Data Stream Stabilization for Optical Coherence Tomography Volumetric Scanning

Guiqiu Liao, Oscar Caravaca-Mora, Benoit Rosa, *Member, IEEE*, Philippe Zanne, Alexandre Asch, Diego Dall'Alba, Paolo Fiorini, *Life Fellow, IEEE*, Michel de Mathelin, *Senior Member, IEEE*, Florent Nageotte, and Michalina J. Gora

Abstract—Optical Coherence Tomography (OCT) is an emerging medical imaging modality for luminal organ diagnosis. The non-constant rotation speed of optical components in the OCT catheter tip causes rotational distortion in OCT volumetric scanning. By improving the scanning process, this instability can be partially reduced. To further correct the rotational distortion in the OCT image, a volumetric data stabilization algorithm is proposed. The algorithm first estimates the Non-Uniform Rotational Distortion (NURD) for each B-scan by using a Convolutional Neural Network (CNN). A correlation map between two successive B-scans is computed and provided as input to the CNN. To solve the problem of accumulative error in iterative frame stream processing, we deploy an overall rotation estimation between reference orientation and actual OCT image orientation. We train the network with synthetic OCT videos by intentionally adding rotational distortion into real OCT images. As part of this article we discuss the proposed method in two different scanning modes: the first is a conventional pullback mode where the optical components move along the protection sheath, and the second is a self-designed scanning mode where the catheter is globally translated by using an external actuator. The efficiency and robustness of the proposed method are evaluated with synthetic scans as well as real scans under two scanning modes.

Index Terms—Optical coherence tomography, video stabilization, non-uniform rotational distortion, convolutional neural network.

Guiqiu Liao is with ICube, UMR 7357 CNRS-University of Strabourg, 67091 Strasbourg, France, and also with the Department of Computer Science, University of Verona, 37129 Verona, Italy (e-mail: liao.guiqiu@etu.unistra.fr).

Oscar Caravaca-Mora, Benoit Rosa, Philippe Zanne, Alexandre Asch, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora are with ICube, UMR 7357 CNRS-University of Strabourg, 67091 Strasbourg, France (e-mail: caravacamora@kuleuven.be; b.rosa@unistra.fr; zanne.philippe@unistra.fr; alexasch3h@gmail.com; demathelin@unistra.fr; nageotte@unistra.fr; gora@unistra.fr).

Diego Dall'Alba and Paolo Fiorini are with the Department of Computer Science, University of Verona, 37129 Verona, Italy (e-mail: diego.dallalba@univr.it; paolo.fiorini@univr.it).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMRB.2021.3093434, provided by the authors.

Digital Object Identifier 10.1109/TMRB.2021.3093434

I. INTRODUCTION

OPTICAL Coherence Tomography (OCT) is an emerging medical imaging modality for diagnosing diseases in cardiovascular system, airways and gastrointestinal tract [1]. The main challenges related to these applications are delivery and collection of light from the volumetric scanning of the area of interest. Fourier domain OCT in a single measurement provides a one-dimensional in-depth information and in order to collect three-dimensional information the optical beam has to be scanned over the tissue. Various designs of specialized micro-optical catheters have been developed to achieve this goal [2].

A circumferential two-dimensional scan can be performed by rotation of an optical beam reflected to the side of the probe using a micro-motor on the distal tip, or by a proximal rotational actuation, which is remotely connected to the distal optical components with a torque coil (see Fig. 1). Volumetric scanning, in both cases, is then typically effectuated by pulling back the rotating optical core to create a helical scan. This approach for volumetric scanning was originally developed for cardiovascular applications where an OCT catheter is inserted using a guidewire into the location of interest and a pullback needs to be synchronized with a blood occlusion [3]. It was then adopted in gastrointestinal imaging both in low-profile and balloon catheters, which are inserted in the digestive system using a working channel of an endoscope [4]. Most recently, an internal pullback scanning system was also developed for a tethered capsule device [5]. A high precision short segment pullback enabled high quality en-face imaging that could not be achieved with a standard tethered capsule devices typically pulled back manually.

Even though motorized rotational and pullback actuation provides good control of the OCT beam position, scanning distortions are still very common [4], [5]. Non-Uniform Rotational Distortion (NURD) is crucial for catheter-based imaging systems such as endoscopic OCT but also intravascular ultrasound [6], [7].

In proximal rotation devices NURD can be observed due to mechanical friction between rotating optical components and the protecting sheath, mostly caused by bending of the catheter inserted in the tortuous path of the digestive, cardiovascular or respiratory system. Miniaturization of motors enabled distal scanning and more compact [1] and more easily miniaturized
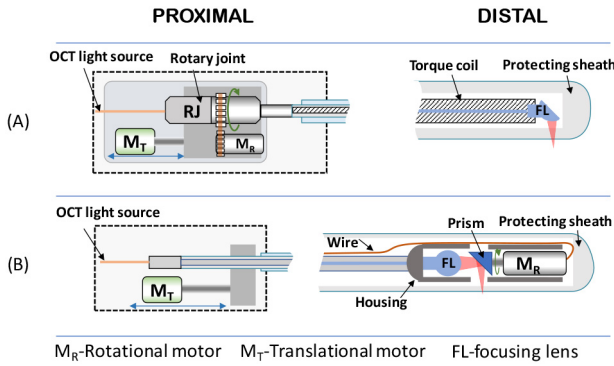
Fig. 1. Schematics of (A) proximal-scanning with a fiber-optic rotary joint; (B) distal-scanning endoscope with a micro-motor.

designs [11], in which instabilities are only related to the quality of the motor itself. However, these designs require electrical connection of the distally located motor, which adds additional safety risks to the device and causes partial shadowing of the tissue by the wires. Non-linear transmission of displacement, and imperfect synchronization between data acquisition and scanning speed also causes distortions of the longitudinal scanning. This is typically observed as drift in the 3D reconstruction.

We have previously investigated the NURD performance of an OCT catheter that was developed by our team for real-time assistance during minimally invasive robotized treatment of colorectal cancer [47]. This robotized OCT catheter can be inserted in one of the two instrument channels of the robotized interventional endoscope and teleoperated using a master controller. In this article, we focus on studying the volumetric imaging instability of proximal scanning endoscopic OCT catheters, and on stabilizing the OCT image by proposing both a new robotic 3D scanning and a software approach. A new scanning method was motivated by redundancy of two degrees of freedom (DoF) in the OCT enhanced robotized interventional endoscope. The OCT catheter once inserted in the robot has two DoF for volumetric imaging (rotation and pullback of the inner optical core, provided by the OCT scanning system) and 3 DoF for maneuverability (rotation, translation and bending, provided by the robotics system [48]). As can be seen both rotational and translation are redundant, but OCT rotational scanning requires at least 3000 rpm, which cannot be achieved by a motorized instrument driver. The required translation speed of 2.5 mm/s, on the other hand, can be achieved by both inner core pullback or outer tool pullback. The outer pullback can reduce the instability of proximal scanning with a moving torque coil, but not perfectly, because the residual friction with the protecting sheath cannot be completely eliminated, and thus we propose a software correction to further decreased distortions.

When using the information within OCT images, two challenges need to be addressed for the software correction of the NURD. The first challenge is to compensate the NURD on-line, which means that the algorithm should only take historical OCT data to stabilize the scanning video. This can allow the catheter to correctly access the cross-sectional structure of tissues in real-time, which can dramatically reduce

the time it takes for medical doctors/physicians to make a diagnostic decision. Unlike the off-line stabilization that uses the entire data after a scan to optimize the warping parameters, on-line NURD correction typically follows an iterative processing pipeline where the estimation error accumulates. In this case, common iterative algorithms fail when a certain frame has a large stabilization error or when the processing time is long. Using optical markers is an alternative choice to stabilize rotational distortion, but it will block the OCT laser beam and degrade the image quality. Thus the second challenge is that the algorithm should not use any type of additional objects to track the orientation shifts.

We formulate the problem as finding a warping vector for each frame, to shift each individual scanning A-line to its correct position. This can be done based on computing the cross-correlation or Euler distance between A-lines in the new uncorrected frame and in previous stabilized frames [8]–[11]. However, in a real application it is hard to ensure that every A-line of OCT images can provide visible features for correlation because of the limited Field of View (FoV) of OCT catheters (part of the OCT image is filled with background noise when the tubular lumen is larger than the catheter FoV). Given the prior knowledge [8], [9] that the warping path should be continuous through the whole OCT frame stream, the problem of searching the path in a computed correlation matrix is highly similar to a continuous boundary line prediction problem. Motivated by this, we seek to leverage the state of art machine learning techniques for the boundary line prediction method [15], [22] as a solution to optimize the warping vector estimation.

Although the warping vector estimation accuracy can be improved, the accumulative drift still exists in long data stream processing. We address this issue by estimating an overall rotation value for each frame, and compensating the drift error of the NURD warping vector using this overall rotation value. To do this, the algorithm exploits the consistency of the OCT protecting sheath as a cue for overall rotation estimation, and this overall rotation is a direct and robust estimation (without time-integration process), although it only provides a low accuracy because it relies on limited information. We deploy a fusion algorithm [29] to integrate the overall orientation and element-wise NURD estimation, which is inspired the fusion of gyroscope and accelerometry for attitude angle estimation [30].

We demonstrate our method in two different 3D scanning modes: internal pullback scanning and outer tool scanning (see Fig. 2 (A) and Fig. 2 (B)). For the internal pullback scanning, we deploy an additional volumetric sheath registration. We train the CNN based NURD estimation algorithm with semi-synthetic data by intentionally distorting OCT images. We evaluate the accuracy of the NURD estimation with semi-synthetic data, and test its robustness on real data. An ablation study is demonstrated by disabling parts of the full algorithm.

II. RELATED WORK

In this section we provide an overview of previous research on NURD correction for catheter-based imaging systems,

followed by an introduction to the state-of-the-art video stabilization research and Convolutional Neural Network (CNN) research for white light camera, which inspired the proposed method for endoscopic OCT stabilization.

A. NURD Correction

Several NURD correction methods have been proposed for catheter-based imaging systems [6], [8]–[12], [14]. A frequency analysis of the texture of the Intra-Vascular Ultrasound (IVUS) image was used in [6] to estimate the rotational velocity. Matching based algorithms by minimizing the Euler distance of warped neighboring frames [8], [9], or by maximizing the cross-correlation [10], [11], [14] have been proposed to track the shift error of A-lines. Specifically, in intravascular OCT imaging, intra-vascular stents may serve as landmarks that help to capture rotational distortion in pullback volumetric scanning [13]. However, matching based methods using cross-correlation or Euler distance typically require highly correlated images. Even with tractable features between successive frame pairs, it is difficult to guarantee that the estimation of NURD is correct, which leads to iterative drift in long data streams processing. Adding structural markers on the OCT sheath provides an extrinsic reference for orientation change [9], but the markers block the OCT light and thus remove information about the tissue.

B. Data Fusion for Rotation Estimation

Similar to iterative NURD estimation, the estimation of attitude angle with integral gyroscope data also has the problem of drift [27], [28]. The integral drift of a gyroscope is usually compensated by another robust angle estimation from sensors such as accelerometers and magnetometers [30]–[33]. While a gyroscope provides excellent information about rapid orientation changes, it only provides relative orientation changes that gradually drift with their lifetime and temperature. An accelerometer or magnetometer, on the other hand, has a direct measurement of orientation but with lower accuracy. Various classes of filters were demonstrated to fuse accurate rapid relative (indirect) measurements with less accurate direct measurements such as with the Extended Kalman Filter (EKF) [33], the complementary filter [29], [31], [32] and a gradient based filter [30]. In a similar way to the role of an accelerometer or a magnetometer, another additional overall rotation can be estimated and fused with the NURD estimation, which compensates the accumulative error.

C. Video Stabilization

In the research field of image processing, CNN and deep learning related approaches stand as state-of-the-art. For OCT image processing, CNN has been applied to tissue layer segmentation [17]–[19], classification [20] and cancer tissue identification [21], but there is no CNN application for OCT stabilization. On the other hand, the literature in video stabilization is richer for white light cameras than for medical imaging systems (including the OCT). We seek to fill in the gap between the common computer vision research field and

that of the OCT imaging, by relying on the CNN to enhance the efficiency of OCT frame stream stabilization.

For white light camera video stabilization, there are two types of approaches to model the problem. One seeks to directly estimate the camera path (position), and the video stabilization can be considered as a camera path smoothing problem [36]. This formulation aims to stabilize homographic distortion cause by camera shaking, and recently a deep learning based method has been developed to learn from data registered by a mechanical stabilizer [37], which shows greater efficiency than traditional algorithms. The other type of approach models the instability of the video (or frame stream) as an appearance change [38]. This modeling methodology can be adapted to different imaging systems beyond the white light camera. To formulate the appearance change, features matching algorithms or optical flow [39], [40] can be used. A recent study uses deep learning techniques to estimate an optical flow field representing shift map of pixels in the video frames, and then applies another CNN regression module to estimate a pixel-wise warping field from the optical flow field [35]. Inspired by this, we deploy a CNN to estimate the element-wise NURD warping vector from a correlation map which roughly interprets the NURD distortion of a single frame.

D. CNN for Path Searching

An essential step of the NURD estimation is to search a continuous optimal path with large correlation value from the correlation map. Solutions applied in previous OCT stabilization studies [10], [11], [14] are mainly based on graph searching (GS), and rely on local features of gradients, maxima, textures and other prior information. This type of technique is also a traditional way of contour tracing [23]. Recently deep learning based contour tracing techniques [22], [24]–[26] has been demonstrated to be faster and more robust than traditional methods.

The state-of-the-art deep learning models for pixel-wise segmentation are based on adaptations of convolutional networks that had originally been designed for image classification. To solve dense prediction problems such as semantic segmentation, which are structurally different from image classification, striding and dilated CNN [41] are proposed to systematically aggregate multiscale contextual information without losing resolution. Path detection can be achieved with a pixel-wise segmentation architecture (e.g., predict a binary map where the path position and background pixels have different values). This is a high-cost approach, which usually adopts a U-shape CNN [22], [25], [26] using up-convolution layers [42], [43]. We deploy a CNN to predict a single vector representing path coordinates from the correlation map. Compared with pixel-wise prediction, the proposed architecture can be more efficient and no post processing is required. Moreover, it is easy to integrate the continuity prior knowledge into the network training, by an additional term in the loss function. Since we consider the situation when part of the image has no feature, a simple cascade layer architecture [45] is not sufficient for spacial exploitation. Therefore we use a parallel scheme [44] which controls striding of convolution to extract
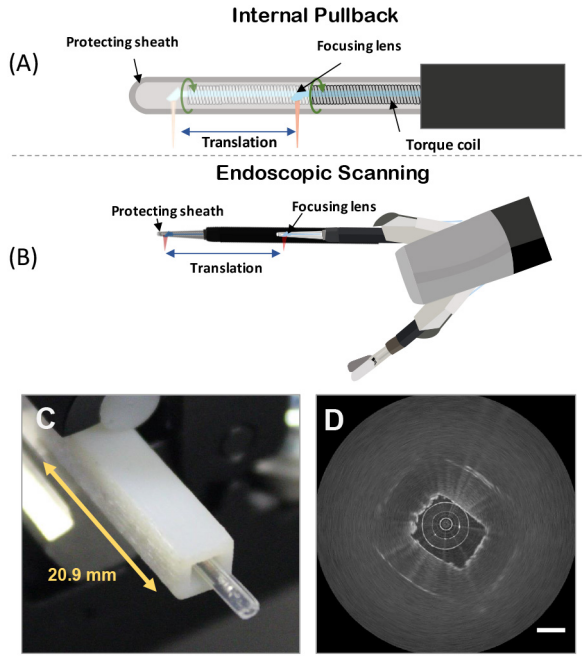
Fig. 2. (A) Schematic of internal pull back for proximal-scanning OCT; (B) Schematic of robotic endoscopic scanning; (C) The rectangular test object; (D) An example of OCT B-scan, the scale bar corresponds to 1 mm.
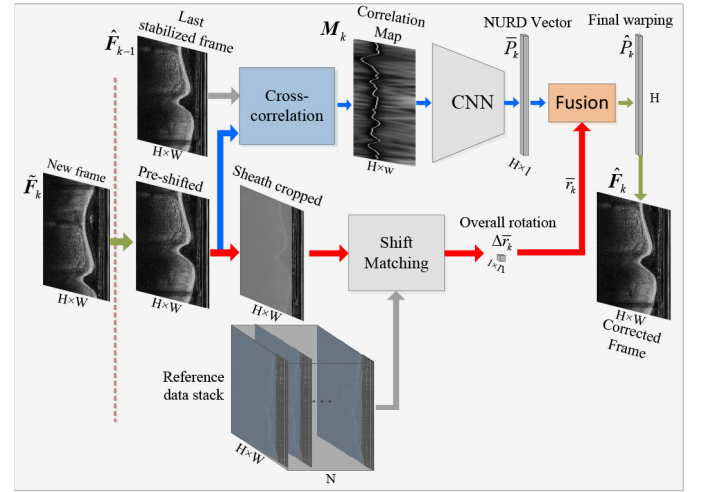

Fig. 3. Algorithm pipeline of the proposed OCT stabilization method. The original new frame is first pre-shifted based on the previous estimated overall rotation. Then the pre-shifted OCT image follows two paths in parallel to estimate two levels of artifacts. In the first path (with blue color) a correlation map between the pre-shifted frame and last stabilized frame is computed, and a CNN estimates an A-line level NURD vector. In the second path (with red color), the sheath image of the pre-shifted frame is used to match the reference sheath stack to estimate an overall rotation value. The final warping vector is a fusion of the A-line level NURD vector (from the first path) and the single rotation value (from the second path), which is used to shift each individual A-line of the new frame to obtain a new corrected frame.

hierarchical abstract information from the correlation matrix. Moreover, the proposed network has shallower layers than VGG [45] or Resnet [46] and thus requires less inference time. Our tests and results indicate that this choice is sufficient for the NURD stabilization task.

III. METHODS

A. OCT Volumetric Imaging

To compare the conventional pullback and robotic 3D scanning performance we use an OCT instrument with an outer diameter of 3.5 mm [47]. The instrument is terminated at the distal end with a transparent sheath on the tip, which allows three-dimensional OCT imaging using an internal rotating side-focusing optical probe with two proximal external scanning actuators. The instrument is connected to an OCT imaging system built around the OCT Axsun engine, with a 1310 nm center wavelength swept source laser and 100 kHz A-line rate. The OCT catheter is compatible with an instrument channel of a robotized flexible interventional endoscope [48]. The distal end of the robotized endoscope can be bent in 2 orthogonal planes, translated and rotated. The OCT catheter can be translated, rotated and bent in one plane. In addition, the rotation and translation of the inner OCT probe can also be controlled by a servo system.

We test two different types of volumetric scans with this system. One is an internal pullback scanning where the torque coil of the catheter moves through the protecting sheath, and another one is a robotic outer pullback of the whole tool where the torque coil delivers only rotational scanning. Fig. 2 shows a schematic of the internal pullback scan and the outer tool scan.

B. Data Stream Stabilization

A typical OCT scanning system makes the assumption that the focusing lens at the distal tip rotates with a constant speed, so the OCT data acquisition system arranges A-lines of each frame with equal spacing to cover a 360 degrees region in the polar domain. However, for various reasons such as friction and non-constant motor speed the angular speed of the distal optical components may not be uniform over time, resulting in a distortion of the reconstructed B-scan in polar domain. The OCT rotational distortion correction algorithm estimates a warping vector representing the angular position error of each independent A-line. This vector is used to place A-lines at new positions in order to limit the NURD artifact.

The architecture of the proposed stabilization algorithm is shown in Fig. 3, and is composed of two parts. The first part iteratively estimates the NURD vector $\bar{P}_k$ with the last stabilized frame $\hat{F}_{k-1}$ and the new frame $\tilde{F}_k$, which is pre-shifted with a previous estimated overall rotation $\bar{r}_{k-1}$. The second part estimates the new overall rotation $\bar{r}_k$ with the pre-shifted $\tilde{F}_k$ and a reference frame stack. The NURD warping vector $\bar{P}_k$ is fused with $\bar{r}_k$, to compensate the iterative drift of NURD estimation. By doing so, a final warping vector $\hat{P}_k$ is estimated. $\hat{P}_k$ is equal in length to the polar image height H, which is the number of A-lines in a single OCT frame. Details on these two parts and the fusion algorithm are presented in the following two subsections.

1) NURD Estimation: The first part of the stabilization algorithm follows two steps to estimate the NURD vector between the previous stabilized $\hat{F}_{k-1}$ and pre-shifted frame $\tilde{F}_k$: First a correlation map M_k is calculated using

Correlation Map $\boldsymbol{M}_k$

1×4 conv,stride(1,2),8	6×6 conv,stride(2,2),32
3×4 conv,stride(1,2),16	4×4 conv,stride(2,2),96
3×4 conv,stride(1,2),32	4×4 conv,stride(2,2),288
2×4 conv,stride(1,2),64	2×2 conv,stride(2,2),864
2×4 conv,stride(1,1),128	2×2 conv,stride(2,2),1728
1×1 conv,stride(1,1),1	2×2 conv,stride(2,2),64
	Reshape

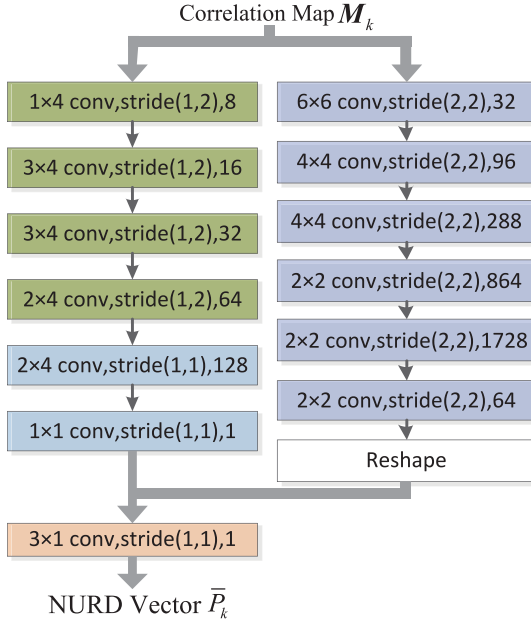3×1 conv,stride(1,1),1

NURD Vector $\overline{P}_k$

Fig. 4. Schematics of the CNN for NURD estimation. This network has two sub-branches that have different convolution kernels and striding.

cross-correlation, and then a CNN estimates the NURD vector taking the $\boldsymbol{M}_k$ as input.

Given two consecutive OCT frames with height H (with H A-lines in one frame) and width W in polar coordinates, we used the Pearson correlation coefficient to reflect the similarity between two local image patches from these two frames:

$$s_{i,j} = \frac{\sum_{l=1}^{n} \boldsymbol{f}_{i,l} \boldsymbol{f}'_{j,l} - n\bar{\boldsymbol{f}}_i \bar{\boldsymbol{f}}'_j}{\sqrt{\sum_{l=1}^{n} \boldsymbol{f}_{i,l}^2 - n\bar{\boldsymbol{f}}_i^2} \sqrt{\sum_{l=1}^{n} \boldsymbol{f}_{j,l}'^2 - n\bar{\boldsymbol{f}}_j'^2}} \quad (1)$$

where $\boldsymbol{f}_i \in \mathbb{R}^{h \times W}$ ($h \ll H$) is a local image patch centered at index position i, $i \in [0, H)$ of the newest frame. Correspondingly, $\boldsymbol{f}'_j \in \mathbb{R}^{h \times W}$ is a local image patch centered at index position j of $\hat{\boldsymbol{F}}_{k-1}$ ($j \in [i - w/2, i + w/2)$, w is the shifting window width). $s_{i,j}$ is one element of the correlation map $\boldsymbol{M}_k \in \mathbb{R}^{H \times w}$. The pixel index l will operate through the patch size $n = h \times W$. $\bar{\boldsymbol{f}}_i$ and $\bar{\boldsymbol{f}}'_j$ are the mean values of patches $\boldsymbol{f}_i$ and $\boldsymbol{f}'_j$ respectively.

The correlation map $\boldsymbol{M}_k$ contains information on the NURD, and we design a CNN to further extract the NURD vector. The detailed diagram of the NURD estimating nets is shown in Fig. 4, where each block gives the convolutional kernel size, the stride step length and the output feature map depth of each convolution layer. Two convolution sub-branches are operated in parallel to capture features and produce hierarchically coarse-to-fine responses. The main difference between these two sub-branches is the strides control. Both the left sub-branch and the right sub-branch of the NURD estimation nets have 6 convolutional layers, and a LeakyReLU activation [49] is used after each convolution layer.

The left sub-branch always keeps the vertical stride as 1. The objective of this design is to emphasize the vertical spacial correspondence. By doing so, the front 5 feature extraction

layers can gradually reduce the feature map width from w to 1, while maintaining the feature map height H as the 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 $5th$ feature map $A_F^5 \in \mathbb{R}^{H \times 1 \times 128}$ 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 $5th$ feature map and decrease channels to a sub-branch output of $\bar{P}$ with size $H \times 1$.

Considering that in certain situations $\boldsymbol{M}_k$ will miss valid information for some row m_i when there is no feature in a patch (window) $\boldsymbol{f}_i$ of $\tilde{\boldsymbol{F}}_k$, the right sub-branch is added. This sub-branch loosens the horizontal stride to 2 to involve more spacial information. This is a more common architecture for many deep learning applications [45]. Compared to the left sub-branch, this sub-branch can extract higher abstract features which are less sensitive to local missing information in the correlation map. We put larger kernels in the front layers to involve more common spatial information for neighboring positions of convolution, which prevents the under-fitting of feature extraction [16]. The kernel size is reduced in deeper layers to ensure a good trade-off between kernel size and computational burden. After the six convolutions, this subbranch outputs a matrix $\bar{\boldsymbol{P}}''_m$ of size $(H/2^6) \times w$. To get an output of warping vector size $H \times 1$, a reshape operation is applied to the output of this sub-branch (i.e., by sequentially connecting each $1 \times w$ row). A final convolution layer with kernel size 3×1, input depth 2 and output depth 1, combines the prediction of two sub-branches to get an estimation of $\bar{P}$. The loss function for training the NURD estimating nets uses the conventional L_2 loss and an additional continuity loss. A standard L_2 loss is described by Eq. (2):

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

where P_i is the element of P, the true NURD vector (ground truth), $n_p = H$ 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 the continuity of the distortion vector [8]–[10], we add the continuity loss as follows:

$$L_c = \frac{1}{n_p - 1} \sum_{i=1}^{n_p - 1} \left(\bar{P}_{k,i} - \bar{P}_{k,i+1} \right)^2 \quad (3)$$

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

$$L_A = \alpha L_c + (1 - \alpha) L_2. \quad (4)$$

2) Drift Compensation: It is possible to correct a short frame stack with just the NURD estimation part presented in the previous subsection. To compensate the accumulative error through iterations (especially for a long data stream), we estimate an overall rotation value $\bar{r}_k$.
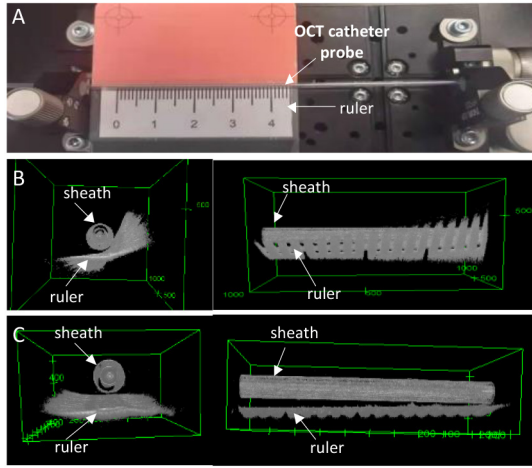
Fig. 5.   (A) Setup for sheath registration; (B) Original registration data in 3D view; (C) Calibrated registration data in 3D view.

In conventional pullback scans only the sheath data is consistent no matter what the environment outside the sheath is. According to this, the overall rotation can be observed by matching real-time sheath images with pre-recorded reference sheath images. However, when using the unstabilized pullback scanning to record the reference images, it still suffers from rotational distortion. To ensure that the orientations of reference frames are correct, we follow a calibration procedure. The setup of the sheath registration is shown in Fig. 5 (A), which relies on a external calibration object (a straight and flat ruler). As shown in Fig. 5 (B), the raw reference data still has the rotational distortion, which leads to both the ruler and sheath images located at wrong direction. We extract the contour surface of the ruler, and align the raw reference frame stack by minimizing the surface distance of all frames. By doing so, the rational error of the raw reference volumetric data is reduced from 59.4° to 2.79° (see Fig. 5 (C)). This calibrated reference data composes one of the inputs of the overall rotation estimation. Another input is composed by real-time B-scans, where image outside the sheath is masked out and only the sheath part is used.

With a recorded sheath image stack $S \in \mathbb{R}^{H \times W \times N}$ (N is the number of frames in the entire reference stack), we compute Euler distances to estimate the rotation deviation $\Delta \bar{r}_k$ of pre-shifted image $\tilde{F}_k$ (pre-shifted by $\bar{r}_{k-1}$):

$$d_r = \mathcal{L}\left(\tilde{S}_k, S_k\right) \qquad (5)$$

where $\tilde{S}_k \in \mathbb{R}^{H \times W \times l}$ is buffered pre-shifted original images, $S_k \in \mathbb{R}^{H \times W \times l}$ is the corresponding selected buffer indexed at k of S ($l \ll N$, and assuming that in every pullback scanning the torque coil moves with equal interval). The function $\mathcal{L}$ calculates Euler distance of two input buffers. For the pullback scan, S_k is the corresponding reference image in the data stack, while to correct an outer robotic tool pullback 3D scan the algorithm can be simplified and just use the first frame F_0 as the reference, because in this scanning condition the relative longitude position between sheath and rotating ball-lens is constant. For both $\tilde{S}_k$ and S_k only the sheath part is used for distance calculation. The index $r \in (-w, w)$ shifts within a given window w to calculate the distance d_r, and the position with the lowest distance value is considered as the matched angular position of $\Delta \bar{r}_k$. Finally, the overall rotation used to compensate the NURD drift is computed by $\bar{r}_k = \bar{r}_{k-1} + \Delta \bar{r}_k$.

We use the concept of a *PI Complementary Filter* [31] to fuse the $\bar{P}_k$ vector with the $\bar{r}_k$ value, which is shown to be efficient and able to solve the warping vector integral drift problem. A discrete form of PI complementary filter is:

$$\hat{P}_k = k_p \bar{P}_k + \left(1 - k_p\right)\bar{r}_k \mathbf{1} + k_i I_k \qquad (6)$$

$$I_k = I_{k-1} + \left(\bar{r}_k \mathbf{1} - \hat{P}_k\right) \qquad (7)$$

where k_p and k_i are PI compensating gains. I_k is the integral component vector. $\mathbf{1}$ is a vector of ones. Each element $\hat{P}_{k,i}$ of $\hat{P}_k$ represents the angular distance between the position of the ith A-line of $\tilde{F}_k$ and its correct position in polar domain. Applying $\hat{P}_k$ to this raw frame $\tilde{F}_k$, a stabilized frame $\hat{F}_k$ is obtained.

C. Experiment

1) Data: We test the proposed stabilization algorithm for both the internal conventional pullback and the outer scanning using a custom made OCT system (see description in Section III-A). We collect videos with the testing object (Fig. 2 (C)) to evaluate the performance for both types of scanning mode, since it has a special geometry. 6 scans were collected for each scanning mode, and each data stream comprises 500 frames. For the internal pullback, one reference scan is collected following the procedure described in Section III-B2, and this reference scan is used in the correction of all internal pullback scans. To perform quantitative analysis, we also use synthetic videos generated from a variety of OCT images [50]–[52] to test the proposed algorithm, since it is almost impossible to manually annotate the NURD in the OCT data. We used the published videos from [50]–[52], and extracted all the OCT images out and transformed them into the same resolution in the polar domain. Approximately 5000 images which cover the catheter based OCT imaging in intravascular, digestive tract and lung respiration airway are obtained from these public videos.

2) NURD Estimation Networks Training: The network is only trained with synthetic videos by intentionally adding NURD artifacts to the Optical Coherence Tomography (OCT) images, and tested on both true OCT scans and synthetic scanning videos. To ensure that the artificially added distortion covers the distribution of artifacts present in real videos, we first compute the correlation map M_k of every neighboring raw image pairs in real videos, and then measure the maximum NURD error with a graph path searching method [11]. The range of the GS-observed NURD was then expanded by 1/3 on each side. For each synthetic image, the synthetic NURD vector was then randomly generated within this expanded range. In addition, we implemented data augmentation that included geometric transformations, noise addition, brightness/contrast modification and OCT speckle/shadow simulation. By doing so the domain of artifacts of synthetic videos should cover the distribution of real artifacts. We randomly split all the

collected images by 1:1:1 into train, validation and test data. For both the CNN training and algorithm deployment, we use the same PC with an Nvidia Qt1000 graphic card and Intel i5-9400H CPU. The code is implemented using the Pytorch framework [53]. For training, Batch Normalization (BN) is used right after convolution and before activation [54], dropout is not used [58] and weight initialization is performed following the method described in [57]. The CNN is trained with the Stochastic Gradient Descent (SGD) weights optimization method (we used a weight decay of 0.0001 and a momentum of 0.9).

3) Metrics: For synthetic videos we calculate the NURD estimation error to evaluate the accuracy. For true videos, no guaranteed ground truth is available. We calculate the normalized Standard Deviation (STD) σ for a certain time window of videos to access the stabilization performance [8]. The definition of STD is:

$$\sigma = \frac{1}{N_{sig}} \sum_{i=1, j=1}^{N_{sig}} \bar{\sigma}\left(f_{i,j}\right) \tag{8}$$

where $\bar{\sigma}\left(f_{i,j}\right)$ is the STD calculated with pixel $f_{i,j}$ in one stacked frame stream, i and j are selected pixel indices on horizontal and vertical axis respectively. N_{sig} is the number of pixels used to calculate $\bar{\sigma}$. Based on the STD calculation, we give a threshold to differentiate unstable pixels from stable pixels, and count the number of unstable pixels for each frame.

To evaluate the rotation error through a complete scan, we compute en-face projections [10], [11] of the OCT videos where each A-line is accumulated to one single value, so that the volumetric data in the 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 to a longitudinal volumetric scanning (3D Scan). By counting the shift of max intensity pixels in a whole stack, we can evaluate the precession angle. We also measure the local fluctuation on the en-face image, which represents the local shaking between neighboring frames.

IV. RESULTS

A. Pullback Actuation Comparison

To compare distortions we effectuated two proposed types of scans within a 21 mm long rectangular tube target 3D printed in polyjet VeroWhite material (Stratasys, Rehovot, Israel), as shown in Fig. 2 (C). Table I summarizes statistical analysis of distortion of these 2 types of scans. We have also compared them to a stationary scanning without moving the optical components on the catheter tip, which is used as a baseline for hardware instability. It can be seen that the stationary scan data has lower precession, local rotation angle and STD (we also separate the sheath image STD from the full image STD) compared to any of the volumetric scanning.

These two volumetric scanning methods introduce instability into the data, and in particular the conventional internal pullback scanning presents larger image rotation between successive 2D radial frames, with a maximum precession angle of 79.6°. When performing volumetric robotic scanning of the OCT tool the image rotation (precession) is reduced

TABLE I
STATISTICS OF DIFFERENT SCANNING METHODS

Method	Preces-sion(°)↓	local (°)↓	STD↓	Sheath STD↓
Stationary	**11.6**	12.9±2.5	16.6±2.2	**13.3±2.5**
Conventional	79.6	13.6±3.1	18.6±2.4	19.8±2.6
Robotic 3D	24.7	12.3±1.9	17.2±2.9	16.9±2.6

TABLE II
COMPARISON WITH BASELINE METHOD

Method	Prec(°) ↓	local (°)↓	MPC(×10³)↓
Original (conventional)	79.6	13.6±3.1	8.72±4.46
Baseline (conventional)	79.1	13.1±3.2	2.94±2.31
Proposed (conventional)	16.0	**2.61±0.4**	**1.61±0.72**
Original (robotic)	24.7	12.3±1.9	10.9±5.90
Baseline (robotic)	24.6	12.1±2.3	3.79±2.83
Proposed (robotic)	**6.05**	**3.02±0.8**	**2.76±1.03**

TABLE III
ABLATION STUDY FOR TWO SCANNING MODES AND SYNTHETIC SCANS

Method	Prec(°) ↓	MPC(×10³) ↓
Proposed full (conventional)	**16.0**	**1.61±0.72**
w/o overall (conventional)	21.6	2.28±0.95
w/o NURD (conventional)	23.4	6.86±4.32
Proposed full (robotic)	**6.05**	**2.67±1.03**
w/o overall (robotic)	61.4	6.02±3.22
w/o NURD (robotic)	13.4	8.91±4.83
original (synthetic)	132	19.8±5.30
Proposed full (synthetic)	**2.61**	**2.60±1.43**
w/o overall (synthetic)	136	3.01±1.72
w/o NURD (synthetic)	12.9	5.45±2.95

to within the range of 24.6°. This constitutes an improvement and reduces the difficulty of software stabilization (see Section III-B). Other metrics such as the local rotation and STD reflect the short term instability. Although the 3 scanning methods share similar short term instability, the stationary scan is sightly better than the internal pullback scan and robotic endoscopic scan.

B. Software Improvement

1) Quantitative Results: The proposed stabilization algorithm is applied to both scanning modes. A pixel-wise stack smoothing is deployed to be the baseline method (average filtering of each pixel using 3 consecutive frames).

Table II shows the precession of the whole data stream, local fluctuation angle and **mean pixel counts (MPC)** of unstable pixels. For both scanning modes, the precession angle is reduced by 75%-80%, and for the robotic scan the precession is stabilized to only 6°. The baseline method can reduce the intensity derivation by blurring the 3D data, but does not improve the rotations and angular fluctuations. The proposed method not only reduces the overall rotations, but also reduces the pixel-wise instability without blurring the image or changing other information within every OCT frame (see the images in the subsection for qualitative results).

2) Ablation Study: We use both the conventional internal pullback and the robotic scan and the synthetic video to study
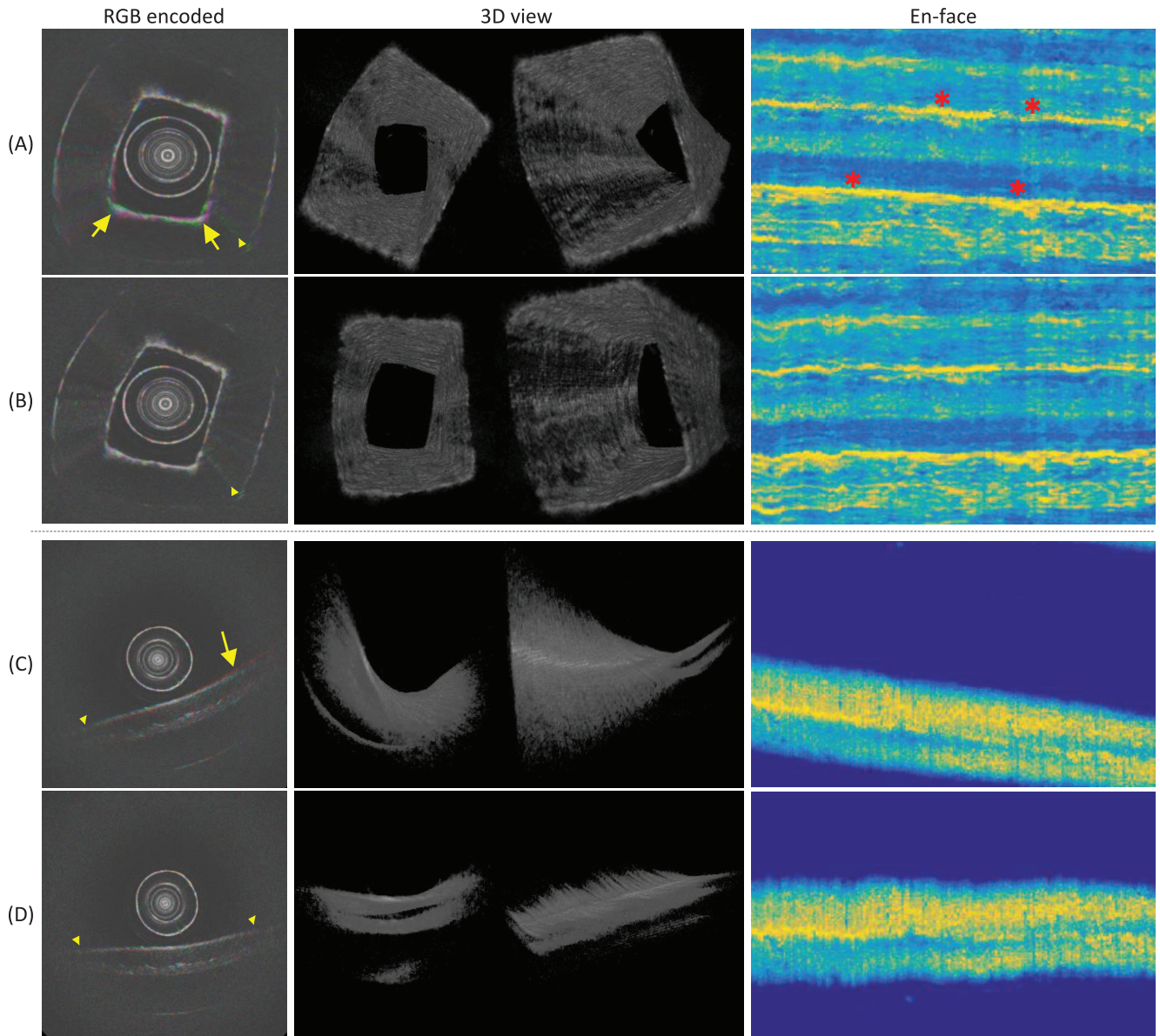
Fig. 6. Result presented in RGB encoded image, 3D view and en-face projection. (A) shows results of an original scan of rectangular phantom and (B) shows the corresponding stabilized data. (C) shows results of an original scan of layered colon tissue phantom and (D) shows the corresponding stabilized data. The yellow arrows point out rotation artifacts represented by colorful pixels in B-scans, the yellow arrowheads point out colorful pixels caused by the features changes of the object itself, and the red asterisks mark out aliasing artifacts in en-face projections.

the functionality of the two parts of the algorithm (the element-wise NURD estimation part, and the overall rotation estimation part). We did a thorough ablation comparison by disabling any of those two parts, and the results are shown in Table III. The most detrimental modification is to disable the overall estimation part. In this case the proposed algorithm can no longer provide a valid angular value to compensate the drift error, and the precession error is no longer corrected. Sometimes it will even introduce a precession larger than the original precession, for example in the case of the real robotic scanning and synthetic videos. Disabling the element-wise NURD estimation also affects the reduction of the precession angle, but not as much as disabling the overall estimation. Compared to our full stabilization algorithm, disabling the NURD estimation obviously increases the unstable pixel count, because the

overall rotation estimation can not align individual A-lines. This is worse for real scans where the overall rotation may have a larger estimation error due to the sparse features in sheath matching.

3) Qualitative Results: Fig. 6 shows qualitative results of pullback scans. To illustrate the NURD instability, we encode each 3 consecutive frames in independent channels of RGB color images [8]. In this RGB encoding, a colorful part of the image indicates the NURD artifact, and a well stabilized frame sequence should have minimal colorful pixels. We transform the data stream into the en-face projection (see description in Section IV-A). All these formats of results are presented in Fig. 6. It can be seen that the "colorful" pixels of the original sequence (RGB encoded) is reduced by the proposed algorithm, and without changing the image quality. The 3D results
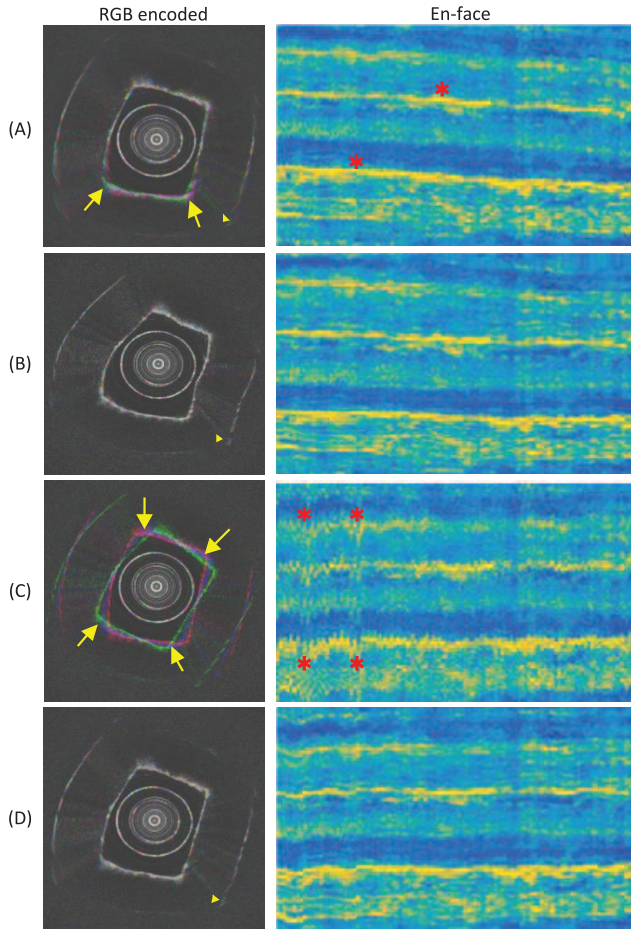
Fig. 7. Ablation comparison results for scans of rectangular phantom. RGB encoded images and en-face projections of (A) original data, (B) stabilized data without (w/o) overall rotation estimation, (C) stabilized data with only overall rotation estimation and (D) stabilized data of the full proposed algorithm. The yellow arrows point out rotation artifacts represented by colorful pixels in B-scans, the yellow arrowheads point out colorful pixels caused by the feature changing of the object itself, and the red asterisks mark out aliasing artifacts in en-face projections.
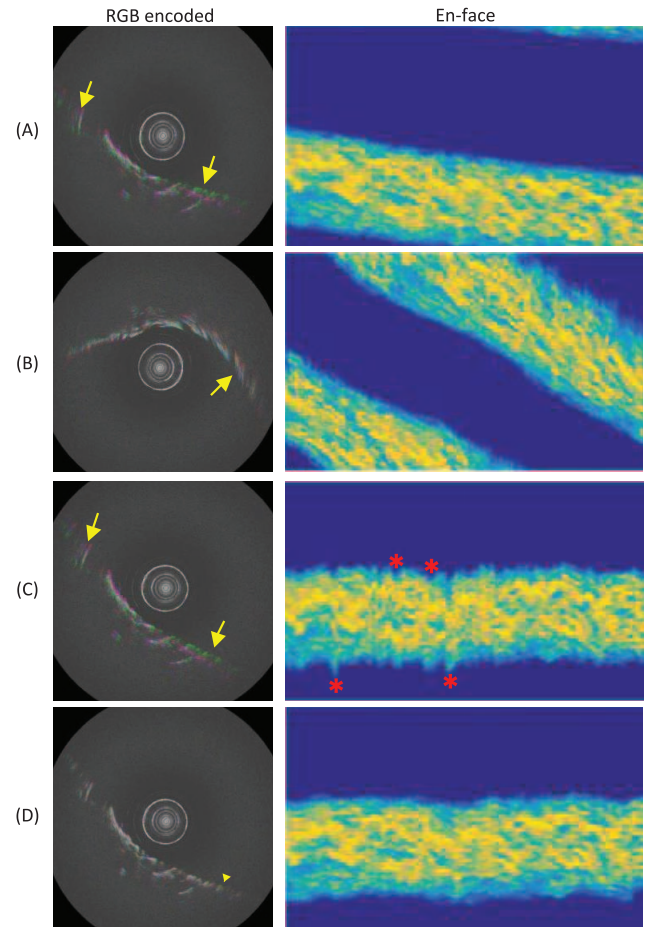

Fig. 8. Ablation comparison results for scans of sponge surface. RGB encoded images and en-face projections of (A) original data, (B) stabilized data without overall rotation estimation, (C) stabilized data with only overall rotation estimation and (D) stabilized data of the full proposed algorithm. The yellow arrows point out rotation artifacts represented by colorful pixels in B-scans, the yellow arrowheads point out colorful pixels caused by the feature changing of the object itself, and the red asterisks mark out aliasing artifacts in en-face projections.

are consistent with the en-face projections, where the angular rotation is reflected by the shift of high intensity pixels of en-face images. It can be seen that the stabilization algorithm straightens the high intensity lines of the original en-face projection, which maintains the intensity distribution of the whole data stream. More qualitative visual comparisons can be seen in Supplementary Movies.

Qualitative results on the ablation study are presented in Figs. 7 and 8. Keeping only the NURD estimation will introduce iterative distortion to the image, and also limit the precession reduction (Fig. 7 (B) and Fig. 8 (B)). Fig. 7 (C) and Fig. 8 (C) show the results with only the overall rotation estimation from the reference data stream. In this case the shaking of the frame sequence can be enlarged due to a large angular estimation error. The fusion stage from the full algorithm can effectively limit these phenomena.

4) Comparison With State of the Art: In addition, we compare the NURD estimation part of the proposed CNN based method and a state-of-the-art approach based on the graph path searching (GS) [11], which is not a learning based method.
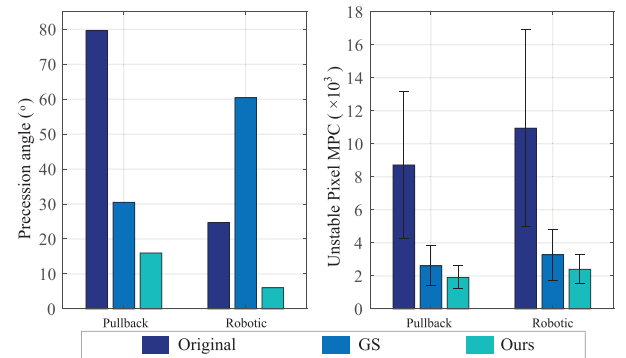

Fig. 9. Comparison with GS method. The left figure shows the max precession angles of difference data output, the right figure shows the in-stable pixel MPC comparison.

Since this online correction method has no drift compensation, we disable the overall rotation estimation to compare the NURD estimation performance, using the indirect metrics mean pixel account (MPC) for in-stable pixels of real

TABLE IV
NURD ESTIMATION EVALUATION

Method	MSE(Deg^2)↓	Prec. Reduction (%) ↑	MPC Reduction (%) ↑
GS [10]	6.47 ± 1.62	12.85	70.26 ± 14.3
Proposed	**0.38 ± 0.09**	**78.86**	**77.78 ± 8.90**

scans. We also compare them for the synthetic scans, using mean square error (MSE) of NURD estimation as a direct metric. Results from original scans and different algorithms are shown in Fig. 9. Table IV shows statistics of these comparisons, the NURD estimation error of the proposed method is significantly lower than graph searching based method and can better reduce the number of unstable pixels (MPC).

V. CONCLUSION

We have developed a new solution to tackle the rotational distortion problem using deep CNN, which can be generalized for scanning situations with different targets and catheters. We proposed a new warping vector estimation net to estimate NURD between adjacent frames, which has a higher accuracy and robustness compared with the conventional approach in situations where the images have few features. Moreover, we solved the problem of drift error accumulation in iterative video process, with a group rotation estimation net. We were able to apply the CNN based algorithm trained on synthetic data to real videos acquired in various scanning conditions. Compared to the A-line level NURD vector estimated with tissue information, the orientation estimation using the sheath is less accurate because the feature information provided by the sheath is even more sparse. Nevertheless, this overall rotation is robust and will not be affected by integral drift. The overall rotation is a redundant complementary estimation besides the NURD estimation. Fusing the NURD warping vector with the overall rotation value can form a robust A-line level warping vector which suppresses the integral drift.

An ablation study has been performed on different scanning data by calculating a quantitative stabilization metric and qualitative analysis, which shows that both the NURD estimation and overall rotation are essential for the stabilization of the OCT volumetric scanning. The proposed algorithm outperforms other state-of-the-art methods [11] that iteratively estimate the warping path using a graphic searching approach without a drift compensation module.

In the literature of OCT stabilization, very little research study can be found, and existing hardware solutions are mainly based on optical marker for calibration [10]. On the other hand, software methods are for stationary scanning without longitudinal translation [8] or did not consider the precession [11]. These image based methods will certainly be affected by accumulative error in the volumetric scanning. The proposed stabilization algorithm is designed for both conventional pullback scan and free 3D scan. For the conventional pullback scan we design a sheath registration and calibration procedure to record reference data. The new proposed 3D scan with robotic tool pullback in a endoscope channel provides two benefits for OCT volumetric imaging. The first is that it reduces the unstable friction and reduces the precession from hardware level. The second benefit is that the sheath in the first frame can be reference for overall rotation estimation, which reduces the procedures of the further stabilization algorithm.

The proposed method assumes that the appearance change between consecutive images is more affected by the rotational artifacts than by real changes of tissue. The performance of the algorithm can be impaired in some cases where the appearance of tissue in the field of view changes rapidly, for example when the probe is brought to the tissue very fast or in the presence of peristalsis. However, in the majority of clinical situations, to allow thorough inspection of the tissue and to avoid its damaging, the probe is maneuvered slowly, and thus the algorithm will achieve good stabilization.

REFERENCES

[1] M. Gora, M. Suter, G. Tearney, and X. Li, "Endoscopic optical coherence tomography: Technologies and clinical applications [Invited]," *Biomed. Opt. Exp.*, vol. 8, no. 5, pp. 2405–2444, 2017.

[2] M. Atif, H. Ullah, M. Y. Hamza, and M. Ikram, "Catheters for optical coherence tomography," *Laser Phys. Lett.*, vol. 8, no. 9, p. 629, 2011.

[3] T. Okamura *et al.*, "*In vivo* evaluation of stent strut distribution patterns in the bioabsorbable everolimus-eluting device: An OCT ad hoc analysis of the revision 1.0 and revision 1.1 stent design in the ABSORB clinical trial," *EuroIntervention*, vol. 5, no. 8, pp. 932–938, 2010.

[4] H.-C. Lee *et al.*, "Circumferential optical coherence tomography angiography imaging of the swine esophagus using a micromotor balloon catheter," *Biomed. Opt. Exp.*, vol. 7, no. 8, pp. 2927–2942, 2016.

[5] K. Liang *et al.*, "Ultrahigh speed EN face OCT capsule for endoscopic imaging," *Biomed. Opt. Exp.*, vol. 6, no. 4, pp. 1146–1163, 2015.

[6] Y. Kawase *et al.*, "Comparison of nonuniform rotational distortion between mechanical IVUS and OCT using a phantom model," *Ultrasound Med. Biol.*, vol. 33, no. 1, pp. 67–73, 2007.

[7] W. Kang *et al.*, "Motion artifacts associated with *in vivo* endoscopic OCT images of the esophagus," *Opt. Exp.*, vol. 19, no. 21, 2011, Art. no. 20722.

[8] G. V. Soest, J. Bosch, and A. V. D. Steen, "Azimuthal registration of image sequences affected by nonuniform rotation distortion," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 3, pp. 348–355, May 2008.

[9] O. O. Ahsen *et al.*, "Correction of rotational distortion for catheter-based en face OCT and OCT angiography," *Opt. Lett.*, vol. 39, no. 20, pp. 5973–5976, 2014.

[10] N. Uribe-Patarroyo and B. E. Bouma, "Rotational distortion correction in endoscopic optical coherence tomography based on speckle de-correlation," *Opt. Lett.*, vol. 40, no. 23, pp. 5518–5521, 2015.

[11] E. Abouei *et al.*, "Correction of motion artifacts in endoscopic optical coherence tomography and auto fluorescence images based on azimuthal EN face image registration," *J. Biomed. Opt.*, vol. 23, no. 1, pp. 1–14, 2018.

[12] S. Sathyanarayana, "Nonuniform rotational distortion (NURD) reduction," U.S. Patent 7 024 025 B2, Apr. 4, 2006.

[13] G. J. Ughi, "Automatic three-dimensional registration of intravascular optical coherence tomography images," *J. Biomed. Opt.*, vol. 17, no. 2, 2012, Art. no. 026005.

[14] C. Gatta, O. Pujol, O. Leor, J. Ferre, and P. Radeva, "Fast rigid registration of vascular structures in IVUs sequences," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 1006–1011, Nov. 2009.

[15] H. Zhang, W. Yang, H. Yu, H. Zhang, and G.-S. Xia, "Detecting power lines in UAV images with convolutional features and structured constraints," *Remote Sens.*, vol. 11, no. 11, p. 1342, 2019.

[16] I. J. Goodfellow, Y. Bengio, and A. Courville, "Part II: Modern practical deep networks," in *Deep Learning*. Cambridge, MA, USA: MIT Press, 2017.

[17] J. V. D. Putten *et al.*, "Tissue segmentation in volumetric laser endomicroscopy data using FusionNet and a domain-specific loss function," in *Proc. Med. Imag. Image Process.*, 2019, Art. no. 109492J.

[18] D. Li *et al.*, "Parallel deep neural networks for endoscopic OCT image segmentation," *Biomed. Opt. Exp.*, vol. 10, no. 3, p. 1126, Jul. 2019.

[19] Y. L. Yong, L. K. Tan, R. A. Mclaughlin, K. H. Chee, and Y. M. Liew, "Linear-regression convolutional neural network for fully automated coronary lumen segmentation in intravascular optical coherence tomography," *J. Biomed. Opt.*, vol. 22, no. 12, p. 1, 2017.

[20] J. V. D. Putten *et al.*, "Deep principal dimension encoding for the classification of early neoplasia in Barretts Esophagus with volumetric laser endomicroscopy," *Comput. Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101701.

[21] Y. Zeng *et al.*, "Real-time colorectal cancer diagnosis using PR-OCT with deep learning," *Theranostics*, vol. 10, no. 6, pp. 2587–2596, 2020.

[22] K.-K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. V. Gool, "Convolutional oriented boundaries: From image segmentation to high-level tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 819–833, Apr. 2018.

[23] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, 2nd ed. Pacific Grove, CA, USA: PWS, 1999.

[24] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3982–3991.

[25] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4380–4389.

[26] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder–decoder network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 193–202.

[27] J. L. Crassidis, F. L. Markley, and Y. Cheng, "Survey of nonlinear attitude estimation methods," *J. Guid. Control Dyn.*, vol. 30, no. 1, pp. 12–28, 2007.

[28] A. Philipp and S. Behnke, "Robust sensor fusion for robot attitude estimation," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2014, pp. 218–224.

[29] R. Mahony, T. Hamel, and J. M. Pflimlin, "Complementary filter design on the special orthogonal group SO (3)," in *Proc. 44th IEEE Conf. Decis. Control*, 2005, pp. 1477–1484.

[30] J. Justa, V. Šmídl, and A. Hamáček, "Fast AHRS filter for accelerometer, magnetometer, and gyroscope combination with separated sensor corrections," *Sensors*, vol. 20, no. 14, p. 3824, 2020.

[31] J. Wu, Z. Zhou, H. Fourati, and Y. Cheng, "A super fast attitude determination algorithm for consumer-level accelerometer and magnetometer," *IEEE Trans. Consum. Electron.*, vol. 64, no. 3, pp. 375–381, Aug. 2018.

[32] D. Gebre-Egziabher, R. Hayward, and J. Powell, "Design of multi-sensor attitude determination systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 40, no. 2, pp. 627–649, Apr. 2004.

[33] Y. S. Suh, "Simple-structured quaternion estimator separating inertial and magnetic sensor effects," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 2698–2706, Dec. 2019.

[34] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, 2011, pp. 1–7.

[35] J. Yu and R. Ramamoorthi, "Learning video stabilization using optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8156–8164.

[36] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 225–232.

[37] M. Wang *et al.*, "Deep online video stabilization with multi-grid warping transformation learning," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2283–2292, May 2019.

[38] S. Liu, L. Yuan, P. Tan, and J. Sun, "SteadyFlow: Spatially smooth optical flow for video stabilization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4209–4216.

[39] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2432–2439.

[40] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2462–2470.

[41] F. Yu and V. Koltun. (2015). *Multi-Scale Context Aggregation by Dilated Convolutions*. [Online]. Available: https://arxiv.org/abs/1511.07122

[42] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2018–2025.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.

[44] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–6.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[47] O. C. Mora *et al.*, "Steerable OCT catheter for real-time assistance during teleoperated endoscopic treatment of colorectal cancer," *Biomed. Opt. Exp.*, vol. 11, no. 3, p. 1231, 2020.

[48] L. Zorn *et al.*, "A novel telemanipulated robotic assistant for surgical endoscopy: Preclinical application to ESD," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 4, pp. 797–808, Apr. 2018.

[49] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML Workshop Deep Learn. Audio Speech Lang. Process. (WDLASL)*, 2013, p. 28.

[50] T. Wang *et al.*, "Heartbeat OCT: *in vivo* intravascular megahertz-optical coherence tomography," *Biomed. Opt. Exp.*, vol. 6, no. 12, pp. 5021–5032. 2015.

[51] M. J. Gora *et al.*, "Tethered capsule endomicroscopy enables less invasive imaging of gastrointestinal tract microstructure," *Nat. Med.*, vol. 19, no. 2, pp. 238–240, 2013.

[52] S. W. Lee *et al.*, "Quantification of airway thickness changes in smoke-inhalation injury using *in-vivo* 3-D endoscopic frequency-domain optical coherence tomography," *Biomed. Opt. Exp.*, vol. 2, no. 2, pp. 243–254, 2011.

[53] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. NIPS-W*, 2017, pp. 1–4.

[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–9.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, p. 9.

[56] Y. Bengio, "Global optimization strategies," in *Learning Deep Architectures for AI*. Hanover, MA, USA: Now, 2009, pp. 99–104.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1026–1034.

[58] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing coadaptation of feature detectors," 2012. [Online]. Available: arXiv:1207.0580.