Unsupervised Multi-Modal Image Registration via Geometry Preserving Image-to-Image Translation

Moab Arar Yiftach Ginger Dov Danon Amit H. Bermano Daniel Cohen-Or
Tel Aviv University

Abstract

Many applications, such as autonomous driving, heavily rely on multi-modal data where spatial alignment between the modalities is required. Most multi-modal registration methods struggle computing the spatial correspondence between the images using prevalent cross-modality similarity measures. In this work, we bypass the difficulties of developing cross-modality similarity measures, by training an image-to-image translation network on the two input modalities. This learned translation allows training the registration network using simple and reliable mono-modality metrics. We perform multi-modal registration using two networks - a spatial transformation network and a translation network. We show that by encouraging our translation network to be geometry preserving, we manage to train an accurate spatial transformation network. Compared to state-of-the-art multi-modal methods our presented method is unsupervised, requiring no pairs of aligned modalities for training, and can be adapted to any pair of modalities. We evaluate our method quantitatively and qualitatively on commercial datasets, showing that it performs well on several modalities and achieves accurate alignment.

1. Introduction

Scene acquisition using different sensors is common practice in various disciplines, from classical ones such as medical imaging and remote sensing, to emerging tasks such as autonomous driving. Multi-modal sensors allow gathering a wide range of physical properties, which in turn yields richer scene representations. For example, in radiation planning, multi-modal data (e.g. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans) is used for more accurate tumor contouring which reduces the risk of damaging healthy tissues in radiotherapy treatment [25, 29]. More often than not, multi-modal sensors naturally have different extrinsic parameters between modalities, such as lens parameters and relative position. In these cases, non-rigid image registration is essential for proper execution of the aforementioned downstream tasks.

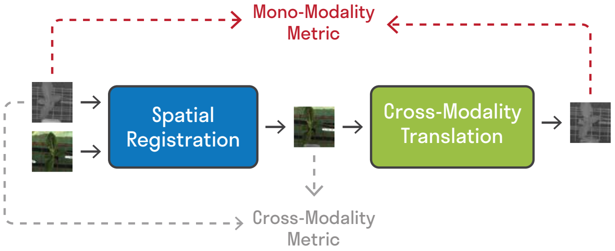


Figure 1: **Method overview.** Conventional methods (faded dashed at the bottom) use cross-modality metrics (e.g., Normalized Cross Correlation) to optimize a spatial transformation function. Our method learns a cross-modality translation, mapping between the two modalities. This enables the use of a reliable accurate mono-modality metric instead.

Classic *multi-modal image registration* techniques attempt to warp a source image to match a target one via a non-linear optimization process, seeking to maximize a predefined similarity measure [40]. Besides a computational disadvantage, which is critical for applications such as autonomous driving, effectively designing similarity measures for such optimization has proven to be quite challenging. This is true for both intensity-based measures, commonly used in the medical imaging [10], and feature-based ones, typically adapted for more detailed modalities (e.g Near Infra-Red (NIR) and RGB) [32].

These difficulties gave rise to the recent development of deep regression models. These types of models typically have lengthy training time, either supervised or unsupervised, yet they offer expeditious inference that usually generalizes well. Since it is extremely hard to collect ground-truth data for the registration parameters, supervised multi-modal registration methods commonly use synthesized data in order to train a registration network [30, 37]. This makes their robustness highly dependent on the similarity between the artificial and real-life data distribution and appearance. Unsupervised registration techniques, on the other-hand, often incorporate a spatial transform network (STN) [14] and train an end-to-end network [7, 19, 16, 36, 8].

Typically, such approaches optimize an STN by comparing the deformed image and the target one using simple similarity metrics such as pixel-wise Mean Squared Error (MSE) [31, 33, 6]. Of course, such approaches can only be used in mono-modality settings and become irrelevant for multi-modality settings. To overcome this limitation, unsupervised multi-modal registration networks use statistics-based similarity metrics, particularly, (Normalized) Mutual Information ((N)MI) [22], Normalized Cross Correlation (NCC) [5], or Structural Similarity Index Metric (SSIM) [21, 22] (see Figure 1, faded dashed path). However, these metrics are either computationally intractable (e.g., MI) [3] and hence cannot be used in gradient-based methods, or are domain-dependent (e.g., NCC), failing to generalize for all modalities.

In this paper, we present an unsupervised method for multi-modal registration. In our work, we exploit the celebrated success of Multi-Modal Image Translation [13, 38, 39, 12], and simultaneously learn multi-modal translation and spatial registration. The key idea is to alleviate the shortcomings of a hand-crafted similarity measure by training an image-to-image translation network T on two given modalities. This in turn will let us use mono-modality metrics for evaluating our registration network R (see Figure 1, vivid path on the top).

The main challenge for this approach is to train the registration network R and the translation network T simultaneously, while encouraging T to be geometry preserving. This ensures that the two networks are task-specific — T performs only a photo-metric mapping, while R learns the geometric transformation required for the registration task. In our work, we use the concepts of generative adversarial networks (GAN [9, 24]) to train T and R. We show that the adversarial training is not only necessary for the translation task (as shown in previous works [13]), but is also necessary to produce smooth and accurate spatial transformation. We evaluate our method on real commercial data, and demonstrate its strength with a series of studies.

The main contributions of our work are:

- An unsupervised method for multi-modal image registration.
- A geometry preserving translation network that allows the application of mono-modality metrics in multi-modal registration.
- A training scheme that encourages a generator to be geometry preserving.

2. Related Works

To deal with the photo-metric difference between modalities, unsupervised multi-modal approaches are forced to find the correlation between the different domains and use it to guide their learning process. In [21] a vanilla CycleGAN architecture is used to regularize a deformation mapping. This is achieved by training a discriminator network to distinguish between deformed and real images. To align a pair of images the entire network needs to be trained in a single pass. Training this network on a large dataset will encourage the deformation mapping to become an identity mapping. This is because the discriminator is given only the real and deformed images. Furthermore the authors use multiple cross-modality similarity metrics including SSIM, NCC and NMI which are limited by the compatibility of the specific modalities used. In contrast, our method learns from a large dataset and bypasses the need for cross-modality similarity metrics.

Wang *et al*. [36] attempt to bypass the need for domain translation by learning an Encoder-Decoder module to create modality-independent features. The features are fed to an STN to learn affine and non-rigid transformations. The authors train their network using a simple similarity measure (MSE) which maintains local similarity, but does not enforce global fidelity.

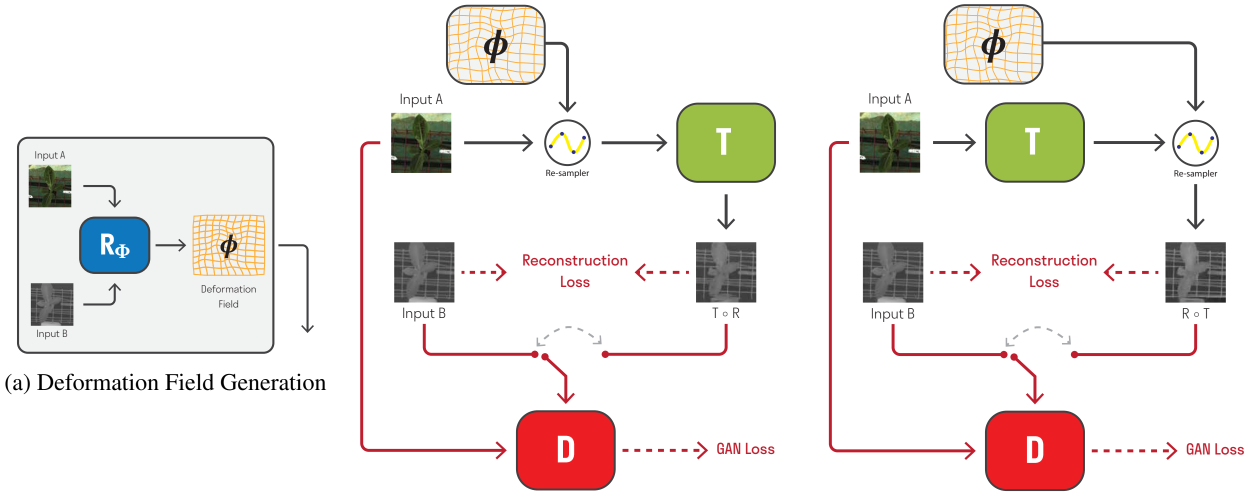
At the other extreme, [8] rely entirely on an adversarial loss function. They train a regular U-Net based STN by giving the resultant registered images to a discriminator network and using its feedback as the STN's loss function. By relying solely on the discriminator network for guiding the training, they lose the ability to enforce local coherence between the registered and target images.

Closest to our work, [27] combines an adversarial loss with similarity measurements in an effort to register the images properly while concentrating on maintaining local geometric properties. They encode the inputs into two separate embedding, one for shape and one for content information, and train a registration network on these disentangled embedding. This method relies on learned disentanglement, which introduces inconsistencies on the local level. Our method directly enforces the similarity in the image space, which leads to a reliable local signal.

3. Overview

Our core idea is to *learn* the translation between the two modalities, rather than using a cross-modality metric. This novel approach is illustrated in Figure 1. The spatially transformed image is translated by a learnable network. The translated image can then be compared to the original source image using a simple uni-modality metric, bypassing the need to use a cross-modality metric. The advantage of using a learnable translation network is that it generalizes and adapts to any pairs of given modalities.

Our registration network consists of two components: (i) a spatial transformation network $R = (R_\Phi, R_S)$ and (ii) an image-to-image translation network T. The two components are trained simultaneously using two training flows



(a) Deformation Field Generation

(b) Register First

(c) Translate First

Figure 2: **Training Flow Overview.** We train two components: (i) a spatial transformation network (STN) $R = (R_\Phi, R_S)$ and (ii) an image-to-image translation network T. The two networks R and T are jointly trained via two different training flows. The two training flows are simultaneously carried-out in each training step. In the first flow, *(b) Register First*, the input image I_a is deformed using ϕ, a deformation field generated by R_Φ, and is then fed to T to map the image onto domain B. The second flow, *(c) Translate First*, is similar with the exception that ϕ is used to transform the *translated* source image. In both cases, the same deformation field ϕ is used.

as depicted in Figure 2. The spatial transformation network takes the two input images and yields a deformation field ϕ. The field is the then applied either before T (Figure 2b) or after it (Figure 2c). Specifically, the field is generated using a network R_Φ and is used by a re-sampling layer R_S to get the transformed image, namely $R_S(T(a), \phi)$ and $T(R_S(a, \phi))$. We will elaborate on these two training schemes in Section 4.2. The key is, as we shall show, that such two-flow training encourages T to be geometry preserving, which implies that all the geometry transformation is encoded in R_Φ.

Once trained, only the spatial transformation network R is used in test time. The network takes two images I_a and I_b representing the same scene, captured from slightly different viewpoint, in two different modalities, A and B, respectively, and aligns I_a with I_b.

4. Method

Our goal is to learn a non-rigid spatial transformation which aligns two images from different domains. Let $\mathcal{A} \subset \mathbb{R}^{H_A \times W_A \times C_A}$ and $\mathcal{B} \subset \mathbb{R}^{H_B \times W_B \times C_B}$ be two paired image domains, where H_D, W_D, C_D are the height, width, and number of channels for domain $\mathcal{D}$, respectively. Pairing means that for each image $I_a \in \mathcal{A}$ there exists a unique image $I_b \in \mathcal{B}$ representing the same scene, as acquired by the different respective sensors. Note that the pairing assumption is a common and reasonable one, since more often than not registration-base applications involve taking an image of the same scene from both modality sensors (e.g satellite images). Throughout this section, we let $I_a \in \mathcal{A}$ and $I_b \in \mathcal{B}$ be a pair of two images such that I_a needs to be aligned with I_b.

To achieve this alignment, we train three learnable components: (i) a registration network R, (ii) a translation network T and (iii) a discriminator D. The three networks are trained using an adversarial model [9, 24], where R and T are jointly trained to outwit D. Below, we describe the design and objectives of each network.

4.1. Registration Network

Our registration network ($R = (R_\Phi, R_S)$) is a spatial transformation network (STN) composed of a fully-convolutional network R_Φ and a re-sampler layer R_S. The transformation we apply is a non-linear dense deformation - allowing non-uniform mapping between the images and hence gives accurate results. Next we give an in-depth description about each component.

$\mathbf{R_\Phi}$ - Deformation Field Generator: The network takes two input images, I_a and I_b, and produces a deformation field $\phi = R(I_a, I_b)$ describing how to non-rigidly align I_a to I_b. The field is an $H_A \times W_A$ matrix of 2-dimensional vectors, indicating the deformation direction for each pixel (i, j) in the input image I_a.

$\mathbf{R_S}$ - Re-sampling Layer: This layer receives the deformation field ϕ, produced by R_Φ, and applies it on a source image I_s. Here, the source image is not necessarily I_a and it could be from either domains - $\mathcal{A}$ or $\mathcal{B}$. Specifically, the value of the transformed image $R_S(I_s, \phi)$ at pixel $\mathbf{v} = (i, j)$ is given by Equation 1:

$$R_S(I_s, \phi)[v] = I_s\left[\mathbf{v} + \phi(\mathbf{v})\right], \tag{1}$$

where $\phi(\mathbf{v}) = (\Delta y, \Delta x)$ is the deformation generated by R_Φ at pixel $\mathbf{v} = (i, j)$, in the x and y-directions, respectively.

To avoid overly distorting the deformed image $R_S(I_s, \phi)$ we restrict R_Φ from producing non-smooth deformations. We adapt a common regularization term that is used to produce smooth deformations. In particular, the regularization loss will encourage neighboring pixels to have similar deformations. Formally, we seek to have small values of the first order gradients of ϕ, hence the loss at pixel $\mathbf{v} = (i, j)$ is then given by:

$$\mathcal{L}_{smooth}(\phi, \mathbf{v}) = \sum_{\mathbf{u} \in N(\mathbf{v})} B(\mathbf{u}, \mathbf{v}) \left\| \phi(\mathbf{u}) - \phi(\mathbf{v}) \right\|, \tag{2}$$

where $N(\mathbf{v})$ is a set of neighbors of the pixel $\mathbf{v}$, and $B\left(\mathbf{u}, \mathbf{v}\right)$ is a bilateral filter [34] used to reduce oversmoothing. Let $O_s = R_S(I_s, \phi)$ to be the deformed image produced by R_S on input I_s, then the bilateral filter is given by:

$$B(\mathbf{u}, \mathbf{v}) = e^{-\alpha \cdot \| O_s[\mathbf{u}] - O_s[\mathbf{v}] \|}. \tag{3}$$

There are two important notes about the bilateral filter B in Equation 3. First, the bilateral filtering is with respect to the transformed image O_s (at each forward pass), and secondly, the term $B\left(I_s, \mathbf{u}, \mathbf{v}\right)$ is a treated as constant (at each backward pass). The latter is important to avoid R_Φ alternating pixel values so that $B(u, v) \approx 0$ (e.g., it could change pixels so that $\|O_s[\mathbf{u}] - O_s[\mathbf{v}]\|$ is relatively large), while the former allows better exploration of the solution space.

In our experiments we look at the 3×3 neighborhood of $\mathbf{v}$, and set $\alpha = 1$. The overall smoothness loss of the network R, denoted by $\mathcal{L}_{smooth}\left(R\right)$, is the mean value over all pixels $\mathbf{v} \in \{1, \ldots, H_A\} \times \{1, \ldots, W_A\}$.

4.2. Geometric Preserving Translation Network

A key challenge of our work is to train the image-to-image translation network T to be **geometric preserving**. If T is geometric preserving, it implies that it only performs photo-metric mapping, and as a consequence the registration task is performed solely by the registration network R. However, during our experiments, we observed that T tends to generate fake images that are spatially aligned with the ground truth image, regardless of R's accuracy.

To avoid this, we could restrict T from performing any spatial alignment by reducing its capacity (number of layers). While we did observe that reducing T's capacity does improve our registration network's performance, it still limits the registration network from doing all the registration task (See supplementary materials).

To implicitly encourage T to be geometric preserving we require that T and R are commutative, i.e., $T \circ R = R \circ T$. In the following we formally define both $T \circ R$ and $R \circ T$:

Translation First - $(\mathbf{R} \circ \mathbf{T})\left(\mathbf{I_a}, \mathbf{I_b}\right)$: This mapping first apply an image-to-image translation on I_a and then a spatial transformation on the translated image. Specifically, the final image is obtained by first applying T on I_a, which generates a fake sample $O_T = T(I_a)$. Then we apply our spatial transformation network R on O_T and get the final output:

$$O_{RT} = R_S\left(O_T, \phi\right) = R\left(T\left(I_a\right), R_\Phi\left(I_a, I_b\right)\right).$$

Register First - $(\mathbf{T} \circ \mathbf{R})\left(\mathbf{I_a}, \mathbf{I_b}\right)$ in this composition, we first apply spatial transformation on I_a and obtain a deformed image $O_R = R(I_a, \phi)$. Then, we translate O_R to domain $\mathcal{B}$ using our translation network T:

$$O_{TR} = T\left(R_S\left(I_a, \phi\right)\right) = T\left(R_S\left(I_a, R_\Phi\left(I_a, I_b\right)\right)\right).$$

Note that in both compositions (i.e., $T \circ R$ and $R \circ T$), the deformation field, used by the re-sampler R_S, is given by $R_\Phi\left(I_a, I_b\right)$. The only difference is in the source image from which we re-sample the deformed image.

To understand why this training scheme gives the desired property, note that the translation network T is fed geometrically different input images. Namely, these are the input image I_a (in the translation first variant) and the spatially transformed version of that image $R_S(I_a, R_\Phi(I_a, I_b))$ (in the registration first scheme). Thus, T is encouraged to be geometry preserving, since we expect similar behavior for different inputs. Additionally, note that T has limited capacity to perform explicit geometric transformation (as it isn't designated for this task). The registration network R, on the other hand, is designed to be exactly the opposite – it naturally supports geometric deformations, and struggles with stylistic and appearance-based alterations.

Throughout this section, we refer to O_{RT} and O_{TR} as the outputs of $R \circ T$ and $T \circ R$, respectively.

4.3. Training Losses

To train R and T to generate fake samples that are similar to those in domain $\mathcal{B}$, we use an $L1$-reconstruction loss:

$$\mathcal{L}_{recon}(T, R) = \left\| O_{RT} - I_b \right\|_1 + \left\| O_{TR} - I_b \right\|_1 \tag{4}$$

where minimizing the above implies that $T \circ R \approx R \circ T$.

We use conditional GAN (cGAN)[24] as our adversarial loss for training D, T and R. The objective of the adversarial network D is to discriminate between real and fake samples, while T and R are jointly trained to fool the discriminator. The cGAN loss for $T \circ R$ and $R \circ T$ is formulated below:

$$\begin{aligned}\mathcal{L}_{cGAN}(T, R, D) = &\mathbb{E}\left[\log\left(D\left(I_b, I_a\right)\right)\right] \\ &+ \mathbb{E}\left[\log(1 - D(O_{RT}, I_a))\right] \\ &+ \mathbb{E}\left[\log(1 - D(O_{TR}, I_a))\right],\end{aligned} \tag{5}$$

The total objective is given by:

$$\begin{aligned}\mathcal{L}(T, R) = &\arg\max_{D} \mathcal{L}_{cGAN}(T, D, R) \\ &+ \lambda_R \cdot \mathcal{L}_{recon}(T, R) + \lambda_S \cdot \mathcal{L}_{smooth}(R),\end{aligned} \tag{6}$$

where we are opt to find T^* and R^* such that $T^*, R^* = \arg\min_{R,T} \mathcal{L}(T, R)$. Furthermore, in our experiments, we set $\lambda_R = 100$ and $\lambda_S = 200$.

4.4. Implementation Details

Our code is implemented using PyTorch 1.1.0 [26] and is based on the framework and implementation of Pix2Pix [13], CycleGAN [38] and BiCycleGAN [39]. The network T is an encoder-decoder network with residual connections [1, 15] and the registration network is U-NET based [28] with residual connections in the encoder and output paths. In all residual connections, we use Instance Normalization Layer [35]. All networks were initialized by the Kaiming [11] initialization method.

The experiments were conducted on single GeForce RTX 2080 Ti. We use Adam Optimizer [17] on a mini-batch of size 12 with parameters $lr = 1 \times e^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train our model for 200 epochs, and activate linear learning rate decay after 100 epochs.

5. Experimental Results

In the following section we evaluate our approach and explore the interactions between R, T and the different loss terms we use.

All our experiments were conducted on a commercial dataset, which contains a collection of images of banana plants with different growing conditions and phenotype. The dataset contains 6100 image frames, where each frame consist of an RGB image, IR Image and Depth Image. The colored images are a 24bit Color Bitmap captured from a high-resolution sensor. The IR images are a 16bit gray-scale image, captured from a long-wave infrared (LWIR) sensor. Finally, the depth images were captured by Intel Real-Sense depth camera. The three sensors were calibrated, and an initial registration was applied based on affine transformation estimation via depth and controlled lab measurements. The misalignment in the dataset is due to depth variation within different objects in the scene, which the initial registration fails to handle. We split the dataset into training and test samples, where the test images were sampled with probability $p = 0.1$.

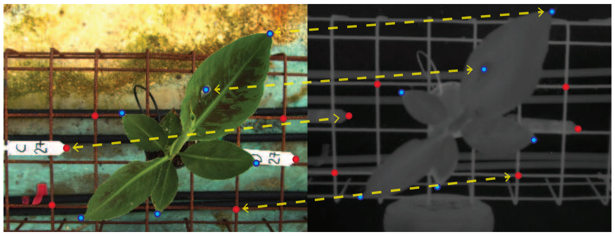


Figure 3: **Annotation sample.** We pick points from both the source image I_a (left) and the target image I_b (right). The blue points are on salient objects and the red points are general points from the scene. We added several arrows to illustrate some matching points. The geometry of each point is with respect to its corresponding image.

STN	SSIM on edges	NCC on edges	NCC	Ours
R_{Affine}	X / X	19.44 / 19.45	20.56 / 13.26	13.53 / 8.5
R_{TPS}	X / X	32.47 / 26.82	28.68 / 26.47	10.01 / 7.02
R	28.41 / 26.12	27.41 / 16.78	29.91 / 15.8	**6.93 / 6.27**

Table 1: **Registration accuracy of several similarity measures.** We report the average registration accuracy of different registration networks (i.e. R_{Affine}, R_{TPS} and ours R). In each table entry, we report two accuracies, one that is measured based on full scene annotation (left) and the other based on salient objects only annotation (right). X denotes cases where the network degenerates.

5.1. Evaluation

Registration Accuracy Metric. We manually annotated 100 random pairs of test images. We tagged 10-15 pairs of point landmarks on the source and target images which are notable and expected to match in the registration process (See Figure 3). Given a pair of test images, I_a and I_b, with a set of tagged pairs. The accuracy of the registration network R is simply the average Euclidean distance between the target points and their matching deformed source points.

Furthermore, we used two types of annotations. The first type of annotation is located over salient objects in the scene (the blue points in Figure 3). This is important because in most cases, down-stream tasks are affected mainly by the alignment of the main object in the scene in both modalities. The second annotation is performed by picking landmark points from all objects across the scene.

Quantitative Evaluation. As the crux of our work is the alleviation of the need for cross-modality similarity measures, we trained our network with commonly used loss terms. In Table 1 we show the registration accuracy of our registration network when trained with differ-

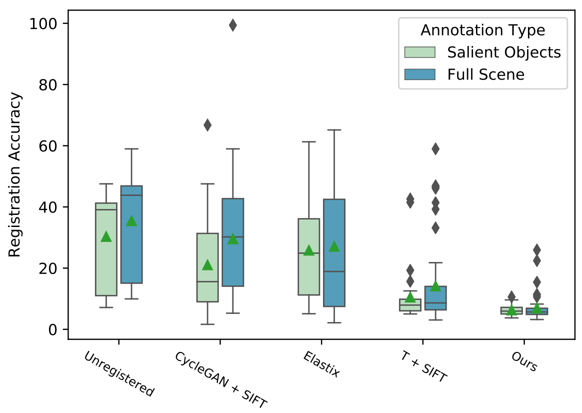


Figure 4: **Registration accuracy of different methods.** Unregistered indicates the misalignment in our dataset. We show the accuracy obtained with Elastix [18, 23] and a feature-based registration based on SIFT [20] and CycleGAN [38] or our translation network (i.e. T).

ent metrics. Specifically, we used the Normalized Cross-Correlation (NCC) metric as it is frequently used in unsupervised multi-modal registration methods. Additionally, we trained our network by maximizing similarity metrics (e.g. Structural Similarity Index Metric (SSIM) and NCC) on edges detected by Canny edge-detector [4] from both the deformed and target image. As can be seen from Table 1, training the registration network R using prescribed cross-modality similarity measures do not perform well. Further, using NCC produces noisy results, while using SSIM gives smooth but less accurate registration accuracy (see supplemental materials).

Also, since our learned metric is generic and can be incorporated with any spatial transformation, we trained different spatial transformations with our metric as well. Specifically, we trained an affine-based STN (R_{Affine}) and a thin-plate-spline based STN (R_{TPS}). The registration accuracy of these networks is reported in Table 1. As can be seen from the table, training these networks with our metric yields substantial improvement over other loss terms.

We also compared our method with three techniques. The first method we considered is SimpleElastix [18, 23], an iterative registration technique based on Mutual Information. The other two methods are feature-based techniques in which the SIFT [20] descriptor is adapted. However, since SIFT [20] is not designed for multi-modal data, then it cannot be directly used on the source and target images.

Instead, we train CycleGAN [38] network to translate between the two modalities at hand, without any supervision to match the ground truth. CycleGAN, like other unsupervised image-to-image translation networks, is not trained to generate images matching ground truth samples, thus, geometric transformation is not explicitly required from the translation network. Once trained, we use one of the generators in the CycleGAN, the one that maps domain $\mathcal{A}$ to domain $\mathcal{B}$, to translate the input image I_a onto modality $\mathcal{B}$. Assuming this generator is both geometry preserving and translates well between the modalities, it is expected that it also match well between feature of the fake sample and the target image. Thus, we extracted SIFT descriptors from the generated images by the CycleGAN translation network, and extracted SIFT features from the target image I_b. We then matched these features and estimated the needed spatial registration. The registration accuracy using this method is significantly better than directly using SIFT [20] features on the input image I_a. Similarly, we used our geometry preserving translation network T along with SIFT descriptor.

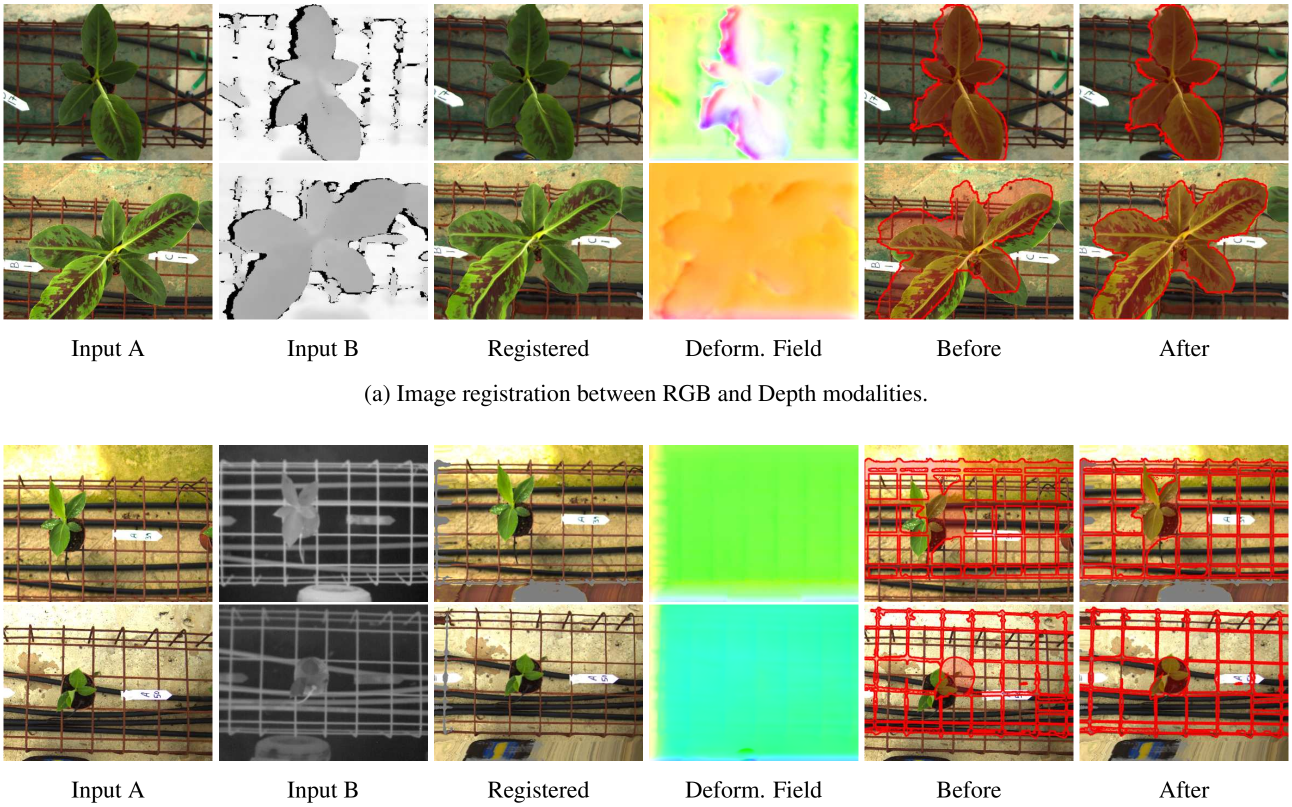
In Figure 4 we show the accuracy of the aforementioned methods. As can be seen, our method gives the best registration accuracy. Moreover, using our translation network T with the SIFT descriptor achieves substantial improvement over CycleGAN [38] + SIFT [20] and Elastix [23, 18]. Thus indicates that our translator T is both geometry preserving and performs accurate photo-metric mapping.

Qualitative Evaluation. Figure 5 shows that our registration network successfully aligns images from different pairs of modalities and handles different alignment cases. For example, the banana leaves in the first raw in Figure 5a are well-aligned in the two modalities. Our registration network maintains this alignment and only deforms the background for full alignment between the images. This can be seen from the deformation field visualization [2], where little deformation is applied to the banana plant, while most of the deformation is applied to the background. Furthermore, in the second row in Figure 5a, most of the image is translated in a certain direction due to the camera shift, but depth-dependent variation can still be seen. To help measuring the alignment success, we overlay (with semi-transparency) the plant in image B on top of both image A before and after the registration. This means that the silhouette has the same spatial location in all images (the original image B, image A before and after the registration). Lastly, we achieve similar success in the registration between RGB and IR images (see Figure 5b).

It is worth mentioning that in some cases, the deformation field points to regions outside the source image. In those cases, we simply sample zero values. This happens because the the target image content (i.e., I_b) in these regions is not available in the source image (i.e., I_a) (see supplemental materials for more results).

5.2. Ablation Study

Next, we present a series of ablation studies that analyze the effectiveness of different aspects in our work. First,



(a) Image registration between RGB and Depth modalities.

(b) Image registration between RGB and IR modalities.

Figure 5: **Qualitative Evaluation.** We show sample results on the registration between two pairs of domains; (a) RGB to Depth registration and (b) RGB to IR registration. In the first two columns we show the corresponding images I_a and I_b. The third column is the registered image, i.e the image I_a after deformation. The deformation field (4th column) is visualized using the standard optical-flow visualization [2]. Finally, we segment the salient object in I_b and overlay it (with opacity 25%) in the same spatial location onto the image before and after registration (last two columns).

we show that training both compositions (i.e our presented two training flows) of T and R indeed encourages a geometric preserving translator T. Additionally, we analyze the impact of the different loss terms on the registration network's accuracy. We further show the effectiveness of the bilateral filtering, and that it indeed improves the registration accuracy. All experiments, unless otherwise stated, were conducted without the bilateral filtering. **Geometric-Preserving Translation Network.** To evaluate the impact of training of T and R simultaneously with the two training flows proposed in Figure 2, we compare the registration accuracy of our method with that of training models with either $T \circ R$ or $R \circ T$. As can be seen from Figure 6, training both combinations yields a substantial improvement in the registration accuracy (shown in blue), compared to each training flow (i.e., $T \circ R$ and $R \circ T$) separately. Moreover, while the reconstruction loss of $T \circ R$ (shown in read) is lowest among the three options, it does not necessarily indicate a better registration. This is because in this setting the translation network T implicitly performs both the alignment and translation tasks. Conversely, when training with $R \circ T$ only (shown in green), the network R is unstable and at some point it starts to alternate pixel values, essentially taking on the role of a translation network. Since R is only geometry-aware by design it fails to generate good samples. This is indicated by how fast the discriminator detects that the generated samples are fake (i.e., the adversarial loss decays fast). Visual results are provided in the supplementary materials.

Loss ablation. It has been shown in previous works [39, 13, 38] that training an image-to-image translation network with both a reconstruction and an adversarial loss yields better results. In particular, the reconstruction loss stabilizes the training process and improves the vividness of the output images, while the adversarial loss encourages the generation of samples matching the real-data distribution.

The main objective of our work is the production of a registration network. Therefore, we seek to understand the

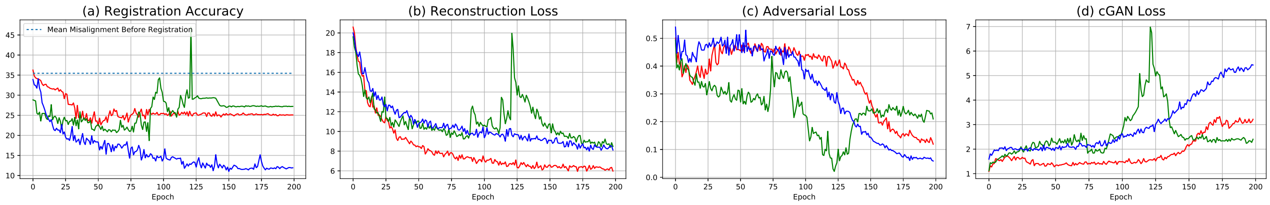


Figure 6: **Composition Ablation Study**. We show the values of the (a) Registration Accuracy, (b) Reconstruction loss, (c) Adversarial Loss and (d) cGAN Loss. The x-axis in all figures is the epoch number. The loss values are shown for $T \circ R$ (red), $R \circ T$ (green) and ours (blue). As can be seen, the registration accuracy is best using our method. In $T \circ R$, the reconstruction loss is the lowest, however, the registration is inaccurate because a significant portion of the registration task is implicitly performed by the translator T. Further, the composition $R \circ T$ is unstable because at some point, the registration network R starts alternating pixels values, which is detected by the discriminator (see the dip in (c)).

GAN loss L1 loss	R	T	Both
R	-	X	28.15
T	29.02	-	22.03
Both	X	X	11.01

Table 2: **Loss ablation results.** Columns denote modules trained with a GAN loss term. Rows denote modules trained with an L1 loss term. We do not report results where only one module is not trained with any loss terms. X denotes cases where the training diverges. For example, the result in the second row and first column represents the registration accuracy achieved when module R's weights are updated with respect to the cGAN loss and module T with respect to the reconstruction loss term.

Method	Test Acc.	Train Acc.
No Registration	35.45	34.96
W/O Bilateral	11.01	9.89
With Bilateral	**6.93**	**6.12**

Table 3: **Smoothness Regularization**. Effect of bilateral filtering on registration accuracy. We show the registration accuracy on annotated test samples, and annotated train samples.

impact of both losses (reconstruction and adversarial) on the registration network. To understand the impact of each loss, we train our model with different settings: each time we fix either R or T's weights with respect to one of the loss functions. The registration accuracy is presented in Table 2. Please refer to the supplementary material for qualitative results. As can be seen in these figures, training R only with respect to the reconstruction loss leads to overly sharp, but unrealistic images where the deformation field creates noisy artifacts. On the other hand, training R only with respect to the adversarial loss creates realistic images, but with inexact alignment. This is especially evident in Table 2 where training R with respect to the reconstruction loss achieves a significant improvement in the alignment, and the best accuracy is obtained when the loss terms are both used to update all the networks weights.

Bilateral Filtering Effectiveness Using bilateral filtering to weigh the smoothness loss allows us, in effect, to encourage piece-wise smoothness on the deformation map. As can be seen in Table 3, this enhances the precision of the registration. These results suggest that using segmentation maps for controlling the smoothness loss term could be beneficial.

6. Summary and Conclusions

We presented an unsupervised multi-modal image registration technique based on image-to-image translation network. Specifically, we developed a geometry preserving image-to-image translation network which allows comparing the deformed and target image using simple mono-modality metrics. The geometric preserving translation network was made possible by a novel training scheme, which alternates and combines two different flows to train the spatial transformation.

We believe that geometric preserving generators can be useful for applications other than image registration. In the future, we would like to continue to explore the idea of alternate training several layers or operators, in different flows, to encourage them being commutative as means to achieve certain non-trivial properties.

Acknowledgments

This research was supported by a generic RD program of the Israeli innovation authority, and the Phenomics consortium.

References

[1] Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[2] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vision*, 92(1):1–31, Mar. 2011. 6, 7

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 2

[4] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986. 6

[5] Xiaohuan Cao, Jianhuan Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen. Deep learning based intermodality image registration supervised by intra-modality similarity. *Machine learning in medical imaging. MLMI*, 11046:55–63, 2018. 2

[6] Adrian V. Dalca, Guha Balakrishnan, John V. Guttag, and Mert R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *MICCAI*, 2018. 2

[7] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Isgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer F. Syeda-Mahmood, João Manuel R. S. Tavares, Mehdi Moradi, Andrew P. Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings*, volume 10553 of *Lecture Notes in Computer Science*, pages 204–212. Springer, 2017. 1

[8] Jingfan Fan, Xiaohuan Cao, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Adversarial learning for mono- or multimodal registration. *Medical Image Analysis*, 58:101545, 2019. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2, 3

[10] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: A survey, 2019. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. 5

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 5, 7

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2017–2025, 2015. 1

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 5

[16] Boah Kim, Jieun Kim, June-Goo Lee, Dong Hwan Kim, Seong Ho Park, and Jong Chul Ye. Unsupervised deformable image registration using cycle-consistent CNN. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI*, volume 11769 of *Lecture Notes in Computer Science*, pages 166–174. Springer, 2019. 1

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 5

[18] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P. W. Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging*, 29(1):196–205, 2010. 6

[19] Matthew C. H. Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-spatial transformer networks for structure-guided image registration. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II*, volume 11765 of *Lecture Notes in Computer Science*, pages 337–345. Springer, 2019. 1

[20] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 6

[21] Dwarikanath Mahapatra, Bhavna J. Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*, pages 1449–1453. IEEE, 2018. 2

[22] Dwarikanath Mahapatra, Zongyuan Ge, Suman Sedai, and Rajib Chakravorty. Joint registration and segmentation of xray images using generative adversarial networks. In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *Machine Learning in Medical Imaging*, volume 11046 of *Lecture Notes in Computer Science*, pages 73–80. Springer, 1 2018. 2

[23] Kasper Marstal, Floris F. Berendsen, Marius Staring, and Stefan Klein. Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016*, pages 574–582. IEEE Computer Society, 2016. 6

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2, 3, 4

[25] Seungjong Oh and Siyong Kim. Deformable image registration in radiation therapy. *Radiation oncology journal*, 35(2):101, 2017. 1

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 5

[27] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging - 26th International Conference, IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings*, volume 11492 of *Lecture Notes in Computer Science*, pages 249–261. Springer, 2019. 2

[28] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 5

[29] Maria A Schmidt and Geoffrey S Payne. Radiotherapy planning using mri. *Physics in Medicine & Biology*, 60(22):R323, 2015. 1

[30] N. Schneider, F. Piewak, C. Stiller, and U. Franke. Regnet: Multimodal sensor registration using deep neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1803–1810, June 2017. 1

[31] A. Sheikhjafari, Michelle Noga, Kumaradevan Punithakumar, and Nilanjan Ray. Unsupervised deformable image registration with fully connected generative neural network. 2018. 2

[32] Xiaoyong Shen, Li Xu, Qi Zhang, and Jiaya Jia. Multimodal and multi-spectral registration for natural images. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pages 309–324, 2014. 1

[33] Chang Shu, Xi Chen, Qiwei Xie, and Hua Han. An unsupervised network for fast microscopic image registration. In John E. Tomaszewski and Metin N. Gurcan, editors, *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 363 – 370. International Society for Optics and Photonics, SPIE, 2018. 2

[34] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 839–, Washington, DC, USA, 1998. IEEE Computer Society. 4

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016. 5

[36] Chengjia Wang, Giorgos Papanastasiou, Agisilaos Chartsias, Grzegorz Jacenkow, Sotirios A. Tsaftaris, and Heye Zhang. FIRE: unsupervised bi-directional inter-modality registration using deep networks. *CoRR*, abs/1907.05062, 2019. 1, 2

[37] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *ECCV*, 2018. 1

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 5, 6, 7

[39] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 2, 5, 7

[40] Barbara Zitov and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003. 1