

# CoMIR: Contrastive Multimodal Image Representation for Registration

Nicolas Pielawski\*, Elisabeth Wetzer\*, Johan Öfverstedt,  
Jiahao Lu, Carolina Wählby, Joakim Lindblad, Nataša Sladoje  
Dept. of Information Technology, Uppsala University, Sweden

## Abstract

We propose contrastive coding to learn shared, dense image representations, referred to as CoMIRs (Contrastive Multimodal Image Representations). CoMIRs enable the registration of multimodal images where existing registration methods often fail due to a lack of sufficiently similar image structures. CoMIRs reduce the multimodal registration problem to a monomodal one, in which general intensity-based, as well as feature-based, registration algorithms can be applied. The method involves training one neural network per modality on aligned images, using a contrastive loss based on noise-contrastive estimation (InfoNCE). Unlike other contrastive coding methods, used for, e.g., classification, our approach generates image-like representations that contain the information shared between modalities. We introduce a novel, hyperparameter-free modification to InfoNCE, to enforce rotational equivariance of the learnt representations, a property essential to the registration task. We assess the extent of achieved rotational equivariance and the stability of the representations with respect to weight initialization, training set, and hyperparameter settings, on a remote sensing dataset of RGB and near-infrared images. We evaluate the learnt representations through registration of a biomedical dataset of bright-field and second-harmonic generation microscopy images; two modalities with very little apparent correlation. The proposed approach based on CoMIRs significantly outperforms registration of representations created by GAN-based image-to-image translation, as well as a state-of-the-art, application-specific method which takes additional knowledge about the data into account. Code is available at: <https://github.com/MIDA-group/CoMIR>.

\* Authors contributed equally.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

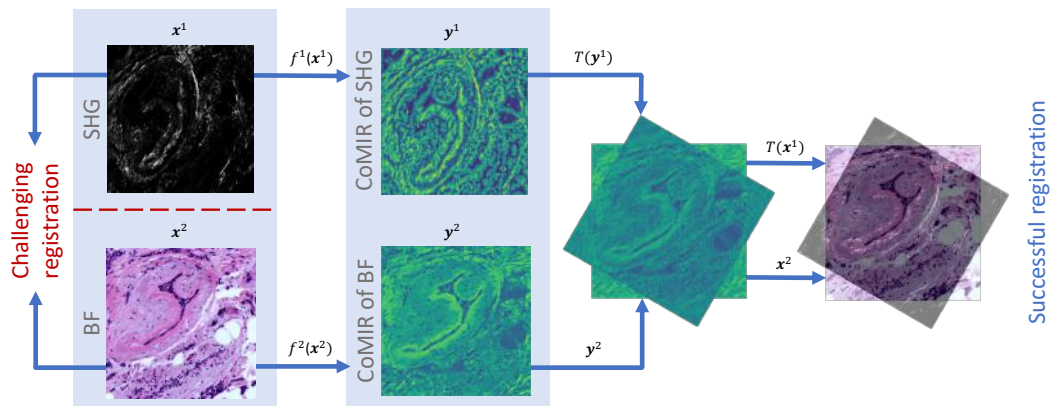


Figure 1: Registration of images of different modalities (here bright-field (BF) and second-harmonic generation imaging (SHG)) may be very challenging. CoMIR successfully estimates the shared representation of these images, and enables their successful registration by monomodal approaches.

## 1 Introduction

Multimodal images refer to images captured with multiple types of sensors, where each sensor outputs information not fully provided by the other sensors. Multimodal image fusion is the process of combining information from multiple imaging modalities. It allows downstream tasks to exploit complementary information as well as relationships between modalities. In order to perform image fusion, the images need to be aligned, either by joint acquisition, or by manual or automated registration. To enable registration, common structures between the modalities need to be found, reflected by the shared or mutual information (MI); this is in general a very difficult task. Our method, *Contrastive Multimodal Image Representation for registration (CoMIR)* reduces the challenging problem of multimodal registration to a simpler, monomodal one, as shown in Fig. 1. While image-to-image translation aims to predict cross-modality, i.e. to predict one modality given the other modality as input, our method directly learns representations called CoMIRs, relying on the MI by maximizing the Mutual Information Noise-Contrastive Estimation (InfoNCE) [20, 25]. InfoNCE can, under certain assumptions, act as a lower bound to MI, and has been successfully applied in connection with contrastive coding in other tasks, such as classification and segmentation. Although contrastive losses (CL) are often used for representation learning [1, 6, 22, 24, 25, 43, 46, 47, 49, 53, 62], to the best of our knowledge, the method presented here is the first to produce dense representations for very different imaging modalities which can be utilized for monomodal registration by existing feature- or intensity-based registration methods. InfoNCE has been previously used to learn embeddings used in classification and segmentation tasks in which the resulting subspace is required to feature properties such as separability between classes. Furthermore, registration requires representations which are both translation and rotation equivariant, as opposed to classification, for which invariance is required. Our proposed method produces image-like, contrastive representations that possess the necessary equivariant properties to find a transformation between the original inputs.

The contributions of this paper are the following: We show that (1) contrastive learning of aligned pairs of images can produce representations that reduce complex (or even non-feasible), rigid multimodal registration tasks to much simpler monomodal registration; (2) the proposed CoMIRs are rotation equivariant due to a modification of a commonly used CL. This modification is model-independent and does not require any architecture modifications, nor any additional hyperparameter tuning. Although the method is supervised, thanks to our sophisticated scheme for generating training patches, as little as one image pair can be sufficient to generate CoMIRs (depending on the nature of the imaging data).

## 2 Background and Related Work

Image registration consists of finding transformations matching unaligned images. This process is of highest importance, notably in computer vision and medical imaging, and has been extensively studied in the last fifty years [19]. In many cases, a modality-independent solution can be obtained by optimizing MI or generating hand-crafted local structural feature descriptors [23]. Unfortunately, optimizing MI has many caveats: it has a slow convergence rate and a narrow catchment basin with many local maxima. Furthermore, the method relies on specific statistical clues that not all datasets contain. Learning-based algorithms can potentially provide a solution, as the algorithm can learn semantic relationships between modalities. In general it is difficult to find suitable image similarities between multimodal images due to their difference in appearance and/or signal expression. In [58] internal similarities in images across modalities are captured by defining structural representations of image patches through Laplacian eigenmaps and entropy images in order to perform multimodal registration. [5] uses sparse coding to learn a mapping between two image modalities such that sum of squared differences or MI can be used as measures for image registration. Deep learning has changed the field of similarity and metric learning from learning distances to learning feature embeddings that fit distance functions [18]. CL [21] and in particular triplet loss [48, 61] have shown great success in deep metric learning for a wide range of applications, such as object retrieval [45], single-shot object classification [55], object tracking [34, 52], classification [8] and multimodal patch matching of aerial images [63]. In [26], keypoint descriptors are learnt using contrastive metric learning to minimize the difference between two feature representations from two corresponding points and maximize the difference of two feature representation from two distant points to register multimodal medical images. The link between CL, classification and MI is an active area of study, with [49] showing that the triplet loss can be extended to make use of  $N$  negative samples instead of one and is an approximation of the softmax function.

**Contrastive Loss and MI** Several recent representation learning approaches [1, 24, 25, 46, 51, 53, 62] are based on the infomax principle which refers to maximizing the MI between input and output, [2, 39]. One approach showing impressive results is Deep InfoMax (DIM) [25] which argues that maximizing MI with high dimensional inputs and outputs is a difficult task. Similarly, Contrastive Predictive Coding (CPC)[46] maximizes MI between global and local representation pairs by sequential aggregation to learn latent features which can be used for classification. Using Noise-Contrastive Estimation (InfoNCE) as a lower bound to MI with flexible critic was found to result in the most useful representations for classification in [25, 46]. It was further used in [1], where InfoNCE is maximized between features from multiple views in a self-supervised manner. The term multiview can refer to augmented versions of one unlabelled image or multiple modalities of one instance. Instead of maximizing MI between features from one single image as in DIM, MI is maximized across multiple feature scales simultaneously as well as across independently augmented copies of each image. However, in [54], Tschannen et al. show that maximizing tighter bounds on MI than InfoNCE can result in worse representations and argue that the success of these methods cannot only be attributed to the maximization of MI. The authors put the success of approximate MI maximization in connection with the usage of the triplet loss or CL by showing that representation learning across different views by maximizing InfoNCE can be equivalent to metric learning using the multi-class K-pair loss presented in [49]. Tschannen et al. point out that the negative samples for the CL have to be drawn independently in order for InfoNCE to yield a lower bound for the MI, an assumption often disregarded. Despite this, [1, 6, 25, 47, 53] report a better performance of the respective downstream task when using many negative samples or hard examples.

**Equivariance** As [25] emphasizes, the usefulness and quality of a learnt representation is not only a matter of information content but also representational characteristics. Features desirable to request from a representation are equivariances. In tasks such as segmentation or registration, translational and rotational equivariance are highly beneficial. Equivariance defines the property of a function to commute with the action of a symmetry group when its domain and codomain are acted on by that symmetry group. A function or operator  $f : \Omega \mapsto Y$  is called equivariant under a family of transformations  $\mathcal{T}$  if for any transformation  $T \in \mathcal{T}$ , there exists  $T' \in \mathcal{T}$  s.t.

$$f(T(X)) = T'(f(X)) \quad \forall X \in \Omega. \quad (1)$$

Feature equivariance can be achieved by three different approaches. Firstly, by data augmentation, where randomly transformed pairs of input and label masks are passed to a model which learns some degree of equivariance. The second approach is model-based, as proposed in [12, 13, 60], where equivariant mappings are achieved by adjusting the convolutional, activation and pooling layers to be applied over groups and sharing weights. Thereby these so-called group equivariant convolutional networks (G-CNNs) ensure that the layers themselves become equivariant. A further in-depth study of the theory of equivariant CNNs is presented in [11] and [59]. Thirdly, [9] encourages rotational invariance through an additional constraint in the loss function which has been adapted by [37] to achieve equivariance. They propose an additional term to the cross-entropy loss used to train the model for segmentation:  $L_{rot} = \frac{1}{2N} \sum_{x_i \in X} \|O(I) - \bar{r}(I)\|_2^2$ , where  $O(I)$  is the feature map of the image at  $0^\circ$  and  $\bar{r}(I)$  the mean feature map of the input rotated by multiples of  $90^\circ$ . This introduces a hyperparameter, impairs training due to the different scale of gradients between the cross-entropy loss and  $L_{rot}$  and cannot guarantee any rotational equivariance apart from the  $\mathcal{C}_4$  symmetry group, i.e. rotations by multiples of  $90^\circ$ . Enforcing equivariance through the loss was used in [33], which modifies a CL proposed in [56] to learn action-equivariance for Markov Decision Processes.

### 3 Proposed Method

We introduce a modality independent approach to map two given images of different modalities to similar representations called CoMIRs. Contrastive learning is applied to aligned pairs of images during training to create dense representations. These learnt representations are sufficiently similar to allow the application of monomodal registration algorithms. As our method does not require any additional knowledge regarding the modalities at hand, there are no limitations with respect to application area. In the following section we introduce the CL used in our method, the sampling scheme needed to form the CL, and our modifications to achieve rotational equivariance. We also discuss the choice of critic used for the downstream task of registration.

**Contrastive Loss** Here, we introduce the CL for two modalities, the general case for  $M$  modalities is given in App. 7.1. Let  $\mathcal{D} = \{(x_i^1, x_i^2)\}_{i=1}^n$  be an i.i.d. dataset containing  $n$  data points, where  $x_i^j$  is an image in modality  $j$ , and  $f_{\theta_j}$  the network processing modality  $j$  with respective parameters  $\theta_j$

for  $j \in \{1, 2\}$ . For an arbitrary datapoint  $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{D}$ , the loss is given by

$$\mathcal{L}^{opt}(\mathcal{D}) = -\mathbb{E}_{(\mathbf{x}^1, \mathbf{x}^2) \sim \mathcal{D}} \left[ \log \frac{\frac{p(\mathbf{x}^1, \mathbf{x}^2)}{p(\mathbf{x}^1)p(\mathbf{x}^2)}}{\frac{p(\mathbf{x}^1, \mathbf{x}^2)}{p(\mathbf{x}^1)p(\mathbf{x}^2)} + \sum_{\mathbf{x}_i \in \mathcal{D} \setminus \{\mathbf{x}\}} \frac{p(\mathbf{x}_i^1, \mathbf{x}_i^2)}{p(\mathbf{x}_i^1)p(\mathbf{x}_i^2)}} \right] \quad (2)$$

Eq. (2) is equivalent to a categorical loss discriminating between negative and positive samples. The ratio distribution  $\frac{p(\mathbf{x}^1, \mathbf{x}^2)}{p(\mathbf{x}^1)p(\mathbf{x}^2)}$  is approximated by the exponential of a critic function  $h(\mathbf{y}^1, \mathbf{y}^2)$  that computes (some arbitrary) similarity between CoMIRs  $\mathbf{y}^1 = f_{\theta_1}(\mathbf{x}^1)$  and  $\mathbf{y}^2 = f_{\theta_2}(\mathbf{x}^2)$  for the scaling parameter  $\tau > 0$

$$\mathcal{L}_{\theta}(\mathcal{D}) = -\frac{1}{n} \sum_{i=1}^n \left( \log \frac{e^{h(\mathbf{y}_i^1, \mathbf{y}_i^2)/\tau}}{e^{h(\mathbf{y}_i^1, \mathbf{y}_i^2)/\tau} + \sum_{\mathbf{y}_j^1, \mathbf{y}_j^2 \in \mathcal{D}_{neg}} e^{h(\mathbf{y}_j^1, \mathbf{y}_j^2)/\tau}} \right) \quad (3)$$

$\mathcal{L}(\mathcal{D})$  is named InfoNCE as described in [46]. Its minimization approximately maximizes a lower bound on the MI, given by  $I(\mathbf{y}^1, \mathbf{y}^2) - \log(n)$ , which gets tighter as  $n \rightarrow \infty$  [46, App. A1], assuming that  $\mathcal{D}_{neg}$ , the set of chosen negative samples, is sampled i.i.d. The MI is defined as  $I(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{X}, \mathbf{y} \sim \mathcal{Y}} [\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}]$ .

**The Critic** The loss function in Eq. (2) contains the ratio  $\frac{p(\mathbf{x}^1, \mathbf{x}^2)}{p(\mathbf{x}^1)p(\mathbf{x}^2)}$  which is an unknown quantity [62], but can be approximated with the exponential of a statistical, often bilinear, model [1, 6, 46, 49, 53, 62]. Our experiments using a bilinear model resulted in CoMIRs less suitable for registration (see section 4.2 and App. 7.4, Fig. 11) than choosing a positive, symmetric critic  $h(\mathbf{y}^1, \mathbf{y}^2)$  with a global maximum for  $\mathbf{y}^1 = \mathbf{y}^2$ . We experiment with both a Gaussian model with a constant variance  $h(\mathbf{y}^1, \mathbf{y}^2) = -\|\mathbf{y}^1 - \mathbf{y}^2\|_2^2$  which uses the mean squared error (MSE) as a similarity function, and a trigonometric model  $h(\mathbf{y}^1, \mathbf{y}^2) = \frac{\langle \mathbf{y}^1, \mathbf{y}^2 \rangle}{\|\mathbf{y}^1\| \|\mathbf{y}^2\|}$  which relates to the cosine similarity.

**Rotational Equivariance** To equip our model with rotational equivariance, we propose to add to our objective function an additional constraint which does not require additional parameter tuning and can be incorporated into the CL. In particular, instead of only maximizing  $h(\mathbf{y}_i^1, \mathbf{y}_i^2)$  within the CL for an aligned pair  $\mathbf{x}_i^1, \mathbf{x}_i^2$ , we also maximize  $h(\mathbf{y}_i^1, T_1'(f_{\theta_1}(T_1(\mathbf{x}_i^1))))$ , and  $h(\mathbf{y}_i^2, T_2'(f_{\theta_2}(T_2(\mathbf{x}_i^2))))$  for  $f_{\theta_i}$  being the model trained on modality  $i$ ,  $h$  the critic between the resulting representations, and  $T_i, T_i' \in \mathcal{T}$ . Here we choose  $\mathcal{T} = \mathcal{C}_4$ , the finite, cyclic, symmetry group of multiples of  $90^\circ$  rotations. While the actions of this symmetry group do not require any interpolation of the input images or CoMIRs, they result in sufficient rotational equivariance for angles beyond multiple of  $90^\circ$ , as shown in section 4.1. In general, any symmetry group could be chosen. Rather than extending our original loss term  $h(\mathbf{y}_i^1, \mathbf{y}_i^2)$  by three separate explicit loss terms, we combine the constraints in a single loss term that implicitly enforces the  $\mathcal{C}_4$  equivariance

$$h(T_1'(f_{\theta_1}(T_1(\mathbf{x}_i^1))), T_2'(f_{\theta_2}(T_2(\mathbf{x}_i^2)))). \quad (4)$$

We randomly sample  $T_1$  and  $T_2$  once per training step for each element in the batch, which iteratively optimizes all combinations of  $\mathcal{C}_4$ -transformations.

**Sampling of negative samples** In [32], the authors argued that the ability to discriminate between signal and noise increases with more negative samples. We sample negative samples  $\mathcal{D}_{neg}$  from random patches during training within the entirety of all training image pairs at random positions with random orientations. The patches can be extracted from the original images without introducing any padding border effects caused by rotations. Every extracted patch is subject to data specific, random augmentation. This sampling scheme guarantees large variation within every batch, which increases the (statistical) efficiency and generalization of the model. The elements of both modalities from the batch are reused as negatives, such that for a given pair of matching samples, there are  $2n - 2$  negative samples available, for a batch size of  $n$  pairs, as done in [6].

## 4 Experiments

We evaluate our proposed method on two multimodal image datasets.

**Zurich Dataset** The open Zurich dataset [57] consists of 20 aerial images of the city of Zurich of about  $930 \times 940$ px. The images are composed of four channels, RGB and Near-Infrared (NIR) and are captured with the same sensor in identical resolution. An example is given in App. 7.2, Fig. 5.

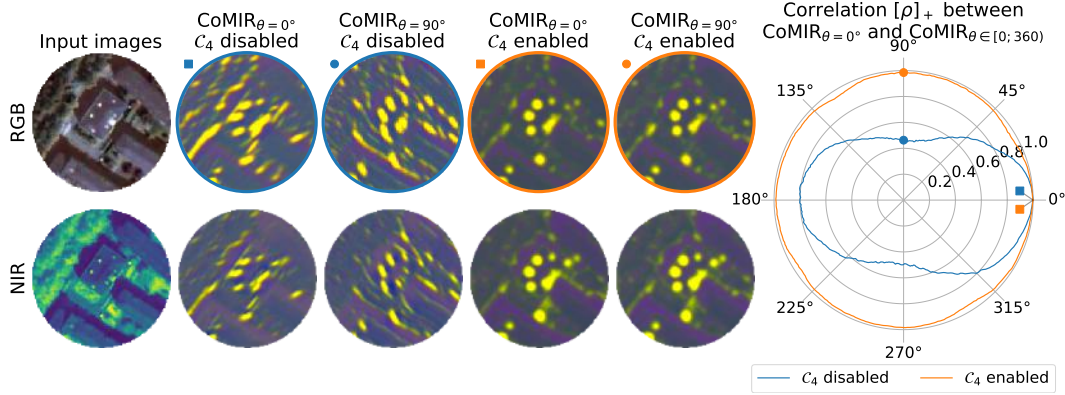


Figure 2: Top row: RGB input patch  $x^1$  cropped from the Zurich test set and the resulting CoMIRs using the cosine similarity; Bottom row: the matching NIR image  $x^2$  and its resulting CoMIRs. The stabilized CoMIRs  $f_{\theta_i}(x^i)$  and  $f_{\theta_i}(T(x^i))$  are shown for  $x^i, T$  a rotation of  $90^\circ$ . Without the  $C_4$ -equivariance constraint in the loss function, the resulting CoMIRs (marked blue) are not identical and display edge artifacts. The polar plot (to the right) shows the positive correlation between the stabilized  $\text{CoMIR}_{\theta=0^\circ} = f_{\theta_i}(x^i)$  and  $\text{CoMIR}_{\theta} = f_{\theta_i}(T(x^i))$  for  $T$  a rotation by  $\theta \in [0, 360)^\circ$ , demonstrating that rotational equivariance beyond multiples of  $90^\circ$  is achieved by the  $C_4$ -constrained CL. Animated version: [https://youtu.be/iN5G1PWFZ\\_Q](https://youtu.be/iN5G1PWFZ_Q).

**Biomedical Dataset** The dataset consists of 206 aligned Bright-Field (BF) and Second-Harmonic Generation (SHG) tissue microarray core image pairs which are  $2048 \times 2048$ px in spatial dimensions and were manually aligned using landmarks [30]. The training set consists of 40 image pairs of size  $834 \times 834$ px which are center-cropped from the corresponding original images. The validation set for the CNN training consists of another 25 such pairs, a tuning set for registration parameters of additional 7 pairs, and the test set for evaluation of another 134 pairs. Fig. 6 in App. 7.2 shows an example image of an original tissue microarray core and the center-cropped patch used in this study. The cropped dataset used in this study is provided in [16].

#### 4.1 Evaluation of Representations

**Rotational Equivariance** While different downstream tasks might require different properties of the learnt representations, for registration, we require the representations to be rotationally equivariant. The impact of the rotational equivariance constraint on the learnt CoMIRs of the Zurich dataset is displayed in Fig. 2. When the rotation equivariance is not explicitly enforced in the model or in the loss, the CoMIRs feature edge artifacts that rotate along with the input. We test the level of equivariance by measuring the positive correlation between the stabilized CoMIR of an input image and the stabilized CoMIR of the rotated input image. The result for  $\theta \in [0, 360)^\circ$  is shown in Fig. 2 and demonstrates that the  $C_4$ -constrained CL achieves rotational equivariance for all angles, not being restricted to only multiples of  $90^\circ$ .

**Reproducibility of CoMIR** Training a model to produce CoMIRs with  $c$  channels leads to multiple solutions reaching the same loss value. One such case is a permutation of the channels, and is likely to occur between two training sessions;  $c!$  different permutations can be achieved. It is possible, however, to consistently obtain very similar CoMIRs between experiments. [10] observe that the convergence of convolutional filters is in direct connection with their initialization. Similarly, we find that initializing the models' parameters randomly with a fixed seed results in similarly trained models and CoMIRs. We compare CoMIRs generated by 50 identical models, all initialized with the same seed and trained on the Zurich dataset. We then compute the mean pairwise correlation between all the CoMIRs to measure the similarity and consistency between runs. The source of stochasticity between runs originates from the minibatch sampling and from the data augmentation scheme. To test the influence of the data we train 20 models on patches from the same single image, and additionally train 19 models on patches augmented from a single training image varying for each experiment. Table 1 shows the results of the experiments. We observe a high consistency between training runs when the training patches are sampled from the same image and the model's parameters are initialized with the same seed. Fig. 10 in App. 7.4 shows the generated CoMIRs used to compute the average correlation.

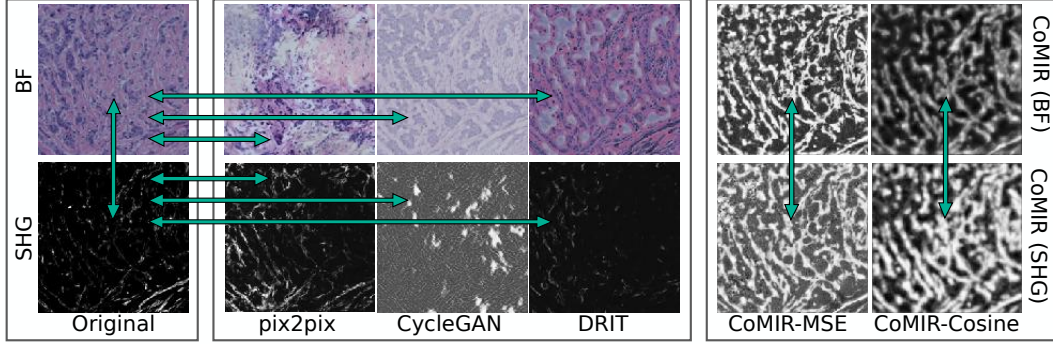


Figure 3: Multiple methods to transform BF and SHG into one common modality. To the left, the original BF (top) and SHG (bottom) images from the test set, followed by image translations of BF and SHG into the respectively other domain. To the right CoMIRs based on MSE and cosine similarity. The arrows indicate the different image pairs to be registered in the experimental setup.

Table 1: Mean pairwise correlation between the CoMIRs during 4 different experiments on the Zurich dataset. The consistency (i.e. the reproducibility of a given training session) between the CoMIRs is highest when the weights are initialized with a fixed seed (most important) and the training distribution is similar (less important). 95% C.I. estimated with empirical bootstrap.

Mean pairwise Corr.	1 unique train image	1 varying train image per model
Fixed seed init.	93.0% [ 90.6%; 95.5%], N=100	78.4% [ 70.4%; 83.9%], N=38
Random seed init.	-2.7% [-22.3%; -1.4%], N=40	1.9% [-18.0%; 5.7%], N=38

**Temperature Analysis** The temperature  $\tau$  is a hyperparameter of the loss function. High  $\tau$  yields a smoother loss. We tested a wide range of values of  $\tau$  and observed that all models converged and produced reasonable CoMIRs. See App. 7.4, Fig. 11, for examples of representations with various  $\tau$  and loss combinations. We found  $\tau = 0.1$  to be a good setting for the Zurich and  $\tau = 0.5$  for the biomedical dataset.

## 4.2 Multimodal Image Registration

**Evaluation Set and Metrics for Registration** For each image pair in the test set of 134 biomedical images, first a random rotation up to  $\pm 30$  degrees, followed by random translations in x and y directions of up to  $\pm 100$ px was applied. To avoid any border effects, the transformed patches were cropped from the original  $2048 \times 2048$  images. The magnitude of transformation was measured by the average Euclidean displacement  $\frac{1}{4} \sum_{i=1}^4 \|C_i^{Ref} - C_i^{Trans}\|_2$  of the corner points  $C_i$  between the reference image  $C_i^{Ref}$ , and the transformed image  $C_i^{Trans}$ . While the transformations themselves were entirely random, they were sampled in a stratified manner, such that 44 pairs were considered to have undergone a "large" transformation, 45 a "medium" and another 45 a "small" transformation. An average displacement was considered small if not exceeding 100px, medium if of a size (100, 200]px, and large if exceeding 200px. The same evaluation metric was used for all registration results:  $err = \frac{1}{4} \sum_{i=1}^4 \|C_i^{Ref} - C_i^{Reg}\|_2 > 100$ px is considered a failure.  $C_i^{Ref}, C_i^{Reg}, i \in \{1, 2, 3, 4\}$ , are the corners of the reference image and the one resulting from the registration, respectively. As the ground truth was obtained by manual registration, we included an independent manual registration task on a subset of our experimental setup by 6 human annotators which showed that a displacement error of up to  $\sim 50$  pixels can be expected; see App. 7.4, Fig. 7.

**Implementation details** We make no assumptions about the properties of the model. We choose two identical dense U-Nets [29], one per modality, which share no parameters. We experiment with both the MSE and cosine similarity as a critic. We set the temperature to  $\tau = 0.5$ . We use 46 negative samples. An important parameter to choose is the field of view of the patches that the model is trained on. This is data-specific and a patch should be large enough to cover significant structures in the images. For both datasets patches of the size  $128 \times 128$ px were chosen. For the Zurich dataset we choose 3-channel CoMIRs, for the biomedical dataset we choose 1-channel CoMIRs as this is suitable for the downstream registration methods considered. The SHG images were preprocessed by



applying a log-transform  $\log(1 + x)$  with  $x \in [0, 1]$ . Implementation details of our method as well as the subsequent registration algorithms can be found in App. 7.3.

**Baseline** To set a baseline performance, we perform multimodal registration using Mattes MI algorithm [42] on the original SHG and BF images with a (1+1) evolutionary strategy [50].

**Intensity-based Registration using  $\alpha$ -AMD** [44] is a general registration method based on distances combining intensity and spatial information [38]. It has been shown to be both accurate and having a large convergence region. One of its limitations, compared to e.g. MI, is that it requires the intensities to be in  $[0, 1]$  and approximately equal (not merely correlated) for corresponding structures.

**Feature-based Registration using SIFT** SIFT is a feature detection algorithm introduced by [40]. It extracts features from both a reference and a floating image which are invariant to scale and rotation, and robust across a large range of affine distortions, additive noises, and changes in illumination. The extracted feature points are matched with Random Sample Consensus (RANSAC [17]).

**Manual Registrations** To obtain a baseline for comparing the machine performance to a human level, a panel of six annotators was selected to perform the registration on a small set of test patches ( $n=10$ , randomly selected, identical for all annotators). The setup for this registration task is the same as for the automatic registration methods:  $834 \times 834$ px patches from the center of the tissue microarray cores are to be aligned. Note that this differs significantly from the setup in which the ground truth (GT) for the biomedical dataset was originally manually acquired where the full cores were available to the annotator as shown in App. 7.2, Fig. 6. The GT was obtained by manually aligning the selected landmarks [30], while the manual registrations in our evaluation is performed by moving the SHG images over the bright-field images to obtain an appropriate fit.

**Generative Methods** Generative models, such as generative adversarial networks (GANs), are often used in image-to-image translation ([28, 64]) in which they have the potential to enable the use of monomodal registration methods by translating one modality into the other. We implement three well-known image translation methods: **pix2pix** [28], **CycleGAN** [64], and **DRIT** [35, 36]. As image translation aims to predict a representation of a BF image given the SHG input and vice versa, the resulting images can be considered to be in a common space in which monomodal registration can be attempted.

**Data-specific State-of-the-Art** The first intensity-based registration method capable of automatically aligning SHG and BF images was proposed as CurveAlign in [30]. CurveAlign relies on data-specific, biomedical knowledge, i.e. that BF images are stained by hematoxylin and eosin (H&E), where eosin stains extracellular matrix components such as collagen, in particular with shades of pink. SHG mainly corresponds to collagen fibers. Using this prior on the data, CurveAlign performs segmentation on the BF image to isolate collagen structures, which are then registered to the SHG using a registration scheme based on MI with a (1 + 1) evolutionary algorithm.

**Choice of Critic** The effect of the choice of critic can be visually inspected on an image from the Zurich test set, in App. 7.4, Fig. 11. The correlation between the CoMIRs of RGB and NIR, for each of the observed critics – bilinear model, cosine similarity, and MSE – was averaged across a set of temperature settings  $\tau$  ( $N=15$ ). The results show that a bilinear critic produces weakly correlated maps w.r.t.  $\tau$ :  $\bar{\rho}_{\text{bilinear}} = 60.9\% [56.7\%; 65.0\%]$ . Using MSE resulted in  $\bar{\rho}_{\text{MSE}} = 85.4\% [83.2\%; 88.1\%]$ , whereas cosine similarity gave  $\bar{\rho}_{\text{cosine}} = 91.5\% [88.4\%; 94.7\%]$ . The 95% Confidence Intervals (C.I.) were computed with empirical bootstrap. Hence, we perform the registration evaluation on the biomedical dataset for MSE and cosine similarity. Training with the MSE gave consistently better CoMIRs w.r.t. registration, see App. 7.4, Fig. 9.

**Results** Fig. 3 shows an image pair of the biomedical test set, together with the representations produced by the image translations utilizing pix2pix, CycleGAN and DRIT, as well as the corresponding CoMIRs. The arrows indicate the pairs of modalities for which the registration task was attempted. Using image pairs resulting from image translation led to poor performance using both intensity-based as well as feature-based registration. The similarities between the GAN generated images and their original counterparts appeared too large for registration by  $\alpha$ -AMD. The corresponding SIFT features were detected for only three image pairs among all combinations of image translations and modalities, but even for these the registration error exceeded our success-threshold of 100px. We report the results for registration by MI for these methods in App. 7.3 Fig. 8.

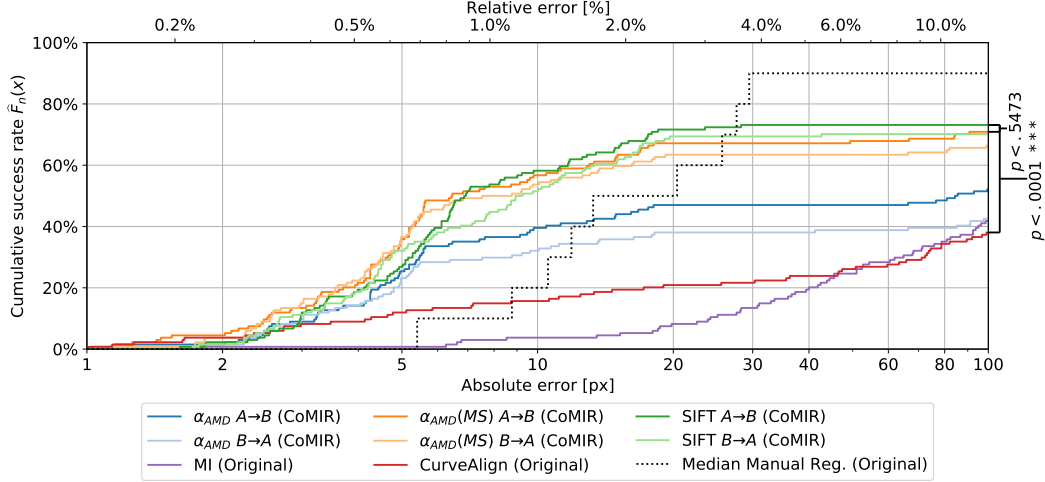


Figure 4: The cumulative number of successful registrations for the increasing error shown as an eCDF of the different registration methods over the biomedical test set. The results are compared to the median of the error of six independent manual registrations on a subset of ten images. The relative error is the proportion to the patch width and height. A Wilcoxon signed-rank test is performed to statistically highlight the differences between CurveAlign, SIFT, and  $\alpha$ -AMD.

Fig. 4 shows the empirical cumulative distribution functions (eCDF) of the successful registrations on the CoMIRs based on MSE, as a function of the error for (i)  $\alpha$ -AMD, (ii)  $\alpha$ -AMD with multiple starts (MS), and (iii) SIFT, using both CoMIR learnt from BF ( $A \rightarrow B$ ) and SHG ( $B \rightarrow A$ ) as the reference image. As a baseline comparison we also show the eCDF of (iv) the MI registration on the original BF and SHG images, (v) CurveAlign which registers BF to SHG, as well as (vi) the median eCDF for the six manual registrations on a subset of ten images. A registration error not smaller than 100 is considered a failure and is not shown. The feature- and intensity-based registration performances on CoMIRs are consistently better than the multimodal registration approaches on the original multimodal images. Wilcoxon signed-rank tests show that SIFT significantly ( $p = 5e-11$ ) outperforms CurveAlign, but show no significant difference between  $\alpha$ -AMD MS and SIFT ( $p < .5473$ ). The median error on the manual registration highlights the difficulty of the task. We further observe that the success of registration by MI is highly dependent on the size of the transformation between the image pair. We perform registration by MI (i) on the original images; (ii) for the corresponding CoMIRs; (iii) the BF and GAN generated BF as well as (iv) SHG and GAN generated SHG for pix2pix, CycleGAN and DRIT and observe linear dependence between the mean registration error and the mean displacement of the corner points for all attempts. The results are shown in App. 7.4, Fig. 8. The choice of critic influences the resulting CoMIRs. MSE encourages the intensities of the representations to be more similar than cosine similarity, which is favorable for the registration task. We observe that for both registration by SIFT, as well as  $\alpha$ -AMD, the results are considerably better for the MSE-based CoMIRs than for the cosine similarity, see App. 7.4 Fig. 9. App. 7.4 Table 2 gives the number of successfully registered image pairs which resulted in less than 1% relative error to the image width and length ( $< 9\text{px}$ ), as well as 5% relative error ( $< 42\text{px}$ ). Both the training (GPU, 1345s) and inference (CPU, 5s/image) of the model are fast, and the process of registration adds between 2s/image with SIFT and 25s/image with  $\alpha$ -AMD MS (see App. 7.4.1). An animated illustration depicting the process of registration of three example images with  $\alpha$ -AMD is available at: <https://youtu.be/zpcgnqcQgqM>.

## 5 Conclusion

The contrastive loss based on InfoNCE was successfully adjusted to the task of image registration and modified to result in rotationally equivariant features. The proposed CoMIRs successfully extract shared content in multimodal images to enable multimodal image registration by reducing it to a monomodal one. Using monomodal intensity- and feature-based registration methods significantly outperforms multimodal registration by MI, as well as a state-of-the-art, data-specific approach. For registration tasks, the generated CoMIRs contain more valuable information than GAN generated images. We show that the training of CoMIRs is stable w.r.t. hyperparameters and reproducible in



connection with weight initialization and training data. We show that the size of the training dataset can be as little as one image pair and give insight to the choice of critic. Future research could explore adding dense aleatoric uncertainty to CoMIRs, learning CoMIRs of more complex and higher dimensional data and more diverse modalities (e.g. volumetric images, audio, videos, and time-series) and extend equivariant properties to other groups, as well as investigate the CoMIRs’ applicability to segmentation and pixel regression tasks.

## 6 Broader Impact

Using CoMIRs for multimodal image registration has a direct application in multimodal image fusion. A wide range of areas benefit from fusing content of images of different sensors. One such area is material science. Early stage anomaly detection is used to characterize newly developed materials w.r.t. physical properties. As a concrete example, carbides along the grain boundary of a material can indicate impairments in material strength, but require different Scanning Electron Microscopy (SEM) sensors which acquire images asynchronously at different spatial resolutions [7]. This results in a multimodal registration problem which could be addressed by the proposed CoMIRs. The implications research in material science based on successful registration and fusion of images of this kind can have on the society are widespread. The Materials Genome Initiative (MGI) which has been launched by the US Federal Government in 2011 for example, aims to address clean energy, national security, and human welfare [41]. While this area of research can have a beneficial impact on society, by developing biocompatible materials for medical advances or materials needed in a variety of settings to reduce the carbon footprint, scientific findings in this area are directly linked to military developments also (see e.g. the US Air Force’s involvement in MGI). Another area of application which makes use of multimodal imaging data is the field of remote sensing. Again, while aerial observation can be used to monitor geological changes like melting glaciers due to climate change or early detection of wildfires, it is also tightly connected to military action (espionage, navigation systems such as Unmanned Aerial Vehicles (UAV), target localization, lethal autonomous war weapons). The area which can profit the most from multimodal image registration and following fusion in a positive way, is biomedicine. It is a broad and active field of research to combine information from modalities such as computed tomography (CT), magnetic resonance imaging (MRI) and Positron emission tomography (PET) or BF, SHG and two-photon-excited fluorescence (TPEF) microscopy. These imaging techniques often provide complementary and clinically relevant information needed for a diagnostic task. For example CT gives good spatial resolution and dense tissue contrast, while MRI yields better soft tissue contrast. BF and SHG are for example studied together in connection with collagen organization in tumor growth. By providing a joint representation between multiple modalities, issues regarding patient data anonymization need to be taken into account, e.g. if only images in one modality were subject to anonymization (e.g. photographs), while the other was not (e.g. CT scan), CoMIRs could potentially facilitate data de-anonymization from one modality to another. Registration by MI is computationally expensive, and as we show in our paper, MI is highly susceptible to an initial starting position close to the global extremum. In order to perform well, a larger number of restarts is needed to overcome being trapped in a local extremum, increasing the computational load even more. Generating CoMIRs for an entire dataset combined with registration by SIFT is much cheaper computationally, especially regarding the small training data needed for CoMIR generation. Hence, in settings where multimodal registration is already in place, CoMIRs can reduce the energy cost required and in consequence the environmental impact, however it may encourage the analyses of multimodal data on a much greater scale than ever before. The CoMIRs’ usage is also not limited to the task of multimodal registration, but could be useful for classification, segmentation or patch retrieval. This is subject of future research, but we foresee a potential in segmentation by training on the label masks as a second modality for example.

## Acknowledgments and Disclosure of Funding

We would like to thank Prof. Kevin Eliceiri (Laboratory for Optical and Computational Instrumentation (LOCI), University of Wisconsin-Madison) and his team for their support and kindly providing the dataset of BF and SHG imaging of breast tissue microarray cores.

The project was financially supported by the Swedish Foundation for Strategic Research (grants SB16-0046, BD150008), the European Research Council (grant 682810), the Wallenberg Autonomous Systems and Software Program, WASP, AI-Math initiative, VINNOVA (MedTech4Health project 2017-02447) and the Swedish Research Council (project 2017-04385).

## References

- [1] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15535–15545. Curran Associates, Inc., 2019.
- [2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
- [3] S. L. Best, Y. Liu, A. Keikhosravi, C. R. Drifka, K. M. Woo, M. A. Guneet S. Mehta, T. N. Thimm, M. Houlihan, J. S. Bredfeldt, E. J. Abel, and W. H. . K. W. Eliceiri. Collagen organization of renal cell carcinoma differs between low and high grade tumors. *BMC Cancer*, 490(19), 2019.
- [4] J. S. Bredfeldt, Y. Liu, M. W. Conklin, P. J. Keely, T. R. Mackie, and K. W. Eliceiri. Automated quantification of aligned collagen for human breast carcinoma prognosis. *J Pathol Inform.*, 5(28), 2014.
- [5] T. Cao, C. Zach, S. Modla, D. Powell, K. Czymmek, and M. Niethammer. Multi-modal registration for correlative microscopy using image analogies. *Medical Image Analysis*, 18(6):914 – 926, 2014. Sparse Methods for Signal Reconstruction and Medical Image Analysis.
- [6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- [7] Y.-H. Chen. *Multimodal Image Fusion and Its Applications*. PhD thesis, University of Michigan, 2016.
- [8] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, May 2018.
- [9] G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.
- [10] J. P. Cohen, H. Z. Lo, and W. Ding. Randomout: Using a convolutional gradient norm to win the filter lottery. *International Conference on Learning Representations (ICLR) Workshops*, 2016.
- [11] T. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant CNNs on homogeneous spaces. In *Advances in Neural Information Processing Systems 32*, pages 9145–9156. Curran Associates, Inc., 2019.
- [12] T. Cohen and M. Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 2016.
- [13] T. S. Cohen and M. Welling. Steerable CNNs. *arXiv preprint:1612.08498*, 2016.
- [14] M. W. Conklin, J. C. Eickhoff, K. M. Riching, C. A. Pehlke, K. W. Eliceiri, P. P. Provenzano, A. Friedl, and P. J. Keely. Aligned collagen is a prognostic signature for survival in human breast carcinoma. *American Journal of Pathology*, 3(178):1221–1232, 03 2011.
- [15] C. R. Drifka, A. G. Loeffler, K. Mathewson, A. Keikhosravi, J. C. Eickhoff, Y. Liu, S. M. Weber, W. John Kao, and K. W. Eliceiri. Highly aligned stromal collagen is a negative prognostic factor following pancreatic ductal adenocarcinoma resection. *Oncotarget*, 7(46):76197–76213, 2016.
- [16] K. Eliceiri. Multimodal Biomedical Dataset for Evaluating Registration Methods (patches from TMA Cores), June 2020.
- [17] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.
- [18] W. Ge. Deep metric learning with hierarchical triplet loss. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [19] A. A. Goshtasby. *Image Registration; Principles, Tools and Methods*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag London, 1 edition, 2012.
- [20] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304. PMLR, 2010.

- [21] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, June 2006.
- [22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning, 2019.
- [23] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423 – 1435, 2012. Special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention.
- [24] O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272, 2019.
- [25] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.
- [26] J. Hu, S. Sun, X. Yang, S. Zhou, X. Wang, Y. Fu, J. Zhou, Y. Yin, K. Cao, Q. Song, and X. Wu. Towards accurate and robust multi-modal medical image registration using contrastive metric learning. *IEEE Access*, 7:132816–132827, 2019.
- [27] M. J. Huttunen, R. Hristu, A. Dumitru, I. Floroiu, M. Costache, and S. G. Stanciu. Multiphoton microscopy of the dermoepidermal junction and automated identification of dysplastic tissues with deep learning. *Biomed. Opt. Express*, 11(1):186–199, Jan 2020.
- [28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.
- [29] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops*, pages 11–19, 2017.
- [30] A. Keikhosravi, B. Li, Y. Liu, and K. W. Eliceiri. Intensity-based registration of bright-field and second-harmonic generation images of histopathology tissue sections. *Biomed. Opt. Express*, 11(1):160–173, Jan 2020.
- [31] D. S. Kerby. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11–IT, 2014.
- [32] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning, 2020.
- [33] T. Kipf, E. van der Pol, and M. Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.
- [34] B. L., V. J., H. J.F., V. A., and T. P.H.S. Fully-convolutional siamese networks for object tracking. *European Conference on Computer Vision (ECCV) 2016 Workshops*, 9914, 2016.
- [35] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018.
- [36] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. K. Singh, and M.-H. Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.
- [37] K. Lin, B. Huang, L. M. Collins, K. Bradbury, and J. M. Malof. A simple rotational equivariance loss for generic convolutional segmentation networks: preliminary results. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3876–3879, 2019.
- [38] J. Lindblad and N. Sladoje. Linear time distances between fuzzy sets with applications to pattern matching and classification. *IEEE Transactions on Image Processing*, 23(1):126–136, 2013.
- [39] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
- [40] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.
- [41] Materials Genome Initiative. <https://www.mgi.gov>. Accessed: 2020-06-01.

- [42] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank. Nonrigid multimodality image registration. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 1609 – 1620. International Society for Optics and Photonics, SPIE, 2001.
- [43] I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations, 2019.
- [44] J. Öfverstedt, J. Lindblad, and N. Sladoje. Fast and robust symmetric image registration based on distances combining intensity and spatial information. *IEEE Transactions on Image Processing*, 28(7):3584–3597, 2019.
- [45] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [46] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [47] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 09–15 Jun 2019.
- [48] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [49] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29*, pages 1857–1865. Curran Associates, Inc., 2016.
- [50] M. Styner, C. Brechbuhler, G. Szckely, and G. Gerig. Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE Transactions on Medical Imaging*, 19(3):153–165, 2000.
- [51] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Contrastive bidirectional transformer for temporal representation learning. *CoRR*, abs/1906.05743, 2019.
- [52] R. Tao, E. Gavves, and A. W. Smeulders. Siamese instance search for tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [53] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- [54] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.
- [55] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems 29*, pages 4170–4178. Curran Associates, Inc., 2016.
- [56] E. van der Pol, T. Kipf, F. A. Oliehoek, and M. Welling. Plannable approximations to MDP homomorphisms: Equivariance under actions, 2020.
- [57] M. Volpi and V. Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.
- [58] C. Wachinger and N. Navab. Entropy and laplacian images: Structural representations for multi-modal registration. *Medical Image Analysis*, 16(1):1 – 17, 2012.
- [59] M. Weiler and G. Cesa. General  $e(2)$ -equivariant steerable CNNs. In *Advances in Neural Information Processing Systems 32*, pages 14334–14345. Curran Associates, Inc., 2019.
- [60] M. Weiler, F. A. Hamprecht, and M. Storath. Learning steerable filters for rotation equivariant CNNs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [61] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, 2006.
- [62] H. Wu, A. Gattami, and M. Flierl. Conditional mutual information-based contrastive loss for financial time series forecasting. *arXiv preprint arXiv:2002.07638*, 2020.
- [63] H. Zhang, W. Ni, W. Yan, D. Xiang, J. Wu, X. Yang, and H. Bian. Registration of multimodal remote sensing image based on deep fully convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):3028–3042, Aug 2019.
- [64] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

## 7 Appendix

### 7.1 Algorithm

In this section, we present a general algorithm for (supervised) training of models for generating CoMIRs given a dataset of aligned images. The CL used in this work is not fundamentally limited to two modalities, but can be extended for  $M$  modalities.

Let  $\mathcal{D} = \{(\mathbf{x}_i^1, \dots, \mathbf{x}_i^M)\}_{i=1}^n$  be an i.i.d. dataset containing  $n$  data points, where  $\mathbf{x}^j$  is an image in modality  $j$ , and  $f_{\theta_j}$  the network processing modality  $j$  with respective parameters  $\theta_j$  for  $j \in \{1, \dots, M\}$ . For an arbitrary datapoint  $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^M) \in \mathcal{D}$ ,  $i \in \{1, 2, \dots, n\}$  the loss is given by

$$\mathcal{L}^{opt}(\mathcal{D}) = -\mathbb{E}_{\mathbf{x}^1, \dots, \mathbf{x}^M \sim \mathcal{D}} \left[ \log \frac{\frac{p(\mathbf{x}^1, \dots, \mathbf{x}^M)}{\prod_{i=1}^M p(\mathbf{x}^i)}}{\frac{p(\mathbf{x}^1, \dots, \mathbf{x}^M)}{\prod_{i=1}^M p(\mathbf{x}^i)} + \sum_{\mathbf{x}_j \neq \mathbf{x}} \frac{p(\mathbf{x}_j^1, \dots, \mathbf{x}_j^M)}{\prod_{i=1}^M p(\mathbf{x}_j^i)}} \right] \quad (5)$$

The pseudo-code for training models using the CL to produce CoMIRs, based on aligned images, is given in algorithm 1.

---

#### Algorithm 1 CoMIR learning algorithm

---

**input:** batch size  $N$ , set of datapoints  $\{\mathbf{x}_i\}_{i=1}^N$  in  $M$  modalities, a group  $\mathcal{G}$  (e.g.  $\mathcal{C}_4$ ), a critic function  $h(\cdot, \cdot)$ , a temperature  $\tau$ .  
initialize all models  $\{f_{\theta_i}\}_{i=1}^M$   
**for** sampled mini-batch  $\mathbf{x} = \{\mathbf{x}_k\}_{k=1}^N$  **do**  
    # Computation of the latent spaces  
    **for all**  $m \in \{1, \dots, M\}$  **do** # for each modality  
        **for all**  $k \in \{1, \dots, N\}$  **do** # for each element in batch  
            draw an operation  $T \sim \mathcal{G}$   
             $\mathbf{y}_k^m = T'(f_{\theta_m}(T(\mathbf{x}_k^m)))$   
        **end for**  
    **end for**  
    # Computation of the similarity matrix  
     $\mathbf{z} = \|\mathbf{y}_i^1\|_{i=1}^M$  # Concatenated latent spaces ( $MN \times 1$ )  
     $\mathbf{S} \in \mathbb{R}^{MN \times MN}$   
    **for all**  $i \in \{1, \dots, MN\}$  **do**  
        **for all**  $j \in \{1, \dots, i\}$  **do**  
             $\mathbf{S}_{i,j} = \mathbf{S}_{j,i} = \log h(\mathbf{z}_i, \mathbf{z}_j) - \log \tau$   
        **end for**  
    **end for**  
    # Computation of the loss  
     $\mathcal{L} = 0$   
    **for all**  $i \in \{1, \dots, MN\}$  **do**  
        **for all**  $m \in \{2, \dots, M\}$  **do**  
             $\mathcal{L} = \mathcal{L} - \mathbf{S}_{(i, mN+i)\%MN} + \log \left( \sum_{j=1}^{MN} e^{\mathbf{S}_{i,j}} - \sum_{j=1}^M e^{\mathbf{S}_{(i, jN+i)\%MN}} + e^{\mathbf{S}_{(i, mN+i)\%MN}} \right)$   
        **end for**  
    **end for**  
     $\mathcal{L} = \frac{1}{MN(M-1)} \mathcal{L}$   
    update all models  $\{f_{\theta_i}\}_{i=1}^M$  to minimize  $\mathcal{L}$   
**end for**  
**return** all models  $\{f_{\theta_i}\}_{i=1}^M$

---

### 7.2 Datasets

**Zurich Dataset** The dataset can be downloaded at <https://sites.google.com/site/michelevoipiresearch/data/zurich-dataset>. Each image was acquired with QuickBird and consists of four channels in the order [NIR,B,G,R]. An example image of the dataset is given in Fig. 5. In all our experiments we train on only one image of the Zurich zh10.tif, except for the experiments in table 1 for which we investigate the impact of the training distribution and vary which image from the Zurich dataset is used as the training image. As a test image to evaluate the representations zh12.tif was chosen.



Figure 5: Example image of the Zurich dataset. To the left the RGB image, to the right the corresponding NIR image.

**Biomedical Dataset** The dataset used in this study was kindly provided by the authors of [30] and modified for this study. The modified dataset used in our experimental setup can be downloaded at <https://zenodo.org/record/3874362>, [16]. Multimodal Image Registration is of particular interest in biomedical tasks such as the registration of bright-field (BF) microscopy images and second-harmonic generation (SHG). The two ways of imaging provide complementary information and are of particular interest in studying collagen organization in cancerous tissue, [3], [15] [14] and [4]. In general the two modalities are often captured in different microscopes which means that the SHG sample must be relocated and registered within a BF scan. This is a challenging task as the modalities result in very different signals and have little appearance in common. The difficulties of this registration task are discussed in detail in [30], which provides the first intensity-based registration method, called CurveAlign, capable of automatically aligning SHG images and BF images that are usually aligned manually. They evaluate their methods on tissue microarray cores as can be seen in Fig. 6 to the left. In this paper we use center-cropped patches within these cores, marked by the green square. This registration task is harder as the overview of larger structures and the tissue boundary are missing. Using these patches allows us to avoid any padding effects after applying random transformations as described in 4.2. The GT alignment used for the evaluation is based on manual landmark annotation provided by [30] and was done on the full tissue microarray cores. The manual registration performed as part of the evaluation in this paper is however performed on the cropped patches and by overlaying the SHG image on to the BF image. The two kinds of manual registrations (GT and manual registration in the evaluation) are hence not one to one comparable. The manual registration in the evaluation was done according to the setup used in the automated methods, to give a baseline and sense of the magnitude of pixel error that can be expected in this task. By enabling registration on patches within the tissue micro arrays, a further downstream task of patch retrieval can be achieved. This is in particular interesting for finding SHG patches within BF whole slide images as is relevant in ongoing research such as [27].

### 7.3 Implementation Details

We chose the same dense U-Net architectures for experiments on the Zurich and the biomedical dataset [29], using 32 convolutional filters for the first convolution, 4 dense blocks of depth 6 as down and up blocks and 4 bottleneck layers. Upsampling was used to avoid grid artifacts. Max pooling, a dropout rate of 0.2, no early transition or activation function in the last layer and a compression rate of the convolutional layers of 0.75 were used. The commonly used non-linear activation in the final layer is omitted. The 1-channel CoMIRs illustrated in 3 are visualized by applying a logistic function with temperature of 0.5. Illustrations done on the Zurich dataset are made by normalizing two corresponding CoMIRs jointly by the maximum of the 1-percentile and the minimum of the 99-percentile of the two representations.

**Experiments on Zurich Dataset** For generating CoMIRs in experiments on the Zurich dataset, Adam optimizer was used with a learning rate of  $1e-3$ , a weight decay of  $1e-4$ . Temperature  $\tau$  was set to 0.1, the batch size to 24, and the steps per epoch to 32. The gradient norm was limited to 1.  $L_1$  and  $L_2$  activation decay were set to 0. To generate a batch, patches of size  $128 \times 128$  were cropped from the training image. Data augmentation consisted of flips ( $p = 0.5$ ) and random rotations from by up to  $\pm 180^\circ$  using either a linear, nearest neighbor or cubic interpolation randomly. With a probability  $p = 0.2$  either additive Gaussian noise ( $\mu = 0$ ,  $\sigma \in (0., 0.05)$  randomly chosen), Gaussian blur ( $\sigma = 0.1$ ), or coarse dropout (rate of dropout of 10% per channel, superpixel



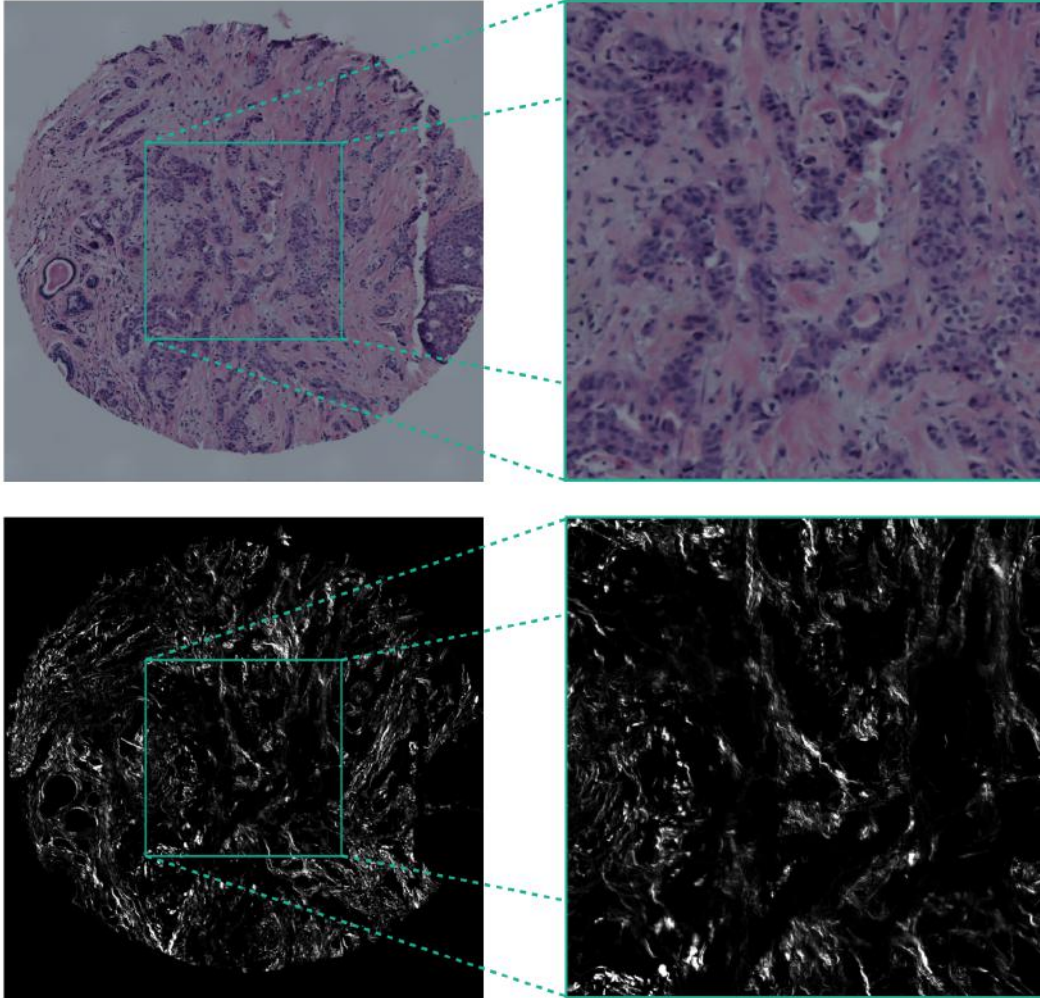


Figure 6: This example image pair of the test set shows the different usage of the data introduced and registered in [30]. The registration problem becomes harder by using the cropped patches marked in green, as larger structures that can aid the registration are not available. The GT of the dataset as obtained in [30] by a manual user choosing landmarks was performed on the large images to the left, while the manual registrations in our evaluation are performed on the patches to the right as described in 4.2.

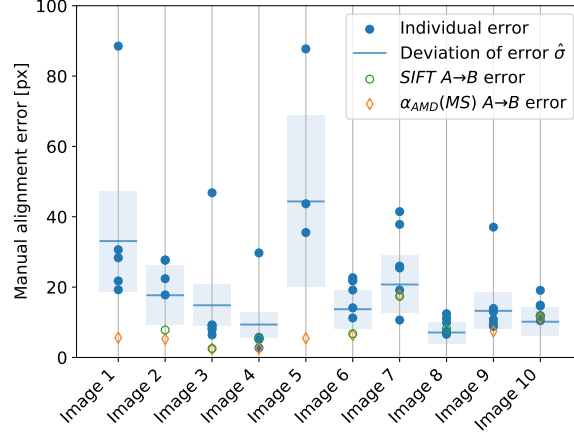


Figure 7: Variation in manual annotation errors across test patches (blue dots represent errors of individual annotators) compared to the best performing registration methods,  $\alpha_{AMD}$  and SIFT, applied on CoMIRs. The parameters  $\hat{\sigma}$  are estimated along with their 95% confidence interval (Rayleigh distribution).

size of 5%) is applied or the edge image is taken. Lastly, each channel is multiplied randomly ( $p=0.3$ ) by a value in  $(0.9, 1.1)$ . The models were trained for 5 epochs.

**Experiments on Biomedical Dataset** For generating 1-channel CoMIRs for the multimodal registration experiments, stochastic gradient descent was used with a learning rate of  $1e-2$ , a weight decay of  $1e-5$  and a momentum of 0.9. Temperature  $\tau$  was set to 0.5, the batch size to 32, and the steps per epoch to 32. The gradient norm was limited to 1.  $L_1$  activation decay was set to  $1e-4$ ,  $L_2$  activation decay to  $1e-4$ . To generate a batch, patches of size  $128 \times 128$  were cropped from the training image and the data augmentation consisted of flips ( $p = 0.5$ ) and random integer rotations by up to  $\pm 180^\circ$  using either a linear, nearest neighbor or cubic interpolation randomly. The models were trained for 23 epochs in the efficient mode of the model implementation.

**Registration using  $\alpha$ -AMD** For  $\alpha$ -AMD, 3 resolution levels were used with sub-sampling factors given by 4, 2, 1 and Gaussian smoothing  $\sigma$  given by 12, 5, 1 respectively. A SGD optimizer was used with a momentum of 0.9, and step-sizes  $(2, 2, 2 \rightarrow 0.2)$  where the transition in the last resolution level is a linear interpolation between 2 and 0.2, and a hard gradient clipping at a magnitude of 1, parameter-wise. The logistic function (without temperature) is applied to the CoMIRs to map them into the range  $[0, 1]$ , which is needed for the method, and then the images were quantized into 7 (non-zero) levels. The number of iterations, per resolution level, are 3000, 1000, 500 respectively. A sampling fraction of 0.005 was used. For the multistart version of the method, three starts were chosen at rotations (in radians) in  $\{-0.3, 0, 0.3\}$ , and the registration with the lowest final distance value was chosen as the output of the algorithm. We used a modified version of an open-source implementation ([https://github.com/MIDA-group/py\\_alpha\\_amd\\_release](https://github.com/MIDA-group/py_alpha_amd_release)) of the method. We provide the modified version in the github repository.

**Registration using SIFT** The SIFT implementation in Fiji 2.0.0 was used based on the `mpicbg.imagefeatures` package. The feature descriptor size was set to 4 samples per row and column, the orientation bins for 8 bins per local histogram. The scale octaves were set to be in  $[128, 1024]$ px with 3 steps per scale octave and an initial  $\sigma$  of each scale octave equal to 1.6.

**Registration by Mutual Information** Registration by Mattes MI was implemented in Matlab 2019b using a (1+1) evolutionary algorithm with an initial size of the search radius of  $1e-5$ , a minimum size of the search radius of  $1.5e-8$ , a growth factor of the search radius of  $(1+1e-4)$  and a maximum number of iterations of 1500. The number of spatial samples for the MI computation was 500 and the number of histogram bins 80. All pixels were included in the overlap region. As registration by MI required 1-channel input, all 3-channel representations (BF and GAN generated BF images) have been reduced to one channel by principal component analysis (PCA).

**Registration by CurveAlign** CurveAlign v5.0 Beta provides HSV color based registration of SHG and BF images. The method was run with the default settings, however it was set to rigid registration, where the default is affine. The code is provided in <https://github.com/uw-loci/curvelets>.

**Manual Registration** The registration was performed using in Fiji 2.0.0 with TrakEM 2. The mean error of the corners after manual registration for ten randomly chosen images of the test set is shown in Fig. 7.

Table 2: Number of successful registrations on the test set (N=134) for multiscale  $\alpha$ -AMD,  $\alpha$ -AMD, SIFT, MI and CurveAlign which resulted in less than 1% pixel error ( $< 9\text{px}$ ) and less than 5% pixel error ( $< 42\text{px}$ ).  $\widehat{\text{BF}}$  and  $\widehat{\text{SHG}}$  denote the fake images produced by a GAN image translation given the respective other modality. 95% C.I. are Clopper-Pearson intervals.

Reg. Method	MS $\alpha$ -AMD		$\alpha$ -AMD		SIFT		MI		CurveAlign
Originals	Input	—	—	—	—	—	$\widehat{\text{BF}} \rightarrow \text{SHG}$	$\text{SHG} \rightarrow \widehat{\text{BF}}$	$\widehat{\text{BF}} \rightarrow \text{SHG}$
	Succ.( $< 1\%$ Err.)	—	—	—	—	—	7 [3; 14]	7 [3; 14]	<b>22</b> [14; 32]
	Succ.( $< 5\%$ Err.)	—	—	—	—	—	30 [21; 41]	29 [20; 40]	<b>33</b> [24; 44]
CoMIR MSE	Input	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A
	Succ.( $< 1\%$ Err.)	72 [60; 84]	70 [58; 82]	49 [38; 61]	43 [33; 55]	<b>75</b> [63; 86]	67 [55; 79]	7 [3; 14]	9 [4; 17]
	Succ.( $< 5\%$ Err.)	90 [78; 101]	87 [75; 98]	63 [51; 75]	54 [43; 66]	<b>98</b> [87; 108]	93 [82; 103]	30 [21; 41]	30 [21; 41]
CoMIR Cos.	Input	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A	A $\rightarrow$ B	B $\rightarrow$ A
	Succ.( $< 1\%$ Err.)	<b>54</b> [43; 66]	53 [42; 65]	19 [12; 28]	18 [11; 27]	47 [36; 59]	52 [41; 64]	12 [6; 20]	11 [6; 19]
	Succ.( $< 5\%$ Err.)	62 [50; 74]	62 [50; 74]	34 [24; 45]	33 [24; 44]	67 [55; 79]	<b>69</b> [57; 81]	29 [20; 40]	27 [18; 37]
CycleGAN	Input	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$
	Succ.( $< 1\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	<b>6</b> [2; 13]	5 [2; 11]
	Succ.( $< 5\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	<b>30</b> [21; 41]	28 [19; 39]
Pix2Pix	Input	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$
	Succ.( $< 1\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	6 [2; 13]	7 [3; 14]
	Succ.( $< 5\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	29 [20; 40]	<b>30</b> [21; 41]
DRIT	Input	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$	$\widehat{\text{BF}} \rightarrow \text{BF}$	$\widehat{\text{SHG}} \rightarrow \text{SHG}$
	Succ.( $< 1\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	5 [2; 11]	<b>6</b> [2; 13]
	Succ.( $< 5\%$ Err.)	—	—	—	—	0 [0; 4]	0 [0; 4]	<b>31</b> [22; 42]	29 [20; 40]

**Alternative methods** pix2pix, CycleGAN and DRIT train on image patches created and augmented in the same way as for the generation of CoMIRs, however the patches are not created during runtime but before training. The code for the competing methods are provided in <https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix> for pix2pix and CycleGAN and <https://github.com/HsinYingLee/DRIT> for DRIT.

The pix2pix and CycleGAN models were trained for 100 epochs with the initial learning rate of  $2e - 4$  and another 100 epochs with a linearly decaying learning rate thereafter. Adam optimizer was used with  $\beta_1 = 0.5$ . The GAN-objective was  $\text{ls}_{\text{gan}}$ , the discriminator architecture was a  $70 \times 70$  PatchGAN and a 9-block ResNet for the generator architecture. 64 filters were used in the last convolutional layer of the generator as well as the first generator of the discriminator. Instance normalization was used and no dropout for the generator. During the test phase, the input image sizes were padded to minimum multiples of 256 (i.e. as the size of input images were 834, they were padded to 1024) due to the network architectures' restriction for pix2pix [28] and CycleGAN [64]. To avoid edge artifacts during inference, the padded borders were filled with the reflection of the vector mirrored on the first and last values of the vector along each axis. The padded areas of translated images are cropped off before further processing.

DRIT [35, 36] was adjusted to fit the registration task, by using the disentangled attribute representation of the image for registration with the corresponding image, rather than using the disentangled attribute representation of a random given image from the other modality. The discriminator had no normalization layers and its scale was set to 3. The model was trained for 1200 epochs, with a learning rate decay for the last 600 epochs.

## 7.4 Results

In table 2 the number of successfully registered image pairs which resulted in less than 1% relative error to the image width and length ( $< 9\text{px}$ ), as well as 5% relative error ( $< 42\text{px}$ ) is given for each applicable registration method. Fig. 7 shows the results of the manual registration on ten image pairs from the test set together with the result of the automated registration based on  $\alpha$ -AMD and SIFT. Fig. 8 shows the error of registration results based on MI with respect to the displacement of the corner points resulting from the transformation applied to the image. To point out the importance of the extent of the transformation which needs to be recovered in case of MI, SIFT results are also shown for comparison. Fig. 9 shows the eCDF of automatic registrations of the biomedical test set on the CoMIRs learnt by using a critic based on MSE and cosine similarity. Fig. 10 shows the influence weight initialization and a particular training image have on the CoMIRs (3 channels) generated of an image patch of image `zh12.tif` in the Zurich test set. Fig. 11 shows the influence of temperature  $\tau$  and the choice of critic on the generated CoMIRs (3-channel) of `zh12.tif`.

### 7.4.1 Time Analysis

The experiments are run on an Intel(R) Core(TM) i9-7980XE CPU @ 2.60GHz (hyperthreading enabled) and a Titan V (12GB) GPU. Table 3 shows the approximate time needed for registration using MI on the CPU versus generating CoMIR and registration using SIFT. The inference includes the loading of the dataset and the time reported includes generation and registration using each modality as a reference image (A  $\rightarrow$  B and B  $\rightarrow$  A).

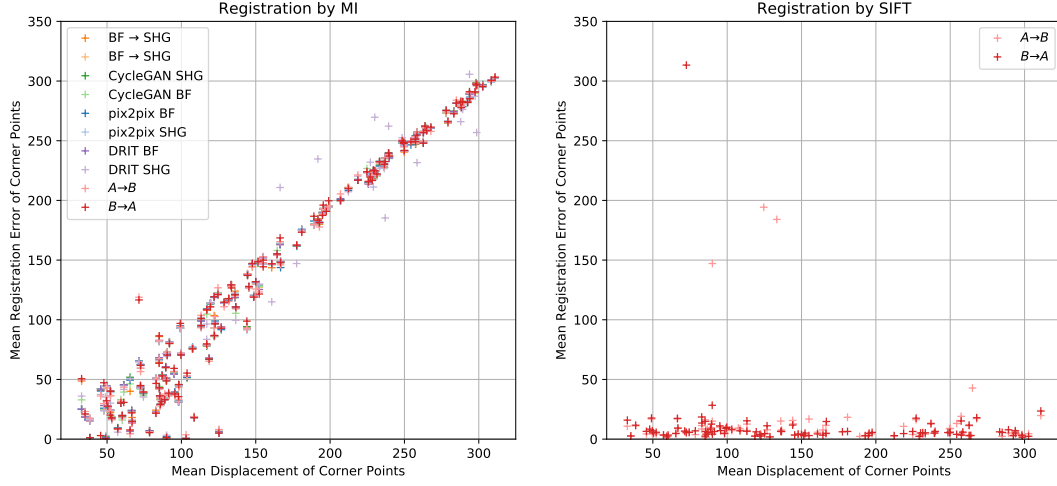


Figure 8: The mean registration error of the corner points vs. the mean displacement of corner points is shown for registration by MI as well as SIFT. It can be seen that MI suffers from large transformations and fails to detect a global extremum if the initial starting position of the optimization is too far away. This dependence of the accuracy on the starting position can be observed for all modalities. In contrast, SIFT does not depend on the extent of the transformation between the images that have to be registered, as can be seen to the right.

Table 3: Approximate time for steps in registration pipelines given in seconds for the entire test set.

Device	Preprocessing		Registration using MI (Matlab)	Preprocessing + Registration
CPU 1 worker	114		9593	9707
CPU 18 worker	—		896	1010
	Training 23 epochs (Pytorch/Python)	Inference (Pytorch/Python)	Registration using SIFT (Fiji/Java)	Inference + Registration
GPU	1345	49	—	324
CPU 1 worker	—	2631	275	2906
CPU 18 worker	—	1020	—	1295

While registration using MI will scale linearly with the number of image pairs which need to be registered, the training of the proposed model can be considered constant overhead and the generation of CoMIRs takes only 0.2 seconds per image on a GPU. After training the inference can also be done on the CPU, albeit is slower, but can be beneficial in a clinical setting where no access to a GPU can be given, or to encode images of large spatial dimensions which do not fit onto a GPU. It should be noted that no parallelization attempt was made for registration by SIFT nor for the preprocessing (turning the images to 1-channel images by PCA), which could speed up the registration further.

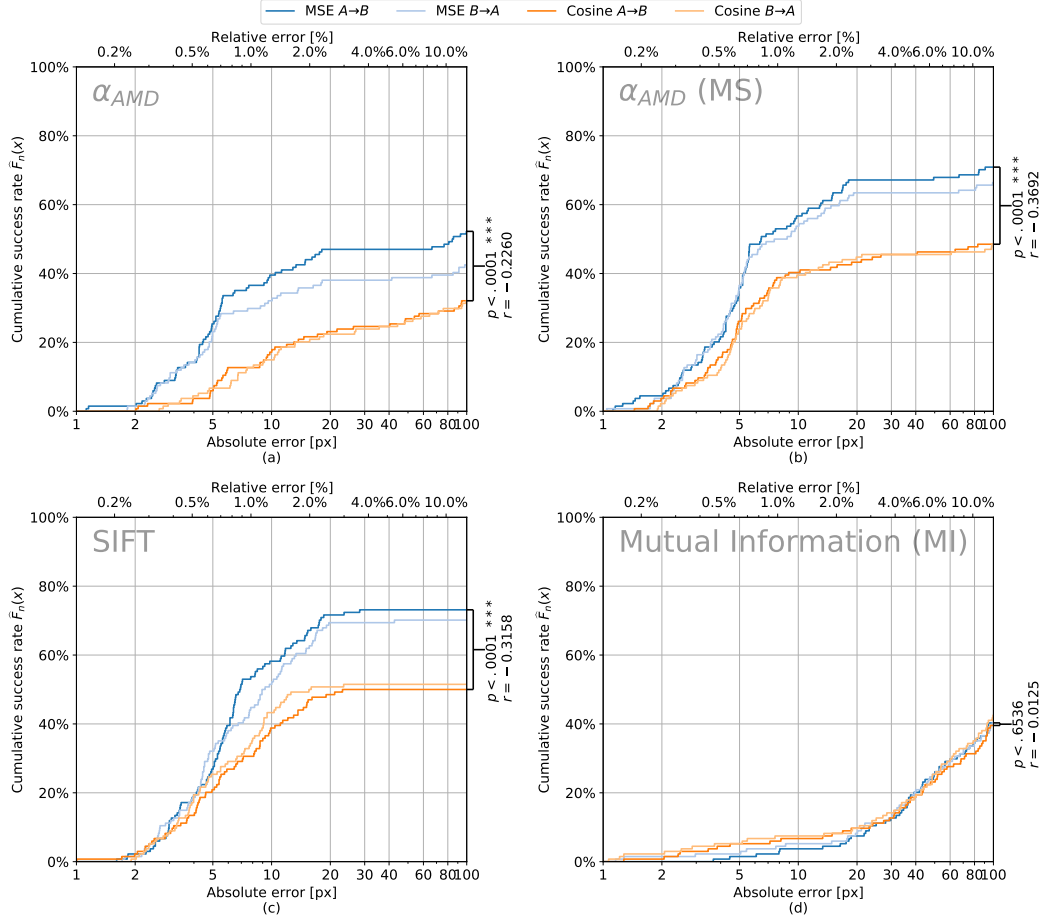
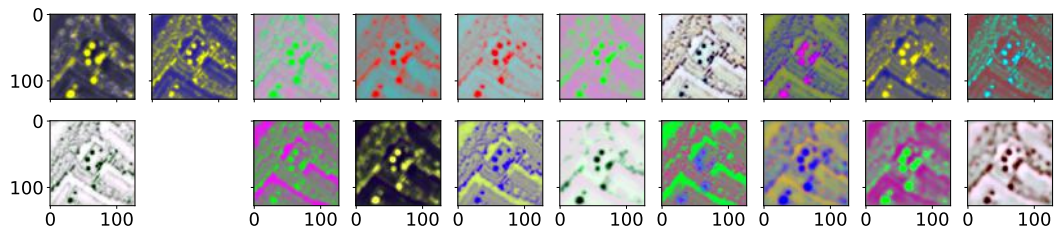
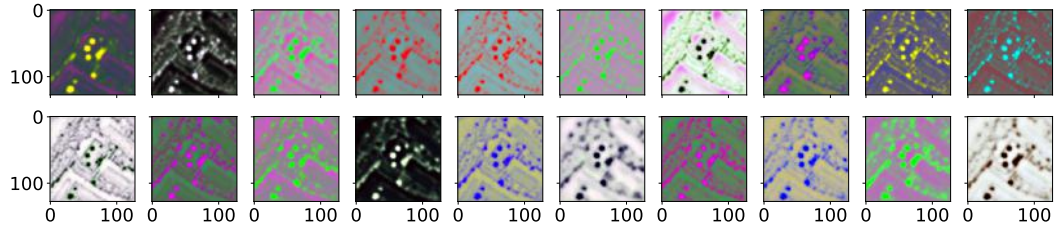


Figure 9: eCDF of the automatic registration methods over the biomedical test dataset for CoMIRs produced by the MSE-based critic as well as the cosine similarity-based critic ((a)  $\alpha$ -AMD, (b)  $\alpha$ -AMD MS, (c) SIFT and (d) MI). Wilcoxon tests between the results based on MSE and cosine similarity are shown on the right side of each plot, along with the matched-pairs rank-biserial correlation  $r$  (effect size)[31].

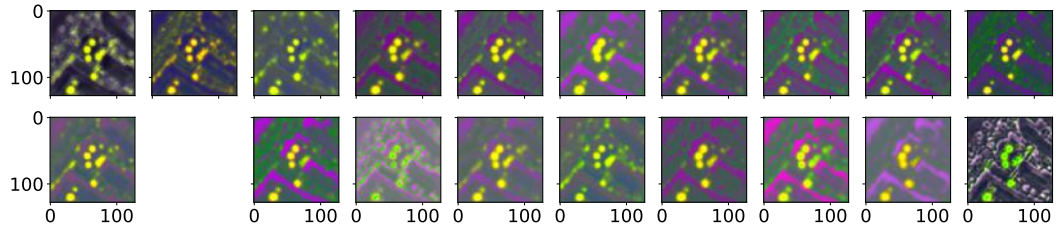




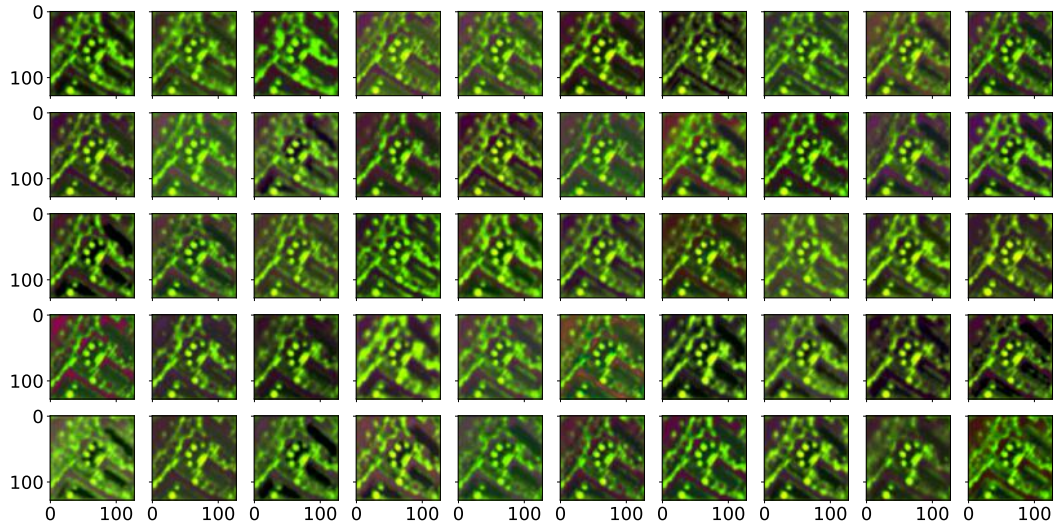
(a) CoMIRs of models trained with different weight initializations and different images.



(b) CoMIRs of models trained with different weight initializations and a unique common image.



(c) CoMIRs of models trained with the same weight initialization and different images.



(d) CoMIRs of models trained with the same weight initialization and a unique common image.

Figure 10: Different experiments were performed to show that a random weight initialization scheme with a fixed seed yields similar models. Using a small training set increases the similarity of the trained models. All models were trained on only one training image, either fixed or varying, depending on the experiment. (a) and (c) have the 12th image omitted as they were trained on the same image as the test image, used to create this figure.



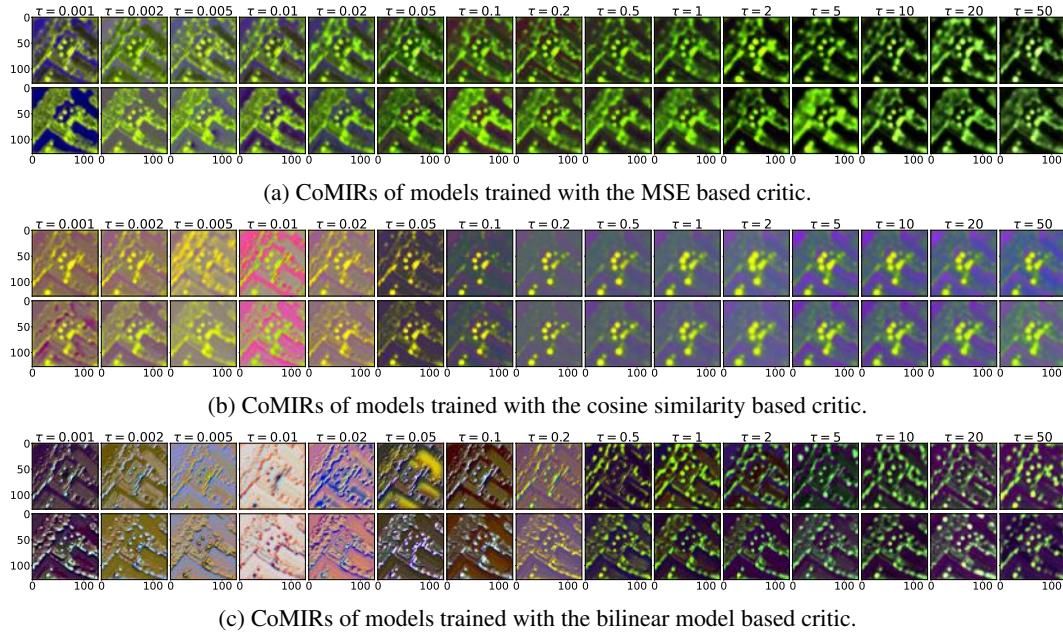


Figure 11: CoMIRs generated from models trained with different losses and temperatures  $\tau$ . Top-rows are generated with modality 1 (RGB), bottom rows with modality 2 (NIR). A high temperature makes the CoMIRs more blurred. The models were initialized with the same seed.