Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15530-7

De-redundancy in wireless capsule endoscopy video sequences using correspondence matching and motion analysis

Libin Lan[1,2,3] · Chunxiao Ye[2,3] · Chao Liao[2,3] · Chengliang Wang[2,3] · Xin Feng[1]

Received: 10 November 2021 / Revised: 1 May 2022 / Accepted: 19 April 2023

Abstract
Handling wireless capsule endoscopy (WCE) de-redundancy is a challenging task. This paper proposes a scheme, called *SS-VCF-Der*, to consider applying a flow field estimation between two successive WCE frames to WCE imaging motion analysis and then address the WCE de-redundancy problem based on the results of the motion analysis. **To this end**, we intend to exploit a self-supervised technique to learn interframe visual correspondence representations from large amounts of raw WCE videos without manual human supervision, and predict the flow field. **Our key idea** is to use the natural spatial-temporal coherence in color and cycle consistency in time in WCE videos as free supervisory signal to learn WCE visual correspondence relations from scratch. We call this procedure self-supervised visual correspondence flow learning (*SS-VCF*). **At training time**, we use three losses: forward-backward cycle-consistency loss, visual similarity loss, and color loss, to train and optimize model. **At test time**, we use the acquired representation to generate a flow field for analyzing pixel movement between two successive WCE frames. Furthermore, according to the resulting flow field estimation, we compute the motion intensity of motion fields between two successive frames, and use our proposed de-redundancy method, namely *SS-VCF-MI*, to select some frames as key ones with distinct scene changes in local neighborhood so as to achieve the purpose of de-redundancy. Extensive experiments on our collected WCE-2019-Video dataset show that our scheme can achieve a promising result, verifying its effectiveness on the visual correspondence representation and redundancy removal for WCE videos.

Keywords Wireless capsule endoscopy · Self-supervised learning · Correspondence matching · Flow estimation · Motion intensity · Redundancy removal

✉ Chunxiao Ye
yecx@cqu.edu.cn

Extended author information available on the last page of the article

Published online: 27 July 2023

1 Introduction

Wireless capsule endoscopy (WCE) is the preferred unparalleled modality for diagnosis and assessment of small bowel diseases due to its many advantages, particularly its painless and noninvasive inspection [22]. However, during one WCE procedure, amounts of images with high similarity are generated, and only a small percentage of video data is useful for diagnosis. Manually reviewing these WCE frames is time-consuming and hard for an experienced clinician, and does not guarantee that some important abnormal information is not missed [3, 29, 54]. Although more recently, magnetic actuation and localization technology has also been developed to accelerate and locate the wireless capsule endoscopy in the human digestive tract [37, 62], this technique does not change the fact that WCE produces large amounts of similar redundant frames for a patient in one examination. Thus, it is extremely needed to explore new computational methods that can help clinician reduce the time spent in the examination.

Much work has been devoted to reducing redundancy and reviewing time [20], These methods used in these work include non-negative matrix factorization (NMF) [21], factorization analysis based on sliding window singular value decomposition (SVD) [4], video summarization [16, 40, 55, 63], feature representation [4, 6, 12, 35], image registration [14, 15, 26, 36, 57], similarity [1, 51], as well as motion analysis [33, 39, 43, 58] most related to our work. But these methods almost not attempt to use deep learning to benefit the de-redundancy problem [11], except for work [4, 6, 36, 58] that uses deep learning techniques to feature extraction. With respect to more work about reducing redundancy, readers can refer to the related review literature [25, 50].

Recently, using convolutional neural networks (CNNs) to find correspondences and to resolve optical flow estimation problem between two input images [10, 23, 60] in natural scene, has attracted more and more attention. In this paper, we introduce this idea into our WCE imaging motion analysis. We **focus on** the problem of how to establish visual correspondences for WCE video representations depicting the movement relations between two or more frames. **Our key idea** is that we can obtain unlimited supervision for the visual correspondence representations from large amounts of unlabeled raw WCE video data by using natural spatial-temporal coherence in color and cycle consistency in time in WCE videos as pretext task, as shown in Fig. 1 (bottom). **Our main aim** is that after learning this representation, this model can be used to compute coordinate-wise correspondences and to generate a flow field between two successive WCE frames. The resulting flow field estimation can be then used to our WCE de-redundancy scheme so as to achieve the task of redundancy removal.

However, it is a challenging task for learning representations for visual correspondence from the raw WCE video due to the non-rigid deformations and poor structural motion information, as well as low-texture context in WCE video. Meanwhile, with respect to WCE video, collecting the large-scale ground truth datasets necessary for high performance of learning visual correspondences, often requires extensive effort that is extremely expensive and even impractical. Furthermore, directly getting optical flow ground-truth labels for realistic video material, particularly for WCE video is known to be extremely difficult [5]. So, attempting to evaluate whether there is a scene change between two WCE frames based on a ground truth dataset in a supervised way, is almost infeasible, as shown in Fig. 1 (top). Moreover, some techniques employing handcrafted features such as SIFT [46] or HOG [7] have been successfully applied to dense semantic correspondence in natural scene. However, they may be limited in effectiveness for WCE video scene. That is mainly due to the fact that there exit

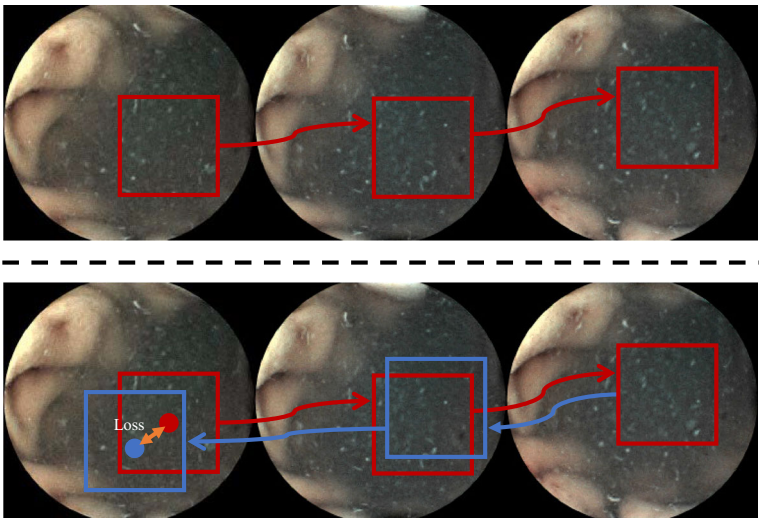


Fig. 1 The comparison between supervised and self-supervised learning. Learning visual correspondences via supervised learning requires ground-truth labels for every frame of the training videos. By utilizing the forward-backward cycle-consistency loss in the feature space and color loss in the *Lab* color space, we train an encoder chaining correspondence in each successive frame. Top: *Supervised* training. Bottom: *Self-supervised* training

some inherent limitations in WCE video [39, 41, 43], such as complicated and chaotic motion of the camera itself, no distinct foreground and background, and non-rigid deformation (local nonrigid motion) of gastrointestinal (GI) tract.

All these problems motivate us to propose a new self-supervised method[1] that can generate a flow field for pixel motion from the correspondences in each successive frame based on raw unlabeled WCE video itself, and then this flow field estimation can be applied to the WCE imaging motion analysis which is further considered applying to the WCE de-redundancy problem. We believe that correspondence flow is a crucial and suitable feature for the analysis of WCE scene changes based on the following experimental results.

In this paper, we propose a visual correspondence flow learning framework for the WCE visual correspondence representations which is trained in a self-supervised manner, referred to as visual correspondence flow learning (SS-VCF), that learns representations by finding these correspondences between successive frames and turn them into a learning signal. This representation can be future used to tackle the problem of redundancy elimination. Our work is inspired by the recent success of using correspondence learning to video representations in natural scene [13, 30, 44, 60]. The main objective of these work is to learn a representation for visual correspondence from raw video, in a self-supervised fashion, and some of them [30, 60] apply the learned representation to flow field estimation. Our work is also inspired by the innovative approach of [32, 59, 65] where colorization is used as a pretext task for learning in a self-supervised way. Common to all the above work is that these models are trained in a self-supervised manner to learn correspondences between observations adjacent in time, and thus our model is also trained in the same manner.

[1] We do not distinguish between the term unsupervised and self-supervised, as both refer to learning without human supervision. But in this paper, we use the term of self-supervised learning for WCE video representation.

Since our work aims to achieve the task of WCE de-redundancy, we refer to the related literatures [33, 39, 41, 43, 56, 58] involving motion analysis applied to WCE de-redundancy, and then divide our whole scheme called SS-VCF-Der into three main procedures. i) We learn visual correspondence representations from large amounts of raw WCE videos, in a self-supervised manner. We name the procedure self-supervised visual correspondence flow learning (SS-VCF); ii) We use this learned model to compute coordinate-wise correspondences and generate a flow field estimation between two successive WCE frames, which is then used to WCE imaging motion analysis; iii) We measure and analyze the degree or intensity of scene change in each successive frame by using motion intensity (MI) of the acquired motion estimation, and then adopt our proposed de-redundancy method, namely SS-VCF-MI, to decide whether there is a scene change between two successive frames and to select some frames as key ones with obvious scene changes. After all the three steps, we obtain the result of de-redundancy in WCE video, consisting of all the key frames. The framework of our de-redundancy scheme is presented in Fig. 2.

Quantitative and qualitative evaluations on our WCE-2019-Video dataset demonstrate that our proposed scheme is effective, and it can achieve the results on par with or even better than the other latest methods.

In summary, the main contributions in this paper are given as follows:

1. We first tentatively use a self-supervised method to learn visual correspondence representations between successive frames in WCE video. We further apply the learned representations to compute coordinate-wise correspondences and generate a flow field for pixel movement. According to the resulting flow field estimation, we compute motion intensity of motion fields between two successive frames as motion features. And then we use the extracted motion features to our proposed de-redundancy scheme so as to achieve the purpose of de-redundancy.
2. Based on the acquired motion features, we proposed a de-redundancy method to select some frames as key ones with distinct scene changes in local neighborhood. Whether there is a scene change between two successive frames will be determined by setting a local maximum value satisfying some conditions.
3. We have conducted extensive comparative studies on our WCE-2019-Video dataset. The experiments consist of two parts. The first part confirms that the proposed SS-VCF model is suitable for flow field estimation in WCE video, and the second verifies that our proposed SS-VCF-MI method is effective for WCE redundancy removal. With respective to the two experiments, we evaluate the performance of the method in each experiment based on different evaluation metrics from both quantitative and qualitative perspectives.

The remainder of this paper is organized as follows. Section 2 reviews prior work. Section 3 specifies our whole scheme, including training process for WCE video representations and de-redundancy approach. Section 4 gives experimental details and results under both the learned representation and de-redundancy approach, and compares with other methods, and finally Section 5 presents our conclusion.

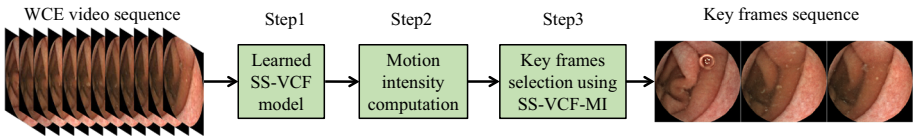


Fig. 2 The workflow of our entire de-redundancy scheme

2 Related work

In this section, we perform a literature review, covering the following several aspects: i) correspondence matching, ii) optical flow, iii) self-supervised representation learning, iv) related techniques applied to WCE redundancy removal.

Correspondence matching Earlier much work mainly focused on estimating correspondences for a pair of images including the same scene or object by using hand-crafted features such as SIFT [46] or HOG [7]. Recently, many researchers have studied the task of correspondence matching in natural scenes using deep CNNs [17, 27, 30, 38, 53, 60]. But different from the work from Rocco et al. [53] which trains CNNs by learning from synthetic transformations in natural scenes, we learn correspondence matching by exploiting the inherent spatial-temporal coherence in realistic medical WCE video data itself instead of synthetic training data.

Optical flow The conventional variational approaches have dominated optical flow estimation. Since the energy minimization framework of Horn and Schunck [19] and coarse-to-fine image warping by Lucas and Kanade [47], much success has been made in optical flow estimation. Recently, CNNs have been applied to solving optical flow estimation problem [10, 23, 34] in natural scenes. We introduce similar techniques [30, 60] into our task of WCE imaging motion analysis. We use CNN to learn the flow estimation from WCE video itself by considering the natural spatial-temporal coherence of time and color in each successive frame.

Self-supervised representation learning from video Our framework closely relates to the self-supervised representation learning. Learning representations from video using various pretext tasks as supervision has been proposed. Some pretext tasks involve colorization [59], cycle-consistency in time [49, 60], or form pseudo-labels by Siamese correlation filter network [61]. Our work is inspired by the colorization, cycle-consistency in time, and pseudo-labels, all of which can be used to provide supervisory signal for training. Following the above three work, we use the natural spatial-temporal coherence of time and color in WCE videos as a pretext task and integrate the representation learning and the tracker into a self-supervised training procedure, in an end-to-end fashion.

Related techniques applied to endoscopy video de-redundancy Various related techniques applied to WCE video de-redundancy have been extensively studied. Some optical flow-based approaches have been explored to reduce the redundancy and reviewing time [33, 39, 43, 56]. Also, there are other methods, e.g., video summarization techniques [16, 40, 55, 63] used to the de-redundancy problem. In these methods, optical flow-based approaches are most related to ours. However, these methods almost not benefit from deep learning techniques, particularly CNNs. Our method learns visual correspondence representations in a self-supervised manner by using a CNN architecture, and generates the flow field in each successive frame. To the best of our knowledge, there is still no related work applying this similar approach of visual correspondence learning to WCE video de-redundancy. Our work should be the first attempting to apply it to the problem of WCE redundancy elimination.

3 Methods

In this section, we will elaborate on our de-redundancy scheme from the following three respects: i) What is the SS-VCF model and how to train it; ii) How to extract motion features

Table 1 Some abbreviations with the corresponding full names (a)

(a) list of abbreviations

Abbr.	Full Name	Abbr.	Full Name
WCE	Wireless capsule endoscopy	MI	Motion intensity
CNN	Convolutional neural network	GI	Gastrointestinal

(b) list of notations

Symbol	Description	Symbol	Description
$\mathcal{T}$	Differentiable tracker	ϕ	Spatial feature encoder
I	Image	p	Image patch
f	Affinity function	A	Affinity matrix
g	Localizer	h	Bilinear Sample
$\mathcal{L}$	Loss function	λ	Balance parameter
θ	Localization parameter	x	Spatial feature

The notations used in our method and corresponding descriptions (b)

between two successive WCE frames using the SS-VCF model; iii) How to analyze the scene changes and select key frames via extracted motion features. Here we also list some abbreviations with the corresponding full names across the full paper, as well as a concise reference describing the notation used throughout our method in Table 1.

3.1 SS-VCF model

We first give the description of our model, and then show how to train our model for visual correspondence representation in WCE video. Finally, we give some implementation details.

3.1.1 Overview of our model

We take the design of TimeCycle [60] as the baseline. Our SS-VCF model consists of a spatial feature encoder ϕ and a differentiable tracker $\mathcal{T}$. The spatial feature encoder can be any form of CNN architecture, and the differentiable tracker is a spatial transformer network [24], performing co-localisation task. In this paper, we follow [60] and adopt a same ResNet-50 architecture [18] as the spatial feature encoder ϕ. The differentiable tracker $\mathcal{T}$ can be inserted into the bottom of ϕ. We take one cycle in time as example to give an illustration of the training procedure, as shown in Fig. 3.

Similarly to [60], our goal is to learn a feature space ϕ by tracking a patch p_t extracted from image I_t forwards and then backwards in time, while minimizing three losses: cycle-consistency loss $\mathcal{L}_{cycle}$, similarity loss $\mathcal{L}_{sim}$, and color loss $\mathcal{L}_{color}$, where $\mathcal{L}_{sim}$ and $\mathcal{L}_{color}$ include two losses in both forward and backward paths, respectively. The cycle-consistency loss $\mathcal{L}_{cycle}$ is the euclidean distance between the spatial coordinates of initial patch p_t and the patch $\hat{p}_t$ found at the end of the cycle in I_t. The similarity loss $\mathcal{L}_{sim}$ explicitly require the current patch p_t and target patch p_{t+1}, as well as p_t and $\hat{p}_t$, to be similar in the feature space, which amounts to the negative Frobenius inner product between spatial feature tensors p_t and p_{t+1}, as well as p_t and $\hat{p}_t$. The color loss is the cross-entropy categorical loss in the Lab space between the current patch p_t and and target patch p_{t+1}, and between current patch p_t

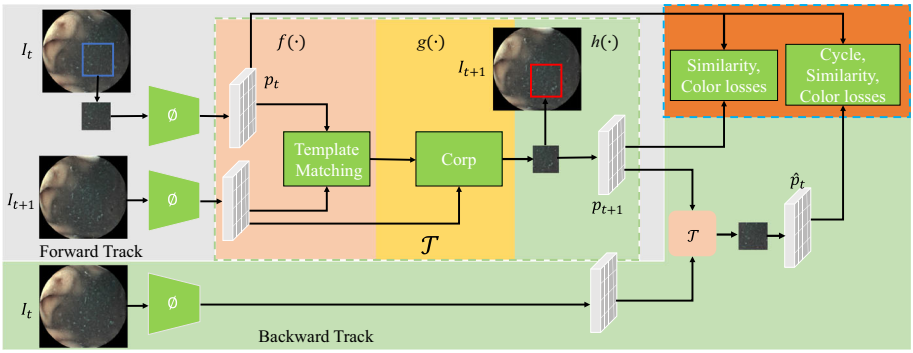


Fig. 3 The overview of SS-VCF model. Note that color loss is performed in *Lab* color space

and the patch $\hat{p}_t$ found at the end of the cycle in I_t. We learn an optimal feature space ϕ by minimizing the sum of three losses.

Training ϕ relies on a differentiable tracking operation $\mathcal{T}$, which takes as inputs the features of a current patch p_t and a target image I_{t+1}, and returns the image feature region with maximum similarity and minimum color loss. The implementation of $\mathcal{T}$ is shown in the green dotted-line box of Fig. 3. $\mathcal{T}$ must match features to localize the next patch and can be iteratively applied forwards and then backwards through time to track along an arbitrarily long cycle. It consists of three main components: affinity function f, localizer g, and and bilinear sampler h, which draws from the work by Wang et al. [60]. Here, we briefly review it.

Affinity function f provides a measure of similarity between coordinates of spatial features x^I and x^p, where $x^I = \phi(I)$ and $x^p = \phi(p)$. We define the affinity function as $f(x^I, x^p) := A$, and denote spatial grid j in feature x^I as $x^I(j)$ and the grid i in x^p as $x^p(i)$, such that

$$A(j,i) = \frac{\exp(x^I(j)^{\mathrm{T}} x^p(i))}{\sum_j \exp(x^I(j)^{\mathrm{T}} x^p(i))}, \tag{1}$$

where the similarity $A(j,i)$ is normalized by the softmax over the spatial dimension of x^I, for each x^p.

Localizer g takes affinity matrix A as input and estimates localization parameters θ corresponding to the patch in feature x^I which best matches x^p. g is composed of two convolutional layers and one linear layer which regresses to the θ parameters, and it is restricted to output 3 parameters for the bilinear sampling grid, corresponding to 2D translation and rotation: $g(A) := \theta$.

Bilinear sampler h uses the image feature x^I and θ predicted by g to perform bilinear sampling to produce a new patch feature $h(x^I, \theta)$ which is the same size as x^p.

3.1.2 Training loss

Given a sequence of video frames $I_{t:t+k}$ and a patch p_t taken from I_t, as well as their corresponding spatial features: $x^I_{t:t+k} = \phi(I_{t:t+k})$ and $x^p_t = \phi(p_t)$, where k denotes the

number of forward frames. Let $\mathcal{T}$ be a differentiable operation $x_s^I \times x_t^p \mapsto x_s^p$, where s and t represent time steps. The role of $\mathcal{T}$ is to localize the patch features x_s^p in image features x_s^I that are most similar to x_t^p. We formulate the forward-backward cycle-consistency tracking as an iterative process below in a forward manner i times from $t+1$ to $t+i$ and in a backward manner i times from $t+i$ to $t+1$, respectively:

$$\mathcal{T}^{(i)}(x_{t+i}^I, x^p)=\mathcal{T}(x_{t+i}^I, \mathcal{T}(x_{t+i-1}^I, ...\mathcal{T}(x_{t+1}^I, x^p))), \tag{2}$$

$$\mathcal{T}^{(-i)}(x_{t+1}^I, x^p) = \mathcal{T}(x_{t+1}^I, \mathcal{T}(x_{t+2}^I, ...\mathcal{T}(x_{t+i}^I, x^p))). \tag{3}$$

Based on the above formulation, similarly to [60], we give the following loss functions to train the SS-VCF model: i) cycle-consistency loss $\mathcal{L}_{cycle}$; ii) feature similarity loss $\mathcal{L}_{sim}$; iii) color loss $\mathcal{L}_{color}$, as illustrated in the blue dotted-line box of Fig. 3.

Cycle-consistency loss $\mathcal{L}_{cycle}^i$. The cycle-consistency loss $\mathcal{L}_{cycle}^i$ is defined as:

$$\mathcal{L}_{cycle}^i=l_\theta(x_t^p, \mathcal{T}^{(-i)}(x_{t+1}^I, \mathcal{T}^{(i)}(x_{t+i}^I, x_t^p))). \tag{4}$$

The tracker attempts to follow features forward and then backward steps i in time to re-arrive to the initial patch. l_θ is Euclidean distance between them.

Feature similarity loss $\mathcal{L}_{sim}^i$ We explicitly require the current patch $\mathcal{T}^{(i)}(x_{t+i-1}^I, x_t^p)$ and localized patch $\mathcal{T}^{(i)}(x_{t+i}^I, x_t^p)$ in the forward path, as well as current patch $\mathcal{T}^{(-i)}(x_{t+1}^I, \mathcal{T}^{(i)}(x_{t+i}^I, x_t^p))$ and localized patch $\mathcal{T}^{(-i)}(x_{t+1}^I, \mathcal{T}^{(i)}(x_{t+i-1}^I, x_t^p))$ in the backward path, to be similar in feature space. This loss amounts to the sum of negative Frobenius inner product between spatial feature tensors in both the forward and backward paths:

$$\begin{aligned}\mathcal{L}_{sim}^i=\mathcal{L}_{sim}^i+\mathcal{L}_{sim}^{(-i)}=-(&< \mathcal{T}^{(i)}(x_{t+i-1}^I, x_t^p), \mathcal{T}^{(i)}(x_{t+i}^I, x_t^p) > \\ &+ < \mathcal{T}^{(-i)}(x_{t+1}^I, \mathcal{T}^{(i)}(x_{t+i}^I, x_t^p)), \mathcal{T}^{(-i)}(x_{t+1}^I, \mathcal{T}^{(i)}(x_{t+i-1}^I, x_t^p)) >)\end{aligned}. \tag{5}$$

Color loss $\mathcal{L}_{color}^i$ Similarly to [59], we cast frame reconstruction as a classification problem. The color for each pixel is quantized into 16 classes with K-means clustering in the Lab space. The objective function is defined as:

$$\mathcal{L}_{color}^i=\alpha_1 \sum_{m=1}^{i} \mathcal{L}_1(I_m, \hat{I}_m)+\alpha_2 \sum_{n=i}^{1} \mathcal{L}_2(I_n, \hat{I}_n), \tag{6}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ refer to the pixel-wise cross entropy between current patch and reconstructed patch in the forward and backward paths, and the loss weights are set as α_1=0.8 and α_2=0.2, respectively in all our experiments. Also, the predicted colors in I_{t+1} are viewed as a linear combination of colors in the past frame:

$$I_{t+1} = \sum_t A_{(t,t+1)} I_t. \tag{7}$$

Overall loss $\mathcal{L}$ The overall learning objective sums over the k possible cycles, with weight λ_1=0.8 and λ_2=0.5:

$$\mathcal{L}=\sum_{i=1}^{k} \mathcal{L}_{sim}^i + \lambda_1 \mathcal{L}_{cycle}^i + \lambda_2 \mathcal{L}_{color}^i. \tag{8}$$

Training SS-VCF The combination of ϕ and $\mathcal{T}$ forms a forward-backward cycle tacker, allowing for end-to-end training:

$$x^I, x^p = \phi(I), \phi(p), \tag{9}$$

$$\mathcal{T}(x^I, x^p) = h(x^I, g(f(x^I, x^p))). \tag{10}$$

This training procedure can be described in Algorithm 1.

Algorithm 1 Training SS-VCF model.

Input: Video frame sequences $I_{t:t+k}$ and a patch p_t taken from I_t, CNN model encoder ϕ, differentiable tracker $\mathcal{T}$
Output: Learned model ϕ, $\mathcal{T}$
1: **for** training epochs **do**
2: % Map to feature space
3: $x^I_{t+i}, x^p_t = \phi(I_{t+i}), \phi(p_t)$
4: % Compute loss according to (4), (5), (6), and (8).
5: $\mathcal{L} = \sum_{i=1}^{k} \mathcal{L}^i_{sim} + \lambda_1 \mathcal{L}^i_{cycle} + \lambda_2 \mathcal{L}^i_{color}$
6: % Update model
7: $\phi = update(\phi, \frac{\partial \mathcal{L}}{\partial \phi})$
8: $\mathcal{T} = update(\mathcal{T}, \frac{\partial \mathcal{L}}{\partial \mathcal{T}})$
9: **end for**

3.1.3 Implementation details

Training We train the model without using any annotations or pre-training on the WCE-2019-Video dataset containing 102,346 frames. We evaluate our model on WCE-2019-Video dataset from the quantitative and qualitative perspectives respectively, verifying its effectiveness on learning visual correspondence representations in WCE video. During training, we set the temporal length as $k = 4$. We train our SS-VCF model with 30 epochs on an Nvidia TitanXp graphics card using Adam [28] optimizer whose learning rates and batch size are set to 2e-4 and 16, respectively. We implement our approach using pytorch [52].

Inference At test time, based on (1), we apply the learned representations to compute coordinate-wise correspondences and generate a flow field for pixel movement.

3.2 Scene changes analysis

The literature [43] simultaneously considers the movement of WCE camera and gastrointestinal tract to model the WCE imaging motion, which it seems to be more complicated to clinical practice, since it is difficult to predict the motion orientation of WCE camera. In this paper, according to the inherent properties of WCE video itself, such as chaotic motion of WCE camera, non-rigid deformation of GI tract, and low-texture context of WCE image, we directly model the WCE video motion and analyze scene changes between two successive WCE video frames from I_t to I_{t+1} in a WCE video. In order to analyze scene changes, we first estimate the temporal motion field between both using the above learned SS-VCF model. Since motion intensity (MI) (viz. magnitude) has power to describe inter-frame motion

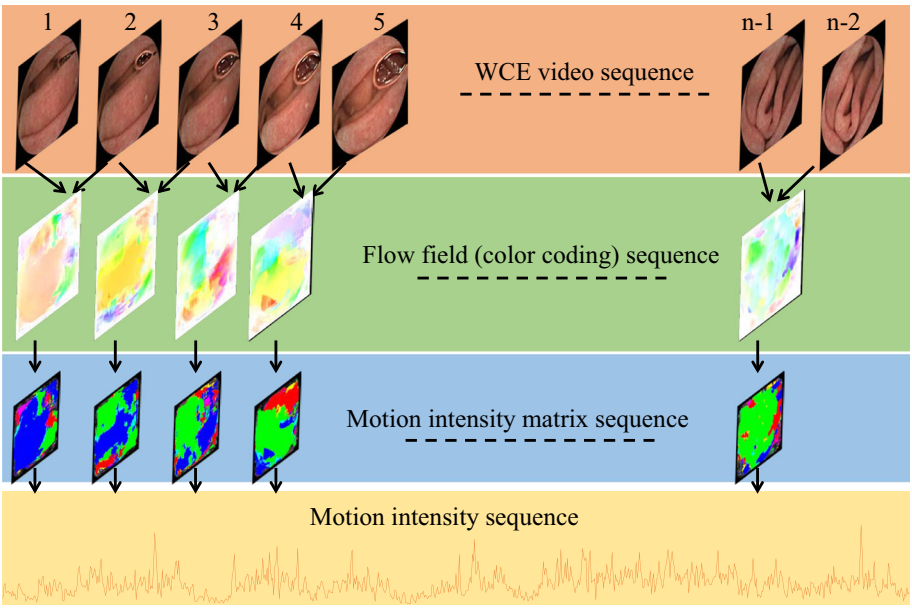


Fig. 4 An illustration of this idea using MI to WCE video scene changes analysis

[8, 45], we compute the motion intensity of motion field between two successive frames as motion features to decide whether scene change occur by using the definition in [8, 9, 19]. We adapted the definition to our experiments, which is presented as follows:

$$MI = \frac{1}{MN} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \sqrt{({U_{k,l}}^2 + {V_{k,l}}^2)}, \tag{11}$$

where U and V are the horizontal and vertical vectors in flow field matrix, respectively. M and N are the width and height of flow field matrix, respectively. (k, l) indicates the indices of flow field matrix. $\sqrt{U^2+V^2}$ denotes motion intensity matrix corresponding to one flow field matrix. These MI values form one-dimensional signal in the time domain, which is convenient to analyze some fluctuations along the one-dimensional signal. This idea of WCE video scene changes analysis using MI is illustrated in Fig. 4.

3.3 WCE de-redundancy process

An initial approach extracting key frames was to choose the first frame of a segment (or shot) as the key-frame. The same kind of method was also used in [43]. It is a reasonable approach and works well for low-motion or stable-motion shots in natural scenes. However, for endoscopy video, selecting the first frame as a key-frame may be not appliable. Similarly to [39, 41], we also adopt a peak value along the curve of motion intensity (MI) to decide whether a large motion occurs between two successive frames, which indicates that there are some significant scene changes between both of the two frames. We select one of two frames adjacent in the peak value as the key-frames to accomplish the process of redundancy removal. Since the peak value is a local activity in the time domain, it is more suitable for deciding whether there is a scene change between two successive frames by setting a local

threshold. So, following [39, 41], we can claim a local maximum as the peak value to decide whether a large motion occurs, and further to consider selecting these frames adjacent in the peak value as key-frames. The local maximum from I_t to I_{t+1} satisfies the following conditions: (i) The local value is the maximum in a symmetric neighbor of size $2m-1$, and (ii) The local maximum is also n times of the second largest value in the neighbor. We set the size of symmetric neighbor to 3, i.e., $m=2$, and times as $n=1.5$ in all our experiments. The whole extraction procedure of key-frames in a segment can be described in algorithm 2.

Algorithm 2 The procedure of our de-redundancy scheme SS-VCF-Der

Input: A WCE video sequence
1: Train the SS-VCF model.
2: Compute coordinate-wise correspondences between two successive frames by the learned SS-VCF model.
3: Generate a flow field for pixel movement between I_t and I_{t+1}.
4: Calculate the motion intensity (MI) according to the flow field generated in step 3.
5: Find peak values according to the conditions of local maximum in main text.
6: Decide which frames are selected as key frames.
7: Generate a short summary.
Output: A summarized sequence consisting of a set of key frames from input sequence

4 Results and discussion

The goal of our work is to process WCE video sequences that contain high redundancy. Our entire experiment consists of two parts. The first part checks whether the proposed SS-VCF model is suitable for flow field estimation of WCE video, and the second part verifies whether our proposed SS-VCF-MI method is effective on the de-redundancy of WCE video.

In the first experiment, i.e., the experiment of modeling the visual correspondence representation (SS-VCF) in WCE video, we conduct a set of experiments on our WCE-2019-Video dataset from both quantitative and qualitative aspects respectively, and compare our method with other approaches. In this experiment, we consider reconstruction error used in [60] as a main evaluation metric to evaluate the performance of SS-VCF model and others. Also, we use F-score and compression ratio (CR) as metric to reflect the representation capability of the generated flow field as motion feature. These metrics are commonly used for video summarization [48, 64]. Based on these metrics, we report our quantitative results. Furthermore, we also report our qualitative results via two visualizing results: flow flied between two succussive frames, and warping results with this generated flow.

In the second one, i.e., the experiment of de-redundancy (SS-VCF-MI), we also give the quantitative and qualitative results. We use F-score and compression ratio as the principal criterion to measure the de-redundancy performance, which are widely used metrics for video summarization. We compare the performance of our de-redundancy scheme: SS-VCF-Der with other methods on our WCE-2019-Video dataset. For the definitions of F-score and compression ratio, please refer to related literatures [48, 58, 64].

4.1 Datasets

WCE-2019-Video dataset is specially collected for the task of WCE video summarization in 2019, with frame-level importance scores. Our built WCE-2019-Video dataset contains 5 categories and 30 videos (6 per category from 6 patients) collected at the first phase of the

task, and other two videos (corresponding 2 categories from the seventh patient) collected at the second phase, totaling 32 videos and 102,346 frames. Each video varies from 600 to 7500 frames. We use these videos from patient ID #3–#7 in WCE-2019-Video dataset, totaling 22 videos with 69,843 frames as training set, and the remaining videos from patient ID #1–#2, totaling 10 videos with 32,503 frames as testing set for evaluating our SS-VCF model and de-redundancy scheme.

The learned representation is evaluated without fine-tuning on our WCE-2019-Video dataset. Since getting optical flow ground truth for such a large-scale dataset is impossible, we did not annotate this dataset for our task of visual correspondence learning, but the frame-level importance scores of each video in WCE-2019-Video dataset have been annotated by 6 experienced clinicians. So, except for the metric of reconstruction error, we also use the F-score and compression ratio to implicitly reflect the representation capability of SS-VCF model.

4.2 Baseline

We compare with the following baselines:

Optical Flow(HS [19]) We use standard optical flow of Horn and Schunck (HS) to compute flow field between two consecutive WCE frames I_t and I_{t+1}. The optical flow estimation is viewed as an energy minimization problem based on brightness constancy and spatial smoothness. We then use this resulting flow field to compute the motion intensity between them. Also, we warp it on I_t to generate a new frame I'_t, which is used to the intuitive comparison of visualization and to computing quantitative result of reconstruction error.

SIFT Flow [42] For a reference frame I_t, we compute the SIFT flow between frames I_t and its following frames I_{t+1} from SIFT images of the two frames via an objective function. The objective function for SIFT flow aims to finding a best match. We then use this resulting flow flied to compute the corresponding motion intensity and further analyze the motion information between two frames, as well as visualize the warping results.

Video Colorization [59] A self-supervised approach using color as a supervisory signal to learn visual representations on WCE-2019-Video dataset from scratch. But the architecture of this method is 3D ResNet-18. We use the same settings from the published papers to train the colorization model, and make good use of it to compute the flow field between two successive frames, and then analyze the motion information. Finally, we report our experimental results.

TimeCycle [60] A self-supervised method which use cycle-consistency in time as free supervisory signal to learn visual representations from scratch. We use the learned visual representations to compute coordinate-wise correspondences and generate a flow field for pixel movement from frame I_t to I_{t+1}. We resize the flow field to the same size as the frame I_t by bilinear interpolation, and warp it to generate a new frame I'_t. We then report the reconstruction error between them and visualize the warping results.

Correspondence Flow (CorrFlow [30]) It is also a self-supervised method using the combination of various methods, such as color dropout, restricted attention, scheduled sampling and cycle consistency. We adopt its open-source implementation to our experiments on WCE-2019-Video dataset, and use the trained model to generate a flow field, and compute the

motion intensity and illustrate the visualization of warping results between two consecutive frames I_t and I_{t+1}.

BAME-SIFTFlow [43] An invalid region is defined for a robust measurement of scene changes between two successive WCE frames. Meanwhile, a reduction scheme is designed to select key frames according to the method of setting both the first frame as key frame and the threshold of the diameter of the max-inscribed circle (DMC) of the defined invalid region. We compared the performance of both our de-redundancy scheme and theirs on WCE-2019-Video dataset.

4.3 Ablation analysis

We conducted several ablation studies to analyze the contribution of each component of the SS-VCF model by removing one component at a time. Ablation experiments are performed on our WCE-2019-Video dataset. We adopt reconstruction error used in [60] as an evaluation metric, which is also used as quantitative evaluation metric in our first experiment. All models are trained from scratch without fine-tuning on our training set from WCE-2019-Video dataset, and evaluated on the testing set. More concretely, we generate the optical flow between two successive frames and warp the first one I_t with this resulting flow to a new frame I'_t, and then compute reconstruction error between the warped frame I'_t and second frame I_{t+1} by using L1 distance. The reconstruction error is an average value on our testing set from WCE-2019-Video dataset, which is statistically meaningful. Quantitative comparisons of ablation studies are shown in Table 2. Depending on which training loss is adopted, we consider the following ablation variants of SS-VCF.

SS-VCF$_{\text{w/o-cycle}}$ This variant indicates that the strategy of cycle-consistency $\mathcal{L}_{cycle}$ is not adopted. This variant uses the loss of feature similarity $\mathcal{L}_{sim}$ and the color loss $\mathcal{L}_{color}$ to train model. It is used to verify the effects of cycle-consistency to the capability of visual correspondence representation learning. This case is equivalent to adding a loss of feature similarity $\mathcal{L}_{sim}$ to colorization model [59] which uses a pointer mechanism to reconstruct a target frame I'_{t+1} by copying pixels from a reference frame I_t. The colorization model is trained such that the predicted colors in I'_{t+1} are close to the true target colors in I_{t+1}.

SS-VCF$_{\text{w/o-sim}}$ The variant denotes that the loss of feature similarity $\mathcal{L}_{sim}$ is not included, which is explicitly required in our SS-VCF model. This variant uses the cycle-consistency loss $\mathcal{L}_{cycle}$ and the color loss $\mathcal{L}_{color}$ to train model. It is used to verify the effects of the loss

Table 2 Performance comparisons on reconstruction errors of SS-VCF model and its ablation variants on WCE-2019-Video dataset for two gaps

Method	1-F	5-F
SS-VCF$_{w/o-cycle}$	38.6	57.6
SS-VCF$_{w/o-sim}$	39.2	58.1
SS-VCF$_{w/o-color}$	38.7	56.3
SS-VCF	31.2	46.2

The gaps are 1 or 5 frames. The lower the reconstruction error, the better the performance

of feature similarity $\mathcal{L}_{sim}$ to the capability of visual correspondence representation learning. This case is similar to CorrFlow model [30].

SS-VCF$_{\text{w/o-color}}$ This variant indicates that the color loss $\mathcal{L}_{color}$ is not adopted. This variant uses the cycle-consistency loss $\mathcal{L}_{cycle}$ and the loss of feature similarity $\mathcal{L}_{sim}$ to train model. It is used to verify the effects of colorization to the capability of visual correspondence representation learning. This case is similar to TimeCycle model [60], but without the loss of skip cycle.

SS-VCF In this case, the overall loss function $\mathcal{L}$ is the final objective for training the SS-VCF model in a self-supervised manner. We show that SS-VCF can yield best results when combined with all the unsupervised losses above.

Comparing reconstruction error of SS-VCF$_{w/o-cycle}$ with SS-VCF$_{w/o-sim}$ in Table 2, it can be seen from Table 2 that the performance of SS-VCF model without the loss of cycle-consistency $\mathcal{L}_{cycle}$ all outperforms that of SS-VCF model without the loss of feature similarity $\mathcal{L}_{sim}$ on both two gaps. This indicates that the loss of feature similarity $\mathcal{L}_{sim}$ than the loss of cycle-consistency $\mathcal{L}_{cycle}$ is more benefit to the reconstruction.

Also, one can see that the performance of SS-VCF model without the color loss $\mathcal{L}_{color}$ outperforms that of SS-VCF model without the loss of feature similarity $\mathcal{L}_{sim}$ on both two gaps. This indicates that the color loss $\mathcal{L}_{color}$ than the loss of feature similarity $\mathcal{L}_{sim}$ makes more contributions on improving the performance of the reconstruction.

Additionally, as one can see, the performance of SS-VCF model without the loss of cycle-consistency $\mathcal{L}_{cycle}$ outperforms that of SS-VCF model without the color loss $\mathcal{L}_{color}$ on 1-frame gap, but not on 5-frame gap. This indicates that the loss of cycle-consistency $\mathcal{L}_{cycle}$ may be helpful for long-range correspondence representation leaning. Furthermore, we also notice that the performances of both them on 1-frame gap are almost the same. This shows that they have almost identical ability for correspondence representation between two frames adjacent in time.

Finally, the performance of SS-VCF model is the best. This shows that the SS-VCF combining with all the unsupervised losses can yield best the performance of visual correspondence representation learning.

4.4 Quantitative results

4.4.1 Experimental comparisons of SS-VCF and other methods

Since our approach mainly aims to self-supervised learning, we compare our learned SS-VCF model with the other five non-supervised methods: Optical Flow (HS [19]), SIFT Flow [42], Video colorization model [59], TimeCycle [60], and CorrFlow [30] on our WCE-2019-Video dataset. Note that here all methods use the same settings from their published papers, some of which are slightly different from the variants of ablation analysis in Section 4.2. After learning, for given two frames I_t and I_{t+1} adjacent in time in a WCE video, we compute coordinate-wise correspondences under each acquired model, and generate a flow field for per-pixel movement between them. We then warp the flow field on frame I_t to generate a new image I'_t similar to I_{t+1}. Following [60], we compare the L1 distance (viz. Manhattan distance, a.k.a., CityBlock distance) between I'_t and I_{t+1} in RGB space and report the reconstruction errors in Table 3.

Table 3 Performance comparisons on reconstruction errors of various models on WCE-2019-Video dataset for both two gaps

Method	1-F	5-F
Optical Flow (HS [19])	41.2	65.2
SIFT Flow [42]	39.3	60.5
Video Colorization [59]	39.2	58.7
TimeCycle [60]	38.7	56.3
CorrFlow [30]	39.5	59.1
SS-VCF	31.2	46.2

The two gaps are 1 frame or 5 frames

We use two gaps: 1-frame and 5-frame in this paper, to show and compare the performance of each model. The reconstruction error is an average value on our testing set from WCE-2019-Video dataset, which is statistically meaningful. The results are presented in Table 3.

One can see from Table 3 that the performance of SS-VCF is best on both two gaps. This shows that all the three losses together can train a best representation model of visual correspondence, that compared with the traditional handcrafted methods, the CNN-based method can learn a better visual correspondence representation.

Additionally, we use F-score and compression ratio (CR) as metric to compare our model with the other five methods to implicitly reflect the representation capability of learned SS-VCF model generating flow field as motion feature. The results are presented in Table 4, all of which are based on our SS-VCF-MI de-redundancy method. It can be seen from Table 4 that the summarization performance of SS-VCF model outperforms that of all other methods on both metrics.

4.4.2 Experimental comparisons of SS-VCF-Der and other methods

In this section, for the sake of space, we take as example 5 video segments of patient ID #2 to demonstrate the details of extracting key frames. Each segment takes the first 50 frames in corresponding video including #2, #8, #14, #20, and #26. Here we report the indices of extracted key frames and compression ratio. The results are based on our SS-VCF model and SS-VCF-MI de-redundancy method, which are shown in Table 5. One can see that all the compression ratios are more than 70%.

Additionally, for a fair comparison with BAME-SIFTFlow [43] and testing whether our SS-VCF-MI de-redundancy method is effective on WCE video, we conduct a comparable experiment for the two de-redundancy schemes. According to the evaluation metrics of F-

Table 4 Comparisons on F-score (%) and compression ratio (CR) (%) of our method and other methods on the testing set, totaling 32,503 frames

Method	F-score	CR
Optical Flow (HS [19])	32.2	71.6
SIFT Flow [42]	34.5	69.6
Video Colorization [59]	37.3	71.8
TimeCycle [60]	36.5	71.4
CorrFlow [30]	35.2	70.1
SS-VCF	38.3	72.4

Table 5 The index and number of key frames in each segment and compression ratio (CR)

Segment	Key frame indices	Number	CR
000172-000221	000179, 000181, 000185, 000191, 000193, 000197,	13	74
Stomach	000199, 000204, 000207, 000210, 000213, 000216, 000219		
001860-001909	001863,001866, 001868, 001870, 001872, 001875,	14	72
Duodenum	001884, 001887, 001890, 001897, 001899, 001901, 001904, 001908		
002939-002988	002941, 002948, 002950, 002955, 002958, 002962,	13	74
Jejunum	002964, 002968, 002970, 002974, 002978, 002983, 002986		
006228-006277	006232, 006237, 006240, 006244, 006250, 006252,	11	78
Ileum	006257, 006261, 006265, 006269, 006273		
018240-018289	018243, 018245, 018249, 018251, 018255, 018257,	14	72
Colon	018262, 018266, 018269, 018272, 018277, 018286, 018288		

Table 6 Comparisons on F-score (%) and compression ratio (CR) (%) of our method and other methods on the testing set, totaling 32,503 frames

Method	F-score	CR
BAME-SIFTFlow [43]	38.7	67.2
SS-VCF-Der	34.5	69.6

score and compression ratio, we report the quantitative results on the testing set. All the results are based on SIFT Flow [42] and presented in Table 6.

Furthermore, we conduct a comparable experiment for the SS-VCF-Der scheme and our another work: Adv-Ptr-Der-SUM [31], and report the results on F-score and compression ratio. For a fair comparison, both models are trained on same training set and evaluated on same testing set. The results are based on SS-VCF model and SS-VCF-MI de-redundancy method, as well as Adv-Ptr-Der-SUM model, respectively, as presented in Table 7.

It can be seen from Tables 6 and 7 that SS-VCF-Der is superior than both BAME-SIFTFlow and Adv-Ptr-Der-SUM on compression ratio, but inferior on F-score, which shows that the SS-VCF-Der may obtain a low recall or precision. The reasons may stem from the following aspects. Firstly, the WCE-2019-Video dataset itself may be not accurately and objectively annotated, which results in an undesirable performance. Although we have set some criteria for annotation, the process of annotation may be subjective due to the clinician's preferences. Furthermore, the setting conditions of selecting key frames may be not reasonable. We believe that an average value of motion intensity should be set, and these frames adjacent in the peak value with higher MI than average value should be selected as key-frames. It may lead to a high F-score. Additionally, the summarization performance of SS-VCF-Der is lower than that of Adv-Ptr-Der-SUM. We believe that this may be because the two work adopts different techniques. Adv-Ptr-Der-SUM is based on learning, while SS-VCF-Der relies on visual representation and local threshold.

4.5 Qualitative results

4.5.1 Experimental comparisons of SS-VCF and other methods

We follow the code in [2] to visualize the flow fields generated by each representation method. The visualization results include two gaps: 1-frame and 5-frame. Also, we give the visualization of warping results. These are shown in Fig. 5(b) and (c), respectively. We take as example frames from #001860 to #001865 to demonstrate the results of visualization. These frames come from a sample video #8 of patient ID #2 in WCE-2019-Video, and are shown in Fig. 5(a).

It can be seen from Fig. 5 that our method can capture more details, which may contain potential diseases or important something. This is very important for an accurate diagnosis.

Additionally, we plot the curves of motion intensity generated by our model using (11) and the other five methods. The curves are presented in Fig. 6. This experiment is conducted

Table 7 Comparisons on F-score (%) and compression ratio (CR) (%) of both Adv-Ptr-Der-SUM and SS-VCF-Der on the testing set, totaling 32,503 frames

Method	F-score	CR
Adv-Ptr-Der-SUM [31]	49.2	69.2
SS-VCF-Der	38.3	72.4

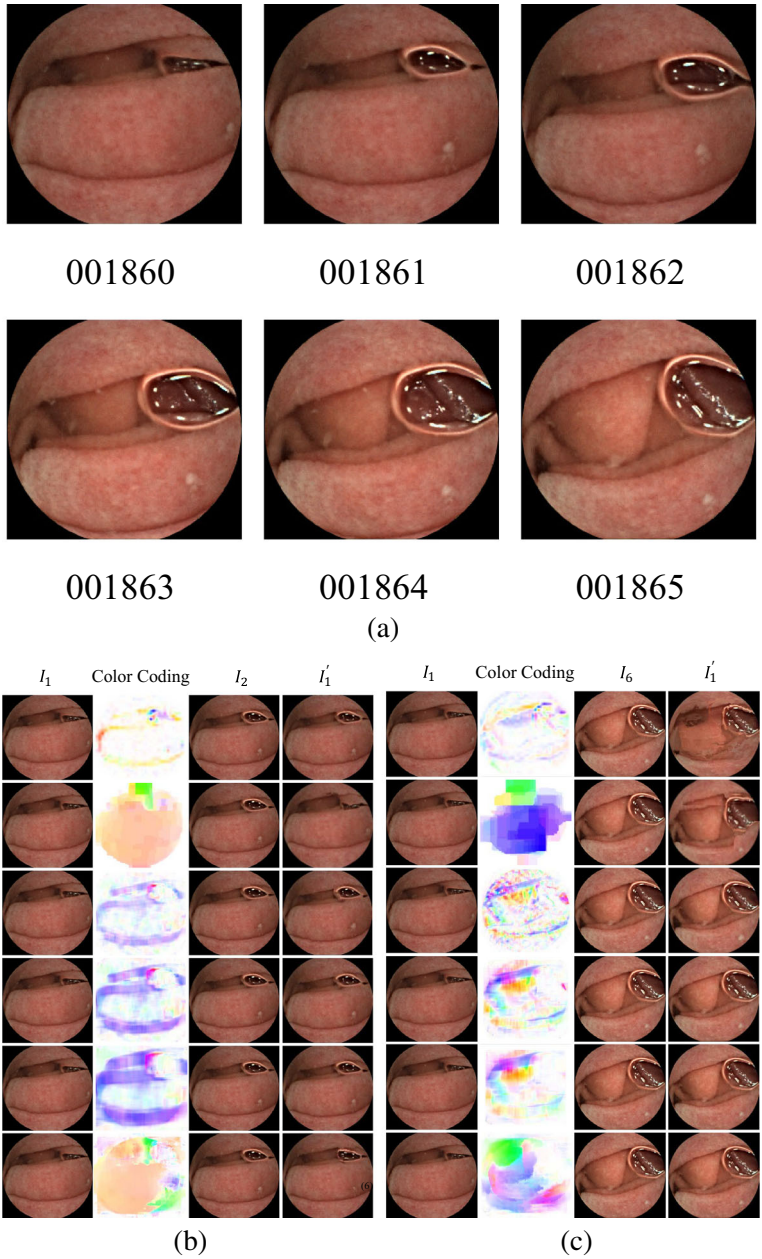


Fig. 5 The visualizations in both flow fields and warping results. (a) Exemplar frames from #001860 to #001865; (b) and (c) Both color coding and warping are conducted between frames #001860 and #001861, between frames #001860 and #001865, respectively. Where each pixel denotes a flow vector, and its hue and saturation represent the orientation and magnitude, respectively. Each result from top to bottom is obtained by Optical Flow (HS [19]), SIFT Flow [42], Video Colorization [59], TimeCycle [60], CorrFlow [30], and SS-VCF

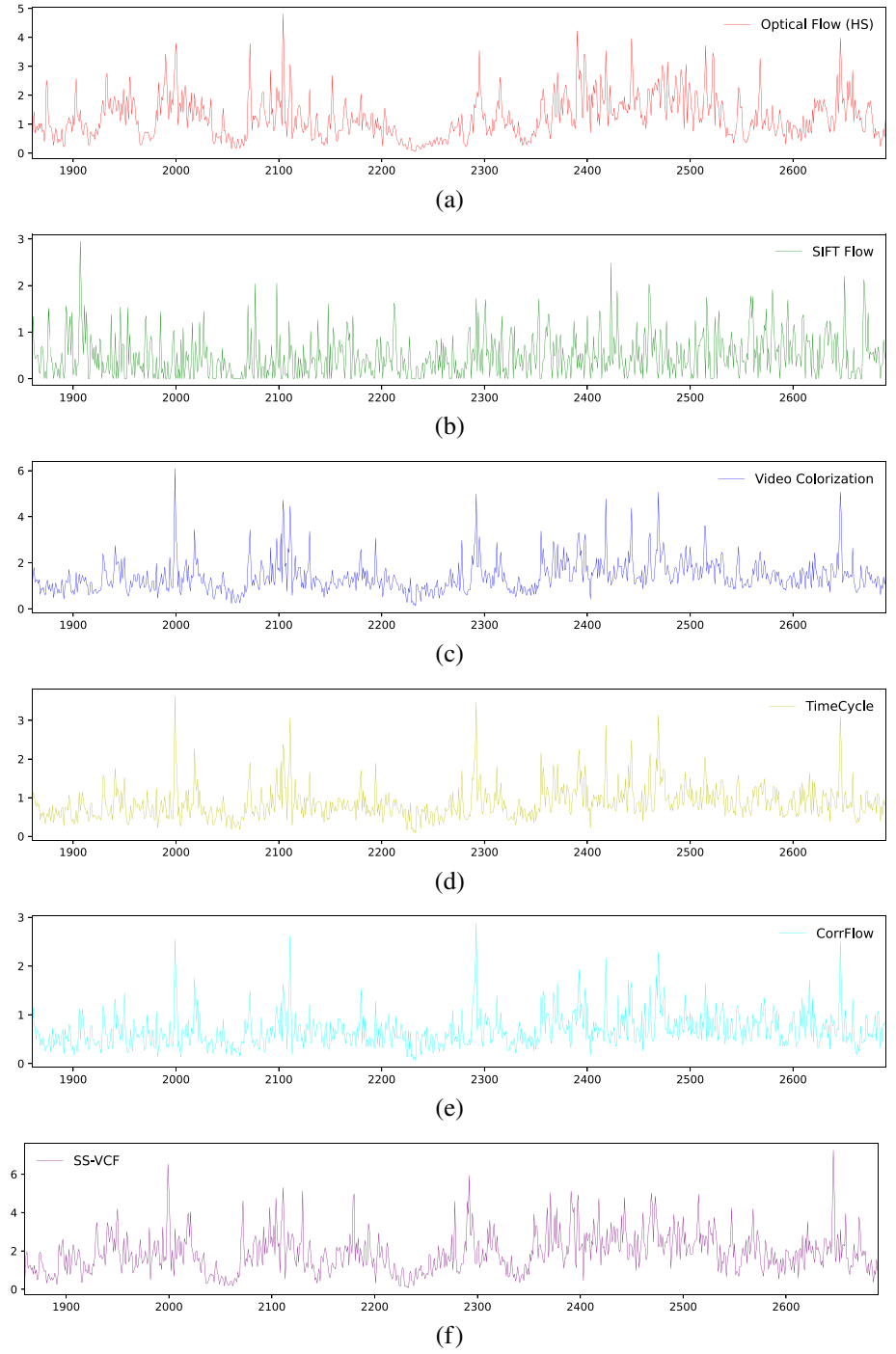
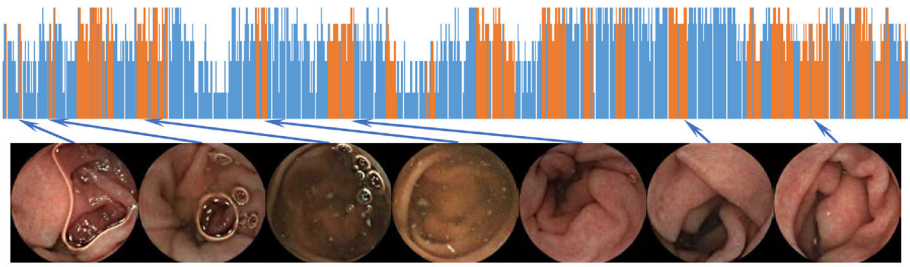


Fig. 6 The curves of motion intensity (MI) of each method. The X-axis and Y-axis indicate frame index and MI value, respectively. The index ranges from 1860 to 2691, totaling 832 frames. (a) Optical Flow (HS [19]), (b) SIFT Flow [42], (c) Video Colorization [59], (d) TimeCycle [60], (e) CorrFlow [30], and (f) SS-VCF

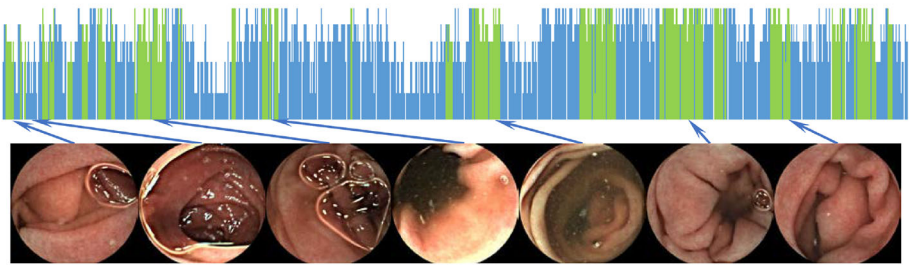
on a sample video #8 of patient ID #2 in WCE-2019-Video dataset, totaling 832 frames. The indices of the frames in video #8 are #001860 to #002691. One can see from Fig. 6 that our method can be suitable for the description of practical motion in WCE video.

4.5.2 Experimental comparisons of SS-VCF-Der and other methods

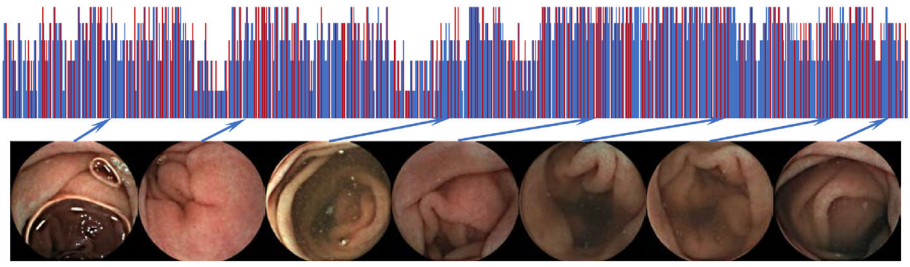
In this subsection, we provide qualitative results to better illustrate how well the SS-VCF-Der scheme selects WCE key frames. Figure 7 demonstrates summarization examples from a sample video #8 of patient ID #2 in WCE-2019-Video dataset, which generated by the three methods including: BAME-SIFTFlow [43], Adv-Ptr-Der-SUM, and SS-VCF-Der. As shown in Fig. 7, our scheme can select some key frames with more distinct scene changes in local neighborhood than the other two methods under the conditions of selecting key frames in Section 3.2.



(a) BAME-SIFTFlow



(b) Adv-Ptr-Der-SUM



(c) SS-VCF-Der

Fig. 7 Exemplar video summaries by three methods. Example summaries from a sample video #8 of patient ID #2 in WCE-2019-Video. The blue bars show the annotation importance scores. The colored segments are the selected subsets using the specified methods

5 Conclusion

In this paper, we propose a self-supervised technique called SS-VCF for learning interframe visual correspondence representations from large amounts of raw WCE videos, and then predicting the flow field. Also, according to the resulting flow field estimation, we compute the motion intensity between two successive frames as extracted motion features, and use our proposed SS-VCF-MI de-redundancy method to select some frames as key ones with distinct scene changes in local neighborhood so as to achieve the task of de-redundancy. Extensive experiments on our collected WCE-2019-Video dataset exhibit that our model and de-redundancy method can achieve a promising result, verifying the effectiveness of our SS-VCF-Der scheme on the visual correspondence representation and redundancy removal for WCE video. As future work, potential extensions can be that: First, seeking for more suitable pretext tasks for self-supervised representation learning of endoscopy video; Second, exploring a combination of motion features and other image features, such as color and texture for a better de-redundancy performance. Our code will be released at https://github.com/lanlbn.

Acknowledgements This work is supported in part by the Scientific Research Foundation of Chongqing University of Technology (0103210650), in part by the National Key Research and Development Program of China (Grant No. 2017YFB0802400), in part by the National Natural Science Foundation of China research fund (61672115), in part by the Chongqing Social Undertakings and Livelihood Security Science and Technology Innovation Project Special Program (cstc2017shmsA30003), and in part by the Humanity and Social Science Youth Foundation, Ministry of Education (Grant No. 17YJCZH043). In addition, we thank Juan Zhou and her colleagues from the Second Affiliated Hospital, Third Military Medical University, for the helpful discussions and suggestions. We also thank the Chongqing Jinshan Science & Technology (Group) Co., Ltd., for providing vital support with raw WCE videos. We would also like to thank the anonymous reviewers for their helpful comments which have led to many improvements in this paper.

Declarations

Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Al-shebani Q, Premaratne P, McAndrew DJ, Vial PJ, Abey S (2019) A frame reduction system based on a color structural similarity (css) method and bayer images analysis for capsule endoscopy. Artif Intell Med 94:18–27. https://doi.org/10.1016/j.artmed.2018.12.008
2. Baker S, Roth S, Scharstein D, Black MJ, Lewis JP, Szeliski R: A database and evaluation methodology for optical ow. In: 2007 IEEE 11th International Conference on Computer Vision, pp 1–8 (2007). https://doi.org/10.1109/ICCV.2007.4408903
3. Beg S, Card T, Sidhu R, Wronska E, Ragunath K, Ching H-L, Koulaouzidis A, Yung D, Panter S, Mcalindon M, Johnson M, Kurup A, Shonde A, San-Juan Acosta M, Sansone S, Simmon E, Thurston V, Healy A, Chetcuti Zammit S, Schembri J, Lau MS, Lam C, Nizamuddin M, Baxter A, Patel J, Archer T, Oppong P, Phillips F, Dorn T, Fateen W, White J, Budihal S, Tan H, Tiwari R (2021) The impact of reader fatigue on the accuracy of capsule endoscopy interpretation. Digestive and Liver Disease 53(8):1028–1033. https://doi.org/10.1016/j.dld.2021.04.024
4. Biniaz A, Zoroo RA, Sohrabi MR (2020) Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis. Biomed Signal Process Control 59:101897. https://doi.org/10.1016/j.bspc.2020.101897
5. Butler DJ, Wul J, Stanley GB, Black MJ: A naturalistic open source movie for optical ow evaluation. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part VI. ECCV'12, pp 611–625. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44

6. Chen J, Zou Y, Wang Y: Wireless capsule endoscopy video summarization: A learning approach based on siamese neural network and support vector machine. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp 1303–1308 (2016). https://doi.org/10.1109/ICPR.2016.7899817
7. Dalal N, Triggs B: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01. CVPR '05, pp 886–893. IEEE Computer Society, USA (2005). https://doi.org/10.1109/CVPR.2005.177
8. Divakaran A, Peker K, Huifang S: A region based descriptor for spatial distribution of motion activity for compressed video. In: Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101), vol 2, pp 287–2902 (2000). https://doi.org/10.1109/ICIP.2000.899359
9. Divakaran A, Sun H: Descriptor for spatial distribution of motion activity for compressed video. In: Storage and Retrieval for Media Databases 2000, vol 3972, pp 392–398. https://doi.org/10.1117/12.373571
10. Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, v. d. Smagt P, Cremers D, Brox T: Flownet: Learning optical ow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2758–2766 (2015). https://doi.org/10.1109/ICCV.2015.316
11. Dray X, Iakovidis D, Houdeville C, Jover R, Diamantis D, Histace A, Koulaouzidis A (2021) Artificial intelligence in small bowel capsule endoscopy - current status, challenges and future promise. J Gastroenterology Hepatology 36(1):12–19. https://doi.org/10.1111/jgh.15341
12. Drozdzal M, Igual L, Vitrià J, Malagelada C, Azpiroz F, Radeva P: Aligning endoluminal scene sequences in wireless capsule endoscopy. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp 117–124 (2010). https://doi.org/10.1109/CVPRW.2010.5543456
13. Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A: Temporal cycle-consistency learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1801–1810 (2019). https://doi.org/10.1109/CVPR.2019.00190
14. Figueiredo IN, Leal C, Pinto L, Figueiredo PN, Tsai R: Dissimilarity measure of consecutive frames in wireless capsule endoscopy videos: A way of searching for abnormalities. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), pp 702–707 (2017). https://doi.org/10.1109/CBMS.2017.18
15. Figueiredo IN, Leal C, Pinto L, Figueiredo PN, Tsai R (2018) Hybrid multiscale affine and elastic image registration approach towards wireless capsule endoscope localization. Biomed Signal Process Control 39:486–502. https://doi.org/10.1016/j.bspc.2017.08.019
16. Fu Y, Liu H, Cheng Y, Yan T, Li T, Meng MQ-: Key-frame selection in wce video based on shot detection. In: Proceedings of the 10th World Congress on Intelligent Control and Automation, pp 5030–5034 (2012). https://doi.org/10.1109/WCICA.2012.6359431
17. Han K, Rezende RS, Ham B, Wong KK, Cho M, Schmid C, Ponce J: Scnet: Learning semantic correspondence. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 1849–1858 (2017). https://doi.org/10.1109/ICCV.2017.203
18. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
19. Horn BKP, Schunck BG (1981) Determining optical ow. Artif Intell 17(1):185–203. https://doi.org/10.1016/0004-3702(81)90024-2
20. Iakovidis DK, Koulaouzidis A (2015) Software for enhanced video capsule endoscopy: challenges for essential progress. Nature Rev Gastroenterology Hepatology 12:172–186. https://doi.org/10.1038/nrgastro.2015.13
21. Iakovidis DK, Tsevas S, Polydorou A (2010) Reduction of capsule endoscopy reading times by unsupervised image mining. Comput Med Imag Graph 34(6):471–478. https://doi.org/10.1016/j.compmedimag.2009.11.005
22. Iddan G, Meron G, Glukhovsky A, Swain P (2000) Wireless capsule endoscopy. Nature 405(6785):417–417. https://doi.org/10.1038/35013140
23. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T: Flownet 2.0: Evolution of optical ow estimation with deep networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1647–1655 (2017). https://doi.org/10.1109/CVPR.2017.179
24. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K: Spatial transformer networks (2015) arXiv:1506.02025
25. Jani KK, Srivastava R (2019) A survey on medical image analysis in capsule endoscopy. Current Med Imag Rev 15(7):622–636. https://doi.org/10.2174/1573405614666181102152434
26. Karargyris A, Bourbakis N: A video-frame based registration using segmentation and graph connectivity for wireless capsule endoscopy. In: 2009 IEEE/NIH Life Science Systems and Applications Workshop, pp 74–79 (2009). https://doi.org/10.1109/LISSA.2009.4906713

27. Kim S, Min D, Ham B, Lin S, Sohn K (2019) Fcss: Fully convolutional self-similarity for dense semantic correspondence. IEEE Trans Pattern Anal Mach Intell 41(3):581–595. https://doi.org/10.1109/TPAMI.2018.2803169
28. Kingma DP, Ba J: Adam: A method for stochastic optimization (2014) arXiv:1412.6980
29. Koulaouzidis A, Dabos K, Philipper M, Toth E, Keuchel M (2021) How should we do colon capsule endoscopy reading: a practical guide. Therapeutic Advances in Gastrointestinal Endoscopy 14:26317745211001984. https://doi.org/10.1177/26317745211001983. (**PMID: 33817637**)
30. Lai Z, Xie W: Self-supervised learning for video correspondence ow (2019) arXiv:1905.00875
31. Lan L, Ye C: Recurrent generative adversarial networks for unsupervised wce video summarization. Knowledge-Based Systems, 106971 (2021). https://doi.org/10.1016/j.knosys.2021.106971
32. Larsson G, Maire M, Shakhnarovich G: Colorization as a proxy task for visual understanding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 840–849. IEEE Computer Society, Los Alamitos, CA, USA (2017). https://doi.org/10.1109/CVPR.2017.96
33. Lee H-G, Choi M-K, Shin B-S, Lee S-C (2013) Reducing redundancy in wireless capsule endoscopy videos. Comput Biology Med 43(6):670–682. https://doi.org/10.1016/j.compbiomed.2013.02.009
34. Lee J, Kim D, Ponce J, Ham B: Sfnet: Learning object-aware semantic correspondence. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2273–2282 (2019). https://doi.org/10.1109/CVPR.2019.00238
35. Li C, Hamza AB, Bouguila N, Wang X, Ming F, Xiao G (2014) Online redundant image elimination and its application to wireless capsule endoscopy. Signal Imag Video Process 8(8):1497–1506. https://doi.org/10.1007/s11760-012-0384-3
36. Liao C, Wang C, Bai J, Lan L, Wu X (2021) Deep learning for registration of region of interest in consecutive wireless capsule endoscopy frames. Comput Method Prog Biomed 208:106189. https://doi.org/10.1016/j.cmpb.2021.106189
37. Lien G, Liu C, Jiang J, Chuang C, Teng M (2012) Magnetic control system targeted for capsule endoscopic operations in the stomach|design, fabrication, and in vitro and ex vivo evaluations. IEEE Trans Biomed Eng 59(7):2068–2079. https://doi.org/10.1109/TBME.2012.2198061
38. Li S, Han K, Costain TW, Howard-Jenkins H, Prisacariu V: Correspondence networks with adaptive neighbourhood consensus. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10193–10202 (2020). https://doi.org/10.1109/CVPR42600.2020.01021
39. Li B, Meng MQ-, Hu C: Motion analysis for capsule endoscopy video segmentation. In: 2011 IEEE International Conference on Automation and Logistics (ICAL), pp 46–51 (2011). https://doi.org/10.1109/ICAL.2011.6024682
40. Li B, Meng MQ-, Zhao Q: Wireless capsule endoscopy video summary. In: 2010 IEEE International Conference on Robotics and Biomimetics, pp 454–459 (2010). https://doi.org/10.1109/ROBIO.2010.5723369
41. Li B, Meng MQ-: Capsule endoscopy video boundary detection. In: 2011 IEEE International Conference on Information and Automation, pp 373–378 (2011). https://doi.org/10.1109/ICINFA.2011.5949020
42. Liu C, Yuen J, Torralba A (2011) Sift ow: Dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994. https://doi.org/10.1109/TPAMI.2010.147
43. Liu H, Pan N, Lu H, Song E, Wang Q, Hung CC (2013) Wireless capsule endoscopy video reduction based on camera motion estimation. J Digital Imag. https://doi.org/10.1007/s10278-012-9519-x
44. Liu X, Lee J, Jin H: Learning video representations from correspondence proposals. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4268–4276 (2019). https://doi.org/10.1109/CVPR.2019.00440
45. Li H, Zhang Y, Yang M, Men Y, Chao H: A rapid abnormal event detection method for surveillance video based on a novel feature in compressed domain of hevc. In: 2014 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6 (2014). https://doi.org/10.1109/ICME.2014.6890212
46. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
47. Lucas BD, Kanade T: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'81, pp 674–679. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1981). https://doi.org/10.5555/1623264.1623280
48. Mahasseni B, Lam M, Todorovic S: Unsupervised video summarization with adversarial lstm networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2982–2991 (2017). https://doi.org/10.1109/CVPR.2017.318
49. Meister S, Hur J, Roth S: Unflow: Unsupervised learning of optical flow with a bidirectional census loss (2017) arXiv:1711.07837

50. Muhammad K, Khan S, Kumar N, Del Ser J, Mirjalili S (2020) Visionbased personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges. Future Generation Comput Syst 113:266–280. https://doi.org/10.1016/j.future.2020.06.048
51. Nie R, Yang H, Peng H, Luo W, Fan W, Zhang J, Liao J, Huang F, Xiao Y: Application of Structural Similarity Analysis of Visually Salient Areas and Hierarchical Clustering in the Screening of Similar Wireless Capsule Endoscopic Images. arXiv e-prints, 2004–02805 (2020) arXiv:2004.02805 [eess.IV]
52. Paszke A, am Gross, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A: Automatic di erentiation in pytorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp 1–4 (2017)
53. Rocco I, Arandjelović R, Sivic J (2019) Convolutional neural network architecture for geometric matching. IEEE Trans Pattern Anal Mach Intell 41(11):2553–2567. https://doi.org/10.1109/TPAMI.2018.2865351
54. Rondonotti E, Pennazio M, Toth E, Koulaouzidis A (2020) How to read small bowel capsule endoscopy: a practical guide for everyday use. Endoscopy Int Open 8(10):1220–1224. https://doi.org/10.1055/a-1210-4830
55. Schoeffmann K, Fabro MD, Szkaliczki T, aszlo Böszörmenyi, Keckstein J (2015) Keyframe extraction in endoscopic video. J Multimed Tools Appl 74:11187–11206. https://doi.org/10.1007/s11042-014-2224-7
56. Spyrou E, Iakovidis DK (2013) Video-based measurements for wireless capsule endoscope tracking. Measure Sci Technol 25(1):015002. https://doi.org/10.1088/0957-0233/25/1/015002
57. Spyrou E, Diamantis D, Iakovidis DK: Panoramic visual summaries for efficient reading of capsule endoscopy videos. In: 2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization, pp 41–46 (2013). https://doi.org/10.1109/SMAP.2013.21
58. Sushma B, Aparna P (2021) Summarization of wireless capsule endoscopy video using deep feature matching and motion analysis. IEEE Access 9:13691–13703. https://doi.org/10.1109/ACCESS.2020.3044759
59. Vondrick C, Shrivastava A, Fathi A, Guadarrama S, Murphy K: Tracking emerges by colorizing videos. In: Computer Vision - ECCV 2018, pp 402–419. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_24
60. Wang X, Jabri A, Efros AA: Learning correspondence from the cycleconsistency of time. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2561–2571 (2019). https://doi.org/10.1109/CVPR.2019.00267
61. Wang N, Song Y, Ma C, Zhou W, Liu W, Li H: Unsupervised deep tracking. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1308–1317 (2019). https://doi.org/10.1109/CVPR.2019.00140
62. Xu Y, Li K, Zhao Z, Meng MQ-: A novel system for closed-loop simultaneous magnetic actuation and localization of wce based on external sensors and rotating actuation. IEEE Trans Autom Sci Eng, 1–13 (2020). https://doi.org/10.1109/TASE.2020.3013954
63. Yuan Y, Meng MQ-: Hierarchical key frames extraction for wce video. In: 2013 IEEE International Conference on Mechatronics and Automation, pp 225–229 (2013). https://doi.org/10.1109/ICMA.2013.6617922
64. Zhang K, Chao W, Sha F, Grauman K: Video summarization with long short-term memory (2016) arXiv:1605.08110
65. Zhang R, Isola P, Efros AA: Colorful image colorization. In: Computer Vision - ECCV 2016, pp 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40

Authors and Affiliations

Libin Lan[1,2,3] · Chunxiao Ye[2,3] · Chao Liao[2,3] · Chengliang Wang[2,3] · Xin Feng[1]

Libin Lan
lanlbn@cqu.edu.cn

Chao Liao
liaochaocqu@outlook.com

Chengliang Wang
wangcl@cqu.edu.cn

Xin Feng
xfeng@cqut.edu.cn

1 College of Computer Science and Engineering, Chongqing University of Technology, No.69 Hongguang Avenue, Banan District, Chongqing 400054, Chongqing, China

2 College of Computer Science, Chongqing University, Chongqing 400044, China

3 Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, No.174 ShaZheng Street, ShaPingBa District, Chongqing 400044,
Chongqing, China