Learning Unsupervised Parameter-specific Affine Transformation for Medical Images Registration

Xu Chen[1,3], Yanda Meng[1], Yitian Zhao[2], Rachel Williams[1], Srinivasa R. Vallabhaneni[1,3], and Yalin Zheng[1, ✉]

[1] Institute of Life Course and Medical Sciences, University of Liverpool, UK
yalin.zheng@liverpool.ac.uk
[2] Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, P. R. China
[3] Liverpool Vascular & Endovascular Service, Royal Liverpool University Hospital NHS Trust, UK

Abstract. Affine registration has recently been formulated using deep learning frameworks to establish spatial correspondences between different images. In this work, we propose a new unsupervised model that investigates two new strategies to tackle fundamental problems related to affine registration. More specifically, the new model 1) has the advantage to explicitly learn specific geometric transformation parameters (e.g. translations, rotation, scaling and shearing); and 2) can effectively understand the context between the images via cross-stitch units allowing feature exchange. The proposed model is evaluated on two two-dimensional X-ray datasets and a three-dimensional CT dataset. Our experimental results show that our model not only outperforms state-of-art approaches and also can predict specific transformation parameters. Our core source code is made available online[1].

1 Introduction

Image registration is a crucial challenge in the field of biomedical image analysis. Image registration aims to align two (or more) given images, namely, a target image $I_{tgt} : \Omega_{tgt} \subset \mathbb{R}^d \mapsto \mathbb{R}$, and a source image $I_{src} : \Omega_{src} \subset \mathbb{R}^d \mapsto \mathbb{R}$, by establishing their spatial correspondences into a common coordinate system.

Affine transformation is commonly used to correct for geometric distortions or deformations that occur with non-ideal camera angles. The planetary of surfaces, parallelism and angles between lines are all preserved in affine transformation. In general, an affine transformation is a composition of rotations, translations, scaling, and shears, which can be expressed as an energy minimization problem: $\mathbf{A}^* = argmin\{S[I_{tgt}, \mathbf{A}(I_{src})]\}$, where $\mathbf{A}$ is the affine transformation matrix[2] and S is the metrics to measure the dissimilarity between I_{tgt} and $\mathbf{A}(I_{src})$.

[1] *https://github.com/xuuuuuuchen/PASTA*

[2] $\mathbf{A}_{2D} = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{bmatrix}$ and $\mathbf{A}_{3D} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Image registration formulated in deep learning settings has shown promising results. Several approaches have been proposed for affine image registration and nonlinear image registration with the convolutional neural networks (CNN). The Spatial Transformer Network (STN) [8] is one of the first CNN-based methods to learn two-dimensional (2D) affine transformation for the classification of distorted MNIST digit in a supervised learning manner. The localization network of STN can regress the outputs from a CNN to produce the 2D transformation matrix $\mathbf{A}_{2D}$. $\mathbf{A}_{2D}$ has six parameters encoded from translation, scaling, rotation and shearing. *Miao et al.* proposed a supervised CNN to regress the three-dimensional (3D) transformation matrix $\mathbf{A}_{3D}$ according to the synthesized transformation parameters as the ground truth for affine registration of X-ray images [11].

On the other hand, unsupervised models for affine and nonlinear transformation learning are more desirable as transformation ground truth is no longer required, which is not always available. In a recent work by *de Vos et al.* [16], an unsupervised *Deep Learning Image Registration* (DLIR) framework for joint affine and nonlinear registration was proposed. The affine transformation framework in the *DLIR* is a multi-stage approach for the multi-temporal MRI and CT 3D image registration. The $\mathbf{A}_{3D}$ matrix is regressed by two separate CNNs. The performance of the *DLIR* outperformed conventional image registration methods when tested on a cine cardiac MRI dataset and a chest CT dataset, respectively. Similar to the DLIR, in *Hu et al.*'s work [6], segmentation labels are used as a type of ground truth and considered in a loss function to help the similarity maximisation for MRI-Ultrasound image scans registration. There are two sub-networks in this model: a CNN regressor as the *Global-Net* and a U-net-like architecture [13] as the *Local-Net* for affine and nonlinear transformation, respectively. For unsupervised 3D affine transformation learning, the AIRNet [4] was proposed to estimate the $\mathbf{A}_{3D}$ for brain MR scan alignments by training a self-supervised CNN.

Despite the recent promising progress in deep learning-based image registration, most of the existing approaches directly regress the affine transformation matrix $\mathbf{A}$, but not explicitly regress specific geometric transformation parameters in the form of translations, rotation, scaling and shearing. For **2D transformation**, there will be seven transformation parameters, namely, one rotation (θ), two translations (t_x, t_y), two scaling (sc_x, sc_y) and two shears (sh_x, sh_y). It is obvious that six parameters in the matrix form $\mathbf{A}_{2D}$ can be easily derived from these seven spatial transformation parameters (θ, t_x, t_y, sc_x, sc_y, sh_x and sh_y), but not vice versa. Similar to 2D transformation, there will be 15 transformation parameters in **3D transformation**, that is, three rotations (θ_x, θ_y, θ_z), three translation (t_x, t_y, t_z), three scaling (sc_x, sc_y, sc_z) and six shears (sh_{xy}, sh_{xz}, sh_{yx}, sh_{yz}, sh_{zx}, sh_{zy}). These 15 parameters can be used to derive the twelve parameters of the matrix $\mathbf{A}_{3D}$, but not vice versa.

Therefore, there is a dilemma: if we only determine the matrix $\mathbf{A}$ (e.g. $\mathbf{A}_{2D}$ and $\mathbf{A}_{3D}$ for 2D and 3D case, respectively) like most of the other existing deep learning-based models, we cannot infer those spatial transformation parameters

as there is no unique solution and cannot explain the effect of each type of transformations. To tackle the above drawbacks and limitations, we propose a novel parameter-specific affine transformation model by explicitly learning all these spatial transformation parameters rather than learning their combinations, namely, the transformation matrix $\mathbf{A}$. Furthermore, cross-stitch units [12] have been developing for multi-task learning [3, 14, 15]. We introduce "cross-stitch" units [12] into our model to effectively learn an optimal combination of shared representations between image pairs.

2 Methods

In this section, we will describe our model in detail. More specifically, it has two unique features: a CNN-based parameter-specific affine transformation (PASTA) framework to formulate affine transformation, and a Cross-stitch Affine Network (CANet) to effectively learn transformation parameters in an unsupervised manner. For simplicity, the 2D affine transformation case will be presented here whist the 3D case will be shown in the supplementary material.

2.1 Parameter-specific Affine Transformation

The affine transformation matrix $\mathbf{A}$ can be formed by composing the rotation matrix $\mathbf{M}_{ro}$, the shearing matrix $\mathbf{M}_{sh}$, the scale matrix $\mathbf{M}_{sc}$ and the translation matrix $\mathbf{M}_t$ in turn[3],

$$\mathbf{A} = \mathbf{M}_t \cdot \mathbf{M}_{sc} \cdot \mathbf{M}_{sh} \cdot \mathbf{M}_{ro} = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

Where, each of the above four 2D transformation types can be represented as: $\mathbf{M}_t = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$ $\mathbf{M}_{ro} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\mathbf{M}_{sh} = \begin{bmatrix} 1 & sh_x & 0 \\ sh_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\mathbf{M}_{sc} = \begin{bmatrix} sc_x & 0 & 0 \\ 0 & sc_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$.
Our PASTA framework aims to optimise each transformation parameters, namely, one for rotation (θ), two for translation (t_x, t_y), two for scaling (sc_x, sc_y) and two for shearing (sh_x, sh_y) instead of directly optimising the transformation matrix $\mathbf{A}$. After optimising each transformation parameters, the matrix $\mathbf{A}$ as expressed in Eq.(2) can be derived according to the order in Eq.(1).

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 = sc_x\cos(\theta) + sh_x sc_x \sin(\theta) & \mathbf{a}_2 = -sc_x\sin(\theta) + sh_x sc_x \cos(\theta) & \mathbf{a}_3 = t_x \\ \mathbf{a}_4 = sh_y sc_y \cos(\theta) + sc_y \sin(\theta) & \mathbf{a}_5 = -sh_y sc_y \sin(\theta) + sc_y \cos(\theta) & \mathbf{a}_6 = t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Note, the range of each of the seven transformation parameters will be empirically known for specific applications. Thus, we can linearly normalise them

[3] $\mathbf{A}$ is subject to the composition order. In this work, we use the order shown in Eq.(1)

into the range of $[0, 1]$ where 0 and 1 corresponds to the minimum λ^i_{min} and maximum λ^i_{max} value of the ith parameter, respectively. This normalisation is beneficial: the outputs from the network will always be within $[0, 1]$, and the gradient in the optimisation will not be too small or too big. When our framework is in action, it will regress each of the seven parameters between $[0, 1]$. Each of them will then be mapped back to the actual values. Following that, these parameters will be used to generate the matrix form $\mathbf{A}$. After that, bilinear interpolation is applied when warping the I_{src} image by the $\mathbf{A}$.

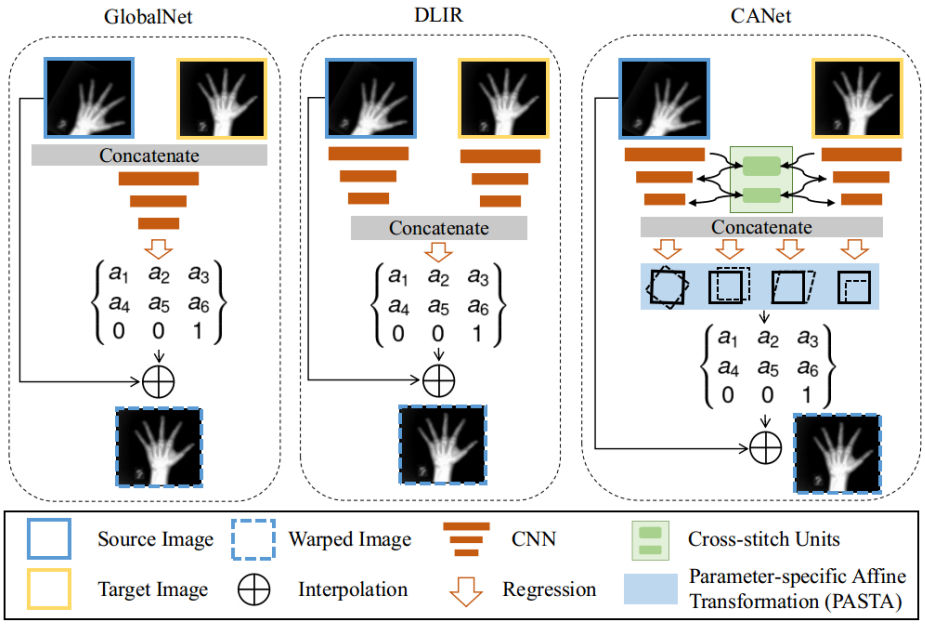

Fig. 1. Overview of our proposed model for unsupervised affine image registration

2.2 Cross-stitch Affine Network

We also propose a novel architecture for unsupervised affine transformation learning, partly motivated by cross-stitch units from multi-task learning [12]. Cross-stitch units intend to allow a model to determine how the task-specific network leverages the other task-specific network's knowledge. Cross-stitch units can learn the optimal linear combination of the output from the previous layers. We integrate these cross-stitch units into our model with our PASTA framework and provide an end-to-end learning model. We refer to it as a Cross-stitch Affine Network (CANet), sketched in Fig. 1. For affine transformation learning, the I_{src} and I_{tgt} are fed into two separate sub-networks S and T, respectively, and the two outputs (activation maps) h_S and h_T from the S and T are concatenated

to estimate the matrix $\mathbf{A}$ to warp the I_{src}. Unlike the previous works [16] that uses the hard parameter sharing, our model can learn the best-shared representations between two separate sub-networks just as in the soft parameter sharing. More specifically, the k-th layer in our CANet shares the representations via the Cross-stitch units C by learning a linear combination of the activation maps $h^{ij}_{S,k}$ and $h^{ij}_{T,k}$ at location (i, j). The outputs of the C are:

$$\begin{bmatrix} \tilde{h}^{ij}_{S,k} \\ \tilde{h}^{ij}_{T,k} \end{bmatrix} = \begin{bmatrix} C_{SS} & C_{TS} \\ C_{ST} & C_{TT} \end{bmatrix} \begin{bmatrix} h^{ij}_{S,k} \\ h^{ij}_{T,k} \end{bmatrix} \tag{3}$$

Where the C_{ST} and C_{TS} are the parameters weighting the activations between the sub-networks h_S and h_T, and the C_{SS} and C_{TT} are the parameters weighting the activations of the same sub-networks. Further, we demonstrate these units' effectiveness for the 2D and 3D affine transformation tasks.

3 Experiments and Results

In this section, we investigate the performance of our proposed model in both 2D and 3D applications. All models were trained on one node of a cluster with sixteen 8-core Intel CPUs, 8 TESLA V100 GPUs and 1TB memory with the spatial transformer module adapted from the open-source code in VoxelMorph [2], implemented in TensorFlow 1.14. All of the models were trained by the Adam optimizer. For a fair comparison of different models, we searched the optimal learning rate between e^{-3} and e^{-6} for each model based on the validation set.

Baseline To evaluate the performance of the proposed model, we compared ours with the GlobalNet [6] and DLIR [16], the two most widely used networks for affine registration. To investigate the robustness and generalizability of the proposed PASTA framework, we plugged in our PASTA model into the GlobalNet [6], the DLIR [16], and our CANet for 2D and 3D tasks. All the above networks are not pre-trained on any image datasets. Two widely used metrics, normalized cross correlation score (NCC) and Dice coefficient score (DSC), were introduced for the image registration dissimilarity assessment. Mean absolute error (MAE) is introduced for evaluating the performance of individual affine transformation parameters compared with the synthetic parameters.

Datasets In order to evaluate the performance of our models, we have applied it to three biomedical image datasets: (1) ChestMNIST of MedMNIST [18]: ChestMNIST[4] contains 10,000 frontal-view chest X-ray images based on the NIH-ChestXray14 dataset [17]. (2) HandMNIST of MedMNIST [18] contains 10,000 hand X-ray images. We used the original size 64x64 available when we downloaded them. (3) Learn2Reg[5]: 2020 MICCAI Registration Challenge (Task 2) [5], this dataset consists of 60 3D CT thorax images taken from 30 subjects (20 for training and 10 for testing). For all the scans an automatic lung segmentation is provided to evaluate the registration methods.

[4] https://medmnist.github.io/#dataset
[5] https://learn2reg.grand-challenge.org/Datasets/

Initialization Because real-world datasets with high-quality annotation is hard to acquire and kept for evaluation only, training on synthetic data is necessary. We generated 20 synthetic transformed X-ray images for each X-ray image. In this work, the range $[\lambda_{min}, \lambda_{max}]$ of 7 transformation parameters are sh_x, sh_y $\in [-0.1, 0.1]$, $\theta \in [-30°, 30°]$, sc_x, $sc_y \in [0.9, 1.1]$ and t_x, $t_y \in [-0.2, 0.2]$. In total, 200,000 pairs of synthetic images were generated and divided 50% (n=120,000) of images for training, 25% (n=40,000) for validation, and the remaining 25% for testing. The corresponding seven transformations parameters were used as 'ground truth' for further evaluation and comparison. For the 3D work, each scan was used to generate 100 synthetic transformed scans to pair itself. The ranges $[\lambda_{min}, \lambda_{max}]$ of the 15 transformation parameters are θ_x, θ_y, $\theta_z \in [-5°, 5°]$, sc_x, sc_y, $sc_z \in [0.90, 1.0]$, sh_{xy}, sh_{xz}, sh_{yx}, sh_{yz}, sh_{zx}, $sh_{zy} \in [0.0, 0.1]$ and t_x, t_y, t_z $\in [-0.1, 0.1]$. We used the official data split and resized the source images into $128 \times 128 \times 128$. There are 6,000 pairs of synthetic CT scans and were divided into 2,400 pairs for training, 1,600 pairs for validation, and the remaining 2,000 pairs for testing.

Ablation Study We investigated the number of cross-stitch units n in our proposed 2D- and 3D-CANet, respectively. We introduced a variable-controlling method to perform this ablation study to investigate the individual impact of different number of cross-stitch units.

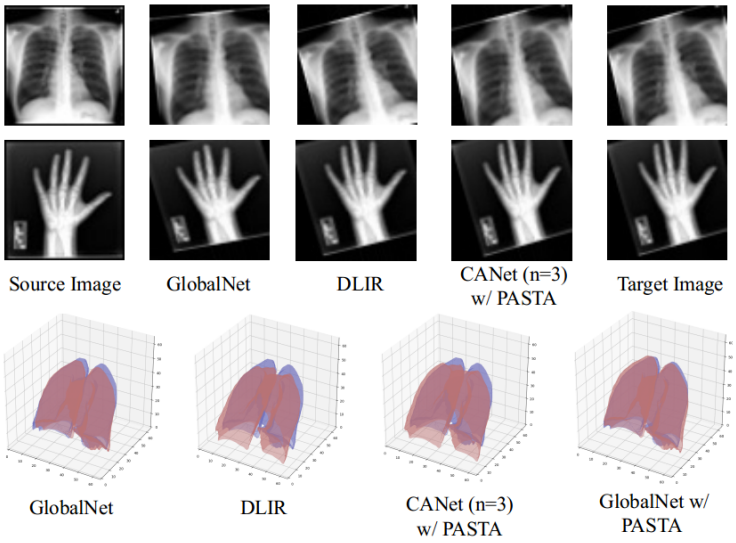

Fig. 2. 2D and 3D registration results of our proposed model for unsupervised affine image registration compared to the previous state-of-art approaches

Results on 2D Datasets We evaluated our models on the HandMNIST dataset and ChestMNIST dataset in comparison to the GlobalNet and DLIR.

Table 1. Quantitative 2D and 3D registration results of our proposed models compared to the others. Standard deviation is provided in the brackets.

Datasets \ Models	PASTA	GlobalNet [8]		DLIR [16]		CANet (n=1)		CANet (n=2)		CANet (n=3)	
		NCC	#Para.	NCC	#Para.	NCC	#Para.	NCC	#Para.	NCC	#Para.
HandMNIST	w/o	0.868 (0.086)	72K	0.918 (0.056)	298K	0.927 (0.051)	298K	0.920 (0.028)	302K	0.964 (0.033)	564K
	w/	0.929 (0.054)	72K	0.933 (0.049)	298K	0.913 (0.057)	298K	0.928 (0.055)	302K	**0.966 (0.032)**	564K
ChestMNIST	w/o	0.859 (0.074)	72K	0.957 (0.029)	298K	0.962 (0.026)	298K	0.947 (0.033)	302K	0.978 (0.015)	564K
	w/	0.972 (0.025)	72K	0.945 (0.038)	298K	0.935 (0.039)	298K	0.970 (0.025)	302K	**0.988 (0.012)**	564K
		DSC	Para.	DSC	Para.	DSC	Para.	DSC	Para.	DSC	Para.
Learn2Reg (Task 2)	w/o	0.903 (0.094)	142K	0.884 (0.030)	682K	0.889 (0.030)	683K	0.908 (0.033)	748K	0.883 (0.045)	17M
	w/	**0.938 (0.029)**	142K	0.890 (0.037)	682K	0.886 (0.042)	683K	0.911 (0.047)	748K	0.910 (0.059)	17M

The quantitative results on the both datasets are presented in Table 1. For the HandMNIST dataset, our CANet (n=3) with PASTA framework achieved better performance in terms of NCC score of 0.966 than the GlobalNet (NCC=0.868) and DLIR (NCC=0.918), followed by the CANet (n=3) (NCC=0.964). The use of PASTA framework improved the performance in terms of NCC score of the GlobalNet from 0.868 to 0.929 and the GlobalNet from 0.918 to 0.933, respectively. On the other hand, for the HandMNIST dataset, compared to the GlobalNet (NCC = 0.859) and the DLIR (NCC = 0.957), our CANet (n=3) with PASTA achieved the best performance in terms of NCC of 0.988, followed by the CANet (n=3) (NCC=0.978). For both datasets, when the number of cross-stitch units is increased from one to three, the performance in terms of NCC score are improved consistently, but the contribution from PASTA is decreasing gradually because the network is more powerful to directly regress the $\mathbf{A}$ matrix. However, the computational cost will proportionally increase.

Results on 3D Dataset We also evaluated our models on the CT lung dataset compared to the existing models, and presented the quantitative results in Table 1. The GlobalNet with our PASTA achieved the best performance among all the other models in terms of DSC score of 0.938, followed by CANet (n=2) with the PASTA (DSC = 0.911) and CANet (n=3) with the PASTA (DSC = 0.910). When our PASTA in action, except the GlobalNet was improved from 0.903 to 0.938, the DLIR was improved from 0.884 to 0.890, and the CANet (n=3) was improved from 0.883 to 0.910, the performance of CANet (n=1,2) were similar. Figure. 2 presents the registration results of 2D and 3D registration, we can observe that the registration results of the proposed PASTA and CANet are more accurate compared with the other existing methods.

Statistical Analysis We performed t-tests for our 2D and 3D results. Except DLIR and CANet (n=1) for the HandMNIST and CANet (n=1) for ChestMNIST and 3D lung dataset, all the other networks using PASTA have shown statistically significant improvements than those without PASTA ($p<0.001$).

On the other hand, when PASTA is used, CANet (n=3) performs significantly better than all the other networks ($p<0.001$) but the GlobalNet for the 3D lung dataset. These results confirmed the value of PASTA and the effectiveness of CANet.

Specific Transformation Parameters Analysis Due to the benefit of the PASTA framework, we also can investigate and evaluate the performance of individual affine transformation parameters for the 2D and 3D affine transformation compared with the synthetic parameters. The experimental results showing that the CANet (n=3) with PASTA performs the best in terms of mean absolute error (MAE) = 0.2%, followed by CANet (n=2) with PASTA (MAE = 0.5%), CANet (n=1) with PASTA (MAE = 0.6%) and DLIR with PASTA (MAE = 0.6%).

Table 2. Quantitative results of 2D transformation in terms of (mean absolute error) of our proposed models compared with the others.

	Translation		Rotation	Shear		Scaling		Avg.
	x	y	θ	x	y	x	y	
GlobalNet w/ PASTA	0.000	0.000	0.013	0.012	0.009	0.001	0.000	0.005
DLIR + PASTA	0.002	0.000	0.011	0.013	0.009	0.004	0.000	0.006
CANet (n=1) w/ PASTA	0.002	0.000	0.012	0.015	0.010	0.004	0.001	0.006
CANet (n=2) w/ PASTA	0.000	0.000	0.010	0.013	0.010	0.000	0.000	0.005
CANet (n=3) w/ PASTA	0.000	0.000	0.005	0.006	0.006	0.000	0.000	**0.002**

Results on Real Datasets Further, we investigated and evaluated the performance between real pairs in HandMNIST dataset. 44,850 unique pairs were generated by randomly chosen from 300 different X-ray images of left hands (a ratio of 60:20:20 for training, validation and testing). The results on the testing set proved that the CANet and PASTA (NCC=0.849) can introduce improvement in terms of NCC compared to the methods without using them or before registration (NCC=0.655). The GlobalNet with PASTA and CANet (n=3) with PASTA achieved better performance in terms of NCC score of 0.853 and 0.849 respectively than the GlobalNet (NCC=0.843), the DLIR with PASTA (NCC=0.833) and the DLIR (NCC=0.829). Fig. 3 presents the real registration results by different models.

Furthermore, we randomly chose 50 pairs of images and annotated the fingertips of thumb, middle finger and pinky finger, because 1) real pairs with true transformation parameters is not easy to acquire and 2) the error of key-points is more reasonable to evaluate the registration performance than the similarity between images only. Euclidean Distance is introduced to measure the registration accuracy. Before applying registration, the errors of thumb, middle finger and pinky are 7.547, 8.325 and 9.041 pixels, respectively. The quantitative results are presented in Table 3. For the thumb, middle finger alignment, our CANet (n=3) with PASTA framework achieved the smallest distance of 4.155 and 3.877 pixels respectively, followed by the GlobalNet with PASTA (4.37 and 4.686 pix-

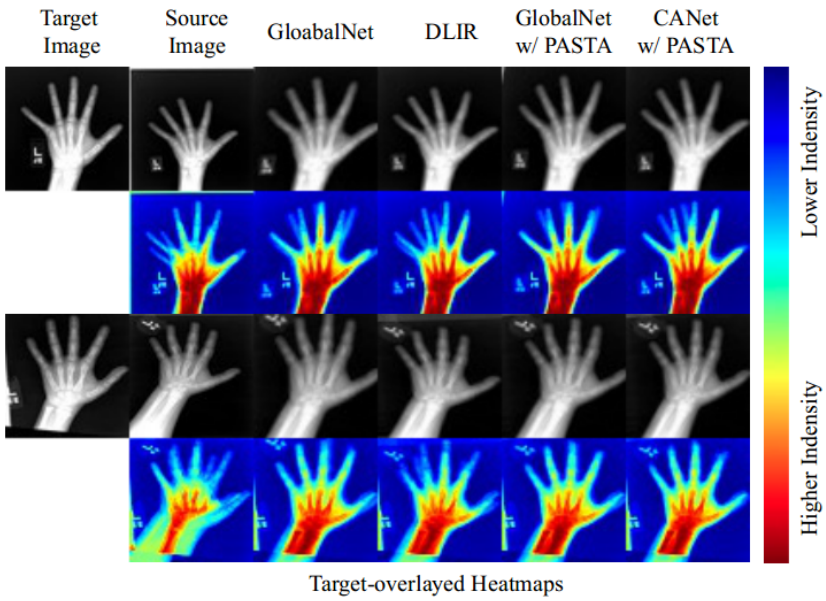

Fig. 3. 2D real registration results of our proposed model for unsupervised affine image registration compared to the previous state-of-art approaches

els) and the GlobalNet (4.47 and 4.458 pixels). For the pinky finger alignment, the GlobalNet with PASTA (4.607 pixel) outperformed the CANet (n=3) with PASTA (5.420 pixel), the GlobalNet (6.604 pixel) and the DLIR (7.398 pixel).

Table 3. Quantitative results of real transformation of our proposed models compared to the others. Standard deviation is provided in the brackets.

Metrics / Models	Image Similarity	Thumb	Middle Finger	Pinky Finger
	NCC	Euclidean Distance of Fingertips		
Before	0.655 (0.139)	7.547 (3.911)	8.325 (4.973)	9.041 (4.515)
GlobalNet [6]	0.843 (0.070)	4.470 (2.158)	4.458 (3.814)	6.604 (3.275)
DLIR [16]	0.829 (0.076)	5.655 (2.877)	6.875 (4.929)	7.398 (3.705)
GlobalNet [6] w/ PASTA	0.853 (0.067)	4.372 (2.278)	4.686 (5.908)	4.607 (2.253)
CANet w/ PASTA	0.849 (0.066)	4.155 (2.489)	3.877 (3.705)	5.420 (2.492)

4 Conclusion

In this work, we propose an unsupervised registration model that can explicitly learn specific geometric transformation parameters in the form of translations, rotation, scaling and shearing for both 2D and 3D transformation. we propose the

CANet that can effectively learn the linear combination between the images pairs via cross-stitch units for affine transformation learning. Three public datasets: 2D ChestMNIST, HandMNIST and 3D Learn2Reg CT (task 2), are used for the evaluation of our proposed models. Our experimental results show that our models in 2D and 3D outperform the state-of-art approaches (GlobalNet and DLIR). PASTA is generic and could be compatible with other networks without increasing the computation cost. In the future, we will extend our models for joint affine and nonlinear image registration [7][16] as well as graph convolutional networks-based [10][9] or atlas-based [1][19] image segmentation problems.

Acknowledgments

Xu Chen is funded by a studentship jointly funded by the Vascular Surgery Research Fund in Liverpool and Institute of Life Course and Medical Sciences, University of Liverpool, and partially funded by The Great Britain-China Educational Trust (no.269944) administered by the Great Britain-China Centre.

References

1. Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D.: Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage **46**(3), 726–738 (2009)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging (2019)
3. Beljaards, L., Elmahdy, M.S., Verbeek, F., Staring, M.: A cross-stitch architecture for joint registration and segmentation in adaptive radiotherapy. In: Medical Imaging with Deep Learning. pp. 62–74. PMLR (2020)
4. Chee, E., Wu, J.: Airnet: Self-supervised affine registration for 3d medical images using neural networks. arXiv preprint arXiv:1810.02583 (2018)
5. Hering, A., Murphy, K., van Ginneken, B.: Lean2reg challenge: Ct lung registration - training data (May 2020)
6. Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T.: Label-driven weakly-supervised learning for multimodal deformable image registration. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1070–1074. IEEE (2018)
7. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis **49**, 1–13 (2018)
8. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
9. Meng, Y., Meng, W., Gao, D., Zhao, Y., Yang, X., Huang, X., Zheng, Y.: Regression of instance boundary by aggregated cnn and gcn. In: European Conference on Computer Vision. pp. 190–207. Springer (2020)

10. Meng, Y., Wei, M., Gao, D., Zhao, Y., Yang, X., Huang, X., Zheng, Y.: Cnn-gcn aggregation enabled boundary regression for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 352–362. Springer (2020)
11. Miao, S., Wang, Z.J., Liao, R.: A cnn regression approach for real-time 2d/3d registration. IEEE Transactions on Medical Imaging **35**(5), 1352–1363 (2016)
12. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3994–4003 (2016)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
14. Ruder, S., Bingel, J., Augenstein, I., Søgaard, A.: Sluice networks: Learning what to share between loosely related tasks. arXiv preprint arXiv:1705.08142 **2** (2017)
15. Tissera, D., Vithanage, K., Wijesinghe, R., Kahatapitiya, K., Fernando, S., Rodrigo, R.: Feature-dependent cross-connections in multi-path neural networks. arXiv preprint arXiv:2006.13904 (2020)
16. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52**, 128–143 (2019)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
18. Yang, J., Shi, R., Ni, B.: Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. arXiv preprint arXiv:2010.14925 (2020)
19. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transforms for one-shot medical image segmentation. arXiv preprint arXiv:1902.09383 (2019)