Motion-Aware Graph Reasoning Hashing for Self-supervised Video Retrieval

Ziyun Zeng[1,3,#], Jinpeng Wang[1,3], Bin Chen[2,3,*], Yuting Wang[1,3], Shu-Tao Xia[1,3]

[1]Tsinghua Shenzhen International Graduate School
[2]Harbin Institute of Technology, Shenzhen
[3]Research Center of Artificial Intelligence, Peng Cheng Laboratory

#zengzy21@mails.tsinghua.edu.cn *chenbin2021@hit.edu.cn

Abstract

Unsupervised video hashing aims to learn a nonlinear hashing function to map videos into a similarity-preserving hamming space without label supervision. In this paper, we propose *Motion-Aware Graph Reasoning Hashing* (MAGRH), an end-to-end framework that utilizes the motion information explicitly while keeping inference efficiency. Specifically, we design a dual-branch architecture consisting of a main and an auxiliary branch. The main (auxiliary) branch receives frame-wise (clip-wise) inputs and produces general (motion) hash codes via graph reasoning modules. We develop a combination of intra- and inter-branch objectives to simultaneously learn branch-specific hashing functions as well as transfer motion knowledge from the auxiliary branch to the main branch. In inference, the auxiliary branch is removed. Benefiting from motion guidance, our MAGRH yields superior performance on FCVID and ActivityNet, even with a small frame rate.

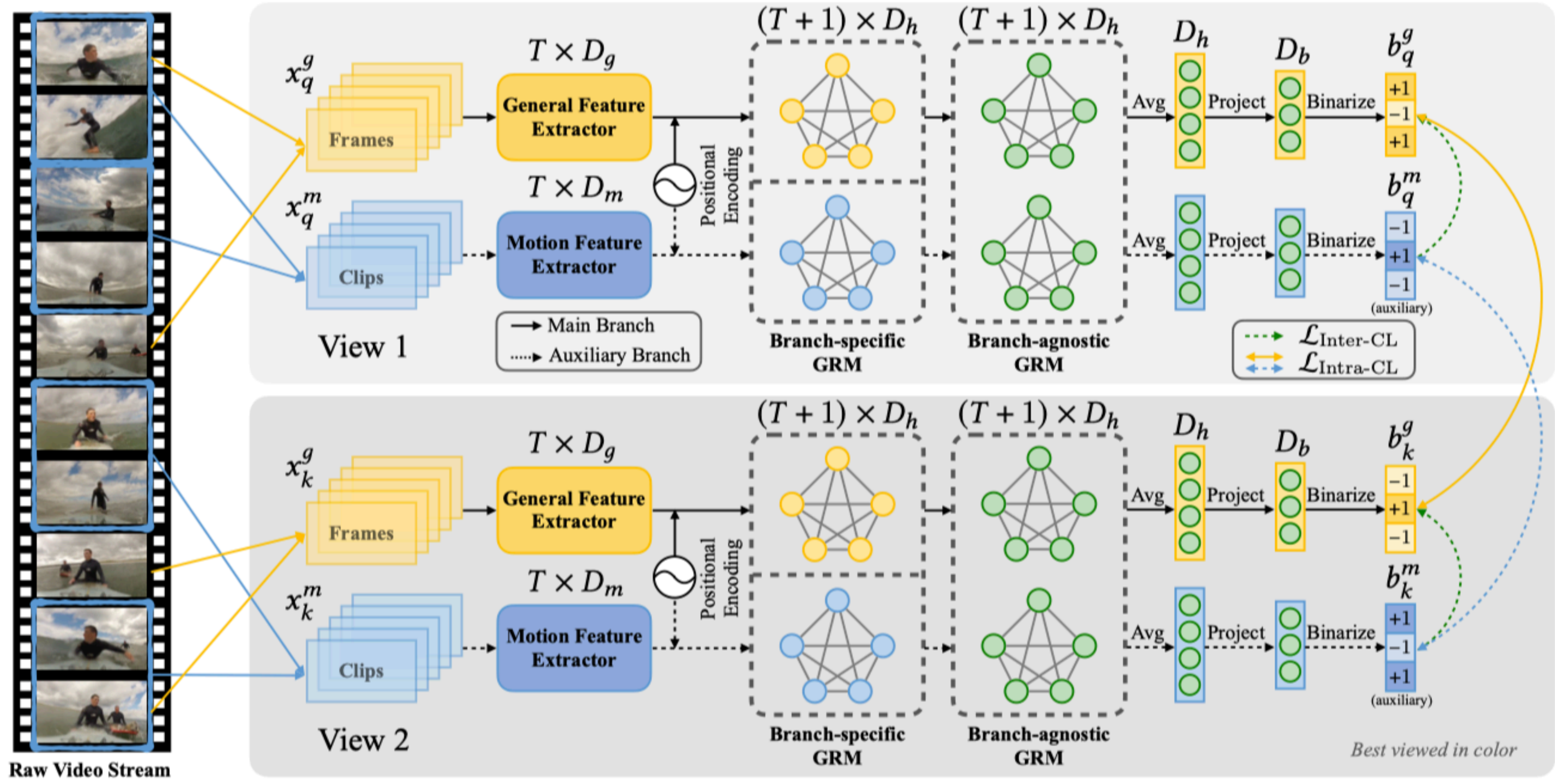


Figure 1: The framework of MAGRH. We first sample two sets of frames and two sets of clips to construct two correlated views. Each view consists of two branches: **(i)** A main branch to encode frame-wise inputs and produce general hash codes. **(ii)** An auxiliary branch to encode clip-wise inputs and produce motion hash codes. We develop two contrastive objectives to learn better representations within each branch and transfer motion knowledge from the auxiliary branch to the main branch. In inference, the auxiliary branch is removed.

Learning Objective

$$\mathcal{L}_{\text{CL}}^{*} = -\sum_{b_q^*, b_k^* \in \mathcal{B}} \log \frac{\exp(s(b_q^*, b_k^*)/\tau)}{\exp(s(b_q^*, b_k^*)/\tau) + \sum_{b_{k-}^* \in \mathcal{B} \setminus \{b_q^*, b_k^*\}} \exp(s(b_q^*, b_{k-}^*)/\tau)}$$

$$\hat{\mathcal{L}}_{\text{CL}}^{*} = -\sum_{b_*^g, b_*^m \in \mathcal{B}} \log \frac{\exp(s(b_*^g, b_*^m)/\hat{\tau})}{\exp(s(b_*^g, b_*^m)/\hat{\tau}) + \sum_{b_{*-}^m \in \mathcal{B} \setminus \{b_*^g, b_*^m\}} \exp(s(b_*^g, b_{*-}^m)/\hat{\tau})}$$

$$\mathcal{L}_{\text{Intra-CL}} = \frac{1}{2}\left(\mathcal{L}_{\text{CL}}^{g} + \mathcal{L}_{\text{CL}}^{m}\right) \quad \mathcal{L}_{\text{Inter-CL}} = \frac{1}{2}\left(\hat{\mathcal{L}}_{\text{CL}}^{q} + \hat{\mathcal{L}}_{\text{CL}}^{k}\right) \quad \mathcal{L}_{\text{reg}} = \sum_{i \in \{q,k\}} \sum_{j \in \{g,m\}} \|b_i^j - h_i^j\|_2^2$$

$$\mathcal{L}_{\text{MAGRH}} = \mathcal{L}_{\text{Intra-CL}} + \gamma \mathcal{L}_{\text{Inter-CL}} + \mathcal{L}_{\text{reg}}$$

Graph Reasoning and Hash Layers

$$\hat{x}^* = [\varphi_{agg}^*;\ \phi^*(x^*) + z_p + z_m^*] \in \mathbb{R}^{(T+1)\times D_h},\ * \in \{g,m\}$$

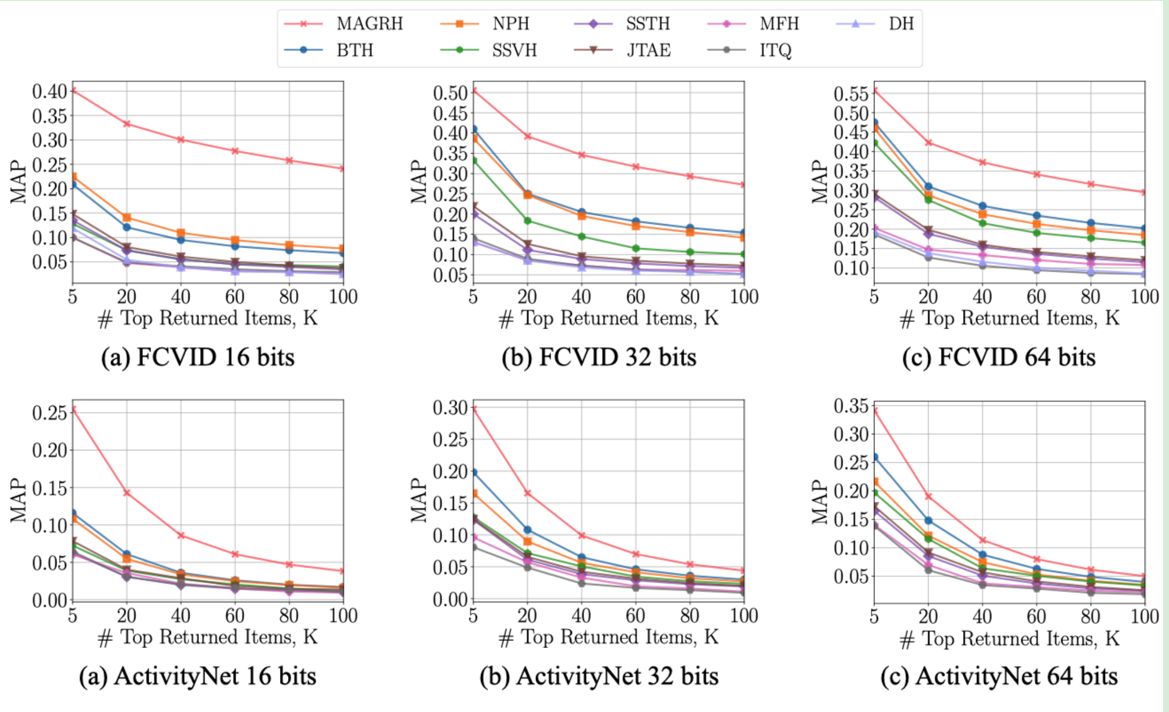
$$A = g_1(\hat{x})^T g_2(\hat{x}), \quad A_{i,j} \to \frac{A_{i,j}^2}{\sum_{t=1}^{T+1} A_{i,t}^2}$$

$$\hat{x}_l = \text{ReLU}(\text{LN}(A\hat{x}_{l-1}W_l)), l \in \{1, \cdots, L\}$$

$$\bar{x}^* = \frac{1}{T+1}\sum_{i=1}^{T+1} \hat{x}_{L,i}^* \in \mathbb{R}^{D_h},\ * \in \{g,m\} \quad \begin{array}{l} h^* = \Phi^*(\bar{x}^*) \in \mathbb{R}^{D_b},\ * \in \{g,m\} \\ b^* = \text{sgn}(h^*) \in \{-1,1\}^{D_b},\ * \in \{g,m\} \end{array}$$

Experiments

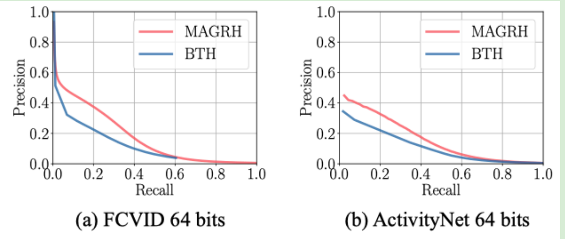
We compared the proposed MAGRH with: BTH, NPH, SSVH, SSTH, JTAE, MFH, ITQ, and DH. Our MAGRH outperforms all previous methods under all code lengths by a large margin in terms of MAP@K.



(a) FCVID 16 bits (b) FCVID 32 bits (c) FCVID 64 bits

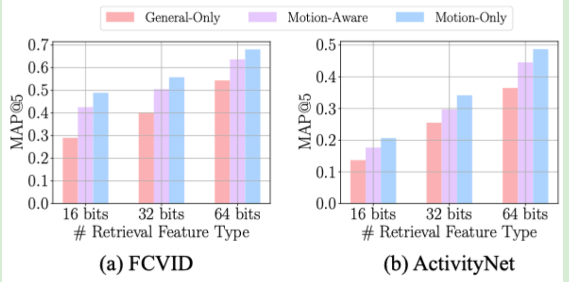
(a) ActivityNet 16 bits (b) ActivityNet 32 bits (c) ActivityNet 64 bits

Figure 2: The MAP@K results of different methods under 16, 32, and 64 bits.



(a) FCVID 64 bits (b) ActivityNet 64 bits

Figure 3: The PR curve of MARGH and BTH under 64 bits.



(a) FCVID (b) ActivityNet

Figure 4: The MAP@5 results *w.r.t.* different training protocols.