**KeypointMorph: Robust Multi-modal Affine Registration via Unsupervised Keypoint Detection**

Evan M. Yu
Alan Q. Wang
Adrian V. Dalca
Mert R. Sabuncu

1 Nancy E. and Peter C. Meinig School of Biomedical Engineering, Cornell University
2 School of Electrical and Computer Engineering, Cornell University and Cornell Tech
3 Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School
4 Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT
5 Department of Radiology, Weill Cornell Medical

Editors: Under Review for MIDL 2022

**Abstract**

Registration is a fundamental task in medical imaging, and recent machine learning methods have become the state-of-the-art. However, these approaches are often not interpretable, lack robustness to large misalignments, and do not incorporate symmetries of the problem. In this work, we propose KeypointMorph, an unsupervised end-to-end learning-based image registration framework that relies on automatically detecting corresponding keypoints. Our core insight is straightforward: matching keypoints between images can be used to obtain the optimal transformation via a differentiable closed-form expression. We use this observation to drive the unsupervised learning of anatomically-consistent keypoints from images. This not only leads to substantially more robust registration but also yields better interpretability, since the keypoints reveal which parts of the image are driving the final alignment. Moreover, KeypointMorph can be designed to be equivariant under image translations and/or symmetric with respect to the input image ordering. We demonstrate the proposed framework in solving 3D affine registration of multi-modal brain MRI scans. Remarkably, we show that this strategy leads to consistent keypoints, even across modalities. We demonstrate registration accuracy that surpasses current state-of-the-art methods, especially in the context of large displacements. Our code is available at [URL](#)

**Keywords:** Image registration, Multi-modal, Keypoint detection, Unsupervised Learning

1. Introduction

Registration is a core problem in biomedical imaging applications. Multiple images, often encompassing a variety of contrasts, are commonly acquired (Uludağ and Roebroeck, 2014). Classical (i.e. non-learning-based) registration methods involve an iterative optimization of a similarity metric over a space of transformations (Oliveira and Tavares, 2014; Sotiras et al., 2013). Recent deep learning-based strategies leverage large datasets of images to solve registration. Given a pair of images \( \mathbf{x}_f, \mathbf{x}_m \), these strategies use neural network architectures that either output transformation parameters (e.g. affine or spline) (Lee et al.,

* Joint senior author
1. Code will be made available after paper acceptance

Figure 1: Proposed framework. Fixed and moving 3D images are passed through the same keypoint detection network composed of convolutional layers and a final center-of-mass (CoM) layer that predicts keypoints useful for registration. The affine matrix is then computed using Eq. 2 and is used to transform the moving image.

2019a; de Vos et al., 2019) or a dense deformation field (Balakrishnan et al., 2019) which aligns them. Since the registration step is achieved via a single feed-forward pass, it is substantially faster than iterative methods.

However, prior learning-based methods often fail when the given image pair has a large misalignment. Existing systems typically require that image pairs are roughly aligned (Balakrishnan et al., 2019; Dalca et al., 2019; de Vos et al., 2019; Qin et al., 2019), or at least in the same orientation (Mocanu et al., 2021). In addition, today’s learning-based registration methods lack interpretability, as they are essentially “black-box models” that output either transformation parameters or a deformation field and provide little insight into what drives the alignment. Finally, the machine learning models used in registration tasks generally do not exploit the symmetries and equivariances present in the problem. For example, if one of the images is translated by a fixed amount, this has a pre-determined effect on the optimal registration solution, and this property is not built into any of the architectures used today for image registration.

**Contribution.** We propose KeypointMorph, an unsupervised end-to-end deep-learning-based framework that aims to address all aforementioned issues. Our main insight is that matched keypoints can be used to derive the optimal transformation in closed-form. This keypoint-based formulation is robust against misalignments as the closed-form solution does not depend on the initial position of the keypoints. Additionally, the model is interpretable since the keypoints that drive the alignment can be visualized. Finally, we also show how to incorporate symmetries into the model design. For example, the architecture we describe in this paper is translation equivariant and leverages a recently-proposed center-of-mass layer (Ma et al., 2020; Sofka et al., 2017). Rather than treating matched keypoint detection as a supervised learning problem requiring human-annotated keypoints, we propose to use an end-to-end unsupervised strategy tailored toward registration. By unifying image registration and keypoint detection, we can train a model that finds matching keypoints useful
for aligning images. We demonstrate this framework in the context of affine registration of 3D multi-modal brain MR scans.

2. Related Works

**Classical Image Registration Methods.** Pairwise iterative, optimization-based approaches have been extensively studied in medical image registration (Hill et al., 2001; Oliveira and Tavares, 2014). These methods employ a variety of similarity functions, types of deformation, transformation constraints or regularization strategies, and optimization techniques. Intensity-based similarity criteria are most often used, such as mean-squared error (MSE) or normalized cross correlation for registering images of the same modality (Avants et al., 2009, 2008; Hermosillo et al., 2002). For registering image pairs from different modalities, statistical measures like mutual information or contrast-invariant features like MIND are popular (Heinrich et al., 2012; Hermosillo et al., 2002; Hoffmann et al., 2020; Mattes et al., 2003; Viola and Wells III, 1997).

Another registration paradigm first detects features or keypoints in the images, and then establishes their correspondence. This approach often involves handcrafted features (Tuytellaars and Mikolajczyk, 2008), features extracted from curvature of contours (Rosenfeld and Thurston, 1971), image intensity (Förstner and Gülich, 1987; Harris et al., 1988), color information (Montesinos et al., 1998; Van de Weijer et al., 2005), or segmented regions (Matas et al., 2004; Wachinger et al., 2018). Features can be also obtained so that they are invariant to viewpoints (Bay et al., 2006; Brown et al., 2005; Lowe, 2004; Toews et al., 2013). These algorithms then optimize similarity functions based on these features over the space of transformations (Chui and Rangarajan, 2003; Hill et al., 2001). This strategy is sensitive to the quality of the keypoints and often suffer in the presence of substantial contrast and/or color variation (Verdie et al., 2015).

**Learning-based Methods.** In learning-based image registration, supervision can be provided through ground-truth transformations, either synthesized or computed by classical methods (Cao et al., 2018; Dosovitskiy et al., 2015; Epplenhofer and Pluim, 2018; Lee et al., 2019b; Uzunova et al., 2017; Yang et al., 2017). Unsupervised strategies use loss functions similar to those employed in classical methods (Balakrishnan et al., 2019; Dalca et al., 2019; de Vos et al., 2019; Fan et al., 2018; Krebs et al., 2019; Qin et al., 2019; Wu et al., 2015). Weakly supervised models employ (additional) landmarks or labels to guide training (Balakrishnan et al., 2019; Fan et al., 2019; Hu et al., 2018a,b).

Recent learning-based methods compute image features or keypoints (Ma et al., 2021) that can be used for image recognition, retrieval, or registration. Learning the keypoints can be done with supervision (Verdie et al., 2015; Yi et al., 2016), self-supervision (DeTone et al., 2018) or without supervision (Barroso-Laguna et al., 2019; Lenc and Vedaldi, 2016; Ono et al., 2018). In contrast, our focus is on robust image registration via corresponding keypoints; these prior works aim explicitly to obtain keypoints that are repeatable under different viewpoints and/or image acquisition conditions. Nevertheless, we build on previous ideas and introduce a framework that output matched keypoints in 3D space regardless of the initial position of the image.
3. Proposed Method

Let $x_m$ and $x_f$ be moving (source) and fixed (target) volumes, which may vary in modality and orientation. In this paper, we focus on 3D affine transformations, where the goal is to find the optimal affine transformation matrix $A^* \in \mathbb{R}^{3 \times 4}$ such that the moved (registered) image $x_r = x_m \circ A^*$ matches the fixed image $x_f$, where $\circ$ denotes the spatial transformation of an image. To efficiently compute $A^*$, we employ corresponding keypoints from $x_m$ and $x_f$. The matching keypoints are computed using a convolutional neural network, and $A^*$ is estimated using a non-learnable computation layer.

3.1. Keypoint Detector Network

To compute $K$ matching keypoints, we use a single neural network $g_\phi$ with parameters $\phi$ to produce $p_m = g_\phi(x_m)$ and $p_f = g_\phi(x_f)$, where $p_m$ and $p_f$ are matrices of shape $d \times K$ (i.e. each column is a corresponding keypoint pair of $d$ dimensions).

In our implementation, the backbone architecture of the keypoint detector consists of convolutional layers, followed by instance normalization (Ulyanov et al., 2016), ReLU activation, and 2x downsampling via strided convolution, as shown in Fig. 1. The output from the backbone is followed by a center-of-mass (CoM) layer (Ma et al., 2020; Sofka et al., 2017), which computes the center-of-mass for each of the $K$ activation maps. This specialized layer is (approximately) translation equivariant and enables precise localization. We provide more details and compare CoM to fully connected layers in Appendix B.

3.2. Affine Computation Layer

Given $K$ corresponding keypoint pairs, we can derive a differentiable closed-form expression for an affine transformation that aligns the keypoints. Let $\tilde{p}_m = [p_m \ 1]^T \in \mathbb{R}^{(d+1) \times K}$, where $1 \in \mathbb{R}^{1 \times K}$ is a vector of ones and $K > d$. We aim to find the optimal affine transformation

$$A^* = \arg \min_A \| A\tilde{p}_m - p_f \|_F,$$

where $\| \cdot \|_F$ denotes the Frobenius norm. This leads to the closed-form solution

$$A^* = A(p_f, p_m) = p_f \tilde{p}_m^T (\tilde{p}_m \tilde{p}_m^T)^{-1}.$$

We provide the derivation in the Appendix A.

3.3. Training

We found the following self-supervised pre-training strategy to be effective for initializing the keypoint detector backbone. We first pick a set of initial keypoints $p$ chosen uniformly at random over the image coordinate grid. For a single subject, we assume that we have access to aligned multi-modal (e.g. T1, T2, PD-weighted MRI) volumes $\{x_i\}$. During

---

2. Although we consider 3D volumes in this work, we stress that our method is agnostic to the number of dimensions. The terms “image” and “volume” are used interchangeably.
pre-training, in each mini-batch, we apply random affine transformations $T$ (drawn from a uniform distribution over the parameter space) to $x_i$ and $p$, and minimize:

$$\arg \min_{\phi} \sum_i \mathbb{E}_T \| Tp - g_{\phi}(x_i \circ T) \|^2,$$

where $\mathbb{E}$ denotes expectation and $Tp$ transforms the list of coordinates in $p$ by the affine transformation $T$.

Following the pre-training strategy, we train KeypointMorph on random image pairs using the entire training dataset. We consider the typical real-world scenario of multi-parametric MRI, where scans of potentially different contrasts (e.g. T1, T2, PD-weighted) need to be registered. We present only within-modality image pairs to the network during training, where a simple loss like MSE may be used. Thus, given a dataset $D$ composed of same-modality image pairs $(x_f, x_m)$, the overall objective is:

$$\arg \min_{\phi} \mathbb{E}_{(x_f, x_m) \sim D} \| x_m \circ A(g_{\phi}(x_f), g_{\phi}(x_m)) - x_f \|^2.$$

During training, we apply random affine transformations to the images as an augmentation strategy. Remarkably, we found that KeypointMorph learns to detect anatomically consistent keypoints across modalities, even though it is trained on same-modality pairs. In our experiments, we demonstrate how this can be used to perform multi-modal registration at test-time.

### 3.4. KeypointMorph Variants

We use KeypointMorph$_{mse}$ to refer to the main unsupervised variant of our model, trained with Equation 4. In addition, we employ a supervised variant, KeypointMorph$_{dice}$, that exploits segmentations during training. During pre-training of this variant, we use the center-of-mass of segmentation labels as ground truth keypoints and do not restrict to a single subject. During subsequent end-to-end training, we use soft-Dice, a loss function that measures volume overlap of the moving and registered label maps (Balakrishnan et al., 2019; Hoffmann et al., 2020; Hu et al., 2018a).

### 4. Experiments

#### 4.1. Dataset

We used the IXI brain MRI dataset\(^3\) for evaluation. Each subject has T1, T2, and PD-weighted 3D MRI scans in spatial alignment. We partitioned the 577 total subjects into sets of 427, 50, and 100 for training, validation, and testing, respectively. We performed standard skull stripping (Kleesiek et al., 2016) on all images.

We used a pre-trained and validated SynthSeg model (Billot et al., 2020) to automatically delineate 23 regions of interest (ROIs)\(^4\). We used these segmentations for training a subset of models, as described below. Furthermore, all performance evaluations were based on examining the overlap of ROIs in the test images.

---

4. ROIs were pallidum, amygdala, caudate, cerebral cortex, hippocampus, thalamus, putamen, white matter, cerebellar cortex, ventricle, cerebral white matter, and brainstem.
Figure 2: Registration results. The x-axis of the second column shows the average absolute displacement in the moving volume after rotation, scaling, or translation. The Dice score is averaged for all test subjects and brain anatomical regions. See Section 4.3 for details on the naming scheme.

4.2. Test-time Performance Evaluation

We used each testing subject as a moving volume $x_m$, paired with a different test volume treated as fixed image $x_f$. For all test volumes, we use the 23-label segmentation maps to quantify alignment. We simulated different degrees of misalignment by transforming $x_m$ using rotation, scaling, or translation. For a given transformation type, we choose 1-3 axes randomly and apply a uniform random transformation (e.g. rotation) up to a given amount (e.g. degree). We use the predicted transformation to resample the moved segmentation labels on the fixed image grid, and compute the Dice score to quantify alignment quality.

We performed registration across all combinations of available modalities (registering T1 to T1, T1 to T2, etc). All pairings and amount of transformations was kept the same across the registration experiments.

4.3. Baselines

Advanced Normalizing Tools (ANTs) is a widely used software package for medical image registration (Avants et al., 2009). We use the “TRSAA” affine implementation, which consist of translation, rigid, similarity and two affine transformation steps. The volumes are registered successively at three different resolutions: 0.25x, 0.5x and finally at full resolution. At 0.25x and 0.5x resolution, Gaussian smoothing with $\sigma$ of two and one voxels is applied, respectively. In addition, we use the ANTs affine initializer, which conducts a grid search over a range of rotations and translations to find a good initialization. We used mutual information as the similarity metric, which is suitable for registering images with different contrasts.

Deep Learning for Image Registration (DLIR) is a recent learning-based method that includes affine image registration (de Vos et al., 2019). DLIR implements a Spatial
KeypointMorph

Figure 3: (a) Sample keypoints learned without supervision by KeypointMorph. Each row shows a different keypoint and each column shows a different subject and/or modality. (b) Mean registration performance for rotation, scaling and translation under different number of keypoints.

Transformer Network (Jaderberg et al., 2015), which was originally used to improve class prediction accuracy and has since become the backbone of many subsequent works in image registration (de Vos et al., 2017; Lee et al., 2019a; Balakrishnan et al., 2019). For a direct comparison, we used the same backbone architecture as KeypointMorph replacing the center-of-mass layer with a fully-connected (FC) layer which outputs 12 parameters for the 3D affine transformation.\(^5\) In Appendix B, we also investigated a KeypointMorph implementation with the same FC layer as DLIR.

We find that DLIR often cannot register image pairs with large misalignments. We alleviate this by using more aggressive augmentation during training, where the images are randomly transformed. We consider two different amounts of rotation for the training of DLIR: maximum ±35° or ±180°. We use the same loss function and training scheme as we used for KeypointMorph. We trained separate DLIR models for each modality as it produces better results than training DLIR across modalities with mutual information. We also trained supervised modality-specific and multi-modal DLIR models using a soft-Dice loss computed on the aligned segmentation maps (Lee et al., 2019b).

Altogether, we implement six different DLIR variants. The naming scheme follows the convention \(<\text{mod}>\text{DLIR}_<\text{loss}>_<\text{degree}>\), where \(<\text{mod}>\) denotes whether the model was trained on a single (uni) or multiple (multi) modalities, \(<\text{loss}>\) denotes the training loss function, and \(<\text{degree}>\) denotes the maximum angle of rotation for augmentation.

5. Results

Fig. 2 and Table A.2 summarize the results. We find that all DLIR models suffer substantially as the rotation angle increases. Training with aggressive augmentation increases performance for test pairs with large misalignments, but reduces the accuracy for those with smaller misalignments. Using supervision (uni\text{DLIR}_\text{dice35}) leads to improved accuracy. For unimodal registration, the DLIR model that was trained with all modalities

\[^{5}\text{We used instance Norms for multiDLIR and Batch Norms for uniDLIR, which we found to work well in practice.}\]
(multiDLIR_dice) did not produce better accuracy than a model that was trained with each modality separately. ANTs yields excellent results when the initial misalignment is small (e.g., less than 45 degrees rotation). However, the accuracy drops substantially when the misalignment exceeds this range.

In contrast to these models, KeypointMorph variants performed well across all types of transformations and ranges, with only marginal drops in accuracy in large misalignments. In the case where we have access to ROIs during training, we find that KeypointMorph trained with Dice outperforms all models in nearly all the tasks. However, the unsupervised variant trained with MSE still yields excellent accuracy across all settings and is only minimally suboptimal compared to its supervised counterpart. We provide qualitative results in Appendix D, and compare the computational time across different models in Appendix C. Overall, the KeypointMorph variants substantially outperform other learning-based baselines, and KeypointMorph performs comparably or often substantially better (at large misalignments) than state-of-the-art ANTs registration, while requiring substantially less runtime.

5.1. Keypoint Analysis

Number of Keypoints. We trained five unsupervised KeypointMorph model variants with 16, 32, 64, and 128 keypoints, respectively. Fig. 3b illustrates that increasing the number of keypoints leads to improved Dice scores (and lower variability), up to 64 keypoints. However, we find that further increasing the keypoints can lead to a drop in performance (at 128 keypoints). We hypothesize that it is harder to find a relatively large number of anatomical landmarks that are consistent across individuals.

Keypoint Visualization. In contrast to existing models that compute the transformation parameters using a “black-box” neural network, we can investigate the keypoints that KeypointMorph learns to drive the alignment. Fig. 3a illustrates a set of representative keypoint examples for a learned unsupervised KeypointMorph model. We find that the final learned keypoints correspond to the same anatomical region in different subjects and modalities. In appendix F, we provide a quantitative study on keypoint consistency across the scans of each subject, and in Appendix G we show an expanded version of Fig. 3a.

6. Conclusion

We introduce KeypointMorph, a robust deep learning-based affine registration framework that employs unsupervised keypoint extraction. Our key insight is that matched keypoints yield a closed-form solution for affine registration, even in the case of large misalignments, and this in turn can be used to drive unsupervised keypoint detection. We showed that state-of-the-art optimization and learning-based methods for image registration struggle to register image pairs that have large misalignment. In contrast to many “black-box” machine learning-based registration methods, KeypointMorph also offers the ability to investigate what drives the registration by visualizing the keypoints. We envision that KeypointMorph can be used in a variety of applications that exhibit large misalignments, and can be extended to compute non-linear deformations.
References

Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric
diffeomorphic image registration with cross-correlation: evaluating automated labeling of

Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants).

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca.
Voxelmorp: a learning framework for deformable medical image registration. *IEEE

Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Key-
point detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF


Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias,


Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised


Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and


KeypointMorph


Appendix A. Derivation of Closed-form Expression

Let \( \mathbf{p}_f \in \mathbb{R}^{d \times K} \) and \( \mathbf{p}_m = [\mathbf{p}_m, \mathbf{1}]^T \in \mathbb{R}^{(d+1) \times K} \), where \( \mathbf{1} \in \mathbb{R}^{1 \times K} \) is a vector of ones and \( K > d \). We wish to find the optimal affine transformation \( A \in \mathbb{R}^{d \times (d+1)} \) which minimizes:

\[
\mathcal{L} = \|A\mathbf{p}_m - \mathbf{p}_f\|_F,
\]

where \( \|\cdot\|_F \) denotes the Frobenius norm. Taking the derivative with respect to \( A \) and setting the result to zero, we obtain:

\[
\frac{\partial \mathcal{L}}{\partial A} = (A\mathbf{p}_m - \mathbf{p}_f)^T\mathbf{p}_m = 0
\]

\[
\implies A\mathbf{p}_m\mathbf{p}_m^T = \mathbf{p}_f\mathbf{p}_m^T
\]

\[
\implies A = \mathbf{p}_f\mathbf{p}_m^T(\mathbf{p}_m\mathbf{p}_m^T)^{-1}.
\]

Appendix B. Center-of-Mass Layer vs Fully Connected Layer

The backbone of our architecture is composed of CNN layers and a center-of-mass (CoM) layer at the end of the network (Ma et al., 2020; Sofka et al., 2017). The CoM layer computes the center-of-mass of the activation map for each channel of the CNN output. In other words, weighted average between the voxel values and the grid coordinates. These centers-of-masses are then used as keypoints in KeypointMorph. CoM layer is shift equivariant as compared to fully connected (FC) layer. As an ablation study, we compared the performance of using a CoM layer and FC layer, which is commonly used in registration (de Vos et al., 2019; Jaderberg et al., 2015; de Vos et al., 2017; Lee et al., 2019a). In Fig. A.1, the model using FC layers follows the same architecture as DLIR model that was used in the baseline. However, instead of outputting 12 affine parameters, it outputs 64 keypoints. We repeated the experiments found in Section 4. We can see that regardless of the type of layer used to compute the keypoints, KeypointMorph provide robustness to large deformation. Models that uses CoM have comparable performance at low deformation compared to models that uses FC layers. However, CoM models provide the best performance at higher degrees of misalignment.
Figure A.1: Performance comparison between KeypointMorph models that use center-of-mass (CoM) and fully connected (FC) layers to predict keypoints. The suffix `mse` and `dice` represent the unsupervised and supervised version of KeypointMorph, respectively.

Appendix C. Computation Time

Table A.1 summarizes runtime at inference of all methods. On a modern CPU, the ANTs baseline, which does not have GPU support, requires more than 60 seconds, whereas KeypointMorph requires roughly 2.7 seconds per subject at test time (about 0.2 sec on GPU). Comparing ANTs and KeypointMorph (CPU) and KeypointMorph (GPU), this represent more than 20x and 300x speed-up, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>CPU Time (s)</th>
<th>GPU Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANTs</td>
<td>66.95±1.56</td>
<td>-</td>
</tr>
<tr>
<td>DLIR</td>
<td>1.49±0.09</td>
<td>0.02±0.001</td>
</tr>
<tr>
<td>KeypointMorph</td>
<td>2.68±0.44</td>
<td>0.21±0.012</td>
</tr>
</tbody>
</table>

Table A.1: Average computation time across different models
Appendix D. Qualitative Results

We present some qualitative results of KeypointMorph trained without supervision. Moving and fixed subjects were picked randomly from the test set. We introduced random affine transformation with $\pm 180$ degrees of rotation, $\pm 20\%$ of scaling, and $\pm 25$ voxels of translation to the moving image. Fig. A.2 presents the registration result of the moved brains in the third column. We overlayed the aligned image (green) to the fixed/target brain (red).

Appendix E. Quantitative of Results

This section provides more details on the quantitative results for Fig. 2 from the main paper. We randomly picked a different fixed volume $x_f$ on the test set for each moving subject. We introduce random amount of misalignment to $x_m$. Each of the 3 axes of $x_m$ had 50% chance of being deformed with a given transformation with at least one axis being perturbed. Since each subject in the IXI dataset has a corresponding T1, T2 and PD weighted MRI scan, we performed registration across all combination of modalities (e.g. registering T1 to T1, T1 to T2 etc..). All pairings and random transformations were kept the same across the registration experiments.
investigated which regions KeypointMorph uses to register the volume pairs. We picked 12
In this section, we provide an extended visualization of keypoints. In the figures below, we
Appendix G. Keypoint Visualization
subject, allowing KeypointMorph to perform accurate multimodal registration.
In other words, the keypoints lie in almost the same location across modalities for a given
different modalities in an unsupervised manner. In Fig. A.3, we plotted the mean absolute error between the keypoints of each test subject across different modalities, as a function of training iteration. We can observe that as the model trains, the keypoints become more consistent across modalities. In other words, the keypoints lie in almost the same location across modalities for a given subject, allowing KeypointMorph to perform accurate multimodal registration.

Appendix F. Keypoint Consistency
Keypoint consistency plays an important role for multimodal registration. In this section, we show that KeypointMorph learns consistent keypoints across different modalities in an unsupervised manner. In Fig. A.3, we plotted the mean absolute error between the keypoints of each test subject across different modalities, as a function of training iteration. We can observe that as the model trains, the keypoints become more consistent across modalities. In other words, the keypoints lie in almost the same location across modalities for a given subject, allowing KeypointMorph to perform accurate multimodal registration.

Appendix G. Keypoint Visualization
In this section, we provide an extended visualization of keypoints. In the figures below, we investigated which regions KeypointMorph uses to register the volume pairs. We picked 12 random subjects and showed 12 of the 64 learned keypoints in each row. These keypoints were learned without supervision of any labeled regions. We can observe that the final keypoints lie within similar region of the brain across different subjects and scans.
Figure A.3: Mean absolute error between keypoint locations of test subject across different modalities. Each training iteration represent a model update after 32 training subject.
Figure A.4: Keypoints for different T1 scans
Figure A.5: Keypoints for different T2 scans
Figure A.6: Keypoints for different PD scans
Figure A.7: Keypoints for different multimodal scans