{J})$ with evidence lower bound (ELBO)

$$\log p(\mathbf{J}|\mathbf{I}) \geq E_{q(\Phi|\mathbf{I},\mathbf{J})} \left[\log p_{\text{noise}}(\mathbf{J}|\mathbf{I} \circ \Phi)\right] - KL(q(\Phi|\mathbf{I}, \mathbf{J}) \| p(\Phi|\mathbf{I})) \ . \tag{1}$$

In contrast to variational autoencoders, the decoder is given by the known application of Φ to $\mathbf{I}$. Thus, the degrees of freedom in this model are in the choice of the encoder, prior, and the noise distribution. Dalca et al. [8] proposed to parameterize Φ as a vector field $V_k^{(d)}$ on the tangent space, which turns application of $\Phi = \exp(V)$ into sampling an image with a spatial transformer module [21]. As a prior for this parameterization, they chose a prior independent of $\mathbf{I}$

$$p(\Phi) = \prod_{d=1}^{D} \mathcal{N}\left(V^{(d)} \mid 0, \Lambda^{-1}\right) \ ,$$

where we used the implicit identification of Φ and V and the precision matrix Λ is chosen as the Graph Laplacian over the neighbourhood graph (see notation). Using an encoder that for each pixel proposes $q(V_k^{(d)}|\mathbf{I}, \mathbf{J}) = \mathcal{N}(\mu_k^{(d)}, v_k^{(d)})$, the KL divergence is derived as

$$\mathrm{KL}\left(q(\Phi|\mathbf{I}, \mathbf{J}) \| p(\Phi|\mathbf{I})\right) = \frac{1}{2} \sum_{d=1}^{D} \sum_{k=1}^{n} -\log v_k^{(d)} + |N(k)| v_k^{(d)} + \sum_{l \in N(k)} \left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 + \text{const} \ . \tag{2}$$

It is worth noting that this equation is invariant under translations of μ. This invariance manifests in rank-deficiency of Λ and as a result, const is infinite. Thus, sampling from the prior and bounding

the objective is impossible. Still training with this term works in practice as images are usually pre-aligned with an affine transformation and thus translations are close to zero. We will present a slightly modified approach, rectifying the missing eigenvalue.

3 Detection of topological differences

The variational approach for learning the distribution of transformations introduced before optimizes an ELBO on $\log p(\mathbf{J}|\mathbf{I})$. This information is enough to detect images that contain topological differences under the assumption that these images will overall have a lower likelihood. However, in our application, we need not only to detect the existence but also the position of outliers in the image. For this, we have to ensure that $\log p(\mathbf{J}|\mathbf{I})$ can be decomposed into a likelihood for each pixel of the image. It is immediately obvious by inspection of the ELBO (1) together with the KL-Divergence (2), that the lower bound on $\log p(\mathbf{J}|\mathbf{I})$ can be decomposed into pixel-wise terms if $\log p_{\text{noise}}(\mathbf{J}|\mathbf{I} \circ \Phi)$ can be decomposed as such. To enforce this, we will introduce a general form of error function, which can be decomposed and includes the MSE as a special case. For this, we first map the images I and J to feature maps over the pixel positions k via a mapping $f_k(\mathbf{I}) \in \mathbb{R}^F$ and define the loss as:

$$p_{\text{noise}}(\mathbf{J}|\mathbf{I} \circ \Phi) = \prod_{k=1}^{n} \mathcal{N}(f_k(\mathbf{J})|f_k(\mathbf{I}) \circ \Phi, \Sigma_f) \ , \tag{3}$$

where $\Sigma_f \in \mathbb{R}^{F \times F}$ is a diagonal covariance matrix with variances learned during training.

The ability to decompose the likelihood is not enough for a *meaningful* metric, as we have to ensure that each term is calculated in the correct coordinate system. This depends on the parameterisation and regularisation of Φ. In the approach by Dalca et al. [8] the parameterization V of Φ is defined on the tangent space and consequently the prior is also on this space. Since the connection between Φ and V is given by integration of the vector field, decomposing (2) for a single pixel k will produce estimates based on the local differential of the transformation, but will not take the full path with starting and endpoints into account. Thus, correct cost assignments require integration of (2) over the computed path, which is expensive and suffers from severe integration inaccuracies. Instead, we will use an alternative approach, where we parameterize Φ directly as a vector field on the image domain. This drops the common topology assumption for the transformation, as transformations parameterized this way are not necessarily invertible anymore, yet smoothness is still encouraged by the prior.

Learnable prior Using this parameterization, we extend the approach by Dalca et al. [8] and introduce a parameterized prior on Φ_k that is learned simultaneously with the model:

$$p(\Phi) = \prod_{d=1}^{D} \mathcal{N}\left(\Phi^{(d)} \mid 0, \Lambda_{\alpha\beta}^{-1}\right), \ \ \Lambda_{\alpha\beta} = \alpha\Lambda + \frac{\beta}{n^2}\mathbb{1}\mathbb{1}^T$$

The second term ensures that $\Lambda_{\alpha\beta}$ is invertible, by adding a multiple of the eigenvector $\mathbb{1} = (1, \ldots, 1)^T$. It can be verified easily that $\Lambda\mathbb{1} = 0$. Unlike adding a multiple of the identity matrix to Λ to rectify this issue, adding the missing eigenvalue does not modify the prior in any other way than regularizing the translations. Further, it ensures that the KL divergence of the resulting matrix can be quickly computed up to a constant as α and β do not modify the same eigenvalues. With this, α and β are tuneable parameters that govern the number of expected variations between pixels as well as the number of expected translations. Recomputing the KL-divergence for n transformation vectors in D dimensions leads to

$$\begin{aligned} 2\,\mathrm{KL}\left(q(\Phi|\mathbf{I},\mathbf{J})\|p_{\alpha\beta}(\Phi)\right) = &-(n-1)D\log\alpha - D\log\beta + \beta\sum_{d=1}^{D}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i^{(d)}\right)^2 \\ &+ \sum_{d=1}^{D}\sum_{k=1}^{n} -\log v_k^{(d)} + \left(\alpha|N(k)| + \frac{\beta}{n^2}\right)v_k^{(d)} + \alpha\sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 + \text{const} \end{aligned} \tag{4}$$

Decomposed error metric We define our pixel-wise error measure for outlier detection based on the ELBO (1) with KL-divergence (4) as follows, where we compute $\mu_k^{(d)}$ and $v_k^{(d)}$ via the proposal

distribution $q(\Phi|\mathbf{I},\mathbf{J})$ and pick $\Phi_k^{(d)} = \mu_k^{(d)}$:

$$L_k(\mathbf{J}|\mathbf{I}) = -\log \mathcal{N}(f_k(\mathbf{J})|f_k(\mathbf{I})\circ\Phi,\Sigma_f) + \frac{\beta\mu_k^{(d)}}{n^2}\sum_{d=1}^{D}\sum_{i=1}^{n}\mu_i^{(d)} + \sum_{d=1}^{D} -\log v_k^{(d)} + \left(\alpha|N(k)| + \frac{\beta}{n^2}\right)v_k^{(d)} + \alpha\sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2 \quad . \tag{5}$$

We will treat the loss over all pixels $L(\mathbf{J}|\mathbf{I}) = (L_1(\mathbf{J}|\mathbf{I}),\ldots,L_n(\mathbf{J}|\mathbf{I}))$ as another image with domain and pixel coordinates the same as $\mathbf{J}$. This measure is not symmetric. If $\Phi_{\mathbf{J}\to\mathbf{I}}$ maps a line in $\mathbf{J}$ to an area in $\mathbf{I}$, this will incur a large visible feature along the line. On the other hand, if an area in $\mathbf{J}$ gets mapped to a line in $\mathbf{I}$, the overall error contribution is smoothed out over the area. Moreover, the prior distribution does not treat the distributions $q(\Phi|\mathbf{I},\mathbf{J})$ and $q(\Phi|\mathbf{J},\mathbf{I})$ equally. To rectify the latter, we will compute a bidirectional measure $L_{\text{sym}}(\mathbf{J}|\mathbf{I}) = L(\mathbf{J}|\mathbf{I}) + L(\mathbf{I}|\mathbf{J})\circ\Phi_{\mathbf{J}\to\mathbf{I}}$, where $\Phi_{\mathbf{I}\to\mathbf{J}}$ is the same as the one used to compute $L(\mathbf{J}|\mathbf{I})$. For this measure it holds that if $\Phi_{\mathbf{J}\to\mathbf{I}} = \Phi_{\mathbf{I}\to\mathbf{J}}^{-1}$, we have $L_{\text{sym}}(\mathbf{I}|\mathbf{J}) = L_{\text{sym}}(\mathbf{J}|\mathbf{I})\circ\Phi_{\mathbf{J}\to\mathbf{I}}$ up to interpolation errors caused by the finite coordinate grid.

Outlier detection We can detect outliers using L_{sym} by contrasting the observed deviations with the observed deviations within a larger set of control images $\mathcal{C}$

$$Q(\mathbf{J}) = \mathbb{E}_{\mathbf{I}\in\mathcal{C}}\left[L_{\text{sym}}(\mathbf{J}|\mathbf{I}) - \mathbb{E}_{\mathbf{K}\in\mathcal{C}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K})\right]\circ\Phi_{\mathbf{I}\to\mathbf{J}}\right]. \tag{6}$$

A classifier could be obtained from this score using a learned sigmoid on $Q(\mathbf{J})$ given a set of annotated outlier pixels. Instead, we will use the AUC to measure alignment with regions where we assume outliers.

3.1 Efficient learning of the prior

Training the model with the KL-Divergence (4), leads to a dependency between α, β and $v_k^{(d)}$ of the proposal distribution q. Thus, a bad initialization can lead to slow convergence. However, the prior parameters enter the ELBO in (1) only through the KL-divergence. Thus, it is possible to compute an estimate of the optimal prior parameters given a batch of samples, similar to batch normalization [20]. Optimizing (4) for α and β as expectation over the dataset and omitting constant terms leads to:

$$\min_{\alpha,\beta} 2\,\mathbb{E}_{\mathbf{I},\mathbf{J}}\left[\mathrm{KL}\left(q(\Phi|\mathbf{I},\mathbf{J})\|p_{\alpha\beta}(\Phi)\right)\right] = D\log\mathbb{E}_{\mu,v}\left[\sum_{d=1}^{D}\left(\sum_{k=1}^{n}\mu_k^{(d)}\right)^2 + \sum_{k=1}^{n}v_k^{(d)}\right] - \mathbb{E}_v\left[\sum_{d=1}^{D}\sum_{k=1}^{n}\log v_k^{(d)}\right] + (n-1)D\log\mathbb{E}_{\mu,v}\left[\sum_{d=1}^{D}\sum_{k=1}^{n}|N(k)|v_k^{(d)} + \sum_{l\in N(k)}\left(\mu_k^{(d)} - \mu_l^{(d)}\right)^2\right] + \text{const} \quad , \tag{7}$$

which we use during training. Here, the expectation $\mathbb{E}_{\mu,v}$ refers to computing $q(\Phi|\mathbf{I},\mathbf{J})$ and taking the expectation over all image pairs in the full dataset, which can be approximated using samples from a single batch. For evaluation, we replace this greedy optimum by a time-average of α,β obtained during training.

4 Evaluation

As there exists no standard dataset of annotated topological differences for image registration, we follow and expand on the evaluation strategies of prior work. Following [23, 25, 26], we perform qualitative evaluation on individual images, combined with a quantitative evaluation on a proxy task. We base our evaluation on registration and segmentation of structural Brain-MRI scans. Individual brains are known to be topologically different, in particular at the cortical surface, where the sulci vary significantly [41], and near ventricles, which can either be open cavities, or partially closed [30]. Even more pronounced differences are found in the presence of tumors, available in annotated public

datasets. Hence, our first validation considers the proxy task of detecting brain tumors and edema (swelling) within the brain. This challenging task is usually solved using contrast MRI [28], as edema can not be detected well via texture changes in T1 MRI. However, edemas change the morphology of the brain and can thus be detected indirectly via the large transformations they cause.

For this, we first train our model using a dataset of healthy images and then use (6) to obtain a score for outlier detection. We train the model using the standard MSE and a semantic similarity metric. See Section 4.1 for more details on the dataset and Section 4.2 for details on model and training. Since we compare supervised and unsupervised models, we plot the receiver operating characteristic curve (ROC curve) and compare the area under the curve (AUC) between the models. AUC estimates are bootstrapped on the subject level to obtain error estimates. As labels, we consider a) only detecting the tumor core and b) the core combined with the enlarged region of brain edema surrounding it. We compare our model to the following baselines:

1. Two unsupervised approaches for topological anomaly detection:
   - Li and Wyatt's [26] intensity difference and image gradient-based approach using a deterministic registration model [5] to obtain the transformations.
   - Using the same model, we devise a method based on the Jacobian Determinant of the transformation field $|J_\Phi|$. We expect strong stretching or shrinkage in areas of topological mismatch, which we measure using use the score $\log(|\text{det}J_\Phi|)^2$.

   We adapt both approaches to the task of tumor detection by subtracting the average scores over healthy patients for each pixel via the scheme presented in (6).
2. The approach by An and Cho [1] for unsupervised anomaly detection in images is based on the local reconstruction error of a variational autoencoder. The error score is $\|\mathbf{J} - \text{dec}(\text{enc}(\mathbf{J}))\|^2$, where $\text{enc}(\mathbf{J})$ maps $\mathbf{J}$ to the mean of the variational proposal distribution and dec is the corresponding learned decoder. As the score does not use registration, we cannot use equation (6).
3. A supervised segmentation model trained for tumor segmentation. Since this model requires annotated data, we withhold 80% of the tumor-annotated brain volumes for training and evaluate on the remaining samples. To increase the amount of data available during training, we perform data augmentation by random affine transformations on the 3d volumes before slicing.

In our second evaluation, we investigate whether regions with known topological variability get assigned higher scores in our model. For this we compute the pairwise average score L_{sym} over multiple healthy subjects and register them all to a brain atlas using $\mathbb{E}_{\mathbf{I},\mathbf{K}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K}) \circ \Phi_{\mathbf{I}\to\text{Atlas}}\right]$. We group the scores by their position on the atlas into partitions: cortical surfaces, subcortical regions, and ventricles.

4.1 Data

We utilize two datasets of brain MRI scans, one of subjects without tumors, which in this context indicates relatively normal anatomy, and one of patients exhibiting tumors. All datasets are anonymized, with no protected health information included and participants gave informed consent to data collection. For the normal anatomy set, we combine T1 weighted MRI scans from the ABIDE I [10][1], ABIDE II [11] and OASIS3 [24] studies for atlas-based alignment of *Brain-MRI* scans. Health conditions of the subjects vary, but no gross anatomical abnormalities are present. For the tumor set we use MRI scans from the BraTS2020 brain tumor segmentation challenge [3, 4, 28], which have expert-annotated tumors and edema. We use the T1 weighted MRI scans, and combine labels of the classes necrotic/cystic and enhancing tumor core into a single tumor class.

We perform standard pre-processing on both datasets, including intensity normalization, affine spatial alignment, and skull-stripping using FreeSurfer [13]. From each 3D volume, we extract a center slice of 160×224 pixels. Scans with preprocessing errors are discarded, and the remaining images of the anatomically normal dataset are split 3665/250/250 for train/validation/test. Of the tumor dataset, 84 annotated images with tumors larger than 5cm^2 along the slice are used for evaluation (17 for the supervised approach due to the training-test split of subjects).

[1] CC BY-NC-SA 3.0, `https://creativecommons.org/licenses/by-nc-sa/3.0/`

4.2 Model and training

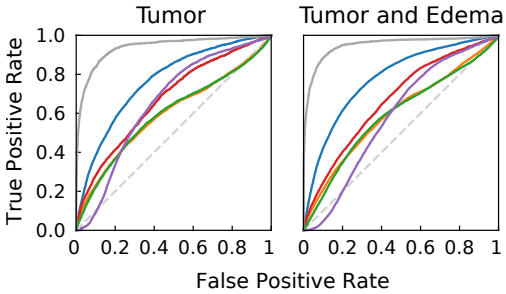
All models evaluated are based on the same U-Net [37] architecture, except An and Cho [1], which we implement using as a spatial VAE following the previously published adaptation to Brain-scans by Venkatakrishnan et al. [42]. The U-Net consists of three encoder and decoder stages of $64, 128, 256$ channels. Each stage consists of a batch normalization [20], a convolutional, and a dropout layer [14]. We experimented with deeper architectures but found they do not increase performance.

In our approach, we use the U-Net to model $p(\Phi|\mathbf{I}, \mathbf{J})$. The output of the last decoder stage is fed through separate convolution layers with linear activation functions to predict the transformation mean and log-scaled variance. Throughout the network, we use LeakyReLu activation functions. The generator step $I \circ \Phi$ is implemented by a parameterless spatial transformer layer [21]. During training of our model, we use the analytical solution for prior parameters α, β (Eq. 7), averaged over the mini-batch of 32 image pairs. For validation and test, we use the running mean recorded during training. The diagonal covariance of the reconstruction loss Σ_f is treated as a trainable parameter.

For all networks, the optimization algorithm used is ADAM [22] with a learning rate of 10^{-4}. We train on two Quadro P6000 GPUs, all models and baselines train in under 12 hours each. Hyperparameters: The network by Venkatakrishnan et al. [42] has $\sigma = 1$ chosen from $\{0.1, 1, 10\}$, based on reconstruction loss on validation set. The deterministic registration model was trained using $\lambda = 0.1$ as in [7]. For [26], the parameters σ of the Gaussian derivative kernel and hyper-parameter K where chosen to maximize the AUC score, selecting $\sigma = 6$, $K = 2$ out of $\{1, \ldots, 9\}^2$.

For the reconstruction loss, we compare two different loss functions. The first is using the MSE as in [8, 26]. The second is a semantic similarity metric similar to [7]. To obtain the semantic image descriptors, we train a similarly structured U-net for segmentation of the anatomically normal set, using labels automatically created with FreeSurfer [13]. From this network, we extract the features of the first three layers, upsample each of the layers to the size of the full image and concatenate them to a single image with $F = 480$ channels. We use this as a feature map in the loss (3). For both the MSE and the semantic loss, we learn the variance parameters while training the variational autoencoder.

4.3 Results



Method	AUC (T)	AUC (T+E)
Ours (Sem. Loss)	$.754 \pm .010$	$.774 \pm .003$
Ours (MSE)	$.582 \pm .011$	$.601 \pm .005$
Li and Wyatt	$.672 \pm .012$	$.668 \pm .013$
Jac. Det.	$.590 \pm .009$	$.595 \pm .010$
An and Cho	$.670 \pm .011$	$.584 \pm .013$
Supervised Seg.	$.933 \pm .016$	$.955 \pm .007$

Figure 2: ROC curves and AUC score on the anomaly detection proxy task. We evaluate with the positive class containing just the tumor core (T) and the tumor core and surrounding edema (T+E). We test models of our method for unsupervised topological difference detection, trained with a semantic loss function and the MSE in the reconstruction term, and compare against unsupervised baselines from image registration (Li and Wyatt [26], Jacobian Determinant) and unsupervised anomaly detection (An and Cho [1]). For reference, we also include a supervised segmentation model.

The ROC curves of all trained models can be seen in Figure 2. For the tumor detection task, the supervised model performed best (AUC 0.93), while our proposed approach with semantic loss performed best among the unsupervised models (AUC 0.75). The two unsupervised approaches by Li and Wyatt [26] (AUC 0.67) and An and Cho [1] (AUC 0.67) performed similarly, but worse than our method. The approaches using the Jacobian determinant (AUC 0.59) and our approach using MSE performed similarly (AUC 0.58) but worse than the other methods. For the combined detection task, all methods performed similarly or slightly better compared to the tumor core detection task, with the exception being the model by An and Cho [1], which performed worse (AUC 0.58).

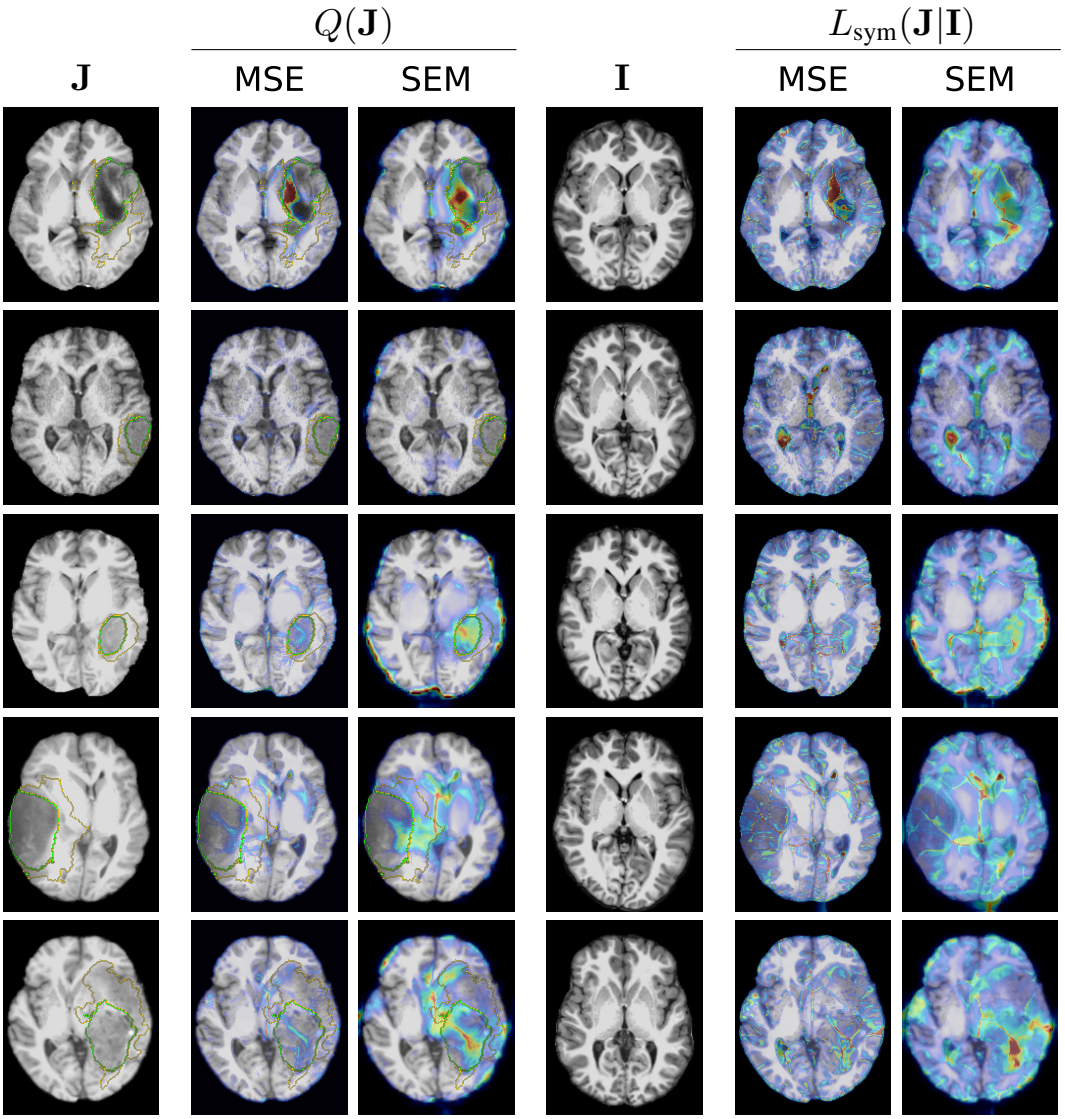


Figure 3: Topological differences detected by our method. Image of tumor brain **J** in column 1, tumor outlined in green, edema outlined in yellow. Heatmaps in columns 2,3 show likelihood of topological differences caused by the anomaly, filtered by (Eq. 6). Heatmaps in columns 5,6 show the unfiltered topological differences between **J** and a randomly selected example image **I** shown in column 4, as measured by L_{sym}. Heatmaps are overlayed on top of image **J** for easier comparison.

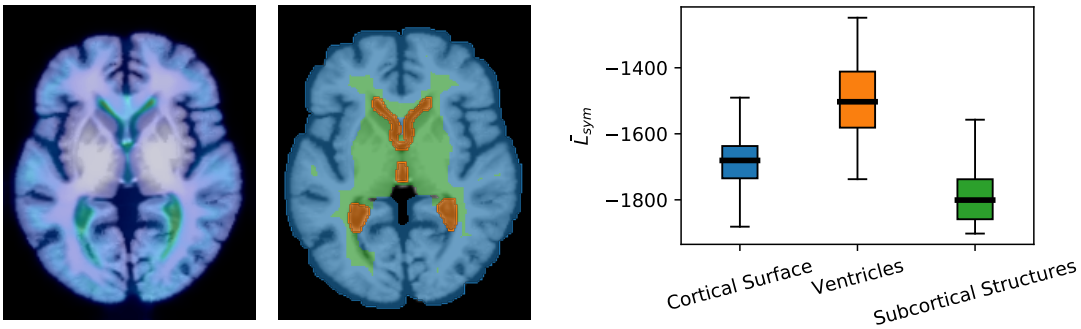


Figure 4: Left: Heatmap of average location of topological differences among the control group, predicted by the semantic model, averaged with $\mathbb{E}_{\mathbf{I},\mathbf{K}}\left[L_{\text{sym}}(\mathbf{I}|\mathbf{K}) \circ \Phi_{\mathbf{I}\to\text{Atlas}}\right]$ using a brain atlas as reference image. Center: We use morphological operations to split the atlas into cortical surface (blue), ventricles (orange) and sub-cortical structures (green). Right: Likelihood of topological differences occurring in each region. Boxplot with median, quartiles, deciles.

When analyzing the ROC curves, the supervised approach performed best for all false positive rates, while our approach with semantic loss performed consistently better than the other unsupervised approaches. Finally, even though both models share the same trained model, the score used by Li and Wyatt [26] performed better than scoring using the Jacobian determinant.

For the qualitative results, we present plots in Figure 3. When looking at the last two columns, we see that L_{sym} detected notable areas with high morphological differences compared to the reference image $\mathbf{I}$. This includes the ventricles (rows 2,3), the cortical areas with the sulci (all rows) as well as tumor areas (rows 1,3,5). There was a clear difference in the behaviour between semantic loss and MSE as the semantic loss highlights broader regions of the surface. When comparing to the $Q(\mathbf{J})$ measure in columns 2 and 3, we can see that our approach filtered most of the ventricles and sulci leaving an area around most tumor regions. Notable exceptions are rows 2 and 4, where the tumor area was not highlighted, as well as row 1 where only part of the tumor core was detected.

In Figure 4, we show the average score on healthy subjects. We see on the brain image and the box plot, that the cortical surfaces and ventricles get assigned higher scores than the subcortical structures.

5 Discussion and conclusion

In this work, we have introduced a novel approach for detection of topological differences. We evaluated our approach qualitatively and compared it quantitatively to previous approaches on an unsupervised segmentation task. Our unsupervised approach performed best among the unsupervised methods for this task, but could not reach the performance of the supervised method. This is expected since the statistics of tumor cells are very different from healthy tissue and the unsupervised models have not been trained on tumor tissue.

Our results are useful in practice as, unlike in tumor segmentation, general topological differences can not be annotated well on real data and there exists no labeled systematic dataset to learn them. While our results are not pixel exact, they indicate where a registration algorithm must be used more carefully to obtain a valid registration. The results obtained on tumor segmentation are reinforced by the distribution of scores obtained on healthy patients in different parts of the brain. The high likelihood of topological differences in ventricles found agrees with previous work [30] and the higher scores in cortical surfaces reflect the fact, that the sulci of the cortical surface exhibit high variability between subjects [6], which was previously difficult to quantify.

Our results also show that using a semantic loss function is advantageous compared to the MSE in this task, as all MSE based methods performed worse than our approach using the semantic loss. This is likely because the contrast between gray matter and white matter is quite small and thus the MSE is dominated by the contrast between brain and black background. This is a good reasoning for the image gradient-based correction introduced by Li and Wyatt [26]. In contrast, the semantic loss incorporates more texture information and thus is capable of detecting some tumor areas, which are not distant from the brain cortex. However, our approach misses tumor cores close to the cortex. We hypothesize, that this is in part caused by the similar appearance of tumors and grey matter, and in part due to the cortical area containing high topological variation among the control group as well.

Our unsupervised results for the method by An and Cho [1] are in line with previously reported results on a comparable dataset [42]. However, our supervised results are not comparable to the results published for the BRATS challenge, as we selected a subset of data for training and because we only use structural MRI images, discarding the other modalities.

Our study has several limitations. We only investigate registrations in 2D and topological differences might vanish if the whole 3D volume is considered. The transformations obtained by our unsupervised method differ from strongly regularised methods, as the hyperparameter-less learned prior under-regularises in order to maximize the likelihood of a topological match in healthy patients. Conversely, the poor performance of the Jacobian determinant might be due to a strong regularisation for good performance in healthy patients as we used the hyperparameters as found in [7].

In conclusion, our approach serves as the first step for unsupervised annotation of topological differences in image registration. Our approach is fully unsupervised and hyperparameter-free, making it a prospective building block in an end-to-end topology-aware image registration model.

References

[1] Jinwon An and Sungzoon Cho. *Variational Autoencoder based Anomaly Detection using Reconstruction Probability*. Tech. rep. 2015.

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. "A log-euclidean framework for statistics on diffeomorphisms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer Verlag, 2006, pp. 924–931.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific Data* 4 (2017).

[4] Spyridon Bakas et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge". In: *arXiv* 124 (2018).

[5] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800.

[6] Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

[7] Steffen Czolbe, Oswin Krause, and Aasa Feragen. "Semantic similarity metrics for learned image registration". In: *Proceedings of Machine Learning Research*. 2021.

[8] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. "Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration". In: *Medical Image Computing and Computer Assisted Intervention* (2018), pp. 729–738.

[9] V Delmon, S Rit, R Pinho, and D Sarrut. "Registration of sliding objects using direction dependent B-splines decomposition Registration of sliding objects using direction dependent B-splines decomposition *". In: *Phys. Med. Bio* 58.5 (2013), pp. 1303–1314.

[10] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular psychiatry* 19.6 (2014), pp. 659–667.

[11] Adriana Di Martino et al. "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II". In: *Scientific Data* 4.1 (2017), pp. 1–15.

[12] Mirza Faisal Beg, Michael I Miller, Alain Trouvétrouv, and Laurent Younes. "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms". In: *International Journal of Computer Vision* 61.2 (2005), pp. 139–157.

[13] B Fischl. *FreeSurfer. Neurolmage, 62 (2), 774–781*. 2012.

[14] Zoubin Gal, Yarin and Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *International Conference on Machine Learning*. 2016, pp. 1050–1059.

[15] Ulf Grenander and Michael I Miller. "Computational anatomy: An emerging discipline". In: *Quarterly of applied mathematics* 56 (1998), pp. 617–694.

[16] Lasse Hansen and Mattias P. Heinrich. "Tackling the Problem of Large Deformations in Deep Learning Based Medical Image Registration Using Displacement Embeddings". In: *Medical Imaging with Deep Learning* (2020).

[17] Søren Hauberg, Oren Freifeld, Anders Boesen, Lindbo Larsen, John W Fisher, Iii Lars, and Kai Hansen. *Dreaming More Data: Class-dependent Distributions over Diffeomorphisms for Learned Data Augmentation*. Tech. rep. 2016, pp. 342–350.

[18] Berthold KP Horn and Brian G Schunck. "Determining optical flow". In: *Artificial intelligence* 1 (1981), pp. 185–203.

[19] Rui Hua, Jose M. Pozo, Zeike A. Taylor, and Alejandro F. Frangi. "Multiresolution eXtended Free-Form Deformations (XFFD) for non-rigid registration with discontinuous transforms". In: *Medical Image Analysis* 36 (2017), pp. 113–122.

[20] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning*. International Machine Learning Society (IMLS), 2015, pp. 448–456.

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. "Spatial Transformer Networks". In: *Advances in neural information processing systems* (2015).

[22] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization". In: *nternational Conference on Learning Representations*. International Conference on Learning Representations, ICLR, 2015.

[23] Dongjin Kwon, Marc Niethammer, Hamed Akbari, Michel Bilello, Christos Davatzikos, and Kilian M. Pohl. "PORTR: Pre-operative and post-recurrence brain tumor registration". In: *IEEE Transactions on Medical Imaging* 33.3 (2014), pp. 651–667.

[24] Pamela J LaMontagne, Tammie L S Benzinger, John C Morris, et al. "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease". In: *medRxiv* (2019).

[25] Xiaoxing Li, Xiaojing Long, Christopher Wyatt, and Paul Laurienti. "Registration of Images with Varying Topology Using Embedded Maps". In: *IEEE Transactions on Medical Imaging* 31.3 (2012), pp. 749–765.

[26] Xiaoxing Li and Chritopher Wyatt. "Modeling topological changes in deformable registration". In: *2010 7th IEEE International Symposium on Biomedical Imaging*. 2010, pp. 360–363.

[27] Lihao Liu, Xiaowei Hu, Lei Zhu, and Pheng-Ann Heng. "Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 346–354.

[28] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2015), pp. 1993–2024.

[29] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. *SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows*. Tech. rep. 2020, pp. 12685–12696.

[30] Rune Kok Nielsen, Sune Darkner, and Aasa Feragen. "TopAwaRe: Topology-Aware Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 364–372.

[31] Danielle F. Pace, Stephen R. Aylward, and Marc Niethammer. "A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs". In: *IEEE Transactions on Medical Imaging* 32.11 (2013), pp. 2114–2126.

[32] Bartłomiej W. Papiez, Mattias P. Heinrich, Jérome Fehrenbach, Laurent Risser, and Julia A. Schnabel. "An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration". In: *Medical Image Analysis* 18.8 (2014), pp. 1299–1311.

[33] Sarah Parisot, William Wells, Stéphane Chemouny, Hugues Duffau, and Nikos Paragios. "Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs". In: *Medical Image Analysis* 18.4 (2014), pp. 647–659.

[34] Gabriel Peyré and Marco Cuturi. "Computational optimal transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 1–257.

[35] Danilo Jimenez Rezende and Shakir Mohamed. *Variational Inference with Normalizing Flows*. Tech. rep. 2015, pp. 1530–1538.

[36] Laurent Risser, François Xavier Vialard, Habib Y. Baluwala, and Julia A. Schnabel. "Piecewise-diffeomorphic image registration: Application to the motion estimation between 3D CT lung images with sliding conditions". In: *Medical Image Analysis* 17.2 (2013), pp. 182–193.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 9351. Springer Verlag, 2015, pp. 234–241.

[38] Dan Ruan, Selim Esedoĝlu, and Jeffrey A. Fessler. "Discriminative sliding preserving regularization in medical image registration". In: *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*. 2009, pp. 430–433.

[39] Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, and Heinz Handels. "Fast explicit diffusion for registration with direction-dependent regularization". In: *Biomedical Image Registration* 7359 (2012), pp. 220–228.

[40] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. *Learning Structured Output Representation using Deep Conditional Generative Models*. Tech. rep. 2015.

[41] Elizabeth R. Sowell, Paul M. Thompson, David Rex, David Kornsand, Kevin D. Tessner, Terry L. Jernigan, and Arthur W. Toga. “Mapping sulcal pattern asymmetry and local cortical surface gray matter distribution in vivo: Maturation in perisylvian cortices”. In: *Cerebral Cortex* 12.1 (2002), pp. 17–26.

[42] Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. “Self-Supervised Out-of-Distribution Detection in Brain CT Scans”. In: *Medical Imaging Meets NeurIPS Workshop* (2020).

[43] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. *Occlusion Aware Unsupervised Learning of Optical Flow*. Tech. rep. 2018, pp. 4884–4893.

[44] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. “Quicksilver: Fast predictive image registration - A deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We believe there are no negative or morally questionable impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We describe our model and the related math in detail, but do not include any theorems, proofs or lemma.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Section 4. Code is included in supplementary material. Pre-trained models are included if file size limits permits it. As we are not allowed to re-distribute the data, data access must be obtained individually.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The details of our model are given in Section 4.2. Hyper parameters of compared methods are given in Section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported bootstrapped AUC values but did not conduct multiple model evaluations.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section 4.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Section 4.1 and Acknowledgements (Acknowledgements redacted during double blind review)
   (b) Did you mention the license of the assets? [Yes] Section 4.1
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code is included in supplementary material. Pre-trained models are included if file size limits permits it. At point of publication, both will be released publicly on github.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Section 4.1. We checked that the dataset publications mentioned informed consent, however we do not know how consent was given and the way it was given might have warried through the different institutions who pooled their data for the referenced studies.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Section 4.1
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]