Hybrid Relation Guided Set Matching for Few-shot Action Recognition Supplementary Materials

Xiang Wang[1] Shiwei Zhang[2*] Zhiwu Qing[1] Mingqian Tang[2] Zhengrong Zuo[1]
Changxin Gao[1] Rong Jin[2] Nong Sang[1*]
[1]Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group
{wxiang,qzw,zhrzuo,cgao,nsang}@hust.edu.cn, {zhangjin.zsw,mingqian.tmq,jinrong.jr}@alibaba-inc.com

Setting	Dataset	1-shot	5-shot
Support-only	SSv2-Full	52.1	67.2
Support&Query (ours)		**54.3**	**69.0**
Support-only	Kinetics	73.4	85.5
Support&Query (ours)		**73.7**	**86.1**

Table 1. Performance comparison with different relation modeling paradigms on SSv2-Full and Kinetics.

A. Splits of Epic-kitchens

Epic-kitchens [5] is a large-scale first-view dataset and contains diverse unedited object interactions in kitchens. In our experiment, we divide the dataset according to the verbs of the actions.

Meta-training set: 'take', 'put-down', 'open', 'turn-off', 'dry', 'hand', 'tie', 'remove', 'cut', 'pull-down', 'shake', 'drink', 'move', 'lift', 'stir', 'adjust', 'crush', 'taste', 'check', 'drain', 'sprinkle', 'empty', 'knead', 'spread-in', 'scoop', 'add', 'push', 'set-off', 'wear', 'fill', 'turn-down', 'measure', 'scrape', 'read', 'peel', 'smell', 'plug-in', 'flip', 'turn', 'enter', 'unscrew', 'screw-in', 'tap-on', 'break', 'fry', 'brush', 'scrub', 'spill', 'separate', 'immerse', 'rub-on', 'lower', 'stretch', 'slide', 'use', 'form-into', 'oil', 'sharpen', 'touch', 'let'.

Meta-testing set: 'wash', 'squeeze', 'turn-on', 'throw-in', 'close', 'put-into', 'fold', 'unfold', 'pour', 'tear', 'look-for', 'hold', 'roll', 'arrange', 'spray', 'wait', 'collect', 'turn-up', 'grate', 'wet'.

Note that there is no overlap between the meta-training set and the meta-testing set.

B. Other relation modeling forms

Previous few-shot image classification methods of learning task-specific features have also achieved promising results [11, 20]. However, many of them use some complex and fixed operations to learn the dependencies between images, while our method is greatly simple and flexible. Moreover, most previous works only use the information within the support set to learn task-specific features, ignoring the correlation with query samples. In our hybrid relation module, we add the query video to the pool of inter-relation modeling to extract relevant information suitable for query classification. As illustrated in Table 1, we try to remove the query video from the pool, *i.e.*, *Support-only*, but we can observe that after removing the query video, the performance of 1-shot and 5-shot on SSv2-Full reduces by 2.2% and 1.8%, respectively. There are similar conclusions on the Kinetics dataset. This evidences that the proposed hybrid relation module is reasonable and can effectively extract task-related features, thereby promoting query classification performance.

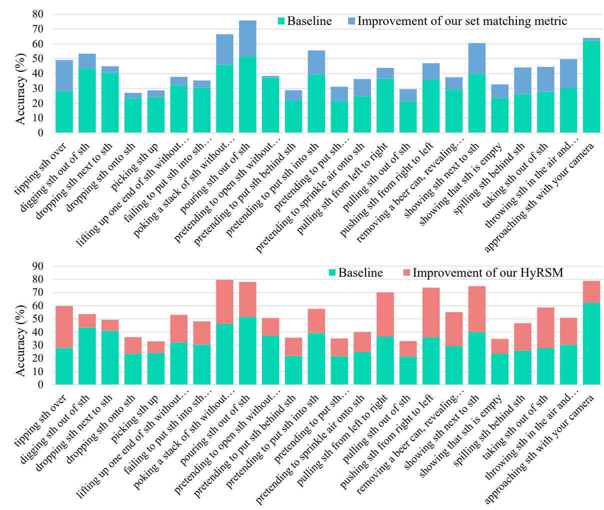

Figure 1. Category gain on the SSv2-Full dataset.

∗ Corresponding authors.

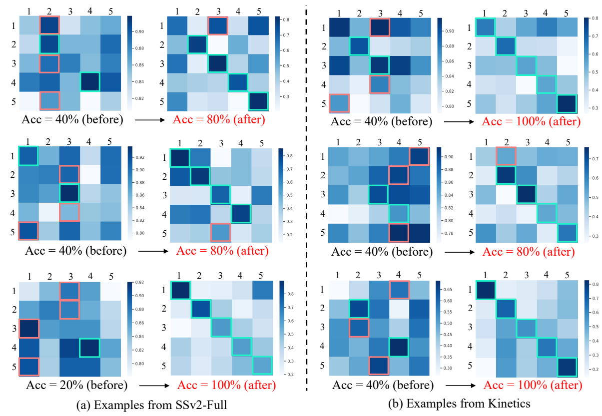

Figure 2. Similarity visualization of how query videos (rows) match to support videos (columns) before and after the hybrid relation module in HyRSM. The boxes of different colors correspond to: correct match and incorrect match.

C. Class improvement

In order to further analyze the performance improvement of each action category, we compare the improvement of the proposed set matching metric and HyRSM compared to the baseline on SSv2-Full, as depicted in Figure 1. For the set matching metric, some action classes have limited improvements, *e.g.*, "drop something onto something" and "pretending to open something without actually opening it", whereas some action classes have more than 20% improvement, *e.g.*, "tipping something over" and "showing something next to something". For our HyRSM, the improvement of each category is more evident than the set matching metric. In particular, "pulling something from left to right" and "pushing something from right to left" do not have significant increases in set matching metric but increase by more than 25% in HyRSM. This suggests that the hybrid relation module and the proposed set matching metric are strongly complementary.

D. Visualization analysis

To further demonstrate the effectiveness of our proposed hybrid relation module, we visualize the similarity maps of features before and after the hybrid relation module in HyRSM in Figure 2. The results indicate that the features are improved significantly after refining by the hybrid relation module. In addition, to qualitatively evaluate the proposed HyRSM, we compare the class activation maps visualization results of HyRSM to the competitive OTAM [1]. As shown in Figure 3 and Figure 4, the features of OTAM usually contain non-target objects since it lacks the mechanism of learning task-specific embeddings for feature adaptation. In contrast, our proposed HyRSM processes the query and support videos with adaptive relation modeling operation, which allows it to focus on the different target objects.

E. Relation modeling operations

In the literature [3, 8, 10, 14–17], there are many alternative relation modeling operations, including multi-head self-attention (MSA), Transformer, Bi-LSTM, Bi-GRU, *etc*.

Multi-head self-attention mechanism operates on the triple query Q, key K and value V, and relies on scaled dot-product attention operator:

$$Attention(Q;K;V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where d_k a scaling factor equal to the channel dimension of key K. Multi-head self-attention obtains h different heads and each head computes scaled dot-product attention representations of triple (Q, K, V), concatenates the intermediates, and projects the concatenation through a fully connected layer. The formula can be expressed as:

$$head_i = Attention(QW_i^q; KW_i^k; VW_i^v) \quad (2)$$

$$MSA(Q;K;V) = concat_i(head_i)W', 1 \leq i \leq h. \quad (3)$$

where the W_i^q, W_i^k, W_i^v and W' are fully connected layer parameters. Finally, a residual connection operation is employed to generate the final aggregated representation:

$$f_{msa} = MSA(f;f;f) + f \quad (4)$$

where f comes from the output of the previous layer. Note that query, key and value are the same in self-attention.

Transformer is a state-of-the-art architecture for natural language processing [4, 6, 14]. Recently, it has been widely used in the field of compute vision [2, 7, 18, 19] due to its excellent contextual modeling ability, and has achieved significant performances. Transformer contains two sub-layers: (a) a multi-head self-attention layer (MSA), and (b) a feed-forward network (FFN). Formulaic expression is:

$$f_{transformer} = FFN(f_{msa}) + f_{msa} \quad (5)$$

where FFN contains two MLP layers with a GELU non-linearity [9].

Bi-LSTM is an bidirectional extension of the Long Short-Term Memory (LSTM) with the ability of managing variable-length sequence inputs. Generally, an LSTM consists of three gates: forget gate, input gate and output gate. The forget gate controls what the existing information needs to be preserved/removed from the memory. The input gate makes the decision of whether the new arrival will be added. The output gate uses a sigmoid layer to determine which

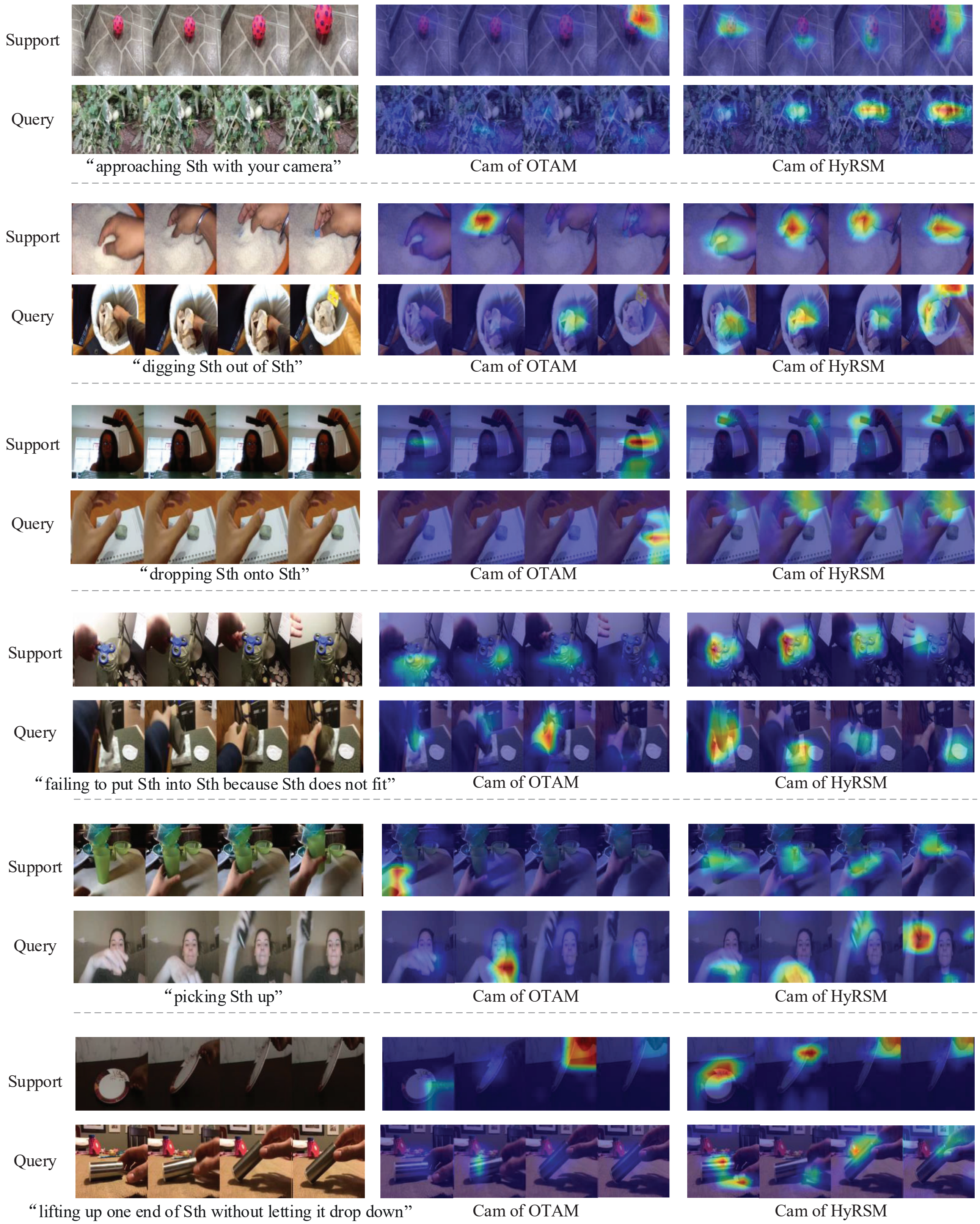

Figure 3. Visualization of class activation maps (Cam) with Grad-CAM [13] on SSv2-Full. Corresponding to: original RGB images (left), Cam of OTAM [1] (middle) and Cam of HyRSM (right).

Figure 4. Visualization of class activation maps (Cam) with Grad-CAM [13] on Kinetics. Corresponding to: original RGB images (left), Cam of OTAM [1] (middle) and Cam of HyRSM (right).

part of memory attributes to the final output. The mathematical equations are:

$$f_t = \sigma(W_{f_h}[h_{t-1}] + W_{f_x}[x_t] + b_f) \quad (6)$$
$$i_t = \sigma(W_{i_h}[h_{t-1}] + W_{i_x}[x_t] + b_i) \quad (7)$$
$$\widetilde{c_t} = \tanh(W_{c_h}[h_{t-1}] + W_{c_x}[x_t] + b_c) \quad (8)$$
$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \quad (9)$$
$$o_t = \sigma(W_{o_h}[h_{t-1}] + W_{o_x}[x_t] + b_o) \quad (10)$$
$$h_t = o_t * tanh(c_t) \quad (11)$$

where f_t is the value of the forget gate, o_t is the output result, and h_t is the output memory. In Bi-LSTM, two LSTMs are applied to the input and the given input data is utilized twice for training (*i.e*., first from left to right, and then from right to left). Thus, Bi-LSTM can be used for sequence data to learn long-term temporal dependencies.

Bi-GRU is a variant of Gated Recurrent Unit (GRU) and have been shown to perform well with long sequence applications [12, 21]. In general, the GRU cell contains two gates: update gate and reset gate. The update gate z_t determines how much information is retained in the previous hidden state and how much new information is added to the memory. The reset gate r_t controls how much past information needs to be forgotten. The formula can be expressed as:

$$z_t = \sigma(W_z[x_t] + U_z[h_{t-1}] + b_z) \quad (12)$$
$$r_t = \sigma(W_r[x_t] + U_r[h_{t-1}] + b_r) \quad (13)$$
$$\widetilde{h_t} = g(W_h[x_t] + U_h[(r_t * h_{t-1})] + b_h) \quad (14)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \quad (15)$$

where x_t is the current input and h_t is the output hidden state.

References

[1] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, pages 10618–10627, 2020. 2, 3, 4

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[4] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649, 2013. 2

[9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 2

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[11] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *CVPR*, pages 1–10, 2019. 1

[12] Peng Peng, Wenjia Zhang, Yi Zhang, Yanyan Xu, Hongwei Wang, and Heming Zhang. Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis. *Neurocomputing*, 407:232–245, 2020. 5

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 3, 4

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[16] Xiang Wang, Zhiwu Qing, Ziyuan Huang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Changxin Gao, and Nong Sang. Proposal relation network for temporal action detection. *arXiv preprint arXiv:2106.11812*, 2021. 2

[17] Xiang Wang, Zhiwu Qing, Ziyuan Huang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Yuanjie Shao, and Nong Sang. Weakly-supervised temporal action localization through local-global background modeling. *arXiv preprint arXiv:2106.11811*, 2021. 2

[18] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. *ICCV*, 2021. 2

[19] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, pages 8741–8750, 2021. 2

[20] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *ICML*, pages 7115–7123, 2019. 1

[21] Rui Zhao, Dongzhe Wang, Ruqiang Yan, Kezhi Mao, Fei Shen, and Jinjiang Wang. Machine health monitoring using local feature-based gated recurrent unit networks. *IEEE Transactions on Industrial Electronics*, 65(2):1539–1548, 2017. 5