# Respiratory Motion Prediction Based on 4D-CT/CBCT Using Deep Learning

by

Xinzhi Teng

Duke Kunshan and Duke University
Medical Physics Graduate Program

Date:_____

Approved:

Lei Ren, Supervisor

_____

James Bowsher

_____

Jackie Wu

_____

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the
Medical Physics Graduate Program of
Duke Kunshan and Duke University

2019

ABSTRACT

Respiratory Motion Prediction Based on 4D-CT/CBCT Using Deep Learning

by

Xinzhi Teng

Duke Kunshan and Duke University
Medical Physics Graduate Program

Date:_____

Approved:

Lei Ren, Supervisor

_____

James Bowsher

_____

Jackie Wu

_____

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the
Medical Physics Graduate Program of
Duke Kunshan and Duke University

2019

# Abstract

Purpose: The purpose is to investigate the feasibility of using Convolutional Neural Network (CNN) to register phase-to-phase deformation vector field (DVF) of lung 4D Computed Tomography (CT) / Cone-Beam Computed Tomography (CBCT).

Methods: A Convolutional Neural Network (CNN) based deep learning method was built to directly register the deformation vector field from the individual phases images of patient 4D CT or 4D CBCT for 4D contouring, dose accumulation or target verification. The input consists of image pairs while the output is the corresponding DVF that registers the image pairs. The centers of patch pairs are uniformly chosen across the lung and the size of the patches was chosen to cover the majority movement of deformable vectors. The network consisted of four convolutional layers, two average pooling layers and two fully connected layers. The loss function was half mean squared error. 11 sets of 4D image volumes from 9 patients with lung cancer were used as material, and the feasibility of the CNN was tested with volumes of intra-patient and inter-patient. In intra-patient study, the image volumes were sorted into different combinations, (1) training and testing samples from the same 4D-CT image volume, (2) training and testing samples from two 4D-CT volumes, (3) training and testing samples from 4D-CBCT volumes simulated by DRR from 4D-CT volumes, (4) training from 4D-CT and testing from 4D-CBCT reconstructed from primary projections, and (5) training

and testing samples from two 4D-CBCT volumes reconstructed from primary projections. In inter-patient study, five 4D-CT volumes from five patients were used as the training set and the sixth patient 4D-CT volume was the testing set. The functionality of a well-trained network adapting new patient's anatomy was tested. The coefficient of correlations between the prediction DVF and the reference DVF was calculated. The deformed images from reference DVF and predicted DVF were reconstructed. The cross correlation between both deformed images were calculated.

Results: Both deformed images show a good match in some major features, such as diaphragm, main vessels and fiducial marker, but some residual motions still exist. The cross correlation between both deformed images in the region of diaphragm is calculated. For all the intra-patient cases, the cross correlation between two deformed images are over 0.9 in the region of diaphragm. For the inter-patient case, this cross correlation is 0.87.

Conclusion: CNN based regression model successfully learn the DVF from one image set, and the trained model can be successfully transferred to another data set, provided the high image quality in training sets and similar anatomic structure between both image sets.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I would like to express my gratitude to all those who helped me during the writing of this thesis.

I gratefully acknowledge the help of my supervisor, Dr. Lei Ren, who offered me such a valuable research opportunity and also offered me tons of suggestions in the academic studies.

I also acknowledge my groupmates, who are willing to discuss my difficulties in research and help me solve the problems, such as Yao Zhao, Yingxuan Chen and Zhuoran Jang.

I'd like to further thank my parents who support me spiritually and financially to finish this program.

Last but not the least, this research is supported by the grand R01-CA184173.

# 1. Introduction

This part presents the background of the thesis, such as the explaining the terminologies, introducing the origin of the ideas, explaining the concepts, and emphasizing the novelty of this work.

## 1.1 Image Registration

Image registration is the process of transforming different sets of data into one coordinate system. The data may be from different sensors, times, depths or viewpoints (Brown, 1992). This technique has wide range of utility, such as medical image, military application, and computer target reorganization. Image registration is a crucial step when comparing or integrating the data which from different measurements.

### 1.1.1 Basic concepts in Image Registration

When performing image registration, we need to register the moving/source image to a fixed /target image. The moving image is the image that will be transformed into the target image coordinate system. When registering image A to image B, image A is referred as the moving or source image, and image B is referred as the fixed or target image. After the image registration, the image A is transformed to image A', which is the registered image or deformed image that matches with image B. (Goshtasby, 2005) Image A' is not necessarily identical to the target image B due to registration errors. However, compared with the source image A, image A' is much closer to image B, as the example shown in fig. 1.

The image registration algorithm generates a deformation vector field (DVF), which specifies how each voxel in the image A moves to deform A to match with the image B. The key component from registration algorithm is to solve the DVF.

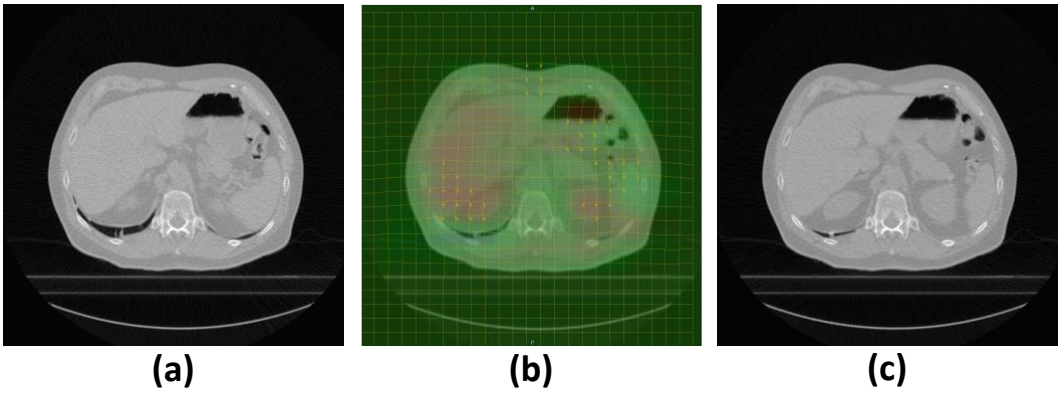


**Figure 1 The demonstration of image registration with coronal slice of lung 4D-CT image in different respiratory phases. (a) is the source image, (b) is the overlap of deformed image and deformable vector field, and (c) is the target image. One could see that after image registration, (b) the deformed image is more similar to (c) the target image.**

*Forward and backward mapping in deformable registration*

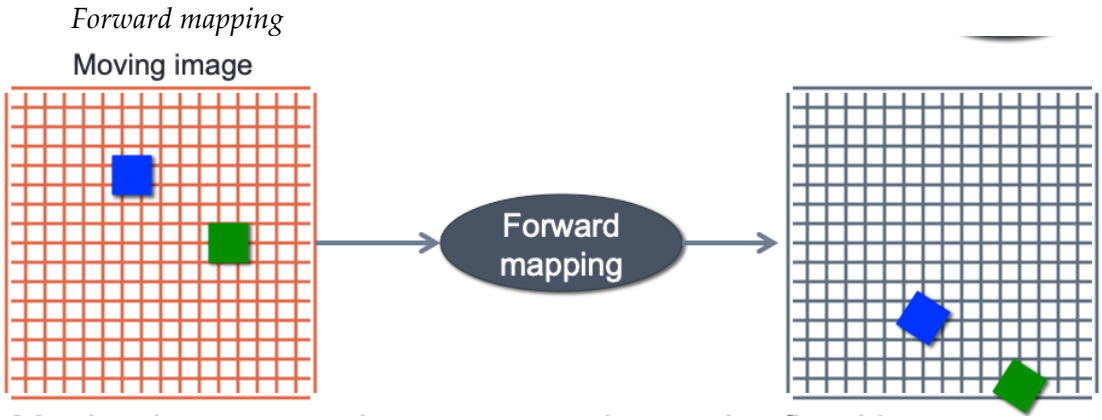


**Figure 2 Adapted from (Tosun-Turgut, 2010). The demonstration of forward mapping method. The left image with red grids is the moving image and the right image is the target. Blue and green square are the corresponding voxel intensities that map from moving image to the target image via forward mapping.**

The DVF can be defined either through forward mapping or backward mapping. This forward mapping method maps the moving image voxels onto the fixed image voxels by defining DVF in the coordinates of the moving image, as shown in fig. 1. If the target image is inflated, one moving voxels could correlate to multiple voxels on the fixed image. If the target image is deflated, multiple moving voxels could correlate to a single voxel on the fixed image. However, in such case, some voxels on the fixed images may have no mappings from moving image. Consequently, there can be holes on registered moving image, which is the major limitation of the forward mapping.

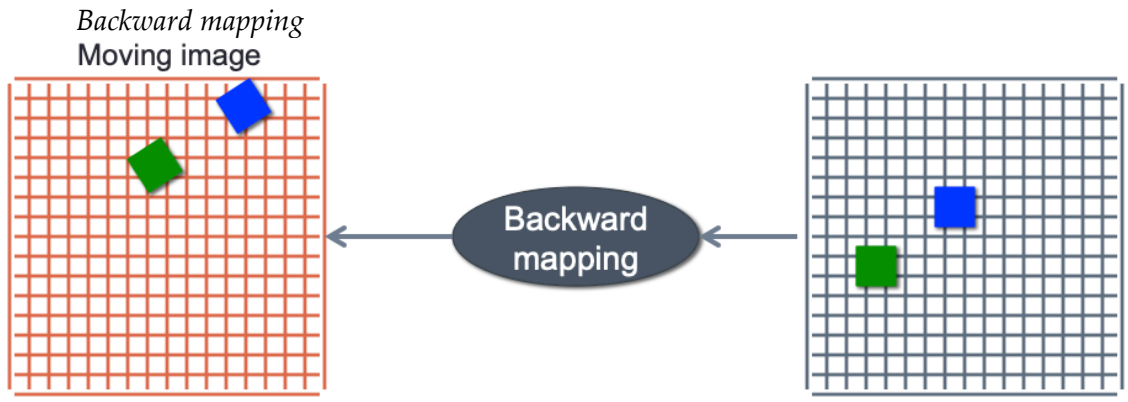


**Figure 3 Adapted from (Tosun-Turgut, 2010). The demonstration of backward mapping method. The left image with red grids is the moving image and the right image is the target. Blue and green square are the corresponding voxel intensities that map from fixed image to the moving image via backward mapping.**

The backward mapping method maps the fixed image voxels onto the moving image voxels by defining DVF on the coordinate of the fixed image, shown in fig. 3. All the voxels on the fixed image are scanned sequentially, thus the hole and overlaps in the output registered image are avoided. The output voxel values must be interpolated in the moving image. The image registration in the medical imaging is commonly done with backward mapping with intensity interpolation using interpolation functions such as nearest neighbor, linear, Spline or Sinc.

The fig.4 demonstrates how backward mapping avoid the hole by sequential mapping back and interpolation.

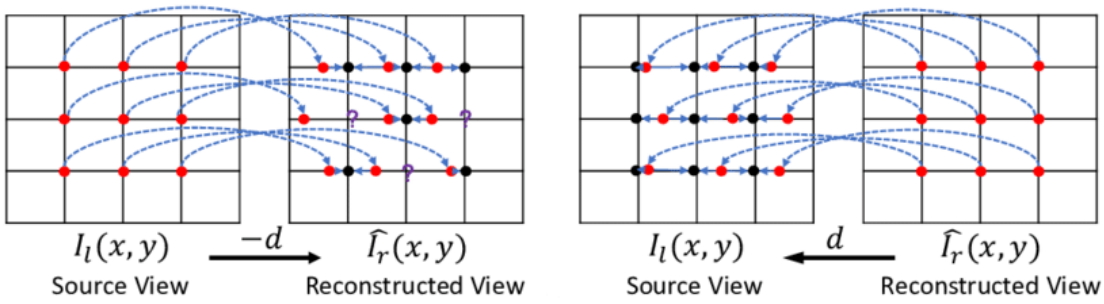


**Figure 4 Adapted from (Chen, Tang, & John, 2018). The $I_l$ presents the intensity of source image, $\widehat{I_r}$ presents the registered image, x y are the coordinates and d is the mapping function. The left one shows the forward mapping with interpolation, and the right one shows the backward mapping with interpolation. The question mark on the left one indicates the hole appearing in forward mapping.**

### 1.1.2 Classification of Image Registration

Image registration algorithms can be classified into two categories, intensity-based and feature-based. Intensity-based methods compares the intensity values of two images, while feature-based methods find the correspondence between features in the source image and target images such as landmarks, contours, or surfaces. Intensity based methods usually uses similarity matrix, such as the squared error, cross correlation or mutual information, to compare the intensity similarity between the target image and registered image during the registration. On the other hand, feature-based methods build a correspondence between a number of distinct points or lines on both images, then the transformation matrix for all voxels is generated based on the feature correspondence to map the source image to the target image. (Papademetris, Jackowski, Schultz, Staib, & Duncan, 2004)

### 1.1.3 Complexity of the Problem

Image registration is to map the moving images to the target image system; therefore, a geometric transformation matrix that minimizing the difference between deformed image and target image is required. This transformation matrix is usually described by its degree of freedom (DoF), which is the number of independent variables in the transformation matrix. Increasing the DoF allows more flexibility in the transformation to match the source images to the target image, but it also increases the complexity of the problem. (Barillot, Haynor, & Hellier, 2004)

If the source image is considered as a rigid body in the image registration, such registration is called rigid image registration (RIR). The transformation matrix for RIR has 6 degrees of freedom, i.e. 3 translations and 3 rotations. If the scaling and shear effects of source image is considered, it is referred as affine transformation and has 12 degrees of freedom. Further increasing the DoF leads to registration in the nonlinear regime, which is referred as deformable image registration (DIR).

For deformable image registration (DIR), the transformation matrix contains all transformations that do not fit into the affine transformation model. It covers the range that from nearly rigid transformation to the most general transformation which each voxel has its own displacement field (3 DoF). One could estimate that there are over a million DoF in a typical image volume. Finally, a vector map is used to show the

displacement of each voxel in the moving image, and this map is often referred as deformable vector field (DVF).

In order to have an idea of the complexity of million-DoF problem, let's see an example with only 18 degrees of freedom, three-body problem. It is a well-known problem in physics regime due to its simplicity while without an analytical solution. It has only 18 degrees of freedom, which include 9 displacements and 9 speeds of three bodies in a 3D space. Noted that each body has 3 degrees of freedom in space. Comparing with unsolvable three-body problems, the problem in DIR is even tougher and there is no analytical solution to it either. (Papademetris et al., 2004)

### 1.1.4 Applications

After obtaining the transformation matrix or the deformable vector field (DVF), one could have the registered moving image with various applications. For example, any image or contour, fluence map or dose distribution attached to the moving image can be registered to the target image. The application in the medical field will be mainly discussed here. The application of DIR in the field of radiotherapy could be categorized in to four fields: dose accumulation, functional imaging, automatic segmentation and mathematical modeling. (Oh & Kim, 2017) The function of DIR is to find the spatial correspondence between two considered image sets and serve the daily medical duties.

*Dose accumulation*

Dose accumulation is the first application of DIR in radiotherapy. Basically, dose accumulation uses the DVF to deform the dose map from planning stage to delivery stage, and the goal is to estimate the real dose delivered. (Yan, Jaffray, & Wong, 1999)

As we know, in planning stage, the dose distribution map was generated by dosimetrist and it is displaced on planning CT image volume. However, due to the inter- and intra-fractional motion or anatomic change during treatment course, the planning dose distribution is not the actual dose delivered. Therefore, the DVF is calculated from moving image (CT image in planning stage) to target image (daily CT image sets), and the real dose distribution across the treatment course can be estimated based on this DVF and the planned dose distribution. Furthermore, when the second treatment is needed for recurrent tumors, the DVF between the CT image of first course and CT image of second course will help deform the dose distribution of one course to the next to determine the accumulated dose to the patient to assess the toxicities.

*Automatic segmentation*

Automatic segmentation uses the DVF between to image sets to map the contour from moving image to target image. For example, during the treatment, usually the contour was done on the planning CT image sets, if the DVF between planning CT image (moving image) and daily CBCT image (target image) is found, all the contours on planning CT image can be transformed to daily CBCT image sets. It is a way to use pre-defined reference segmentation data. (Christensen et al., 2001) Furthermore, this

technique could be used on 4DCT/CBCT image sets to transform the contour on one phase to other phases to improve the efficiency of the process.

### 1.1.5 Limitation of Current DIR Methods

Deformable image registration (DIR) is regarded as a multi-dimensional (over a million degrees of freedom) optimization problem, which aimed at building the best correspondence between source image and target image (i.e. maximizing the similarity between both images) by optimizing the deformable vector field (DVF), regardless of intensity-based method or feature-based method (Cao et al., 2017) (H. J. Johnson & G. E. Christensen, 2002) (Ou, Sotiras, Paragios, & Davatzikos, 2011). The commonly used method requires iterative optimization and parameter tuning to estimate the deformation field between images (Cao et al., 2017).

In the case that the source image and target image has larger anatomic change, the performance of DIR declined significantly. It is because that the deformation from one image to another is large and the optimization procedures are more likely to be stuck in local minimums. If the optimization to find the DVF from scratch, it is problematic with many current DIR algorithms (Wu, Kim, Wang, & Shen, 2014).

The current DIR methods encounters certain issues:

1. Long computational time to estimate the deformation field and escape from local minima.
2. High vulnerability being stuck in local minima.

3. Performance depending on user tuning the optimization parameters.

The reference DVF of the learning is produced by VelocityAI, which uses B-spline method. The free-form deformation is regularized by introducing a smoothness term based on derivatives of the free-from deformation along with the image similarity term in the cost function. And the cost function is to be minimized to find the optimal transformation. The problem is that such regularization further limits the ability to model small localized warping.

### 1.1.6 The Novelty of This Work

This thesis aimed at tackling the problems presented in previous section by using the deep learning or convolutional neural network. The goals are (a) reducing the time of deformable image registration, especially for large anatomic difference between source and target image, (b) avoiding being stucking into local minimum by learning the registration from ground truth DVF, and (c) making the registration process user-independent and automatic without parameter tuning. Taking the advantage of deep learning, such as feature learning, several research groups has applied this method on DIR in different part of the body. For example, Cao et. al. tried to train CNN model on deformable image registrations on one set of brain image and successfully transferred to another data set, despite the high variability of brain appearance across different data sets (Cao et al., 2017).

In this work, similar concept is used for 4DCT and 4DCBCT lung image set. The CNN model is trained on one 4DCT lung image sets to learn the respiratory motion, and the feasibility of transfer such CNN model to another 4DCT or 4D-CBCT lung image sets are tested.

## 1.2 Convolutional Neural Network (CNN)

In deep learning, convolutional neural network (CNN) is a class of neural networks, most commonly applied to analyzing visual imagery (Wikipedia). Based on its literal meaning, CNN simulates some actions in human visual cortex by an input layer, output layer and various hidden layers. It can take an input image, assign weighting/importance to different aspects/features in the image, and therefore understand the input image better. Consequently, the CNN model can be applied in various aspects, such as image and video recognition, recommender system, image classification etc. The following figure shows a standard CNN architecture.

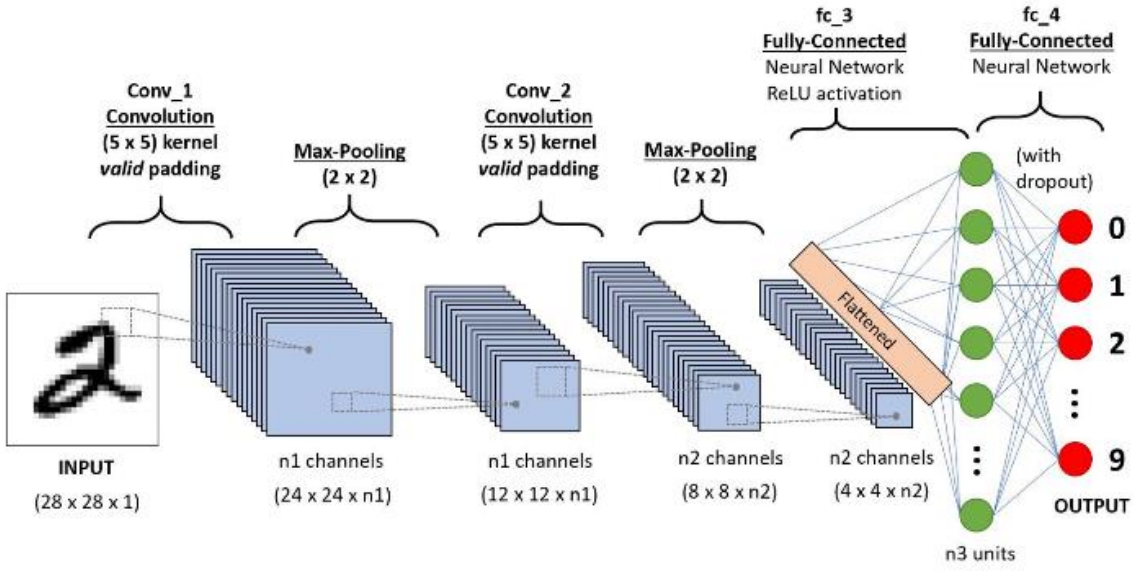


**Figure 5 CNN model to classify the handwritten digits.**

## 1.2.1 The Design of the networks

A convolutional neural network consists of an input layer, an output layers and several hidden layers, as shown in Fig.5. The input layer as shown in Fig.5 is the beginning of the network. It brings the prepared data into system for further processing by other hidden layers. The output layer is the last layer of the neural network.

The hidden layers of a CNN typically consist of convolutional layer, ReLu layers, pooling layers, fully connected layers and loss layers.

### 1.2.1.1 Convolutional layers

The definition of convolve in Latin words is to roll together. Mathematically, a convolution is the integral measuring how much two functions overlap with each other as one passes over the other. One could take convolution as a way to mix two functions by multiplying them. It is the core building of a CNN as the name of CNN suggests.

A set of learnable filters forms layer's parameters, and filters has a small receptive filed. As shown in fig.5, the first convolutional layer Conv_1 has n1 filters, meaning that the first convolutional layers in this CNN performed n1 convolution with input volume.

The convolutional layer works in this way. During the volume was passed from previous layer (could be input layer or other layers), each filter is convolved across the height and width of the input volume, computing the dot product between filter and the entry of the input images, and producing a new result called feature map. The feature maps shown in Fig.5 in the n1 channels. The following figure shows an example convolution.

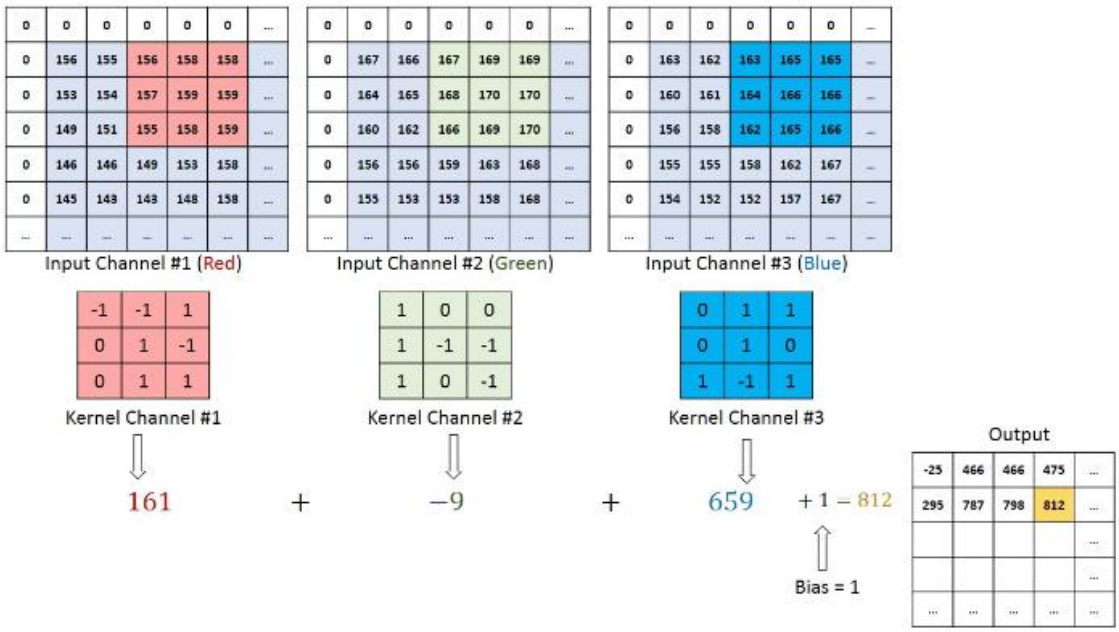


**Figure 6 The convolution operation on a three channels image with 3 x 3 filters on each channel. The output is the feature map.**

It is to be noted that here are two modes of convolution. As shown in figure 5, the input image has the size of 28 x 28. After convolutional layer with 5 x 5 filter with valid padding, it becomes 24 x 24. It is easy to understand that it is because the edge of the input image cannot be set as the center of the convolution. Therefore, the output size after convolution can be represented as

$$size_{input} - size_{filter} + 1$$

. If we want to size of image remain, we could use SAME mode to perform the operation. In this mode, the output has the same size as the input by padding the input image with number 0. The size of the input is enlarged; therefore, the size of output is the same as input. The following figure shows the example of SAME mode convolution.

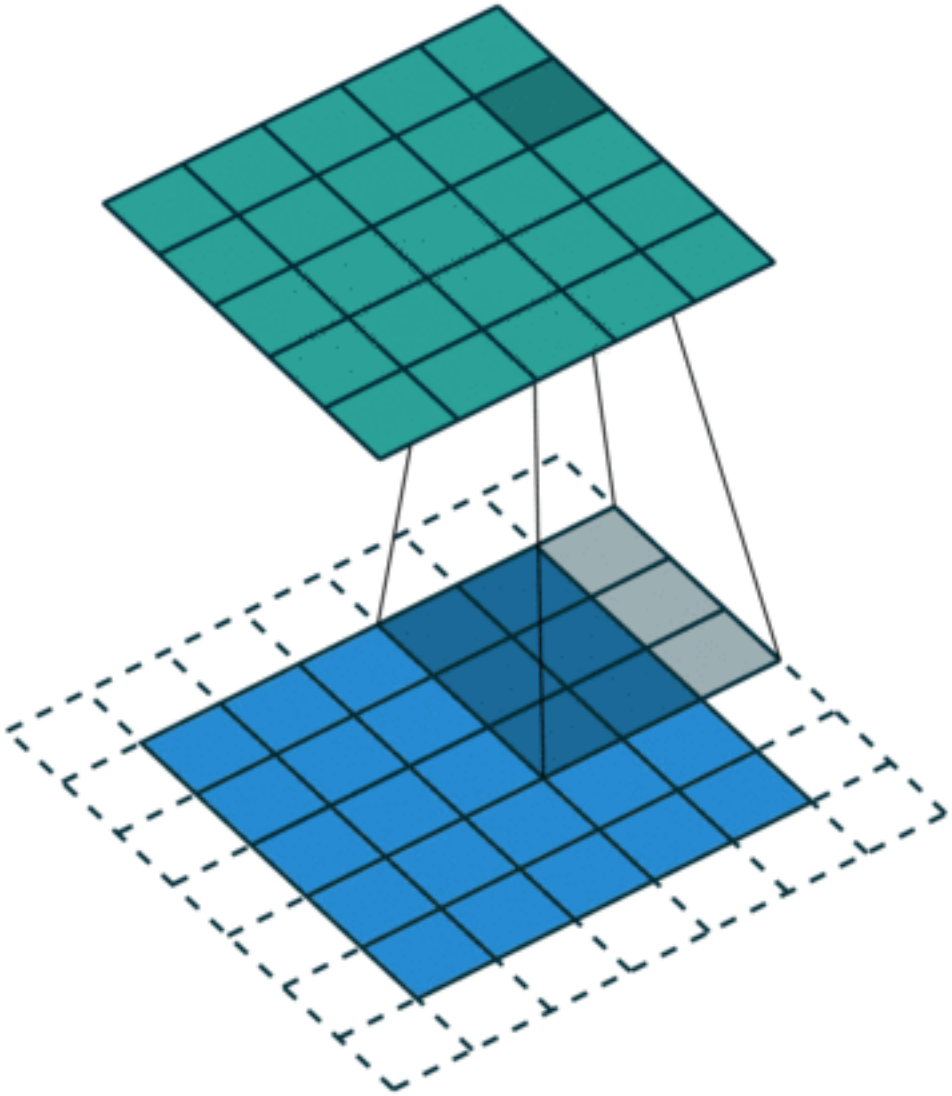

**Figure 7 Adapted from (Ian, Yoshua, & Aaron, 2016)The convolutional operation with SAME mode. The green patch is a 5 x 5 output feature map, the blue patch is a 5 x 5 input image, and the filter is the shadowed are with size of 3 x 3. The dashed line around the input images are the padding, the edge of the input image, therefore, can locate at the convolve center.**

This procedure could be seen as a feature extraction mechanism. The

convolutions of different filters were executed on input image, and at every location, the

dot product happened and the sum of the result forms the feature map. This feature extraction procedures are the key to CNN model.

The first convolutional layer is responsible to capture or extract the low-level features such as edges, gradient orientation, color etc. With more convolutional layers involved, high dimensional features could be extracted.

#### 1.2.1.2 Pooling layers

Pooling layers are responsible for reduce the dimensions of the extracted features. It is a form of non-linear down sampling. It divides the input image into a set of non-overlapping region, and for each sub-region, yields an output based on the pooling criteria. Generally, there are two ways of pooling, max pooling and average pooling.

Max pooling yields the maximum value of that sub-region. It extracts the dominant feature in each sub-region. The average pooling yields the average value of the sub-region. Max pooling can also perform as a noise suppressor, since it selects the dominant value in each region. On the other hand, the average pooling simply reduces dimension to suppress the noise. The following figure shows the example of max pooling and average pooling.

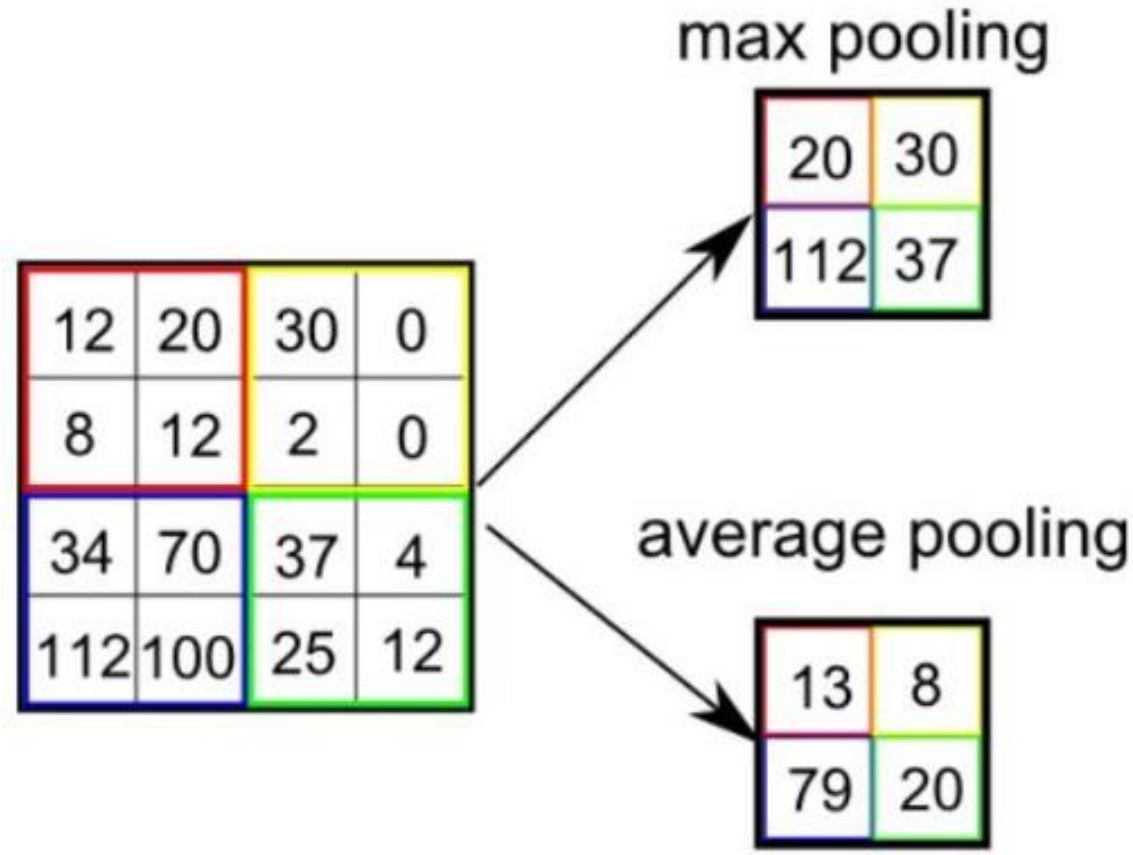


**Figure 8 Adapted from (Ian et al., 2016). The example of max pooling and average pooling. The left image with size of 4 x 4 is the input, the pooling layer size is 2 x 2, the right top image is the output after max pooling and the right bottom image is the output after average pooling.**

The pooling layer emphasizes the feature by blurring its exact location, since intuitively, the relative location of features is more important than the absolute location of the feature. Taking car recognition as an example, the CNN will recognize the car in several photos, the car will not be at the same position of all photos. If the car can be seen as the combination of several features, the relative position of these features is more important.

The main feature of pooling layer is to reduce the size of the input image. With the reduction of the dimension, the number of parameters and the computational time can be reduced. Therefore, it can control the overfitting problem and shorten the training time.(Hutchison et al., 2010)

#### 1.2.1.3 ReLU layers

ReLU is the abbreviation of rectified linear unit. It is the activation function in CNN model and it is often followed with convolutional layer to add the non-linearity on the feature map. The non-linearity features could be observed in lots of images, such as the color transition in pixels and the borders. The function of ReLU layer is to further introduce such non-linearity.

The ReLU function represents $\psi(x) = \max(x, 0)$, and the graph of the function shows in figure 9. It replaces the negative value with 0, therefore increases the non-linearity.


**Figure 9 The graph of ReLU function. It is a half linear function. All the input values which is smaller than 0 are set to be 0 and the values greater than 0 remains.**

With the increased non-linearity, the linearity that introduced by convolutional operation is compensated. In other words, with increased non-linearity, the feature on the image is more outstanding and easier to be captured. Therefore, the training time will decrease and the performance of the model will increase. (Krizhevsky, Sutskever, & Hinton, 2017)

There are some functions such as hyperbolic tangent and sigmoid function to increase the non-linearity, but the ReLU is often preferred over the others due to its faster in training speed without compromising the generalization accuracy. (Krizhevsky et al., 2017)

#### 1.2.1.4 Fully connected layers

Fully-connected layers is usually the final layer before the output layer. Fully-connected layers learns the non-linear combinations of high-level features. The high-level features are the output of several convolutional layers and pooling layers. As shown in Fig.5, the line of the green circles presents the fully-connected layers. Each green circle is a neuron which connects to all the activations in previous layers.

In classification problem, the fully-connected layer takes an input volume and generate an N-dimensional vector, where the number N is the number of classes we want to classify. If we want to have a *yes*/*no* model, it will yield a 1x2 vector, and the first value in this vector represents the probability of having a yes and the second value in this vector represents the probability of having a no.

Basically, the fully-connected layers search the correlation between high-dimensional features and the class in the output. If there is only class in the output, then it functions like a regression model to find the only relationship between the input and the only output.

### 1.2.2 Model training

The training part is one of the most important part in any neural network, since it shows how the model work. As mentioned in the beginning of this section, the weights or the importance is assigned to each part of the image. Thinking in this way, if it is about to tell the difference between the dog and bird, the weights may heavier at the

parts that distinguish them, such as the beak and wings of the bird and the tail of dog. If the model could find such features and assign them with an important weight, the classification can be done. The backpropagation is used to determine the importance of the features. .

#### 1.2.2.1 Backpropagation process

Backpropagation has four parts, forward pass, the loss function, the backward pass and the weight update. Let us look at each part.

*Forward pass*

As suggested in the name, during the forward pass, the training image was sent to the network. The weights at this stage is initialized, which means it has no preferences to any features in the image. At this moment, the model cannot classify anything or draw any conclusion about the input image, even though there will be an output based on initial weights. Then, it passes to the next stage.

*Loss function*

Even though the image goes through the network and the network cannot determine anything yet, besides the training input data, each input image has a corresponding label. This label is the target that we want our network to learn. Intuitively, the loss function is defined as to minimize the difference between the output and the target by adjusting the weightings. The most commonly used loss function is the half-mean-square-error (HMSE), and it can be described as the following equation, $Error = \sum \frac{1}{2}(target - output)^2$, the graphs of the lost function is shown in next figure.

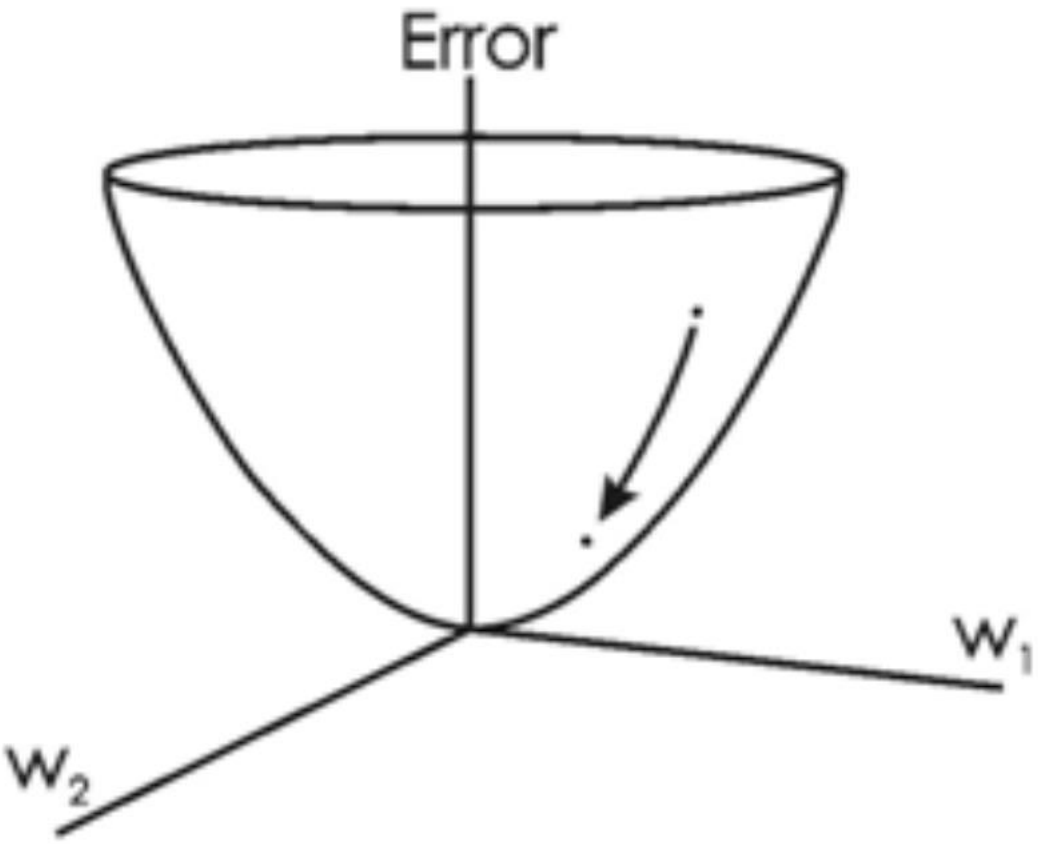


**Figure 10 Adpated from (Ian et al., 2016). The graph of an example visualizing a loss function. The error axis is the value of the error function, the $w_1$ and $w_2$ are the weights of a specific layer, and the axis of w represents the different layers in the network. The arrow indicates that the goal is to minimize the error.**

It is expected that the loss is large at the beginning of the training process since the predicted output is yielded from some random initial weights.

*Backward pass*

Backward pass is to find the weights that contribute most to the loss function. It is the common optimization problem in calculus, and the change of the loss function along each layer can be obtained with the derivative, represented in $\frac{\mathrm{dError}}{dw_i}$, where the

Error is the loss function and the $w_i$ is the ith layer. The backward pass is basically to go backward and find the contribution of each layers.

Taking dog and bird classification problem as an example, the backward pass process will find that the layers representing the dogs' tail and birds' beak having the larger contribution to the loss function.

*Weight updates*

Weight updates replaces the old weight by the new weight that has been found in backward pass. The amount of change of the weight is determined by the learning rate. The to-be-updated weight has the form

$$w = w_{initial} - \eta \frac{\partial Error}{\partial w}$$

where w is the final weight, $w_{initial}$ is the initial weight, η is the learning rate, and Error is the loss function.

The learning rate is a parameter set by users. Smaller value means slow training but better precision on optimal point. If the learning rate is too large, the loss function may change dramatically and miss the optimal point (it is also called overshoot).

**1.2.2.2 Iteration**

The training stage of the network essentially is an optimization problem, which minimizes the difference between target label and output label by adjusting the weights of each layer. The four parts of backpropagation process is called one iteration. During the real training, there will be a group of training data input in the network, and such

group of data is called a batch. A fixed number of iterations are performed in one batch of data, and at the moment, the network updated the weights for this batch. The networks actually learn the weights. Then, the second batch of data is fed into the system. Based on the updated weights in the first batch training, the weights of layers updated once again to best fit the second batch. With larger batch size, the network's performance is more general over all samples. (Hinton, Osindero, & Teh, 2006)

### 1.2.3 Network Testing

The testing stage is important to see whether our network works. The testing data is another set of data which different from training data, and it also has the images and labels, and this label is the ground truth which verifies the prediction result. The testing images are fed into the network and it yields the predicted label. By comparing the predicted label and ground truth label, the performance of the network could be evaluated.

### 1.2.4 Advantage of CNN

There are lots of advantages of CNN comparing with Neural Network (NN) or other conventional methods for a certain task. The most fundamental advantage is automatic feature extraction for a given task (i.e. distinguishing the birds and dogs), provided that the input can be represented as a tensor where the local elements are correlated with each other. (Ian et al., 2016) In other words, the CNN is a feature extractor which can learn the features from the input data such as image, audio and

video, and it can serve as a classifier. This is the main reason that the CNN has a wide application on several fields such as military, medical etc.

Besides the feature learning ability, CNN is also effective and accurate in certain task such as classification problem with the proper training to the network.

## 1.3 Respiratory Motion Prediction with CNN

The respiratory motion has long been a tough problem in treatment planning and irradiation delivery to lung cancer patient. Due to the breath-in and breath-out, it is very difficult to only irradiate the tumor without hurting surrounding normal tissue. The probability of having normal tissue complexity is high. Therefore, four-dimensional computed tomography (4D-CT) and four-dimensional cone-beam computed tomography (4D-CBCT) are proposed to provide the position of tumor and inner organs, and provide the information on respiratory pattern for easier tracking of the target in radiotherapy.

### 1.3.1 The Principle of 4DCT

The 4D-CT procedures usually require that the machine works around the patient at different angle for several breath circles. (Vedam et al., 2003) Besides having the CT images, it also requires a simultaneous measurement of internal or external agents that represents the breath patter. Such agents representing the breath patter could be the amplitude of the image data, or the organ position. Then, the images taken in several breath circles are mapped to the breath patter, sorted into different phases (usually 10 phases) and reconstructed to conventional CT volumes.(Ford, Mageras, Yorke, & Ling, 2002) Basically, 4D-CT volumes are ten sets of 3D-CT volumes, and each set corresponds to a respiratory phase. Usually, phase 1 is the end-of-inhalation and phase 6 is end-of-exhalation.

### 1.3.2 Application of DIR on 4D-CT Volumes

With the 4D-CT volumes representing the respiratory motion, several things could be done with DIR, such as the modeling of the respiratory motion which can further serve the online image guided radiotherapy (IGRT) and auto-segmentation from one phase to others.

Yang et. al. (Yang et al., 2008) uses 4D-CT volume data and B-spline based deformable image registration method to compute the respiratory motion and finally to build a motion model. With the motion model of the patient, one could tract the tumor and organ at risk in an effective manner, deliver irradiation more accurate and alleviate the normal tissue toxicity.

Ehler et. al. (D Ehler, Bzdusek, & Tome, 2009) uses 4D-CT volumes and rigid image registration as well as deformable image registration to automatically contour the gross tumor volume (GTV) on all the phases of 4D-CT with the help of Pinnacle (3) v8.1. With the accurate auto segmentation method, the physician could save time on patient.

The above two examples show the application of DIR and 4D-CT volumes. With further development of DIR techniques by increasing its speed and accuracy, the real-time target tracking could be done and real-time image-guided radiation therapy is possible.

### 1.3.3 Advantages of using CNN for DIR

With the advanced development of hardware in the computer, the computational power of computer has dramatically increased. The general training process of CNN can be done within dozens of hours. The applications of CNN have been exploited in different fields. Medical imaging is one of those fields. More importantly, the efficiency of prediction of CNN is extremely fast. As discussed in section 1.2.2.1, the prediction process is only the first step of backpropagation, forward pass, and there is no iteration process involved in

We have discussed the current issues and limitations in conventional deformable image registration in section 1.1.5. One of the major limitation is the efficiency of deformable image registration. The performance of the conventional registration relies on several factors such as the anatomic and inter-modality difference between source and target image, the experience of user, and parameter tuning. And the registration process is also a time-consuming process since the iteration and optimization has to be done for every change in parameters.

For CNN, on the other hand, the optimization and iteration has been done in the training stage. The speed to do the registration is, therefore, fast and efficient. Furthermore, image registration simply is to map one image on the other, and the registered image need to be similar with target images. In other words, for clinic images, the land marks such as the bone, the artery, the boundary of the organ etc. has to be

match well after registration. Such feature learning and extraction is exactly the fundamental advantage of the CNN. Therefore, CNN has the potential to perform deformable image registration in an effective manner.

### 1.3.4 How to Apply CNN on DIR

As discussed in section 1.1.1, we realized that the DIR is an ill-posed problem with millions of degrees of freedom, which next to impossible to solve with an analytical solution. The current methods such as Demons-based method and optical-flow based method are done in an iterative way, which is time consuming. (Li et al., 2017) The key component of deformable image registration is the deformation field, which tell each pixel where to go and match with the target image. Each pixel could be understood having three degrees of freedom. If the DVF is represented in the form as $\phi$, and at position $i$, the DVF is $\phi_i = [d_x, d_y, d_z]$, where $d_x, d_y, and\ d_z$ are the displacement vector of pixel $i$. If we only consider one pixel, the million-dimensional problem were reduced to the one with only three degrees of freedom. The problem has been extremely simplified. And it is also possible to obtained the DVF at each pixel, which can serve as the label. This DVF, unfortunately, has to be obtained from conventional DIR algorithms, even though they are not accurate. The further question will be, by only knowing one point, it is still impossible for CNN to learn the features, therefore we need a patch to provide the corresponding feature between source image and target images. An example shows below,

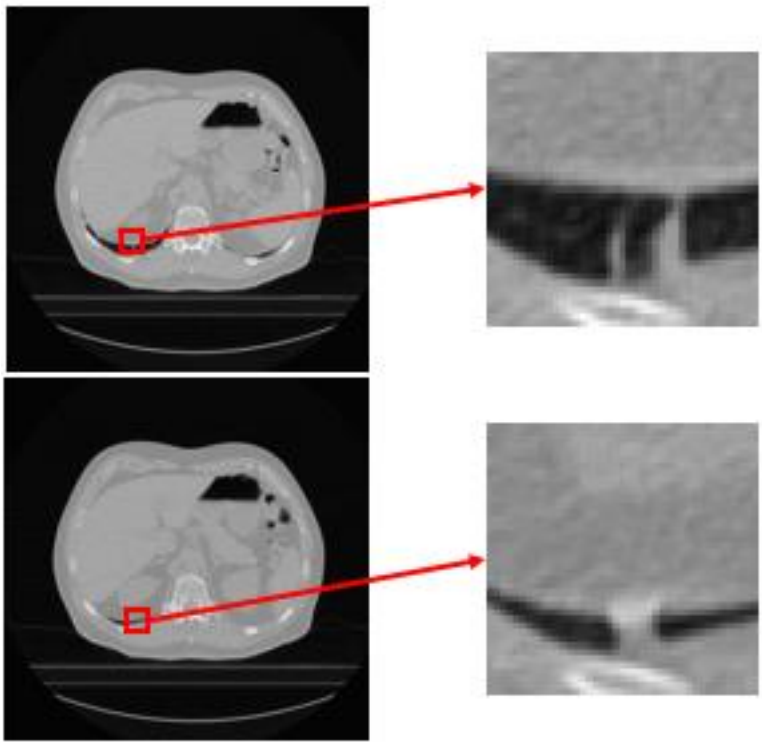

**Figure 11 The left two graphs are the axial slice of 4D-CT volume. The top one is the phase 1 and the bottom one is phase 6. The local patches are extracted from the left side. The local diaphragm movement is represented clearly on the patch.**

As shown in Fig. 11, a 4D-CT volume could be cut into millions of small patches, and it will provide sufficient training samples feeding the CNN. Each patch pair will also provide the network with feature movement information.

The next question is what we want the CNN to learn. The answer is respiratory motion in 4D image volumes. Therefore, the registration is to map one phase in 4D-CT to other phases. For example, if we had two sets of 4D-CT for the same patient. One set of the volume can serve as training sample, and another set could serve as the testing sets. Intuitively, the application if this could be that physician has contoured the tumor

and OAR in one phase on the testing set, and these contours could be transferred to other phases via the DVF predicted by the CNN.

# 2. Method

This section presents the framework predicting phase-to-phase DVF of lung 4D-CT/CBCT via the training of convolutional neural network (CNN). First, the samples for both training and testing were made. Each sample consists of a patch pair of moving and target image, and their corresponding DVF at the center of the patch. Then, a CNN architecture was built and trained with training samples. Finally, with the well-trained network, we predicted the DVF of testing set and compared with ground truth DVF for evaluation. Besides, the dense deformation field is interpolated with thin-plate spline interpolation, and the registered images from both predicted DVF and ground truth DVF are reconstructed and compared. The third-party commercial software *VelocityAI* was used to generate the ground truth DVF. The *MATLAB2018a* was used to process the data and train the network.

All the coding applications such as training, testing, and sample preparing are implemented with MATLAB2018a on a PC (Intel Xeon CPU, 32GB RAM) using a single NVIDIA Tesla K40s GPU with 12GB memory.

## 2.1 Sample Sets Preparation

This section presents the detailed methods preparing the samples.

### 2.1.1 Mathematical Description

In image registration, there are moving image $M$, target image $T$, registered image $M'$ and the deformable vector field $\phi$. The patches extracted represent as a group $\{P_M(u), P_T(u)\}$, where $P_M$ is the patch of moving image, $P_T$ is the patch of target image and $u$ is patch extraction points. The deformation field at the position $u$ represents as the displacement vector $\phi(u) = [d_x, d_y, d_z]$ for position u. Combining of both is one sample at position $\{P_M(u), P_T(u)|\phi(u)\}$.

### 2.1.2 Patch Extraction Position Selection

The patch extraction position determines the feature included in the patch and magnitude of the deformation vector. It is preferred selecting the points with large displacement or on the feature position such as the diaphragm, vessels, the lumen and the tumor boundary, because these features are the representation of the respiratory motion, such as the contraction and relaxation of the diaphragm. With the input of meaningful samples and supervision of corresponding DVF, the network is able to learn the respiratory motion.

In order to simplify the process, the patch extraction point was chosen to be a uniform 3-dimensional grid across the lung. The density of the grid determines the number of sample. In order to make sure the sufficient training samples, we chose the density such that the number of samples to be around 20,000. The density for the superior-inferior (SI) direction was chosen to be denser than anterior-posterior (AP)

direction and medial-lateral (ML) direction, because the magnitude of deformation field along SI direction is generally larger than the other two and we want to emphasize the learning of the CNN along this direction.

### 2.1.3 Patch Size Selection

The patch is a 3-dimensional volume around the extraction point. A pair of patches from moving and target images must contain the enough information for the network to learn. If the patch size is too small, the CNN will not learn the feature movement from moving patch to target patch. The registration, therefore, will fail. If the patch size is too large, the feature extraction process of the CNN will be slow and therefore low learning efficiency. Besides, the physical memory will be occupied more, compromising the number of samples. Therefore, we chose the patch size that at least covers the majority of the feature movement.

## 2.3 The CNN Architecture

The CNN architecture is shown in figure below. There are four convolutional layers, the kernel number for each layer is 64, 128, 256, and 256, and the kernel size is 3 x 3 with stride [1 1]. The convolution between kernel and input is in the 'same' mode, thus the size of the output kept the same as the input after each convolutional layer. Each convolutional layer is followed by a batch normalization layer and a ReLU activation. The first two convolutional layers are followed by an average pooling 2-D layer with size of 2 x 2 and stride of [2 2]. The fully connected layer with output size 3 were set

behind, connecting the output deformation vector $\phi(u) = [d_x, d_y, d_z]$. The loss function is half mean squared error.

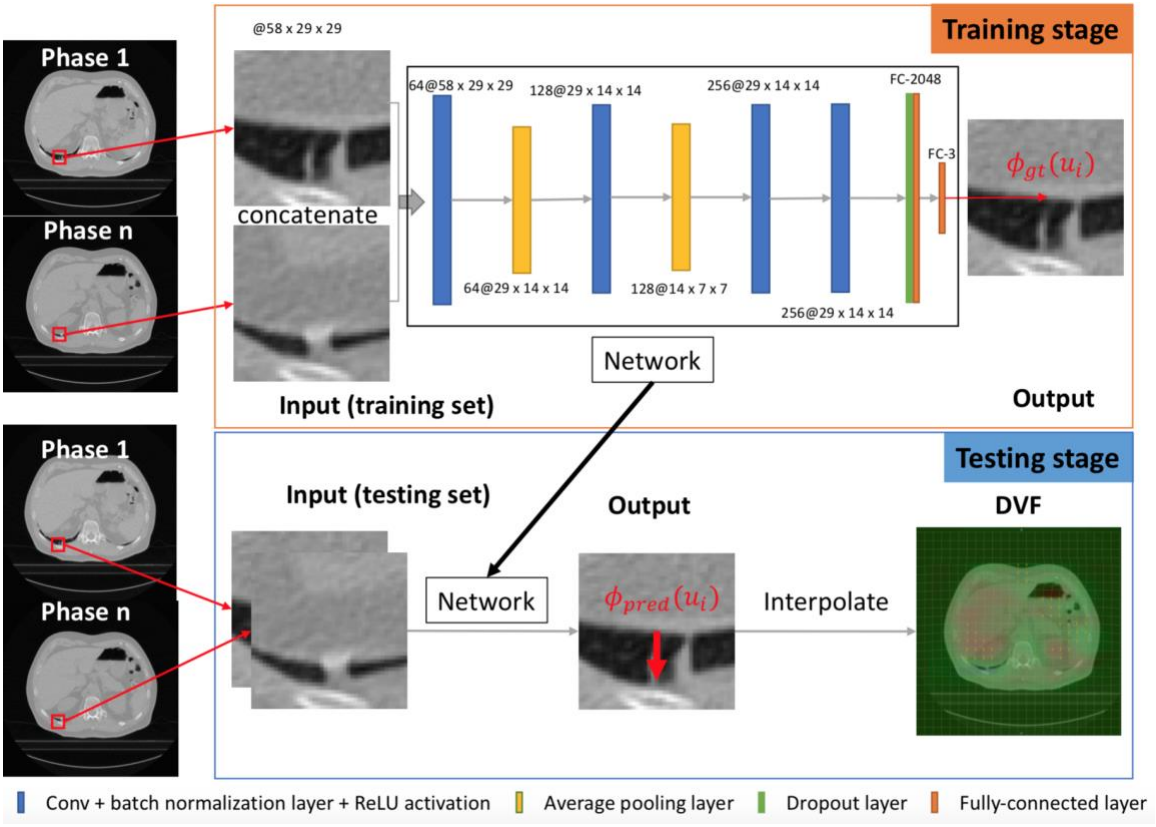


**Figure 12 The CNN architecture and the general workflow of training and testing. The two images on the left is the coronal view of phase 1 and phase n of the 4D-CT volume. The orange box shows the training stage. The patch pair were made as the input of the CNN. The network architecture is in the black box and the ground truth DVF $\boldsymbol{\phi_{gt}(u)}$ supervised the training. The blue box shows the testing stage. A testing data set is the input of the network, and the output is the predicted DVF $\boldsymbol{\phi_{pred}(u)}$. Then, the deformable vector field map could be interpolated.**

## 2.4 The Experimental Design

The training samples feeding the CNN determine the feature movements network is going to learns. The feature movement in 4D image data sets are represented by the anatomic movements mainly due to patient's respiratory. The image quality of

the anatomy, therefore, determines the learning (predicting) quality of the network. It is because with high image quality, the anatomic movement is more outstanding, the network can extract the feature easier, and our ground truth DVF is more precise. Therefore, different 4D image volumes with different image quality (i.e. 4D-CT and 4D-CBCT volumes) were used as the training set and the performance of prediction is evaluated.

The testing on different image volumes also yield different result and the performance of the CNN is different. Intuitively, sharing the similar anatomy in training and testing sets could guarantee a decent performance on predicting DVF. For example, the training and testing images are from the same patient and taken at the same time.

The combination of the testing and training set is, therefore, of importance. One the one hand, the performance of prediction must maintain to a certain level, since fast predicting the DVF is the main goal of this work. On the other hand, the clinic application must be considered. Training and testing on the same image sets may have decent prediction performance, but it loses the practicability.

### 2.4.1 One 4D-CT volume

We started with the most ideal situation to initial assess whether or not the idea works. In this case, the training and testing were done with only one set of 4D-CT volume. 4D-CT volume provided us with high image quality with clear anatomic structures and features in the lung. Using one data set as both training and testing

prevented from the inter-fractional anatomic changes (i.e. mainly the respiratory motion), and the network will be tested on the images with identical anatomy but different respiratory phases. This experiment will show if the network has the ability to learn the respiratory pattern.

In this case, the training sample can be represented in the form

$$\{P_{phase\ 1}(u), P_{phase\ n}(u) | \phi_{1\ to\ n}(u)\}$$

, where $n$ is the other phases.

### 2.4.2 Double 4D-CT volumes

We have two sets of 4D-CT volumes for the same patient in this case. One set of the 4D-CT was used as the training set, and another one is testing set. This case is more challenging than previous case because the inter-fractional anatomic changes were introduced. The bright side is more clinic usage of this could be exploited. For example, the CNN learnt patients respiratory pattern with training on the 4D-CT in planning stage, this network could also be used for automatic delineation on second 4D-CT volume. 4D-CT volumes also provide us with high image quality and is easier for the network to learn.

### 2.4.3 Double DRR Simulated 4D-CBCT

We now pushed our network to not only adapt the inter-fractional anatomic change but the cone-beam modality. In this case, two 4D-CBCT volumes were reconstructed with FDK method using 340 the DRR simulated projections per phase

from the double 4D-CT volumes in previous section. The training and testing were followed the same procedures as section 3.4.2.

### 2.4.4 4D-CT and 4D-CBCT

In considering double 4D-CT is not the common practice in clinic and the 4D-CBCT is the one used in daily procedures for verifying the position of the patient and the target, the testing set was changed to 4D-CBCT volumes for wider practicability. Due to the lack of 4D-CBCT volume in our facility, the volumes and projections provided by Spare challenge were used. In this case, the 4D-CT volume (i.e. The Spare Challenge ground truth volume) is the training set and 4D-CBCT volume is the testing set. The 4D-CBCT volume is reconstructed with FDK algorithm using 680 primary projections per phase.

### 2.4.5 Double FDK-Reconstructed 4D-CBCT

The performance of the network with different training and testing image quality was tested, in order to understand the impact of image quality to the network performance. The training 4D-CBCT volumes for the same patient are reconstructed with FDK using 680 projections per phase, while the testing 4D-CBCT was reconstructed with 170 projections. In this case, besides the respiratory motion, the network needs to deal with two challenges: the first one is the image quality and the second one is inter-fractional anatomic change. Finally, we found that image quality limits the learning

efficiency of the network, and the inter-fractional anatomic change limits the prediction efficiency.

### 2.4.6 Multiple 4D-CT Volumes from Different Patient

For further application of the network, a data base that contains multiple patients' respiratory pattern is proposed and built. In other words, the network learnt the different respiratory motion from multiple patients, and it can be used to predict the phase-to-phase DVF of new patient. It requires that the network could 'conclude' a general respiratory pattern fitting to new patient. The most challenging task for the network is to adapt to a new patient's anatomic features. There are total six patient 4D-CT volumes, five of them were used as training and the left one is the test.

## 2.5 The Evaluation methods

The direct result is the output of the network after feeding with the testing sample sets. This output is the predicted deformable vectors in the selected points $\{\phi_{pred}(u_i)\}$, where $u_i$ is the $i^{th}$ patch center and $i$ is the index of the patch. For the testing sets, we still have the ground truth label, in other words, we have the ground truth deformable vectors in the selected points $\{\phi_{gt}(u_i)\}$, where $u_i$ is the $i^{th}$ patch center and $i$ is the index of the patch. The coefficient of correlation between the prediction and the ground truth were calculated. The coefficient of correlation is a value between -1 and +1, which measures the strength and direction of a linear relationship between two variables on a scattered plot. In this work, the value closer to 1 means better prediction

result. Noted that only the phase 7 was reconstructed and shown in this paper, because phase 7 is generally in end-of-exhalation phase which is out of phase to phase 1. The magnitude of deformable vectors from phase 1 to phase 7 is the greatest.

The dense deformable vector field is interpolated from the predicted and ground truth $\{\phi_{1to7}(u_i)\}$ with thin-plate spline interpolation, and both applied on the moving image which is the phase 1 of each image set. The registered images with predicted DVF and ground truth DVF are compared.

# 3. Result and Discussion

This section presents the performance of our network in predicting the phase-to-phase deformable vectors, the coefficients of correlation for each testing are tableted, and the registered images with ground truth and predicted deformable vector field is presented. All the results are discussed in this part.

## 3.1 One 4D-CT Volume

Several combinations of training set are tested in this case.

### 3.1.1 Coefficient of Correlation

All the combinations along with the coefficients of correlation in predicting the deformable vectors of phase 7 are listed in table below.

**Table 1 The training set combination and the coefficient of correlation in predicting phase 7 for the cases of single 4D-CT volume.**

| Training set Combination | Coefficient of correlation* | | |
|---|---|---|---|
| | AP# | ML# | SI# |
| A | 0.917 | 0.932 | 0.984 |
| B | 0.871 | 0.886 | 0.972 |
| C | 0.807 | 0.848 | 0.959 |
| D | 0.876 | 0.883 | 0.968 |
| E | 0.904 | 0.916 | 0.968 |

* The value is between -1 and 1. In this work, larger value means better prediction result.

# AP: Snterior-posterior, ML: Medial-lateral, SI: Superior-inferior

A: using the phase 2, 4, 6, 8, 10 as the training sets

B: using the phase 2, 3, 4, 5 as the training sets

C: using the phase 2, 3, 4 as the training sets

D: using the phase 2, 6, 10 as the training sets

E: using the phase 6 as the training sets

The best predicting of deformable vectors is the use of more training set, which is the combination A, as shown in table 1. This is reasonable since there are sufficient samples for network to learn. And the respiratory amplitude in phase 7 are quite similar with phase 6 and phase 8. Both phases were in the training sets, which provided the network with comparable reference in predicting phase 7. Overall, the coefficient of correlation for all directions shows very high correlation and it is expected that the registered image with the prediction DVF will comparable with the image registered with ground truth DVF.

The training sets is reduced in phases for combination B, and C, and phase 7, the prediction phase, has larger respiratory phase different from training set. The AP and ML directions showed a noticeable decrease in the correlation between ground truth and prediction, while the SI direction maintained a high correlation. The training samples in combination C has larger divergence to testing set comparing to combination B, the prediction performance, therefore, is a little bit worse than combination B.

The combination E, which the training samples are only from phase 6, outperformed combination D in AP and ML direction, which the training samples are from phase 2, 6 and 10. It can be concluded that the sample number is not the deterministic factor for the prediction power of the network but the quality of the samples. The coefficients of correlation of other phases in combination E are also been

calculated, and it also shows the trend that larger difference in respiratory phase between training and testing samples leads to poorer performance of the network.

However, the general result for prediction are pretty good for these cases. The general respiratory pattern can be predicted with feeding the proper samples to the network. In this case, both the training and testing samples are from same 4D-CT image set, and the network only needs to deal with the respiratory motion. In predicting the large respiratory motion (i.e. end-of-inhalation to end-of-exhalation), the high correlation was found in prediction and ground truth for all combinations. Furthermore, the SI direction, which has the largest movement in respiratory motion, maintained a very high correlation in all combination of training sets.

### 3.1.2 Comparison of the Registered Images

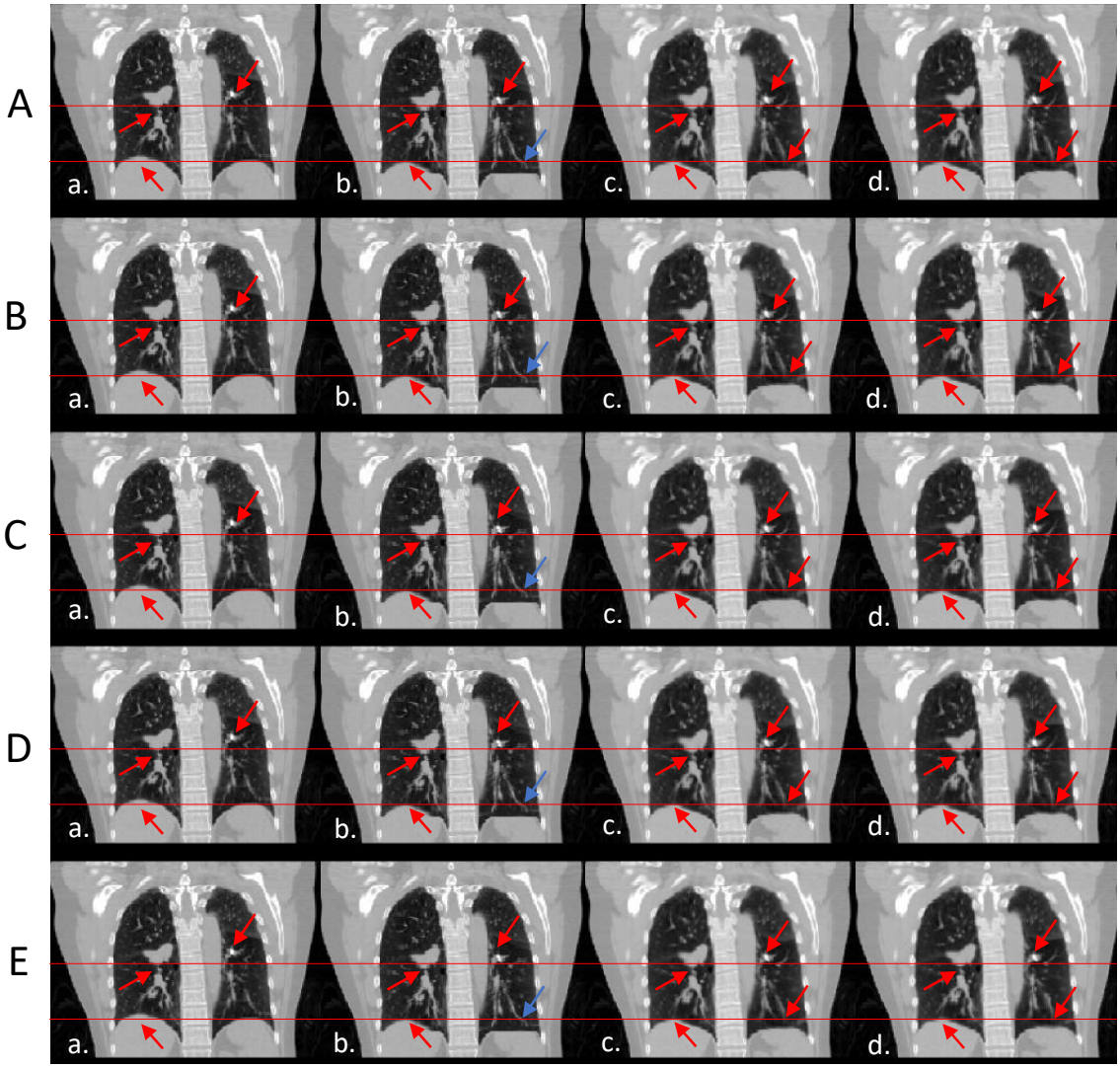


**Figure 13 Showing a coronal slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with prediction DVF (d). The horizontal lines were drawn for reference purposes, red arrows indicate several main structures, and the blue arrows indicates the structures that are not matched well. The A, B, C, D and E are the combinations in table 1.**

As shown in fig 13, the fiducial markers, the bottom boundary of the mass and the right diaphragm shows a very good match for all the training combinations. It is

further proved that the respiratory motion can be captured and learnt by the network. Besides, in figure 13, the blue arrows indicate the disagreement in target image and ground truth image by the inaccuracy of conventional deformable image registration algorithm. The target image also shows disagreement with predicted image, while the prediction agrees with the ground truth. It is because the training label or the training target is mis-registered in blue arrow area at the first place, and the network learnt the wrong pattern here. When it was used as prediction, the network persisted the wrong pattern. This disagreement in target image and registered image, on the one hand, indicates that our network did learn the feature pattern from the training samples. On the other hand, it revealed that the quality of ground truth affects the prediction accuracy of the network. It is a flaw of this work, but due to the limitation of the existing DIR algorithm, the perfect ground truth is impossible.

### 3.1.3 Discussion

Interphase learning and prediction of respiratory motion didn't have much practicability in clinic. However, its promising result did encourage the continue of this work. Furthermore, it did show a key characteristic of DVF prediction with CNN. High similarity in learning and testing samples has better performance in prediction DVF. It is intuitive that the network will have the best performance when training and testing samples are identical.

## 3.2 Double 4D-CT Volumes

In this case, the training volume and the testing volume are from intra-patient image sets of different 4D-CT volumes.

### 3.2.1 Coefficient of Correlation

The coefficient of correlation is shown in table 2.

**Table 2 The coefficient of correlation between the predicted deformable vectors and the ground truth deformable vectors in predicting phase 7.**

| Coefficient of correlation | | |
|---|---|---|
| AP | ML | SI |
| 0.581 | 0.673 | 0.884 |

The coefficient of correlation showed a decent predicting performance in end-of inhalation (phase 1) to end-of-exhalation (phase 7) motion. The predicting performance in SI direction is the best, and in AP direction is the worst. The reason could be that the magnitude of DVF in SI direction is the largest, and in AP direction is the smallest. With larger movement, the input patch pair showed a larger contrast in respiratory motion, and this contrast is captured easier by the network. Therefore, the network showed greater learning efficiency and predicting power.

### 3.2.2 The Registered Image Comparison

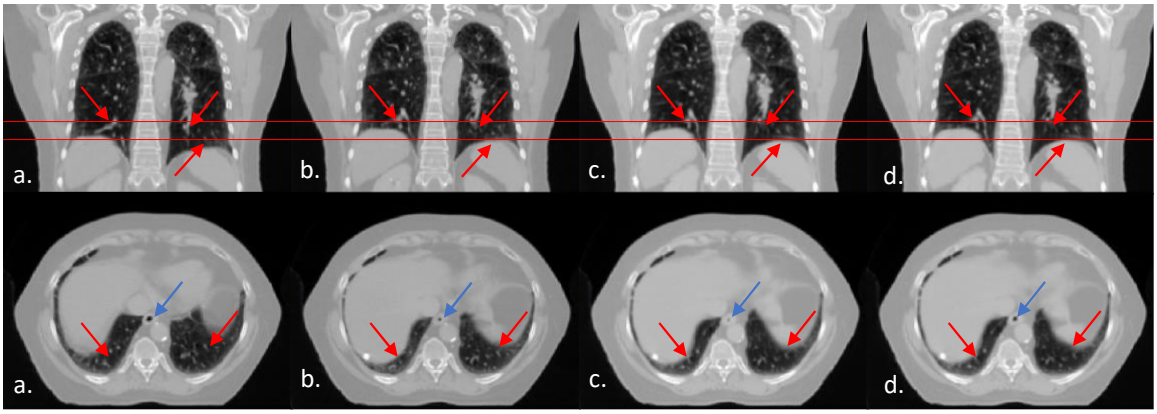


**Figure 14 Showing a coronal slice and axial slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with predicted DVF (d). The horizontal lines were drawn for reference purposes, red arrows indicate several main structures and blue arrows indicates the structure where the residual motions remain.**

As shown in figure 14, the diaphragm, vessel and mass showed a decent match in registered image from ground truth DVF and registered image from predicted DVF. But, some residual motions can still be found such as the lumen indicated by blue arrows in the axial slice.

### 3.2.3 Discussion

The network is required to deal with the difference in intra-fractional difference (respiratory phase motion) and inter-fractional anatomic difference in this case. The date taken both volumes are separated by 20 days, which means there is not many difference in anatomic change for the patient.

The feasibility of this experiment implies some potential clinical applications. It suggests that the CNN is able to be trained to learn the respiratory motion from the 4D-

CT volumes in the beginning of a treatment course, and the network could learn the respiratory motion of this patient. With more 4D-CT image volumes are taken during the course or in the end of the course, the well-trained CNN model could be transformed to perform the inter-phase registration on the new 4D-CT volume, automatic delineate the organs, and build target tracking model based on respiratory motion.

The following figure showed the automatic target delineation using the network. In testing volume, the liver has been contoured by Dr. Ren on phase 1, the network is used to predict the DVF from phase 1 to phase 7 in testing set, and the predicted DVF is used to register the phase 1 liver contour to phase 7. The axial slice of target image and registered contour is shown in figure below,

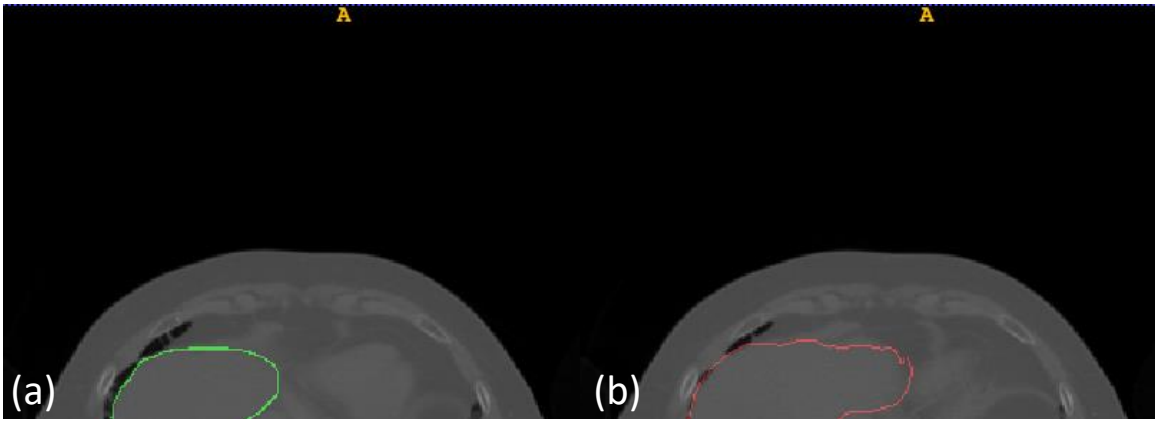


**Figure 15 The testing set phase 1 (source image) with manual liver contour (a), and the testing set phase 7 (target image) with predicted contour (b).**

As shown in figure 15, after physician has contoured liver in one phases (the end-of-inhalation in this case), the liver contour in rest phases could be registered to other phases with the respiratory motion learnt by the CNN.

It is fast for network to predict the DVF. Without considering the data reading and writing time, the prediction of the DVF and constructing registered image and contours took less than 1 min for each phase, and this performance can be further improved with better hardware.

## 3.3 Double DRR Simulated 4D-CBCT with Double 4D-CT Volumes

This case is similar to double 4D-CT case, the same image set were used as material to FDK-reconstructed the 4D-CBCT with 340 simulated DRR projections per phase. The main difference to double 4D-CT case is that the image quality. In DRR reconstructed 4D-CBCT, the structures are blurred and the streaks due to insufficient projections are obvious. Comparing this work with double 4D-CT case, the robustness of the network to low image quality is tested.

### 3.3.1 Coefficient of Correlation

The coefficient of correlation in predicting phase 1 to phase 7 DVF is shown in table 3.

**Table 3 The coefficient of correlation between the predicted deformable vectors and the ground truth deformable vectors in predicting phase 7.**

| Coefficient of correlation | | |
|---|---|---|
| AP | ML | SI |
| 0.136 | 0.383 | 0.655 |

The performance network has a sharp and noticeable decline in terms of correlation. The AP and ML direction showed week correlation between ground truth and prediction. The correlation in SI direction is stronger than the other two. Up to now, it is obvious that the image quality impacts the performance of the network strongly. The reasons are several. First, the ground truth DVF are not accurate enough due to poor image quality, and it will directly affect the learning accuracy of the network. Second, the learning effectiveness of the network is affected by poor image quality, since the blurred structures and streaks prevented the network from extracting the features. The poor image quality limits the ground truth accuracy and feature learning ability of the network, the predictability, therefore, has a significant degradation.

### 3.3.2 The Registered Image Comparison

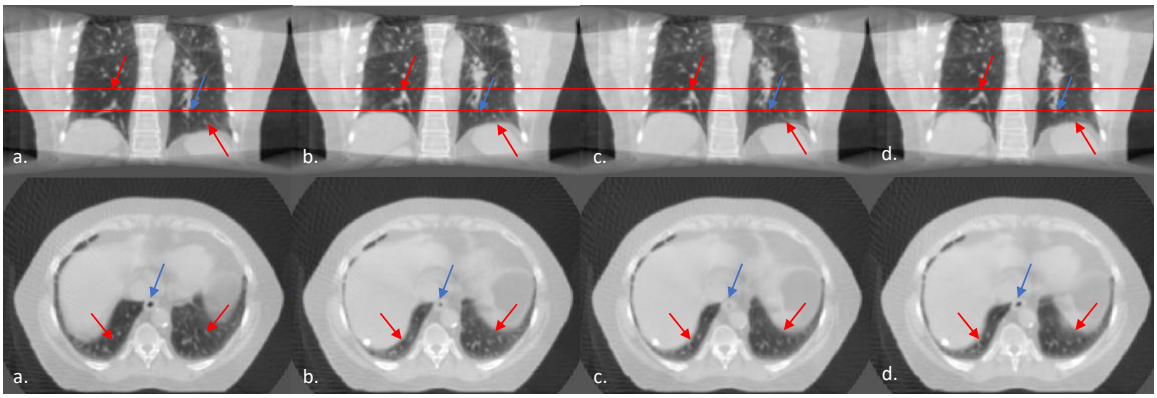


**Figure 16 Showing a coronal slice and axial slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with predicted DVF (d) for the case of double DRR simulated 4D-CBCT. The horizontal lines were drawn for reference purposes, red arrows indicate several main structures and blue arrows indicates the structure where the residual motions remain.**

As shown in figure 16, both diaphragms and one feature indicated by red arrows matched well in both registered image shown in (c) and (d). However, there are more residual motions in this case comparing with double 4D-CT case, which is expected and has been discussed in section 4.3.1.

### 3.3.3 Discussion

The robustness of the network against the poor image quality is tested with changing the training and testing samples to DRR simulated 4D-CBCT volumes. Due to the image quality degradation of the CBCT and streaks from insufficient projections, the performance of the network is expected to be worse than the double 4D-CT case. Both coefficient of correlation and comparison in registered images suggests more residual motion in registered image with the CNN. Despite residual motions, the main structures such as the diaphragm and some main vessels is still been registered, showing the potential of the network used as a tool doing the initial deformable image registration. As discussed in section 1.1.5, the large difference in source and target images limits the performance of conventional deformable image registration algorithms. Therefore, the CNN could register the image at the beginning and the residual motions being registered with conventional methods later. Noted that our proposed method could predict DVF and produce registered image very fast, the overall speed, therefore, will be faster than using the conventional DIR algorithm only.

## 3.4 4D-CT and 4D-CBCT

The case of double DRR simulated 4D-CBCT reveals that the image quality of training samples may impact on learning effectiveness of the network severely. Therefore, the training set was changed to 4D-CT this time. Besides, considering some wider applications in clinic, the testing sets are still 4D-CBCT volumes, since CBCT images are taken for daily verification purposes. In order to get rid of the streaks, the 4D-CBCT volumes are reconstructed with iterative method minimizing the total variation (TV) term using 170 primary projections per phase. The CT volume and CBCT projections are from Spare Challenge data set.

### 3.4.1 Coefficient of Correlation

**Table 4 The coefficient of correlation between the predicted deformable vectors and the ground truth deformable vectors in predicting phase 7.**

| Coefficient of correlation | | |
|---|---|---|
| AP | ML | SI |
| 0.516 | 0.385 | 0.686 |

In this case, the coefficient of correlation showed improvement comparing with double DRR-simulated 4D-CBCT. This improvement suggests the importance of image quality of both training and testing samples.

### 3.4.2 The Registered Image Comparison

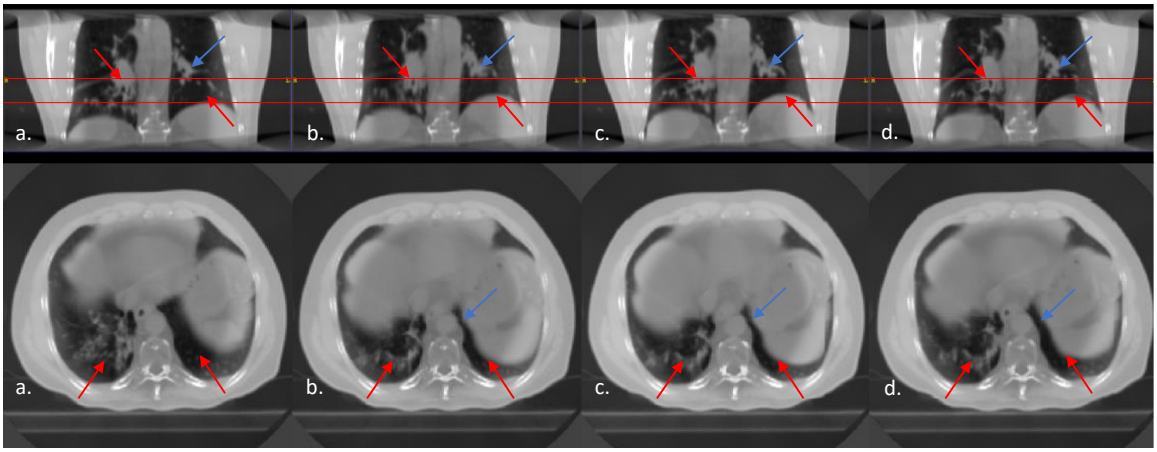


**Figure 17 Showing a coronal slice and axial slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with predicted DVF (d) for the case of learning with CT and predicting with 4D-CBCT. The horizontal lines were drawn for reference purposes, red arrows indicate several main structures and blue arrows indicates the structure where the residual motions remain.**

The image quality of this testing set is much better than FDK-reconstructed volume. The diaphragms, lumen and some features are predicted. However, there are still some residual movements indicated by blue arrows.

### 3.4.3 Discussion

With the improvement in image quality, the coefficient of correlation did show some improvement but not significant one, and the image also showed some residual motions failed to be registered by the network. The reason could be that the time interval between training image volume and target image volume are too long, the comparison of two image sets are shown in figure 18.

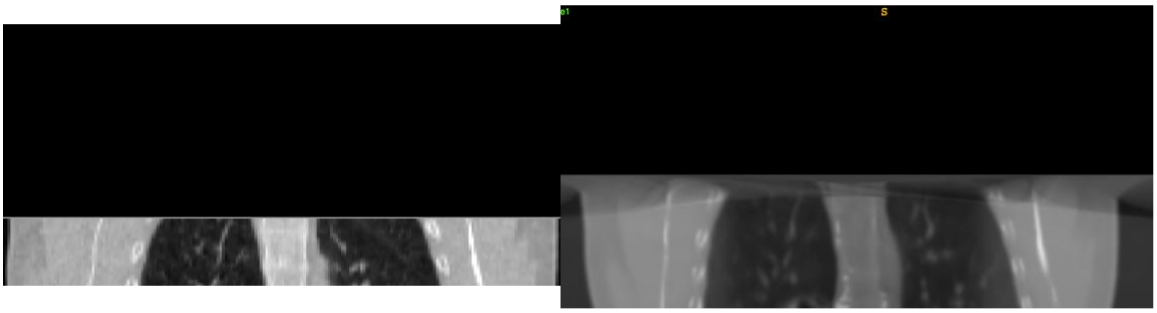

**Figure 18 The comparison of coronal slice in training and testing volumes. The major difference is the tumor above right diaphragm.**

Due to such large difference between training and testing samples, the undesirable predictability of the network might be explained. Again, this result shows the high similarity between the training and the testing leads to better performance in prediction.

## 3.5 Double FDK-reconstructed CBCT

In this case, both training and testing image sets were reconstructed with FDK method. In order to improve the image quality, get rid of the streaks, and enhance the effectiveness of learning process, 680 primary projections (fully-sampled) were used to reconstruct the training set images. 170 primary projections were used to reconstruct the testing image sets. Two image sets were taken from different days.

### 3.5.1 Coefficient of Correlation

**Table 5 The coefficient of correlation between the predicted deformable vectors and the ground truth deformable vectors in predicting phase 7.**

| Coefficient of correlation | | |
|---|---|---|
| AP | ML | SI |
| 0.764 | 0.719 | 0.875 |

As shown in table 5, predicted DVF and ground truth DVF showed a good correlation. Based on our previous analysis, higher image quality in training set will provide a more accurate ground truth deformable vector field as learning target, and also helps the network to learn and extract the feature in an effective manner. The low image quality of testing set has less impact on prediction performance this time. Furthermore, the inter-fractional different between both image sets are not significant. It might be concluded that the image quality is more important for training set and the anatomic different between two image sets determines the predictability of the network.

### 3.5.2 The Registered Image Comparison

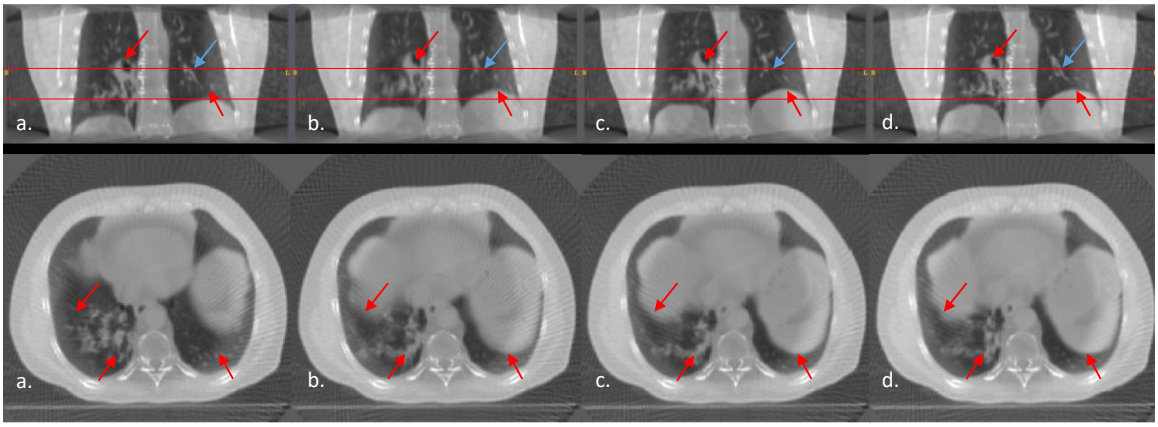


**Figure 19 Showing a coronal slice and axial slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with predicted DVF (d) for the case of double FDK reconstructed 4D-CBCT volumes. The horizontal lines were drawn for reference purposes, red arrows indicate several main structures and blue arrows indicates the structure where the residual motions remain.**

As shown in figure 19, the image quality for testing set are obviously low. However, despite the low image quality, the main structures inside the lung such as

diaphragms and mass were registered pretty well. As indicated by blue arrows, the residual motions still exist. But, with conventional deformable image registration algorithm, the residual motion can be registered efficiently, since the main structures have been registered with our network.

### 3.5.3 Discussion

In this case, a fully sampled 4D-CBCT is used for training and an under-sampled 4D-CBCT is used for testing. Noted that the projections used for reconstruction are primary beam only. The result shows that the learning effectiveness of the network is significantly improved with high image quality in training samples. The predictability of the network is mainly affected by the anatomic similarity between training and testing samples, and is affected by image quality of testing set insignificantly. The following figure shows the comparison of the training and testing images.

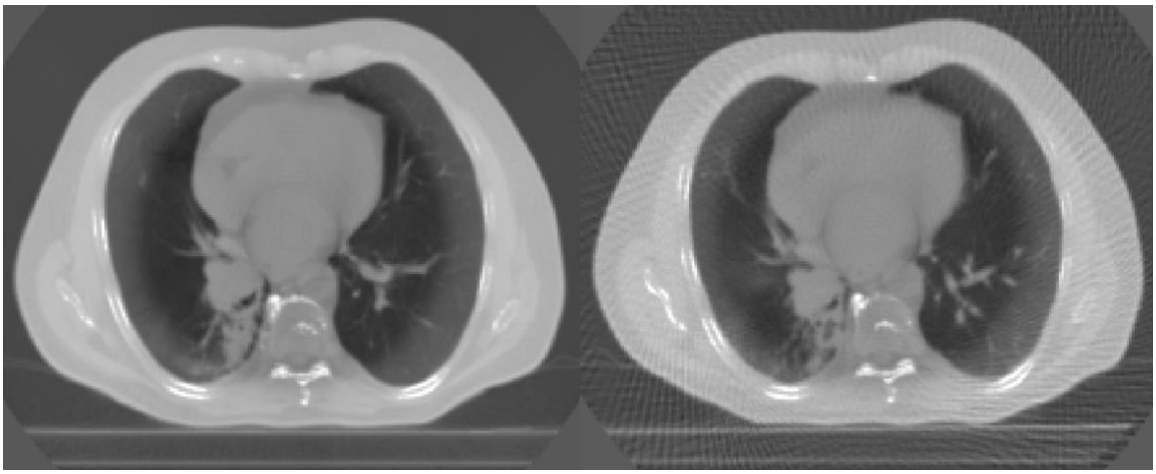

**Figure 20 The image quality comparison of the training and testing images. The left is the fully-sampled training image and the right is the under-sampled testing image.**

Noted that the volumes used for this test are reconstructed by primary projections and it is not clinical one. It limits the application of the network on a daily basis.

## 3.6 Interpatient 4D-CT

This case is the most challenging one since interpatient prediction unavoidably introduced significant anatomic difference between training and testing samples. Therefore, several 4D-CT volumes from different patients were used as training samples to build a data base containing a variety respiratory pattern and feature structure.

### 3.6.1 Coefficient of Correlation

**Table 6 The coefficient of correlation between the predicted deformable vectors and the ground truth deformable vectors in predicting phase 7.**

| Coefficient of correlation | | |
|---|---|---|
| AP | ML | SI |
| 0.229 | 0.709 | 0.769 |

As shown by the correlation, the network performed well in the ML and SI direction. Less correlation is shown in AP direction. The reason could be that the trained data base did not cover enough AP direction feature movement, and more samples are suggested.

### 3.6.2 The Registered Image Comparison

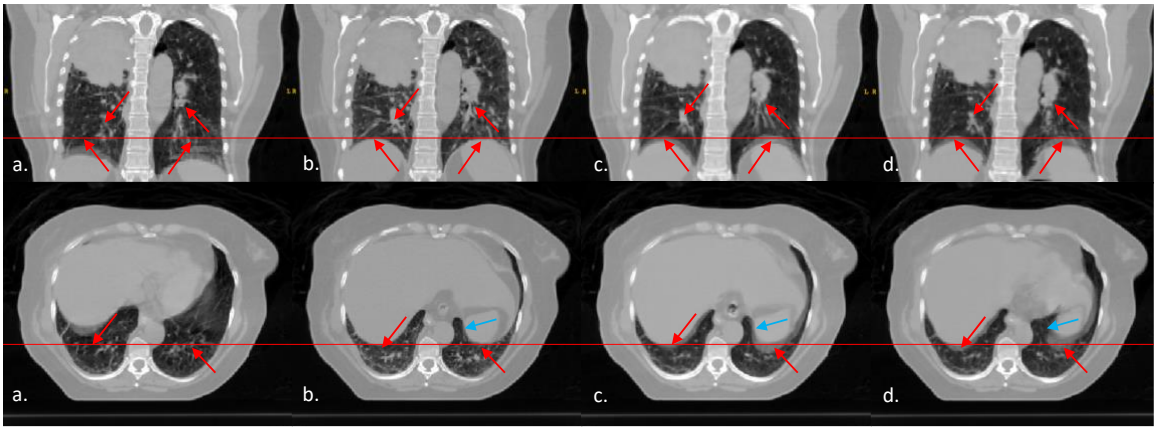


**Figure 21 Showing a coronal slice and axial slice of the source images (a), the target image (b), the registered image with ground truth DVF (c), and the registered image with predicted DVF (d) for the case of double FDK reconstructed 4D-CBCT volumes. The horizontal lines were drawn for reference purposes, red arrows indicate several main structures and blue arrows indicates the structure where the residual motions remain.**

As shown in figure 21 the red arrows, the main structures showed a good match in ground truth and prediction, such as the diaphragm and mass. However, the residual motions are quite significant indicated by blue arrows.

### 3.6.3 Discussion and Application

This result indicates a potential application of the network in clinic. A network containing the respiratory motion patter for multiple patients, and the diaphragms and some main structures of respiratory motion of new patient could be initially registered with this network in an efficient manner. The residual motion, then, can be registered with existing DIR algorithms. The overall registration time will be less than direct register of the image from scratch with conventional DIR algorithm, because our

network could at least register the main feature of the image and reduce the difference between source and target images quickly.

## 3.7 Clinic Double 4D-CBCT

This case is a failed one. Two clinic 4D-CBCT volumes for the same patient are used for training and testing. The network barely learnt nothing from it due to the limited image quality of 4D-CBCT.

## 3.8 Summary

All the results have been shown and discussed. For all the successful cases, the main feature inside the lung such as the diaphragms and main vessels can be registered with the network. There are some residual motions remaining after registration.

### 3.8.1 Advantages and Applications

A main advantage of using CNN to predict the respiratory motion is the efficiency. The deformable vectors were predicted within few seconds and the registered image was produced within half minute. This advantage can be used for initial registering the images, and matching the main features between source and target images, then the conventional DIR algorithms deals with less-challenging residual motion. The overall register time will be reduced and the efficiency of routine clinic work could be improved. With further reducing the registration time in the future, the on-board target tracking and treatment planning could be achieved.

In addition, the CNN based approach is fully automatic and user independent. In contrast, many traditional registration algorithms require user input for tuning the parameters, and the registration results are very much dependent on the user experience.

### 3.8.2 Characteristics of Learning and Predicting

Image quality, amplitude of respiratory, inter-fractional difference and accuracy of the ground truth will affect the learning and predicting process.

High image quality is crucial for the network to effectively learn the respiratory motion. Besides, image quality will also affect the ground truth DVF, since the ground truth DVF in this work is produced with the *VelocityAI.* With high image quality, it is easier for the network to extract the features and learn the respiratory pattern in training process. And the accurate ground truth will supervise the network to learn the respiratory motion accurately. On the other hand, the image quality has less effects to testing stage as shown in table 5 and figure 20.

Larger respiratory motion is easier to be learnt and predicted by the network. The respiratory motion is more significant in SI direction, then the ML direction, then the AP direction. The general coefficient of correlation showed that the predicted DVF in SI direction has stronger correlation with the ground truth.

The inter-fractional anatomic difference impacts the predictability of the network. The result shows that larger inter-fraction difference (i.e. larger difference in

training and testing samples) leads to poorer prediction performance. It is understandable that the network will poorly react to unfamiliar features, and therefore the performance is not good.

# 4. Limitation

This section will introduce the limitation of this work from the angles of the method and the future application.

## 4.1 Limitation in Methods

This work used patch-based CNN to learn the deformable vectors with the supervision of the ground truth deformable vectors. Then, the prediction results are evaluated with the coefficient of correlation and compare the registered image with predicted DVF and ground truth DVF.

### 4.1.2 Limitation in Samples

Each sample contains a patch pair from source and target images associated with the deformable vector at the control points. The selection of control points, the choosing of the patch size, and the accuracy of the ground truth deformable vectors affect the quality of the samples. The sample quality directly affects the entire learning efficiency and prediction accuracy.

The control points, in this work, were chosen with a uniform grid across the lung, which the density of the grid is directly related to the sample numbers. Considerably, the meaningless patches such as some ribs, which has very small movement, were chosen. The network will process more unnecessary samples. A more developed control points selection method can be proposed, for example the training samples were chosen by the correlation between two patches. In fact, for inter-patient

DVF prediction section, this method has been used but showed insignificant improvement. Therefore, it was not shown in the result part. Yet, this improvement can be made to improve the learning effectiveness.

The ground truth definitely will impact on the performance of the network, because it supervised the learning process of the network. Due to the limitations of the existing DIR algorithm, the perfect ground truth DVF is next to impossible, and the imperfection of the ground truth will affect the learning accuracy of the prediction.

### 4.1.3 Limitation in Evaluation Methods

There is no a standard way to evaluate the deformable image registration performance. In this work, the coefficient of correlation is used. This coefficient evaluates the global performance of the prediction. However, the goal of the registration for us is to match the main features from one image to another, and the global performance is not representative enough for this purpose. Therefore, it is only a reference of the performance.

Dice similarity coefficients (DSCs) are often used in evaluating the similarity of two images with comparing the contours for certain structure. However, in our case, there is no attached contour on both source and target images. This method is not feasible for us.

## 4.3 Limitations in Applications

This DVF prediction method cannot be applied in daily 4D-CBCT volumes, because the excessive scatters and defects on the image prevent the network from extracting the features. Therefore, in clinic, the application of this method is limited to 4D-CT volumes.

In order to apply the network to the daily basis clinic work, the image quality improvement methods for the 4D-CBCT images is proposed. (Zhao et. al. 2019) With effective image quality improvement for CBCT volumes in clinic, this method will gain a wider range of applications in the future.

# 5. Conclusion

Deformable image registration has long been an ill-posed problem. Most conventional algorithms are time-consuming, user-dependent and in need of parameter-tuning. The convolutional neural network is good at feature extraction and pattern learning, and the speed of the prediction using CNN is very fast. Therefore, a patch-based CNN model is built to predict the deformable vector of the input patch pair. The network is supervised with the ground truth deformable vectors generated by *VelocityAI*. Different image sets combinations were trained and tested. The registration performance was evaluated, the robustness against poor image quality, inter-fractional anatomic difference, and interpatient anatomic different was evaluated, and the applications in clinic were exploited and analyzed. The general performance of this method is decent, provided with the high image quality and smaller inter-fractional anatomic differences. Under all combinations, the main structures such as diaphragms, tumor and main vessels can be registered well with the network. The residual motions exist in all combinations. Smaller inter-fractional anatomic difference between training and testing samples showed less residual motions in the prediction..

## References


Barillot, C., Haynor, D. R., & Hellier, P. (Eds.). (2004). *Medical image computing and computer-assisted intervention: MICCAI 2004, 7th international conference, Saint-Malo, France, September 26-29, 2004: proceedings*. Berlin ; New York, N.Y: Springer.

Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys*, *24*(4), 325–376. https://doi.org/10.1145/146370.146374

Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., & Shen, D. (2017). Deformable Image Registration Based on Similarity-Steered CNN Regression. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, & S. Duchesne (Eds.), *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017* (Vol. 10433, pp. 300–308). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-66182-7_35

Chen, L., Tang, W., & John, N. (n.d.). Self-Supervised Monocular Image Depth Learning and Confidence Estimation, 19.

Christensen, G. E., Carlson, B., Chao, K. S. C., Yin, P., Grigsby, P. W., Nguyen, K., … Williamson, J. F. (2001). Image-based dose planning of intracavitary brachytherapy: registration of serial-imaging studies using deformable anatomic templates. *International Journal of Radiation Oncology*Biology*Physics*, *51*(1), 227–243. https://doi.org/10.1016/S0360-3016(01)01667-4

D Ehler, E., Bzdusek, K., & Tome, W. (2009). A Method to Automate the Segmentation of the GTV and ITV for Lung Tumors. *Medical Dosimetry : Official Journal of the American Association of Medical Dosimetrists*, *34*, 145–153. https://doi.org/10.1016/j.meddos.2008.08.007

Ford, E. C., Mageras, G. S., Yorke, E., & Ling, C. C. (2002). Respiration-correlated spiral CT: A method of measuring respiratory-induced anatomic motion for radiation treatment planning. *Medical Physics*, *30*(1), 88–97. https://doi.org/10.1118/1.1531177

Goshtasby, A. (2005). *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications*.

H. J. Johnson, & G. E. Christensen. (2002). Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, *21*(5), 450–461. https://doi.org/10.1109/TMI.2002.1009381

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., … Behnke, S. (2010). Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In K. Diamantaras, W. Duch, & L. S. Iliadis (Eds.), *Artificial Neural Networks – ICANN 2010* (Vol. 6354, pp. 92–101). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15825-4_10

Ian, G., Yoshua, B., & Aaron, C. (2016). *Deep Learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Li, X., Zhang, Y., Shi, Y., Wu, S., Xiao, Y., Gu, X., … Zhou, L. (2017). Comprehensive evaluation of ten deformable image registration algorithms for contour propagation between CT and cone-beam CT images in adaptive head & neck radiotherapy. *PLOS ONE*, *12*(4), 1–17. https://doi.org/10.1371/journal.pone.0175906

Oh, S., & Kim, S. (2017). Deformable image registration in radiation therapy. *Radiation Oncology Journal*, *35*(2), 101–111. https://doi.org/10.3857/roj.2017.00325

Ou, Y., Sotiras, A., Paragios, N., & Davatzikos, C. (2011). DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, *15*(4), 622–639. https://doi.org/10.1016/j.media.2010.07.002

Papademetris, X., Jackowski, A. P., Schultz, R. T., Staib, L. H., & Duncan, J. S. (2004). Integrated Intensity and Point-Feature Nonrigid Registration. In C. Barillot, D. R. Haynor, & P. Hellier (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004* (pp. 763–770). Springer Berlin Heidelberg.

Tosun-Turgut, D. (2010). *Rigid Image registration*.

Vedam, S. S., Keall, P. J., Kini, V. R., Mostafavi, H., Shukla, H. P., & Mohan, R. (2003). Acquiring a four-dimensional computed tomography dataset using an external

respiratory signal. *Physics in Medicine and Biology*, *48*(1), 45–62. https://doi.org/10.1088/0031-9155/48/1/304

Wu, G., Kim, M., Wang, Q., & Shen, D. (2014). S-HAMMER: Hierarchical attribute-guided, symmetric diffeomorphic registration for MR brain images. *Human Brain Mapping*, *35*(3), 1044–1060. https://doi.org/10.1002/hbm.22233

Yan, D., Jaffray, D. A., & Wong, J. W. (1999). A model to accumulate fractionated dose in a deforming organ. *International Journal of Radiation Oncology*Biology*Physics*, *44*(3), 665–675. https://doi.org/10.1016/S0360-3016(99)00007-3

Yang, D., Lu, W., Low, D. A., Deasy, J. O., Hope, A. J., & El Naqa, I. (2008). 4D-CT motion estimation using deformable image registration and 5D respiratory motion modeling: 4D-CT motion estimation and modeling. *Medical Physics*, *35*(10), 4577–4590. https://doi.org/10.1118/1.2977828