Convolutive Attention for Image Registration

Tim J. Parbs*, Philipp Koch†, and Alfred Mertins*†
*Institute for Signal Processing, University of Lübeck, Germany
†German Research Center for Artificial Intelligence (DFKI), AI in Biomedical Signal Processing, Lübeck, Germany
{t.parbs, alfred.mertins}@uni-luebeck.de, philipp.koch@dfki.de

***Abstract*—Elastic registration of deformed images is a vital component of many computer vision tasks, especially when considering medical image data. Deep learning techniques, particularly U-Nets, offer state-of-the-art performance, but do not yet use the rich spatial information context available in natural images. We propose an augmentation based on the recently introduced attention mechanism to allow a U-Net to use spatial image context. A dedicated convolutive attention scheme has been developed by calculating local similarity scores of the multidimensional inputs. Additionally, a dedicated composite error function based on common image similarity measures is introduced in order to further improve the registration results. To evaluate our approach, we conducted several experiments on an augmented real-world dataset containing cardiac cine MRI scans. The comparison with state-of-the-art registration schemes highlights the potential of our approach.**

***Index Terms*—image registration, deep learning, attention, local similarity score, U-Net, composite loss**

I. Introduction

In many computer vision tasks, correctly aligning different images is an important early step. This task is especially demanding when considering arbitrary, elastic deformation and not restricting the solution space to affine transformations. While numeric algorithms based on convex optimisation exist, they are often rather slow or very sensitive to the presence of noise. These problems motivate research into alternative registration approaches using deep learning. In recent years, convolutional neural networks (CNNs) became popular in research fields dealing with multidimensional data. These networks offer several advantages over fully connected networks, namely the possibility of sharing learned weights across data dimensions, leading to translation invariance. This inductive bias enables strong feature extraction and generalisation on natural images. In many applications, a multi scale approach to extracting data information is highly desirable. Thus, the most common variant of CNNs in computer vision tasks is currently the so-called U-Net, which became the de-facto gold standard in various computer vision disciplines, including image segmentation [1] and object detection [2]. These networks consist of a contractive encoding path and an expansive path which formulates the output of the network. Besides extracting low-level features, the contractive part condenses image information while discarding superfluous details. In the expansive pathway, the low-resolution feature maps are upsampled and combined with spatial information from the contractive path at every scale via so-called skip connections. However, since spatial information is extracted through network depth, the beginning stages of the network do not have access to much of the spatial information from every image region. The recently introduced concept of attention [3] already helped to revolutionise various fields in deep learning. Originally invented to overcome long gradient paths and the exploding/vanishing gradients problem prevalent in recurrent networks for natural language processing tasks, it was quickly adopted in many different fields and often outperformed previous state-of-the-art approaches. And, in the last months, attention mechanisms are finding their way into deep learning architectures tackling multidimensional problems. Modified transformer architectures have been used for image classification (in e.g. [4] and [5]), medical image segmentation (in e.g. [6] and [7]) or denoising [8]. However, these networks focus mainly on channel-wise dependencies. While some image registration networks based on the attention mechanism exist, they utilise a similarity score based on tokenised discrete image patches [9] or on whole feature maps [10], which might not be ideal to capture spatial context.

In this work, we explore the viability of the insertion of convolutive attention cells into the U-Net architecture to enable usage of spatial context information from natural medical images. We aim to keep most of the tried-and-true U-Net architecture intact, but allow the network to access structural image information on every stage by modifying its skip connections with the aforementioned cell.

II. Dataset

We use the popular Automatic Cardiac Diagnosis Challenge (ACDC) [11] training dataset. It consists of cardiac cine magnetic resonance imaging sequences from 100 clinical examinations, each consisting of a time resolved 3D image volume with a maximum image size of 428×512 and up to 31 slices. The dataset was created to develop and evaluate pathology prediction systems as well as segmentation methods and thus contains sequences of healthy heart activity as well as sequences from known pathologies. Furthermore, the volume data for each measurement contains 3D label maps for both end systole and end diastole created by medical experts. The label maps show the position of myocardium and both ventricles in the volume.

In this work, we focus on 2D image registration. By slicing the 4D image data into 2D images and discarding unlabeled images as well as corrupt data, we end up with a dataset of 772 image pairs. More images from the dataset could be used in training, but we wanted to establish comparability to similar

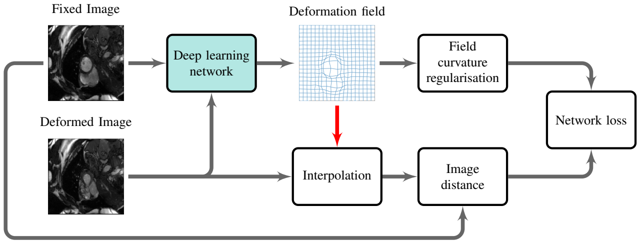

Figure 1: An overview of the task at hand.

deep learning approaches, which only used labeled slices. The images were normalised for pixel values between 0 and 1 and rescaled to a common size of 112 x 112 using bi-cubic interpolation.

III. Methods

Given a pair of images I_r, $I_m \in N_1 \times N_2$ taken from a motion sequence, image registration in general is the search of a transformation field $y : \mathbb{R}^2 \to \mathbb{R}^2$ which warps the moved image I_m to resemble the reference I_r. Thus, in a best case scenario, the deformation yields $I_m(y) = I_r$, where off-grid values in $I_m(y)$ are interpolated using a suitable interpolation function. We use a deep learning network to create y as a response to an input image pair I_r, I_m, so that an interpolated image $I_y = I_m(y)$ is close to I_r. A convex loss function is utilised to minimise the difference between I_y and I_r. As no ground truth deformation exists or would necessarily be generated by a classical registration scheme, our network is trained solely on image data in a unsupervised fashion. Our basic setup is schematically shown in Fig. 1.

A. Loss Functions

The proposed objective function of our network is a composite loss function consisting of two commonly used differentiable image quality metrics and a straight forward deformation regulariser. We define it using the weighted sum

$$\mathrm{E}(I_r, I_y, y) = \lambda_S \psi(I_r, I_y) + \lambda_L \ell_2(I_r, I_y) + \lambda_R R(y), \quad (1)$$

where λ_S, λ_L, λ_R are positive scalar weightings. Using $N = N_1 \cdot N_2$ we define the parts of E by

$$\psi(I_r, I_y) = 1 - \mathrm{SSIM}(I_r, I_y), \quad (2)$$

$$\ell_2(I_r, I_y) = \frac{1}{2} \sum\nolimits_{n=0}^{N-1} \|I_r - I_y\|_2^2 \quad \text{and} \quad (3)$$

$$R(y) = \frac{1}{2} \sum\nolimits_{n=0}^{N-1} (\nabla_1 y_n)^2 + (\nabla_2 y_n)^2 . \quad (4)$$

The Structural Similarity Index Measure (SSIM) [12] is a popular image metric which captures subjective image quality. It has been shown that a 'pure' SSIM loss is unsuitable for computer vision tasks, as CNNs trained on such a loss struggle to correctly align image edges [13]. Therefore, we pair it with a conventional ℓ_2-loss from the difference between I_r and I_y. Since our images contain large uniform patches (mostly regions of low intensity outside the region of interest), we have to impose regularisation on the deformation field created by the network. We penalise abrupt change in the vector field by defining the regulariser $R(y)$ in (4) as the sum of the squared spatial differences at every pixel using finite difference operators ∇_1, ∇_2 for both spatial dimensions.

B. Model Architecture

Our entire network architecture is schematically shown in Fig. 2. The basic network design can be understood as an augmented U-Net. We use a three stage contractive part in which we use a repeated concatenation of 2D convolution, batch normalisation and rectified linear unit (the (Convolve-Batchnorm-ReLu)^2-cell, or CBR^2-cell, pictured in Fig. 2) for feature extraction. We use standard Max-Pooling-Layers between levels of the U-Net, which reduce height and width of feature maps by a factor 2. In the expansive network part, feature map size is restored through consecutive transposed convolution of stride size two.

We propose a new module which we call the Residual Self-Attention cell (ReSAtt, shown schematically in Fig. 3) which modifies each skip connection of conventional U-Nets. The input to the attention cell is concatenated with a positional encoding, for which we use a normalised Cartesian grid, and then used to generate Key (K), Query (Q) and Value (V) tensors of size $H \times W \times C$ through three CBR^2 cells. Following the core idea of attention mechanism, we aim to generate an attention score as a similarity measure between entries of K and Q. Where attention cells in e.g. Natural Language Processing calculate this similarity on the features of embedded tokenised words and use this similarity to establish context between input sequence parts, we do the same on small image regions. For this reason, we partition K and Q into tensors of filter patches. Let $R_{r,k}(\cdot)$ be an operator that extracts a neighborhood of size $k \times k \times C$ centered around the pixel r, we extract

$$\begin{aligned} q_r &= R_{r,k}(Q) \\ v_r &= R_{r,k}(V). \end{aligned} \quad (5)$$

Sampling r on a Cartesian grid with step size u, we compose the neighbourhood tensors $\hat{Q} = [q_0, q_1, ...q_F]$ as well as $\hat{V} = [v_0, v_1, ...v_F]$ with both $\hat{Q}, \hat{V}$ of size $F \times k \times k \times C$. We create a set of transposed filters $\hat{V}^T$ of size $C \times k \times k \times F$ through permutation. Neighbourhood patch size k and step size u are set to 5 and 3, respectively. With $\star$ denoting correlation, we generate an attention score matrix A by

$$A = [a_0, a_1, ...a_F] \quad (6)$$

$$a_f = K \star q_f \ , \ \ f \in \{0, 1, F-1\} \quad (7)$$

and, using the convolution operator $*$, the attention output B of size $H \times W \times C$ with

$$B = [b_0, b_1, ...b_C] \quad (8)$$

$$b_c = \mathrm{softmax}(A) * \hat{V}_c^T \ , \ \ c \in \{0, 1, C-1\}. \quad (9)$$

This can be seen as an extension of the scaled dot-product attention in the 'conventional' attention mechanism where instead of a matrix product, we calculate similarities using

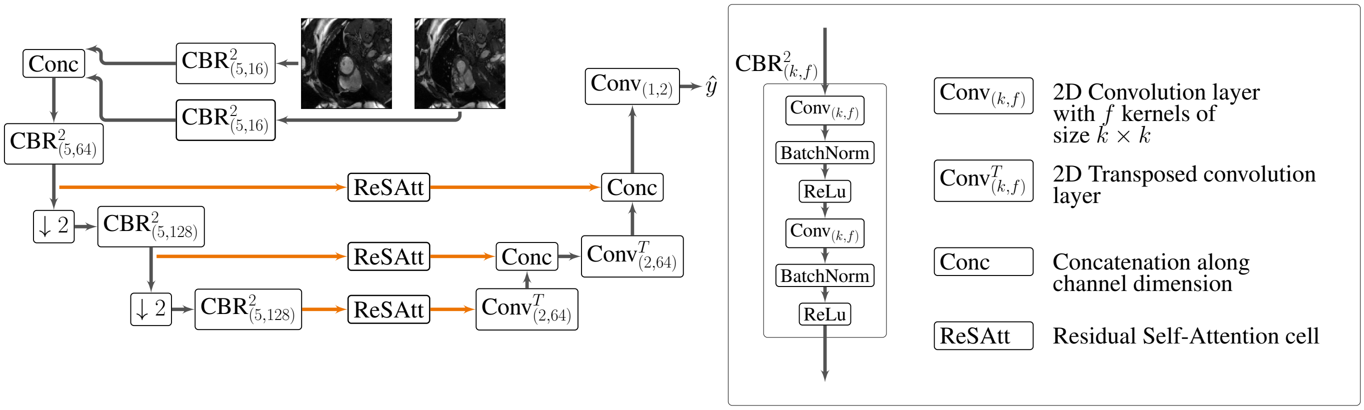
Conc
$\mathrm{CBR}^2_{(5,16)}$
$\mathrm{CBR}^2_{(5,16)}$
$\mathrm{CBR}^2_{(5,64)}$
ReSAtt
$\downarrow 2$
$\mathrm{CBR}^2_{(5,128)}$
ReSAtt
Conc
$\downarrow 2$
$\mathrm{CBR}^2_{(5,128)}$
ReSAtt
$\mathrm{Conv}^T_{(2,64)}$
$\mathrm{Conv}^T_{(2,64)}$
Conc
$\mathrm{Conv}_{(1,2)}$
$\hat{y}$
$\mathrm{CBR}^2_{(k,f)}$
$\mathrm{Conv}_{(k,f)}$
BatchNorm
ReLu
$\mathrm{Conv}_{(k,f)}$
BatchNorm
ReLu
$\mathrm{Conv}_{(k,f)}$ 2D Convolution layer with f kernels of size $k \times k$
$\mathrm{Conv}^T_{(k,f)}$ 2D Transposed convolution layer
Conc Concatenation along channel dimension
ReSAtt Residual Self-Attention cell

Figure 2: The network architecture, schematic of the CBR^2 cell, and legend.

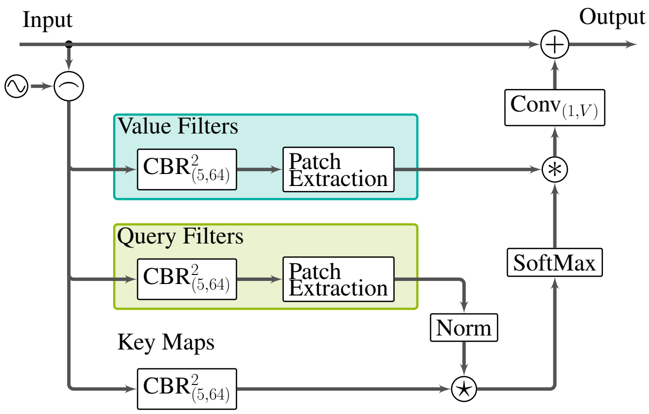

Figure 3: The residual attention cell. V is dynamically set to match the input. ∿ is the positional encoding appended to the input channels.

correlation and convolution. Using a convolutional layer with a kernel size of one, we adapt the number of features to the input. This way, we can keep the number of channels inside each attention cell constant. Finally, we combine the attention output with the input using a residual connection. As the output $\hat{y} \in N_1 \times N_2 \times 2$ of the network is considered as the perturbation of the the identity grid x (with $I(x) = I$), the dense vector field used for deformation is $y = x + \hat{y}$. The registered image I_y is then created using y by bilinear interpolation.

C. Data Augmentation

Image augmentation was used to compensate for our small training set. In addition to randomly switching the order of the input images I_r and I_m, we employed standard image augmentations consisting of random flipping and rotation as well as random image cropping to between 80% and 100% of the original image size. We also slightly deformed the input images using low-pass filtered Gaussian noise [14]. All image augmentation was done at run time and imposed virtually no overhead.

IV. Experimental Setup

A. Data, Training and Model Parameters

Our model was implemented in Tensorflow 2.4.1 using an ADAM optimiser and a learning rate of $1\mathrm{e}{-5}$. We empirically found values of $\lambda_S = 1\mathrm{e}6$, $\lambda_L = 5\mathrm{e}2$, and $\lambda_R = 150$ to yield the best registration results. In all experiments, we trained using k-fold cross validation. We used a pseudorandom split to partition the image data along the 100 available sequences into 10 bins of roughly equal size. The bins were then used in a ten-fold cross validation procedure to train our architecture. This way, we used a roughly 9:1 split for our training and validation procedure and every image pair was represented in the validation set once. All networks were trained with a minibatch size of 10 and for 700 epochs, even though convergence was typically reached much sooner. In each training epoch, every image pair in the training set was augmented and shown to the network once.

B. Baseline Systems

We will compare registration results with a state-of-the-art U-Net architecture of Hering et al. [15] trained on the same dataset. Their goal was not only to optimise image registration, but to minimise the error in label map alignment as well. Comparison was not straight forward, as their original loss function penalised misalignment of region maps, which we did not involve in network training. We implemented their network and trained with both their originally proposed loss function and training parameters as well as our loss function defined in (1). The data used for training was the same in all networks, and we used the same data augmentation in all cases. In the results, we denote the U-Net trained with the original loss function as $\mathrm{Unet}_{\mathrm{orig}}$, and the U-Net trained with our loss function as $\mathrm{Unet}_{\mathrm{mod}}$. We also will compare with a more traditional diffeomorphic elastic image registration algorithm (Diffeomorphic Demons [16]) and the unregistered case, in which we set $y = x$.

C. Quality Measures

The region maps that were contained in the data set were used to compute quality metrics for the evaluation of registra-

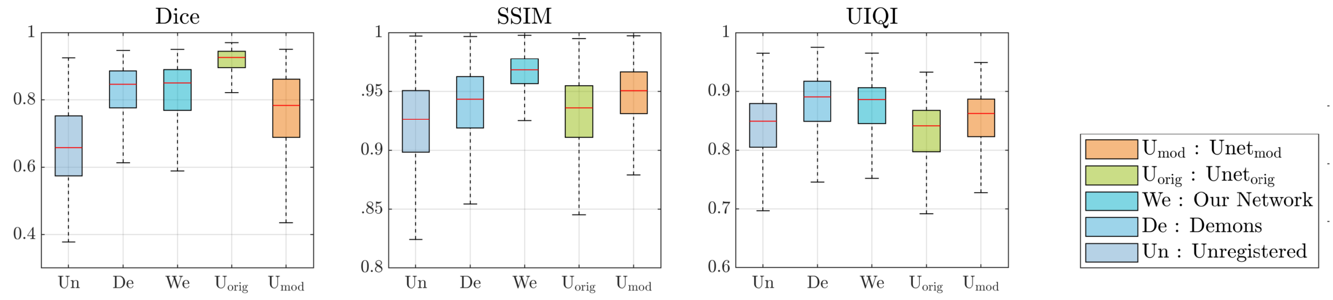

Figure 4: Results for evaluation metrics for the unregistered case, Demons registration, our proposed network and both U-Nets used for comparison. Median is shown as red line, the box marks 25th and 75th percentile, whiskers mark extremes of range.

tion performance. We used the Universal Image Quality Index (UIQI) [17] as a metric for combined image similarity and quality of the registered image. The UIQI captures changes in image correlation, luminance, and contrast between I_r and I_y. The one-hot encoded label map of I_m was deformed using the estimated y and compared to the label map of I_r. We calculated the mean of the Dice coefficient for the maps of every anatomical label. As we do not explicitly restrict the output of our network to the space of diffeomorphisms, we also monitor the folding that occurs during deformation. We use the determinant of the Jacobian of the vector field, which is used to measure change in volume for elements in the image. Negativity of the Jacobian determinant for some volume element would denote a change in orientation of this volume element, causing an overlap in the deformed volume. Therefore, using Iverson bracket notation, a suitable measure for the admissibility of the deformation can be defined with the Jacobian of the approximated deformation $J(y_n)$ with

$$\delta_{J(y)} = \frac{1}{N} \sum_{n=0}^{N-1} [\det\left(J(y_n)\right) \ < \ 0]. \tag{10}$$

This represents the mean occurence count of a negative Jacobian determinant throughout the discrete vector flow field. We approximated the spatial derivatives with a basic forward difference across the spatial dimensions. In the following section, we evaluate registration performance using UIQI, SSIM and Dice Score. All three metrics are upper bounded by one, with one being the optimal result.

V. Results

After training our deep learning models in each of the validation runs, we generated results by feeding each validation data point through the network and recording the metrics of its output. We combined results for all of the 10 validation runs, which yields representative results for the whole dataset. The combined metrics for our network and its two comparisons are shown in Fig. 4. As is visible in the results, our architecture offers a Dice coefficient and UIQI performance similar to a conventional diffeomorphic elastic registration algorithm. Our architecture beats the others based on SSIM, and the network proposed in [15] excels in Dice coefficient performance. This was expected as their loss function is designed to minimise normed label difference. However, when comparing metrics between Unet$_{\text{mod}}$ and our network, it becomes apparent that our architecture achieves significantly better median and metric spread on SSIM and UIQI metrics. A single image pair from a validation set including the deformation field approximated by the network is shown in Fig. 5. Anatomical structures are subjectively well registered in I_y, while the registration was not visibly perturbed by image noise. In our architecture, inference of a registration field took around 0.15s per image pair using a NVIDIA Geforce RTX 2080 GPU.

Table I: Field Folding Results.

λ_R	$\delta_{J(y)}$	Dice	UIQI	SSIM
0.15	3.3e−2 (3e−4)	0.745	0.897	0.986
1.5	1.0e−2 (5e−5)	0.796	0.899	0.988
15	1.9e−3 (4e−6)	0.830	0.892	0.985
150	3.2e−6 (8e−10)	0.803	0.866	0.972
1500	0 (0)	0.678	0.842	0.957

A. Model Parameters and Vector Field Folding

As mentioned, our network is not designed to necessarily generate diffeomorphic deformation fields. Instead, we opted to avoid local folding by penalising the deformation gradient. Varying the weighting term λ_R shows the influence of the regularisation on the invertibility of the created transform and the registration quality. The results are shown in Table I, in which we present the mean value of $\delta_{J(y)}$ (as its median is 0 everywhere) and its variance in parentheses for the whole validation dataset as well as the median of the other performance metrics. In these experiments, the other parts of the loss function were kept the same as in Section IV-A. Smaller values of λ_R result in divergence during training. As expected, a decrease in λ_R caused an increase in occurrences of vector field folding, but also an increase in image quality (measured with SSIM and UIQI) as well as Dice coefficient. With minimal regularisation strength, we observe a decrease in Dice coefficient score, which is caused by heavy deformations as a response to image noise. However, this did not cause a subjective degradation in image quality. We note that even

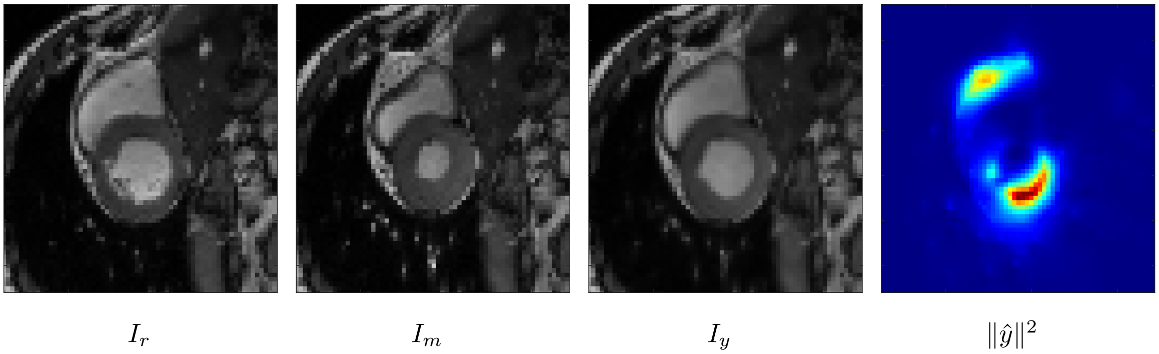

Figure 5: A sample of the registration, showing reference image and deformed image before and after deformation. Also shown is the intensity of deformation in every pixel in the shown image region. Image is zoomed in to only show cardiac region.

for small values of λ_R, the occurrence of folding all but disappears. With the chosen regularisation of $\lambda_R = 150$, we detect 31 occurrences of folding across the whole data set, which we deem acceptable. Varying λ_S and λ_D by an order of magnitude did not negatively affect the registration outcome, and mostly changed the rate of convergence during training.

VI. Discussion

We have presented a deep learning convolutive attention module for the utilisation of spatial information embedded in medical images. The efficacy of the method was evaluated using a data set of image slices from cardiac cine MRI scans. The rather small size of the data set was alleviated by use of strong data augmentation techniques, though further experiments could benefit from using larger data sets. We used common image metrics for unsupervised training and evaluation of image registration performance. We showed that our augmented network architecture has advantages in registered image quality in comparison to other machine learning based registration algorithms. Our architecture achieved better results on metrics capturing subjective image qualities in comparison with unaugmented U-Nets trained on the same loss function. Even though the output from the architecture was not explicitly restricted to the space of diffeomorphisms, we could show that with a suitable choice of regularisation, the created vector field exhibits no local folding. However, explicitly formulating the vector field as a diffeomorphism might be a potential avenue for improvement.

References

[1] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, 2015, pp. 234–241.

[2] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is all you need,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, *An Image is Worth 16x16 Words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929.

[5] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu and W. Gao, “Pre-trained image processing Transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 294–12 305.

[6] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth and D. Xu, “UNETR: Transformers for 3D medical image segmentation,” in *Proc. IEEE/CVF Winter Conf. Appl. of Comput. Vis.*, 2022, pp. 574–584.

[7] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia and Z. Wang, “Attention Unet++: A nested Attention-aware u-net for liver CT image segmentation,” in *Proc. IEEE Int. Conf. Image Process*, 2020, pp. 345–349.

[8] Q. Lyu, D. Xia, Y. Liu, X. Yang and R. Li, “Pyramidal convolution Attention generative adversarial network with data augmentation for image denoising,” *Soft Comput.*, vol. 25, no. 14, pp. 9273–9284, 2021.

[9] Z. Wang and H. Delingette, *Attention for image registration (air): An unsupervised transformer approach*, 2021. arXiv: 2105.02282.

[10] S. Li, Y. Ma and H. Wang, “3d medical image registration based on spatial Attention,” in *Proc. Int. Conf. Video and Image Process.*, 2020, pp. 98–103.

[11] O. Bernard, A. Lalande, C. Zotti *et al.*, “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?” *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, May 2018.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “Image Quality Assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[13] H. Zhao, O. Gallo, I. Frosio and J. Kautz, “Loss functions for image restoration with neural networks,” *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[14] P. Y. Simard, D. Steinkraus and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2003, pp. 958–963.

[15] A. Hering, S. Kuckertz, S. Heldmann and M. P. Heinrich, “Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking,” in *Proc. Bildverarbeitung für die Medizin*, ser. Informatik Aktuell, 2019, pp. 309–314.

[16] T. Vercauteren, X. Pennec, A. Perchant and N. Ayache, “Non-parametric diffeomorphic image registration with the Demons algorithm,” in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, 2007, pp. 319–326.

[17] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.