VT-ReID: Learning Discriminative Visual-Text Representation for Polyp Re-Identification

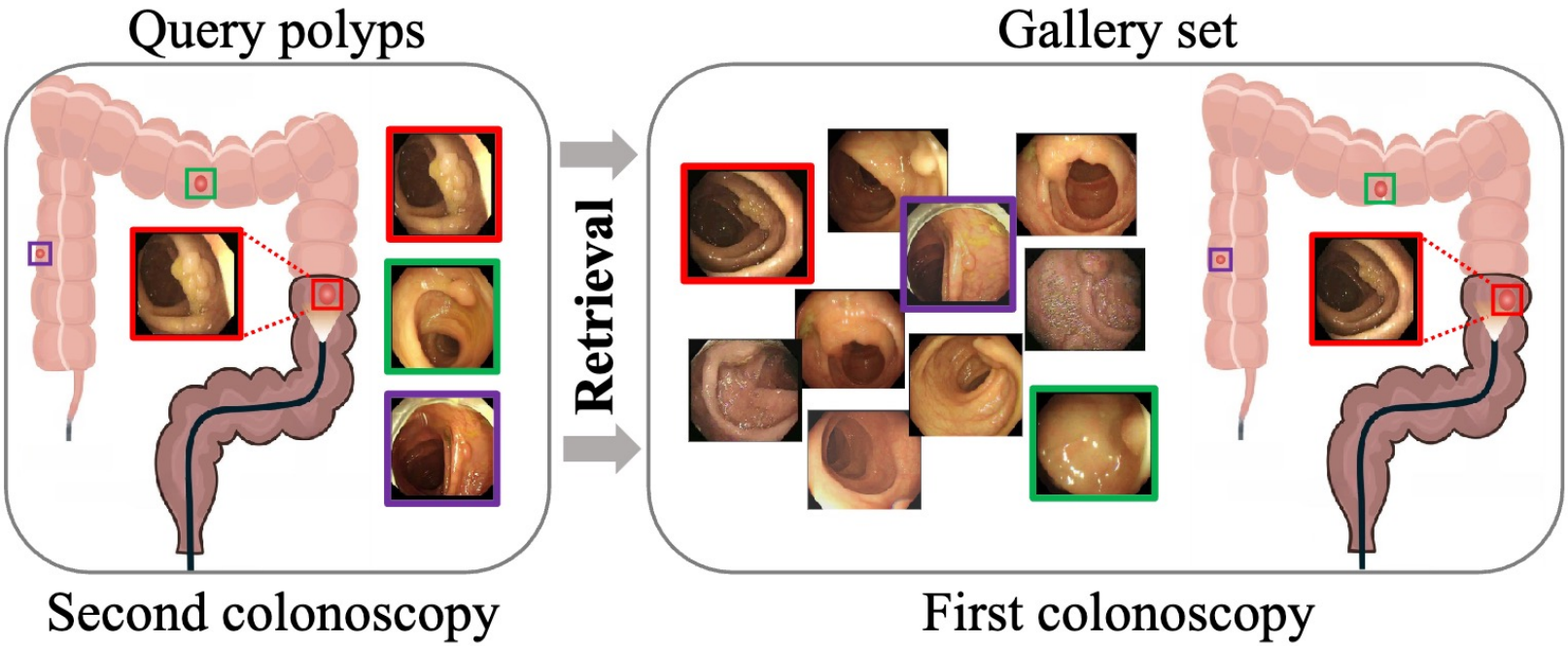
Suncheng Xiang[1]*, Cang Liu[2], Jiacheng Ruan[3], Shilun Cai[1], Sijia Du[1], Dahong Qian[1]

[1]School of Biomedical Engineering, Shanghai Jiao Tong University
[2]National Key Laboratory of Electromagnetic Energy, Naval University of Engineering
[3]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[4]Zhongshan Hospital of Fudan University

https://JeremyXSC.github.io/
https://github.com/JeremyXSC/VT-ReID

Introduction

Task: Colonoscopic Polyp Re-Identification

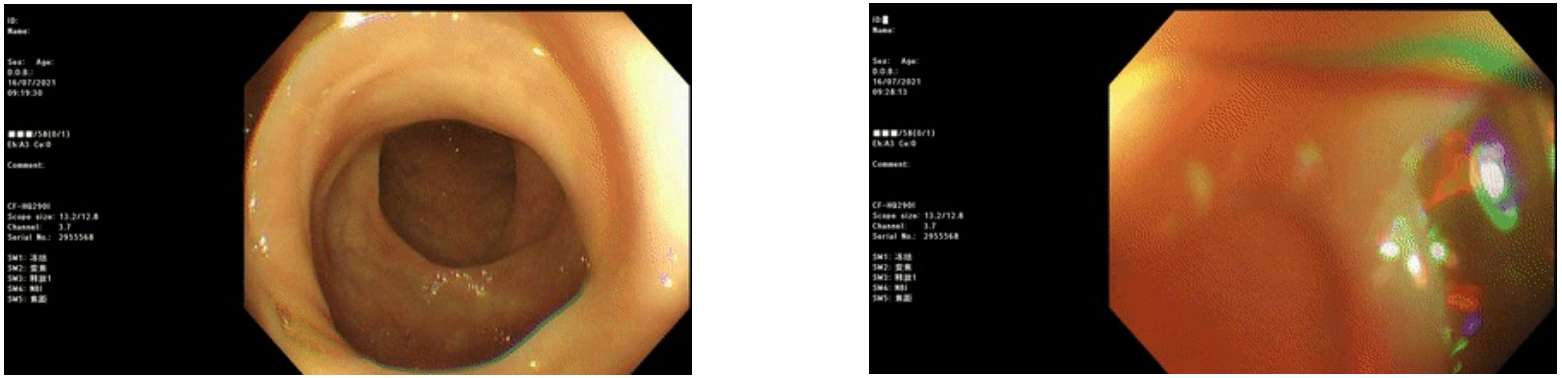


Introduction

Similar to Person Re-Identification

https://arxiv.org/pdf/1610.02984.pdf

Challenge



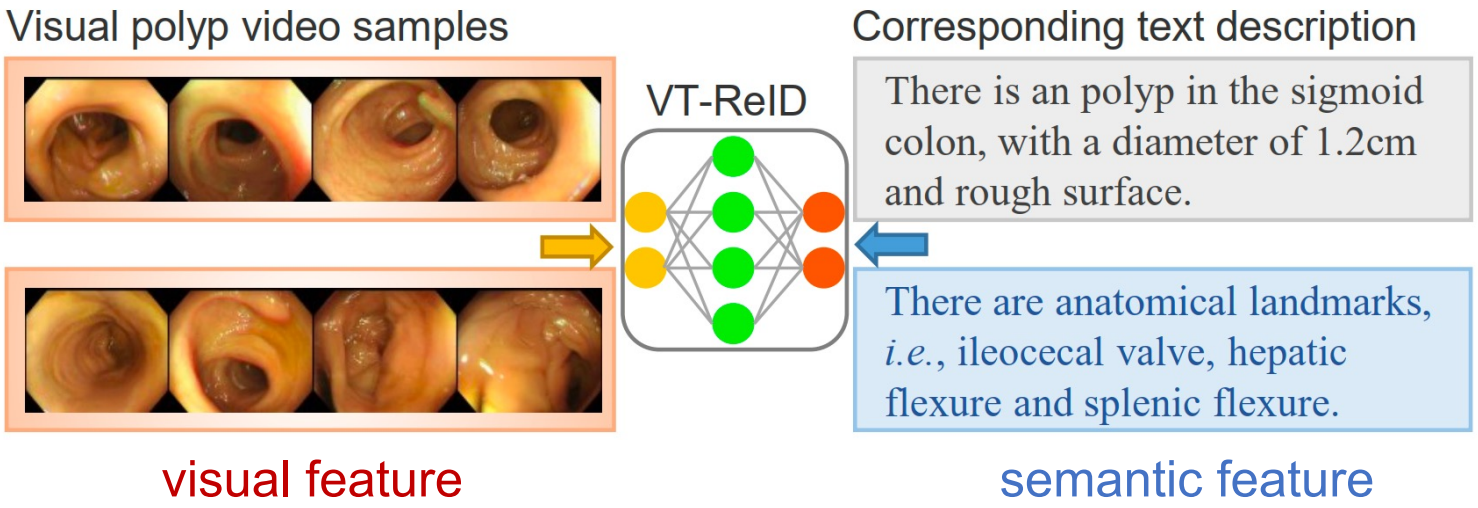
Same polyp video with different viewpoints

Background

- Manually labeling pairwise polyp area data is labor-intensive
- Existing methods rely heavily on visual feature
- Semantic information is always ignored during training

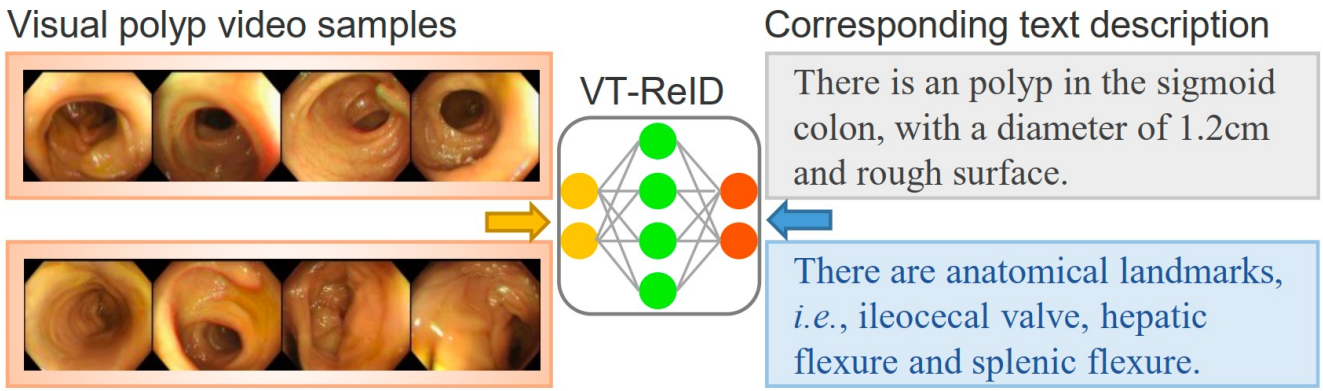
Motivation

- Compared to the conventional person ReID, Polyp ReID is confronted with more challenges, such as variation in terms of backgrounds, viewpoint, and illumination, etc.
- The performance deeply relies on the visual feature of training dataset, other rich information in semantic level is always ignored.

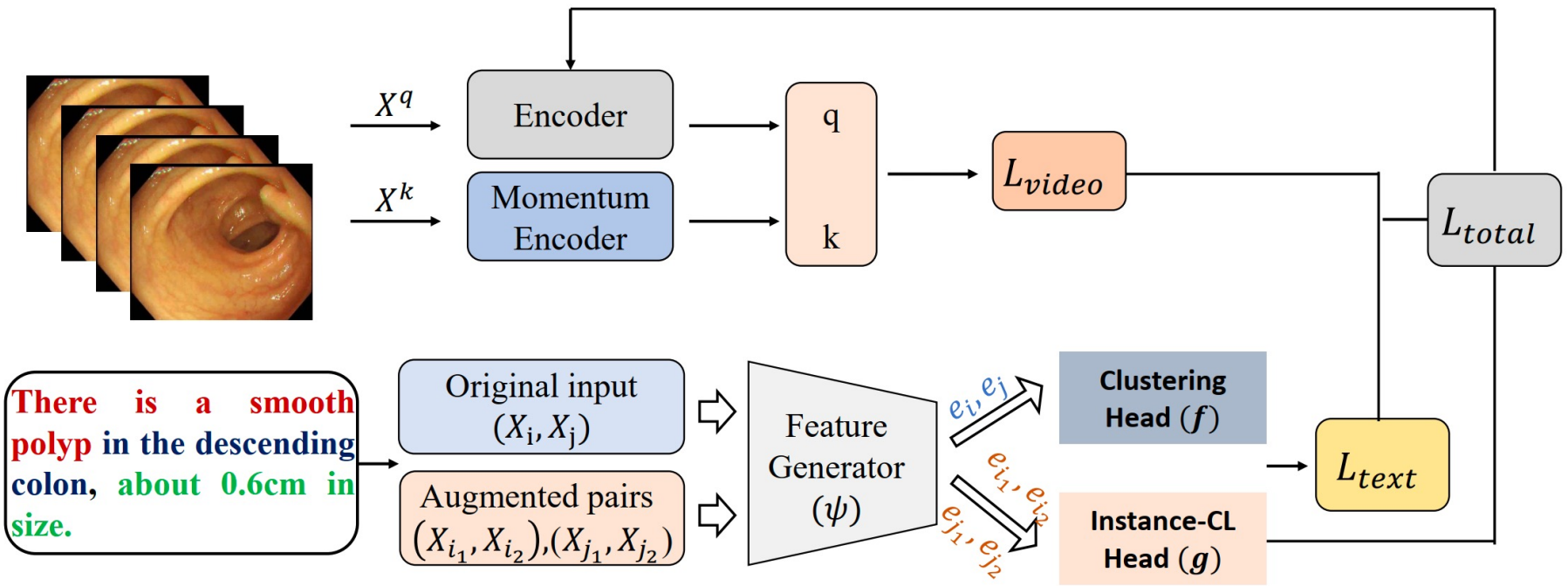


Contribution

- A VT-ReID method is proposed to help model learn general visual-text representation based on the multimodal feature.
- Based on it, a dynamic clustering mechanism is introduced to further enhance the clustering performance of text data in an unsupervised manner.
- Comprehensive experiments demonstrate the effectiveness of our method, surpassing existing methods with a clear margin.



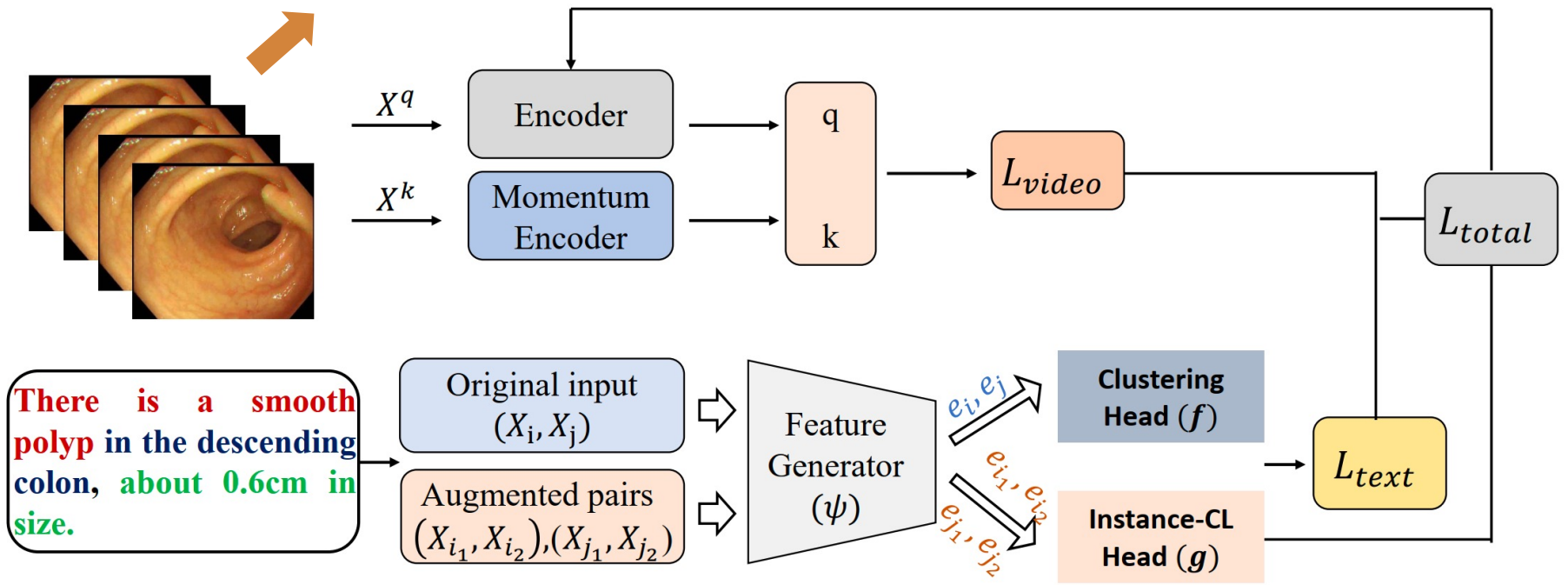
Method



Our VT-ReID Network architecture

Method

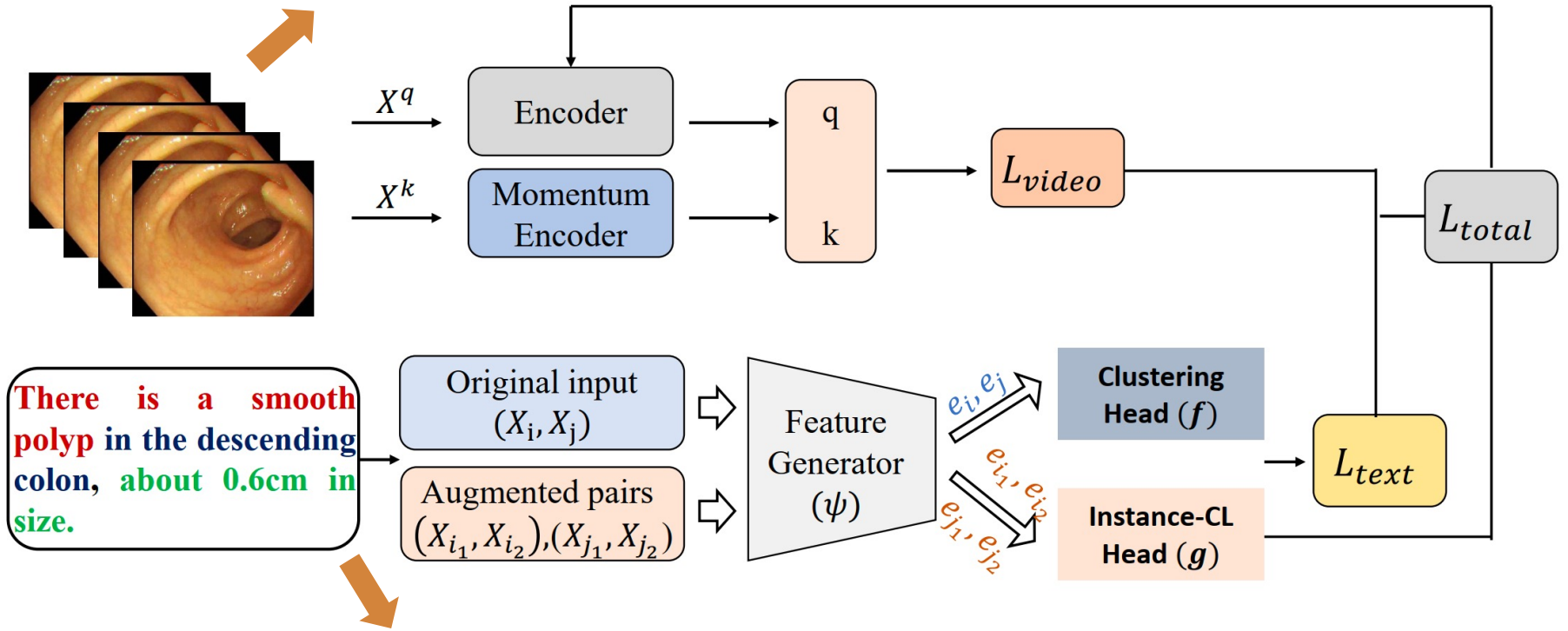
Visual feature backbone is composed of a vision transformer.



Our VT-ReID Network architecture

Method

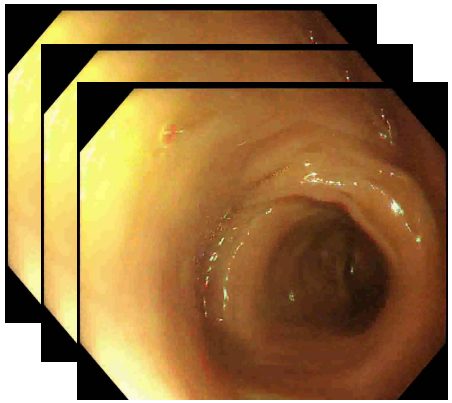
Visual feature backbone is composed of a vision transformer.



Texture feature backbone is composed of deep CNN

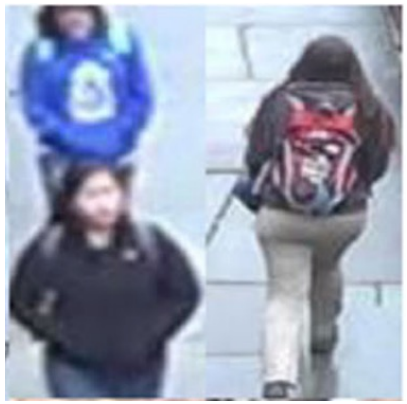
Implementation Details

- Vision-Transformer is regarded as the backbone.
- Random flipping and random cropping for data augmentation.
- L_{video} and L_{text} functions to train the model for ***180*** iterations.
- Nvidia GeForce RTX 2080Ti GPU & Intel Xeon Gold 6130T CPU.



Colo-Pair

Market-1501



DukeMTMC-reID

CUHK03

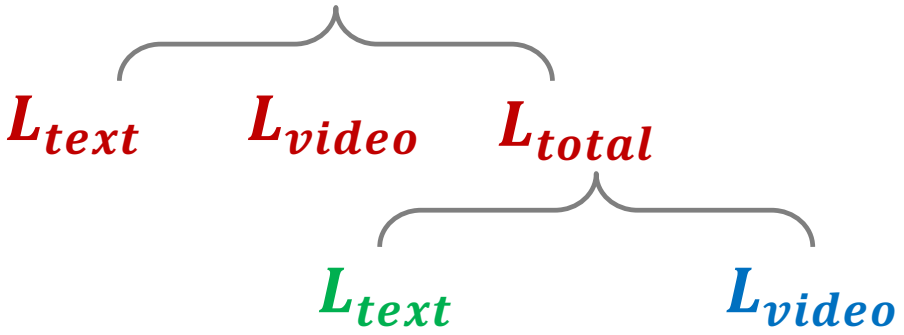
Polyp Re-ID

Person Re-ID

Ablation Study

Methods	Text data	Image data	mAP↑	Rank-1↑	Rank-5↑
Baseline	×	×	25.9	17.5	37.1
w/ $\mathcal{L}_{text}$	✓	×	27.3	17.4	38.4
w/ $\mathcal{L}_{video}$	×	✓	31.6	21.7	41.2
w/ $\mathcal{L}_{total}$	✓	✓	37.9	23.4	44.5

Ablation study of different components of our VT-ReID.



Comparison with SOTAs

Method	Venue	Video Retrieval ↑			
		mAP	Rank-1	Rank-5	Rank-10
ViSiL [17]	ICCV 19	24.9	14.5	30.6	51.6
CoCLR [18]	NIPS 20	16.3	6.5	22.6	33.9
TCA [19]	WCAV 21	27.8	16.1	35.5	53.2
ViT [15]	CVPR 21	20.4	9.7	30.6	43.5
CVRL [20]	CVPR 21	23.6	11.3	32.3	53.2
CgS^c [16]	IJCV 22	21.4	8.1	35.5	45.2
$FgAttS^f_A$ [16]	IJCV 22	23.6	9.7	40.3	50.0
$FgBinS^f_B$ [16]	IJCV 22	21.2	9.7	32.3	48.4
Colo-SCRL [4]	ICME 23	<u>31.5</u>	<u>22.6</u>	<u>41.9</u>	<u>58.1</u>
VT-ReID	**Ours**	**37.9**	**23.4**	**44.5**	**60.1**

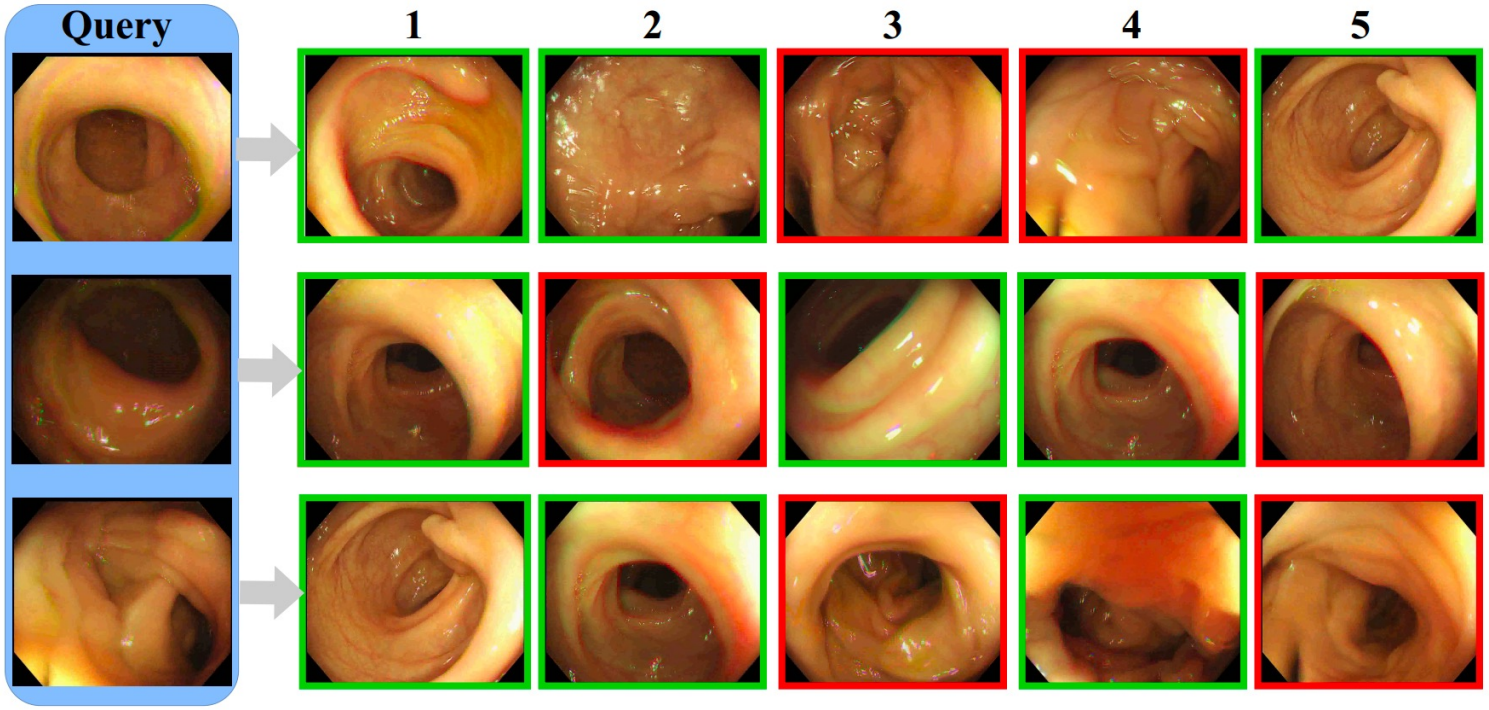
Performance comparison with SOTAs on Colo-Pair dataset

Comparison with SOTAs

Method	Market-1501		DukeMTMC-reID		CUHK03	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
MHN [23]	85.0	95.1	77.2	77.3	76.5	71.7
CBDB [24]	85.0	94.4	74.3	87.7	72.8	75.4
C2F [21]	87.7	<u>94.8</u>	74.9	87.4	<u>84.1</u>	81.3
MGN [25]	86.9	**95.7**	78.4	88.7	66.0	66.8
SCSN [22]	**88.3**	92.4	<u>79.0</u>	<u>91.0</u>	81.0	<u>84.7</u>
VT-ReID	<u>88.1</u>	93.8	**79.2**	**92.6**	**85.3**	**88.3**

Performance comparison with SOTAs on Person ReID dataset

Qualitative Analysis



Top-5 ranking list for some query images on Colo-Pair dataset

From left to right, a query image, a **true match**, and a **false match (distractor).**

Conclusion and Future Work

- We propose a simple but effective multimodal training method VT-ReID.
- A dynamic clustering-based mechanism called DCM is introduced to further boost the performance of colonoscopic polyp ReID task.
- Comprehensive experiments also demonstrate the effectiveness of our method.

License MIT python Star 3

Open-sourced !

Learning Discriminative Visual-Text Representation for Polyp Re-Identification

Introduction

In this work, we propose a simple but effective training method named \textbf{VT-ReID}, which can remarkably enrich the representation of polyp videos with the interchange of high-level semantic information. Moreover, we elaborately design a novel clustering mechanism to introduce prior knowledge from textual data, which leverages contrastive Learning to promote better separation from abundant unlabeled text data.

THANKS