Medical Image Analysis 68 (2021) 101900

Contents lists available at ScienceDirect

Medical Image Analysis

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

A deep learning framework for quality assessment and restoration in video endoscopy

Sharib Ali[a,d,e,*], Felix Zhou[b,d], Adam Bailey[c,d,e], Barbara Braden[c,d,e], James E. East[c,d,e], Xin Lu[b,d,e], Jens Rittscher[a,d,*]

[a] Institute of Biomedical Engineering and Big Data Institute, Oxford, UK
[b] Ludwig Institute for Cancer Research, Oxford, UK
[c] Translational Gastroenterology Unit, Experimental Medicine Div., John Radcliffe Hospital, Oxford, UK
[d] University of Oxford, Old Road Campus, Oxford, UK
[e] Oxford NIHR Biomedical Research Centre, Oxford, UK

ARTICLE INFO

Article history:
Received 10 June 2019
Revised 31 October 2020
Accepted 9 November 2020
Available online 13 November 2020

Keywords:
Video endoscopy
Multi-class artifact detection
Multi-class artifact segmentation
Convolution neural networks
Frame restoration

ABSTRACT

Endoscopy is a routine imaging technique used for both diagnosis and minimally invasive surgical treatment. Artifacts such as motion blur, bubbles, specular reflections, floating objects and pixel saturation impede the visual interpretation and the automated analysis of endoscopy videos. Given the widespread use of endoscopy in different clinical applications, robust and reliable identification of such artifacts and the automated restoration of corrupted video frames is a fundamental medical imaging problem. Existing state-of-the-art methods only deal with the detection and restoration of selected artifacts. However, typically endoscopy videos contain numerous artifacts which motivates to establish a comprehensive solution.

In this paper, a fully automatic framework is proposed that can: 1) detect and classify six different artifacts, 2) segment artifact instances that have indefinable shapes, 3) provide a quality score for each frame, and 4) restore partially corrupted frames. To detect and classify different artifacts, the proposed framework exploits fast, multi-scale and single stage convolution neural network detector. In addition, we use an encoder-decoder model for pixel-wise segmentation of irregular shaped artifacts. A quality score is introduced to assess video frame quality and to predict image restoration success. Generative adversarial networks with carefully chosen regularization and training strategies for discriminator-generator networks are finally used to restore corrupted frames.

The detector yields the highest mean average precision (mAP) of 45.7 and 34.7, respectively for 25% and 50% IoU thresholds, and the lowest computational time of 88 ms allowing for near real-time processing. The restoration models for blind deblurring, saturation correction and inpainting demonstrate significant improvements over previous methods. On a set of 10 test videos, an average of 68.7% of video frames successfully passed the quality score (≥ 0.9) after applying the proposed restoration framework thereby retaining 25% more frames compared to the raw videos. The importance of artifacts detection and their restoration on improved robustness of image analysis methods is also demonstrated in this work.

1. Introduction

Originally used to image the esophagus, stomach and colon, miniaturization of hardware and improvement of imaging sensors now enable endoscopy of the ear, nose, throat, heart, urinary tract, joints, and abdomen. Despite recent hardware improvements of clinical endoscopes allowing high definition and high frame rate image capture, the quality of endoscopic videos are still compromised. This is mostly due to non-optimal reflection of light, unavoidable tissue movements, large variabilities in organ shape and surface texture, occlusions due to bodily fluids and debris. Most common imaging artifacts include the over- and under-exposure of image regions due to variability in illumination and organ topology (termed as “saturation” and “low contrast”, respectively), blur

* Corresponding authors.
E-mail addresses: sharib.ali@eng.ox.ac.uk (S. Ali), jens.rittscher@eng.ox.ac.uk (J. Rittscher).

https://doi.org/10.1016/j.media.2020.101900

due to unsteady hand motion of endoscopists and local organ motion, and specularity due to light reflection from smooth organ surfaces. The presence of fluid and bubbles also influence the visual interpretation of the examined mucosal surface. Detection, and localization of artifact regions can provide quantitative analysis of the actual surveillance of the mucosal surface significant for assessing the quality of the clinical endoscopy procedure. The online restoration of endoscopy frames from inevitable artifacts obscuring the underlying tissue and degrading the video frame quality can improve the mucosal surface visualization and hence minimize the risk of misdiagnosis or requirement for the repetition of endoscopic procedure and can even shorten the relative time of endoscopic surveillance required today.

A systematic approach to identifying bespoken image artifacts is required to improve the utility of endoscopy videos. By identifying artifacts in real-time it will be possible to provide the acquisition quality feedback to human operators during training as well as routine clinical operation. Additionally, video frames affected with artifacts adversely affects any computer assisted endoscopy methods such as video mosaicking (Ali et al., 2016a; 2016b), segmentation (Prasath, 2016), and automated detection (Zhang et al., 2018; Urban et al., 2018). Zhang et al. (2018) showed that the detected bounding boxes were vulnerable to artifacts consequently affecting the detector performance for polyp tracking in colonoscopy data. Similarly, Urban et al. (2018) utilized additional training samples from videos to reduce the false-positive detections due to random artifacts in capsule endoscopy. Also, Prasath (2016) reported that most published methods had to take additional care to avoid the pitfalls of artifacts to make their methods work on endoscopy videos. While, Ali et al. (2016a,b) manually selected the video frames in bladder endoscopy for effective mosaicking by ignoring frames with evident artifacts such as blur, floating objects and pixel saturation. Thus, to maximize the usability of the acquired data, the detection of multi-class artifacts should be combined with context specific image enhancement and restoration techniques. To do so, different methods should be used for correcting motion artifacts, pixel saturation, specularities, and the presence of bubbles or debris.

1.1. Related work

Chikkerur et al. (2011) and Menor et al. (2016) have introduced global quality metrics but these do not provide any additional information about the artefact itself. Such global quality scores only allow for the removal of corrupted frames and do not support any context specific image enhancement, i.e., neither information regarding the cause of frame quality degradation nor the degraded regions could be identified for frame restoration. Such a naïve removal of corrupted frames can severely deplete the information content of videos and affect their overall temporal continuity. Here, video mosaicking methods that can require an overlap of 60% or more can easily fail (Ali et al., 2016a; 2016b). Various research groups have studied the detection and correction of specific artifacts in endoscopic imaging (Akbari et al., 2018; Liu et al., 2011; Stehle, 2006; Tchoulack et al., 2008). For example, deblurring of wireless capsule endoscopy images utilizing a total variational (TV) approach was proposed in (Liu et al., 2011). TV-based de-blurring is however parameter sensitive and requires geometrical features to perform well. Endoscopic images have very sparse features and lack geometrically prominent structures. Both classical image processing methods (Stehle, 2006; Tchoulack et al., 2008; Mohammed et al., 2018) and machine learning approaches were applied (Akbari et al., 2018; Rodrguez-Snchez et al., 2017; Isabel et al., 2018) for endoscopy frame restoration. A major drawback of classical methods is that heuristically chosen image intensities are compared with neighboring (local) image pixels. For machine learning approaches, hand-crafted features (Akbari et al., 2018) as well as neural networks (Rodrguez-Snchez et al., 2017; Isabel et al., 2018) have been used to restore specular reflections. However, all these methods select to restore endoscopic images with one single artifact class while they are simultaneously corrupted with multiple artifacts. For example, both 'specularities' and a water 'bubble' can be present in the same frame. Moreover, the surrounding pixel appearances of these artifacts can also vary in different modalities (see Fig. 1) which can influence artifact detection and restoration methods. Finally, inter-patient variation is significant even when viewed under the same modality. Existing methods fail to adequately address all of these challenges. In addition to addressing one type of imaging artifact, only one imaging modality and a single patient video sequence are considered in most of endoscopy-based image analysis literature mentioned above.

The use of small size data sets and use of only small image patches in these studies also raise concern regarding method generalization to image variabilities often present in endoscopic data. For example, Akbari et al. (2018) used only 100 randomly selected images to train the Support Vector Machine (SVM) for detecting specular regions. Similarly, Rodrguez-Snchez et al. (2017) proposed to train an encoder-decoder network for segmentation of specularity regions for which 160 endoscopic images were used to train the network and 40 images were used for testing. A post processing scheme based on Euclidean distance was used to obtain the restoration and was not trained in an end-to-end fashion. While, Isabel et al. (2018) trained an end-to-end generative adversarial CNN for restoration of specularity in endoscopy frames utilizing 100 image patches per video from 10 different endoscopic videos. In general, both local and global information is required for realistic frame restoration. No addition information war provided on how patches were detected and extracted.

To overcome the limitations of previous methods, a complete framework for multi-class artifact detection, localisation, frame quality scoring and frame restoration of partially corrupted frames is proposed. Here, multi-modal and large dataset sizes are used to train and validate each module of the proposed framework. Literature on current trends in detection and restoration which are relevant to the implemented modules in this work have been discussed below.

Today, with the advancement of GPUs, convolutional neural networks (CNN) show tremendous capability to learn and generalize features for accurate detection compared to hand-crafted features. Current state-of-the-art detection methods includes Faster R-CNNs (Ren et al., 2015), You Only Look Once (YOLO, (Redmon et al., 2016)), and RetinaNet (Lin et al., 2017b). Faster R-CNNs (Ren et al., 2015) introduced a fully trainable end-to-end network yielding an initial region proposal network and successive classifications of the proposed regions without intermediate processing. Since region proposal generation precedes bounding box detection sequentially, this architecture is known as a two-stage detector. Though very accurate, the major drawback is its slow inference and extensive training. YOLO (Redmon et al., 2016) simplifies Faster R-CNNs to predict simultaneously class and bounding box coordinates using a single CNN and a single loss function with good performance and significantly faster inference time. This simultaneous detection is known as a one-stage detector. Compared to two-stage detectors, single-stage detectors mainly suffer two issues: high false detection due to 1) presence of varied size objects and 2) high initial number of anchor boxes requirement that necessitates more accurate positive box mining. The former is corrected by predicting bounding boxes at multiple scales using feature pyramids (He et al., 2014; Lin et al., 2017a). To address the latter, RetinaNet (Lin et al., 2017b) introduced a new focal loss which ad-

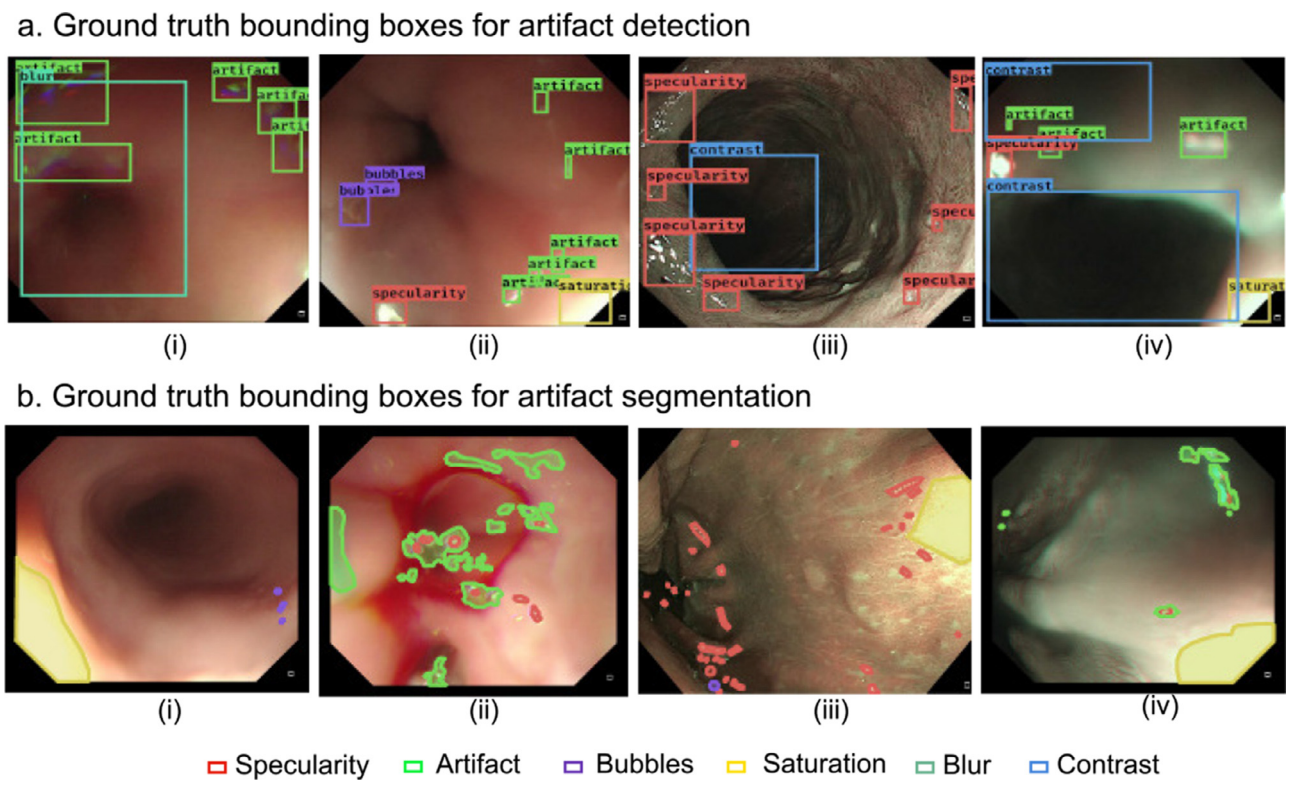

Fig. 1. Multitude of artifacts present in gastroesophageal endoscopy images. (i-ii) represent BF modality and (iii-iv) represents NBI. Annotations of both detection (a) and segmentation (b) are shown. Detection boxes are labelled for all six classes while segmentation labels are used only for four indefinable artifact classes that include specularity, misc. artifacts, bubbles and saturation.

justs the propagated loss to focus more on hard, misclassified samples. Recently, YOLOv3 (Redmon and Farhadi, 2018) simplified the RetinaNet architecture with further speed improvements. Bounding boxes are predicted only at three different scales (unlike five in RetinaNet) utilizing objectness score and an independent logistic regression to enable the detection of objects belonging to multiple classes unlike focal loss in RetinaNet. Collectively, Faster R-CNN, RetinaNet and YOLOv3 define the current state-of-the-art detection envelope of accuracy vs speed on the popular natural images benchmark COCO data set (Lin et al., 2014). Faster R-CNN has been widely used for polyp detection in colonoscopy videos (Mo et al., 2018; Shin et al., 2018).

Most deep learning-based biomedical image segmentation methods (Zhou et al., 2019) employ U-Net-based encoder-decoder architecture (Ronneberger et al., 2015). Fully convolutional network (FCN) has also proven to perform well especially when training datasets are small. Ben-Cohen et al. (2016) used FCN with a VGG16 backbone to perform liver segmentation on dataset with only 20 patients. Adding a mask head on top of existing bounding box regression layer, Faster R-CNN network has been used to perform object segmentation, commonly referred to as "Mask R-CNN' (He et al., 2017). However, a major disadvantage of such network is the heavy reliability on the region proposal network that may miss small objects. Similarly, U-Net-based architectures have shown promise in the medical imaging field in most cases, but such network can perform poorly to identify small and variable size objects. To tackle varied shapes including small size objects as in case of artifacts, feature pyramids or dilated convolutions or a combination of both can boost algorithm performances (Ali et al., 2020). Considering the variable spatial size, texture and locations of artifacts, and requirement of faster inference for clinical use; DeepLabv3+ (Chen et al., 2018) is the current state-of-the-art segmentation architecture.

Image restoration is the process of generating realistic and noise free image pixels from corrupted image pixels. In endoscopic frame restoration, depending upon the artifact type, the goal is either the generation of an entire noise-free image or pixel inpainting of undesirable pixels using surrounding pixel information (Barcelos and Batista, 2007). Convolutional neural networks have also been used as the backbones in designing generative adversarial networks (GANs, (Goodfellow et al., 2014)) for image reconstruction and restoration. GANs have been successfully applied to image-to-image translation problems using limited training data. Various modifications have been made in past to enable GANs to produce realistic images (Bang and Shim, 2018; Isola et al., 2017; Mirza and Osindero, 2014). For restoration of blur images of natural scenes, deblurGAN (Kupyn et al., 2017) and SRN deblurNet (Tao et al., 2018) have been proposed for blind deconvolution. It has been shown that adversarial networks (Kupyn et al., 2017; Tao et al., 2018) surpass computational speed and restoration quality of classical methods (Getreuer, 2012; Tong et al., 2004; Xu et al., 2013). For restoration of missing pixels and corrupted pixels, inpainting methods have been used in literature. Classical methods such as TV-inpainting methods are popular for restoring images with geometrical structures (Shen and Chan, 2002) and patch-based methods (Efros and Freeman, 2001) for texture synthesis. However, these methods are computationally expensive. Deep neural networks have proven to recover visually plausible image structures and textures (Köhler et al., 2014) with near real-time performance. However, they are limited to the size of the mask or the number of unknown pixels in an image. Similar to deblurring, GANs (Iizuka et al., 2017; Pathak et al., 2016; Yu et al., 2018)

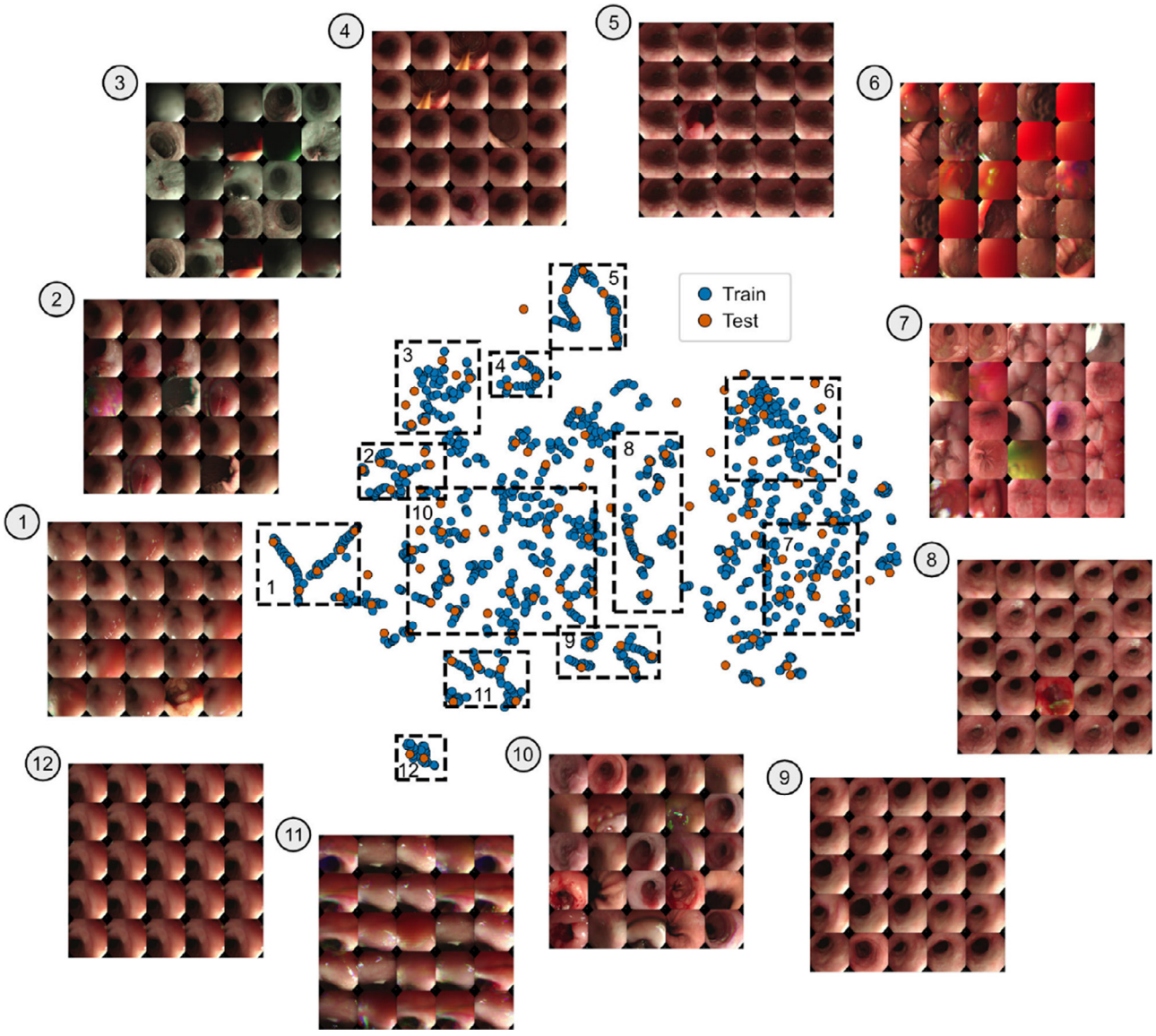

Fig. 2. t-SNE plot for train-test data distribution. The numbered dotted rectangles correspond to the test-train samples shown in the corresponding numbered image matrix. It can be observed that there is a good overlap between train and test samples.

Table 1
Artifact classes identified for detection and restoration of endoscopy frames.

Artefact type	Description
blur	fast camera motion (cyan box in Fig. 1(a))
bubbles	bubbles that distorts tissue appearance (purple box in Fig. 1(b))
specularity	mirror-like reflection (red box in Fig. 1(b-d))
saturation	overexposed bright pixel areas (yellow box in Fig. 1(b, d))
contrast	low contrast areas from underexposure (blue box in Fig. 1(c-d))
misc. artifact	chromatic aberration, debris etc. (green box in Fig. 1(a, b, d))

have been shown to be more successful in providing faster and more coherent reconstructions even with larger masks. However, for over-exposure correction, *i.e., saturation*, recently (Abebe et al., 2018) proposed an approach based on exploitation of correlation between RGB channels. Several other strategies such as inverse tone mapping and reshaping of brightness were also employed to adjust the variability in the natural scene data. To model such an heuristic correction of over-exposure in endoscopy data is however complex and tedious due to their large variability and presence of sparse texture.

1.2. Contributions

A fully automatic, systematic and comprehensive approach for detection of multi-class artifacts, segmentation of indefinable artifact instances, quality assessment and subsequent restoration of partially corrupted frames is presented. The presented framework addresses the detection and localization of six different artifacts and introduces artifact type specific restoration. To avoid the need for parameter adjustment and to overcome the limitations of hand-crafted features only suitable for specific artifacts, multiple class

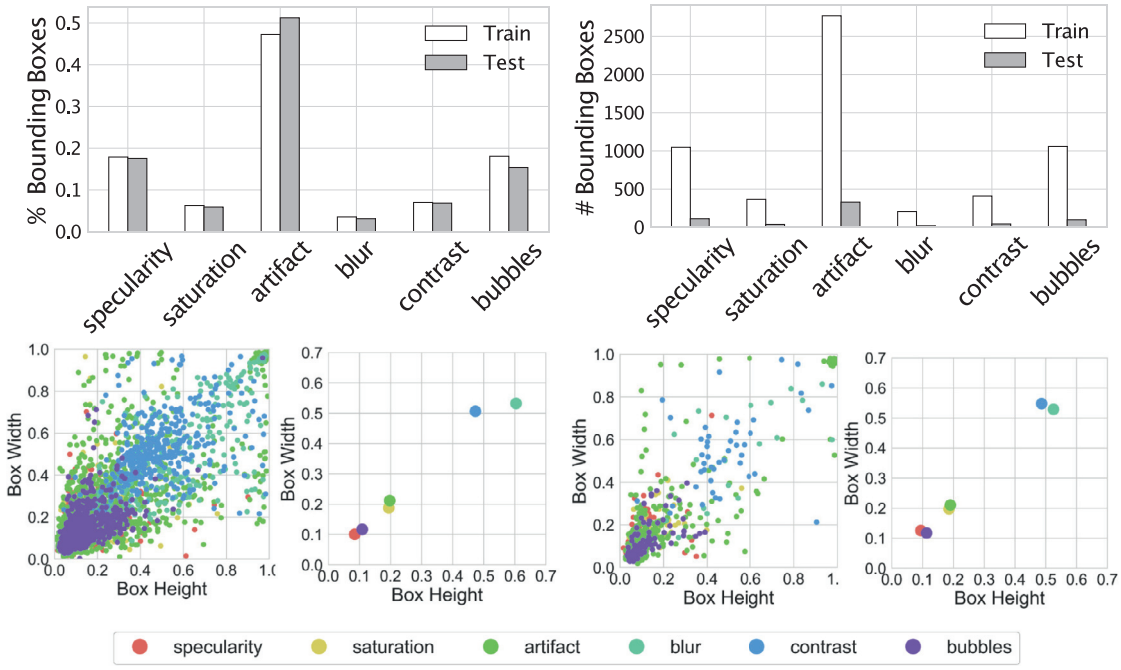

Fig. 3. Top row: artifact type distribution in the training and testing of endoscopy image dataset in terms of number of bounding boxes (left) and percentage of the total number of bounding boxes (right). Bottom row: the first plot on left represents normalised width vs height relative to source image width and height of annotated ground-truth bounding boxes (per small dot) colored by class for the training data while the second plot on the left corresponds to their mean width and mean height pair of each class for the training set. Similarly, the right plots represent that for the test set.

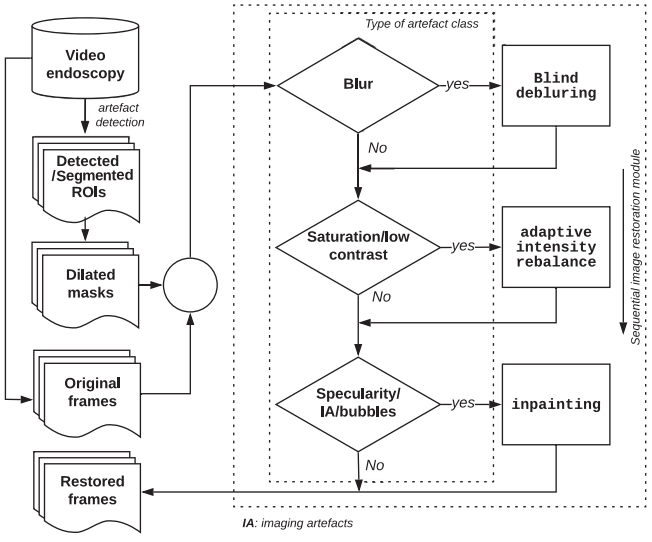

Fig. 4. Sequential processes for endoscopic image restoration from detected region-of-interests (ROI) of 6 different artifacts. Alternatively, masked regions from segmentation areas of indefinable shapes can be used for restoration. First, masks of generarted ROIs are dilated and then only these regions are used for restoration. Unlike, in case of blur, the entire image is used.

artifact are detected and localized and restoration methods are designed utilizing data taken from multi-patient and multi-modality.

A multi-scale single-stage deep convolutional neural network-based object detection is trained on cross patients and cross modality endoscopic data for multi-class artifact detection and localization. Multi-scale YOLOv3 architecture utilizing 4 scales and an additional spatial pyramid pooling layer (SPP) is proposed. Similarly, pixel-wise segmentation of artifact instances whose area cannot be approximated with a bounding boxes is also presented. For this, we have exploited an encoder-decoder model with atrous convolution and SPP layer to handle varied sized artifacts. To quantify the presence of multi-class artifacts in endoscopy image (location, size), and the complication associated with each artifact class for image restoration, a quality metric is proposed which scores each individual frame. Realistic frame restoration is then achieved for partially corrupted images using GANs (Goodfellow et al., 2014). Substantial work has been necessary to adopt these existing techniques to this specific setting and to avoid the introduction of additional artifacts and disruptions to the overall visual coherence. A novel contextual loss on both image and edge profile is used as regularization along with a dual discriminator-generator network for deblurring. Restoration of large saturated pixel areas using bidirectional training of GAN for achieving cycle-consistent learning is used. The saturation correction using deep learning have never been addressed in the literature. Novel discriminator-generator paired training strategies are introduced for both deblurring and saturation removal networks. In addition, a novel color-transfer technique is introduced to handle shifts in color profiles during saturation correction. Complete restoration based on global contextual regularization scheme is used to restore pixels associated to debris, bubbles, and other miscellaneous clutter.

1.3. Outline

The remainder of this article is organized as follows. Section 2 introduces the endoscopy dataset for artifact detection, segmentation, frame restoration, and quality assessment. In Section 3, the details of the proposed approaches for artifact detection and endoscopic video frame restoration are presented. Section 4 provides the experiments and results for each step of the framework showing the efficacy of individual method. This section also illustrates the importance of endoscopy frame restoration for robustness of image analysis methods. Finally, Section 5 concludes the paper and outline directions for future work.

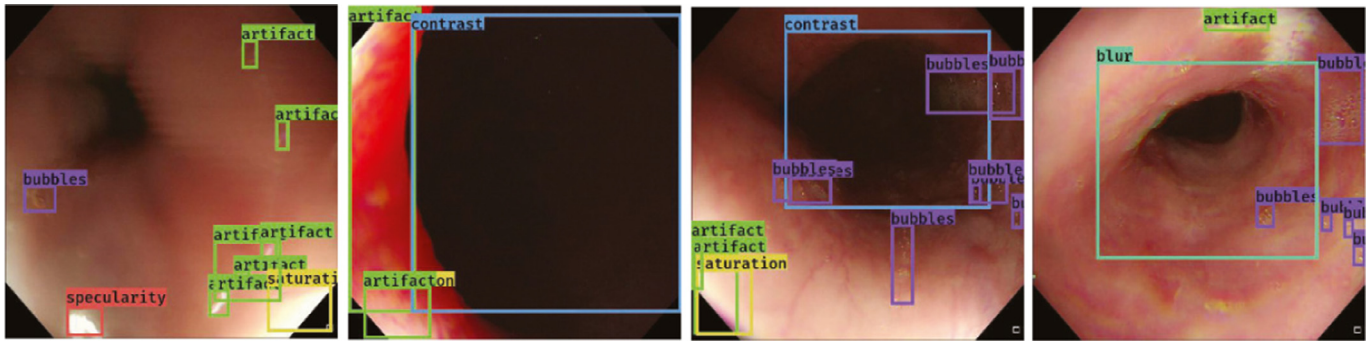

Fig. 5. Examples of detected bounding boxes for some artifact class labels using YOLOv3-spp.

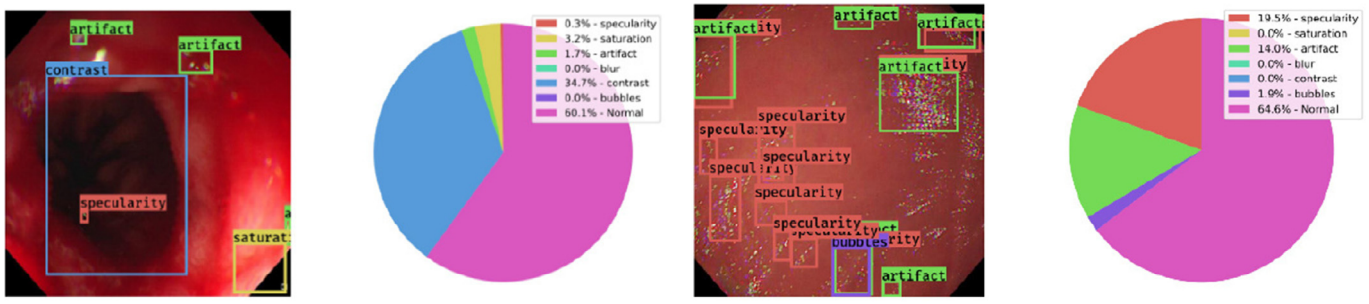

Fig. 6. Quality assessment based on class weight, area and location. Images with detection boxes and their corresponding area fraction are shown. On left: shows image with mostly contrast problem and on right: shows that with multiple misc. artifacts and specularities. Below are their calculated quality scores.

2. Materials

2.1. Artifact detection

Table 1 defines the multi-class artifacts in gastroesophageal endoscopy images and Fig. 1(a) show the samples corresponding to each artifact class that has been annotated. Both normal bright field (BF) and narrow-band imaging (NBI) modalities in the artifact detection dataset are shown. It can observed that even though artifacts appear similar in both modalities the tissue appearance varies drastically.

The artifact detection data set consists of a total of 1290 endoscopy images (resized to 512×512 pixels) from two operating modalities; normal bright field (BF, 1229 images), and narrow-band imaging (NBI, 61 images) from 7 unique patient videos sampled manually from 200 videos (based on artifact types, modality, and texture variability) for training data. The selection was based on number of representative artifacts present in these videos and texture variability of the underlying esophagus. Two experts annotated a total of 6504 artifacts using bounding boxes where each annotation is classified as in Table 1 and corresponding artifact class sample is provided in Fig. 1. A 90%-10% random split was used to construct the train-test set for object detection resulting in 1161 and 129 images (see Fig. 2) and 5860 and 644 bounding boxes, respectively. In general, the training and testing data exhibits the same class distribution (see Fig. 3 (top row)) and similar bounding boxes (roughly square) but either small with average widths less than 0.2 or large with widths greater than 0.5 (see Fig. 3 (bottom row)). Multiple annotations are used in case a given region contains multiple artifacts.

We have also included 184 out-of-sample test data from 3 different HD endoscopy videos (sequence of 74, 50 and 60 frames) acquired from different patients. Two endoscopy video sequences consisted of oesophagus while the third consisted of the pyloric region in the stomach. Due to the least number of bounding boxes in for blur class in our previous test data split, algorithm evaluation could have been affected. To mitigate this issue and to test the reliability of the detection methods for blur classes effectively, we chose the first sequence where the majority of samples consisted of blur (65/74). The dataset consisted of in total 1479 boxes for specularity class, 74 saturation labels, 111 blur, 178 contrast and 618 bubble labels. It is to be noted that large size artifacts (e.g., saturation, blur, contrast) have lower number of samples as they can be labeled by a single large bounding box (see Fig. 1(a)).

2.2. Artifact segmentation

From Fig. 1(b) it is evident that some artifact classes such as specularity, saturation and misc. artifacts appear as an indefinable irregular shapes that might not be sufficiently described by bounding boxes. Similarly, bubbles can accumulate together to form similar irregular shapes. In these circumstances, segmentation of these artifacts are more desirable. We curated 431 training images curated from more than 20 different patient videos and 110 frames were annotated for test dataset from another 10 different patient videos. All frames used in this work were mostly acquired from HD Olympus endoscopes and some included that from capsule endoscopy. The number of annotated artifact class regions for specularity, misc. artifact, bubbles and saturation are 311, 315, 210 and 247, respectively, for training, and 78, 85, 44 and 67 respectively for the test data. It is to be noted that for segmentation task, usually a large region is segmented for small locally scattered specularities or misc. artifacts (see Fig. 1(b)).

2.3. Frame restoration

For frame restoration task, the same 7 patient videos used in the artifact detection task was used to create training datasets

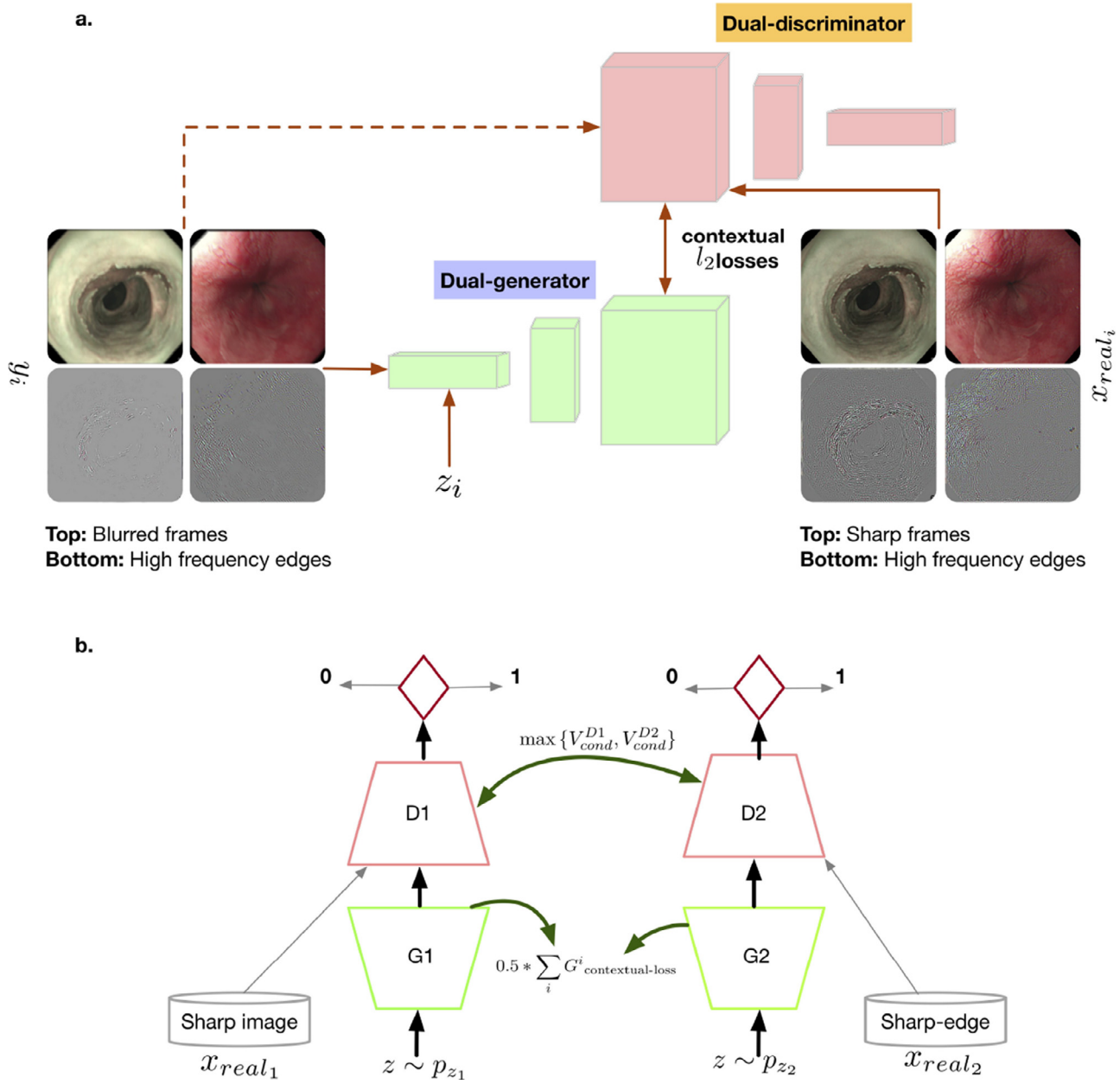

Fig. 7. Blind deblurring using CGAN with added contexual high-frequency feature loss. Dual discriminators D1, D2 and generators G1, G2 are used to leverage the information content for improved accuracy. Here, maximum loss of discriminator obtained from either D1 or D2 is used while an average of the contextual loss which minimises the sharp image (x_{real_1}) and sharp edge (x_{real_2}) w.r.t. generated deblur outputs.

for each frame restoration tasks. Due to difference in the nature of each restoration task the number of samples created for each module varied, for e.g., for deblurring task 15 simulated sequences were created to mimic the hand motion of endoscopists from selected set of images. To make the evaluation consistent, each task was however evaluated on an additional set of 100 high quality sampled images ($QS > 0.95$, see Section 3.4). *Blind deblurring* A paired blur-sharp dataset consisting of 10,710 (715 unique sharp images) multi-patient and multi-modal images with 15 different simulated motion trajectories for blur (Kupyn et al., 2017) were used for training and 5 different motion simulations were used to create a test dataset with 100 samples. *Saturation removal* Due to lack of any ground truth data for two different illumination conditions, a fused dataset was created that included: 200 natural scene image pairs containing diffuse (scattered light) and ambient (additional illumination to natural light giving regions with pixel saturation) illuminations [1]; and 200 endoscopic image pairs were generated using cycleGAN-based style transfer (Zhu et al., 2017)[2]. Saturated frames were generated using the same trained cycleGAN model for 100 samples in the evaluation set. *Specularity and other misc. artifacts removal* 1000 high quality endoscopy video frames ($QS > 0.95$) were used as the 'clean' images (see Section 3.4) for which randomly selected regions were cropped to train the inpainting network. Nearly, 20% of these images were also used as validation set during the training. All 100 images in the evaluation set were randomly cropped for 26×26 and 62×62 size patches (unknown pixels) to measure the efficacy of the inpainting module.

[1] https://engineering.purdue.edu/RVL/Database/specularity_database/index.html.

[2] The cycleGAN network was trained separately using a training data with 200 saturated image samples and 600 normal endoscopy images.

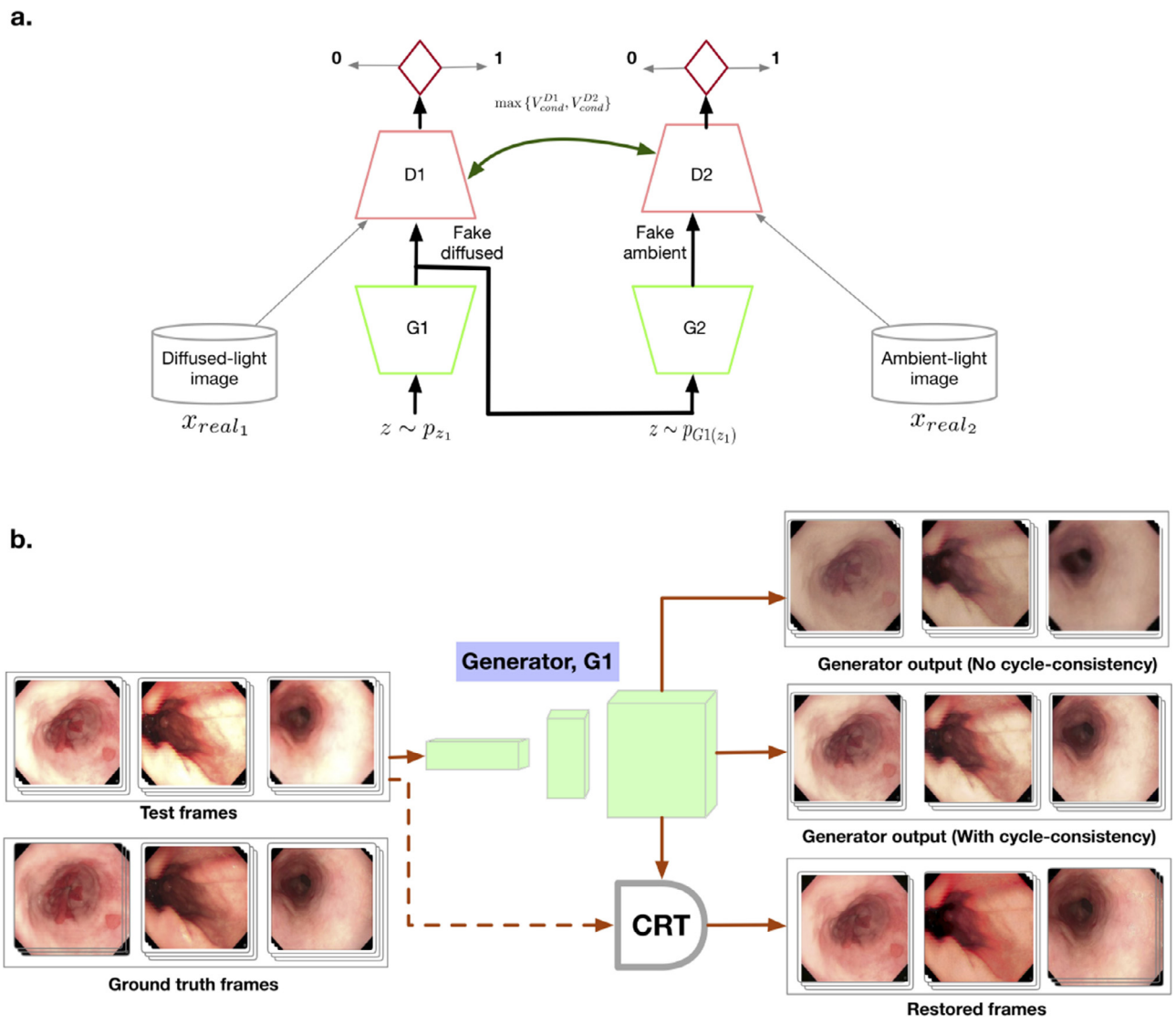

Fig. 8. Cycle-consistent saturation removal is proposed with additional CRT color correction. The generator G1 is trained to generate diffused light image from ambient light image z_1 (*that is usually the cause of saturation of image pixels*) while generator G2 is used to predict ambient light image for the generated diffused light conditioned image $G1(z_1)$. This is done to avoid large deviations ((see (b) right, middle-row) from the natural pixel color in contrast to one-directional generator (see (b) right top-row). Saturation corrected frames generation by the color transfer (bottom right) provides visually comparable result w.r.t. ground truth (bottom left). Heavily saturated frames (see top-left) are used to illustrate the efficacy of the saturation reduction method.

2.4. Video recovery, clinical relevance test and impact on algorithm robustness

The complete pipeline was evaluated with both artifact detection and restoration frameworks on another set of 10 gastroesophageal videos each comprising of more than 10,000 frames for analysis of video recovery. This allowed to quantify the efficacy of video recovery by the proposed framework.

In order to quantify addition of distortion or loss of clinically significant information, restored images of 20 partially corrupted frames were rated independently by two clinical experts. A short video sequence (≈ 50 frames) corresponding to each frame was provided to the clinical experts to assist them with the information regarding of the underlying mucosal surface.

Robustness test of classical image analysis methods such as feature matching and optical flow estimation, before (presence of artifact) and after restoration (removal of artifact) was conducted on 100 paired frames.

3. Method

3.1. Proposed framework

The main aim of this work is to recognize a range of different artifacts and tailor the image restoration accordingly. It is possible that that a single frame can be corrupted by multiple artifacts. Fig. 4 provides an overview of the process and illustrates how detection links with the artifact specific restoration methods. More importantly, in case of corruption of frames with multiple classes of artifact types, a sequential process identified in this work is required for realistic restoration and to avoid further corruption of frames.

Multiple instance object detection is used to discriminate between the six different types of artifacts (see Section 2.1) and normal appearance. For each frame a quality score (QS, refer Section 3.4) is computed based on the type, area and location of the identified artifacts to reflect the feasibility of complete image

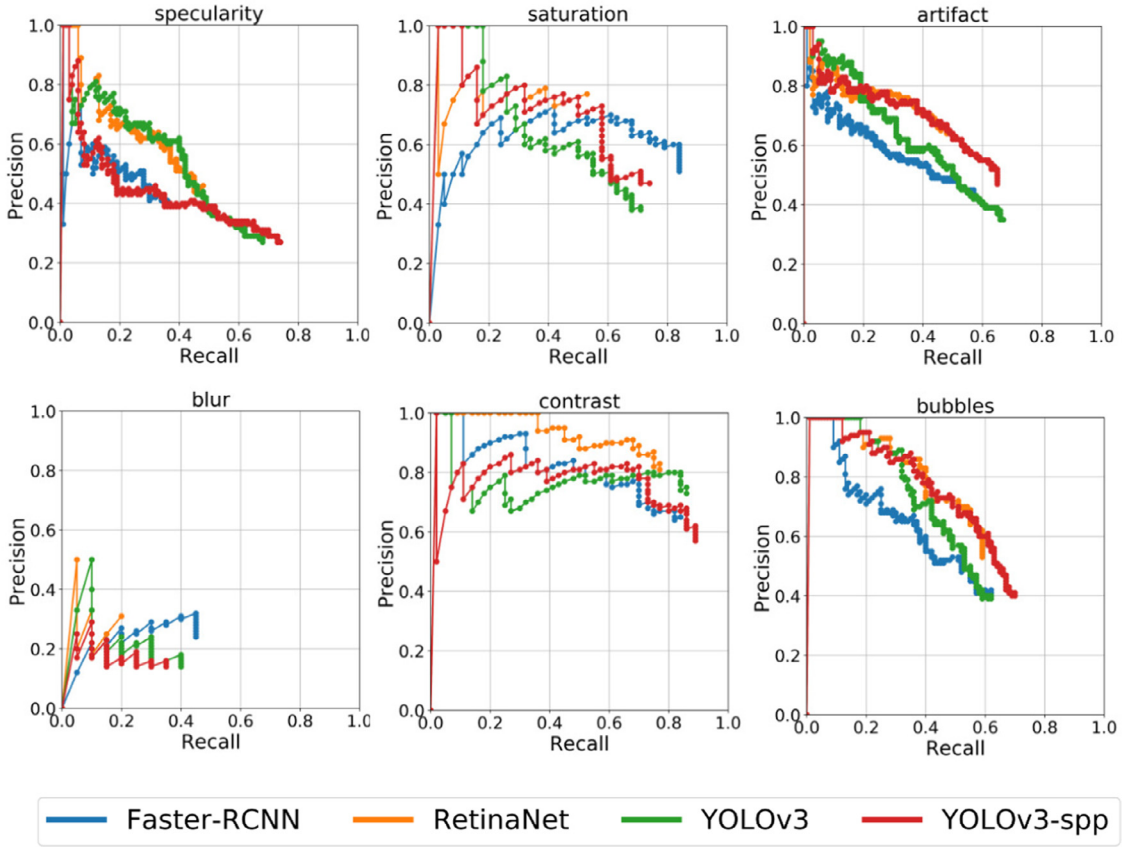

Fig. 9. Class specific precision-recall curves for artifact detection.

restoration via the sequential restoration process depicted in Fig. 4. The scaling of the proposed QS score is set such that it differentiate between severely corrupted frames ($QS < 0.5$), partially corrupted frames ($0.5 \leq QS \leq 0.95$), and frames of high quality ($QS > 0.95$). Severely corrupted frames are discarded without any further processing. The proposed image restoration methods are applied to partially corrupted frames only. In order to guarantee a faithful restoration, partially corrupted frames go through the proposed sequential framework. All remaining frames are directly concatenated into the final list without any processing.

3.2. Artifact region detection

Recent research in computer vision provides object detectors that are both robust and suitable for real-time applications. Here, this paper proposes to use a multi-scale deep object detection and localisation model for identifying the different artifacts in near real-time.

Faster R-CNN (Ren et al., 2015), RetinaNet (Lin et al., 2017b) and YOLOv3 (Redmon and Farhadi, 2018) architectures for artifact detection are investigated. Validated open source codes are available for all of these architectures. Experimentally, YOLOv3 with spatial pyramid pooling (YOLOv3-spp) is chosen for robust detection and improved inference time for endoscopic multi-class artifact detection. Spatial pyramid pooling allowed to pool features from sub-image regions utilizing computed single-stage CNN features at four scales from YOLOv3 architecture. In addition to the boost in the inference speed, incorporating spatial pyramid pooling decreased false positive detections compared to classical YOLOv3 method (see Section 4.2). YOLOv3-spp provided an excellent feature for accuracy-speed trade-off compared to Faster R-CNN and RetinaNet. This is critical for usage in clinical settings. However, Faster R-CNN and RetinaNet were also trained for comparison without any architecture changes. Examples of the detected boxes using YOLOv3-spp are shown in Fig. 5.

3.3. Artifact region segmentation

Experimentally, we found that encoder-decoder architecture together with atrous spatial pyramid pooling of DeepLabv3+ can effectively learn to distinguish the varied shapes of artifacts better than the other state-of-the-art methods. In addition, the 1D separable convolutions of the network improves the inference time that is vital in our video processing pipeline. To obtain a trade-off between speed and accuracy we have used a shallow ResNet50 backbone architecture. We also used widely used U-Net based architecture with ResNet34 and PSPNet with ResNet50 backbone to compare the chosen DeepLabv3+ architecture. Here, choice of backbones are based on our experimental results.

3.4. Quality score

A frame quality score (QS) is proposed based on: a) type, b) area and c) location of the detected artifacts. Weights are assigned to each of these categories and a mean weight is computed as the quality score. Weights are assigned to each type based on the ease of restoration and visual disruption, e.g., an entire blurred image can still be restored but the same would not apply with misc. artifacts, similarly artifacts located centrally are much visually unpleasant to the endoscopists. Thus, misc. artifacts are assigned a higher weight than blur. Similarly, the area and location of detected artifacts in each frame are important. A centrally located imaging artifact with large area detrimentally degrades image information beyond restoration. Below weighting scheme is

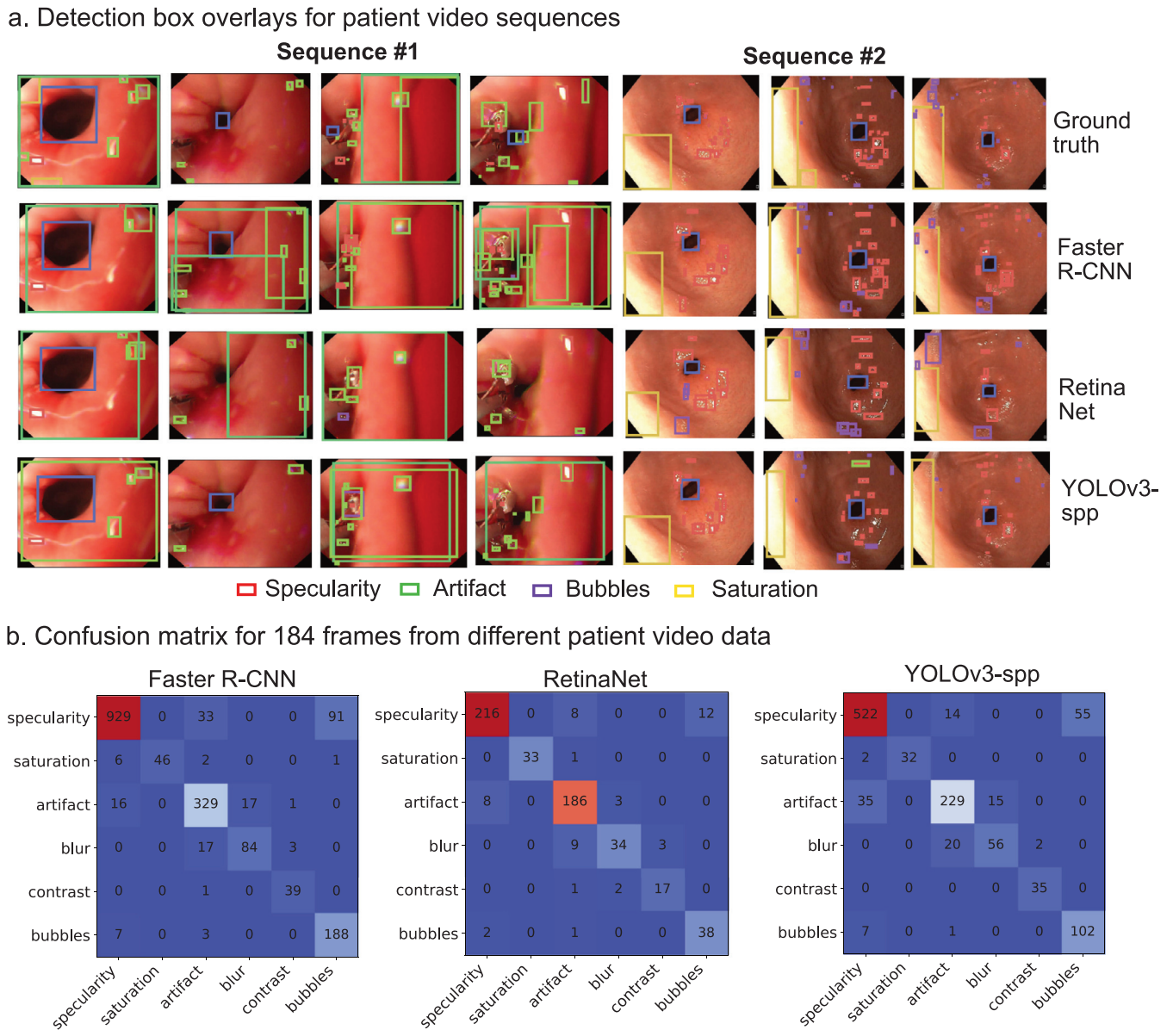

Fig. 10. Artifact detection on video sequence data. a) Overlayed bounding boxes for two different video sequences (each frame taken approximately after 10 frames). b) Confusion matrix showing the true positive detections and confused classes in all 184 sequence frames.

described which is based on three main factors: 1) impact of location on hindrance to visual information, 2) computational difficulty in removal of each artifact class type, and 3) % of area covered.

- Class weight (W_C): misc. artifact (0.50), specularity (0.20), saturation (0.10), blur (0.05), contrast (0.05), bubbles (0.10)
- Area weight (W_A): percentage of the total image area occupied by all detected artifact areas and normal areas
- Location weight (W_L): center (0.5), left (0.25), right (0.25), top (0.25), bottom (0.25), top-left (0.125), top-right (0.125), bottom-left (0.125), bottom-right (0.125).

The presented weights have been heuristically chosen after consultation with expert endoscopists and generalized from their provided quality scores. Such an approach reduces the effort of retraining any machine learning model with new sets of training data and the obtained QS-values will not depend on expert annotation each time minimizing the risk of faulty scoring. In addition, manual scoring do not take into account the physical and computational complexities that have been incorporated in the presented weighting scheme. The final QS is computed as:

$$QS = \lfloor 1 - \sum_{\mathcal{B}} (\lambda_A W_C W_A + \lambda_L W_C W_L) \rfloor_0, \quad (1)$$

where $\mathcal{B}$ denotes the set of bounding boxes associated to each detected artifact, λ_A, λ_L are constants that weight the relative contributions of area and location. $\lambda_A = 0.5, \lambda_L = 0.5$ has been used in the experiments. However, for frames with few detected artifacts (less than 5) such weighting scheme underscores (especially if large area artifacts are present) thus $\lambda_A = 1, \lambda_L = 1$ is used for these cases. Note that QS score in Eq. (1) is lower-bounded by 0.

Examples of the proposed quality score applied to real data are shown in Fig. 6. The video frame in Fig. 6(left) has mostly a contrast problem (i.e., low W_C) so despite its central location (see blue box) and large area the frame intensity can be restored ($\therefore$QS = 0.75). However Fig. 6(right) has many misc. artifacts (high W_C) and specular areas located centrally centrally (c.f. green, red boxes, $\therefore$ QS=0.23) which inhibits realistic frame restoration so the frame is discarded. Although we have used rectangular bounding

Table 2
Computational models used for individual artifact classes.

Artifact type	Restoration method
Motion blur	CGAN (dual) + $l2$-contextual + high-frequency losses
Specularity/Bubbles/ Misc. artifacts	CGAN + $l1$-contextual loss
Saturation	CGAN + $l2$-contextual loss (bi-direction) + CRT transform
Low contrast	same as saturation (reversed training set)

box detection areas only, pixel-wise segmentation mask areas can be directly used for indefinable classes for more precise quality score analysis.

3.5. Image restoration

Formulating the reconstruction of the true signal given the noisy and corrupted input image I as an optimization or estimation problem demands a well-motivated mathematical model. Unfortunately, the various different types of artifacts induce a level of complexity that make this endeavor very challenging. Assuming image noise to be additive and approximating motion blur as a linear convolution with an unknown kernel is reasonable and in line with previous attempts to the problem. In addition, contrast and pixel saturation problems can be formulated as a non-linear gamma correction. Other remaining artifacts (e.g., specularities, bubbles and imaging artifacts) which are due to combined processes of these phenomena can be assumed as a function of the entire process. The corrupted noisy video frame can thus be approximated as:

$$I(t) = F[(h * f(t) + \eta)^{\gamma}], \tag{2}$$

where η denotes the additive noise induced by the imaging system, the convolution with h the approximation to the induced motion blur, γ captures the over- and under-exposed regions and F is a generalized non-linear function that models capturing other artifacts as well (including specularities, bubbles and imaging artifacts) or a combination of them. This model motivates why the restoration of the video frames is structured into separate processing steps, which are implemented as deep learning models.

For multi-class endoscopic artifact restoration, it is required to perform 1) frame deblurring when $h(.)$ is unknown, i.e. a blind deblurring task, 2) minimise the effect of contract imbalance (correction for over- and under-exposed regions) in frames, i.e. γ correction, 3) replace specular pixels and those with imaging artifacts or debris with inpainting, i.e. correction for additive noise $\eta(.)$ or a combined non-linear function $F(.)$. Due to the higher likelihood of the presence of multiple artifacts in a single frame, unordered restoration of these artifacts can further annihilate frame quality. Thus, an adaptive sequential restoration process that account for the nature of individual artifact types is proposed (see Fig. 4). In this work, artifact class dependent contextual losses have been embedded in generative models (see Table 2) for effective restoration. The sequential restoration pipeline (see Fig. 4) is applied to partially corrupted frames. To ensure the stability of the generator output for realistic image restoration, the presented work exploits conditional generative adversarial framework (CGAN) and complimented with artifact specific contextual weighted losses and novel discriminator-generator training strategies.

3.5.1. Motion blur

Unlike static images, motion blur is often non-uniform with unknown kernels $h(.)$ (see Eq. (2)) in video frame data. Let generator G 'generates' a sample $G(z)$ from a prior blurred noise distribution ($p_{noise}(z)$ with $z \sim h(.)$) while a separate discriminator network tries to distinguish between the real target images ($p_{data}(x)$ with assumed $x \sim$ deblurred real image) and the fake image generated by the generator and y be the class label, then the objective function V_{cond} for CGAN can be written as:

$$\min_G \max_D V_{cond}(D,G) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x|y)] + \mathbb{E}_{y\sim p_y, z\sim p_z}[\log(1 - D(G(z|y), y))] \tag{3}$$

In this work, CGAN with a $l2$-contextual loss with 4th layer of VGG16 (squared difference between generated and target/sharp image applied on VGG16 inference) and an additional $l2$ high-frequency loss as regularization on the same. The high-frequency data are also generated by another complementary generator model (see Fig. 7)). As discriminator converges rapidly in this deblurring framework, the discriminator loss is constrained by the maximum discriminator loss at each step, $\max\{V^{D1}_{cond}, V^{D2}_{cond}\}$, where $V^{D}_{cond} = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x|y)]$. In the proposed dual-generator strategy, both generators $G1$ and $G2$ are trained independently. However, their contextual losses are averaged. This is motivated by the fact that motion blur primarily affects image edges, a few discriminative image pixels compared to the entire image. The high-frequency images are first computed both for blurred and sharp images in the training data using iterative low pass-high pass filtering at 4 different scales (Buades et al., 2011). These images are then used to provide additional information to the discriminator regarding the generator's behavior (also see Fig. 7). Eq. (3) thus becomes:

$$\min_{G1} \max_{D1} V'_{cond}(D1,G1) = V_{cond} + \sum_i \lambda \parallel VGG16(x_{real_i}) - VGG16(Gk(z_i|y_i) \parallel_l, \tag{4}$$

where $i = [0,1]$, refer to an original and high-frequency image pair respectively, $k = [1,2]$ refer to generator, $\lambda = 0.5$ is the averaging weight, and $l = 2$ is the norm used. x_{real} is the ground truth image for restoration (i.e. sharp images in this case). Minimization of Eq. (3) using Jensen-Shannon (JS) divergence as in (Goodfellow et al., 2014) can lead to problems like mode collapse, vanishing gradients. Consequently, Arjovsky et al. (2017) proposed to use Wasserstein distance with gradient penalty (WGAN-GP). Thus, CGAN with a critic network is based on WGAN-GP (Kupyn et al., 2017) and an added $l2$ high-frequency regularizer as in Eq. (4). The proposed model was trained for 300 epochs on a paired blur-sharp dataset (refer Section 2.3).

3.5.2. Saturation or low contrast

The variable distances between the light source and the imaged tissue can lead to large illumination changes which can result in saturation or low contrast problems. This motivates the role of the variable γ in Eq. (2). Saturated or low contrast image pixels often occur across large image areas and affect the entire image globally. In addition, these illumination changes are more prominently observed in normal brightfield (BF) modality compared to other modalities. Compensation of affected image pixels is a difficult problem depending on the size of the affected image area.

The saturation restoration task is posed as an image-to-image translation problem and the same end-to-end CGAN approach used for motion deblur is used but unlike previous technique here the generated output of $G1$ is fed to the second generator $G2$ (i.e., see Fig. 8). Thus, a cycle-consistent approach ("bi-direction") is taken, where the first $D1G1$ network is trained to predict diffused light image from a domain of ambient light and vice-versa for $D2G2$ network. The generator output is minimized as the weighted average in Eq. (4). Here also, described l_2 contextual loss is used to train a generator-discriminator network for saturation reduction. l_2 contextual loss is more suitable as it allows to capture the deviation between normal illumination condition w.r.t. saturation and

Table 3
Artifact detection results on test set with different neural network architectures. All run times are reported on a single 6GB NVIDIA GTX Titan Black GPU and is the average time for a single 512x512 image (possibly rescaled on input as indicated) evaluated over all 129 test images. Total number of ground truth boxes = 644 boxes. IoU scores are averaged for both predicted and non-predicted labels. Best scores are in bold.

Method	Backbone	Input Size	mAP_{25}	mAP_{50}	IoU_{25}	IoU_{50}	Overall mAP	Overall IoU	Predict Boxes	t (ms)
Faster R-CNN	ResNet50	600^2	40.4	29.5	28.33	24.50	27.24	22.75	835	555[a]
RetinaNet	ResNet50	608^2	41.2	34.7	**38.87**	**35.61**	29.64	**32.38**	**576**	103[b]
YOLOv3	darknet53	512^2	44.3	**35.1**	24.20	22.14	29.75	19.63	1252	95[c]
YOLOv3	darknet53	608^2	45.2	33.2	21.40	18.98	29.65	16.80	1300	116[4]
YOLOv3-spp	darknet53	512^2	**45.7**	34.7	24.43	22.60	**30.63**	20.01	1120	**88**

[a] Python Keras 2.0, (Tensorflow 1.2 backend) Code.
[b] PyTorch 0.4 Code.
[c] Python call of Darknet trained network.

Table 4
Class-specific average precision (AP) at IoU thresholds 0.25 and 0.5 of different object detection networks (two best scores are highlighted in bold).

Method	Specularity		Saturation		Artifact		Blur		Contrast		Bubbles	
	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
Faster R-CNN	20.7	8.41	**71.00**	38.55	35.10	20.93	**14.50**	**9.42**	58.70	70.99	42.40	28.84
RetinaNet	33.10	20.95	42.90	38.21	39.8	27.40	7.20	4.17	**73.60**	**73.56**	**50.60**	**43.63**
YOLOv3	**40.00**	**22.38**	50.40	**45.62**	**44.30**	**30.53**	**11.60**	**8.25**	70.80	**64.56**	48.90	39.07
YOLOv3-spp	**34.70**	**21.20**	**55.70**	**38.85**	**48.00**	**32.97**	7.50	5.70	**72.10**	61.42	**55.90**	**48.13**

Table 5
Artifact detection results on out-of-sample sequence test set with different neural network architectures. IoU scores are averaged for both predicted and non-predicted labels. Total number of ground truth boxes = 2460 boxes. Best scores are in bold. Overall **AP** for contrast (64.70), blur (50.42) and saturation (39.84) were highest for YOLOv3-spp.

Method	Backbone	Input Size	mAP_{25}	mAP_{50}	IoU_{25}	IoU_{50}	Overall mAP	Overall IoU	Predict Boxes
Faster R-CNN	ResNet50	600^2	**52.16**	**30.91**	35.19	29.30	**31.20**	27.82	3617
RetinaNet	ResNet50	608^2	30.02	18.05	37.93	30.48	17.67	**28.36**	1167
YOLOv3-spp	darknet53	512^2	40.20	28.32	**45.50**	**39.05**	26.51	**36.51**	**1872**

low contrast conditions. A single direction network (only $D1G1$) was also trained to quantify the efficacy of the proposed strategy.

To correct coloration shift due to the incorporation of natural images in the training set, color transfer (CRT) is applied to the generated frames. Given a source image, I_s and a target image, I_t to recolor, the mean (μ_s, μ_t) and covariance matrix $(\mathbf{\Sigma}_s, \mathbf{\Sigma}_t)$ of the respective pixel values (in RGB channels) can be matched through a linear transformation (Hertzmann, 2001):

$$I_t' = \mathbf{\Sigma}_s^{1/2}\mathbf{\Sigma}_t^{-1/2}(I_t - \mu_t) + \mu_s, \tag{5}$$

where I_t' is the recolored output. To avoid re-transfer of color from saturated pixel areas in the source, the mean and covariance matrix are computed from image intensities <90% of the maximum intensity value. Fig. 8 shows the generated results using the trained GAN-based network (on the top right) and after color shift correction (bottom right) showing very close to ground-truth (bottom left). It can also be observed that bidirectional approach (with cycle-consistency) does not have large shifts in color imbalance while reducing optimally over saturation problem compared to one direction approach (no cycle-consistency).

3.5.3. Specularity, and other misc. artifacts removal

Illumination inconsistencies and view point changes cause strong bright spots due to reflections from bubbles and shiny organ surfaces. Water-like substances can also create multi-colored chromatic artifacts (referred to as 'imaging or mixed artifact' in this paper). These inconsistencies appear as a combination of linear (e.g., additive noise η) and non-linear noise (function $F(.)$) in Eq. (2). Presence of specularities, and other misc. artifacts is posed as missing pixel problem and CGAN (see Eq. (4)) based restoration is applied with added l_1−discounted contextual (reconstruction) loss using a distance-based weight mask as by Yu et al. (2018) along with edge-aware loss. Both contextual and generative losses are used in the implemented model along with discontinuity preserving term (edge-aware) required for avoiding over-smoothing during missing pixel generation. A bottleneck approach is used to retrain the model initialized with the pretrained weights of the places2 data set (Zhou et al., 2018). During training and validation masks of different patch sizes $\{(5\times5), (7\times,7), (11\times11), (13\times13), ..., (33\times33)\}$ were randomly generated and were used for restoration. A single image can have one or multiple generated masks for restoration.

4. Results and discussion

In this Section, evaluation metrics used for quantitative assessment of multi-class artifact detection and restoration are first introduced. Utilizing these metrics artifact detection, quality score estimation, and different steps of frame restoration pipeline are quantified. Efficacy of the proposed framework is then analyzed on 10 long endoscopic videos. A clinical relevance test is performed to provide a visual quantification from clinical experts on the visual quality of the frame restoration. An experimental evidence test is finally presented to illustrate the significance of video frame quality improvement required for guaranteeing robustness of image analysis methods.

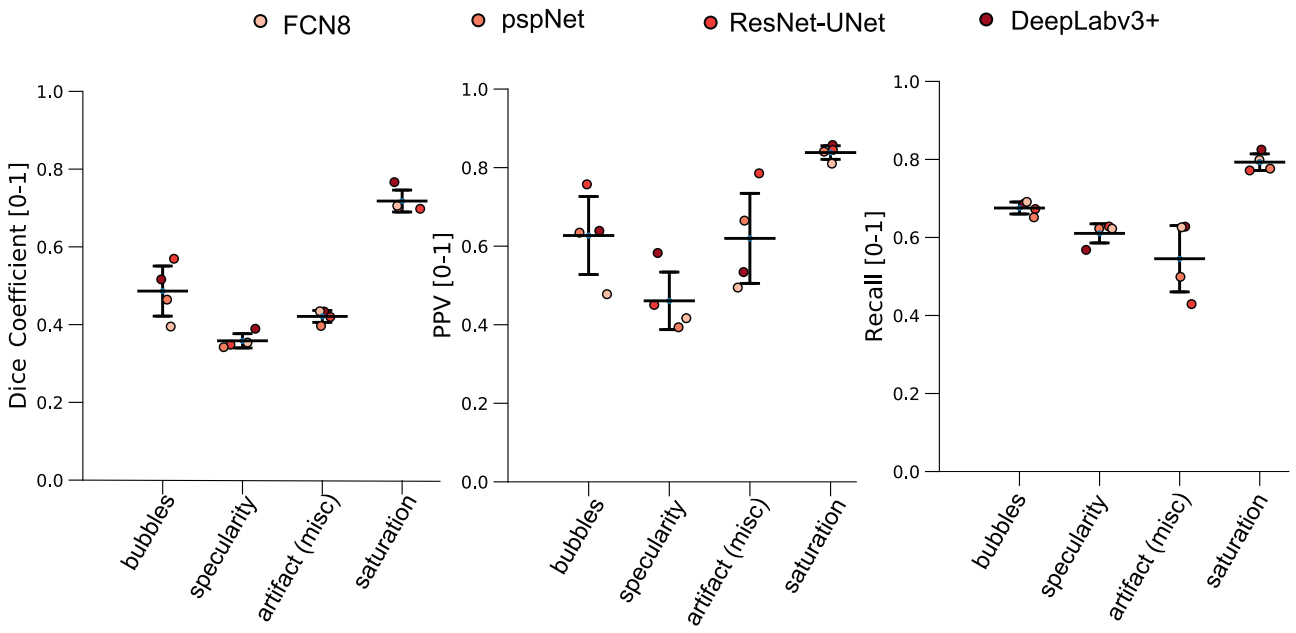

Fig. 11. Dice coefficient, precision (PPV) and recall for each method for specific artifact class (ordered according to their increasing mean size). Smaller deviations in specularity and misc. artifacts in DSC values denote that most methods have similar performances on these two classes. However, for bubbles and saturation DeepLabv3+ and pspNet performs relatively better in DSC than FCN8 and ResNet-UNet. PPV and recall plots show similar trend for bubbles and saturation.

Table 6
Segmentation of artifact instances with irregular shapes. Mean values across four artifact classes (bubbles, specularity, misc. artifact, and saturation) are presented. All methods used imageNet pretrained bacbone network and simple data augmentation (includes flip, rotation, and scaling). Training was obtained using 90-10% split for 500 epochs with learning rate of 0.01. Top segmentation metric values are in bold.

Method	Backbone	JC	DSC	F2	PPV	Rec	**Overall Acc**
FCN8	VGG16	0.4058± 0.1388	0.4723 ± 0.1377	0.4816 ±0.1294	0.5498 ±0.1531	**0.6847** ±0.0714	0.9687
resnet-Unet	ResNet50	0.4527±0.1475	0.5088 ±0.1353	0.5065±0.1323	**0.7096** ±0.1529	0.6254±0.1247	0.9703
pspNet	ResNet50	0.419 ±0.1436	0.4761 ±0.1372	0.4781±0.1274	0.6334 ±0.1593	0.6375 ±0.0985	0.9653
Deeplabv3+	ResNet50	**0.4649**±0.1529	**0.5265** ±0.1459	**0.5262** ±0.139	0.6533 ±0.1236	0.6768 ±0.0953	**0.9755**

Table 7
Correlation coefficients and disagreement scores from paired analysis 100 sample images. Spearman rank and Kendall tau are provided as a measure of correlation with their corresponding p-value to measure the significance of these correlations. A disagreement mean score is provided which is computed as the mean value of the difference in paired quality score for each score.

Comparison	**Spearman rank correlation [-1, +1]**		**Kenadall tau [-1, +1]**		**Disagreement score**
	Corr. value	**p-value**	**Corr. value**	**p-value**	**Mean score**
Proposed vs Expert #1	0.6369	3.64e-13	0.4741	1.48e-11	2.0452
Proposed vs Expert #2	0.5923	3.50e-11	0.4507	3.36e-10	2.4419
Proposed vs Expert #3	0.5449	2.23e-09	0.3887	3.17e-08	2.3560
Intra Expert QS (mean)	0.6151	-	-	-	1.4743

4.1. Evaluation metrics

To evaluate the artifact detection methods, the standard mean average precision (mAP) and intersection-over-union (IoU) metrics are used. The detection results of all architectures are quantified using the mAP at IoU thresholds for a positive match of 25% and 50% denoted as mAP_{25} and mAP_{50}, respectively. The mean IoU between positive matches, the number of predicted boxes relative to the number of annotated boxes and the average inference time for one image are also used as quantitative measures. We also provide overall mAP computed as an average mAPs for IoU from 0.25 to 0.75 with a step-size of 0.05, i.e., an average over 11 IoU levels are used for 6 artifact classes (mAP @[.25 : .05 : .75]) (Everingham et al., 2012). To evaluate the segmentation methods we have used standard metrics: Jaccard similarity coefficient (JC), dice similarity coefficient (DSC), F2 score, precision or positive predictive value (PPV), recall or sensitivity (Rec) and overall accuracy. Quality score estimation is evaluated against three independent expert quality opinions. For the quality assessment of deblurring methods well-known peak signal-to-noise ratio (PSNR), (multi-scale) structural similarity SSIM (Wang et al., 2003) are used. Unlike mean-squared error (MSE), PSNR scales the MSE according to image range. However, it is not always ideal to reveal the improved quality as they entirely depend on intensity range. To overcome the limitations of PSNR for quantification of saturation and specularity restoration tasks, more sophisticated visual information fidelity (Sheikh and Bovik, 2006) and relative edge coherence (Baroncini et al., 2009)) quality assessment metrics which are referred as VIF and RECO,respectively in this paper are included. Statistical analyses are also presented to validate the significance of the proposed methods in the framework.

4.2. Artifact detection

Artifact detection training data consisted of 1161 endoscopy images from 7 unique patient videos which was then split into 70-

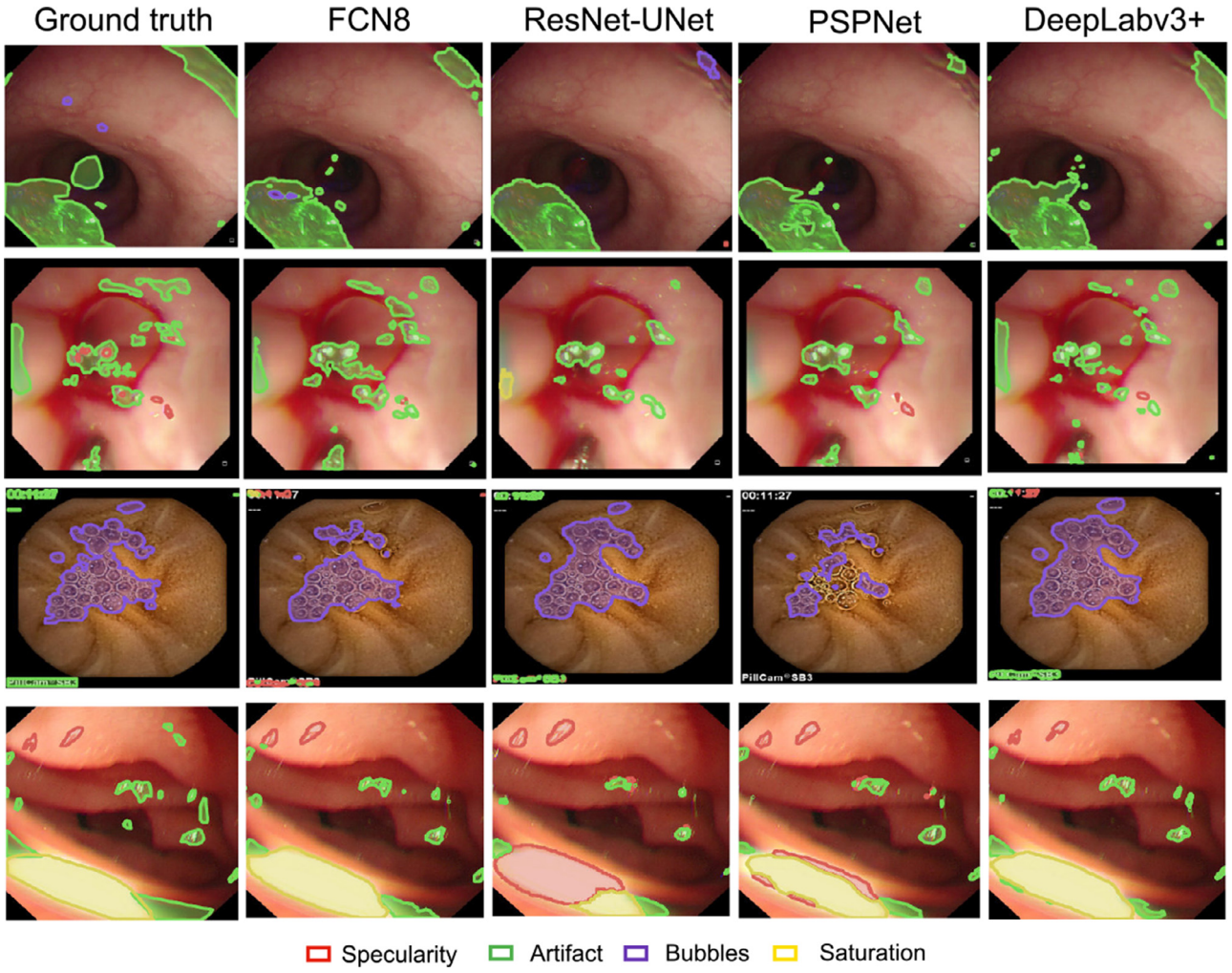

Fig. 12. Qualitative results for multi-class artifact segmentation for irregular shaped artifact instances.

Table 8
Peak signal-to-noise ratio (PSNR) and the structural similarity measure (SSIM) for randomly selected 10 images with different motion blur. The mean values for 100 blurred images is also presented μ_{100}. Corresponding stratistical significance is presented in Fig. 14.

Method	Metrics	Images with varying motion blur										
		#80	#99	#102	#113	#116	#510	#652	#10163	#10450	#1135	μ_{100}
Proposed	PSNR	**25.22**	**28.14**	**27.28**	**23.41**	**24.81**	**25.67**	**30.30**	**24.37**	**27.00**	**27.21**	**25.72**
	SSIM	**0.998**	**0.997**	**0.993**	**0.980**	**0.992**	**0.900**	**0.974**	**0.904**	**0.935**	**0.938**	**0.913**
deblur GAN	PSNR	25.17	27.93	26.96	23.40	**24.81**	23.91	25.37	22.65	25.04	25.34	24.25
	SSIM	**0.998**	**0.997**	0.992	0.979	**0.992**	0.886	0.962	0.894	0.921	0.915	0.911
deblurNet	PSNR	24.61	27.50	25.02	22.23	22.00	23.05	24.92	21.41	24.97	24.58	23.67
	SSIM	0.995	0.996	0.990	0.970	0.970	0.881	0.964	0.895	0.916	0.911	0.894
TV-deconv	PSNR	24.25	26.72	24.75	21.69	22.20	23.97	25.37	22.65	25.33	25.40	24.01
	SSIM	0.966	0.994	0.988	0.966	0.983	0.890	0.965	0.870	0.896	0.912	0.892

30% respectively for as train and validation set during training and 129 images were used for test. Data augmentation with flips (left, right, up and down), changes in HSV values ($\{\pm 20, \pm 60\}$), rotation $\{\pm 5^{\cdot}, \pm 20^{\cdot}\}$ changes were applied randomly to each frame during training. A thorough description of dataset is provided in Section 2.1.

Table 3 shows that YOLOv3 variants outperform both Faster R-CNN and Retinanet. YOLOv3-spp (proposed) yields the best mAP of 45.7 at IoU thresholds 0.25 and overall mAP of 30.63 at a detection speed $\approx 6\times$ faster than Faster R-CNN (Ren et al., 2015). Even though Retinanet exhibits the best mean IoU of 38.87 at threshold of 0.25, it is to be noted that IoU is sensitive to annotator variances in bounding box annotation which might not resemble the performance of detectors. YOLOv3-spp provides a good balance between mAP-IoU tradeoff at threshold of 0.25. For this, it can be observed that mAP_{25} is 4% higher than RetinaNet and 5% higher than Faster R-CNN. It is to be noted that $\text{mAP}_{25} > 45.0$ with $\text{IoU}_{25} > 63.0$ (when only detected boxes are considered) obtained by the proposed YOLOv3-spp can be considered as acceptable high accuracy for both detection and localization in computer vision. Considering the complexity of endoscopy multi-class artifacts due to their huge appearance variability this score is acceptable for clinical usage.

In terms of class-specific performance, proposed YOLOv3-spp is the best across detecting misc. artifacts and bubbles (both are predominantly present in endoscopic videos) with average precision

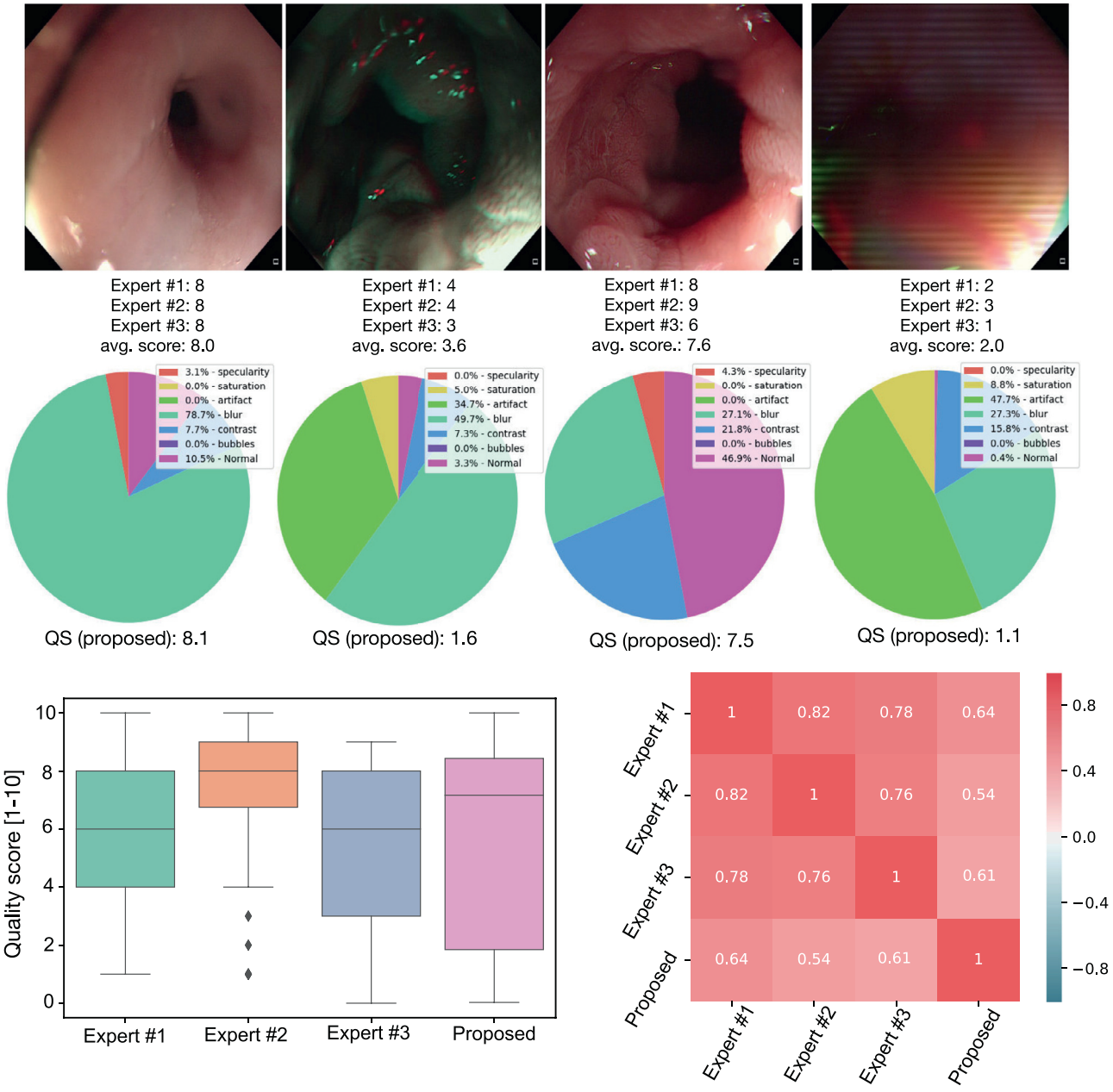

Fig. 13. Quality score evaluation. a) Quality scores corresponding to three experts and an average score between them (top), and quality score determined by the proposed scheme for the detected artifact regions and classes from YOLO-v3 spp detector (bottom). b) Quality scores with median (line) and lower- and upper-quartiles represented with the box plot (left), and the heat-map showing correlation of quality score between different expert raters and the proposed scheme (right).

Table 9
Evaluation (average values) metrics (PSNR, SSIM, VIF and RECO) are used to assess the restoration quality for 100 saturated images. Here, *l*2-contexual CGAN trained in one-direction (onedir., single discriminator and single generator) and bi-direction (bidir., two discriminators and two generators), and final proposed method (with color retransfer as post-processing applied on bidirection method) are presented. Two best results are highlighted in bold.

Metrics	Simulation	*l*2-**contexual CGAN**		Proposed
		onedir.	bidir.	
$\overline{PSNR}$	18.18	17.82	**20.92**	**22.34**
$\overline{SSIM}$	0.843	0.760	**0.854**	**0.952**
$\overline{VIF}$	**0.413**	0.212	0.294	**0.365**
$\overline{RECO}$	0.830	0.914	**0.988**	**0.950**

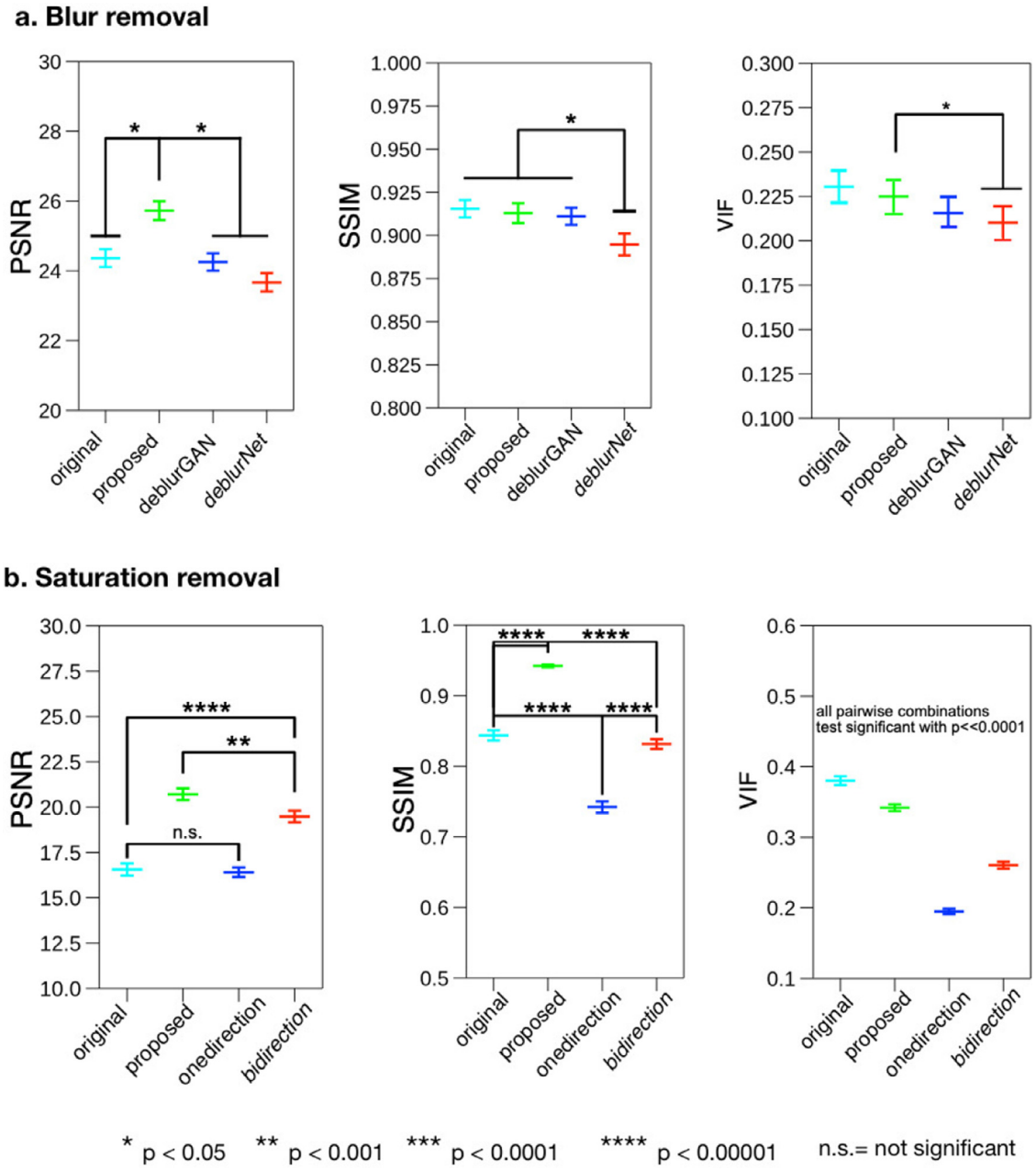

Fig. 14. Statistical analysis on 100 endoscopy video frames. PSNR, SSIM and VIF are used as evaluation metrics and paired t-test is performed to identify the statistical significance of the obtained results. a) Blur removal, and b) saturation removal or correction method. Solid overhead lines join which groups have been compared with stars (see legend) to indicate level of statistical significance.

of 48.0 and 55.9, respectively (see Table. 4, Fig. 9). Faster R-CNN yielded the highest average precision for saturation (71.0) and blur (14.5) while RetinaNet and YOLOv3 outperformed respectively for contrast (73.6) and specularity detection (40.0). It is worth noting that proposed YOLOv3-spp yielded second best average precision scores for specularity (34.7), saturation (55.7) and contrast (72.1). Blur is missed by most networks which is also under-represented in the training dataset (see Fig. 3).

Table 5 presents detection results for the out-of-sample sequence data which consisted of 60% blur samples. On this data, Faster R-CNN obtained the highest mAP_{25} and mAP_{50}, while YOLOv3-spp has the highest IoUs (i.e., better localisation) and second best mAPs. Upon looking at the number of bounding box predictions, Faster RCNN yielded 1157 more bounding boxes. YOLOv3-spp has only 2% lower mAP_{50} but nearly 10% gain in the IoU_{50} than Faster R-CNN. Simiarly, the inference speed is 6× faster than Faster R-CNN. Our per class mAP also revealed that YOLOv3-spp performed the best on contrast, blur and saturation while second best performance was recorded for specularity, bubbles and misc. artifact.

Fig. 10 shows the qualitative results of sequence out-of-sample test data. For sequence #1 in Fig. 10(a), it can be observed that Faster R-CNN detects more bounding boxes compared to ground truth for most frames. However, YOLOv3-spp is able find optimal number of boxes describing the presence of artifact classes compared to other methods. For sequence #2, most methods exhibit similar results and very close to ground truth annotations. Fig. 10(b) presents the confusion matrix for all three methods for each class. It can be observed that specularity class is confused with bubbles and blur is con-

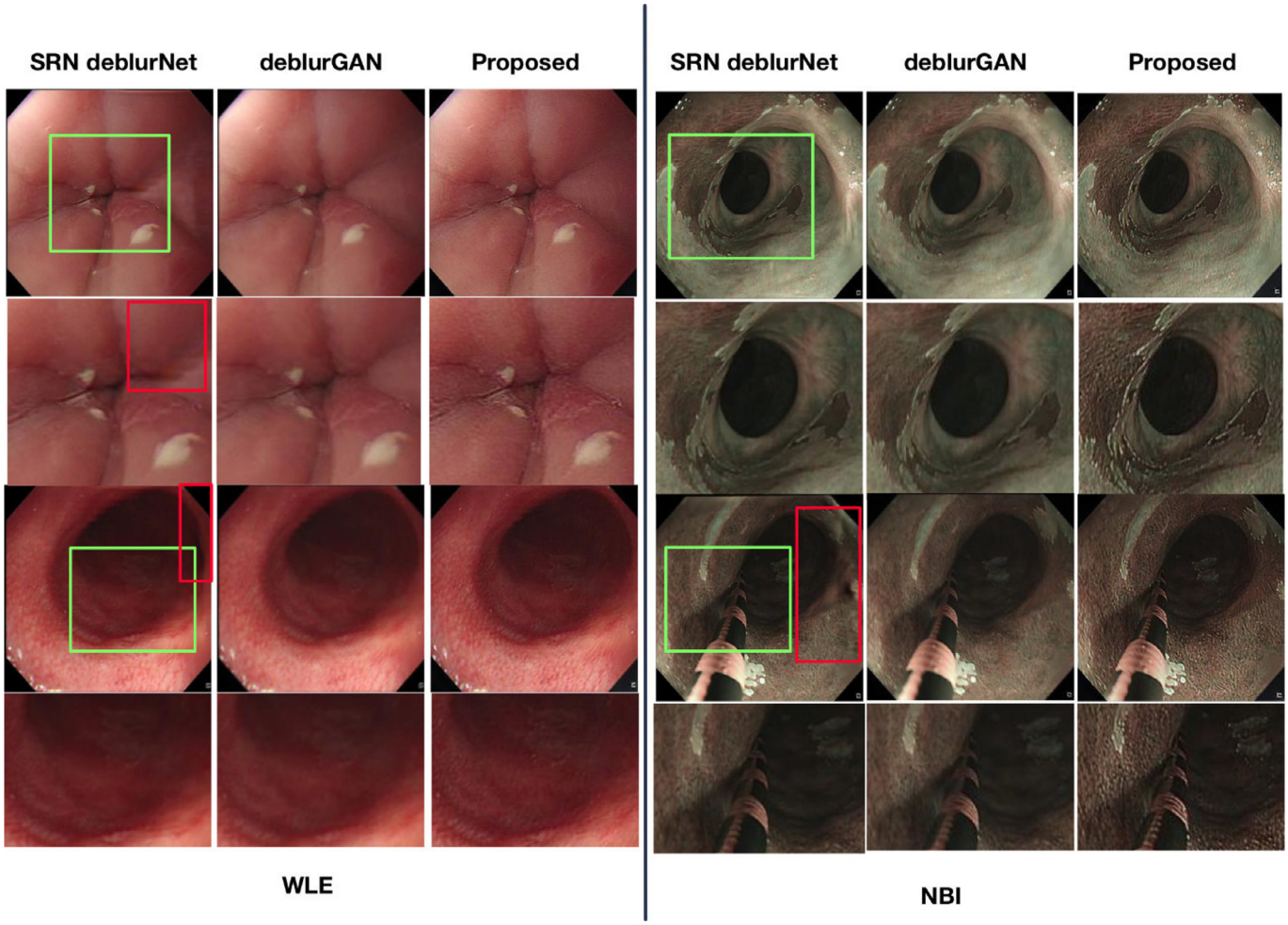

Fig. 15. Qualitative results for different de-blurring methods on WL and NBI frames. Unrealistic reconstrution are marked by red rectangle area while areas with green rectangles are zoomed and presented just below each method. It can be observed that image sharpness is highly improved in the proposed method compared to other state-of-the-art deep learning based methods.

fused with misc. artifact for both Faster R-CNN and YOLOv3-spp while only limited boxes are detected for RetinaNet architecture showing these confusion relatively less. In general, unlike nature scene images, there is not much unclear distinction between artifact classes in images. As a result, there can occur large variability in bounding box annotations which can adversely affect the mAP estimation when taken at different IoU thresholds.

4.3. Artifact segmentation

Table 6 represents mean evaluation metric scores for segmentation of four indefinable artifact class instances that included bubbles, specularity, misc. artifacts, and saturation. It can be observed that DeepLabv3+ with ResNet50 backbone provided the highest DSC, F2 and overall accuracy of 0.5265, 0.5262 and 0.9755 respectively. Similarly, the second best performance was obtained by ResNet-UNet with ResNet34 backbone. While, FCN8 have the least performance scores for almost all the metrics, the number of relevant instances retrieved (Rec) is the highest with 0.6847. A good trade-off between PPV and Rec is desired based on which it can be understood that DeepLabv3+ is competitive in these metrics as well (PPV of 0.6533, Rec of 0.6768).

Fig. 11 represents DSC, PPV and recall for each class for different implemented methods. It can be observed that for most classes DeepLabv3+ achieved the highest DSC value for specularity, misc. artifact and saturation classes. FCN8 has the least PPV for almost all the classes but achieved high recall. However, DeepLabv3+ (bubbles: 0.64, specularity: 0.58, misc. artifact: 0.53, saturation: 0.86) and ResNet-UNet (bubbles: 0.63, specularity: 0.40, misc. artifact: 0.53, saturation: 0.86) have relatively higher PPV for most artifact classes. The inference time for both ResNet-UNet and DeepLabv3+ on RTX 2080 Ti GPU was an average of 0.35 s per image for an image size of 512×512.

Qualitative results for four different video frames in the test dataset is shown in Fig. 12. It can be observed that DeepLabv3+ provided more accurate results than other methods. For ResNet-UNet and PSPNet, saturation confused with specularity. Similarly, some parts of misc. artifact in the the first and second rows were missed by both.

While segmentation is mostly encouraged for irregular shaped artifacts, a combined approach can benefit in improving the delineation of area of interest, for instance, detection information can be used to correct the class instance of segmented pixels. Similarly, the area of bounding box predictions can help to determine the morphological operations required for the improvement of segmented area.

4.4. Quality score

The proposed quality score scheme was evaluated against three expert quality rankings for 100 endoscopy images of the test

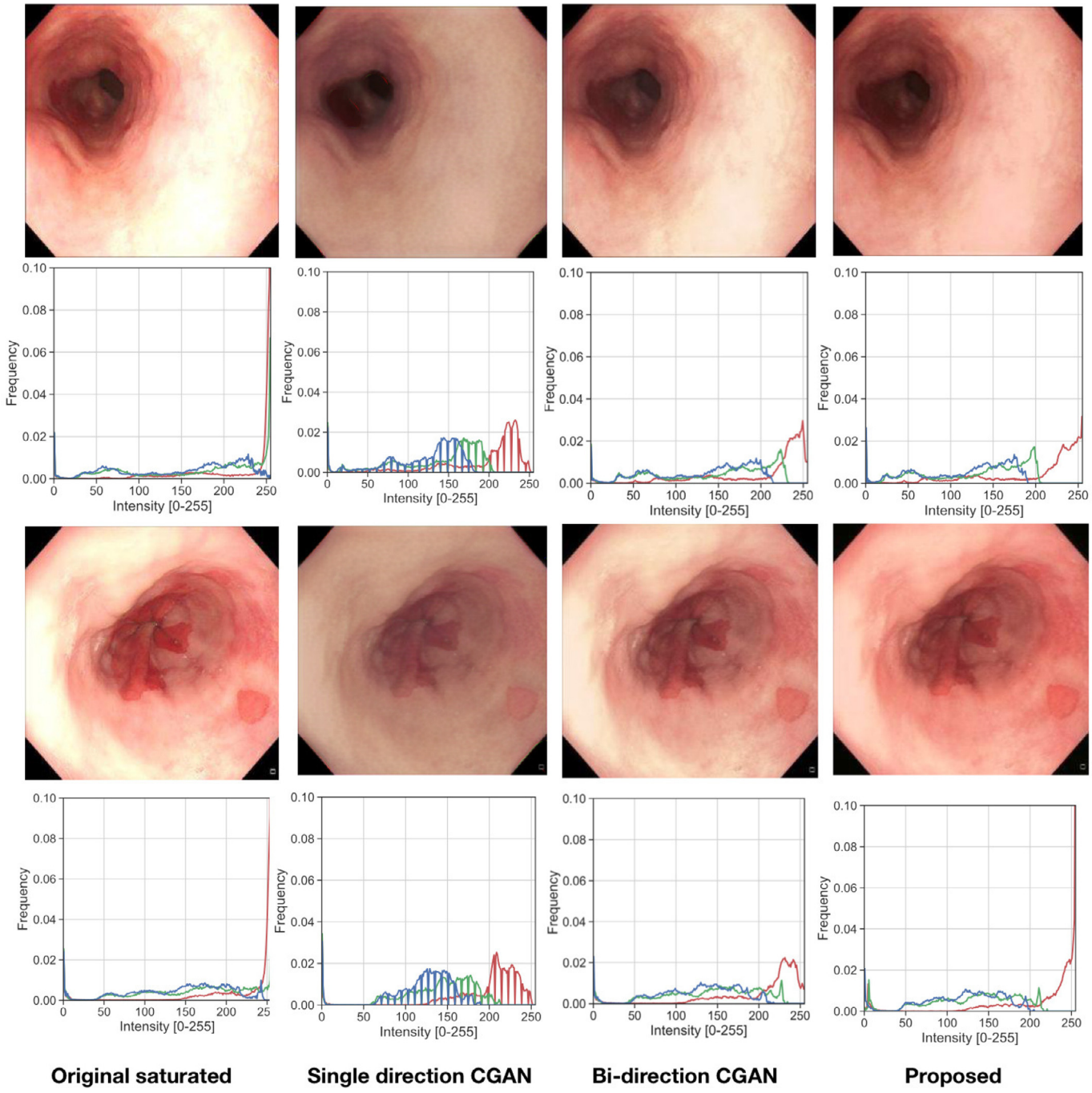

Fig. 16. Saturation correction. 1st column: Saturation of pixels in the region near to the light source (blue arrows mark the saturation regions, left). 2nd– 4th columns represent saturation correction methods one directional, bi.-directional and proposed bi-directional with CRT processing, respectively. For each, corresponding RGB histograms are provided in the consequetive rows.

data used in artifact detection evaluation. It can be observed in Table 7 that the proposed QS scheme has a positive correlation ($>$ 0.54 for Spearman rank correlation) with all three expert scores with very low p-value $<< 10^{-8}$. The Spearman rank correlation coefficient between all three expert ranking is 0.6151 which is very close to that obtained between proposed QS scheme vs the Expert #1 and that with the Expert #2. In addition, the mean disagreement score is also very close to that for between the experts. Fig. 13(a) shows four example images with expert quality scores and the estimated QS from the proposed scheme. Fig. 13(b, left) shows that median line (center middle line in box plot) and the spread of the QS values of the proposed scheme shows its capability to capture the wider range of quality score aligned with Expert #1 and Expert #3. On contrary, Expert #2 quality score suffers from a very narrow range with outlier scores. Also, Fig. 13(b, right) represents a correlation heat-map showing a positive correlation with that of the proposed QS and the experts.

4.5. Frame restoration

4.5.1. Blind deblurring

The proposed (dual) conditional generative adversarial network with added contextual and high-frequency feature losses

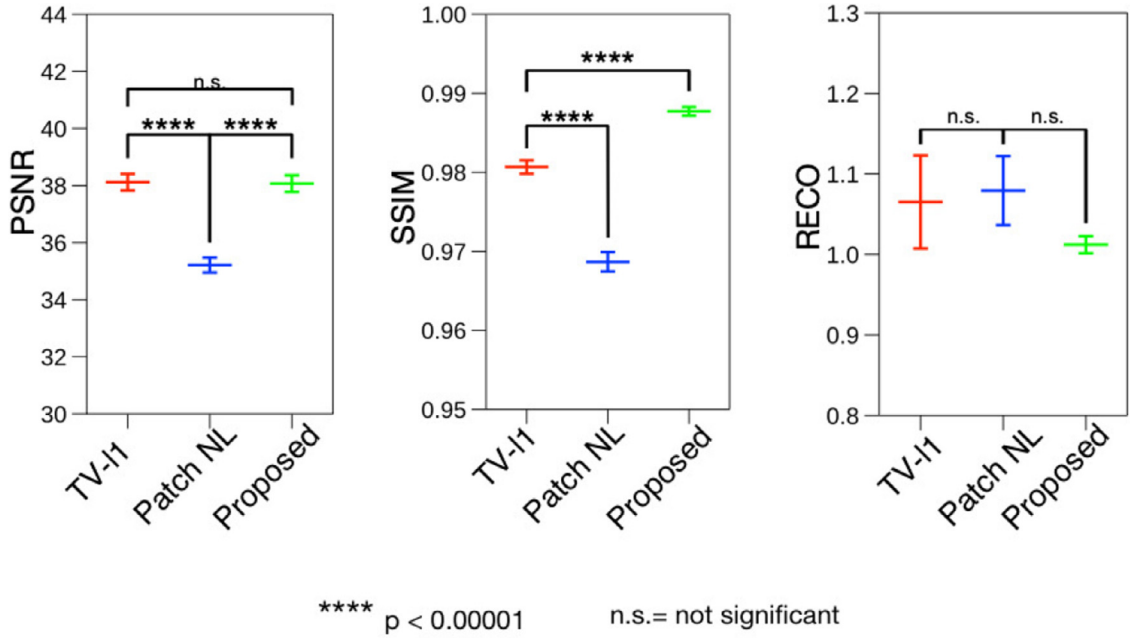

Fig. 17. Statistical analysis on 100 endoscopy video frames. PSNR, SSIM and RECO are used as evaluation metrics and paired t-test is performed to identify the statistical significance of the obtained results for inpainting. Solid overhead lines join which groups have been compared with stars (see legend) to indicate level of statistical significance.

with deblurGAN (Kupyn et al., 2017), scale-recurrent network-based SRN-DeblurNet (Tao et al., 2018), and traditional TV-based method (Getreuer, 2012). TV regularizion weight λ and the blur kernel r affects the quality of recovered deblurred images (Getreuer, 2012). $\lambda = 10^3$ and $r = 2.3$ are chosen after a few iterative parameter setting experiments for the endoscopy data set. Retraining was performed for SRN-DeblurNet (Tao et al., 2018) and deblurGAN (Kupyn et al., 2017) on the same data set.

The quantitative evaluation of the frame deblurring methods using 100 test images with visually large blur. Table 8 shows that proposed (dual) CGAN with $l2$-contextual loss and added high-frequency (HF) feature losses score the highest PSNR and SSIM values for all blurred frames while TV-based deconvolution method (Getreuer, 2012) resulted in the least PSNR and SSIM values over most frames. Moreover, the mean PSNR and SSIM scores are also the highest for 100 test samples compared to all state-of-the-art methods (for PSNR $>$ 1dB compared to deblurGAN and $>$ 2 dB compared to deblurNet). Importantly, significantly large improvements are observed for #652 ($\approx$ 5 dB), #510 and #10163 ($\approx$ 2 dB) compared to competitive deblurGAN method. Qualitative results are presented in Fig. 15, which shows that the proposed dual method leveraging edge information is far more sharper than other approaches. This is illustrated by a zoomed areas across green rectangles. It can be also observed that deblurNet deforms the image at upper right locations in both WL and NBI frames (see red rectangle areas in Fig. 15).

A statistical test is also conducted to show the significance of the proposed network over the state-of-the-art deep learning based deblurring models in Fig. 14(a). It can be observed that the proposed method provides a significant PSNR boost compared to original and other model outputs. The result is consistent for SSIM, however, due to very small change of values the difference is not captured well. The VIF metric also shows an improvement over both state-of-the art methods. In addition, it can be seen that deblurNet achieves worst result which can be due to their hallucinating effect (added distortions) observed in Fig. 15 (see red rectangle regions).

4.5.2. Saturation removal

Quantitative results are provided in Table 9 for 100 simulated test samples from the best selected frames (QS$>$ 0.95). One directional (onedir.) model with a single *DG* network and the proposed bi-directional with two *DG* cycle-consistent networks (bidir.) are compared alongside the final proposed method with added CRT correction. It can be observed that the proposed models with bidirectional network performs better than the one directional network. Further, the proposed CRT post-processing significantly improves the restoration quality. It can be observed that the proposed scheme improves the mean PSNR from 18.18 dB to 22.34 which is more than 4dB improvement. Similarly, a significant boost in the mean SSIM and mean RECO is observed. On contrary, one directional model decreases the image quality which is in line with the observation in the statistical analysis Fig. 14(b). In Fig. 14(b) highly significant improvements for the proposed method compared to both one directional and bi-directional models can be observed with p-values $<<$ 0.001 for most cases. Similarly, a qualitative evaluation presented in Fig. 16 illustrated the visual quality after restoration of the saturated image regions. It can be noted that the bi-directional (proposed) saturation removal technique does not degrade the image (no large color shifts) unlike one directional model. In additional, the proposed CRT post-processing further enhances the saturation restoration.

4.5.3. Specularity and other misc. artifacts removal

Specularity and other local artifacts are removed based on inpainting (Section 3.5.3 for details). To validate the inpainting methods, 100 test images of image size 512×512 were used. Each image was then masked with 10 randomly selected patches of 26×26 and 62×62. CGAN-based model with $l1$-contextual loss model is compared with widely used traditional TV-based and patch-based inpainting methods. It can observed in Table 10 that proposed CGAN-based approach has best score for structure similarity metric SSIM while very consistent PSNR values to that of rigorous TV-$l1$ based inpainting method (Getreuer, 2012). The statistical significance test shows (see Fig. 17) a significant improvement of PSNR and SSIM values over patch-based non-local inpainting

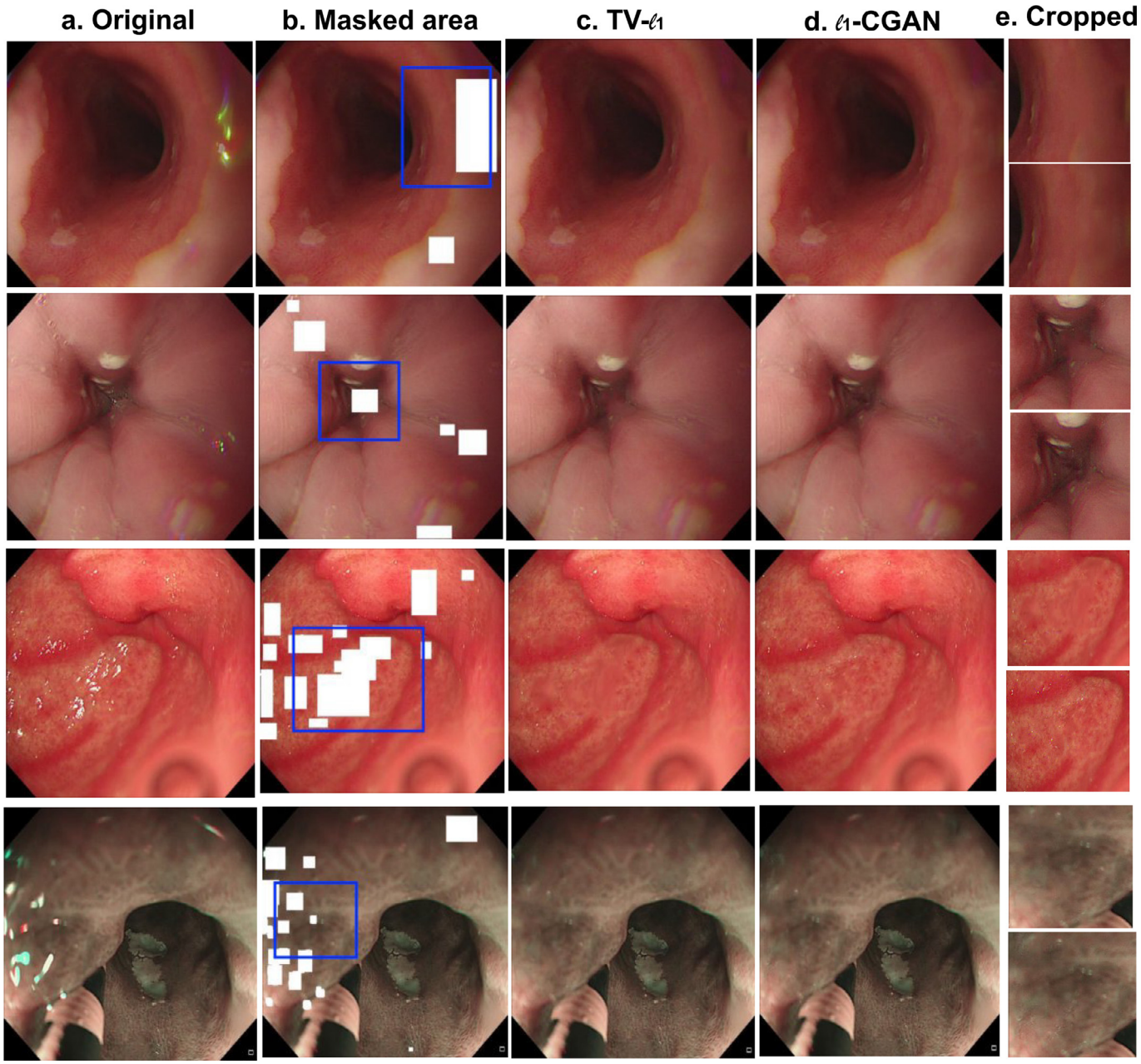

Fig. 18. Image restoration result using inpainting of corrupted areas (specularity, imaging artifacts) detected by the proposed detection method. a) Original corrupted image, b) detected bounding boxes, c) inpainting result using recent TV-based method, d) l1-contexual CGAN, e) top, bottom: Restored area marked with blue rectangle in (b) using TV-based and generative model using l1-contexual CGAN, respectively.

Table 10
Peak signal-to-noise ratio (PSNR) and the structural similarity measure (SSIM) for randomly selected 10 images for inpainting method. The mean values for 100 masked images is also presented μ_{100}. Corresponding stratistical significance is presented in Fig. 17.

Method	Metrics	Image inpainting for combined patches of 26×26 and 62×62 pixels											
		#99	#101	#105	#106	#123	#126	#144	#205	#11439	#11796	μ_{100}	t
Proposed	PSNR	**38.78**	**40.55**	**39.07**	**39.26**	41.70	**38.03**	**39.79**	34.80	**33.92**	**39.98**	38.07	2.5
	SSIM	**0.993**	**0.989**	**0.984**	**0.985**	**0.992**	**0.984**	**0.991**	**0.987**	**0.981**	**0.987**	**0.988**	
TV-l1	PSNR	38.45	38.65	38.75	38.71	**41.85**	36.29	39.40	**35.12**	33.48	40.88	**38.12**	392.0
	SSIM	0.989	0.979	0.980	0.981	**0.988**	0.976	0.983	0.977	0.972	0.988	0.980	
Patch NL	PSNR	37.23	34.58	35.28	35.07	37.31	32.09	36.32	32.65	30.90	37.65	35.21	35.0
	SSIM	0.984	0.963	0.966	0.964	0.973	0.959	0.966	0.964	0.961	0.979	0.968	

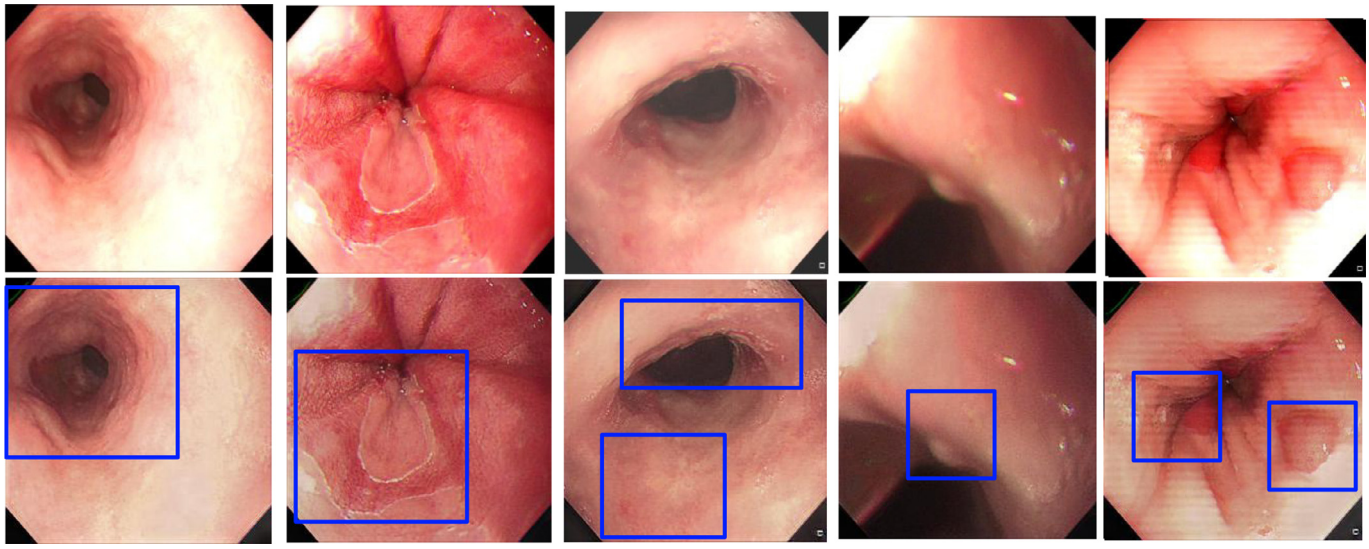

Fig. 19. Simultaneous application of deblurring and bidirectional saturation removal technique on real patient data. Blue rectangles (bottom row) indicate the enahnced local features after deblurring.

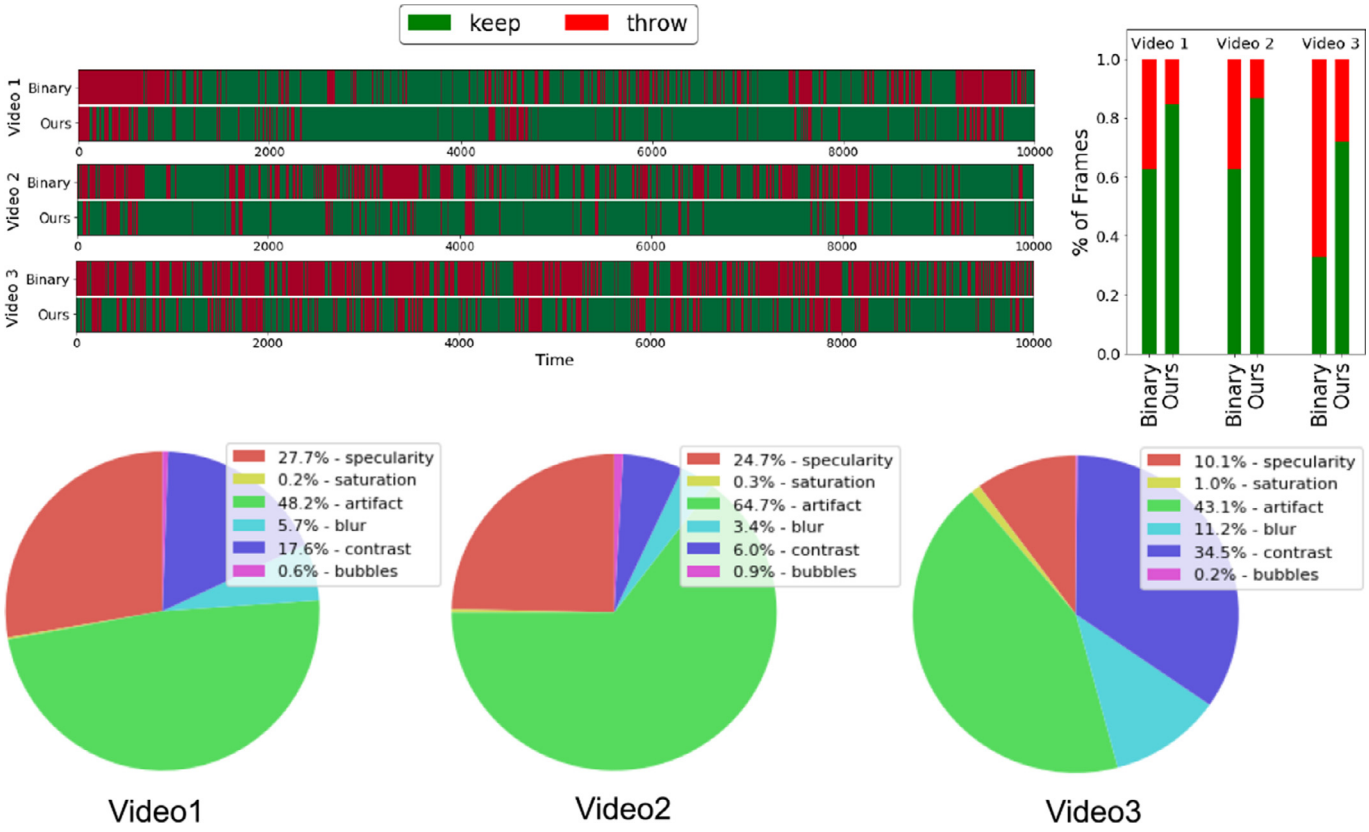

Fig. 20. Frame recovery in clinical endoscopy videos. Top: frames and proportion deemed recoverable over a sequence of and the over a sequence of using a binary deep classifer and the proposed QS score. Bottom: the proportion of each artifact type present in each video.

method (Newson et al., 2017). While no significant improvements were observed in RECO metric. However, it can be observed that the standard deviation is relatively smaller for the CGAN-based approach.

Qualitative results for the proposed specularity and local artifact removal on real problematic gastro-oesophageal endoscopic frames are shown in Fig. 18. In Fig. 18(a), both imaging artifacts (first and fourth rows) and specularities (second and third rows) introduce large deviations in pixel intensities both locally with respect to neighouring pixels and globally with respect to the uncorrupted image appearance. Using inpainting methods (see Fig. 18(c) and (d)), the images have been restored based on the bounding box detections of the proposed artifact detector. The second best TV-based method in Fig. 18(c) produces blurry and non-smooth patches during the reconstruction of unknown pixels (refer to Fig. 18(b)) compared to CGAN generative model (see Fig. 18(d)). A

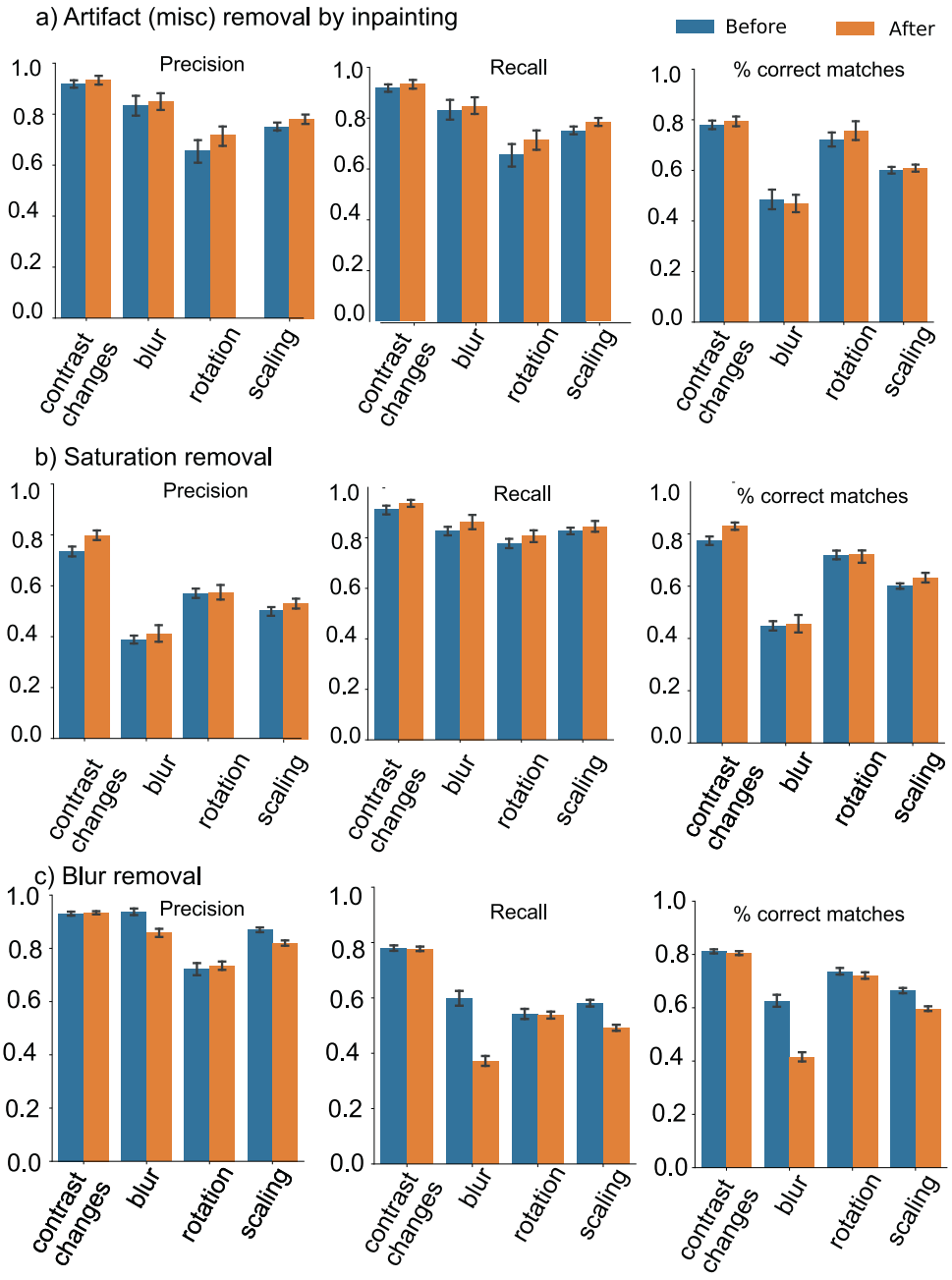

Fig. 21. Invariance tests for feature matching: before and after frame restoration. Different geometric and photometric changes were applied and the mean precision, recall and percentage of correct matches were estimated.

closer look around the unknown regions indicated by blue rectangular boxes in Fig. 18(b), (e) shows that local image structures are well preserved and smoother transition from reconstructed pixels to the surrounding pixels is present. An immediate noticeable ghost effect can be observed in the second row, Fig. 18(e) top using the TV-based method.

4.6. Video recovery and quality assessment

We evaluated the proposed artifact detection and recovery framework on 10 gastroesophageal videos comprising with nearly 10,000 frames each. For artifact detection, an objectness threshold of 0.25 was used to reduce duplication in detected boxes and QS value for restoring the frame was set to ≥ 0.5. As a baseline, a sequential 6-layer convolution neural network is separately trained (layer with 64 filters of sizes $3, 5 \times 5$, ReLU activation function and batch normalization) with a fully connected last layer for binary classification on a set of 6000 manually labeled positive and negative images to decide whether to discard or keep a given input video frame. A threshold of 0.75 was set for the binary classifier to keep only frames of sufficient quality. Our framework successfully retains the vast majority of frames compared to a binary decision, Fig. 20. The quality enhanced video was again fed to the CNN-based binary classifier which resulted in lower number of frame rejection than on raw videos. Consequently, the resultant video is more continuous compared to the equivalent binary cleaned video utilizing raw videos. For example, in video 3, the video after frame removal based on the bi-

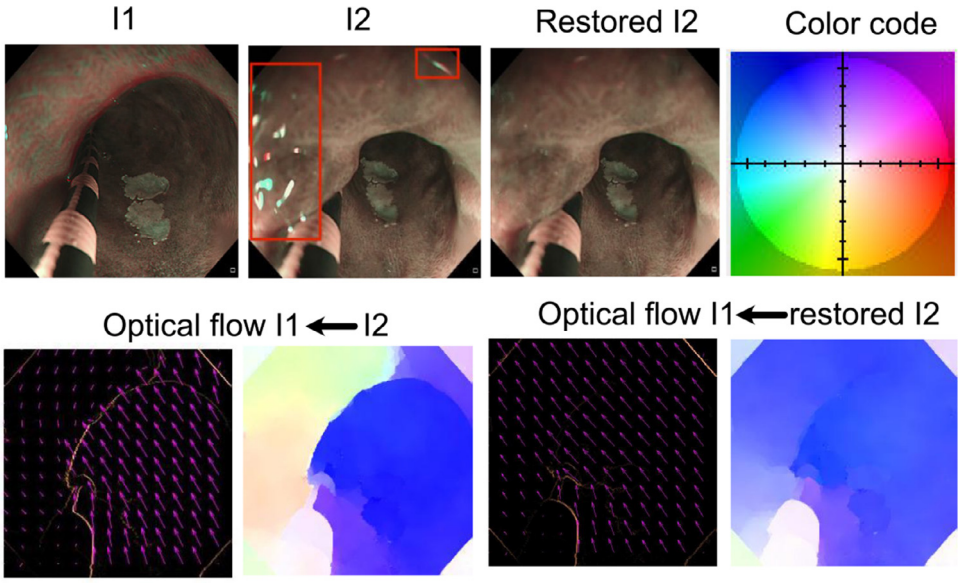

Fig. 22. Optical flow computation (Ali et al., 2016a) between two frames I1 and I2 and I1 and restored I2. Here, frame I2 is corrupted with misc. artifacts (shown with red rectangles). 2nd row (on left) shows the optical flow field computed before the restoration and (on right) after the restoration. The target frame I1 used in the flow field computation is shown in 1st row, left. Both arrowed and flow field color plots are presented for each computed flow-field.

nary classifier directly lead to many distinct abrupt transitions that can be detrimental for post-processing algorithms as only 30% is kept. Comparatively, the proposed framework retains 70% of frames, i.e. a frame restoration of nearly 40%. Quantitatively across all 10 endoscopic videos tested, the framework restored 25% more video frames, retaining on an average of 68.7% of 10 videos considered.

4.7. Clinical relevance test

20 high-quality images with artifacts were selected from 10 test videos that included blur, specularity, saturation and misc. artifacts (refer Section 3.5). Restoration methods were then applied to these images. Two expert endoscopists independently were asked to score these restoration results. Each endoscopist was provided with a partially corrupted frame, restored frame and a corresponding high-quality sequence data depicting mucosal area without artifact. The scores were set-up using the following criteria:

- < 5: restoration added unnatural distortions that can affect clinical outcome
- > 5: restoration succeeded but with clinically insignificant minor distortions

The obtained mean score were blur: 7.87, specularity or misc. artifacts: 7.7, and saturation: 1.5 (one directional plus CRT), 6.5 (bidirectional plus CRT). A remarkable restoration was obtained for blur and specularity or misc. artifacts. However, saturation correction was not pleasant for one directional method to clinicians which was mostly due to loss of 3D information (according to feedback comments) even though visual coherence was improved. However, the bidirectional approach plus CRT was rated much higher with no such remark.

Visual quality improvement on real patient data where sequential deblurring and saturation removal methods were applied is shown in Fig. 19. It can be observed that while saturation removal diminished the intense local and global effect of saturation in the frames, deblurring allowed to improve sharpness in local structures (see blue rectangle areas).

4.8. Significance of frame restoration for algorithmic robustness

While it is intuitively clear that the various artifacts impact any downstream processing, this section presents a concrete experimental evidence on how artifacts impact feature matching as well as optical flow based methods. To demonstrate the significance of frame restoration, 100 partially corrupted frames (with blur, saturation and specularity) are used to assess the algorithm robustness test. *Feature extraction and matching (invariance tests)*

While, most feature based state-of-the-art methods (Bay et al., 2008; Lowe, 2004) are invariant to illumination changes, these methods can be affected in presence of occlusions due to artifacts or extreme saturation conditions in evident in endoscopic video frames. We computed mean precision, mean recall and percentage of correct feature matching for each pair applying geometric (36 rotation angles with 10° spacing)) and photometric (brightness changes, $-127 : 10 : 127$ in gray pixel values) transformations (Ali et al., 2018). The experiment also includes 8 scaling ($[0.25, 2]$) and 9 Gaussian blur ($[1, 9]$). SURF (Bay et al., 2008) feature extraction was used and a brute force feature matching technique was applied. Our experiments (see Fig. 21) showed that while there were almost no change in the overall mean precision and recall for restored images after blind deblurring, saturation removal provided a boost in precision and recall by more than 2-5% compared to original saturated images. Similarly, a small rise in precision and recall were observed after removal of artifacts by inpainting as well. Even though, deblurring resulted in decreased precision and recall for blur invariance test and scaling invariance test, it has almost no affect on contrast and rotation changes.

Optical flow estimation Fig. 22 demonstrates that the artifact removal is critical for estimating pixel-wise flow field widely used for motion estimation. Here, optical flow with an illumination invariant optical flow (OF, (Ali et al., 2016a)) is first computed between a pair of clean image I1 and image with artifact I2 and then the same method is used to compute flow field using the restored image of I2 and the clean image I1. It can observed that the flow field computation is affected by the presence of imaging artifacts (Fig. 22, 2nd row) while on contrary an accurate and smooth flow field is estimated using restored image (Fig. 22, 2nd row, right). An improvement in both divergence ($\nabla \cdot \mathbf{u}$) and curl ($\nabla \times \mathbf{u}$) of the

vector field was noted with standard deviation drop by 0.160 and 0.017, respectively.

In order to quantify further the effect of restoration on optical flow estimation, we used 100 frames on which known (randomly generated) homographies were applied on both corrupted and restored frames which was used as ground truth flow field. We used rotation ($\theta = [1, 5]$), translation ($t_x = [0, 15],\ t_y = [0, 15]$), scale ($s = [0.99, 1.0]$), shear ($S_x = S_y = [0.001, 0.010]$).

TV-l1 primal dual approach (Zach et al., 2007) implementation (Bradski, 2000) was applied to compute the optical flow field. Average angular error (AAE) and average end-point error (AEPE) were estimated for quantification. For saturation, we obtained nearly 2° improvement in AAE and 0.355 improvement for AEPE. Similarly, for deblurring and inpainting approaches we obtained AAE improvement of nearly 1° and over 0.50 improvement in AEPE.

5. Conclusion

The need to improve the diagnostic quality in endoscopy motivates the need for a systematic approach to artifact detection, segmentation, and video restoration. The proposed novel end-to-end framework is capable of identifying a range of different artifacts and provides a context-based approach to frame restoration. The experimental results demonstrates that the presented approaches meet the challenging demands of the real-world application setting. Since all of the modules of the proposed framework are formulated as neural networks, it is possible to achieve near real-time performance when taking full advantage of modern GPU architectures.

Critical to the quality of restored frames are the edge-based (high-frequency) loss for recovering blurred images in a dual *DG* network architecture with contextual *l*2 losses, and a cycle-consistent *DG* network with a color re-transfer scheme to deal with color shifts in generated frames for saturation correction. The highest mAP_{25} with the modulated YOLOv3-spp and the least inference time (88 ms) was achieved for near real time frame quality scoring. Similarly, a good trade-off between predicted boxes, mAP score and the highest IoU at 0.5 IoU threshold was achieved by YOLOv3-spp on out-of-sample patient video sequence data. Pixel-wise segmentation is also investigated for indefinable classes that included specularity, misc. artifacts, bubbles and saturation. A frame quality scoring based on predicted class, area, and location in image was proposed which showed a good correlation with expert ratings.

The quantitative and qualitative improvements for frame restoration tasks showed notable improvements in both PSNR and SSIM metrics for blur and saturation using the proposed models are significant. For specularity and other misc. artifacts removal, improvements on similarity metrics was also achieved compared to non-local inpainting technique with notably nearly 20× faster inference time and more than 150× faster than TV-*l*1 image inpainting method. Importantly, the proposed complete framework was able to restore an average of 25% of the video frames in 10 randomly selected videos from the database. It is worth noting that for 3 videos used for illustration of the importance of the proposed framework, 40% of frames which otherwise would be discarded for downstream analysis were rescued. The work also illustrates that the frame restoration in video endoscopy is vital for accuracy and robustness of image analysis methods. A clinical relevance test has been conducted and results indicate that the proposed restoration techniques do not introduce or remove clinically relevant information. Thus, high quality performance on real clinical endoscopy videos for both intra- and inter-patient variabilities and multimodality has been demonstrated in this work. Future work will focus on further improving the artifact detection and implementing the entire framework as a single end-to-end trainable neural network.

Declaration of Competing Interest

The authors have no conflicts of interest.

CRediT authorship contribution statement

Sharib Ali: Conceptualization, Methodology, Investigation, Data curation, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Felix Zhou:** Data curation, Investigation, Software, Formal analysis, Writing - original draft. **Adam Bailey:** Conceptualization, Resources, Data curation. **Barbara Braden:** Conceptualization, Resources, Data curation. **James E. East:** Conceptualization, Resources, Data curation. **Xin Lu:** Funding acquisition. **Jens Rittscher:** Conceptualization, Funding acquisition, Writing - review & editing.

Acknowledgments

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. S. Ali is supported by the NIHR Oxford BRC. BB, AB, JE and XL are partly funded by NIHR Oxford BRC. F. Zhou and X. Lu are supported by the Ludwig Institute for Cancer Research. J. Rittscher is supported through the EPSRC funded Seebibyte programme (EP/M013774/1) and is adjunct professor of the Ludwig Oxford Branch.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2020.101900.

References

Abebe, M., Booth, A., Kervec, J., Pouli, T., Larabi, M., 2018. Towards an automatic correction of over-exposure in photographs: application to tone mapping. Comput. Vis. Image Underst. 168, 3–20.

Akbari, M., Mohrekesh, M., Najariani, K., Karimi, N., Samavi, S., Soroushmehr, S.M.R., 2018. Adaptive specular reflection detection and inpainting in colonoscopy video frames. In: International Conference on Image Processing, pp. 3134–3138.

Ali, S., Axer, M., Amunts, K., Eils, R., Rohr, K., 2018. Evaluating local features in high-resolution 3D-PLI data. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 729–733.

Ali, S., Daul, C., Galbrun, E., Blondel, W., 2016. Illumination invariant optical flow using neighborhood descriptors. Comput. Vis. Image Underst. 145, 95–110. doi:10.1016/j.cviu.2015.12.003.

Ali, S., Daul, C., Galbrun, E., Guillemin, F., Blondel, W., 2016. Anisotropic motion estimation on edge preserving Riesz wavelets for robust video mosaicing. Patt. Recog. 51, 425–442. doi:10.1016/j.patcog.2015.09.021.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Öksüz, I., Ning, Q., Yang, S., Khan, M.A., Gao, X., Realdon, S., Loshchenov, M., Schnabel, J.A., East, J., Wagnieres, G., Loschenov, V., Grisan, E., Daul, C., Blondel, W., Rittscher, J., 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. 10, 2748.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223.

Bang, D., Shim, H., 2018. Improved training of generative adversarial networks using representative features. In: Proceedings of the 35th International Conference on Machine Learning, ICML 2018. In: Proceedings of Machine Learning Research, vol. 80, pp. 442–451.

Barcelos, C.A.Z., Batista, M.A., 2007. Image restoration using digital inpainting and noise removal. Image Vis. Comput. 25 (1), 61–69. doi:10.1016/j.imavis.2005.12.008.

Baroncini, V., Capodiferro, L., Claudio, E.D.D., Jacovitti, G., 2009. The polar edge coherence: a quasi blind metric for video quality assessment. In: European Signal Processing Conference, pp. 564–568.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 110, 346–359.

Ben-Cohen, A., Diamant, I., Klang, E., Amitai, M., Greenspan, H., 2016. Fully convolutional network for liver segmentation and lesions detection. In: LABELS/DLMIA@MICCAI, pp. 77–85.

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Buades, A., Le, T., Morel, J.-M., Vese, L., 2011. Cartoon+Texture image decomposition. Image Process. On Line 1. doi:10.5201/ipol.2011.blmv_ct.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 801–818.

Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L.J., 2011. Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans. Broadcast. 57 (2), 165–182. doi:10.1109/TBC.2011.2104671.

Efros, A.A., Freeman, W.T., 2001. Image quilting for texture synthesis and transfer. In: Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). ACM New York, pp. 341–346.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. Online http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Getreuer, P., 2012. Total variation deconvolution using split Bregman. Image Process. On Line 2, 158–174.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets. In: Conference on Neural Information Processing Systems, pp. 2672–2680.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision. Springer, pp. 346–361.

Hertzmann, A.P., 2001. Algorithms for rendering in artistic styles. New York University, Graduate School of Arts and Science PhD dissertation.

Iizuka, S., Simo-Serra, E., Ishikawa, H., 2017. Globally and locally consistent image completion. ACM Trans. Graph. 36 (4), 107:1–107:14.

Isabel, F., Sebastian, B., Carina, R., Jürgen, W., Stefanie, S., 2018. Generative adversarial networks for specular highlight removal in endoscopic images. In: Proc. SPIE, vol. 10576 doi:10.1117/12.2293755. pp. 10576 –10576 – 9

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition.

Köhler, R., Schuler, C., Schölkopf, B., Harmeling, S., 2014. Mask-specific inpainting with deep neural networks. In: German Conference on Pattern Recognition. Springer, pp. 523–534. LNCS

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J., 2017. Deblurgan: blind motion deblurring using conditional adversarial networks. arXiv preprint arXiv:1711.07064.

Lin, T.-Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 936–944.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, H., Lu, W.S., Meng, M.Q.H., 2011. De-blurring wireless capsule endoscopy images by total variation minimization. In: Pacific Rim Conference on Communications, Computers and Signal Processing. IEEE, pp. 102–106.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vis. 60 (2), 91–110.

Menor, D.P., Mello, C.A., Zanchettin, C., 2016. Objective video quality assessment based on neural networks. Procedia Comput. Sci. 96, 1551–1559.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR. abs/1411.1784.

Mo, X., Tao, K., Wang, Q., Wang, G., 2018. An efficient approach for polyps detection in endoscopic videos based on faster R-CNN. In: 24th International Conference on Pattern Recognition (ICPR), pp. 3929–3934.

Mohammed, A., Farup, I., Pedersen, M., Hovde, Ø., Yildirim Yayilgan, S., 2018. Stochastic capsule endoscopy image enhancement. J. Imag. 4 (6).

Newson, A., Almansa, A., Gousseau, Y., Pérez, P., 2017. Non-local patch-based image inpainting. Image Process. On Line 7, 373–385. doi:10.5201/ipol.2017.189.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2536–2544.

Prasath, V.B.S., 2016. Polyp detection and segmentation from video capsule endoscopy: a review. J. Imaging 3, 1–10. doi:10.3390/jimaging3010001.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 779–788.

Redmon, J., Farhadi, A., 2018. Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Conference on Neural Information Processing Systems, pp. 91–99.

Rodrguez-Snchez, A., Chea, D., Azzopardi, G., Stabinger, S., 2017. A deep learning approach for detecting and correcting highlights in endoscopic images. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6. doi:10.1109/IPTA.2017.8310082.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. Image Process. 15 (2), 430–444. doi:10.1109/TIP.2005.859378.

Shen, J., Chan, T.F., 2002. Mathematical models for local nontexture inpaintings. SIAM J. Appl. Math. 62 (3), 1019–1043.

Shin, Y., Qadir, H.A., Balasingham, I., 2018. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 6, 56007–56017. doi:10.1109/ACCESS.2018.2872717.

Stehle, T., 2006. Removal of specular reflections in endoscopic images. Acta Polytechnica 46 (4).

Tao, X., Gao, H., Shen, X., Wang, J., Jia, J., 2018. Scale-recurrent network for deep image deblurring. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8174–8182.

Tchoulack, S., Langlois, J.P., Cheriet, F., 2008. A video stream processor for real-time detection and correction of specular reflections in endoscopic images. In: Workshop on Circuit and Syst. and TAISA Conf.. IEEE, pp. 49–52.

Tong, H., Li, M., Zhang, H., Zhang, C., 2004. Blur detection for digital images using wavelet transform. In: International Conference on Multimedia and Expo. IEEE, pp. 17–20.

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155 (4), 1069–1078.e8. doi:10.1053/j.gastro.2018.06.037.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multi-scale structural similarity for image quality assessment. In: Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar), pp. 1398–1402.

Xu, L., Zheng, S., Jia, J., 2013. Unnatural L0 sparse representation for natural image deblurring. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1107–1114.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. CoRR. abs/1801.07892.

Zach, C., Pock, T., Bischof, H., 2007. A duality based approach for realtime TV-L1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (Eds.), Pattern Recognition, pp. 214–223.

Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y., 2018. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recognit. 83, 209–219. doi:10.1016/j.patcog.2018.05.026.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhou, T., Ruan, S., Canu, S., 2019. A review: deep learning for medical image segmentation using multi-modality fusion. Array 3-4, 100004. doi:10.1016/j.array.2019.100004.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International Conference on Computer Vision. IEEE, pp. 2242–2251.