Medical Image Analysis 77 (2022) 102355

Contents lists available at ScienceDirect

Medical Image Analysis

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

Distortion and instability compensation with deep learning for rotational scanning endoscopic optical coherence tomography

Guiqiu Liao[a,b,*], Oscar Caravaca-Mora[a], Benoit Rosa[a], Philippe Zanne[a], Diego Dall'Alba[b], Paolo Fiorini[b], Michel de Mathelin[a], Florent Nageotte[a], Michalina J. Gora[a]

[a] ICube, UMR 7357 CNRS-University of Strabourg, Strasbourg, France
[b] Department of Computer Science, University of Verona, Verona, Italy

ARTICLE INFO

Article history:
Received 28 June 2021
Revised 22 December 2021
Accepted 6 January 2022
Available online 22 January 2022

MSC:
41A05
41A10
65D05
65D17

Keywords:
Optical coherence tomography
Endoscopic catheter
Image correction
Video stabilization
Convolutional neural network

ABSTRACT

Optical Coherence Tomography (OCT) is increasingly used in endoluminal procedures since it provides high-speed and high resolution imaging. Distortion and instability of images obtained with a proximal scanning endoscopic OCT system are significant due to the motor rotation irregularity, the friction between the rotating probe and outer sheath and synchronization issues. On-line compensation of artefacts is essential to ensure image quality suitable for real-time assistance during diagnosis or minimally invasive treatment. In this paper, we propose a new online correction method to tackle both B-scan distortion, video stream shaking and drift problem of endoscopic OCT linked to A-line level image shifting. The proposed computational approach for OCT scanning video correction integrates a Convolutional Neural Network (CNN) to improve the estimation of azimuthal shifting of each A-line. To suppress the accumulative error of integral estimation we also introduce another CNN branch to estimate a dynamic overall orientation angle. We train the network with semi-synthetic OCT videos by intentionally adding rotational distortion into real OCT scanning images. The results show that networks trained on this semi-synthetic data generalize to stabilize real OCT videos, and the algorithm efficacy is demonstrated on both *ex vivo* and in vivo data, where strong scanning artifacts are successfully corrected.

1. Introduction

Optical coherence tomography (OCT) (Huang et al., 1991) is increasingly used in biomedical and clinical imaging because of its high-speed and high resolution optical sectioning (Yonetsu et al., 2013). A one-dimensional (1D) image, called A-line, is obtained by pointing an OCT light beam onto the tissue. The OCT light propagates up to few millimeters within the tissue and is reflected back by the internal tissue structure to the imaging system. A standard two-dimensional (2D) OCT frame, called B-scan, is created by moving the light beam in a plane. In ophthalmology, which is the most common application of OCT, the OCT beam is typically raster scanned over a square field of view to create a three-dimensional (3D) volume. 2D images are displayed in real time and the volume is also typically visualized as an en-face projection to provide orientation and to follow disease progression longitudinally (Costello, 2017). When combined with a miniaturized optical catheter, OCT light can also be delivered into the cardiovascular, respiratory or digestive systems for imaging of internal structures (Gora et al., 2017). Such catheters usually require an outer diameter smaller than 2 mm and a length of up to 2 m. To enable volumetric imaging of tubular organs, in the majority of the designs, a side-viewing micro-optics is simultaneously rotated and pulled back within a surrounding static sheath to create a helical scan. In cardiology, 2D radial OCT frames are displayed in real-time during the longitudinal pullback to assist cardiologists in intravascular stent strut placement (Nam et al., 2016). In gastroenterology, OCT frames are also reviewed in real-time to find suspicious lesions and consequently to guide biopsy collection (Suter et al., 2014). Recently, real-time OCT guidance during endoscopic submucosal dissection has been proposed (Mora et al., 2020).

The quality of beam scanning in endoscopic OCT strongly depends on the actuation mechanism. To effectuate the helical motion of the probe, a scanning device can be placed either at the proximal side (outside of the patient) (Nam et al., 2016; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015) or at the distal end (Tran et al., 2004; Wang et al., 2013; Herz et al., 2004). Compared with distal-scanning OCT systems, proximal-scanning probes are more

* Corresponding author.
E-mail address: liao.guiqiu@etu.unistra.fr (G. Liao).

https://doi.org/10.1016/j.media.2022.102355

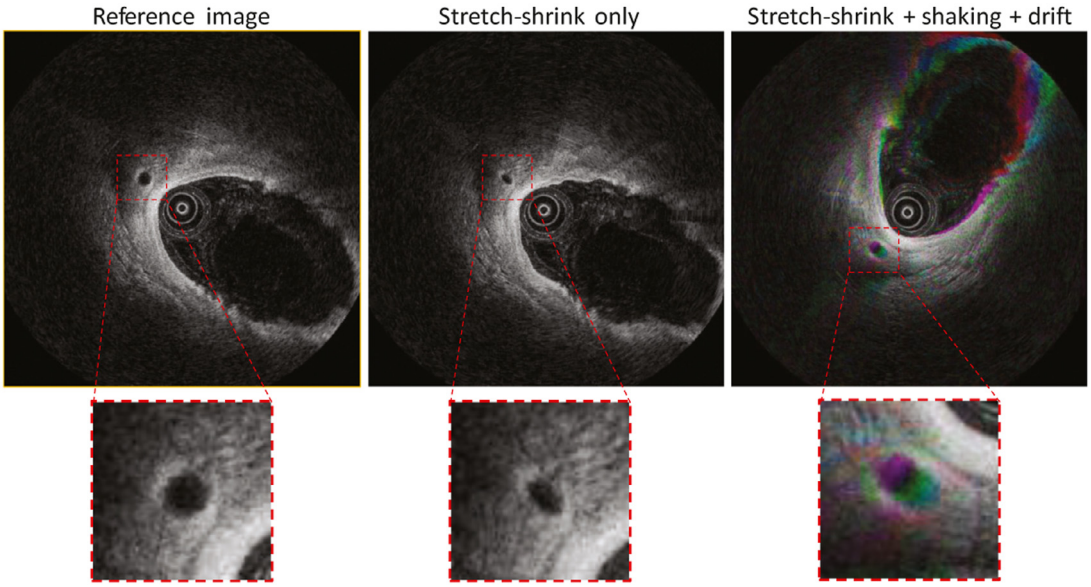

Fig. 1. Illustration of distortion and instability in endoscopic OCT systems. First column: a selected reference IVOCT frame (Wang et al., 2015) with considerable geometry accuracy. Middle column: An OCT frame distorted by stretch-shrink A-line level orientation error. Third column: Situation when both distortion, shaking and drift artifacts exist. To highlight presence of artifacts, three consecutive frames were assigned to one of three channels of the Red, Green& Blue (RGB) image and overlapped (third column).

compact (Gora et al., 2013) and easier to be miniaturized (Abouei et al., 2018).

Both scanning approaches typically suffer from image distortions, which hamper image reconstruction and interpretation. Such distortions are often referred to as Non-Uniform Rotational Distortion (NURD) in the literature, while in fact NURD encompasses several distinct phenomena. Within-frame *stretch and shrink* distortions are an A-line level rotation non-linearity within a B-scan image in the polar domain (Mavadia-Shukla et al., 2020; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015). In proximal scanning OCT, they are usually caused by mechanical friction during bending of the catheter, which in turn affects the transmission of rotation from the proximal actuator to the distal focusing optics typically realized using a torque coil. In distal scanning, it is usually much less prominent and is typically linked to the mechanical design and short term stability of the motor speed. Between-frames *shaking* and *drift* distortions are present in both proximal and distal scanning approaches, and are caused by variations of the motor speed (both in the proximal actuator or at the distal tip), and/or by synchronization errors between the acquisition of images and the scanning speed. Such synchronization problems are also common in raster scanning systems (Ricco et al., 2009).

Within-frame and between-frames distortion/artifacts reduce the image quality and introduce geometry changes (see Fig. 1), which impair correct recognition and diagnosis of anatomical structures of interest. Because it is almost impossible to eliminate all these artifacts by hardware improvements (i.e. the friction between the rotational optical components and the protecting sheath cannot be completely eliminated), computational approaches are required to correct the raw images acquired by OCT systems. Earlier than for OCT, NURD was investigated in Intra-vascular ultrasound (IVUS) (Sathyanarayana, 2006; Kawase et al., 2007; Gatta et al., 2009). IVUS is a standard of care for cardiovascular imaging that also requires rotational scanning. In the work of Kawase et al. (2007) frequency analysis of the texture of the IVUS image was used to estimate the rotational speed. Cross-correlations between image blocks in different IVUS frames was used to track image appearance changes caused by NURD (Gatta et al., 2009). This local feature, marker-free matching based method for IVUS was eventually adapted to OCT, using A-line distance (van Soest et al., 2008) or image block correlation (Uribe-Patarroyo and Bouma, 2015; Abouei et al., 2018). These iterative matching based methods, however, suffer from accumulating residual error. Therefore they cannot track the A-line level position error for long scans and are not applicable to the drift problem. However, the between-frames distortion can be solved by providing a physical reference point in each B-scan of the frame stream. Ahsen et al. (2014) achieved that by adding extrinsic markers on the OCT sheath, and tracking the overall shifting with image features of the makers. However, the markers block the OCT light and thus remove information about tissue. Intra-vascular stents can also be used as landmarks that help to register the rotational distortion in OCT pullback videos, which makes this method only applicable in stent strut assessment tasks (Ughi et al., 2012). Recently, a correction algorithm based on space-frequency analysis was proposed for endoscopic OCT to remove repeated A-lines caused by an extreme occurrence of the stretch-shrink distortion, called stick-slip effect of the torque coil (Mavadia-Shukla et al., 2020). However, this algorithm is not designed for stretch and shrink distortion when the rotation non-linearity is not so strong and no repeated A-lines can be seen.

In the computer vision field, deep learning based methods have been applied to solve off-line or on-line white light camera video instability problems (Wang et al., 2018; Huang et al., 2017; Gast and Roth, 2019), with state of the art efficiency. Deep learning has been recently applied to OCT image processing, by using Convolutional Neural Network (CNN) for tissue layer segmentation (van der Putten et al., 2019; Li et al., 2019; Yong et al., 2017), classification (van der Putten et al., 2020) and cancer detection (Zeng et al., 2020), but not for OCT video stabilization.

In this article, a CNN based method is proposed to reduce *shaking* and *drift* NURD artifacts in OCT videos. While it is not focused on *stretch-shrink*, such artifacts may also be eliminated if they are transient. We introduce a dual-branch architecture to estimate the A-line level positions errors with respect to a given reference frame (see Fig. 2). In the first branch, to estimate a A-line level shifting vector, a correlation matrix between axial scanning lines in the latest image and the previous one is calculated (van Soest et al., 2008; Abouei et al., 2018; Gatta et al., 2009).

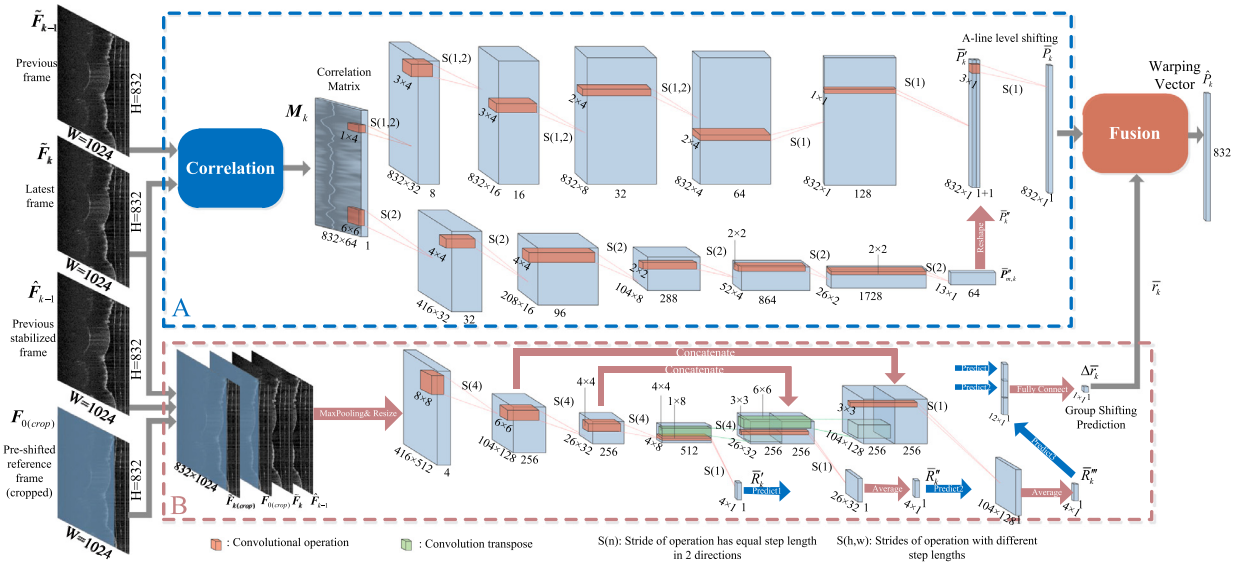

Fig. 2. Scheme of the proposed two-branch algorithm architecture for rotational distortion warping vector estimation. Branch (A) in blue dashed block estimates the shifting vector with an input of image pair, and branch (B) in red dashed block estimates the group rotation from the newest frame to reference with an image array as input.

Inspired by the boundary contour detection algorithms based on CNN (Maninis et al., 2017), we designed a network to find an optimal path within the computed correlation matrix, which represents the shifting angle of each individual A-line. A similar problem can be found in the inertial navigation field, where the rotation angle is iteratively computed with data from a gyroscope. The gyroscope provides a type of relative measurement and introduces accumulating error. A typical solution for this problem is to fuse direct angular measurements (coming from an accelerometer) with the indirect measurements (gyroscope) (Mahony et al., 2005). Inspired by this, another CNN branch estimating overall orientation is separated from the shifting vector estimation. The network design of this orientation/group rotation estimation is also inspired by a method that applied deep neural network to estimate homographic transformation for sports camera video stabilization (Wang et al., 2018). A multi-scale estimation strategy using both local and global features is applied, which has been designed for estimating optical flow between frames in video sequences (Ilg et al., 2017; Dosovitskiy et al., 2015). The shifting vector and the group rotation estimation branches are running in parallel and are deployed to correct the OCT images online: at a given latest time step k, only past information from time steps $[0, \ldots, k]$ is needed.

Training the proposed networks requires a dataset with annotated, clinically relevant OCT images. Such a dataset is however not readily available, since it is almost impossible to manually annotate the non-uniform shifting for each frame of OCT videos. Few reliable approaches exist for generating complex, realistic synthetic OCT images. Therefore, we trained our networks with semi-synthetic OCT videos generated by randomly adding realistic warping vectors and group rotation values to real OCT images. We then deployed the networks for real OCT videos stabilization.

A summary of our contributions is as follows:

- We propose a stabilization method to correct geometry information on the fly when the OCT system is capturing scanning data, that is beneficial for efficient online diagnosis.
- A robust deep CNN architecture is designed to estimate the A-line level distortion error.
- A drift compensation method inspired from inertial navigation is developed for rotational scanning stabilization.
- We trained the networks on semi-synthetic scans generated by adding distortion to real images, which avoids the need of manual annotation.
- We assessed the performance of the proposed method with unseen in vivo pre-clinical and clinical data.

2. Methods

A rotational scanning OCT catheter captures a continuous stream of A-lines. To reconstruct full images (i.e. B-scans), one typically makes the assumption that the optical components at the distal tip of the fiber are rotating with an ideal constant speed. Under this assumption, the OCT data acquisition system arranges H equally-spaced A-lines to cover a 360 degrees region in polar coordinates. We consider a reference frame $\boldsymbol{F}_0$ acquired at the start of the correction algorithm.

The newest frame $\tilde{\boldsymbol{F}}_k$ (in this article k indicates the index of newest data or results) is composed of H A-lines A_k^i ($i \in [0, H)$). The index i represents the position of a given A-line A_k^i in the image in polar domain. Because of the scanning artifacts, A_k^i differs from its correct position which should be aligned to A_0^j in frame $\boldsymbol{F}_0$. The position error of A-line A_k^i is expressed as $\epsilon_k^i = j - i$, and composes one element of an error vector $P_k = [\epsilon_k^0 \cdots \epsilon_k^i \cdots \epsilon_k^H]^T$.

OCT video stabilization consists in minimizing the position error P_k of A-lines in the latest raw frame $\tilde{\boldsymbol{F}}_k$. Note that throughout this article, the tilde ˜, the bar ¯ and the hat ˆ are used to denote a raw value (original measurement), a prediction and an estimation respectively. Given a position error vector P_{k-1} for the previous frame and A-line level shifting vector $\bar{P}_k$ between the two raw frames $\tilde{\boldsymbol{F}}_{k-1}$ and $\tilde{\boldsymbol{F}}_k$, each element of the latest A-line position error P_k can be obtained with an iterative computation operation Φ, as follows:

$$P_k^i = \Phi^{(i)}(\bar{P}_k, P_{k-1}) = \bar{P}_k^i + P_{k-1}^j \tag{1}$$

$$j = \bar{P}_k^i + i \tag{2}$$

Using these definitions, the previously mentioned *stretch-shrink, shaking* and *drift* problems can be described in terms of values in the relative/indirect between-frame shifting vector $\bar{P}_k$ (instead of using the direct error vector P_k). One can write $\bar{P}_k = \Delta\bar{r}_k \mathbf{1} + \bar{P}_{a,k}$, where $\mathbf{1}$ is a vector of ones, $\bar{r}_k$ is an overall rotation error with respect to the reference frame. The scalar $\Delta\bar{r}_k$ contributes to the frame level dynamic shift with respect to the first frame, and the vector $\bar{P}_{a,k}$ is a non-uniform A-line level shifting part. The *stretch-shrink* distortion is represented by $\bar{P}_{a,k}$, and constitutes nonlinear

displacement of individual A-lines in the polar domain within one frame. On the other hand, the *shaking* and *drift* is linked to the between-frames shifting $\Delta\bar{r}_k$. One should note that it is the variation of $\bar{r}_k$ in time (i.e. between frames) that constitutes the *shaking and drift* phenomenon. Eventually, the position error of each A-line in one frame can be expressed as $P_k = \sum_{n=1}^{k} \Delta\bar{r}_n \mathbf{1} + \Phi(\bar{P}_{a,k}, P_{a,k-1})$. Similarly to equation (1), $\Phi(\bar{P}_{a,k}, P_{a,k-1})$ is computed from $\bar{P}_{a,1}$. The accumulation of successive non-zero values will provoke a drift, while quick variations of individual values of $\Delta\bar{r}_k$ from one image to the next model the *shaking* phenomenon. Finally, note that computing P_k from the estimated $\bar{P}_k$ could accumulate estimation errors, which could lead to an even more notable drift. This type of issue also exists when iteratively computing the shifting error vector between latest frame and previous corrected frame, due to the residual correction error. In the following subsection we introduce a solution for estimating the A-line level shifting error considering these problems.

2.1. Algorithm pipeline

The proposed distortion and instability compensation algorithm has a two-branch architecture. As shown in Fig. 2, the upper branch (A) is designed to estimate the non-uniform warping vector between two consecutive frames. In each iteration of the algorithm, the latest original OCT image $\tilde{\boldsymbol{F}}_k$ and the previous buffered original frame $\tilde{\boldsymbol{F}}_{k-1}$ enter a correlation module, and a correlation matrix $\boldsymbol{M}_k$ is calculated. Then a CNN estimates the shifting vector $\bar{P}_k$ from $\boldsymbol{M}_k$. One direct way to correct the distortion is to calculate the position error vector P_k by the iterative computation Φ (see eq. 1), and then apply each element of P_k to shift each A-line of OCT frame $\tilde{\boldsymbol{F}}_k$. This works for a temporary period, but the estimation error accumulates along the processing time.

Similar to how the accelerometers are used to solve the accumulative error of the gyroscope iterative computation, another CNN branch (B) (shown in red dashed block of Fig. 2) is proposed to estimate a direct group rotation value $\bar{r}_k$. Running in parallel with branch (A), the input of the lower branch (B) is composed with the newest frame $\tilde{\boldsymbol{F}}_k$, previous corrected frame $\hat{\boldsymbol{F}}_{k-1}$ and the reference frame $\boldsymbol{F}_0$. $\boldsymbol{F}_0$ is cropped to remove the area outside the OCT sheath. This allows to take into account only the constant features corresponding to the sheath, which will not be affected by the outside environment. The relation between $\tilde{\boldsymbol{F}}_k$ and $\hat{\boldsymbol{F}}_{k-1}$ can also reflect the group rotation and these complete frames provide more features than sheath images. However, using only these two frames will introduce an iterative drift. Alternatively, by combining the 3 frames as an input, branch (B) can estimate a robust and smooth group rotation value.

After each algorithm iteration, the group rotation value $\bar{r}_k$ is fused with the warping vector $\bar{P}_k$, and a new estimation of warping vector $\hat{P}_k$ is obtained. $\hat{P}_k$ is applied to shift each specific axial line of $\tilde{\boldsymbol{F}}_k$ to get a corrected frame $\hat{\boldsymbol{F}}_k$. Details of the two-branch CNNs and fusion are presented in subsections 2.2, 2.3 and 2.4.

2.2. A-Line level shifting error interpretation

To reflect the angular mismatch between the latest frame $\tilde{\boldsymbol{F}}_k$ and the previous frame $\tilde{\boldsymbol{F}}_{k-1}$, we compute the correlation between local image rectangular patches from the latest frame and the previous one.

As shown in Fig. 3, the correlation matrix is obtained in the polar domain. One image patch $\boldsymbol{f}_i$ with dimension $h \times W \times 1$ ($h \ll H$, W is the width of the OCT frame, and h depends on the noise level of image, for example $h = 3$ is a practical value) centered at index position i ($i \in [0, H)$) of the newest frame $\tilde{\boldsymbol{F}}_k$ is used for shifting correlation with w image patches $\boldsymbol{f}'_{i-w/2+j}$ in a window of the previous frame $\tilde{\boldsymbol{F}}_{k-1}$, where $j \in [0, w)$. Each shifting operation outputs one array m_i, which composes one row of a correlation matrix $\boldsymbol{M}_k$. $\boldsymbol{M}_k$ has width w that is equal to the shifting window, and height H equal to the height of $\tilde{\boldsymbol{F}}_k$ in polar coordinates. The value of w is a parameter that depends on the maximum shifting error, which is discussed in the experiment section. For display reasons, the correlation matrices shown in this article are transformed by $255 \times (1 - \boldsymbol{M}_k)$ (the warped "valley" in the center of the demonstration correlation matrix is marked out with a white line in Fig. 3). If there is no rotational artifact in data stream, $\boldsymbol{M}_k$ should have a straight "valley-like" minimum region in the centre. We used the Pearson correlation coefficient $o_{i,j}$ to reflect the similarity between two image patches $\boldsymbol{f}_i$ and $\boldsymbol{f}'_j$:

$$o_{i,j} = \frac{\sum_{l=1}^{n} \boldsymbol{f}_{i,l} \boldsymbol{f}'_{j,l} - n\bar{\boldsymbol{f}}_i \bar{\boldsymbol{f}}'_j}{\sqrt{\sum_{l=1}^{n} \boldsymbol{f}_{i,l}^2 - n\bar{\boldsymbol{f}}_i^2} \sqrt{\sum_{l=1}^{n} \boldsymbol{f}_{j,l}'^2 - n\bar{\boldsymbol{f}}_j'^2}} \tag{3}$$

where the pixel index l operates through the rectangular patch $n = h \times W$. $\bar{\boldsymbol{f}}_i$ and $\bar{\boldsymbol{f}}'_j$ are the mean values of patch $\boldsymbol{f}_i$ and $\boldsymbol{f}'_j$ respectively. To get one element $o_{i,j}$ of correlation matrix $\boldsymbol{M}_k$, $3 \times w \times n^2$ multiplications are operated, thus the correlation matrix calculation for one frame needs $3 \times H \times W^2 \times h^2 \times w$ multiplications. Converting the correlation operation into matrix (or, equivalently, tensor) operations (Jia et al., 2014) is a standard way for computation acceleration, and is for the convenience of CNN input as well.

Before the operation of shifting correlation, 2 stacks (or, equivalently, 2 tensors) $\boldsymbol{S}, \boldsymbol{S}' \in \mathbb{R}^{H \times w \times h \times w}$ are created for correlation acceleration. $\boldsymbol{S}$ and $\boldsymbol{S}'$ stack the image patches of current frame and previous frame as shown in Eq. (4) and Eq. (5).

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{f}_H & \boldsymbol{f}_H & \cdots & \boldsymbol{f}_H \\ \boldsymbol{f}_{H+1} & \boldsymbol{f}_{H+1} & \cdots & \boldsymbol{f}_{H+1} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{f}_{2H} & \boldsymbol{f}_{2H} & \cdots & \boldsymbol{f}_{2H} \end{bmatrix} \tag{4}$$

$$\boldsymbol{S}' = \begin{bmatrix} \boldsymbol{f}'_{H-w/2} & \boldsymbol{f}'_{H-w/2+1} & \cdots & \boldsymbol{f}'_{H+w/2} \\ \boldsymbol{f}'_{H+1-w/2} & \boldsymbol{f}'_{H+2-w/2} & \cdots & \boldsymbol{f}'_{H+1+w/2} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{f}'_{2H-w/2} & \boldsymbol{f}'_{2H-w/2+1} & \cdots & \boldsymbol{f}'_{2H+w/2} \end{bmatrix} \tag{5}$$

Since the OCT image stream is acquired by a continuous circular scanning, the generation of $\boldsymbol{S}'$ covers 2 areas with $w/2$ A-lines from the edge of $\tilde{\boldsymbol{F}}_{k-2}$ and $\tilde{\boldsymbol{F}}_k$ respectively, in addition to $\tilde{\boldsymbol{F}}_{k-1}$. So $\boldsymbol{f}'_i$ in Eq. (5) is sampled from an extended image $\boldsymbol{F}'_L = [\tilde{\boldsymbol{F}}_{k-2}, \tilde{\boldsymbol{F}}_{k-1}, \tilde{\boldsymbol{F}}_k]$ which concatenates $\tilde{\boldsymbol{F}}_{k-2}, \tilde{\boldsymbol{F}}_{k-1}$ and $\tilde{\boldsymbol{F}}_k$. The strategy is similar for $\boldsymbol{S}$. Because one frame is corresponding to one cycle of circular scanning, the image patch in the bottom can copy the top A-lines of $\tilde{\boldsymbol{F}}_k$ when $\boldsymbol{f}_i$ exceeds the boundary, which means that $\boldsymbol{f}_i$ in Eq. (4) is sampled from $\boldsymbol{F}_L = [\tilde{\boldsymbol{F}}_{k-1}, \tilde{\boldsymbol{F}}_k, \tilde{\boldsymbol{F}}_k]$, where $\tilde{\boldsymbol{F}}_k$ is reused in the concatenation. This way, $\boldsymbol{M}_k$ is obtained by 7 multiplications and additions between tensors.

2.3. Shifting vector estimation

The correlation matrix provides a general interpretation of the angular matching likelihood between image patches at different positions. We propose a CNN based approach to finally estimate the shifting vector for image correction.

As shown in the blue dashed block (A) in Fig. 2, first $\boldsymbol{M}_k$ is computed with a predefined shifting window (in OCT videos the estimated maximum error value is 15 pixels in the polar domain, but we increased the margin to ensure the robustness and set the correlation window as $w = 64$). Then two convolution sub-branches with different strides extract features from $\boldsymbol{M}_k$ in parallel and produce hierarchically coarse-to-fine responses.

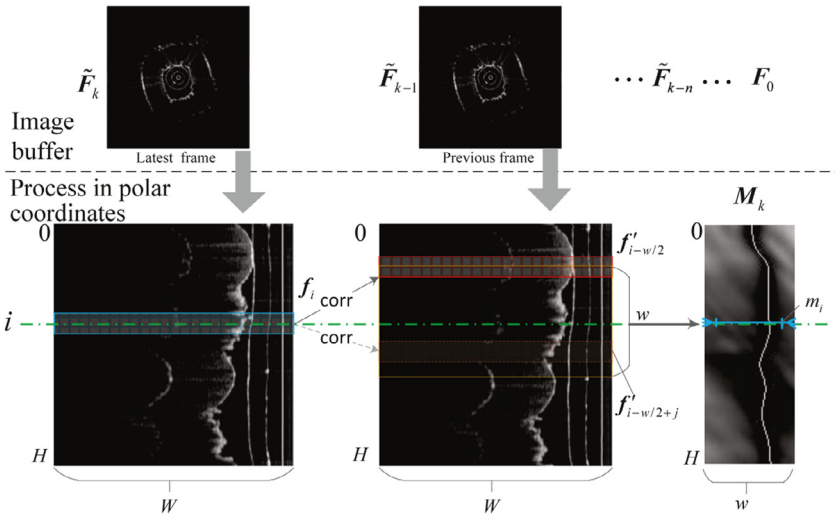

Fig. 3. Correlation operation between adjacent frames. In the upper part of the figure, the images are shown in the Cartesian coordinate system for intuitive visualization. For the angular distortion correction, the images of the sequence are buffered and processed in polar domain.

Both the upper sub-branch and the lower sub-branch of shifting vector estimation nets have 6 convolutional layers, and a LeakyReLU activation (Maas et al., 2013) is used after each convolution layer.

The upper sub-branch has unequal strides size and rectangular convolution kernels (from 1*st* layer to 5*th* layer), to involve more information in horizontal direction than the vertical direction. Importantly, this sub-branch always keeps the vertical stride as 1, which emphasizes the spatial correspondence (information at/around each row of $\boldsymbol{M}_k$ represents the angular shift information of $\tilde{\boldsymbol{F}}_k$ at the same A-line position). By doing so, the front 5 feature extraction layers can gradually reduce the feature map width from 64 to 1, while maintaining the feature map height H as input's height. The depth of each convolution operation's output is twice as deep as its input (here we set the output depth of the first layer as 8). The 5*th* feature map $\boldsymbol{A}_F^5 \in \mathbb{R}^{832\times1\times128}$ extracts 128 local features, which could include the minimal value position, edge, and boundary position. A final layer with kernel size 1×1 and channel depth 128, reorganizes the 5*th* feature map and decrease channels to a sub-branch output $\tilde{P}'$ with size 832×1.

In the ideal situation where the correlation matrix has a good quality (when calculated with images having dense features), $\tilde{P}'$ can represent the azimuthal mismatching between $\tilde{\boldsymbol{F}}_k$ and $\tilde{\boldsymbol{F}}_{k-1}$. However, sometimes $\boldsymbol{M}_k$ can miss valid information for some row m_i when there is no feature in a patch (window) $\boldsymbol{f}_i$ of $\tilde{\boldsymbol{F}}_k$. In this situation, since the estimation $\tilde{P}'$ has low spatial correlation in the vertical direction, the azimuthal distortion estimation at point i of $\tilde{P}'$ can have a significant error. Inspired by the inception module of GoogLeNet (Szegedy et al., 2015), we introduce another sub-branch that loosens the stride step length in the vertical direction to 2, expanding the involved vertical spatial information in every convolution. In each convolution operation of this sub-branch, the output depth is 3 times the input depth. This form of design has been widely used in CNN to extract high-level abstract features from images (Simonyan and Zisserman, 2014). Compared with the upper sub-branch, this lower sub-branch will extract a high-level feature map $\boldsymbol{A}_{F2}^5 \in \mathbb{R}^{26\times1\times1728}$, which is less sensitive to noise and high intensity speckle artifacts. A final layer with kernel size 2×2 re-organizes this feature map, and outputs a matrix $\tilde{\boldsymbol{P}}_m''$ of size 13×64. This matrix contains 13 groups of *path position* information, which represent the warping paths of 13 connected small patch areas (size 64×64) of $\boldsymbol{M}_k$.

The lower sub-branch output $\tilde{\boldsymbol{P}}_m'' \in \mathbb{R}^{13\times1\times64}$ is reshaped to $\tilde{P}'' \in \mathbb{R}^{832\times1}$ with less dimensions by connecting all 1×64 rows. $\tilde{P}''$ is concatenated to the upper sub-branch output $\tilde{P}'$, and then it is operated by a 3×1 convolution kernel (with zero padding on the edges), to provide the final estimation vector $\tilde{P}$ of adjacent frames.

The loss function for training the shifting vector estimating nets uses the conventional L_2 loss function and a self-designed *continuity loss* function. A standard L_2 loss is described by:

$$L_2 = \frac{1}{n_p}\sum_{i=1}^{n_p}\left(P_i - \tilde{P}_i\right)^2 \tag{6}$$

where P_i is an element of the true shifting vector P (ground truth), and $n_p = 832$ is the vector length. The L_2 loss function is commonly used for value estimation, while for this estimation task, to take into account the prior knowledge on continuity of distortion vector (van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015), a continuity loss is added as follows:

$$L_c = \frac{1}{n_p - 1}\sum_{i=1}^{n_p-1}\left(\tilde{P}_{k,i} - \tilde{P}_{k,i+1}\right)^2 \tag{7}$$

By calculating L_c, and combining it with L_2 in the network training, the attraction towards local minima with discontinuous vector estimation will be suppressed. The final loss for branch (A) is:

$$L_A = \alpha L_c + (1-\alpha)L_2 \tag{8}$$

where α gradually decreases from a large value to a smaller value in the training process (see training details in Section 3).

2.4. Accumulative error compensation

2.4.1. Group rotation estimation

The CNN branch (B) (red dashed box in Fig. 2) estimates an overall rotation from an image array.

This branch consists of a contracting path, an expansion path, and a fully connected layer. There are two encoder layers (indicated by convolution in red color) in the contracting path and two decoder layers (indicated by convolution transpose in green color) in the expansion path, and both the encoder and decoder layers are connected with LeakyRelu activation.

The encoder layers are used for learning the contextual feature hierarchy. On the other hand, the decoder layers use transpose convolution (also referred as up-convolution (Long et al., 2015)) to

perform the refinement, and they are concatenated with the corresponding encoder blocks. In this way, the multi-scale information passed from low level local feature maps to high-level coarser feature maps is preserved. The difference in OCT videos is that the rotation (flow) only occurs in one dimension. Our method of reorganizing the three multi-scale feature maps is to apply three small kernels with 1×1 strides to reduce their channel depth from 512 to 1, and then apply average pooling to each fine local estimation to get equally resized $4 \times 1 \times 1$ estimations. By doing so, higher scale estimation $\bar{R}''$ and $\bar{R}'''$ are aligned to coarser estimation $\bar{R}'$. A fully connected layer is used to interpret the estimation from three scale levels to get a final robust estimation $\Delta\bar{r}_k$, and the overall rotation is obtained by $\bar{r}_k = \bar{r}_{k-1} + \Delta\bar{r}_k$.

The loss function for training the group rotation estimating nets in branch (B) is a multi-scale loss, because it should not only ensure the estimation accuracy in the final output of $\Delta\bar{r}_k$, but also maintain the accuracy of higher scale estimation in a certain level:

$$L_B = \beta_1 |\Delta r - \Delta\bar{r}_k| + \beta_2 \left|\Delta r - \Delta\bar{r}'_k\right| + \beta_3 \left|\Delta r - \Delta\bar{r}''_k\right| + \beta_4 \left|\Delta r - \Delta\bar{r}'''_k\right| \quad (9)$$

where $\Delta\bar{r}'_k$ is the mean of the 4×1 estimation vector $\bar{R}'$ extracted from the final encoder result, $\Delta\bar{r}''_k$ and $\Delta\bar{r}'''_k$ are the mean of estimation vectors resized from 26×32 map and 104×128 map respectively. The weights β_i are adjustable during the training process, but β_1 remains predominant (see training details in Section 3).

2.4.2. Fusion and online correction

The fusion of $\bar{P}_k$ and $\bar{r}_k$ can be considered as the problem of fusion between an integral indirect variable with high accuracy and another robust direct variable. Advanced filtering techniques to solve this kind of problem can rely on a form of probabilistic fusion like the extended Kalman filter, or alternatively use complementary filters (Allgeuer and Behnke, 2014). For computational efficiency and robustness, we use the concept of a *PI Complementary Filter* (Mahony et al., 2005) to fuse the $\bar{P}_k$ vector with the $\bar{r}_k$ value. The complementary filter has been widely used as an efficient way to fuse the data of gyroscopes and accelerometers, which combines high-pass easily drifting measurements with low-pass stable measurements to form a robust high bandwidth estimate of the rotational attitude (Mahony et al., 2005).

A discrete form of PI complementary filter for algorithm implementation can be expressed as:

$$\hat{P}_k = k_p \Phi(\bar{P}_k, \hat{P}_{k-1}) + (1 - k_p)\bar{r}_k \mathbf{1} + k_i I_k \quad (10)$$

$$I_k = I_{k-1} + (\bar{r}_k \mathbf{1} - \hat{P}_k) \quad (11)$$

where k_p and k_i are PI compensating gains. I_k is the integral component vector. $\mathbf{1}$ is a vector of ones. $\Phi(\hat{P}_{k-1}, \bar{P}_k)$ is the element-wise operation in formula (1). Each element $\hat{P}_{k,i}$ of the final warping vector $\hat{P}_k$ represents the angular shift between the position of the i^{th} A-line of $\tilde{\boldsymbol{F}}_k$ and its correct position in polar domain.

3. Dataset and training

In medical image processing, there is limited availability of open-source training sets due to ethical and practical reasons. It is even more complicated for the OCT artifacts, since it is hard to label the A-line level shifting within videos, and no public data set with ground truth is available. Using a calibration phantom might increase the accuracy of ground truth annotation. However, it will be difficult to manufacture a variety of such calibration phantoms covering different tissue or material types that allows to afterwards generalize to real tissues. For these reasons, we trained the networks of the proposed framework with semi-synthetic OCT videos by intentionally shifting each A-line in real OCT images (see details in subsection 3.2). In this way, the distribution of rotational distortion in the data can be adjusted to cover the real distribution, but the distribution of scanning noise is not simulated. To solve this, we used a variety of image augmentation strategies to mimic the real scanning noises (details in subsection 3.3.1).We test the trained networks on both semi-synthetic videos and real videos. Additionally, we collected in vivo pre-clinical and clinical OCT videos, which are not included in the training dataset, to evaluate the generalization and robustness of the framework to previously unseen data. This section describes the experimental setup, data generation and network training.

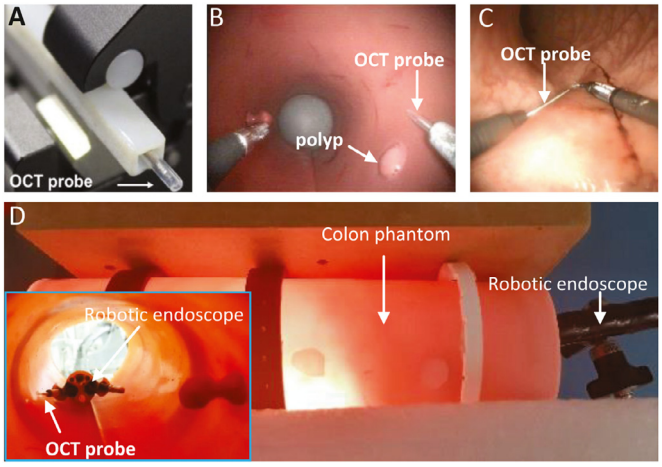

Fig. 4. Endoscopic OCT data acquisition. (A) The OCT probe inserted in the rectangular phantom. (B) A steerable OCT catheter is inserted in an instrument channel of a robotized interventional flexible colonoscope, and it is applied to scan a colon model. (C) The steerable OCT catheter is applied to in vivo testing of a swine colon. (D) The experimental setup of the colon model.

3.1. OCT Data sources

We have applied a data set synthesis strategy to generate training image sequences by intentionally distorting real OCT images. We used previously published data obtained with low-profile OCT catheters in the cardiovascular system (Wang et al., 2015) and the respiratory system (Lee et al., 2011), as well as with a capsule OCT catheter in the digestive tract (Gora et al., 2013) (5K images in total). In addition, OCT videos are also collected using a custom endoscopic OCT system with a proximal scanning (Mora et al., 2020). Volumetric OCT data was collected using an internal pullback of the probe (1K images) or by pulling back the whole sheath during 2D rotational scanning (1K images) in a rectangular phantom tube with a known geometry (Fig. 4(A)). The self-developed probe was also used for endoscopic examination of a colon phantom made with optical mimicking tissue (Zulina et al., 2021) (shown in Fig. 4(B)), where a continuous stream of 2D images (3K) with no pullback was displayed in real-time for inspection. We split all the OCT images (including published and self-collected videos) by 7: 2: 1 into training, validation, and testing data.

3.2. Semi-synthetic OCT for training

To train the *warping vector estimation nets* in branch-A, we generated image pairs, while to train *group rotation estimation nets* in branch-B, we generated image arrays.

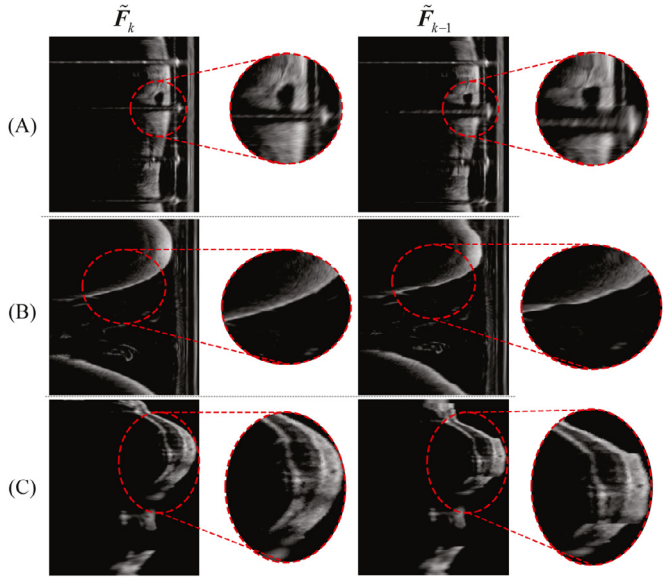

Fig. 5. OCT image pairs of the generated data set in polar domain. Local areas of each pair are enlarged to highlight the distortion. (A) An image pair generated from OCT image with fiducial markers (Wang et al., 2015), so that horizontal strips can be screened in the OCT image. (B) Ordinary OCT image pair without the marker. (C) In the source images of this pair, the sheath has been cropped out (Lee et al., 2011), but these images are still useful for algorithm training.

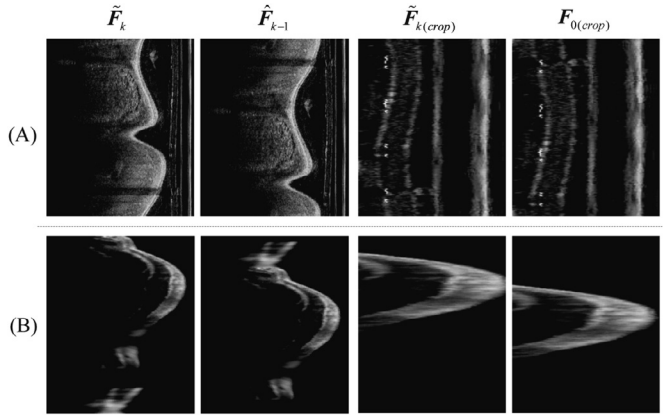

Fig. 6. OCT image arrays generated for training of group rotation nets, on each row from left to right are: latest raw frame $\tilde{\boldsymbol{F}}_k$ and previous stabilized frame $\hat{\boldsymbol{F}}_{k-1}$, cropped and resized latest frame $\tilde{\boldsymbol{F}}_k$, cropped and resized reference frame $\boldsymbol{F}_0$. (A) is an example of images with OCT sheath, so a normal cropping is used; (B) is an example of lung airway OCT images (Lee et al., 2011).

3.2.1. Image pairs with element wise shifting

To generate one training pair (two images) for the warping vector estimation network (branch-A), we first take one OCT image from the database as the initial image. Then each individual A-line within this initial image is shifted by a warping vector P_s. The distorted image is paired with the initial one as network input, while P_s performs as ground truth in training. Fig. 5 shows several training pair samples generated from public and self-acquired original OCT images. P_s is randomly drawn from a distribution that should be representative of distortions in real situations. This distribution is estimated by applying the Graphic Searching (GS) algorithm (Abouei et al., 2018) to real videos and measuring the warping vector $\bar{P}_t$. By doing so, an estimated maximum value m_t of rotational shifting is obtained. In our case, $m_t = 15$ pixels in the polar domain. To ensure proper coverage of extreme cases, we chose a maximum value $m_s = 25$ pixels.

Each element of the synthetic warping vector P_s is uniformly sampled in the $[-m_s, m_s]$ range. To guarantee the continuity of the synthetic warping vector, a 1D Gaussian filter is applied to smooth P_s, and the filtering parameter (sigma) is randomly chosen from 3, 5 or 7.

3.2.2. Image arrays with group rotation

The training set for group rotation contains image arrays and corresponding group rotation ground-truth values r_s. One input image array for the pure group rotation estimation nets is built from 3 images that are cropped and resized: the reference image $\boldsymbol{F}_0$, algorithm stabilized image $\hat{\boldsymbol{F}}_{k-1}$, and newest distorted image $\tilde{\boldsymbol{F}}_k$ (see Fig. 6). To generate such image array, first, one reference frame is directly selected from the original image database, the left part of $\boldsymbol{F}_0$ is cropped out to keep rightmost region of shape $H \times 0.2W$ of the image. As for mimicking the newest unstable frame, the reference frame $\boldsymbol{F}_0$ is distorted to $\tilde{\boldsymbol{F}}_0$ with a random warping vector $P_a = P_s - p_m$, where the mean value p_m of P_s is removed. Then the distorted $\tilde{\boldsymbol{F}}_0$ is rotated by a group rotation value r_s to get $\tilde{\boldsymbol{F}}_k$. In the acquired videos the estimated maximum rotation between two adjacent frames is 15 pixels in the polar domain image, and considering the estimation error, we set the rotation limitation to 35 pixels to cover the distribution and ensure robustness. r_s serves as the ground truth in the learning process. In the ideal situation, $\hat{\boldsymbol{F}}_{k-1}$ could be a copy of $\boldsymbol{F}_0$, however, $\hat{\boldsymbol{F}}_{k-1}$ is taken from the algorithm output where residual correction errors are expected. To prevent the networks from "over-trusting" the stabilized frame, a small random correction error value δ is used to shift the synthetic stabilized frame $\hat{\boldsymbol{F}}_{k-1}$ (the tuning of δ in training process is presented in subsection 3.3).

3.3. Training process

The training pipeline is implemented with Nvidia Qt1000 graphic card and Intel i5-9400H CPU. The code is implemented using the Pytorch framework (Paszke et al., 2017) for tensor operation and gradient backward propagation. We adopt the following implementation choices: Batch Normalization (BN) is used right after convolution and before activation (Ioffe and Szegedy, 2015), dropout is not used (Hinton et al., 2012) and weight initialization is performed following the method described in (He et al., 2015). The final result is hardly affected by the optimization method, both Adam (Kingma and Ba, 2014) and Stochastic Gradient Descent (SGD) solvers can fine tune the networks' weights. The results presented in this article are trained with the SGD weights optimization method (we used a weight decay of 0.0001 and a momentum of 0.9). We first pre-trained the networks on a small dataset to improve the efficiency of determining hyper-parameters and reducing time consumption (Bengio, 2009). We created two small training sets in order to train branch A and branch B, respectively. 16 images were randomly selected (4 from each of cardiovascular, digestive, lung, and colon phantom images), and 500 warping vectors P_s and shifting scalars r_s were randomly generated. In total, both sets feature 8000 image pairs and 8000 image arrays for warping vector learning and group rotation learning respectively. After the networks of the two branches converge on this small data set, training pairs and arrays are generated on-line - an image pair or array is never seen twice during training.

3.3.1. Data augmentation

Data augmentation is vital for machine learning algorithms to avoid over-fitting and to enhance robustness. We enable data augmentation on-line for training. Geometric transformations (shift in 2 directions, and scaling in polar domain) are applied equally to each image within image pairs or image arrays. For the group rotation training array's translation augmentation, the rightmost part

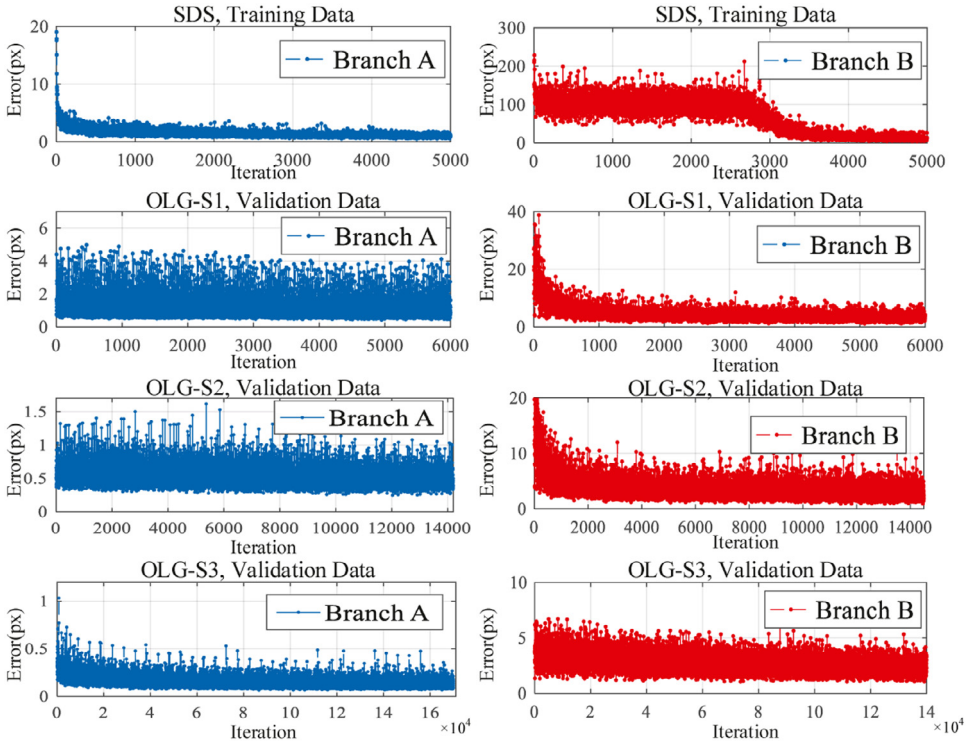

Fig. 7. The estimation error in the different training stages. The error is reported in pixel, and each pixel in polar coordinates corresponds to 0.432° The sub-plots in the top row are the training errors of the two branches with small data set (SDS). The average validation error in 3 stages (S1: stage1, S2: stage2, S3: stage3) of on-line data generation (OLG) are presented in the sub-plots below.

Table 1
Parameters values for the different training stages. (SDS: Small Data Set, OLG: On-Line Generating, S1: Stage 1 of OLG, S2: Stage 2 of OLG, S3: Stage3 of OLG; LR: Learning Rate, BS: Batch Size).

		SDS	OLG		
			S1	S2	S3
LR A		3×10^{-4}	3×10^{-5}	3×10^{-6}	1×10^{-8}
BS A		50	20	8	2
α		0.2	0.1	0.1	0.02
LR B		5×10^{-4}	5×10^{-5}	5×10^{-6}	1×10^{-8}
BS B		20	10	6	2
δ_m		0	0	0	±4.32◦
$\boldsymbol{\beta}$	β_1	0.25	0.3	0.4	0.5
	β_2	0.25	0.3	0.3	0.3
	β_3	0.25	0.2	0.15	0.1
	β_4	0.25	0.2	0.15	0.1

in the polar domain (the central part in Cartesian coordinates) is kept, to ensure that mainly sheath features exist in this area. Noise addition, and brightness and contrast modification are also applied to OCT images. This kind of pixel intensity modification is applied differently to each image of a generated pair or array.

3.3.2. Gradual parameter tuning

Besides the training mode switching strategies, several parameters are gradually changed from the beginning to the final fine-train stage. Table 1 gathers the parameters used initially for the small data set (SDS) and online data generation (OLG). The fine-training on data with on-line generating (OLG) is divided into 3 stages, where the learning rate, data batch size, max/min limitation δ_m of additional rotation δ, continuity loss weight and multi-scale loss weight $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4]$ are gradually modified. The training on the small data set takes 2 hours to converge, whereas the training with on-line data generation approximately takes 48 hours to flatten the variation of loss value. Fig. 7 shows estimation loss in different training stages. The sub-windows in the top row present the training loss of the two branches on the small limited data set, where the group rotation learning of branch (B) takes more time to converge compared with warping learning of branch (A). In the on-line data generating mode, we calculate the average estimating error after each iteration using generated image pairs and images arrays from the validation database, where the validation data batch size is equal to the training batch size. Each time when the average validation error converges to a small value, the parameters are tuned and the training pipeline switches to another training stage. The whole process reduces the average validation error of branch (A) and branch (B) to approximately 0.1 and 3 pixels respectively (1 pixel in polar domain represents 0.432° in Cartesian domain), and at the end of training stage 3 the gradient of the loss function is close to zero.

4. Results

All the trained CNN are deployed with Python codes on Ubuntu 18.04.4 system with the same computer used for training. The networks in branch (A) and (B) take 40 ms and 10 ms respectively in parallel mode, the correlation costs 96 ms with parallelization, and the fusion and warping process additionally take 9 ms. The processing time of an entire algorithm iteration is therefore 145 ms. After the network training, the correction algorithm is tested on both synthetic videos and real videos (on phantom and in vivo) to assess its performance.

4.1. Accuracy assessment

Half-synthetic videos for testing are generated with individual original OCT images, and each of them contains 501 frames. To generate one semi-synthetic video, one image is selected from the validation database to be the first frame, and then 500 warping

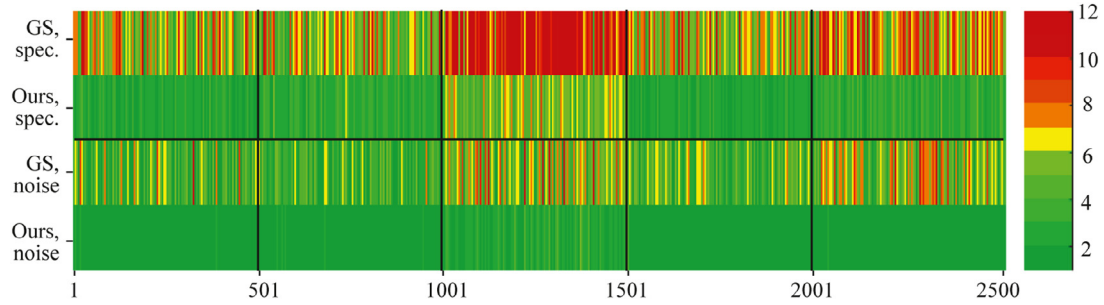

Fig. 8. Heatmap of warping vector estimation mean error (the unit of scale bar is pixel). The columns from left to right are from 5 groups of semi-synthetic videos generated with: cardiovascular, lung air way, digestive tract, rectangular phantom, and tissue phantom OCT images. The proposed method is compared to GS method against two conditions: mimicking high intensity A-line speckles, or adding noise (including Gaussian, pepper&salt and shot noise).

Table 2
Mean square error value in different synthetic video tests. The unit of all values is Pixel^2, each pixel in Polar coordinates represents 0.432°.

	GS		Proposed	
	Noise	Noise+Spec.	Noise	Noise+Spec.
Vascular	20.05±18.50	48.68±72.93	1.88±1.04	6.43±3.59
Air Way	35.39±64.89	58.27±105.2	3.98±2.45	9.24±5.60
Digestive	66.61±224.8	354.8±461.8	6.44±4.50	28.5±13.2
Phantom	19.57±16.51	41.40±39.86	1.21±0.73	5.23±2.56
Model	36.26±25.90	71.09±81.11	1.68±1.00	9.31±5.30

vectors P are randomly sampled with limit value of 8.65° (corresponding to 20 pixels), and 500 group rotation deviation values Δr are randomly sampled with varying limit values (for a period of a synthetic video, the group shift variation is limited to a positive value; while for another period, it is limited to a negative value). Then the first frame is iteratively rotated with Δr and then distorted with P to simulate a video stream.

In the state-of-the-art rotational artifacts correction algorithms, tracking based approaches (Abouei et al., 2018) are more suitable for the scenario when both stretch-shrink and shaking artifacts exist. Tracking based algorithms are less threshold sensitive in comparison to within-frame space frequency analysis based algorithms (Mavadia-Shukla et al., 2020), especially if there is no visible repeated A-lines. Based on these factors, we compare our proposed method to the GS based method (Abouei et al., 2018), that is capable of A-line level error estimation and correction.

The estimation Mean Square Error (MSE) value of each frame in videos is calculated by using true vectors as references, and the results are shown in Table 2. The proposed method is compared to GS under two conditions: adding noise (including Gaussian, pepper&salt and shot noise), and mimicking high intensity A-line speckles in every B-scan. The deep learning based algorithm surpasses the GS based method in all of these situations, and estimation errors are one or two orders of magnitude lower than the GS based method. Among these videos, the performance in digestive tract OCT suffers more from speckle artifacts due to the limited features in capsule OCT images, and also due to the reduced resolution in the available public videos. But still, the proposed method has lower MSE than GS method (9.24±5.60 vs. 354.8±461.8 pixel^2). A mean error heatmap of 5 videos in different scenarios are shown in Fig. 8, where estimation error of every individual frames (2500 frames in total) are presented. The GS method is affected by the addition of speckle artifacts, and more occasionally has significant estimation errors (larger than 12 pixels) in comparison to the proposed method, which maintains estimation errors under 3 pixels in most cases.

Fig. 9 shows examples of warping vector estimation within the 832 × 64 correlation matrices. The vector estimated by the proposed algorithm (red line) is closer to the ground truth vector (white line) than the vector obtained by the GS algorithm (green line). In the yellow dashed circles in Fig. 9, significant estimation error of the GS algorithm can be seen. The reason for this is that in the correlation matrix the "valley-like" feature which the GS algorithm highly relies on is not obvious. Cases (C) and (D) are more problematic for path searching, since some part of the original OCT image does not have adequate features for correlation. In these situations, the value of path searching diverges frequently from the true value. Nevertheless, the CNN estimated warping vector can still follow the ground truth.

We obtained mean value en-face projections (Abouei et al., 2018) of the OCT videos where each A-line is accumulated to one single value, so that the OCT data stream in polar domain are projected into 2D images. In this case the vertical Y axis corresponds to a circumferential scanning (B-Scan) and the horizontal X axis

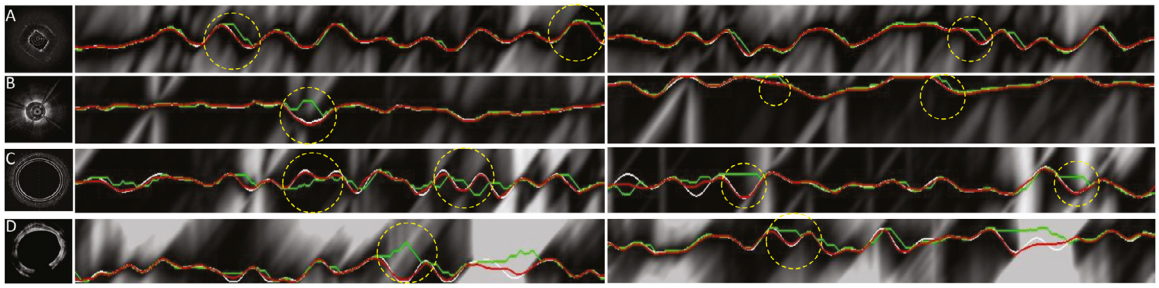

Fig. 9. Comparison of warping vector estimations. Images in each row (from left to right) are: the source image for video synthesis and two 832 × 64 correlation matrices of adjacent frames. In each correlation matrix, the white line indicates the ground truth vector, green line indicates the result of a GS algorithm(Abouei et al., 2018), and the red line indicates the estimation of the proposed algorithm. The dashed yellow circles highlight situations when the GS based method has larger error than the proposed method. Images from top to bottom are results of synthetic videos generated with different original images: (A) Rectangular phantom, (B) Cardiovascular system, (C) Digestive tract and (D) Lung air way OCT images.

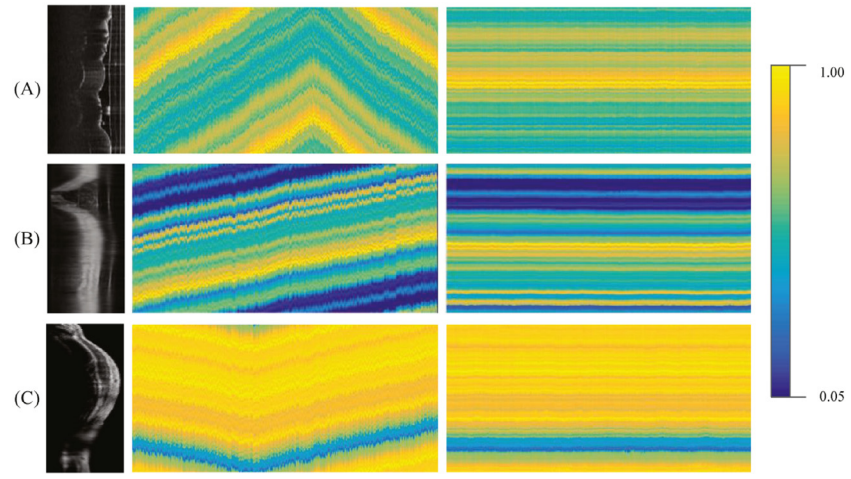

Fig. 10. En-face image comparison of synthetic videos before correction and after algorithm correction. The colorbar indicates the intensity scale normalized by the maximum value. Images in each row (from left to right) are: the source image for video synthesis (in polar coordinates), en-face image of synthetic video and corresponding en-face image of stabilized video. Images from top to bottom are results of synthetic videos generated with different original images: (A) Rectangular phantom, (B) Digestive tract and (C) Lung air way OCT images.

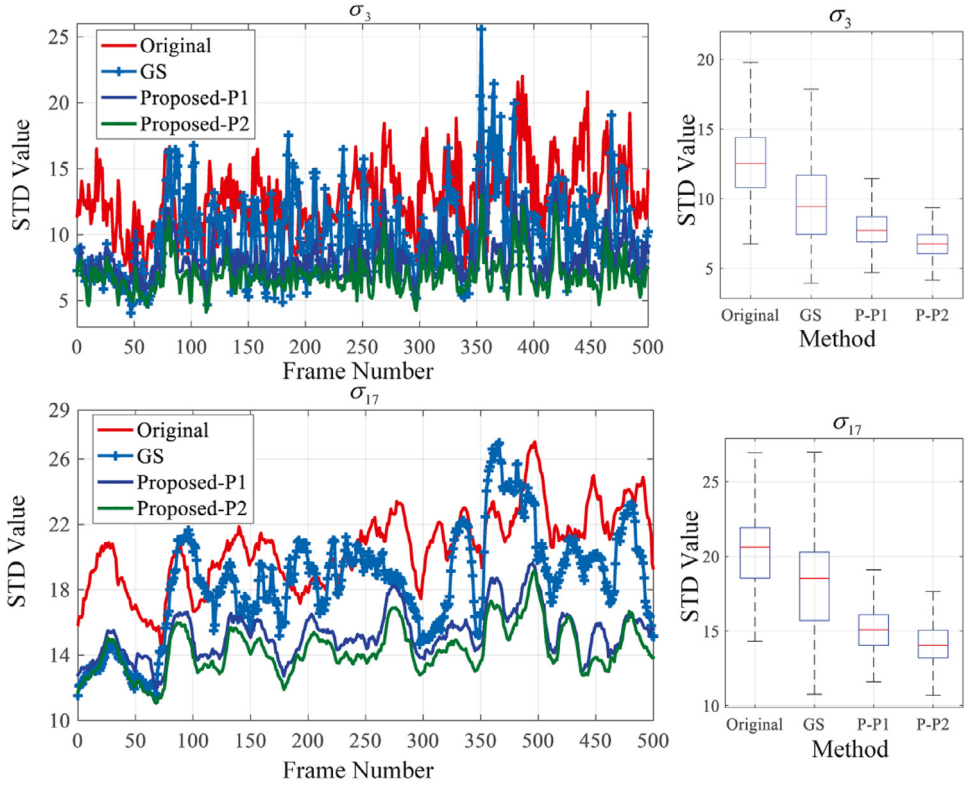

Fig. 11. The STD value of videos from different algorithms' output. The top row and bottom show curves of σ_3 and σ_{17} and corresponding statistical box-plots respectively.

corresponds to a longitudinal volumetric scanning (3D Scan) or time. Fig. 10 shows results of the proposed two-branch networks with fusion parameters k_p=0.85 and k_i=0.0001. Before the algorithm correction, the rotational artifacts existing in the synthetic video are visualized by a combination of overall intensity shift and local fluctuation along the longitudinal direction of en-face projections. After the algorithm correction, the overall shift is eliminated, so that horizontal straight lines patterns can be seen in the en-face images. Moreover, the local fluctuation is significantly reduced by 86% in polar domain (measured by the deviation of max intensity points between 2 adjacent frames).

4.2. Robustness assessment

We assess the robustness of the proposed method by qualitatively evaluating the drift reduction, geometric distortion reduction, as well as quantitative metrics. Synthetic videos provide direct ground truth for validation, while in real OCT videos only objects with significantly distinguishable geometries can provide reliable reference value/ground truth. When no guaranteed distortion ground truth value is available, we calculate the normalized Standard Deviation (STD) σ to estimate the correction performance

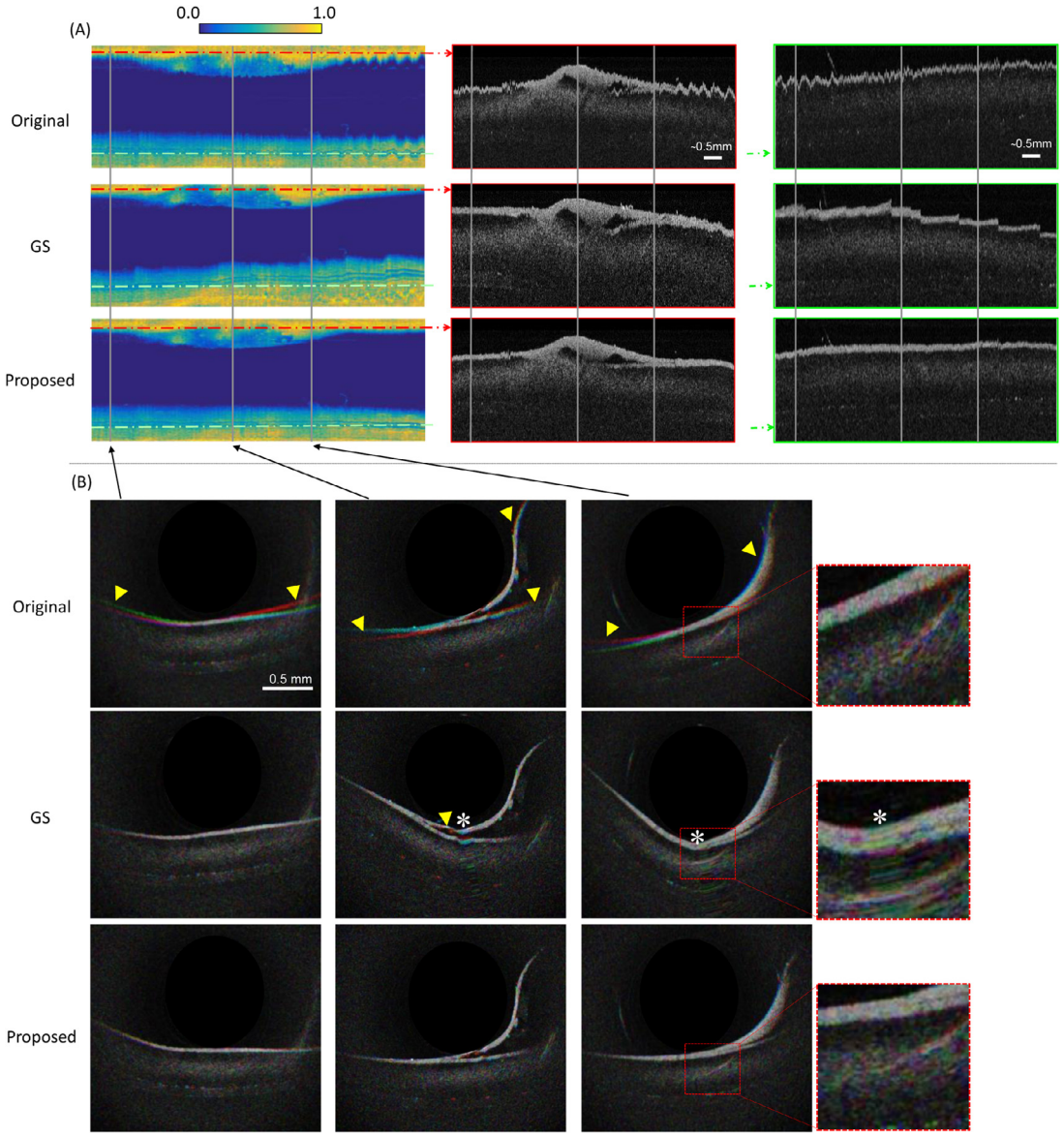

Fig. 12. Results from the anatomical colon model by robotic displacement of the catheter with the experimental setup shown in Fig. 4. (A) shows mean value en-face projections of a scan around a polyp, and two exemplary cross-sections re-sliced along the translation axis. Re-slices from two locations (indicated by red and light-green dashed dot lines) are presented. The intensity scale of en-face projection images is normalized by its maximum value. (B) Exemplary rotational cross-sections obtained from three positions marked by gray lines. In each position three consecutive frames are encoded in RGB. The presence of significant colorful pixels caused by artifacts is pointed out by yellow arrow heads. Asterisks mark out over-stretched images, that appear in the GS output.

Table 3
The mean value and variance of STD value of different algorithm's output in rectangular phantom video.

		Original	GS	Proposed Algorithm							
				Branch-B	$k_p = 0.55$ k_i=10^{-3}	(P1)	$k_p = 0.85$ k_i=10^{-4}	(P2)	$k_p = 0.95$ k_i=10^{-5}	(P3)	Branch-A
σ_3	mean	13.06	9.375	11.82	8.108		**7.152**		7.277		8.407
	variance	9.571	7.612	4.925	2.306		1.685		**1.682**		3.192
σ_{10}	mean	19.18	15.35	16.74	13.68		**12.61**		12.78		14.26
	variance	9.470	8.873	5.021	3.377		2.720		**2.673**		3.780
σ_{17}	mean	21.21	17.90	18.89	16.07		**15.02**		15.22		16.30
	variance	8.448	9.076	5.389	4.143		3.181		**2.978**		3.662

(van Soest et al., 2008). The definition of STD is:

$$\sigma_n = \frac{1}{N_{sig}} \sum_{i=1,j=1}^{N_{sig}} \bar{\sigma}\left(f_{i,j}\right) \tag{12}$$

where n is the number of frames in stack for calculation. $\bar{\sigma}\left(f_{i,j}\right)$ is the standard deviation calculated with pixels $f_{i,j}$ in one stacked frame stream, i and j are selected pixel indices in horizontal and vertical axis respectively. N_{sig} is the number of pixels used to calculate $\bar{\sigma}$. Since different noises and uncorrelated high intensity speckles occur in different frames, alignment algorithms are expected to decrease the STD value, but not to zero (van Soest et al., 2008).

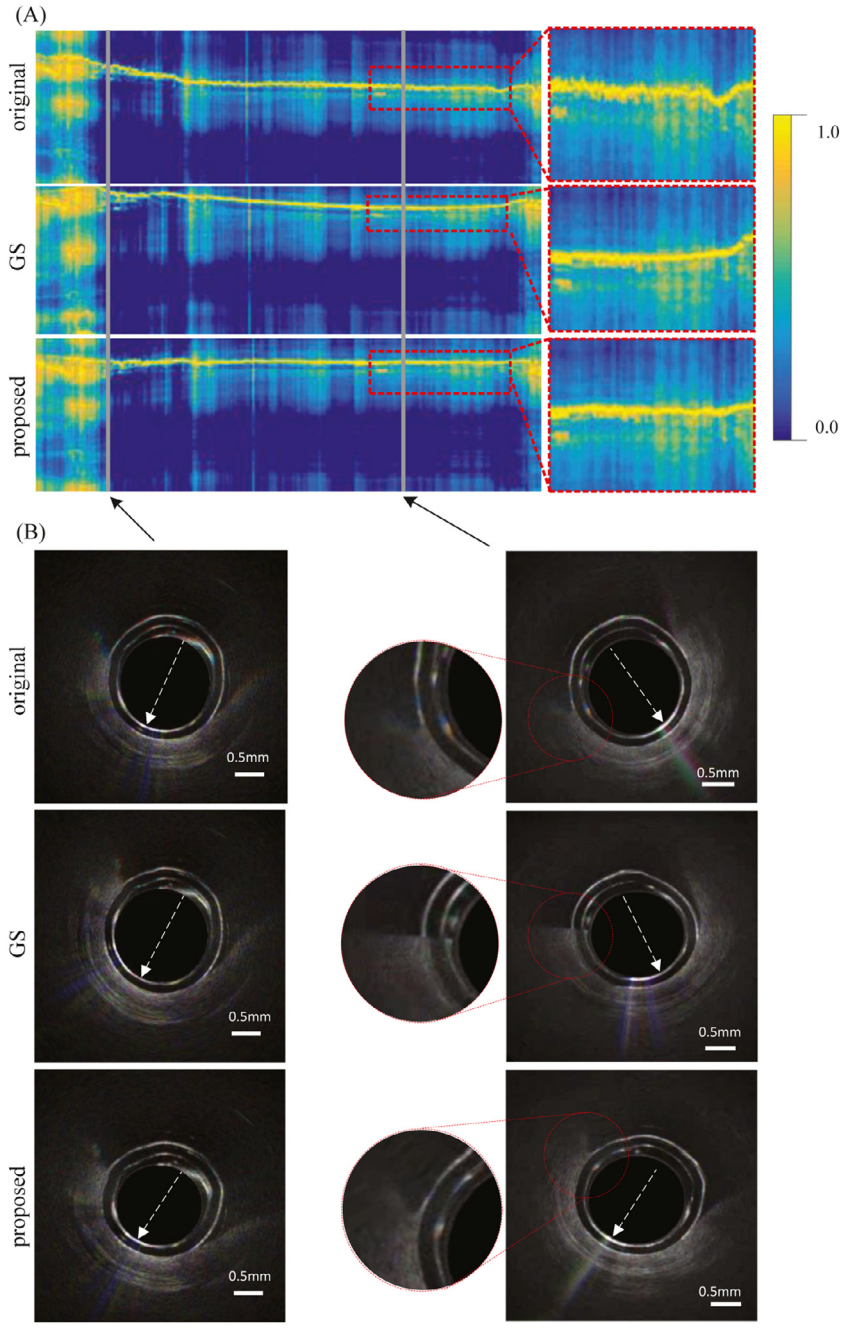

Fig. 13. Results obtained for unseen in vivo data. (A) En-face projection comparison (normalized intensity scale), and images from top to bottom row are original projections, results of the GS algorithm and results of the proposed algorithm. (B) 2D cross-sectional images corresponding to the gray lines in en-face projections; The red dashed circles enlarge the area where additional distortion is introduced by the GS method, while the proposed method correct the geometric orientation without affecting the image quality. The dashed white arrow lines point to the directions of the tissue.

4.2.1. Bench-top quantitative evaluation

For quantitative evaluation of the performance of the CNN based algorithm we use a stream of 2D frames obtained by pulling back the catheter in the rectangular phantom with a known geometry while maintaining a constant orientation of the catheter (Fig. 4 (A)).

Fig. 11 shows the results of STD values with different frame stack lengths. We analyze the instability over both a short term period with σ_3 and a longer period with σ_{17}. Here STD curves of the original video, the video corrected by the conventional GS algorithm, and videos corrected with two parameter combinations, referred to as "P1" and "P2", are shown. The "P1" parameters combination is given by k_p=0.55 and k_i=0.001 and it is introduced to assess the behavior of the algorithm when relying more on the group rotation estimation branch. Parameter combination "P2" is the same as the one used for accuracy assessment in Section 4.1. Both the parameter combinations obtain better correction results than the GS algorithm in both σ_3 and σ_{17}. Detailed statistic analysis of STD is presented in Table 3, which shows the mean value and variance of STD with different stack lengths σ_3, σ_{10} and σ_{17} of different algorithms outputs. Under these metrics, the proposed algorithm has better performances compared with graphic path searching algorithm regardless the choice of fusion parameters, except when disabling branch (A). Generally, compared with the fu-

sion parameters k_p=0.55, k_i=0.001 (combination "P1"), which already have a considerable video correction ability, larger k_p and smaller k_i make the fusion algorithm rely more on branch (A), which can improve the correction in the short term, reducing the short term STD mean value significantly (combinations 'P2' and 'P3' in Table 3). However, if the weight of branch (A) is tuned up to over-rely on the warping vector estimation branch (when k_p=1.0, k_i=0, last column of the table), not only the geometry of individual images will be distorted due to the drift error, but also the performance on STD reduction will be affected because of lacking compensation of branch (B).

4.2.2. Qualitative tissue phantom evaluation

We collected OCT data stream during translation of the OCT probe inside an anatomical colon model using the robotized interventional endoscope (4(D)). The probe scanned the colon lumen lengthwise near a polyp (Fig. 4(B)). The en-face projections and exemplary cross-sections re-sliced along the translation axis show the instability of the original scan. Although the GS based algorithm reduces high frequency instabilities, some instabilities are still visible (Fig. 12 (A)). In comparison, the proposed algorithm reduces the fluctuations and æsmoothsg the tissue surface, and also keeps the intensity distribution of the original en-face image. To qualitatively analyze the influence of stabilization method on A-line distribution per frame and in adjacent frames, three consecutive frames were assigned to one of three channels of the RGB image and overlapped (Fig. 12(B)). Compared with the initial image sequences with rotational artifacts (represented by the colorful pixels and the non-uniform orientations), the proposed method stabilized well the image sequences, while maintaining information about the tissue characteristic and the relative distance between the scanning center and tissue surface. A side-by-side comparison shows that the GS method works fine in the beginning of the scanning (colorful pixels are reduced), but the drift error grows when the OCT probe moves and introduces an extra distortion to the original image. When estimation error is large, the OCT image will be over-stretched and repeated A-lines can be targeted in the correction results (seen from the tissue surface marked by asterisks in the middle rows of Fig. 12(B)).

4.2.3. Generalization to unseen in vivo data

To evaluate the generalizability of the proposed method on unseen data, we collected OCT data using a steerable OCT catheter compatible with a robotized interventional colonoscope (Mora et al., 2020) in in vivo swine experiments (Fig. 4(C)). The animal test was approved by the Institutional Ethical Committee on Animal Experimentation (MESR: #2016072209464427).

In the in vivo animal test the catheter was placed at one position upon the colon tissue, and thus the tissue image should remain at a constant orientation. Overall rotation of the original animal test video is visible in en-face images (see the shift of max intensity position in the first row of Fig. 13 (A)), which has a max vertical shift of 219 pixels within the longitudinal scan (measured by the shift of the max intensity point through the whole en-face projection). Compared with conventional GS algorithm, which still has a orientation shift of 139 pixels, the proposed algorithm can better warp the "curve of max intensity" to a straighter line with only a small variation of 10 pixels, which reduces 91% of the rotational error. Each row of Fig. 13 (B) shows cross-sectional OCT images taken from this data stream at different positions, where rotational artifacts can be targeted. The proposed method corrects the angular errors without changing the quality or other information of the images.

To test the proposed method in clinical OCT images, following data reuse agreement we applied the correction algorithm to OCT images collected previously in two subjects with a tethered capsule endomicroscopy (TCE) in a human trial approved by Institutional Review Board (IRB: #2011P002619). In the TCE technology a rotational scanning OCT probe is enclosed in a distal capsule and a tether that connects it to an external OCT system (Gora et al., 2013). After the capsule is swallowed, typically up to four volumetric OCT images of the esophagus are collected when the capsule descends to the stomach and is pulled up in the esophagus. Bending and tension applied to the tether can add image artifacts. Fig. 14 shows results of correction with the proposed algorithm and GS algorithm of three scans acquired in the same subject. The first column shows en-face projections of 200th to 500th frames obtained during the first descending scan. The en-face image of the original data shows strong in between frame instability visible as a wavy pattern (red arrowheads in Fig. 14). After correction with the proposed algorithm an irregular lesion with lower intensity can be noted (red arrows in Fig. 14). A similar lesion shape can be also seen in en-face images of original 50th to 450th frames of the ascending scan (middle column in Fig. 14) where the capsule stability was very good. As can be observed the proposed algorithm also corrected small instabilities still present in the original data set of the first ascending scan (red starts in the middle column of Fig. 14). On the other hand, graph search algorithm introduced lesion deformation in both descending and ascending scans (red asterisks in the third row of Fig. 14). The right column in Fig. 14 shows 250th to 500th frames of the second ascending scan where a strong drift of the OCT data can be seen. The drift is visualised as a continuous diagonal shift in the en-face image that is almost completely removed by the proposed method. The GS algorithm corrects the scanning data but introduces distortion of the shape of objects in en-face image (red asterisk in the third row of Fig. 14). In Fig. 15 we present a volumetric reconstruction of three dimensional TCE data obtained in another subject. The 3D reconstruction is rendered with ImageJ software (Schindelin et al.,

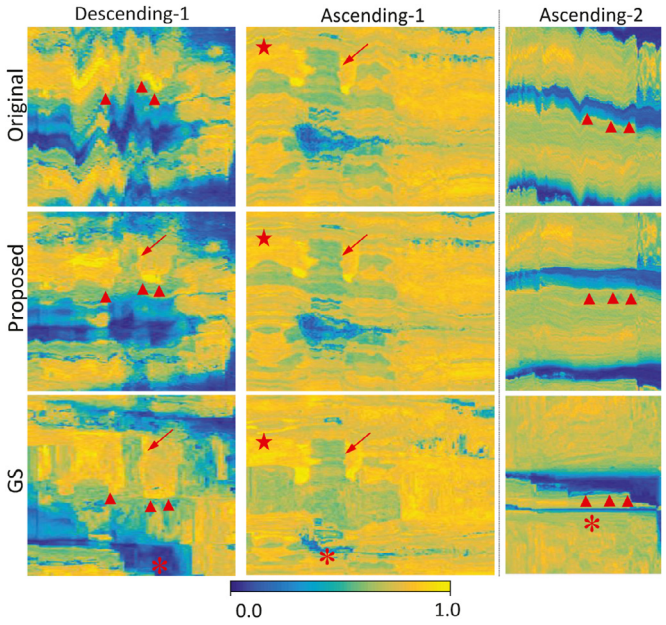

Fig. 14. En-face images (normalized intensity scale) of OCT data collected in the clinical trial with the tethered capsule OCT catheter. The first and second columns show the same region from 2 scans on the same patient, one with the capsule descending the esophagus and the other with the capsule being pulled up. The third column shows a section from the second ascending scan in the distal part of the esophagus where the original scan has strong drift artifacts. Red arrow heads point to large non-alignment caused by artifacts. Red star marks out small instabilities. Red arrows point to the same visible lesion. Asterisks mark out incorrectly deformed parts of the en-face images, that appear in the GS output.

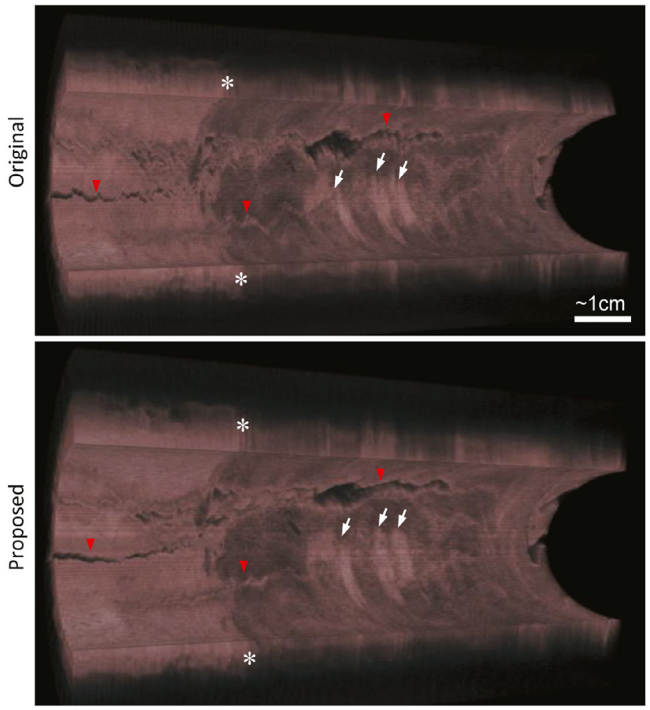

Fig. 15. 3D reconstructions of OCT data collected in the clinical trial with the tethered capsule OCT catheter in another subject. The red arrow heads point to wavy patterns caused by the artifacts. The white asterisks mark out the conjunction area between healthy esophagus and Barrett's esophagus. The white arrows point to higher intensity patches in the Barrett's esophagus segment.

2015). The comparison of the reconstructed data before and after correction shows that the proposed algorithm removes instabilities present in the original data set that are especially noticeable in the areas of loss of contact visible as the darkest areas (red arrow heads in Fig. 15). After correction, typical irregularities of the junction between the tissues with features of the normal esophagus on the left and of Barrett's esophagus on the right can be well appreciated (white asterisks in Fig. 15). In addition, patchy areas of higher intensity in the Barrett's segment (white arrows in Fig. 15) have more regular contours, which helps with their visual assessment.

5. Discussion and conclusion

We developed a new solution to tackle the distortion and instability problem using deep CNN, which can be generalized for scanning situations in different targets and with different catheters. We proposed a new A-line level shifting error vector estimation network to extract optimal path from a correlation matrix, which has higher accuracy and robustness compared with the conventional approach in situations where the images have few features. Moreover, we solved the problem of error accumulation in iterative video processing, with a group rotation estimation net. This CNN based algorithm was trained on semi-synthetic data and applied to real videos acquired in various scanning conditions. A full validation on in vivo data is nearly impossible, due to the fact that annotating rotational distortions on such data is very complex. The results presented, however, suggest that the proposed algorithm generalizes well over relevant in vivo pre-clinical data and clinical data from another modality of rotational scanning OCT, which was never seen during the training.

The proposed image based solution relies on the assumption that the appearance change caused by rotational artifacts is faster than the appearance change of tissue itself. This assumption is valid in most standard cases, as shown in the results section. Nevertheless, the algorithmic reduction of distortion may be affected in some pathological cases, where the screened tissue appearance changes very quickly, especially at the conjunction between two different types of tissue. Note that the proposed method needs a reference frame for correcting drift and accumulative error. In the beginning of a scanning the drift is small and the stretch and shrink distortion happens less occasionally than the shaking, which means a visually correct reference frame can be chosen from a small period at the beginning of a scan. The current implementation presented in this paper is not adapted for a conventional pullback scanning that moves the rotating lens along the protective sheath. Indeed, for this type of pullback the initial frame cannot be used for drift suppression because of possible changes of appearance of the sheath along the pullback. To apply the proposed method to a pullback scanning a sheath registration and calibration will be needed, which means the reference should be a pre-recorded sheath image stack instead of a single B-scan.

Although branch B could also affect the accuracy of A-line level correction, the fusion of the two branches can still compensate a sudden stretch-shrink distortion that would emerge in a B-scan. It is worth mentioning that in the algorithm testing we disabled branch-A (warping vector estimation) or branch-B (group rotation estimation), and the results show that the performance is degraded with only one of the two branches. Correction accuracy may be improved by other probabilistic fusion filters, or by optimizing the parameters of the PI complementary filter based on objective functions.

Another motivation of this work is to follow our previous work on integration of OCT with robotic endoscope (Mora et al., 2020), and online image processing is crucial in this scenario because robot positioning and displacement could be guided by the OCT images. It is however possible only if images are geometrically correct. The on-line correction algorithm can also enable the use of en-face projection images in gastrointestinal applications, which could help, for example, in assessment of the length of Barretts esophagus or localization of suspicious lesions (Liang et al., 2016). The proposed algorithm is designed for on-line video processing with historical data as input only. The current implementation of the algorithm has an update rate around 7 FPS. It is not fast enough for correcting every frame of a real-time OCT imaging system which could have a framerate of 60 FPS due to hardware limitations and large input size. An immediate solution to reduce computational consumption could be down-sampling the input image or shortening the shifting window w, but it will negatively affect the quality of correlation matrix and angular registration range. Alternatively, code and algorithmic optimizations, especially in the correlation stack, could also accelerate the computation. We plan to work on algorithm optimization and testing on a more recent hardware setup as part of our future works, which may help speed-up the image correction and meet the requirements of on-line diagnosis (i.e. with an update rate of 10–20 FPS).

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Acknowledgements

This work was supported by the ATIP-Avenir grant, the ARC Foundation for Cancer research, the University of Strasbourg IdEx, Plan Investissement dAvenir and by the ANR (ANR-10-IAHU-02 and ANR-11-LABX-0004-01) and the ATLAS project that received funding from the European Union's Horizon 2020 research and innova-

tion program under the Marie Sklodowska-Curie grant agreement No 813782. The authors would like to acknowledge Prof. Guillermo Tearney and Catriona Grant from Massachusetts General Hospital for sharing the data obtained in tethered capsule endomicroscopy clinical trial (IRB: #2011P002619).

References

Abouei, E., Lee, A.M., Pahlevaninezhad, H., Hohert, G., Cua, M., Lane, P., Lam, S., MacAulay, C., 2018. Correction of motion artifacts in endoscopic optical coherence tomography and autofluorescence images based on azimuthal en face image registration. J Biomed Opt 23 (1), 016004.

Ahsen, O.O., Lee, H.-C., Giacomelli, M.G., Wang, Z., Liang, K., Tsai, T.-H., Potsaid, B., Mashimo, H., Fujimoto, J.G., 2014. Correction of rotational distortion for catheter-based en face oct and oct angiography. Opt Lett 39 (20), 5973–5976.

Allgeuer, P., Behnke, S., 2014. Robust sensor fusion for robot attitude estimation. In: 2014 IEEE-RAS International Conference on Humanoid Robots. IEEE, pp. 218–224.

Bengio, Y., 2009. Learning deep architectures for AI. Now Publishers Inc.

Costello, F., 2017. Optical coherence tomography in neuro-ophthalmology. Neurol Clin 35 (1), 153–163.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2758–2766.

Gast, J., Roth, S., 2019. Deep video deblurring: The devil is in the details. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, p. 0.

Gatta, C., Pujol, O., Leor, O.R., Ferre, J.M., Radeva, P., 2009. Fast rigid registration of vascular structures in ivus sequences. IEEE Trans. Inf. Technol. Biomed. 13 (6), 1006–1011.

Gora, M.J., Sauk, J.S., Carruth, R.W., Gallagher, K.A., Suter, M.J., Nishioka, N.S., Kava, L.E., Rosenberg, M., Bouma, B.E., Tearney, G.J., 2013. Tethered capsule endomicroscopy enables less invasive imaging of gastrointestinal tract microstructure. Nat. Med. 19 (2), 238–240.

Gora, M.J., Suter, M.J., Tearney, G.J., Li, X., 2017. Endoscopic optical coherence tomography: technologies and clinical applications. Biomed Opt Express 8 (5), 2405–2444.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

Herz, P., Chen, Y., Aguirre, A., Schneider, K., Hsiung, P., Fujimoto, J., Madden, K., Schmitt, J., Goodnow, J., Petersen, C., 2004. Micromotor endoscope catheter for in vivo, ultrahigh-resolution optical coherence tomography. Opt Lett 29 (19), 2261–2263.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al., 1991. Optical coherence tomography. Science 254 (5035), 1178–1181.

Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., Liu, W., 2017. Real–time neural style transfer for videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 783–791.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462–2470.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR, pp. 448–456.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678.

Kawase, Y., Suzuki, Y., Ikeno, F., Yoneyama, R., Hoshino, K., Ly, H.Q., Lau, G.T., Hayase, M., Yeung, A.C., Hajjar, R.J., et al., 2007. Comparison of nonuniform rotational distortion between mechanical ivus and oct using a phantom model. Ultrasound in medicine & biology 33 (1), 67–73.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lee, S.-W., Heidary, A.E., Yoon, D., Mukai, D., Ramalingam, T., Mahon, S., Yin, J., Jing, J., Liu, G., Chen, Z., et al., 2011. Quantification of airway thickness changes in smoke-inhalation injury using in-vivo 3-d endoscopic frequency-domain optical coherence tomography. Biomed Opt Express 2 (2), 243–254.

Li, D., Wu, J., He, Y., Yao, X., Yuan, W., Chen, D., Park, H.-C., Yu, S., Prince, J.L., Li, X., 2019. Parallel deep neural networks for endoscopic oct image segmentation. Biomed Opt Express 10 (3), 1126–1135.

Liang, K., Ahsen, O.O., Lee, H.-C., Wang, Z., Potsaid, B.M., Figueiredo, M., Jayaraman, V., Cable, A.E., Huang, Q., Mashimo, H., et al., 2016. Volumetric mapping of barretts esophagus and dysplasia with en face optical coherence tomography tethered capsule. Am. J. Gastroenterol. 111 (11), 1664.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml, Vol. 30. Citeseer, p. 3.

Mahony, R., Hamel, T., Pflimlin, J.-M., 2005. Complementary filter design on the special orthogonal group so (3). In: Proceedings of the 44th IEEE Conference on Decision and Control. IEEE, pp. 1477–1484.

Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., Van Gool, L., 2017. Convolutional oriented boundaries: from image segmentation to high-level tasks. IEEE Trans Pattern Anal Mach Intell 40 (4), 819–833.

Mavadia-Shukla, J., Zhang, J., Li, K., Li, X., 2020. Stick-slip nonuniform rotation distortion correction in distal scanning optical coherence tomography catheters. J Innov Opt Health Sci 13 (06), 2050030.

Mora, O.C., Zanne, P., Zorn, L., Nageotte, F., Zulina, N., Gravelyn, S., Montgomery, P., De Mathelin, M., Dallemagne, B., Gora, M.J., 2020. Steerable oct catheter for real-time assistance during teleoperated endoscopic treatment of colorectal cancer. Biomed Opt Express 11 (3), 1231–1243.

Nam, H.S., Kim, C.-S., Lee, J.J., Song, J.W., Kim, J.W., Yoo, H., 2016. Automated detection of vessel lumen and stent struts in intravascular optical coherence tomography to evaluate stent apposition and neointimal coverage. Med Phys 43 (4), 1662–1675.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.

van der Putten, J., van der Sommen, F., Struyvenberg, M., de Groof, J., Curvers, W., Schoon, E., Bergman, J.J., et al., 2019. Tissue segmentation in volumetric laser endomicroscopy data using fusionnet and a domain-specific loss function. In: Medical Imaging 2019: Image Processing, Vol. 10949. International Society for Optics and Photonics, p. 109492J.

van der Putten, J., Struyvenberg, M., de Groof, J., Scheeve, T., Curvers, W., Schoon, E., Bergman, J.J., de With, P.H., van der Sommen, F., 2020. Deep principal dimension encoding for the classification of early neoplasia in barrett's esophagus with volumetric laser endomicroscopy. Computerized Medical Imaging and Graphics 80, 101701.

Ricco, S., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J., 2009. Correcting motion artifacts in retinal spectral domain optical coherence tomography via image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 100–107.

Sathyanarayana, S., 2006. Nonuniform rotational distortion (nurd) reduction. US Patent 7,024,025.

Schindelin, J., Rueden, C.T., Hiner, M.C., Eliceiri, K.W., 2015. The imagej ecosystem: an open platform for biomedical image analysis. Mol. Reprod. Dev. 82 (7–8), 518–529.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

van Soest, G., Bosch, J.G., van der Steen, A.F., 2008. Azimuthal registration of image sequences affected by nonuniform rotation distortion. IEEE Trans. Inf. Technol. Biomed. 12 (3), 348–355.

Suter, M.J., Gora, M.J., Lauwers, G.Y., Arnason, T., Sauk, J., Gallagher, K.A., Kava, L., Tan, K.M., Soomro, A.R., Gallagher, T.P., et al., 2014. Esophageal-guided biopsy with volumetric laser endomicroscopy and laser cautery marking: a pilot clinical study. Gastrointest. Endosc. 79 (6), 886–896.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Tran, P.H., Mukai, D.S., Brenner, M., Chen, Z., 2004. In vivo endoscopic optical coherence tomography by use of a rotational microelectromechanical system probe. Opt Lett 29 (11), 1236–1238.

Ughi, G.J., Larsson, M., Dubois, C., Sinnaeve, P.R., Desmet, W., D'Hooge, J., Adriaenssens, T., Coosemans, M., 2012. Automatic three-dimensional registration of intravascular optical coherence tomography images. J Biomed Opt 17 (2), 026005.

Uribe-Patarroyo, N., Bouma, B.E., 2015. Rotational distortion correction in endoscopic optical coherence tomography based on speckle decorrelation. Opt Lett 40 (23), 5518–5521.

Wang, M., Yang, G.-Y., Lin, J.-K., Zhang, S.-H., Shamir, A., Lu, S.-P., Hu, S.-M., 2018. Deep online video stabilization with multi-grid warping transformation learning. IEEE Trans. Image Process. 28 (5), 2283–2292.

Wang, T., Pfeiffer, T., Regar, E., Wieser, W., van Beusekom, H., Lancee, C.T., Springeling, G., Krabbendam, I., van der Steen, A.F., Huber, R., et al., 2015. Heartbeat oct: in vivo intravascular megahertz-optical coherence tomography. Biomed Opt Express 6 (12), 5021–5032.

Wang, T., Wieser, W., Springeling, G., Beurskens, R., Lancee, C.T., Pfeiffer, T., van der Steen, A.F., Huber, R., van Soest, G., 2013. Intravascular optical coherence tomography imaging at 3200 frames per second. Opt Lett 38 (10), 1715–1717.

Yonetsu, T., Bouma, B.E., Kato, K., Fujimoto, J.G., Jang, I.-K., 2013. Optical coherence tomography–15 years in cardiology–. Circulation Journal CJ–13.

Yong, Y.L., Tan, L.K., McLaughlin, R.A., Chee, K.H., Liew, Y.M., 2017. Linear-regression convolutional neural network for fully automated coronary lumen segmentation in intravascular optical coherence tomography. J Biomed Opt 22 (12), 126005.

Zeng, Y., Xu, S., Chapman, W.C., Li, S., Alipour, Z., Abdelal, H., Chatterjee, D., Mutch, M., Zhu, Q., 2020. Real-time colorectal cancer diagnosis using pr-oct with deep learning. In: Optical Coherence Tomography. Optical Society of America, pp. OW2E–5.

Zulina, N., Caravaca, O., Liao, G., Gravelyn, S., Schmitt, M., Badu, K., Heroin, L., Gora, M.J., 2021. Colon phantoms with cancer lesions for endoscopic characterization with optical coherence tomography. Biomed Opt Express 12 (2), 955–968.