Semantic similarity metrics for image registration

Czolbe, Steffen; Pegios, Paraskevas; Krause, Oswin; Feragen, Aasa

Published in:
Medical Image Analysis

DOI:
10.1016/j.media.2023.102830

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Semantic similarity metrics for image registration

Steffen Czolbe, Paraskevas Pegios, Oswin Krause, Aasa Feragen

A B S T R A C T

Image registration aims to find geometric transformations that align images. Most algorithmic and deep learning-based methods solve the registration problem by minimizing a loss function, consisting of a similarity metric comparing the aligned images, and a regularization term ensuring smoothness of the transformation. Existing similarity metrics like Euclidean Distance or Normalized Cross-Correlation focus on aligning pixel intensity values or correlations, giving difficulties with low intensity contrast, noise, and ambiguous matching. We propose a semantic similarity metric for image registration, focusing on aligning image areas based on semantic correspondence instead. Our approach learns dataset-specific features that drive the optimization of a learning-based registration model. We train both an unsupervised approach extracting features with an autoencoder, and a semi-supervised approach using supplemental segmentation data. We validate the semantic similarity metric using both deep-learning-based and algorithmic image registration methods. Compared to existing methods across four different image modalities and applications, the method achieves consistently high registration accuracy and smooth transformation fields.

1. Introduction

Deformable registration, or nonlinear image alignment, is a fundamental tool in medical imaging to capture local deformations or changes between images. Applications include tracking disease progression (Yang et al., 2020; Castillo et al., 2013; Nielsen et al., 2019), population analysis (LaMontagne et al., 2019), co-registration of image modalities (Song et al., 2021; Lee et al., 2019b), object tracking (Ulman et al., 2017), and guiding of medical machinery (Trofimova et al., 2020). The registration model finds correspondences between a set of images and derives a geometric transformation to align them. Most algorithmic and deep learning-based methods solve the registration problem by the minimization of a loss function consisting of a similarity metric and a regularization term ensuring smoothness of the transformation. The similarity metric is essential to the optimization; it judges the quality of the match between registered images and has a strong influence on the result.

Pixel-based similarity metrics like euclidean distance and patchwise cross-correlation are well explored within algorithmic and deep-learning-based image registration. These metrics assume that if the image intensities are aligned, or strongly correlated, the images are well aligned. Each choice of metric adds additional assumptions on the characteristics of the specific dataset. Thus, a common methodological approach is to trial registration models with multiple different pixel-based metrics, and choose the metric performing best on the dataset (Balakrishnan et al., 2019; Hu et al., 2019b).

The shortcomings of pixel-based similarity metrics have been studied substantially in the image generation community (Hou et al., 2017; Zhang et al., 2018), where they have been superseded by deep similarity metrics approximating human visual perception. Here, image representations are commonly extracted by neural networks pre-trained on image-classification tasks (Deng et al., 2009). Performance can be further improved by fine-tuning the representation to human perception (Czolbe et al., 2020; Zhang et al., 2018). These representation-based deep similarity metrics have improved the visual quality of images generated with variational auto-encoders considerably. As image registration is a conditional generative problem (Dalca et al., 2018; Czolbe et al., 2021b), we propose to apply deep similarity metrics within image registration to achieve a similar increase in performance for registration models.

Contributions. We propose a data-driven similarity metric for image registration based on the alignment of learned, task-specific semantic features. The experimental results illustrate that the method is robust toward image noise and achieves consistently favorable trade-offs between registration accuracy and transformation smoothness. We evaluate the method using deep-learning based image registration with...
U-Nets (Ronneberger et al., 2015; Balakrishnan et al., 2019) and Transformers (Chen et al., 2022), and classical registration using the SyN algorithm of the ANTS package (Avants et al., 2008b).

To learn filters of semantic importance to the dataset, we present both an unsupervised approach using auto-encoders, and a semi-supervised approach using a segmentation model. We use the learned features to construct a similarity metric used for training a registration model, and validate our approach on four biomedical datasets of different image modalities and applications. For both methods and across all datasets, our method achieves consistently high registration accuracy and smooth transformation fields.

Finally, we perform an extensive ablation study to evaluate the influence of individual feature layers, model architectures, order of operations of the proposed loss, the possibility of using transfer learning in the absence of a dataset-specific semantic model, and the robustness toward noise in the images.

**Previous publications.** Part of this work has been published at the Medical Imaging with Deep Learning (MIDL) conference (Czolbe et al., 2021c). This journal release contains an extended experimental evaluation, a fourth dataset, deeper discussion, and a broadened background section. We demonstrate the applicability of our method using the recently published state-of-the-art transformer TransMorph, and the well-established SyN algorithm. In addition, the popular similarity metrics of mutual information (Studholme et al., 1999) and MIND-SSC (Heinrich et al., 2013b), further called MIND, have been included as baselines in all experiments. A new ablation study section discusses multi-task pre-training, demonstrates transfer learning in the absence of a dataset-specific semantic feature extractor, and provides insights into multi-level feature learning by evaluating how different levels contribute to the registration accuracy. A new experiment confirms the robustness of our metric towards image noise.

### 2. Background & related work

#### 2.1. Image registration

Intensity-based image registration frameworks model the problem as finding a transformation \( \Phi : \Omega \rightarrow \Omega \) that aligns a moving image \( I : \Omega \rightarrow \mathbb{R} \) to a fixed image \( J : \Omega \rightarrow \mathbb{R} \). The morphed source image, obtained by applying the transformation, is expressed by function composition as \( J \circ \Phi \). The domain \( \Omega \) denotes the set of all coordinates \( x \in \mathbb{R}^d \) within the image.\(^1\) Images record intensity at discrete pixel-coordinates \( p \) but can be viewed as a continuous function by interpolation. The optimal transformation is found by minimizing a similarity metric \( D \) and a \( \lambda \)-weighted regularizer \( R \), expressed via the loss function

\[
    L(I, J, \Phi) = D(I \circ \Phi, J) + \lambda R(\Phi) .
\]

The choice of similarity metric \( D \) is the main objective of this paper, and common choices are discussed later. The regularizer \( R \) is necessary as many non-linear transformation models are over-parametrized, leading to many potential solutions. Smooth transformation fields, that avoid folds or gaps, are assumed to be physically plausible and encouraged by the regularizer (Leow et al., 2007; Kabus et al., 2009). Implicit regularizers achieve these properties by measuring the inverse consistency of the transformation (Greer et al., 2021; Shen et al., 2019b), while explicit regularizers operate on the displacement vector

\[
    \Phi(p) = p + u(p) ,
\]

and the diffusion regularizer thereon is defined as

\[
    R(\Phi) = \sum_{p \in \Omega} \| u(p) \|^2 .
\]

with \( u(p) \) approximated via finite differences over the pixel coordinates.

#### 2.2. Registration methods

Many methods of optimizing Eq. (1) have been proposed, and finding improved registration methods continues to be an active area of research. The field can be grouped into 1. algorithmic methods and 2. deep-learning-based methods.

1. **Algorithmic methods** optimize the objective for each pair of images individually, resulting in slow registration when many images have to be registered, for example in real-time applications or large population studies. Yet, this approach does not require a large up-front investment into training datasets and resources. Most algorithms follow an iterative, gradient-descent-based approach. Some methods optimize the transformation directly, such as elastic models (Bajcsy and Kovacic, 1989; Davatzikos, 1997; Shen and Davatzikos, 2002), sparse parameterizations with b-splines (Rueckert et al., 1999), and Demons (Thirion, 1998; Vercauteren et al., 2007). Others parameterize intermediate transformation steps to offer diffeomorphic guarantees on the transformation field, such as the Large Diffeomorphic Distance Metric Mapping (LDDMM) algorithm (Faisal Beg et al., 2005) and standard symmetric normalization (SyN) (Avants et al., 2008b). Recent approaches follow a discrete optimization scheme (Heinrich et al., 2013a, 2015) while in Siebert et al. (2021), convex global optimization is combined with a local gradient-based instance refinement using an adaptive optimizer.

2. **With the emergence of deep neural networks**, model- and learning-based techniques for image registration are an area of active research. Compared to the algorithmic approach, deep-learning-based registration models are trained on a large dataset, necessitating a longer training time and a large collection of training images. However, after training is completed, inferring a transformation from the model is much faster than the algorithmic counterparts. Early works use supervised approaches, requiring ground-truth transformation fields (Yang et al., 2017; Krebs et al., 2017; Haskins et al., 2019). As these are often infeasible to attain, most contemporary works employ unsupervised or semi-supervised approaches by optimizing objective (1) directly. The dominant network architecture are fully-convolutional neural networks (CNNs), often in a U-Net configuration (Balakrishnan et al., 2019; Hu et al., 2019a; Hoopes et al., 2021). Various modifications, such as multi-level architectures (de Vos et al., 2019; Liu et al., 2019; Hu et al., 2019b; Mok and Chung, 2020, 2021; Shen et al., 2019a; Zhao et al., 2019), probabilistic models (Dalca et al., 2018; Czolbe et al., 2021b), discretized architectures with a correlation layer (Dosovitskiy et al., 2015; Heinrich and Hansen, 2020), and fluid-diffeomorphism based transformations (Dalca et al., 2018) have been proposed. Alternative approaches use vision transformers (Wang and Delingette, 2021; Chen et al., 2021, 2022; Mok and Chung, 2022; Shi et al., 2022; Song et al., 2022; Wang et al., 2022; Pegios and Czolbe, 2022) or graph-based networks (Hansen and Heinrich, 2021).
2.3. Similarity metrics for image registration

Similarity metric $D$ measures the distance between the warped moving (morphed) image $\text{I} \circ \Phi$ and the fixed image $\text{J}$. Pixel-based metrics are well explored within algorithmic image registration, a comparative evaluation is given by Avants et al. (2011). We briefly recall four popular choices used as baselines in our evaluation: mean squared error (MSE), normalized cross correlation (NCC), normalized mutual information (NMI), and modality independent neighborhood descriptor (MIND), and discuss how these can be combined with supervised labels to obtain a semi-supervised similarity metric.

2.3.1. Mean squared error

The pixel-wise MSE is intuitive, computationally efficient, and easy to reason about. It is derived by maximizing the negative log-likelihood of a Gaussian normal distribution, making it an appropriate choice under the assumption of Gaussian noise. On a grid of discrete points $p$ from the domain $\Omega$, the MSE is defined as

$$\text{MSE}(\text{I} \circ \Phi, \text{J}) = \frac{1}{|\Omega|} \sum_{p \in \Omega} ((\text{I} \circ \Phi) (p) - \text{J}(p))^2.$$  \hspace{1cm} (4)

2.3.2. Normalised cross correlation

Patch-wise NCC is robust to variations in brightness and contrast, making it a popular choice for images recorded with different acquisition tools and protocols, or even across image modalities. For two image patches $\text{A}, \text{B}$, represented as column-vectors of length $N$ with patch-wise means $\bar{\text{A}}, \bar{\text{B}}$ and variance $\sigma^2_\text{A}, \sigma^2_\text{B}$, it is defined as

$$\text{NCC}_{\text{patch}}(\text{A}, \text{B}) = \sum_{n=1}^{N} \frac{(\text{A}_n - \bar{\text{A}})(\text{B}_n - \bar{\text{B}})}{\sigma_\text{A}\sigma_\text{B}}.$$  \hspace{1cm} (5)

The patch-wise similarities are then averaged over the image as

$$\text{NCC}(\text{I} \circ \Phi, \text{J}) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \text{NCC}_{\text{patch}}(\text{I} \circ \Phi(p), \text{J}(p)).$$  \hspace{1cm} (6)

where $\text{I}_p, \text{J}_p$ denote the square image patch around pixel $p$ (Gee et al., 1993; Avants et al., 2008b). Patches are centered around each pixel, leading to overlapping patches. Note that a slightly altered but computationally more efficient variant of NCC is used in some image registration works (Avants et al., 2011).

2.3.3. Normalised mutual information

Normalised mutual information (Studholme et al., 1999) models the probabilistic relation between the voxel intensities of the images. It is suitable for applications where no linear relation between the image intensities is present, making it the primary similarity metric used in multi-modal image registration. The relation between the morphed and fixed image is modelled as

$$\text{NMI}(\text{I} \circ \Phi, \text{J}) = \frac{H(\text{I} \circ \Phi) + H(\text{J})}{H(\text{I} \circ \Phi, \text{J})}.$$  \hspace{1cm} (7)

with marginal entropy $H(\text{I}) = -\int_{0}^{1} p_{\text{I}}(x) \log(p_{\text{I}}(x)) \, dx$ and mutual information $I(x, y) = -\int_{0}^{1} \int_{0}^{1} p_{\text{I}, \text{J}}(x, y) \log(p_{\text{I}, \text{J}}(x, y)) \, dx \, dy$.

To calculate the joint probability $p_{\text{I}, \text{J}}(x, y)$ and marginals $p_{\text{I}}(x), p_{\text{J}}(y)$, the intensity distributions of both images are approximated by histograms, making them non-differentiable. For adaptation in gradient-based deep-learning frameworks, Parzen-window estimates with gaussian kernels are used to approximate the distributions. Given a Parzen window function $w_i$, the joint histogram for discrete bucket means $x, y \in \mathbb{R}$ is calculated as

$$h_{\text{I,J}}(x,y) = \sum_{p \in \Omega} w((\text{I} \circ \Phi)(p) - x) w(\text{J}(p) - y).$$  \hspace{1cm} (8)

from which the joint $p_{\text{I}, \text{J}}(x, y)$ is obtained by normalization and $p_{\text{I}}(x), p_{\text{J}}(y)$ by marginalization thereof de Vos et al. (2020) and Qiu et al. (2021).

2.3.4. Modality independent neighborhood descriptor

The MIND-SSC image descriptor (Heinrich et al., 2012, 2013b) extracts representations from images based on their self-similarity context (SSC). It is used as a loss function by comparing the extracted representations of images.

The self-similarity of two patches centered on $x, y$ with local variance $\sigma^2$ is calculated as

$$S(x, y) = \exp\left(-\frac{\|I_x - I_y\|^2}{\sigma^2}\right).$$  \hspace{1cm} (9)

Given this equation, the image descriptor of a pixel coordinate $p$ is then calculated by evaluating the above equation on all pixels $x, y \in \mathcal{N}$ from the neighborhood of $p$, only using pairs $x, y$ are adjacent to each other (euclidian distance of $\sqrt{2}$). Notably, the intensity of the center pixel $p$ has no direct influence on the descriptor of $p$. The descriptor is dependent on the choice of the patch size as well as the dilation and shape of the neighborhood $\mathcal{N}$, which have to be tuned for each application.

2.3.5. Semi-supervised measures

If additional information is available, the unsupervised similarity measures can be extended by a supervised component to align either ground-truth segmentation masks, pre-defined reference points, or reproduce a pre-determined reference transformation field. However, by adding a supervised component, the registration model is incentivized to be biased towards this component. Balakrishnan et al. (2019) study this in detail: as the strength of a supervised loss term is increased, the accuracy on unobserved regions and overall accuracy is decreasing. Thus, in the absence of perfect annotations, it is common practice to combine metrics operating on different representations of the image (Avants et al., 2008a). We compare to a semi-supervised metric by fusing an intensity-based loss $L_{\text{intensity}}$ with a semi-supervised metric $L_{\text{seg}}$ operating on segmentation class annotations as

$$L = L_{\text{intensity}}(\text{I} \circ \Phi, \text{J}) + \gamma L_{\text{seg}}(\text{S} \circ \Phi, \text{T})$$  \hspace{1cm} (10)

for segmentation masks $\text{S}, \text{T}$ of images $\text{I}, \text{J}$ and weighting factor $\gamma$.

For our experiments, we use a supervised version of the NCC metric, formulated as $\text{NCC}_{\text{seg}}(\text{I}, \text{S}, \text{J}, \text{T}, \Phi) = \text{NCC}(\text{I} \circ \Phi, \text{J}) + \text{MSE}(\text{S} \circ \Phi, \text{T})$ for one-hot encoded $\text{S}, \text{T}$ (Balakrishnan et al., 2019).

2.4. Deep similarity metrics in image registration

While deep-learning-based image registration has received much interest recently, similarity metrics utilizing the compositional and data-driven advantages of neural networks remain under-explored. Some works explore how to incorporate scale-space into learned registration models, but similarity metrics remain intensity-based (Hu et al., 2019b; Li and Fan, 2018). Learned similarity metrics are proposed by Haskins et al. (2019) and Krebs et al. (2017), but both approaches require ground truth registration maps for training, which are either synthetically generated or manually created by a medical expert. Lee et al. (2019a) propose to learn annotated structures of interest as part of the registration model to aid alignment, but the method discards sub-regional and non-annotated structures.

Closest to our work is the approach by Wu et al. (2016), who learn a representation of the input images via a stacked autoencoder and use the resulting representations for the downstream task of algorithmic image registration. This is similar to our autoencoder-based approach combined with Syn registration. While their experimental evaluation has limitations, such as the patch-based training on small $21^3$ patches, a model of 2 layers, and a small dataset of 66 images, their observations of increased accuracy and flexibility over handcrafted features are similar to ours. Majumdar et al. (2017) further investigates deep-learning-based features for algorithmic image registration. They find that on comparatively small datasets of less than 30 images, hand-crafted features can outperform learned ones.
localized, concrete features, and global, abstract ones, we calculate a similarity metric. We see that \( \mathcal{P} \) is pre-trained on a surrogate task. To capture alignment of both feature descriptors instead.

3. Method

3.1. A discussion of NCC

Our design of a semantic similarity metric starts by examining the popular NCC metric. We see that NCC between image patches \( A \) and \( B \) is equivalent to the cosine-similarity between the corresponding mean-centered vectors \( f(A) = A - \bar{A} \) and \( f(B) = B - \bar{B} \):

\[
\text{NCC}_{\text{patch}}(A, B) = \frac{\langle f(A), f(B) \rangle}{\| f(A) \| \| f(B) \|}.
\]

with scalar product \( \langle \cdot, \cdot \rangle \) and euclidean norm \( \| \cdot \| \). Thus, an alternative interpretation of the NCC similarity measure is the cosine-similarity between two feature descriptors in a high-dimensional space. The descriptor is given by the intensity values of a centered image patch centered at a pixel \( p \). We will construct a similar metric, using semantic feature descriptors instead.

3.2. A semantic similarity metric for image registration

To align areas of similar semantic value, we propose a similarity metric based on the agreement of semantic feature representations of two images. Semantic feature maps are obtained by a feature extractor, which is pre-trained on a surrogate task. To capture alignment of both localized, concrete features, and global, abstract ones, we calculate the similarity at multiple layers of abstraction. Given a set of feature-extracting functions \( F^i : \mathbb{R}^{L \times C} \rightarrow \mathbb{R}^{L \times C} \) for \( L \) layers, we define

\[
\text{DeepSim}(I \circ \Phi, J) = \frac{1}{L} \sum_{i=0}^{L-1} \sum_{|\Omega|} \frac{\langle F^i(I \circ \Phi)_p, F^i(J)_p \rangle}{\| F^i(I \circ \Phi)_p \| \| F^i(J)_p \|}.
\]

where \( F^i_p(J) \) denotes the \( i \)th layer feature extractor applied to image \( J \), at the spatial coordinate \( p \). It is a vector of \( C \) output channels, and the spatial size of the \( i \)th feature map is denoted by \( |\Omega| \).

Just as for NCC, the neighborhood of the pixel is considered by the similarity metric, as \( F^i \) is composed of convolutional filters with increasingly large receptive field sizes. In contrast to NCC, it is not necessary to zero-mean the feature descriptors, as the semantic feature representations are trained to be robust to variances in image brightness present in the training data.

3.3. Feature extraction

To aid registration, the functions \( F^i(\cdot) \) should extract features of semantic relevance for the registration task, while ignoring noise and artifacts inherent in image acquisition methods. To achieve these properties we extract features from the encoding branch of networks trained on two surrogate tasks:

\[
F^i(I \circ \Phi)
\]
1. Semi-Supervised measure: If segmentation masks are available, we can learn features on a supplementary segmentation task. Segmentation models excel at learning relevant kernels for the data while attaining invariance towards non-predictive features like noise, but require an annotated dataset for training. We denote the proposed similarity metric with feature extractors conditioned on this task as DeepSim$_{seg}$.

2. Unsupervised measure: We can learn an abstract feature representation of the dataset in an unsupervised setting with auto-encoders. Auto-encoders learn an efficient data encoding by training the network to ignore signal noise. A benefit of this approach is that no additional annotations are required. While variational methods for encoding tasks have several advantages, we choose a deterministic auto-encoder for its simplicity and lack of hyperparameters. We denote the similarity metric with feature extractors conditioned on this task as DeepSim$_{ae}$.

The choice of depth and receptive field size of the feature extracting functions has further impact on the metric. Deeper feature extractors can model more complex datasets, but increase computation time and memory requirements during training. Exclusively using high-level features, such as the last layer of a segmentation network, might only align the borders of anatomical regions and has the potential to ignore finer structures within those regions. On the contrary, too shallow features can behave similarly to intensity-based metrics. We evaluate different depth configurations as an ablation study, and use kernels up to the bottleneck of the segmentation network for our main experiments, effectively building a feature pyramid as visualized in Fig. 2.

4. Experimental setup

We evaluate our method using both deep-learning-based and algorithmic image registration. We train deep registration models with the proposed unsupervised DeepSim$_{ae}$ and semi-supervised DeepSim$_{seg}$, and compare to baselines MSE, NCC, NCC$_{sup}$ (NCC with supervised information), NMI, and MIND. Our implementation of the baseline metrics follows Avants et al. (2011), Balakrishnan et al. (2019), Qiu et al. (2021), Hou et al. (2017), and Heinrich et al. (2013b). To show that our method is also applicable to algorithmic image registration, we compare intensity-based registration using SyN (Avants et al., 2008b) to SyN registration of images augmented with semantic features learned by DeepSim. To ensure reproducibility, all code and experiments are available at github.com/SteffenCzolbe/DeepSimRegistration.

4.1. Data

To show that our approach applies to a variety of registration tasks, we validate it on four 2D and 3D datasets of different modalities:

1. T1 weighted Brain-MRI scans from the ABIDE-I, ABIDE-II (Di Martino et al., 2014) and OASIS2 (LaMontagne et al., 2019) studies for atlas-based alignment of Brain-MRI scans. Acquisition details, subject age ranges, and health conditions differ for each dataset, but no large anatomical anomalies are present. We perform standard pre-processing as in Balakrishnan et al. (2019), including intensity normalization, affine spatial alignment, skull-stripping and segmentation for each scan using FreeSurfer (Fischl, 2012) and crop the resulting images to 160 × 192 × 224 voxels. Anatomical regions labeled separately on each hemisphere and smaller regions such as on the sub-structures of the cingulate cortex are combined, resulting in 24 distinct segmentation classes. After scans with preprocessing errors are discarded, we split the data 3665/250/250 for train-, validation-, and test-set, and register images to an atlas.

2. T1 weighted MR scans of the hippocampus from the 2022 Learn2Reg challenge (Hering et al., 2022). The dataset was originally introduced in Jafari-Khouzani et al. (2011) and included in the Medical Segmentation Decathlon (Antonelli et al., 2022). It contains images from 90 healthy adults and 105 adults with a non affective psychotic disorder. Images are cropped to 64 × 64 × 64 voxels. We split the data into 156 train-, 52 validation-, and 52 test-images, and perform inter-subject registration, giving 24000 unique training pairs.

3. Slices of human blood cells from the Platelet-EM dataset (Quay et al., 2018). Images are recorded using serial block-face scanning electron microscopy. The dataset contains 74 slices manually annotated with three classes (Cytoplasm, Organelle, Background). Images are affinely pre-aligned and the dataset is split 50/12/12 for train-, validation-, and test-set. We register neighboring 2d slices.

4. Cell tracking video of the PhC-U373 dataset from the ISBR cell tracing challenge (Maška et al., 2014; Ulman et al., 2017). The video sequence contains 230 2d images and is annotated with two classes (Cells, Background). We split the data 115/68/67 for train-, validation-, and test-set and register images of adjacent time steps.

4.2. Deep learning models

For the registration model, we trial both well-established 2D and 3D U-Net (Ronneberger et al., 2015) architectures as popularized through VoxelMorph (Balakrishnan et al., 2019), and the recent state-of-the-art transformer model TransMorph (Chen et al., 2022).

We use the same U-Net architecture for the image registration model and segmentation-based feature extraction networks. We use a similar architecture for the auto-encoder feature extractor but without the shortcut connections. Each network consists of three encoder and decoder stages. Each stage consists of one batch normalization (Ioffe and Szegedy, 2015), two convolutional, and one dropout layer (Gal and Ghahramani, 2016). After the final decoder step, we smooth the model output with three more convolutional layers. We experimented with deeper architectures but found they do not increase performance. The activation function is LeakyReLU throughout the network, Softmax for the final layer of the segmentation network, Sigmoid for the final layer of the auto-encoder, and linear for the final layer of the registration network. The stages have 64, 128, 256 channels for 2d datasets, and 32, 64, 128 channels for 3d.

In our experiments with TransMorph, we tried different model variants and sizes. We use the original TransMorph version which consists of 4 stages with {2,2,4,2} number of Swin Transformer (Liu et al., 2021) blocks and {4,4,8,8} number of heads in each stage respectively but set the embedding dimension to C = 64 because it performed better. As suggested we set the window size to be the same as at the input size after 32-fold downsampling while we zero-pad the images for the PhC-U373 dataset to make the spatial dimensions divisible by 32.

The segmentation model is trained with a cross-entropy loss function, the auto-encoder with the mean squared error. Both U-Net and

---

### Table 1

Parameters of the baseline similarity metrics used in our experiments.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain-MRI</td>
<td>NCC / NCC$_{sup}$</td>
<td>window size</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>number of bins</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>radius</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>dilation</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>neighborhood size</td>
</tr>
<tr>
<td>Hippocampus MR</td>
<td>NCC</td>
<td>window size</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>number of bins</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>radius</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>dilation</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>neighborhood size</td>
</tr>
<tr>
<td>Platelet-EM</td>
<td>NCC</td>
<td>window size</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>number of bins</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>radius</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>dilation</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>neighborhood size</td>
</tr>
<tr>
<td>PhC-U373</td>
<td>NCC</td>
<td>window size</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>number of bins</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>radius</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>dilation</td>
</tr>
<tr>
<td></td>
<td>MIND</td>
<td>neighborhood size</td>
</tr>
</tbody>
</table>
transformer-based registration networks are trained with the loss given by Eq. (1). The optimization algorithm for all models is ADAM (Kingma and Ba, 2015), the initial learning rate is $10^{-4}$, decreasing by a factor of 10 each time the validation loss plateaus. All models are trained until convergence. Training images are augmented with random affine transformations. Due to the large 3D volumes involved, the choice of batch-size is often limited by available memory. We sum gradients over multiple passes to arrive at effective batch-sizes of 3–5 samples.

4.3. Hyperparameter selection

The characteristics of deformable image registration methods are strongly influenced by the strength of regularization that is applied. Additionally, some baseline metrics have further hyperparameters, e.g., the number of bins in NMI or radius and dilation in MIND. For a fair comparison, we tune all parameters on the validation split of each dataset. The parameter choices used in our experiments can be found in Table 1. For hyperparameter $\lambda$, we trial values $\lambda = 2^n$ for some $n \in \mathbb{Z}$ for each U-Net model and hyperparameter selection, and plot the validation mean dice overlap for further evaluation.

4.4. Algorithmic image registration

We further investigate whether algorithmic image registration benefits from the semantic image representations used for DeepSim. As a baseline, we choose to register the intensity images using the well-establish SyN algorithm (Avants et al., 2008b) from the ANTS software package (Avants et al., 2009), using the default registration parameters. For the semantic similarity metrics, we augment the intensity images by registering semantic feature maps obtained from either the auto-encoder or the segmentation feature extractor as additional modalities. We use channel-wise normalization, so that $\| F_l^c \|_2 = 1$ for each channel and layer, and up-scale all feature maps to image size using bi-/tri-linear interpolation. All modalities contribute equally to the objective function.
Fig. 4. Qualitative comparison of U-Net based deep-learning registration models. We register the moving image \( \mathbf{I} \) (1st column) to the fixed image \( \mathbf{J} \) (2nd column). Morphed images \( \mathbf{I} \circ \Phi \) obtained from registration models trained with baseline similarity metrics MSE, NCC, NCC\(_{sup}\), NMI, and MIND in columns 3–7. Morphed images obtained from our methods DeepSim\(_{ae}\) and DeepSim\(_{seg}\) in columns 8 and 9. Rows: Datasets Brain-MRI, Hippocampus MR, Platelet-EM, PhC-U373. Select segmentation classes annotated in color. The transformation is visualized by morphed grid-lines.

Fig. 5. Detail view of transformation grids on the highlighted spot of the Platelet-EM dataset in Fig. 4. The regularity of the transformation fields on this noisy image patch varies considerably between methods. Models trained with NCC, NCC\(_{sup}\), and MIND exhibit the most irregular transformation fields. Models trained with DeepSim\(_{ae}\) and DeepSim\(_{seg}\) show the smoothest transformation fields. The cell-boundary is annotated in blue. The transformation is visualized by morphed grid-lines.

Fig. 6. For the datasets Brain-MRI, Hippocampus MR, Platelet-EM and PhC-U373, we trial multiple registration models and algorithms and record their test mean dice overlap. U-Net based deep-learning models trained with similarity metrics MSE, NCC (Gee et al., 1993), NCC\(_{sup}\) (Balakrishnan et al., 2019), NMI (Studholme et al., 1999), MIND (Heinrich et al., 2012), DeepSim\(_{ae}\) (ours), DeepSim\(_{seg}\) (ours) on the left side of each plot. On the right side of each plot is algorithmic registration with the SyN algorithm (Avants et al., 2008b), and the SyN algorithm augmented with semantic features from DeepSim\(_{ae}\) and DeepSim\(_{seg}\). Boxplot with median, quartiles, deciles and outliers. Labels of our methods in bold.

5. Results

5.1. Qualitative results

We plot the fixed and moving images \( \mathbf{I}, \mathbf{J} \) and the morphed image \( \mathbf{I} \circ \Phi \) warped by transformations obtained from the U-Net registration models trained with each similarity metric model in Fig. 4. The transformation is visualized by grid-lines and segmentation classes are overlaid for guidance.

5.2. Registration accuracy

We measure registration accuracy by the mean Sørensen Dice overlap of the annotated segmentation masks on the unseen test-set of each dataset. Results are presented in Fig. 6. U-Net registration models trained with our proposed DeepSim\(_{ae}\) and DeepSim\(_{seg}\) metrics achieve higher accuracy than all baselines on the Brain-MRI and Platelet-EM datasets. On the PhC-U373 dataset, only the NCC\(_{sup}\) baseline performs better. On the Hippocampus MR dataset, DeepSim\(_{ae}\) and DeepSim\(_{seg}\) outperform MSE, but fall behind the other baselines.
of the transformation $\sigma^2(\log |J_\delta|)$ on the y-axis. Higher dice overlaps indicate a better alignment, lower variance indicates smoother transformation fields and fewer deformations.

Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Baseline</th>
<th>Effect size</th>
<th>Effect size</th>
<th>Effect size</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain-MRI</td>
<td>DeepSim</td>
<td>&lt;0.001</td>
<td>0.14</td>
<td>&lt;0.001</td>
<td>0.43</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>DeepSim</td>
<td>&lt;0.001</td>
<td>0.30</td>
<td>&lt;0.001</td>
<td>0.60</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hippocampus MR</td>
<td>DeepSim</td>
<td>&lt;0.003</td>
<td>0.15</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>DeepSim</td>
<td>&lt;0.001</td>
<td>0.24</td>
<td>0.006</td>
<td>0.07</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Platelet-EM</td>
<td>DeepSim</td>
<td>&lt;0.010</td>
<td>0.12</td>
<td>&lt;0.001</td>
<td>0.11</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>DeepSim</td>
<td>&lt;0.034</td>
<td>0.10</td>
<td>&lt;0.001</td>
<td>0.12</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PhC-U373</td>
<td>DeepSim</td>
<td>&lt;0.001</td>
<td>0.10</td>
<td>&lt;0.001</td>
<td>0.11</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>DeepSim</td>
<td>&lt;0.001</td>
<td>0.12</td>
<td>&lt;0.001</td>
<td>0.13</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 3

| Method     | Dataset     | $\sigma^2(\log |J_\delta|)$ | $|J_\delta| < 0(\%)$ | $\sigma^2(\log |J_\delta|)$ | $|J_\delta| < 0(\%)$ | $\sigma^2(\log |J_\delta|)$ | $|J_\delta| < 0(\%)$ | $\sigma^2(\log |J_\delta|)$ | $|J_\delta| < 0(\%)$ |
|------------|-------------|-----------------------------|------------------------|-----------------------------|------------------------|-----------------------------|------------------------|-----------------------------|------------------------|
| MSE        | Brain-MRI   | 0.21                        | 0.42                   | 1.84                        | 8.75                   | 0.24                        | 0.51                   | 0.29                        | 0.40                   |
|            | Hippocampus MR | 0.29                        | 0.93                   | 1.15                        | 4.08                   | 0.28                        | 1.36                   | 1.08                        | 4.15                   |
|            | Platelet-EM | 0.16                        | 0.28                   | 1.14                        | 3.99                   | 0.16                        | 0.24                   | 0.12                        | 0.40                   |
|            | PhC-U373    | 0.14                        | 0.20                   | 0.76                        | 1.62                   | 0.12                        | 0.20                   | 0.19                        | 0.14                   |

In Fig. 7, we contrast registration accuracy with transformation regularity (Leow et al., 2007; Kabus et al., 2009). We see that DeepSim and DeepSim$_{seg}$ are placed in the bottom right corner for three out of four datasets, indicating very smooth transformation fields combined with high registration accuracy.

Using algorithmic registration with SyN, the semantic features of DeepSim$_{seg}$ improve the registration accuracy over the baseline on all four datasets. The auto-encoder-based features of DeepSim$_{ae}$ fall short of just intensity-based registration.

We perform statistical significance testing of the model’s results with the Wilcoxon signed rank test for paired samples. A significance level of 5% gives a Bonferroni-adjusted significance threshold $p = 0.002$. We further measure the effect size with Cohen’s $d$ and show the results in Table 2. We see that most results are statistically significant. On the Platelet-EM dataset, the performance difference between models trained with MSE and our proposed metrics falls below the statistical threshold, yet our method outperforms the baselines with at least small effect sizes. On the PhC-U373 dataset, the baseline NCC$_{sup}$ outperform DeepSim with very small effect sizes.

5.3. Regularity of the transformation

To highlight the differences in transformation fields between methods, we display a noisy background patch of the Platelet-EM dataset in Fig. 5. The patch has been registered with transformations obtained from the U-Net registration models trained with each similarity metric in Fig. 5. The patch has been registered with transformations obtained from the U-Net registration models trained with each similarity metric in Fig. 6 on the x-axis, variance of the log Jacobian determinant of the transformation $\sigma^2(\log |J_\delta|)$ on the y-axis. Higher dice overlaps indicate a better alignment, lower variance indicates smoother transformation fields and fewer deformations.

We perform a quantitative analysis of the regularity of the transformations produced by the U-Net models in Table 3, measuring transformation irregularity by the variance of the log-determinant of the Jacobian of the transformation field $\sigma^2(\log |J_\delta|)$, and domain folding by the percentage of transformation voxels with a negative determinant.
Fig. 8. Model performance on noisy data. We add Gaussian noise sampled from $\mathcal{N}(0, \sigma^2)$ to the input data, and measure the dice overlap on the test set. The $x$-axis shows the noise level, and the $y$-axis the test Dice overlap. Images below the plot show image patches under the noise levels.

Table 4
Relative training time of U-Net registration models, based on measurements of 1 epoch of training. MSE = 1.00. Time measurement includes feed-forward and back-propagation through the model and loss function, as well as weight update for the model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Brain-MRI</th>
<th>Hip. MR</th>
<th>Plat.EM</th>
<th>PhC-U373</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>NCC</td>
<td></td>
<td>1.09</td>
<td>1.22</td>
<td>1.05</td>
<td>1.02</td>
</tr>
<tr>
<td>NCC$_{sup}$</td>
<td></td>
<td>1.12</td>
<td>1.24</td>
<td>1.05</td>
<td>1.03</td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td>1.07</td>
<td>1.05</td>
<td>1.08</td>
<td>1.04</td>
</tr>
<tr>
<td>MIND</td>
<td></td>
<td>1.02</td>
<td>1.05</td>
<td>1.03</td>
<td>1.02</td>
</tr>
<tr>
<td>DeepSim$_{ae}$</td>
<td></td>
<td>1.04</td>
<td>1.52</td>
<td>1.17</td>
<td>1.15</td>
</tr>
<tr>
<td>DeepSim$_{seg}$</td>
<td></td>
<td>1.03</td>
<td>1.49</td>
<td>1.16</td>
<td>1.15</td>
</tr>
</tbody>
</table>

5.4. Noise resistance

We further evaluate registration performance in the presence of noise in the input data. Without retraining the models, we measure the mean dice overlap on the test set of the Platelet-EM dataset with added Gaussian noise. We sample the noise from $\mathcal{N}(0, \sigma^2)$, and test noise levels of $\sigma = 0, 0.05, 0.1, \ldots, 0.35$. We show results and examples of the noisy image patches in Fig. 8. The performance of all models decreases as noise is added. However, the models trained with the baselines loose performance quicker then model trained with Deepsim.

5.5. Convergence and speed

We monitor the mean training and validation dice overlap of the U-Net based deep-learning models during training in Fig. 9. The training accuracy is, with few exceptions, similar to the test accuracy, indicating that results generalize well. The relative time per epoch of models trained with each loss function is given in Table 4. Training models with DeepSim adds between 4 – 52% time per epoch.

5.6. Anatomical regions

The Brain-MRI dataset contains annotations of the brain’s anatomical regions. We plot the dice overlap per region in a boxplot in Fig. 10, and highlight regions where both of our metrics perform better than all baselines bold. Baseline methods (blue) perform very similar, despite NCC$_{sup}$ as a supervised metric requiring more information over the unsupervised MSE and NCC.

5.7. Image registration using transformers

We further evaluate the flexibility of the proposed method using the recent state-of-the-art transformer-based model TransMorph on the 2d datasets. As in the previous experiments, we perform hyperparameter tuning both for DeepSim and baseline loss functions and select the transformer model with the highest validation dice overlap.

Given the best parameter choices, we evaluate the tradeoff between dice-overlap and transformation smoothness on the test sets in Fig. 11. Results are similar to ones obtained with the U-Net model in Fig. 7, albeit slightly better overall. TransMorph registration networks trained with DeepSim achieve favorable accuracy-smoothness tradeoffs for both the 2d datasets, placing in the bottom right corner of the plot on both datasets.

6. Ablation studies

After establishing the DeepSim similarity metric and comparing it to established choices, we now focus on investigating decisions made in the design of the metric. We investigate the effect of different levels of extracted features, assess if a dedicated feature extractor needs to be trained for each dataset, and inquire about the order of operations within the metric. These experiments are performed on the 2d dataset only.

6.1. Levels of extracted features

The abstraction levels at which semantic features are extracted can have an impact on the proposed metric. We investigate how different levels of features contribute to the registration accuracy. We trial DeepSim loss functions using deep features extracted from multiple combinations of layers, using only features extracted feature extraction layers 1, 2, 3, 1+2, 1+3, 2+3, and all layers. The level of the feature extraction layer is denoted with superscript, e.g., DeepSim$^1$ compares shallow features extracted only from the first layer of a deep feature extractor, while DeepSim$^{12}$ combines features only from the two first layers.
We re-tune the regularization hyperparameter $\lambda$ on the 2d datasets for the different layer configurations of DeepSim both for our unsupervised and semi-supervised approach, using U-Net based registration models. We plot the registration accuracy on the validation sets for DeepSim$_{ae}$ and DeepSim$_{seg}$ in Figs. 13(a) and 13(b) respectively. We observe that for the Platelet-EM dataset, which contains noisy images, using high-level features such as level 3 of the deep feature extractors improves accuracy. Most notably, disregarding the shallowest layer, DeepSim$_{seg}^{23}$ achieves slightly better performance than DeepSim$_{seg}^{3}$. On the other hand, on the non-noisy PhC-U373 dataset, level 1 contributes the most to the registration accuracy. This is in line with previous results, where the intensity-based baselines performed competitively on the PhC-U373 dataset. In general, it is evident that including both low-level, concrete features and high-level, more abstract ones in the loss, is beneficial to the performance of the registration model in almost all of the cases.

We further plot heatmaps of the loss occurred under different loss functions in Fig. 12(a), and a per-layer view of the loss occurred under DeepSim in Fig. 12(b). The moving and fixed images (top left) have been registered using U-Net models trained with the presented similarity metrics, and the occurred loss at each spatial coordinate is plotted on a color scale normalized to each model. The results show that, compared to the baselines, the DeepSim loss is more evenly distributed around non-matching image parts. It also does not occur a loss in the noisy background area between the cells.
6.2. Transfer learning

One drawback of DeepSim is that a feature extractor has to be trained for each dataset. We investigate whether this is necessary, or if we can use feature-extractors trained on related or similar data instead. This approach is commonly referred to as Transfer Learning.

We trial three separate configurations: for the Platelet-EM dataset, we use the auto-encoder and segmentation model trained on the PhC-U373 dataset as the feature extractor. Vice versa, for the PhC-U373 dataset, we use the auto-encoder and segmentation model trained on the Platelet-EM dataset. We denote these similarity metrics with transferred features as DeepSim\textsubscript{ae}–TL and DeepSim\textsubscript{seg}–TL. Finally, we investigate the performance of our method using a universal feature extractor. To this end, we extract features from a VGG (Simonyan and Zisserman, 2014) classification network trained on ImageNet (Deng et al., 2009), and we denote this variant of the method as DeepSim\textsubscript{VGG}.

For each configuration, we train U-Net based registration networks with different regularization parameters $\lambda$ and plot the mean validation dice overlap in Fig. 14. We observe that for PhC-U373 all transfer learning approaches (DeepSim\textsubscript{ae} – TL, DeepSim\textsubscript{seg} – TL) not only improve the performance compared to the default setup of our method, but also surpass in performance all the baselines from the previous experiments. This might indicate that the PhC-U373 dataset does not have sufficient complexity and size to train a good feature extractor on it. For the Platelet-EM dataset, the performance of DeepSim transfer learning variants falls short of the original method but is still comparable with other baseline loss functions.

6.3. Feature extraction and transformation

As defined in Eq. (12), the DeepSim metric first applies the transformation to the moving image, and then extracts a semantic representation from the morphed image (Transform before Extraction, TbE). A recently used alternative approach (Czolbe et al., 2021a) first extracts a semantic representation from the moving image, and then transforms the semantic representation (Extraction before Transformation, EbT). We empirically compare both variants, using both an auto-encoder and a segmentation model as the feature extractor.

We re-tune the regularization hyperparameter $\lambda$ for the alternative implementation DeepSim (EbT). The necessary transformation of lower resolution feature maps is implemented by down-sampling and scaling the transformation before warping the feature map. Registration accuracy on the validation sets is displayed in Fig. 15. We observe that the optimal choice for $\lambda$ differs between the variations of the loss function, with the optimal value for the EbT version being consistently lower than for the unaltered TbE version across all datasets and feature extractors. The loss functions achieving the highest dice overlap are inconsistent, with the TbE version performing better on...
Fig. 13. Effect of different layer configurations on registration accuracy with (a) DeepSim$_{ae}$ and (b) DeepSim$_{seg}$. We trial loss functions using only features from feature extraction layers 1, 2, 3, 1+2, 1+3, 2+3 and all layers. For each configuration, we train U-Nets with different regularization parameters $\lambda$ (x-axis) and observe the validation dice overlap (y-axis).

Fig. 14. We trial transferring feature extractors between datasets. Models trained with DeepSim$_{ae}$ and DeepSim$_{seg}$ use a feature extractor trained on their dataset. Models trained with DeepSim$_{ae}$ – TL and DeepSim – TL use a feature extractor trained on the opposite dataset. The model trained with DeepSim$_{VGG}$ uses features form the VGG image classification network in the feature extractor. For each loss function, we train models with a range of regularization parameters $\lambda$ (x-axis) and observe the mean dice overlap on the validation set (y-axis).

the Brain-MRI dataset, both versions achieving similar scores on the Platelet-EM dataset, and the EbT version performing better on the PhC-U373 dataset.

7. Discussion

The experimental results show that registration methods optimized with the proposed semantic similarity metric achieve small improvements in accuracy. Additionally, they are robust to noise and produce smoother transformations, resulting in consistent improvements in the accuracy-smoothness tradeoff and more plausible transformations. The trend holds true across four diverse datasets and registration with SyN, U-Nets, and Transformers, showing the general applicability of the results.

We see the largest performance increase on the Platelet-EM dataset, which we hypothesize is caused by the significant noise present in the dataset. The intensity-based metrics incentivize the model to align the noise, producing the observed unsmooth transformations and high loss values across the image, overall hindering registration. The proposed semantic similarity metric has instead learned that the noise is of no semantic importance, thus ignoring it in the registration.
Similarities metrics are independent of the registration method used. To show the general applicability of our metric and its independence from the underlying registration framework, we conducted experiments with U-Nets, transformer-based architectures, and algorithmic registration using the SyN algorithm from the ANTS package. The observed results are similar, especially between the two deep-learning based approaches. This is an indication that the choice of the registration model matters less compared to the metric used during training. Our method is robust and behaves consistently across registration methods.

A drawback of our method is that it requires a feature extractor to obtain features of semantic importance to the dataset. The experimental evaluation shows that the availability of annotated anatomical regions can help with learning semantic features, particularly if the dataset is large enough to support the training of such models. However, labeling a dataset is expensive and time-consuming, especially in biomedical settings.

To alleviate this issue, we investigated two alternatives that do not require labeled data or the training of a dedicated feature extractor. The need for labeled data can be removed by using semantic features extracted from an auto-encoder. This metric outperformed the baselines in both registration accuracy and transformation smoothness when registering images with the deep-learning-based models. However, using algorithmic registration, the unsupervised approach underperformed the baselines, particularly on the 3D Brain-MRI dataset. This could be caused by shortcomings of the auto-encoder, which yielded blurry reconstructions on the large brain volumes. On the other hand, for the 3D Hippocampus MR dataset, our proposed similarity metric provides noticeable improvement when using the SyN registration algorithm.

To remove the requirement of having to train a dedicated feature extractor completely, one can use transfer learning to use an extractor trained on a different dataset. We used extractors trained on other medical datasets, and a general computer vision feature extractor, trained on ImageNet. Both worked especially well when the target dataset was small, even outperforming feature extractors trained directly on the data for the PhC-U373 dataset. This approach could be further expanded by using networks pre-trained on a large range of medical imaging tasks (Zhen et al., 2019).

We focused on mono-modal image registration. The presented method could be extended to multi-modal registration in two different ways: (1) Through the use of modality-specific feature extractors to map each input modality to a common semantic representation (Pielawski et al., 2020), followed by alignment thereof. (2) Alternatively, separate feature extractors can be trained on each modality, and their semantic representations compared with a multi-modal metric such as MIND, NMI, or NCC.

7.1. Weaknesses

A weakness of our method is the need to include a separate model for the extraction of semantic features. While we have shown that no dedicated model is required – other models can be reused with only slight decreases in performance – the design, training, and testing of a second model takes additional resources.

In the absence of ground-truth transformation fields, evaluation of deformable image registration is performed through proxy tasks. We measured accuracy by segmentation dice-overlap, but this evaluation technique only measures the overlap of larger areas, while discarding the alignment of sub-structures inside the annotated regions and does not evaluate point-to-point matches. We further evaluated the smoothness of the transformation fields and balanced this with the dice overlap in our evaluation, but no conclusive way of combining these metrics exists in the image registration literature. We welcome that recent registration challenges focus increasingly on measures besides segmentation dice overlap.

As our similarity metric depends on an auxiliary task, there is also a risk that the metric is biased by this choice of task, as well as by the segmentation masks that are used to train the auxiliary segmentation network. This label bias is perhaps most of all a problem in that its potential downstream effects are hard to foresee. However, we also note that the annotations often used to validate registration algorithms come with similar risks. Registration algorithms are often validated using annotated landmarks or Dice overlap of segmentation masks. We argue that these validation methods, which also affect which models are eventually chosen and published as state-of-the-art, come with a similar risk of label bias.

Any method is based on a large number of choices, decisions, and hyperparameters. While we did extensive trials of some of them in the ablation studies, there is always more that can be tested. We weighted all semantic features evenly in our method, and only considered features extracted from the encoding branch of the feature extraction networks. Tuning the individual weight of each feature is computationally expensive, but can be achieved in the presence of dedicated datasets, as Zhang et al. (2018) show for perceptual similarity metrics in image generation, or through hyperparameter learning strategies as shown by Hoopes et al. (2021) and Mok and Chung (2021) for image registration. While we did tune the regularization hyperparameter for the deep-learning-based models, we did not tune the parameters of SyN, instead using the same default parameters for each method. Due to technical constraints, we did not use the exact formulation of DeepSim for the SyN registration experiment but instead treated the semantic representations from DeepSim as additional modalities during registration with SyN. Because of practical issues, NMI is not used for TransMorph on the PhC-U373 dataset. Due to limited hardware availability, we do not include the 3d datasets in some of the ablation studies and our experiments with TransMorph.

8. Conclusion

We designed a semantic similarity metric for image registration. The new metric measures image similarity via the agreement of semantic
and hierarchical image representations. The semantic representations can be extracted either in an unsupervised approach using an autoencoder or in a supervised approach using supplemental segmentation data. In the absence of both, we have shown that features trained on related datasets can also be used.

The proposed metric achieves robust performance across four diverse datasets and three different registration models, using both deep-learning-based and algorithmic image registration. Image registration optimized with our method shows improved accuracy and smoother transformation fields compared to metrics such as MSE, NCC, NMI, and MIND.

The method is applicable to image registration tasks of all modalities and anatomies. Beyond the diverse range of datasets presented here, our good results in the presence of noise let us hope that our method will improve registration accuracy in domains such as low-dose CT, ultrasound, or microscopy, where details are often hard to identify, and image quality is poor.

We further emphasize that the application of semantic similarity metrics is not limited to the image registration domain. Semantic similarity metrics have the potential to improve methods in other image regression tasks, such as image synthesis, -translation, and -reconstruction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability

Code is public on github. Datasets are public, but we do not redistribute these.

Acknowledgments

We thank Matthew Quay, the Cell Tracking Challenge, and the Cancer Imaging Archive for the provision of the datasets, and Huaqi Qi for providing the implementation of the NMI metric. This work was funded in part by the Lundbeck Foundation, Denmark (grant no. R218-2016-883), in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (grant no. 0062606), and in part through the Danish National Research Foundation through the Danish Pioneer Centre for AI (grant no. P1).

References


