Magnetic Resonance Imaging 64 (2019) 142–153

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.elsevier.com/locate/mri

Review article

Learning image-based spatial transformations via convolutional neural networks: A review

Nicholas J. Tustison[a,*], Brian B. Avants[a], James C. Gee[b]

[a] Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA, United States of America
[b] Department of Radiology, University of Pennsylvania, Philadelphia, PA, United States of America

ARTICLE INFO

Keywords:
ConvNets
Deep learning
Diffeomorphisms
Image registration
Spatial normalization

ABSTRACT

Recent methodological innovations in deep learning and associated advancements in computational hardware have significantly impacted the various core subfields of quantitative medical image analysis. The generalizability, computational efficiency and open-source availability of deep learning algorithms and related software, particularly those utilizing convolutional neural networks, have produced paradigm shifts within the field. This impact is evident from topical prevalence in the literature, conference and workshop themes and winning methodologies in relevant competitions. In this work, we review the various state-of-the-art approaches to learning and prediction and/or optimizing image transformations using convolutional neural networks. Although of primary importance within the quantitative imaging domain, image registration algorithmic development, in the context of these deep learning strategies, has received comparatively less attention than its counterparts (e.g., image segmentation). Nevertheless, significant progress has been made in this particular subfield which has been presented in various research venues. We contextualize these contributions within the broader scope of deep learning advancements and, in so doing, attempt to facilitate the leveraging and further development of such techniques within the medical imaging research community.

1. Introduction

Determining the spatial correspondence between imaging domains is frequently a critical component in quantitative image analysis workflows. The trajectory of image registration theoretical and technological development has led to increasingly high quality transformational mappings that have significantly improved performance in related processing tasks (e.g., image segmentation via joint label fusion [1]) and imaging-based statistical analysis involving template-based normalization (e.g., voxel-based morphometry [2] and sparse canonical correlation analysis [3]). Several reviews [4-10] have charted this chronology and provided insight into related issues such as algorithmic classification, available implementations, evaluation strategies and speculation concerning possible future directions of the field. While prescient in many respects, such speculation vis-à-vis the resurgence of deep learning is understandably limited due to its recent explosion in popularity and research focus.

The foundational concepts that form the basis for contemporary deep learning research dates back decades (e.g., [11]). Since this early seminal work, major developmental milestones include the *Neocognitron*, an early neural network for character recognition [12], and convolutional neural networks ("CNNs" or "ConvNets") utilized in speech [13] and visual signal processing [14], largely inspired by the visual cell types of the feline visual cortex [15]. Historical neural networks are differentiated from their modern progeny by the deep, or "hidden," layering that characterizes current architectures and is one of the principal reasons for the extreme performance gains seen in the contemporary literature. The training of modern architectures is made computationally tractable with innovations such as gradient-based optimization using backpropagation (first performed in [14]) and advances in hardware [16]. Uptake by both industry and academia alike is further facilitated through various neural network open-source platforms (e.g., Tensorflow [17] and Keras [18]).

A key event in the widespread adoption of CNNs was the 2012 ImageNet Large Scale Visual Recognition Challenge for object classification [19]. The winning entry, a CNN-based architecture colloquially known as *AlexNet* [20], reduced the error rate by almost half over other entries. Subsequent years' competitions were dominated by CNN variants such as VGG [21], GoogLeNet [22] and ResNet [23] with performance ultimately exceeding human performance in 2015 [24].

[*] Corresponding author.
E-mail address: ntustison@virginia.edu (N.J. Tustison).

https://doi.org/10.1016/j.mri.2019.05.037
Received 4 January 2019; Received in revised form 22 May 2019; Accepted 26 May 2019

Additional competition outlets, including conference-based venues (e.g., NeurIPS[1]) and community-based platforms, such as Kaggle[2], continue to highlight the utility of CNNs as comprehensive approaches to computer vision problems. This is in addition to the sheer number of formal research reports discussed in the same conferences, published in dedicated journals and hosted in online technical repositories. Notable reviews by key figures in the field include those of LeCun et al. [16], and Schmidhuber [26].

Early CNN-based research tailored to medical imaging dates back to the 1990s with classification tasks providing the majority of use cases (e.g., lung nodule classification [27,28] and breast tissue differentiation [29,30]). Despite the early adoption by certain research groups, widespread uptake did not occur until much later. Several deep learning overviews specific to medical imaging have been presented in the recent literature

- in editorial form [31],
- specific to generative adverserial networks (GANs) [32],
- related to MRI [33] with specific focus on neuro applications [34],
- for issues related to radiation therapy [35],
- surveying various medical imaging applications [36], and
- as general reviews [37-42].

Despite the thorough treatment of the topic contained in these reviews, discussion of chronological adoption within the community is limited. Regardless, one can informally gauge this evolution from utilization of alternative machine learning techniques to predominately CNN-based approaches from the various competitions held simultaneously with medical imaging conferences. For example, the annual Multimodal Brain Tumor Segmentation (BraTS) Challenge has taken place under the auspices of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) since 2012 wherein large sets of training data are provided to the competitors who attempt to perform a voxelwise labeling of the constituent components of brain tumors from multimodal MR image data. The winning entries from the first two years employed random forest classifiers for segmentation [43]. Although variations of the traditional random forest scheme continued to be well represented in the 2014 Challenge, CNN-based image segmentation algorithms made an appearance [44]. By 2018, CNN-based pipelines were, by far, the most common [45] with specific preference being that of the U-net architecture [46,47] which, as we describe below, features prominently in deep learning-based image registration.

Conspicuously, coverage of the topic of deep learning-based image registration, relative to the related algorithmic categories of image classification and segmentation, has not been as extensive in the reviews mentioned above, despite its prominence in the broader research literature. This disparity seems to be similarly reflected in the quantity of published research for those respective categories [32, 39]. This review is meant to address this disparity and thus provide an overview of the current state-of-the-art of this burgeoning subfield.

Given the prospective readership of this review, we only briefly sketch background material traditionally associated with image registration. This permits a follow-up introduction to generic differences (traditional vs. deep learning) including supervised vs. unsupervised optimization and various key network elements that characterize certain deep learning-based image registration architectures. Within this subsection, we also include additional network innovations which might find utility in future architectures. Based on these considerations of readership and topical familiarity, we organize discussion of current techniques based on an architectural classification (versus a more traditional categorization based on, for example, similarity metric). Finally, we briefly editorialize concerning the present and continued future confluence of deep learning and image registration.

2. Background

2.1. Image registration and loss functions

Given two images, denoted by I and J, pairwise image registration is the process of establishing spatial correspondence between the two domain representations. Such transformations are frequently generated through the optimization of an objective function, $E(I,J,T)$, typically of the form:

$$E(I, J, T) = \mathcal{S}(I, J \circ T) + \mathcal{R}(T) \qquad (1)$$

where T represents the spatial transformation between "fixed" (or "source") image I and "moving" (or "target") image J, $\mathcal{S}$ corresponds to a similarity measure between the fixed and warped moving images, and $\mathcal{R}$ denotes a regularization term quantifying certain constraints on transformation flexibility. Image registration, as a field of methodological development, can be largely characterized as a thorough exploration of various aspects of this functional relationship.

Casting the optimization function described by Eq. (1) within a deep learning paradigm is fairly straightforward with the so-called deep learning "loss function" often comprising similar terms as more traditional formulations. For example, the following are a sampling of common similarity measures which have a long history within the community [4-10]:

- Normalized cross correlation (NCC): $1 - \left\langle \frac{I-\mu_I}{\| I-\mu_I \|_2}, \frac{J-\mu_J}{\| J-\mu_J \|_2} \right\rangle$,
- L1 or photometric difference: $\|I - J\|$,
- Mean squared intensity error (MSQ): $\|I - J\|^2$,
- Normalized mutual information (NMI) [6, 48]: $\frac{H(I)+H(J)}{H(I,J)}$ where H is the Shannon entropy,
- Structural similarity index (SSIM) [49]: $\frac{(2\mu_I\mu_J + c_1)(2\sigma_{IJ} + c_2)}{(\mu_I^2 + \mu_J^2 + c_1)(\sigma_I^2 + \sigma_J^2 + c_2)}$, and
- Local cross correlation (LCC) [50]: $\frac{\sum_{x_i \in \mathcal{N}} ((I(x_i) - \mu_{I(x)})(J(x_i) - \mu_{J(x)}))^2}{\sum_{x_i \in \mathcal{N}} (I(x_i) - \mu_{I(x)})^2 \sum_{x_i \in \mathcal{N}} (J(x_i) - \mu_{J(x)})^2}$, and
- Dice label overlap [51]: $\frac{\sum_l |I_l \cap J_l|}{\sum_l |I_l \cup J_l|}$.

and continue to be relevant as loss functions with current deep learning trends (see Table 1). This is in addition to the incorporation of common explicit regularization terms derived from physical or geometric models.

Deep learning also provides opportunities for new perspectives such as training on loss functions which penalize deviations from the "true" transformations contained within the training data. This includes possibilities ranging from the L1-norm on the set of quaternary points defining a homology to the L2-norm of the true and predicted vectors defining a dense displacement field. There is also significant potential for leveraging the learned feature maps generated within the network architectures for optimizing these spatial transforms.

2.2. Supervised vs. unsupervised image registration

The distinction between supervised and unsupervised methods is particularly salient within the deep learning community, the former characterized by training data that contains both inputs and desired outputs which is lacking in the latter. Traditional image registration can be generally viewed as an unsupervised approach to transform optimization with deep learning expanding possibilities to include both supervised and unsupervised learning of spatial correspondence. Both

[1] https://nips.cc.

[2] Following the 2017 ImageNet challenge, in which the vast majority of teams surpassed the 5% classification error rate threshold, the ImageNet organizers ceded management to the Kaggle community which maintains a running performance assessment in ostensible perpetuity [25].

Table 1
Deep learning-based image registration methods organized in terms of basic network architecture.

Reference	Year	n-D	Transform[a]	Loss[b]
Feature localization				
Sergeev et al. [69]	2012	3-D	Affine	–
Weinzaepfel et al. [71]	2013	3-D	Deformable	–
Simonovsky et al. [73]	2016	3-D	Deformable	–
Wu et al. [75]	2016	3-D	Deformable	–
Two channel				
DeTone et al. [78]	2016	2-D	Homography	MSQ$_T$
Nguyen et al. [79]	2018	2-D	Homography	L1, MSQ$_T$
Rohé et al. [80]	2017	3-D	Diffeomorphic	MSQ$_T$
Eppenhof et al. [81]	2018	3-D	TPS	MSQ$_T$
Cao et al. [83]	2017	3-D	Deformable	NCC + ER
Hu et al. [84]	2018	3-D	Affine/deformable	Multiscale Dice
de Vos et al. [85]	2019	3-D	Affine + B-spline	NCC + ER
Shan et al. [87]	2018	2-D	Deformable	L1 + ER
Balakrishnan et al. [89]	2018	3-D	Deformable	LCC + ER
Dalca et al. [90]	2018	3-D	Diffeomorphic	MSQ
Krebs et al. [95]	2018	3-D	Diffeomorphic	LCC
Siamese/pseudo-siamese				
Dosovitskiy et al. [88]	2015	2-D	Optical flow	MSQ$_T$
Nowruzi et al. [99]	2017	2-D	Homography	MSQ$_T$
Rocco et al. [100]	2017	2-D	Affine + TPS	MSQ$_T$
Sloan et al. [101]	2018	2-D	Rigid	MSQ$_T$
Sokooti et al. [102]	2017	3-D	Deformable	L1$_T$
Zhang [103]	2018	3-D	IC deformable	MSQ + ER
Yang et al. [111]	2018	3-D	Diffeomorphic	L1$_T$
Generative adverserial networks				
Mahapatra et al. [113]	2017	3-D	Deformable	NMI + SSIM + VGG
Hu et al. [115]	2018	3-D	Deformable	Multiscale Dice + ER
Fan et al. [117]	2018	3-D	Deformable	ER
Other				
Miao et al. [118]	2016	2-D/ 3-D	Rigid	MSQ$_T$
Sheikhjafari et al. [119]	2018	2-D	Deformable	L1

[a] TPS: thin-plate spline, IC: inverse consistent.

[b] ER: explicit regularization, NCC: normalized cross correlation, MSQ: mean squared intensity error, MSQ$_T$: mean squared transformation error, LCC: local cross correlation, NMI: normalized mutual information, SSIM: structural similarity index, L1: L1 intensity error, L1$_T$: L1 transformation error, and VGG: VGG feature-based.

are represented in the proposed techniques surveyed below.

Supervised image registration within the context of deep learning entails the employment of sufficiently large training data sets of input fixed and moving image pairs with their corresponding transformations. These data are used to train a designated network to learn those transformation parameters based on features discovered through the training process. The loss function quantifies the discrepancy between the predicted and input transformation parameters. Possibilities for obtaining the desired transformations used in the training data include output from traditional image registration algorithms as well as synthetically derived data sets.

Unsupervised deep learning-based approaches are more closely related to their traditional analogs in that they lack of the use of input transformation data. Optimization is driven via loss functions which incorporate intensity-based similarity quantification in learning the correspondence between the fixed and moving images. This is conceptually analogous to the classic neural network example of unsupervised learning—the autoencoder (cf [52])—where differences between the input and the network-generated predicted version of the input are used to learn latent features characterizing the data. In the case of unsupervised image registration, the optimal transformation is that which maximizes the similarity (as determined by the user-selected similarity loss function) between the input, specifically the fixed image, and the network-generated predicted version of the input, specifically the warped moving image as determined by the concomitantly derived transform.

2.3. *Relevant network architectures*

Prior to describing the various image registration algorithms that have been recently proposed in the literature which incorporate elements of deep learning, we first describe some basic architectural components specifically relevant to such discussion which include:

- convolutional neural networks [13,14],
- siamese, pseudo-siamese [53,54] and two-channel networks,
- U-net [46,47],
- spatial transformer networks [55],
- diffeomorphic transformer networks [56],
- generative adverserial networks [57], and
- CoordConv [58].

Although not exhaustive, many of these core networks feature prominently in the research reviewed below. Other networks and/or network components were chosen based on their potential future application in image registration. For additional information, we refer the interested reader to the deep learning reviews cited earlier in addition to pertinent textbooks (e.g., [59]).

2.3.1. *Convolutional neural networks*

The grid-like informational content of certain data structures, such as 2-D and 3-D images, is perfectly suited to CNN-based training. The major elements of CNNs are localized convolutions, connections and pooling [16]. As indicated by nomenclature, the distinguishing characteristic of CNNs is the use of convolution instead of matrix operations in one or more of its constituent layers where the output are feature maps. These feature maps are typically generated in a hierarchical fashion synthesizing simple geometric features at the base convolutional layers (lines, corners, etc.) and progressing to more abstract features at the apical layers. The localized connections and weight-sharing provide a form of regularization while simultaneously reducing memory requirements [59]. The size of the convolution kernel, known as the "receptive field," determines the degree of connectivity. The accompanying pooling layers are used to subsample the convolutional feature maps in a way that statistically summarizes voxel neighborhoods within the feature maps. An illustration of a bare-bones CNN configuration is provided in Fig. 1 which depicts the core components of convolution and max pooling. Architectural novelty derives from innovative arrangements of these core (and other) network components and the connections between them. Although traditional CNNs are characterized by fixed rectilinear receptive fields, recent extensions to CNNs include deformable convolutional networks [60] which permit sampling at non-grid locations dictated by trainable transformation parameters internal to the CNNs themselves.

2.3.2. *Siamese, pseudo-siamese and two-channel networks*

Zagoruyko et al. thoroughly discuss architectural categorization for directly learning similarity functions from images (and patches) via CNNs [61], some of which are well-represented in the works reviewed below. So-called *siamese networks* [53, 54], illustrated in Fig. 2, have identical input branches which feed into a decision layer involving some form of distance measure, oftentimes calculated from the fully connected encoding of input images. Closely related is the pseudo-siamese architecture which uncouples the weights between the two input branches. In addition to the siamese and pseudo-siamese configurations, a perhaps more basic architecture involves arrangement of the image pair as two channels in the input layer of the network. This two

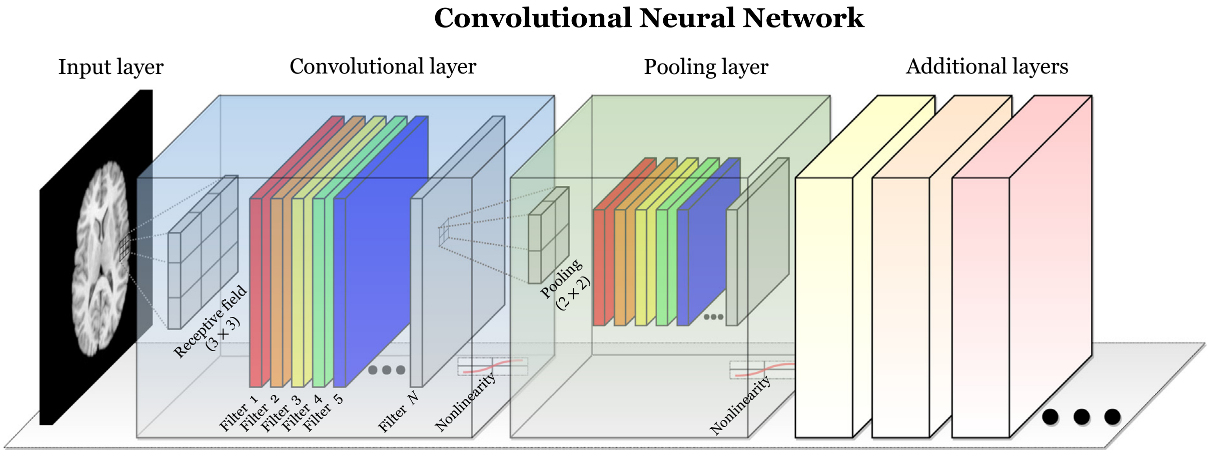

Fig. 1. The basic elements of the CNN. The convolutional layer comprises several filters which are optimized in terms of their responses to various features derived from the input layer (or previous layers in the case of multiple convolutional layers). Pooling is used to extract salient features and reduce computational complexity for further processing by subsequent layers.

channel network is reportedly fast to train but can be more computationally burdensome for testing [61].

2.3.3. U-net

An innovative extension to early CNN architectures is the fully convolutional network (FCN) described in [62] in the context of (semantic) image segmentation. The FCN replaces the traditional fully connected layers at the end of the network with convolutional layers to produce low-resolution heatmaps which indicate the presence of class-specific objects. These are then upsampled using bilinear interpolation and deconvolutional layers to produce the dense output corresponding in size to the input.

U-net [46,47] is a popular variant on the basic encoding/decoding strategy of FCNs but differs significantly in terms of the decoding branch of the network and the leveraging of skip connections inherited from ladder networks [63]. In the case of U-net, the decoder subnetwork mirrors its encoder counterpart with skip connections concatenating corresponding encoder/decoder levels such that the decoder can learn features lost at subsequent encoding levels. It has proved remarkably successful at medical image segmentation tasks (e.g., [45]) and its ability to project learned (i.e., encoded) feature vectors back into input space makes it suitable for learning registration-related data structures such as dense displacement fields.

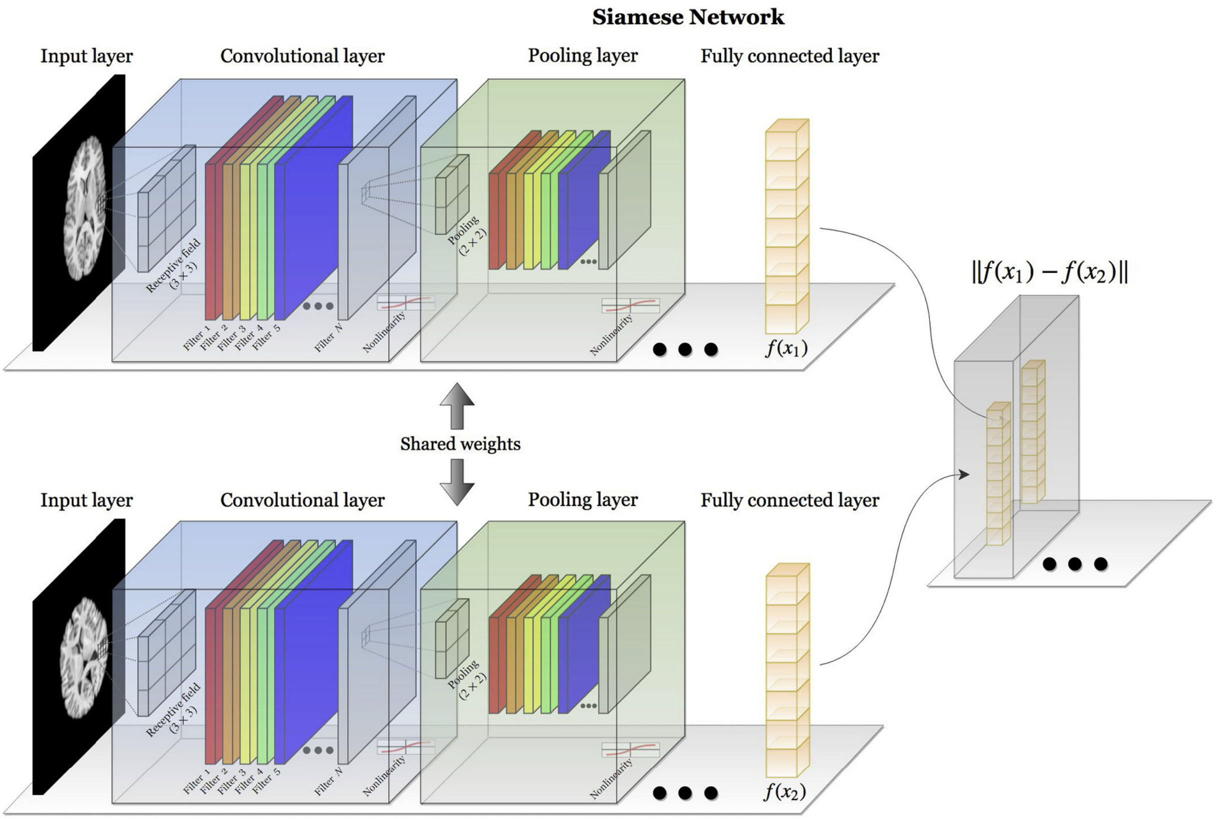

Fig. 2. Illustration of a siamese architecture. Two identical convolutional branches, which traditionally share weights, are used to learn a similarity distance while simultaneously optimizing the input encoding captured in the fully connected layers.

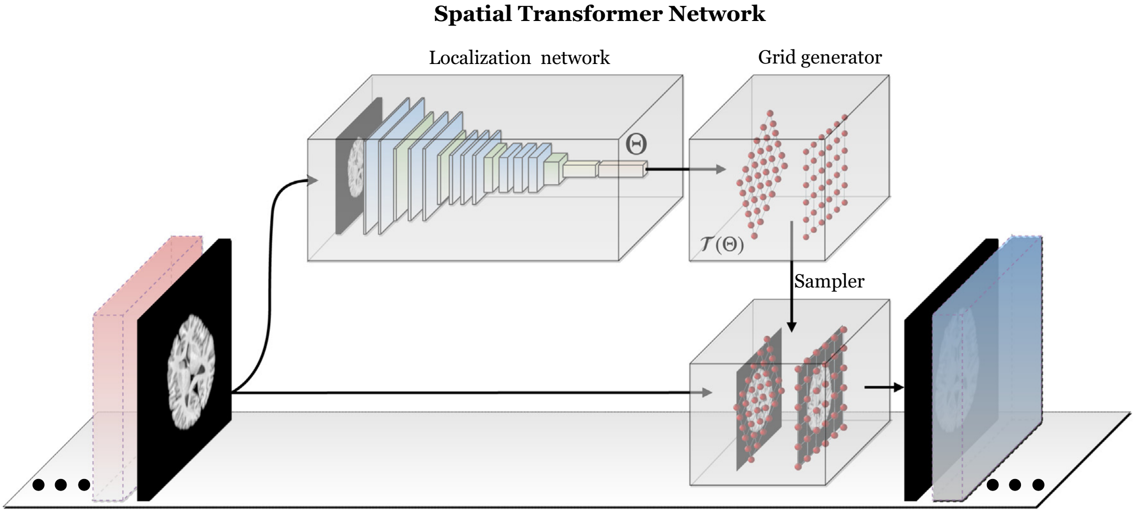

Fig. 3. Diagrammatic illustration of the spatial transformer network. The STN can be placed anywhere within a CNN to provide spatial invariance for the input feature map. Core components include the localization network used to learn/predict the parameters which transform the input feature map. The transformed output feature map is generated with the grid generator and sampler.

2.3.4. *Spatial transformer networks*

In 2015 Jaderberg and his fellow co-authors described a powerful new module, known as the spatial transformer network (STN) [55] which figures prominently in many of the image registration methods that we review below. Generally, STNs enhance CNNs by permitting a flexibility which allows for an explicit spatial invariance that goes beyond the implicitly limited translational invariance associated with the architecture's pooling layers. In many image-based tasks (e.g., localization or segmentation), designing an algorithm that can account for possible pose or geometric variation of the object(s) of interest within the image is crucial for maximizing performance. The STN is a fully differentiable layer which can be inserted anywhere in the CNN to learn the parameters of the transformation of the input feature map (not necessarily an image) which renders the output in such a way so as to optimize the network based on the specified loss function. The added flexibility and the fact that there is no manual supervision or special handling required makes this module an essential addition for any CNN-based toolkit.

An STN comprises three principal components: 1) a localization network, 2) a grid generator and 3) a sampler (see Fig. 3). The localization network uses the input feature map to learn/regress the transformation parameters which optimize a specified loss function. In many examples provided, this amounts to transforming the input feature map to a quasi-canonical configuration. The actual architecture of the localization network is fairly flexible and any conventional architecture, such as a fully connected network (FCN), is suitable as long as the output maps to the continuous estimate of the transformation parameters. These transformation parameters are then applied to the output of the grid generator which are simply the regular coordinates of the input image (or some normalized version thereof). The sampler, or interpolator, is used to map the transformed input feature map to the coordinates of the output feature map.

Since Jaderberg's original STN formulation, extensions have been proposed such as the inverse compositional STN (IC-STN) [64] and the diffeomorphic transformer network [56]. We defer discussion of the latter to the next subsection but briefly describe the former. Two issues with the STN include: 1) potential boundary effects in which learned transforms require sampling outside the boundary of the input image which can cause potential learning errors for subsequent layers and 2) the single-shot estimate of the learned transform which can compromise accuracy for large transformation distances. The IC-STN address both of these issues by 1) propagating transformation parameters instead of propagating warped input feature maps until the final transformation layer and 2) recurrent usage of the localization network for inferring transform compositions in the spirt of the inverse compositional Lucas-Kanade algorithm [65].

2.3.5. *Deep diffeomorphic transformer networks*

Although discussion of transform generalizability was included in the original STN paper [55], discussion was limited to affine, attention (scaling + translation) and thin-plate spline transforms which all comply with the requirement of differentiability. This work was extended to diffeomorphic transforms in [56]. The computational load associated with generating traditional diffeomorphisms through velocity field integration [66] motivated the use of continuous piecewise affine-based (CPAB) transformations [67]. The CPAB approach utilizes a tesselation of the image domain which translates into faster and more accurate generation of the resulting diffeomorphism. Although this does constrain the flexibility of the final transformation, the framework provides an efficient compromise for use in deep learning architectures. Analogous to traditional image registration, the deep diffeomorphic transformer layer can be placed in serial following an affine-based STN layer for a global-to-local total transformation estimation. This is demonstrated in the experiments reported in [56]. Similar to the many publicly available implementations of STN, the authors provide their own Tensorflow implementation of the diffeomorphic transformer network.[3] The authors employ CUDA-based calculations for evaluating the CPAB gradients and transforms due to speed considerations.

2.3.6. *Generative adverserial networks*

Goodfellow et al. introduced generative adverserial networks (GANs) in 2014 [57] and have increasingly found traction in addressing many types of deep learning problems in the medical imaging domain [32] including image registration. GANs are a special type of network composed of two adverserial subnetworks known as the *generator* (usually characterized by deconvolutional layers) and the *discriminator* (usually a CNN). These work in a minimax fashion to learn data distributions in the absence of extensive sample data. Seeded with a random noise image (e.g., sampled from a uniform or Gaussian distribution), the generator produces synthetic images which are then evaluated by the discriminator as belonging either to the true or synthetic data distributions in terms of some probability scalar value. This back-and-forth results in a generator network which continually

[3] https://github.com/SkafteNicki/ddtn.

improves its ability to produce data that more closely resembles the true distribution while simultaneously enhancing the discriminator's ability to judge between true and synthetic data sets. Since the original "vanilla" GAN paper, the number of proposed GAN extensions have exploded in the literature (see the GAN Zoo[4]). Initial extensions included architectural modifications for improved stability in training which have since become standard (e.g., deep convolutional GANs [68]).

2.3.7. Enhancing CNNs with CoordConv

Although not discussed, let alone used, in any of the papers reviewed below, the insight provided in [58] deserves consideration due to the subject matter of encoding spatial coordinates in CNN layers and its relevance to image registration. The authors describe a perplexing issue encountered during the course of their research. Reducing the core issue to toy examples, the authors demonstrate that training CNNs to regress cartesian coordinates from sparse, feature map pixel encodings (and vice versa) is highly problematic for conventional CNNs. In order to remedy this deficiency, the authors propose *CoordConv* which involves the modification of the conventional CNN layer with the concatenation of additional coordinate channels to the input. By explicitly encoding spatial information at each grid point in the input layer of the CNN, the authors improve performance not only in the toy examples but also in detection with the MNIST data set and in reinforcement learning scenarios involving video game play. Although not explicitly tested in the image registration problem domain, it is possible that such straightforward modifications to current architectures would substantially improve performance.

3. Image registration with deep learning

The following overview of deep learning image registration methods is categorized based on the discussion of network architectures given in the previous section with subgranularity provided in terms of loss function. We first discuss early work in which transformations were derived from CNN-based identification and localization of corresponding features in image pairs. We then review two channel approaches in which fixed and moving images are concatenated channelwise in the input layer. This segues to methods involving the related siamese and pseudo-siamese architectures. The final category concerns those adverserial approaches employing GANs. Other methods which do not fit in any of the above categories are also discussed. Additional categorization in terms of publications, application areas, and software libraries used are provided in (Fig. 4).

3.1. Image registration via feature localization

Much of the early work incorporating deep learning into solving image registration problems involved the detection of corresponding features and then using that information to determine the correspondence relationship between spatial domains. For example, at the start of the current era of deep learning in image-related research, the authors of [69] proposed point correspondence detection using multiple feedforward neural networks each of which is trained to detect a single feature. These neural networks are relatively simple consisting of two hidden layers each with 60 neurons where the output is a probability of it containing a specific feature at the center of a small image neighborhood. These detected point correspondences are then used to estimate the total affine transformation with the RANSAC algorithm [70]. Similarly, *DeepFlow* [71] uses CNNs to detect matching features (called *deep matching*) which are then used as additional information in the large displacement optical flow framework [72]. A relatively small architecture, consisting of six layers, is used to detect features at different convolution sizes which are then matched across scales.

A similarity measure for multimodal registration is formulated in terms of CNNs in the work of [73]. A two channel network is developed for input image patches (T1- and T2-weighted brain images). A B-spline image registration algorithm developed from the Insight Toolkit [74] is used to leverage the output CNN-based similarity measure for comparison with an identical registration set-up employing mutual information. Finally, in the category of feature learning, Wu et al. use nested auto-encoders (AE) to map patchwise image content to learned feature vectors [75]. These patches are then subsampled based on the importance criteria outlined in [76] which tends towards regions of high informational content such as edges. The AE-based feature vectors at these image patches are then used to drive a HAMMER-based registration [77] which is inherently a feature-based, traditional image registration approach.

3.2. Two channel architectures for image registration

3.2.1. Homography estimation

Two algorithms for more traditional computer vision applications are proposed in [78] and [79] where both are based on the VGG architecture [21] for 2-D homography estimation. The former framework includes both a regression network for determining corner correspondence and a classification network for providing confidence estimates of those predictions. The work in [79], which is publicly available[5], uses image patch pairs in the input layer and the L1 photometric loss between them to remove the need for direct supervision.

3.2.2. Training loss on ground truth transformations

Instead of training with a loss function based on similarity measures between fixed and moving images, the works of [80] and [81] formulate the loss in terms of the squared difference between ground-truth and predicted transformation parameters. In terms of network architecture, [80] employs a variant of U-net for training/prediction based on reference deformations provided by registration of previously segmented ROIs for cardiac matching where priority is alignment of the epicardium and endocardium. Displacement fields are parameterized by stationary velocity fields [82]. In contrast, [81] uses a smaller version of the VGG architecture to learn the parameters of a $6 \times 6 \times 6$ thin-plate spline grid.

3.2.3. Training loss on similarity metrics

Intermodality transformations involving CT and MRI are learned by training on the intramodality image pairs in [83]. The basic U-net architecture, using input patches of size $68 \times 68 \times 68$ voxels, incorporates a loss function combining normalized cross correlation (NCC) and explicit regularization for enforcing smoothness of the displacement field. A related idea is developed in [84] which uses labeled data and intensity information during the training phase such that only unlabeled image data is required for prediction. The latter architecture is a densely connected U-net architecture with three types of residual shortcuts [23]. For the loss function, the authors use a multiscale Dice function with an explicit regularization term for estimating both global and local transformations.

The unsupervised approach described in [85], denoted as *Deep Learning Image Registration* (*DLIR*), uses NCC to optimize a B-spline transform for 3-D images. This extension of the methodology first described in [86] complements a patch-based B-spline two-channel network with a pseudo-siamese affine registration network. For the deformable component, image patches from the fixed and moving images are passed through a CNN regression network to infer voxelwise displacement vectors which are then converted to B-spline control point parameters through a STN layer. Average pooling instead of max

[4] https://github.com/hindupuravinash/the-gan-zoo.

[5] https://github.com/tynguyen/unsupervisedDeepHomographyRAL2018.

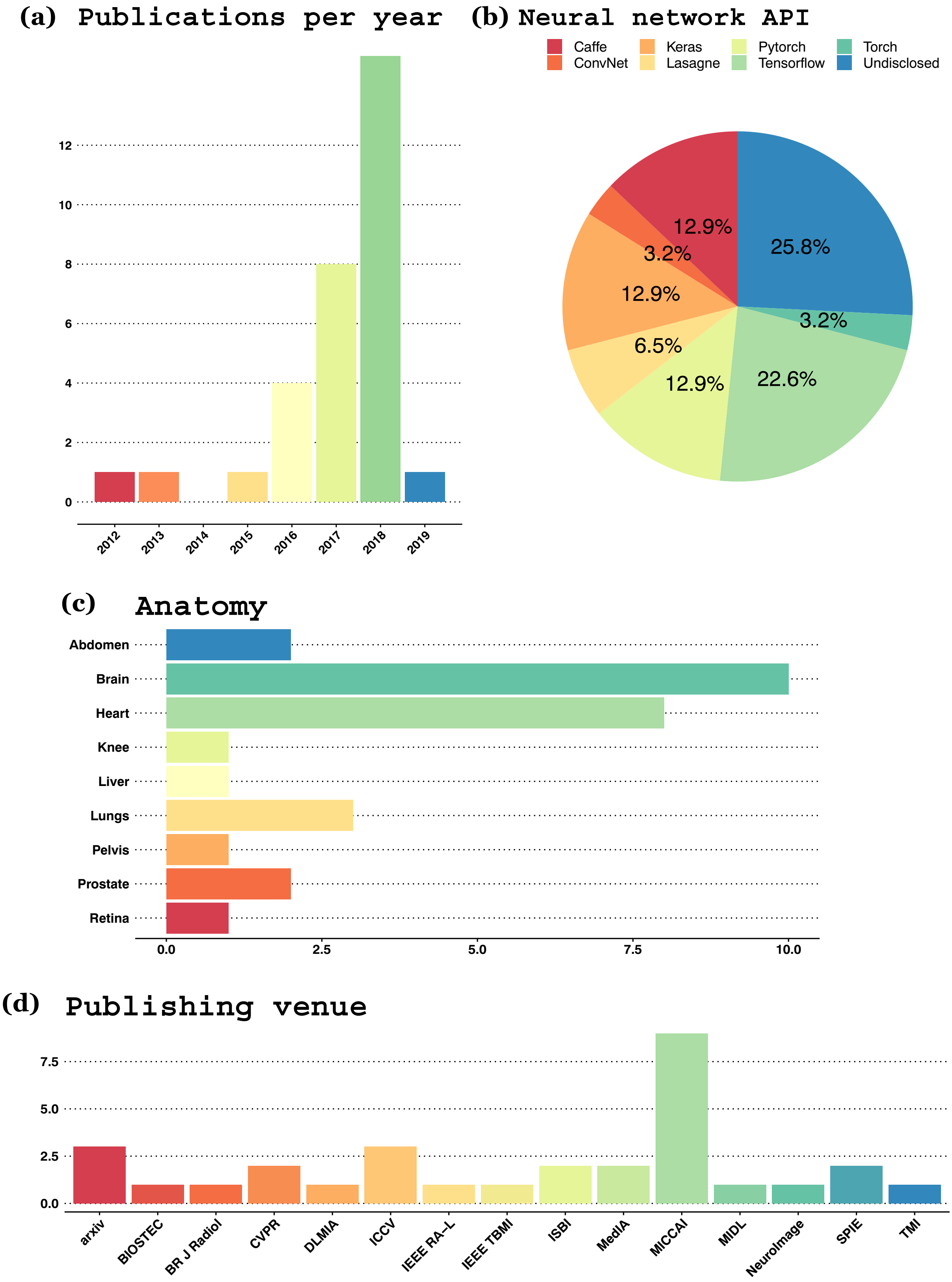

Fig. 4. Graphical overview of the works reviewed including (a) publications per year, (b) choice of neural network API, (c) anatomy (where applicable) and (d) publishing venue.

pooling is used to reduce the slight translation invariance associated with the latter. [87] is a 2-D approach for unsupervised medical image registration which also exploits a STN layer within the previously proposed FlowNet architecture [88] (discussed in the next section). An explicit penalty on the displacement field gradient promotes smoothing which is combined with an L1 photometric intensity error for the loss function.

3.2.4. Probabilistic generative models

Voxelmorph was first introduced in [89]. This method incorporates a U-net architecture with a STN where the input layer consists of the concatenated full fixed and moving image volumes resized and cropped to $160 \times 192 \times 224$ voxels. The output consists of the voxelwise displacement field of the same size as the input (times three—one for each vector component). The loss function for training combines cross correlation and a diffusion regularizer on the spatial gradients of the displacement field. This was extended to a generative approach in [90] to yield diffeomorphic transformations based on SVFs [82] using novel scaling and squaring network layers. The U-net architecture is used to estimate the distribution parameters of the velocity fields encapsulated by training data. A new imaging pair can then be registered by sampling from this learned distribution, computing the resulting diffeomorphic transformation, and then warping the moving image. The underlying code has been made available[6] which has facilitated independent evaluations such as [91] to compare performance with traditional algorithms (i.e., IRTK [48], AIR [92], Elastix [93], ANTs [50] and NiftyReg [94]).

Another generative image registration approach is that of [95] which uses a conditional variational autoencoder [96], an extension of the variational autoencoder [97] which permits incorporation of additional information for latent inference modeling. This multi-scale generative framework encodes the SVFs which are ultimately converted to the total transformation field in a similar fashion as [90]. A comparison with LCC-Demons [98] and Voxelmorph [90] is performed.

3.3. Siamese and pseudo-siamese architectures for image registration

3.3.1. Homography estimation

The homography estimator of [99] uses a hierarchical composition of subnetwork modules to determine final correspondence. The basic architecture is similar to [78] although the initial layers employ a Siamese structure to process the images in parallel. Each successive layer is meant to correct the residual transformation error produced by the previous layer. Similar to other homography estimators, the loss function is based on the mean-squared error of the homography parameters.

3.3.2. Training loss on ground truth transformations

An early seminal paper introduced *FlowNet*, a 2-D CNN-based approach to optical flow optimization [88]. Two encoding/decoding configurations are actually proposed for comparison for alignment of real world RGB images where the architectural distinction lies in the encoding component of the network. *FlowNetSimple* is a two channel architecture characterized by a concatenated series of contracting convolutional layers. Alternatively, the recommended pseudo-siamese *FlowNetCorr* separates the initial contracting layers to find meaningful corresponding features across the image pair which are subsequently combined using a correlation layer. Although the simple variant reportedly generalizes better for one of the data sets, the pseudo-siamese construction outperforms on the other two evaluation data sets. Another unique aspect of this work is the data augmentation performed by synthetic image generation which literally involves the addition of flying chairs to existing image scenes.

Consistent with typical workflows in medical image registration, a two-step transform hierarchy is proposed using deep learning in [100] where the results of an affine CNN regression network are fed into a deformable thin-plate spline network. Supervised training uses the mean squares difference between predicted and ground truth transformations as the loss function. Unlike other methods, the transformation-based loss is actually calculated by transforming a uniform grid based on the predicted and ground-truth transforms and calculating the mean squares distance between corresponding grid points. A Siamese network for regressing rigid transformation parameters on brain images is described in [101]. Inverse consistency considerations are made by swapping fixed and moving image pairs during training. Similar to [86], max pooling is avoided to minimize translational invariance of the operation. Model loss is quantified via the mean square error of the transformation parameters. *RegNet* [102] is a single-shot transformation estimation approach which is trained using a large set of simulated displacement fields. 3-D input patches at multiple scales are employed to determine the patchwise displacement field. The network architecture combines the multiscale patches downstream where the loss is the mean residual distance between predicted and ground truth displacements.

3.3.3. Training loss on similarity metrics

ICNet [103] is motivated by traditional concerns of inverse consistency in deformable transformations [104]. Two parallel U-net structures are used to determine the initial forward and inverse displacements which are then propagated through an inverse network to refine the respective mappings. The loss function comprises multiple regularization terms to prevent topological folding and promote displacement field smoothness in addition to a mean squared intensity term combining both forward and inverse mapped images to their respective counterparts.

3.3.4. Geodesic shooting with Quicksilver

The large deformation diffeomorphic metric mappings (LDDMM) framework for image matching derives from the theoretical foundations underlying diffeomorphic *flows* [105-107]. Such diffeomorphisms are sufficiently differentiable bijective mappings, or transformations, which have sufficiently differentiable inverses. Specifically, the set of possible diffeomorphic mappings, $\phi(\mathbf{x}, t)(\mathbf{x} \in \Omega, t \in [0, 1])$, between two images I and J can be described as the collection of paths connecting the two images determined by the equation

$$\int_0^1 \|v(t)\|_L^2 \, dt + \int_\Omega |I \circ \phi^{-1}(x, 1) - J|^2 d\Omega \tag{2}$$

where v is a time-dependent smooth field dictated by the functional norm L and determines the mapping via the ordinary differential equation

$$\frac{d\phi(\mathbf{x}, t)}{dt} = v(\phi(\mathbf{x}, t), t), \phi(\mathbf{x}, 0) = \mathbf{Id}. \tag{3}$$

The optimal diffeomorphic transformation between I and J can be described as a geodesic path [66] connecting the two images. Traditionally, computational approaches to determining this optimal geodesic path involve discretization of the velocity field followed by numerical integration. This is performed for a given number of iterations where, presumably, convergence implies arrival at the optimal solution (i.e., geodesic path). Alternatively, based on the work of [108], the Euler-Lagrange equations for Eq. (2) can be written as a system incorporating a “momentum” term. This work further demonstrated that the initial momentum determines the entire geodesic path. This alternative perspective engendered a new approach to determining the diffeomorphic solution between two images known as *geodesic shooting* (e.g., [66, 109]). Although initially formulated in terms of scalar momenta [109], a vector formulation was proposed in [110] which tends towards superior numerical behavior.

[6] https://github.com/voxelmorph/voxelmorph.

The supervised deep learning technique of Yang et al. [111], known as *Quicksilver*, leverages this geodesic shooting/vector momentum optimization approach for determining optimal diffeomorphic transformations. The network architecture consists of two parallel encoders for separate fixed/moving image patches (15 × 15 × 15 voxels). The output of these parallel branches is concatenated and sent through three identical decoder branches (one for each dimension). Thus, the output consists of the predicted vector momentum map which, as described above, determines the total transformation. In order to improve accuracy of the predicted momentum maps, a follow-on correction network is also proposed. This correction network, trained by inverting the mapping produced by the predicted momentum and computing the residual error, is meant to account for large deformations across patch boundaries. Of note, Quicksilver, written in PyTorch [112], is one of the handful of algorithms surveyed which has been made publicly available.[7]

3.4. Adverserial image registration

In order to constrain the mapping between moving and fixed images, the GAN-based approach outlined in [113] combines a content loss term (which includes subterms for normalized mutual information, structural similarity [49] and a VGG-based filter feature L2-norm between the two images) with a “cyclical” adverserial loss. This is constructed in the style of [114] who proposed this GAN extension, viz., CycleGAN, to ensure that the normally underconstrained forward intensity mapping is consistent with a similarly generated inverse mapping for “image-to-image translation” (e.g., converting a Monet painting to a realistic photo or rendering a winter nature scene as its summer analog). However, in this case, the cyclical aspect is to ensure a regularized field through forward and inverse displacement consistency.

The work of [115] employs discriminator training between finite-element modeling and generated displacements for the prostate and surrounding tissues to regularize the predicted displacement fields. The generator loss employs the weakly supervised learning method proposed by the same authors in [116] whereby anatomical labels are used to drive registration during training only. The generator is constructed from an encoder/decoder architecture based on ResNet blocks [23]. The prediction framework includes both localized tissue deformation and the linear coordinate-system-changes associated with the ultrasound imaging acquisition.

In [117], the discriminator loss is based on quantification of how well two images are aligned where the negative cases derive from the registration generator and the positive cases consist of identical images (plus small perturbations). Explicit regularization is added to the total loss for the registration network which consists of a U-net type architecture that extracts two 3-D image patches as input and produces a patchwise displacement field. The discriminator network takes an image pair as input and outputs the similarity probability.

3.5. Other CNN-based approaches for image registration

Early work [118] employed CNN-based regression for estimation of 2-D/3-D rigid image alignment of 3-D X-ray attenuation maps derived from CT and corresponding 2-D digitally reconstructed (DRR) X-ray images. The transformation space is partitioned into distinct zones where each zone corresponds to a CNN-based regressor which learns transformation parameters in a hierarchical fashion. The loss function is the means squares error on the transformation parameters.

A novel deep learning perspective is given in [119] where displacement fields are assumed to form low-dimensional manifolds and are represented in the proposed fully connected network as low-dimensional vectors. From the input vector, the network generates a 2-D displacement field used to warp the moving image using bilinear interpolation. The absolute intensity difference is used to optimize the parameters of network and latent vectors. Instead of explicit regularization of the displacement field, the sum of squares of the network weights are included with the intensity error term in the loss function.

4. Discussion

The rich history of image registration illustrates the significant role that it has played in the field of quantitative medical image analysis. This history is punctuated by many research developments, both small and great, which have resulted in transformations of greater accuracy, improved computational efficiency and enlightening theoretical novelty—all of which improves the community's ability to do science. In addition, the open-source emphasis of the current scientific environment has improved the didactic quality and availability of these contributions which democratizes such technologies and, in effect, acts as a positive feedback loop by leading to future methodological advancements.

The recent resurgence of deep learning and, in particular, its CNN-based applications in medical imaging appears to be of paradigmatic significance. The ability to train networks to perform complicated tasks efficiently without the need of hand-tailoring image features has positively disrupted the current research landscape as evidenced by deep learning representation in conference presentations and manuscripts. This significance appears to extend to the domain of image registration, although it is still too early to determine the precise nature of this impact.

Many of the additional challenges which concern traditional image registration methodologies and their introduction into the community persist with the deep learning expansion of the field. These challenges have been discussed at length in various articles, reviews, and editorials but it is worth reiterating due to their importance. Historically-rooted evaluation issues such as the use of public and/or private data sets and reproducibility concerns (e.g., published parameters, code availability) continue to be relevant for the deep learning shift. In addition, new concerns are salient such as the distinction in training and prediction speeds as well as possible hardware issues including GPU advancements and hardware availability.

4.1. Improving evaluation strategies

Concrete advancement in other domains of machine learning has been driven by definitive, public evaluation challenges. In contrast, the field of medical image registration has not arrived at a consensus forum through which to benchmark progress year to year. This is partly due to the lack of high-resolution ground truth datasets generated specifically for image registration evaluation goals. This shortcoming, combined with other factors discussed below, prevent a clear assessment of the specific deep learning contributions to performance gains, computational speed notwithstanding.

New strategies for generating gold or silver standard data specifically for the purposes of evaluating biomedical image registration are needed. The majority of medical image registration evaluation papers remain focused on measuring the degree of overlap between anatomical structures. A good registration result should improve such metrics. However, given the success of deep learning-based segmentation methods—which directly solve this problem without the need for registration—one may question whether registration is necessary at all. Indeed, augmentation strategies used when training segmentation networks typically *add* transformation-based variability into datasets. This is the inverse of how transformations are typically employed in the more traditional biomedical image analysis paradigm.

[7] https://github.com/rkwitt/quicksilver.

4.2. Rethinking methodological reporting in the literature

An additional, related challenge to assessing the literature of biomedical image registration is that the majority of technical papers do not report enough methodological detail to enable readers' understanding of performance differences. In the context of public challenges hosted by Kaggle, participants work off of common baseline datasets, share all code as a prerequisite to involvement and are evaluated against hidden datasets provided by challenge organizers. Data, preprocessing, networks, postprocessing and results are transparent. In the context of the biomedical image registration literature, such transparency in terms of evaluation and development source code—and use of truly hidden data—is rarely present [120].

A recent review of "deep regression" [121] provides guidance on how such issues might be resolved in published work. The paper uses three public ground truth datasets that represent different forms of correspondence problems. The authors evaluate well-known VGG and ResNet regression architectures on these reference datasets. Notably, the authors of this paper are not promoting any particular architecture or method under evaluation. Common parameter variations of these networks are carefully explored and results are reported in terms of the impact on confidence intervals. This study suggests that differences in performance due to preprocessing may exceed differences attributable to changes in network architecture. This finding, the objective approach and the reporting methods in this paper should be kept in mind when researchers and reviewers are considering new methodological efforts.

4.3. Tailoring deep learning tools for medical imaging

An additional challenge is the relatively immature state of medical imaging focused deep learning software frameworks. For instance, many Tensorflow layers that work effortlessly in 2-D are not yet translated to 3-D. Furthermore, the traditional, carefully-defined concepts of patient orientation, patient physical space and well-defined transformations of these spaces in two, three and four dimensions are lacking in existing deep learning frameworks. Facile workarounds to these issues exist. However, it is our hope that some of the deep testing and software engineering from medical imaging focused frameworks such as the Insight ToolKit eventually influence the construction of deep learning systems.

Despite these technical issues, low barrier to entry, medical-imaging focused collections of pre-trained networks and reusable code are beginning to emerge. Packages such as NiftyNet [122] (in python) and ANTsRNet [123] (in R) seek to build a bridge between deep learning knowledge domains that goes beyond segmentation or registration alone to solve a collection of problems via common underlying architectures, consistent interfaces and with the systematic use of best practices known to medical imaging researchers. While powerful, these systems, like most work in deep learning for medical image registration, have yet to run the gauntlet of use in real-world systems which necessitates adaptation, testing and debugging in broad ranges of application areas. However, despite the number of issues which remain to be addressed by the community, deep learning has opened an entirely new vista for exploration by current and future generations of medical imaging scientists.

Acknowledgments

We gratefully acknowledge the support of the NVIDIA Corporation with the donation of two Titan Xp GPUs used in support of associated research.

References

[1] Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. Med Image Anal 2015;24(1):205–19. https://doi.org/10.1016/j.media.2015.06.012.

[2] Ashburner J, Friston KJ. Voxel-based morphometry-the methods. Neuroimage 2000;11(6):805–21. https://doi.org/10.1006/nimg.2000.0582. Pt 1.

[3] Avants BB, Cook PA, Ungar L, Gee JC, Grossman M. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. Neuroimage 2010;50(3):1004–16. https://doi.org/10.1016/j.neuroimage.2010.01.041.

[4] Brown LG. A survey of image registration techniques. ACM Comput Surv 1992;24(4):325–76. https://doi.org/10.1145/146370.146374.

[5] Maintz JB, Viergever MA. A survey of medical image registration. Med Image Anal 1998;2(1):1–36.

[6] Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based registration of medical images a survey. IEEE Trans Med Imaging 2003;22(8):986–1004. https://doi.org/10.1109/TMI.2003.815867.

[7] Gholipour A, Kehtarnavaz N, Briggs R, Devous M, Gopinath K. Brain functional localization a survey of image registration techniques. IEEE Trans Med Imaging 2007;26(4):427–51. https://doi.org/10.1109/TMI.2007.892508.

[8] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration a survey. IEEE Trans Med Imaging 2013;32(7):1153–90. https://doi.org/10.1109/TMI.2013.2265603.

[9] Viergever MA, Maintz JBA, Klein S, Murphy K, Staring M, Pluim JPW. A survey of medical image registration - under review. Med Image Anal 2016;33:140–4. https://doi.org/10.1016/j.media.2016.06.030.

[10] Keszei AP, Berkels B, Deserno TM. Survey of non-rigid registration tools in medicine. J Digit Imaging 2017;30(1):102–16. https://doi.org/10.1007/s10278-016-9915-8.

[11] Ivakhnenko AG. Polynomial theory of complex systems. IEEE Trans Syst Man Cybern Syst 1971;SMC-1(4):364–78.

[12] Fukushima K. Neocognitron a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 1980;36(4):193–202.

[13] Waibel A. Phoneme recognition using time-delay neural networks. Meeting of the Institute of Electrical Information and Communication Engineers (IEICE). 1987.

[14] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1(4):541–51.

[15] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 1962;160:106–54.

[16] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.

[17] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous systems. 2015.

[18] Chollet F. Keras. 2007.

[19] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.

[20] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. 2012. p. 1097–105.

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409 2014:1556.

[22] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. CoRR Abs/1512.00567. 2015: http://arxiv.org/abs/1512.00567 Available at.

[23] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CoRR, Abs/1512.03385. 2015: http://arxiv.org/abs/1512.03385 Available at.

[24] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers surpassing human-level performance on imagenet classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[25] Available athttps://www.kaggle.com/c/imagenet-object-localization-challenge. 2018

[26] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117.

[27] Lo SC, Freedman MT, Lin JS, Mun SK. Computer-aided detection of mammographic calcifications pattern recognition with an artificial neural network. Proc SPIE: Medical imaging: Image processing 1992;1898:859–69.

[28] Lo SC, Freedman MT, Lin JS, Mun SK. Automatic lung nodule detection using profile matching and back-propagation neural network techniques. J Digit Imaging 1993;6(1):48–54.

[29] Chan HP, Lo SC, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications pattern recognition with an artificial neural network. Med Phys 1995;22(10):1555–67. https://doi.org/10.1118/1.597428.

[30] Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 1996;15(5):598–610. https://doi.org/10.1109/42.538937.

[31] Greenspan H, Ginneken BV. Summers, r. m. Deep learning in medical imaging overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35(5):1153–9.

[32] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review ArXiv preprint 2018.

[33] Mazurowski MA, Buda M, Saha A, Bashir Mustafa R. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J Magn Reson Imaging 2018.

[34] Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Marti R, Lladó X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. 2018. https://doi.org/10.1016/j.artmed.2018.08.008.

[35] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. 2018. https://doi.org/10.1002/mp.13264.

[36] Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89.

[37] Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol 2017;10(3):257–73. https://doi.org/10.1007/s12194-017-0406-5.

[38] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.

[39] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[40] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks a review. J Med Syst 2018;42(11):226. https://doi.org/10.1007/s10916-018-1088-1.

[41] Biswas M, Kuppili V, Saba L, Edla DR, Suri HS, Cuadrado-Godia E, et al. State-of-the-art review on deep learning in medical imaging. Front Biosci (Landmark Ed) 2019;24:392–426.

[42] Altaf F, Islam SMS, Akhtar N, Janjua NK. Going deep in medical image analysis: concepts, methods, challenges and future directions ArXiv preprint 2019.

[43] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

[44] Conference proceedings of the 3rd MICCAI BraTS challenge. 2014.

[45] Pre-conference proceedings of the 7th MICCAI BraTS challenge. 2018.

[46] Ronneberger O, Fischer P, Brox T. U-Net convolutional networks for biomedical image segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 9351. 2015. p. 234–41.

[47] Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net deep learning for cell counting, detection, and morphometry. Nat Methods 2019;16(1):67–70. https://doi.org/10.1038/s41592-018-0261-2.

[48] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations application to breast MR images. IEEE Trans Med Imaging 1999;18(8):712–21. https://doi.org/10.1109/42.796284.

[49] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.

[50] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A Reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 2011;54(3):2033–44. https://doi.org/10.1016/j.neuroimage.2010.09.025.

[51] Crum WR, Camara O, Hill DLG. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging 2006;25(11):1451–61. https://doi.org/10.1109/TMI.2006.880587.

[52] Hinton GE, Zemel RS. Autoencoders minimum description length and Helmholtz free energy. Advances in Neural Information Processing Systems 1994:3–10.

[53] Bromley J, Guyon I, LeCun Y, Sckinger E, Shah R. Signature verification using a 'siamese' time delay neural network. Neural Information Processing Systems; 1994.

[54] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. Proceedings of the IEEE International Conference on Computer Vision. 2005.

[55] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. Neural Information Processing Systems; 2015.

[56] Detlefsen NS, Freifeld O, Hauberg S. Deep diffeomorphic transformer networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[57] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014.

[58] Liu R, Lehman J, Molino P, Such FP, Frank E, Sergeev A, Yosinski J. An intriguing failing of convolutional neural networks and the Coordconv solution arXiv preprint 2018.

[59] Goodfellow I, Bengio Y, Courville A. Deep learning. 2016.

[60] Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks arXiv preprint arXiv:1703.06211 2017.

[61] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks. Proceedings of the IEEE, Conference on Computer Vision and Pattern Recognition 2015.

[62] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit 2015:3431–40. https://doi.org/10.1109/CVPR.2015.7298965.

[63] Valpola H. Chapter 8 - from neural PCA to deep unsupervised learning. Advances in independent component analysis and learning machines 2015:143–71. https://doi.org/10.1016/B978-0-12-802806-3.00008-7.

[64] Lin CH, Lucey S. Inverse compositional spatial transformer networks. Proceedings of the IEEE, Conference on Computer Vision and Pattern Recognition 2017.

[65] Baker S, Matthews I. Lucas-Kanade 20 years on: a unifying framework. Int J Comput Vis 2004;56(3):221–55. https://doi.org/10.1023/B:VISI.0000011205.11775.fd Available at.

[66] Beg MF, Miller MI, Trouvé A, Younes L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vis 2004;61(2):139–57.

[67] Freifeld O, Hauberg S, Batmanghelich K, Fisher JW. Transformations based on continuous piecewise-affine velocity fields. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2496–509. https://doi.org/10.1109/TPAMI.2016.2646685.

[68] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. Proceedings of the International Conference on Learning Representations 2016.

[69] Sergeev S, Zhao Y, Linguraru MG, Kazunori Okada. Medical image registration using machine learning-based interest point detector. Proceedings of the SPIE 2012.

[70] Fischler MA, Bolles RC. Random sample consensus a paradigm for model fitting with applications to image analysis and automated cartography, comm. ACM 1981;24(6):381–95.

[71] Weinzaepfel P, Revaud J, Harchaoui Z, Schmid C. Deepflow Large Displacement optical flow with deep matching. Proceedings of the IEEE International Conference on Computer Vision 2013:1385–92. https://doi.org/10.1109/ICCV.2013.175.

[72] Brox T, Malik J. Large displacement optical flow descriptor matching in variational motion estimation. IEEE Trans Pattern Anal Mach Intell 2011;33(3):500–13. https://doi.org/10.1109/TPAMI.2010.143.

[73] Simonovsky M, Gutierrez-Becker B, Mateus D, Navab N, Komodakis N. A deep metric for multimodal registration. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2016.

[74] Yoo TS, Metaxas DN. Open Science-combining open data and open source software: medical image analysis with the insight toolkit. Med Image Anal 2005;9(6):503–6. https://doi.org/10.1016/j.media.2005.04.008.

[75] Wu G, Kim M, Wang Q, Munsell BC, Shen D. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Trans Biomed Eng 2016;63(7):1505–16. https://doi.org/10.1109/TBME.2015.2496253.

[76] Wang Q, Wu G, Yap PT, Shen D. Attribute vector guided groupwise registration. Neuroimage 2010;50(4):1485–96. https://doi.org/10.1016/j.neuroimage.2010.01.040.

[77] Shen D, Davatzikos C. HAMMER hierarchical attribute matching mechanism for elastic registration. IEEE Trans Med Imaging 2002;21(11):1421–39. https://doi.org/10.1109/TMI.2002.803111.

[78] DeTone D, Malisiewicz T, Rabinovich A. Deep image homography estimation, arXiv 2016;1606:03798.

[79] Nguyen T, Chen SW, Shivakumar SS, Taylor CJ, Kumar V. Unsupervised deep homography. A fast and robust homography estimation model: proceedings of IEEE Robotics and Automation Letters; 2018.

[80] Rohe and Datar, M and Heimann, T and Sermesant, M and Pennec, X́ MM. SVF-net learning deformable image registration using shape matching. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 2017:266–74.

[81] Eppenhof KAJ, Lafarge MW, Moeskops P, Veta M, Pluim JPW. Deformable image registration using convolutional neural networks. Proceedings of the SPIE Image Processing: Medical Imaging; 2018.

[82] Arsigny V, Commowick O, Pennec X, Ayache N. A log-Euclidean framework for statistics on diffeomorphisms. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 9. 2006. p. 924–31.

[83] Cao X, Yang J, Zhang J, Nie D, Kim MJ, Wang Q, et al. Deformable image registration based on similarity-steered CNN regression. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 10433. 2017. p. 300–8. https://doi.org/10.1007/978-3-319-66182-7_35.

[84] Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 2018;49:1–13. https://doi.org/10.1016/j.media.2018.07.002.

[85] de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal 2019;52:128–43. https://doi.org/10.1016/j.media.2018.11.010.

[86] Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-end unsupervised deformable image registration with a convolutional neural network. Deep learning in medical image analysis and multimodal learning for clinical decision support. 2017. p. 204Ű-212.

[87] Shan S, Yan W, Guo X, Chang EIC, Fan Y, Xu Y. Unsupervised end-to-end learning for deformable medical image registration, arxiv. 2018.

[88] Dosovitskiy A, Fischery P, Ilg E, Hausser P, Hazirbas C, Golkov V, et al. Flownet: learning optical flow with convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015. p. 2758-Ű2766. https://doi.org/10.1109/ICCV.2015.316.

[89] Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. An unsupervised learning model for deformable medical image registration. Proceedings of the IEEE, Conference on Computer Vision and Pattern Recognition. 2018.

[90] Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning for fast probabilistic diffeomorphic registration. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018. p. 729–38.

[91] Nazib A, Fookes C. And Perrin, D. A comparative analysis of registration tools: traditional vs deep learning approach on high resolution tissue cleared data, arXiv preprint; 2018.

[92] Woods RP, Mazziotta JC, Cherry SR. MRI-pet registration with automated algorithm. J Comput Assist Tomogr 1993;17(4):536–46.

[93] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196–205. https://doi.org/10.1109/TMI.2009.2035616.

[94] Modat M, Ridgway GR, Taylor ZA, Lehmann M, Barnes J, Hawkes DJ, et al. Fast free-form deformation using graphics processing units. Comput Methods Programs Biomed 2010;98(3):278–84. https://doi.org/10.1016/j.cmpb.2009.09.002.

[95] Krebs J, Mansi T, Mailhé B, Ayache N, Delingette H. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. Proceedings of the 4th International Workshop, DLMIA and 8th International Workshop ML-CDS. 2018.

[96] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems 2015;28:3483–91.

[97] Kingma DP, Welling M. Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). 2014.

[98] Lorenzi M, Ayache N, Frisoni GB, Pennec X. Alzheimer's disease neuroimaging initiative (ADNI). LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm. Neuroimage 2013;81:470–83. https://doi.org/10.1016/j.neuroimage.2013.04.114.

[99] Nowruzi FE, Laganiere R, Japkowicz N. Homography estimation from image pairs with hierarchical convolutional networks. Proceedings of the International Conference of Computer Vision. 2017.

[100] Rocco I, Arandjelović R, Sivic J. Convolutional neural network architecture for geometric matching. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[101] Sloan JM, Goatman KA, Siebert JP. Learning rigid image registration - utilizing convolutional neural networks for medical image registration. Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - volume 2: Bioimaging. 2018. p. 89–99.

[102] Sokooti H, Vos B, de Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2017. p. 232-Ű239.

[103] Zhang J. Inverse-consistent deep networks for unsupervised deformable image registration. CoRR abs/180903443. 2018: http://arxiv.org/abs/1809.03443 Available at.

[104] Christensen GE, Johnson HJ. Consistent image registration. IEEE Trans Med Imaging 2001;20(7):568–82. https://doi.org/10.1109/42.932742.

[105] Trouvé A. Diffeomorphic groups and pattern matching in image analysis. Int J Comput Vis 1995;28:213–21.

[106] Christensen GE, Rabbitt RD, Miller MI. Deformable templates using large deformation kinematics. IEEE Trans Image Process 1996;5(10):1435–47. https://doi.org/10.1109/83.536892.

[107] Dupuis P, Grenander U, Miller MI. Variational problems on flows of diffeomorphisms for image matching. Quarterly of Applied Mathematics LVI 1998:587–600.

[108] Miller MI, Trouvé A, Younes L. Geodesic shooting for computational anatomy. J Math Imaging Vis 2006;24(2):209–28. https://doi.org/10.1007/s10851-005-3624-0.

[109] Vialard FX, Risser L, Rueckert D, Cotter CJ. Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. Int J Comput Vis 2012;97:229–41.

[110] Singh N, Hinkle J, Joshi S, Fletcher PT. A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. Proc IEEE Int Symp Biomed Imaging 2013;2013:1219–22. https://doi.org/10.1109/ISBI.2013.6556700.

[111] Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver fast predictive image registration–a deep learning approach. Neuroimage 2017;158:378–96. https://doi.org/10.1016/j.neuroimage.2017.07.008.

[112] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lerer A, et al. Automatic differentiation in PyTorch. NIPS-w; 2017.

[113] Mahapatra D, Antony B, Sedai S, Garnavi R. Deformable medical image registration using generative adversarial networks. Proceedings of IEEE 15th International Symposium on Biomedical Imaging. 2018.

[114] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. IEEE International Conference on Computer Vision. 2017.

[115] Hu Y, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, et al. Adversarial deformation regularization for training image registration neural networks. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018.

[116] Hu Y, Modat M, Gibson E, Ghavami N, Bonmati E, Moore CM, et al. Label-driven weakly-supervised learning for multimodal deformable image registration. Proceedings of IEEE 15th International Symposium on Biomedical Imaging. 2018.

[117] Fan J, Cao X, Xue Z, Yap PT, Shen D. Adversarial similarity network for evaluating image alignment in deep learning based registration. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018.

[118] Miao S, Wang ZJ, Liao R. A CNN regression approach for real-time 2D/3D registration. IEEE Trans Med Imaging 2016;35(5):1352–63. https://doi.org/10.1109/TMI.2016.2521800.

[119] Sheikhjafari A, Noga M, Punithakumar K, Ray N. Unsupervised deformable image registration with fully connected generative neural network. Proceedings of Medical Imaging with Deep Learning. 2018.

[120] Tustison NJ, Johnson HJ, Rohlfing T, Klein A, Ghosh SS, Ibanez L, et al. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. Front Neurosci 2013;7:162. https://doi.org/10.3389/fnins.2013.00162.

[121] Lathuilière S, Mesejo P, Alameda-Pineda X, Horaud RA. Comprehensive analysis of deep regression. CoRR abs/180308450 2018: http://arxiv.org/abs/1803.08450 Available at.

[122] Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, et al. Niftynet: a deep-learning platform for medical imaging. Comput Methods Prog Biomed 2018. https://doi.org/10.1016/j.cmpb.2018.01.025: https://www.sciencedirect.com/science/article/pii/S0169260717311823 Available at.

[123] Tustison NJ, Avants BB, Lin Z, Feng X, Cullen N, Mata JF, et al. Convolutional neural networks with template-based data augmentation for functional lung image quantification. Acad Radiol 2018. https://doi.org/10.1016/j.acra.2018.08.003.