MOFO: MOtion FOcused Self-Supervision for Video Understanding

Mona Ahmadian, Frank Guerin, Andrew Gilbert
University of Surrey, UK
{m.ahmadian, f.guerin, a.gilbert}@surrey.ac.uk

Abstract

Self-supervised learning (SSL) techniques have recently produced outstanding results in learning visual representations from unlabeled videos. However, despite the importance of motion in supervised learning techniques for action recognition, SSL methods often do not explicitly consider motion information in videos. To address this issue, we propose MOFO (MOtion FOcused), a novel SSL method for focusing representation learning on the motion area of a video for action recognition. MOFO automatically detects motion areas in videos and uses these to guide the self-supervision task. We use a masked autoencoder that randomly masks out a high proportion of the input sequence and forces a specified percentage of the inside of the motion area to be masked and the remainder from outside. We further incorporate motion information into the finetuning step to emphasise motion in the downstream task. We demonstrate that our motion-focused innovations can significantly boost the performance of the currently leading SSL method (VideoMAE) for action recognition. Our proposed approach significantly improves the performance of the current SSL method for action recognition, indicating the importance of explicitly encoding motion in SSL.

1 Introduction

Action recognition is an essential task in video understanding and has been extensively investigated in recent years Liu et al. [2022], Wei et al. [2022], Girdhar et al. [2022a]. In video action recognition, supervised deep learning techniques have made significant progress Tran et al. [2015], Feichtenhofer et al. [2019], Lin et al. [2019]; However, due to the lack of labels, which must be manually collected, learning to recognise actions from a small number of labelled videos is a difficult task as data collection will be expensive and challenging. It is especially inappropriate for long-tail open vocabulary object distributions across scenes, such as a kitchen. Furthermore, getting annotations for videos is much more difficult due to the large number of frames and the temporal boundaries of when actions begin and end. Therefore, SSL has gained attention due to the problems above.

Supervised methods Wang and Gupta [2018], Kwon et al. [2020], Patrick et al. [2021] have recognised the importance of motion to understand actions because often, key objects are moving in the scene. However, most SSL methods do not explicitly consider motion or use hand-crafted features Escorcia et al. [2022], limiting their effectiveness. In SSL literature, masked autoencoder models Tong et al. [2022] have been proposed to learn the underlying data distribution but without directly emphasising motion autonomously. Even though this model can perform spatiotemporal reasoning over content, the encoder backbone is ineffective in capturing motion representations (we show this later in Fig. 2). Incorporating motion information is not trivial, especially in egocentric videos. Some previous approaches utilized both RGB frames and optical flows Han et al. [2020], Ni et al. [2022] to strengthen learning of features but the primary issue lies in the stability of the results, which can be significantly impacted by camera movement. When the camera moves rapidly, static objects or background pixels exhibit high movement velocities in optical flow. Several existing methods leveraged object detection to improve egocentric video recognition Wang et al. [2020,?], Wu et al.

37th Conference on Neural Information Processing Systems (NeurIPS 2023). Workshop: Self-Supervised Learning-TheoryandPractice.

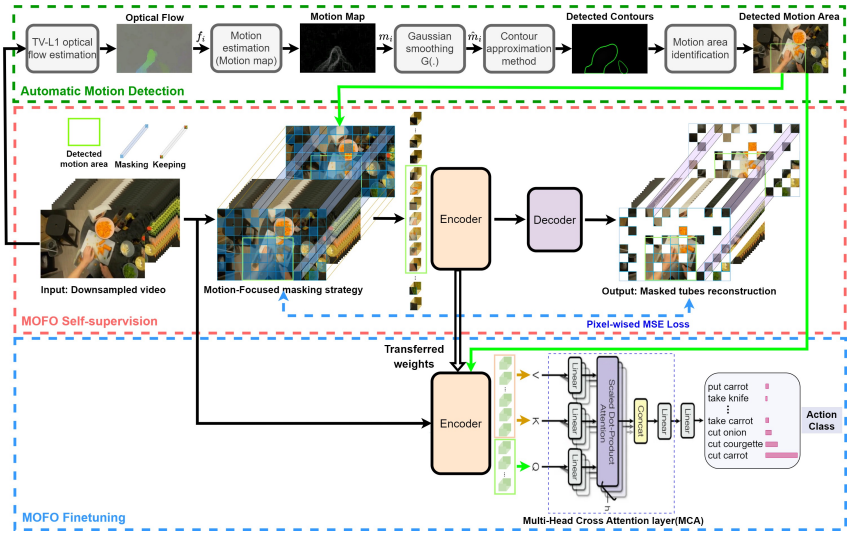

Figure 1: MOFO is a motion-focused self-supervised framework for action recognition.

[2019], Ma et al. [2016], among which Wu et al. [2019] also incorporate temporal contexts to help understand the ongoing action. These approaches may have limited uses in real-world systems since they demand time-consuming, labour-intensive object detection annotations and are computationally expensive. In contrast, our framework does not depend on costly object detectors.

Fig. 1 overviews our method, with three parts: First, our automatic motion area detection using optical flow input to create a motion map to remove camera motion. Second, we propose our new strategy for the SSL pretext task, a reconstruction task focusing more on masking 3D patches on the motion area in the video called MOFO (Motion Focused). Thirdly, the downstream task adaptation step emphasises motion further by integrating motion information during the finetuning training. A key contribution of our work is to detect salient objects and motion in the video based on motion boundaries from optical flow. Using the motion boundaries instead of a direct optical flow output mitigates the challenge of camera motion and creates salient areas of movement or interest without a pretrained network. Given the motion identification, we suggest extending the self-supervised masking Tong et al. [2022] to include motion understanding. A further contribution is that, during the finetuning stage, MOFO prioritises the motion areas in video data identified as a self-supervision pretext task. Since motion areas contain more information, such as moving objects, actions, and interactions, our proposed model gives them a higher priority by emphasising the masking strategy to be more in the motion area.

2 Motion-focused Self-supervised Video Understanding

2 .1 Automatic motion area detection

To identify the motion areas without pretrained object detectors, we propose using classical computer vision features, the optical flow vectors. however, these vectors will be affected by camera motion, with static objects or background pixels exhibiting high movement velocities in optical flow when the camera moves rapidly. To mitigate the problem above, we calculate the motion boundaries Dalal et al. [2006] and use these to define a motion map Li et al. [2021]. Therefore, given a video with T frames and a $H \times W$ dimension, we first extract the optical flow vectors representing $\left\{f_i \in \mathbb{R}^{H \times W}\right\}_{i=1}^{T}$ pixel-level motion between two consecutive frames in a video using the TV-L1 algorithm Zach et al. [2007] that offers increased robustness against illumination changes, occlusions, and noise. Then, given the horizontal and vertical displacements of each pixel between the ith frame and the $(i+1)$th frame represented by the flow maps $u_i, v_i \in \mathbb{R}^{H \times W}$, any kind of local differential or flow difference cancels out most of the effects of the camera rotation. The resulting motion map is defined as:

$$m_i = \sqrt{(\frac{\partial u_i}{\partial x})^2 + (\frac{\partial u_i}{\partial y})^2 + (\frac{\partial v_i}{\partial x})^2 + (\frac{\partial v_i}{\partial y})^2} \tag{1}$$

where every component denotes the corresponding x- and y-derivative differential flow frames contributing towards computing m_i, representing moving velocity in the i-th frame while ignoring the camera motion. As a result, $m_i \in \mathbb{R}^{H \times W}$ is less influenced by camera motion and considers the moving salients in the i-th frame. A low-pass Gaussian filter is used to smooth areas of the image with high-frequency components to further reduce the unwanted noise effect. The Gaussian Smoothing Operator computes an average of the surrounding pixels weighted according to the Gaussian distribution (G).

After noise reduction, the next step is to find the boundaries of the motion. To do so, we create contours Suzuki et al. [1985], which are short curves that connect points of the same hue or intensity. We select the two most significant contours in each frame to create a mask that indicates the motion area in a frame of a specific video. The main reason for choosing two contours is that in our datasets, an action is defined by hands and the corresponding object. We create a bounding box around the resulting area that precisely represents the motion in each video. In Fig. **??**(a), we qualitatively compare our automatic box predictions and the provided supervised annotation for Epic-Kitchens-100 for several sample frames and provide further examples in the Appendix in Fig **??**.

2 .2 Motion-focused self-supervised learning

MOFO uses 3D tube volume embeddings for the self-supervised pretext stage to obtain 3D video patches from frames as inputs. It encodes these with a vanilla ViT Dosovitskiy et al. [2020] with joint space-time attention as a backbone. We segmented each video into N non-overlapping tubes $\mathbf{p}_i \in \mathbb{R}^{H_t \times W_t \times T_t}$. Then, we use a high-ratio tube masking approach to perform masked autoencoder (MAE) pretraining with an asymmetric transformer-based encoder-decoder architecture reconstruction task. Unlike other random masking methods, we explicitly integrate the motion information computed in subsection 2 .1 into our masking strategy, resulting in a motion-guided approach to encoding motion for our MAE. Once the motion area is detected, our novel tube masking strategy enforces a mask to be applied on a high portion of the tubes inside the motion area. In other words, a fixed percentage of the tubes (generally 75%) inside the motion area is always randomly masked to ensure the model is attending more to the motion area at reconstruction time. Therefore, we apply an extremely high masking ratio at random (90%) while always masking a fixed percentage of the tubes (75%) inside the motion area. The encoder produces a latent feature representation of the video using input frames with blacked-out regions. The decoder uses the latent feature representation from the encoder. It estimates the missing region using the mean squared error (MSE) loss, computed in pixel space between the masked patches and trained reconstructed outputs. Our design encourages the network to capture more useful spatiotemporal structures, making MOFO a more meaningful task and improving the performance of self-supervised pretraining. All models only use the unlabelled data in the training set of each dataset for pertaining.

2 .3 Motion-focused finetuning

Recall that the self-supervised learning protocol is split between a pretraining and finetuning stage. We propose a new approach to focus on the motion area at both the pretext and the finetuning of the model. The model is trained end-to-end during finetuning, using the weights of the pretrained network as initialisation for the downstream supervised task dataset.

As the area inside the motion box has more semantic motion information, we wish to exploit this information for our task by leveraging the detected motion box. On the other hand, the video's setting and any nearby items could provide context for categorising the video clips for the action recognition task. For instance, in the case of washing dishes, the hands can be seen in the sink, but the dishes beside the sink may indicate that the person is washing them. Therefore, we propose to use multi-cross attention (MCA) Nagrani et al. [2021] in our encoder. MCA is an attention mechanism that mixes two different embedding sequences; the two are from the same modality. Unlike self-attention, where inputs are the same set, during cross-attention, they differ; MCA's main objective is to determine attention scores using data from various information sources. This module resides between the encoder and MLP classifier layers, takes the inner and outer motion box embeddings, and outputs the fused embedding (see details in Appendix **??**).

3 Experiments

We use two well-known and large datasets to evaluate our proposed approach: **Something-Something V2 (SSV2)** Goyal et al. [2017] and **Epic-Kitchens-100** Damen et al. [2022]. Using

Table 1: Human activity recognition on **Epic-Kitchens** and **Something-Something V2 (SSV2)** in terms of Top-1 and Top-5 accuracy.

Method	Backbone	Param	SSV2 Action Top-1	SSV2 Action Top-5	Epic-Kitchens Verb Top-1	Epic-Kitchens Noun Top-1	Epic-Kitchens Action Top-1
VIMPAC Tan et al. [2021]	ViT-L	307	68.1	-	-	-	-
BEVT Wang et al. [2022]	Swin-B	88	70.6	-	-	-	-
VideoMAE Tong et al. [2022]	ViT-B	87	70.8	92.4	71.6	66.0	53.2
ST-MAE Feichtenhofer et al. [2022]	ViT-L	304	72.1	-	-	-	-
OmniMAE Girdhar et al. [2022a]	ViT-B	87	69.5	-	-	-	39.3
Omnivore(Swin-B) Girdhar et al. [2022b]	ViT-B	-	71.4	93.5	69.5	61.7	49.9
MOFO (Proposed)	ViT-B	102	**75.5**	**95.3**	**74.2**	**68.1**	**54.5**

egocentric videos to predict first-person activity faces many challenges, including a limited field of view, occlusions, and unstable motions, and there is a relative scarcity of labelled data.

Results and analysis We finetune the learned model for action classification based on our proposed MOFO finetuning approach to evaluate the pretrained model and train on a new downstream task with the learned representation. The entire feature encoder and a linear layer are finetuned end-to-end with cross-entropy loss, with recognition accuracy reported in Table 1. We demonstrate significant performance improvement over the other self-supervised approaches, increasing 2.6%, 2.1%, and 1.3% accuracy over the best-performing methods on Epic-Kitchens verb, noun, and action classification and 4.7% on Something Something V2 action classification, respectively. In terms of masking ratio, variants are presented in the Appendix, but we found that the 75% inside masking ratio worked the best. Our strategy outperforms approaches like OmniMAE Girdhar et al. [2022a], trained jointly on images and videos by 3.2% in Top-1 accuracy. On Something Something V2, our method outperforms VIMPAC Tan et al. [2021] and ST-MAE Feichtenhofer et al. [2022], which both use ViT-Large as a backbone, whereas our backbone is vanilla ViT-Base with over 3x fewer parameters. Compared to VideoMAE Tong et al. [2022], our approach achieves significantly better results while the number of backbone parameters remains the same.

Visualizing self-supervised representation To further understand the representations learned by MOFO, we utilise GradCAM Selvaraju et al. [2017] to create a saliency map highlighting each pixel's importance to show how each pixel contributes to the discrimination of the video clip. Fig. 2 visualises the middle frame of a video clip, the motion map of the VideoMAE and our MOFO from the fifth attention layer of the ViT-Base backbone. It is interesting to note that for similar actions: *knead dough*, *cut carrot*, and *cut-in tomato*, MOFO is sensitive to the location that is the most significant motion location as detected by our automatic algorithm.

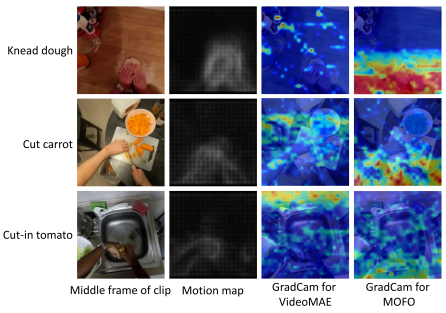

Figure 2: Visualisation of the learned features

4 Conclusion

MOFO introduces a Motion-Focused technique which explores motion information for enhancing motion-aware self-supervised video action recognition. We propose an innovative strategy, an effective self-supervised pretext task, and a modification to masked autoencoding, which focuses masking on the motion area in the video (Motion Focused). Extensive experiments on two challenging datasets demonstrate that this context-based SSL technique improves performance in action recognition tasks, and the public code will guide many research directions.

Acknowledgement The work was partially funded by a Leverhulme Trust Research Project Grant: RPG-2023-079 "How humans understand video".

References

Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

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: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022.

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. *ICLR*, 2020.

Victor Escorcia, Ricardo Guerrero, Xiatian Zhu, and Brais Martinez. Sos! self-supervised learning over sets of handled objects in egocentric action recognition. In *ECCV*, 2022.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

Christoph Feichtenhofer, haoqi fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022a.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022b.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *NeurIPS*, 2020.

Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *ECCV*, 2020.

Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *ICCV*, 2021.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, June 2022.

Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 2021.

Jingcheng Ni, Nan Zhou, Jie Qin, Qian Wu, Junqi Liu, Boxun Li, and Di Huang. Motion sensitive contrastive learning for self-supervised video representation. In *ECCV*, 2022.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 2021.

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*, 2017.

Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 1985.

Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022.

Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *AAAI*, 2020.

Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019.

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, 2007.