Busy-Quiet Video Disentangling for Video Classification

Guoxi Huang and Adrian G. Bors

Department of Computer Science
University of York, York YO10 5GH, UK

{gh825, adrian.bors}@york.ac.uk

Abstract

In video data, busy motion details from moving regions are conveyed within a specific frequency bandwidth in the frequency domain. Meanwhile, the rest of the frequencies of video data are encoded with quiet information with substantial redundancy, which causes low processing efficiency in existing video models that take as input raw RGB frames. In this paper, we consider allocating intenser computation for the processing of the important busy information and less computation for that of the quiet information. We design a trainable Motion Band-Pass Module (MBPM) for separating busy information from quiet information in raw video data. By embedding the MBPM into a two-pathway CNN architecture, we define a Busy-Quiet Net (BQN). The efficiency of BQN is determined by avoiding redundancy in the feature space processed by the two pathways: one operating on Quiet features of low-resolution, while the other processes Busy features. The proposed BQN outperforms many recent video processing models on Something-Something V1, Kinetics400, UCF101 and HMDB51 datasets. The code is available at: https://github.com/guoxih/busy-quiet-net.

1. Introduction

Video classification is a fundamental problem in many video-based tasks. Applications such as autonomous driving technology, controlling drones and robots are driving the demand for new video processing methods. An effective way to extend the usage of Convolutional Neural Networks (CNNs) from the image to the video domain is by expanding the convolution kernels from 2D to 3D [3, 16, 40, 41]. Since the progress made by I3D [3], the main research effort in the video area has been directed towards designing new 3D architectures. However, 3D CNNs are more computationally intensive than 2D CNNs. Some recent works [11, 29, 32, 42, 43, 48] increase the efficiency of 3D CNNs by reducing the redundancy in the model parameters. However, these works have ignored another important factor that

Figure 1: Motion Band-Pass Module (MBPM) distills busy information from the frame sequences. For every three consecutive RGB frames, the MBPM generates a one-frame output, which substantially reduces the redundancy.

causes the heavy computation in video processing: natural video data contains substantial redundancy in the spatio-temporal dimensions. In video data, busy information describes fast-changing motion happening in the boundaries of moving regions which is crucial for defining movement in video. Meanwhile, the quiet information, such as smooth background textures whose information is shared by neighboring locations, contains substantial redundancy. For efficient processing, we disentangle a video into busy and quiet components. Subsequently, we would separately process the busy and quiet components, by allocating high-complexity processing for the busy information and low-complexity processing for the quiet information.

In this study we propose a lightweight, end-to-end trainable motion feature extraction mechanism called Motion Band-Pass Module (MBPM), which can distill the motion information conveyed within a specific frequency bandwidth in the frequency domain. As illustrated in Figure 1, by applying the MBPM to a video of 3 segments the number of representative frames is reduced from 9 to 3 whilst re-

taining the essential motion information. Our experiments demonstrate that by simply replacing the RGB frame input with the motion representation extracted by our MBPM, the performance of existing video models can be boosted. Secondly, we design a two-pathway multi-scale architecture, called the Busy-Quiet Net (BQN), the processing pipeline of which is shown in Figure 2. One pathway, called Busy, is responsible for processing the busy information distilled by the MBPM. The other pathway, called Quiet, is devised to process the quiet information encoded with global smooth spatio-temporal structures. In order to fuse the information from different pathways, we design the Band-Pass Lateral Connection (BPLC) module, which is set up between the Busy and Quiet pathways. During the experiments, we demonstrate that the BPLC is the key factor to the overall model optimization success.

Compared with the frame summarization approaches [1, 44], MBPM retains the strict temporal order of the frame sequences, which is considered essential for long-term temporal relation modeling. Compared with optical flow-based motion representation methods [9, 22, 45, 50, 52], the motion representation captured by MBPM has a smaller temporal size (*e.g.* for every 3 RGB frames, the MBPM encodes only one frame), and can be employed on the fly. Meanwhile, efficient video models such as Octave Convolution [5], bL-Net [4] and SlowFast networks [12] only reduce the input redundancy along either the spatial or temporal dimensions. Instead, the proposed BQN reduces the redundancy in the joint spatio-temporal space.

Our contributions can be summarized as follows:

- A novel Motion Band-Pass Module (MBPM) is proposed for busy motion information distillation. The new motion cue extracted by the MBPM is shown to reduce temporal redundancy significantly.

- We design a two-pathway Busy-Quiet Net (BQN) that separately processes the busy and quiet information in videos. By separating the busy information using MBPM, we can safely downsample the quiet information to further reduce redundancy.

- Extensive experiments demonstrate the superiority of the proposed BQN over a wide range of models on four standard video benchmarks: Kinetics400 [3], Something-Something V1 [15], UCF101 [36], HMDB51 [26].

2. Related Work

Spatio-temporal Networks. Early works [7, 14, 34, 45, 49] attempt to extend the success of 2D CNNs [18, 21, 25, 35, 38, 39] in the image domain to the video domain. Representatively, the two-stream model [34] and its variants [14, 45] utilize optical flow as an auxiliary input modality for effective temporal modeling. Other works [3, 16, 40, 41], given

the progress in GPU performance, tend to exploit the computationally intensive 3D convolution. Meanwhile, some studies focus on improving the efficiency of 3D CNN, such as P3D [32], R(2+1)D [43], S3D [48], TSM [29], CSN [42], X3D [11]. Non-local Net [46] and its variants [2, 20] introduce self-attention mechanisms to CNNs in order to learn long-range dependencies in the spatio-temporal dimension. Our study is complementary to these methods: our Busy-Quiet Net (BQN) can benefit from the efficiency of these CNNs by simply adopting them as backbones.

Motion Representation. Optical flow as a short-term motion representation has been widely used in many video applications. However, the optical flow estimation in large-scale video datasets is inefficient. Some recent works use deep learning to improve the optical flow estimation quality, such as FlowNet [8, 22]. Some other methods aim to explore new end-to-end trainable motion cues, such as OFF [37], TVNet [9], EMV [51], Flow-of-Flow [31], Dynamic Image [1], Squeezed Image [19] and PA [52]. Compared with these approaches, our MBPM is rather as a basic video architecture component which results in higher accuracy while requiring less computation.

Enforcing Low Information Redundancy. In the image field, bL-Net [4] adopts a downsampling strategy that operates at the block level aiming to reduce the spatial redundancy of its feature maps. Octave Convolution [5] replaces the convolutions in existing CNNs to decompose the low and high-frequency components in images, representing the former with lower resolution. In the video field, bLV-Net [10] extends the idea of bL-Net [4] to the temporal dimension. SlowFast networks [12] introduce two pathways for slow and fast motion decomposition along the temporal dimension. However, the generalization of SlowFast to existing CNN architectures is poor, as it requires two specially customized CNNs to be its backbones. Unlike the previous methods, BQN reduces the feature redundancy in the joint spatio-temporal space by using a predefined trainable filter module, MBPM, to disentangle a video into busy and quiet components. Unlike SlowFast, BQN architecture shows excellent generalization to existing CNNs.

3. Motion Band-Pass Module (MBPM)

Firstly, we introduce a 3D band-pass filter, which can distill the video motion information conveyed within a specific spatio-temporal frequency bandwidth. A video clip of T frames can be defined as a function with three arguments, $\boldsymbol{I}^{(t)}(x, y)$, where x, y indicate the spatial dimensions, while t is the temporal dimension. The value of $\boldsymbol{I}^{(t)}(x, y)$ corresponds to the pixel value at position (x, y) in t-th frame of an arbitrary channel in the video. When considering the multi-channel case, we repeat the same procedure for each color channel, which is omitted here for the sake of simplification. The output $\boldsymbol{\Gamma}$ of the 3D band-pass filter is given
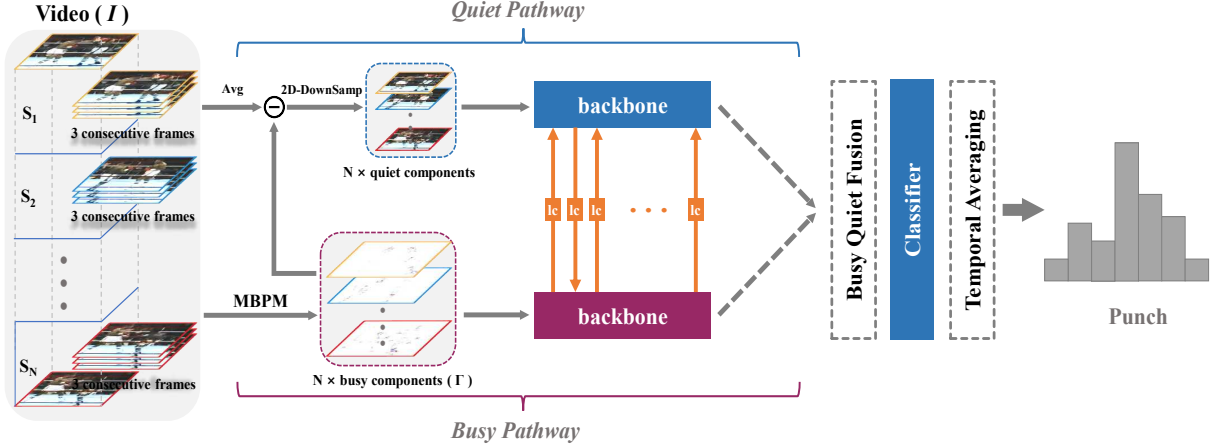
Figure 2: BQN is made up of two parallel pathways: Busy and Quiet. 'lc' indicates Band-Pass Lateral Connection. The backbone networks from the two pathways respectively take as inputs two complementary data components (*i.e.* busy and quiet), which are disentangled by the MBPM. The outputs of the two pathways are fused at various processing stages and the final prediction is obtained by averaging the prediction scores across multiple segments.

by:

$$
\begin{aligned}
\mathbf{\Gamma}(x, y, t) &= \frac{\partial^2}{\partial t^2}\Big[\mathbf{I}^{(t)}(x, y) * LoG_\sigma(x, y)\Big], \\
&\approx \sum_{t-1 \le i \le t+1} h_{(i)} \cdot [\mathbf{I}^{(i)}(x, y) * LoG_\sigma(x, y)],
\end{aligned}
$$

$$
h_{(i)} = \begin{cases} \frac{2}{3} & \text{if } i = t, \\ -\frac{1}{3} & \text{otherwise}, \end{cases} \tag{1}
$$

for $t = 1, \ldots, T$ and '$*$' represents the convolution operation. In equation (1), the second derivative with respect to t is numerically approximated by finite differences, literally implemented by function $h_{(i)}$. Meanwhile, $LoG_\sigma(x, y)$ is a two-dimensional Laplacian of Gaussian with the scale parameter σ:

$$
LoG_\sigma(x, y) = \triangledown^2 G_\sigma(x, y) = -\frac{e^{-\frac{x^2+y^2}{2\sigma^2}}}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^2}\right]. \tag{2}
$$

From equations (1) and (2) we can observe that the 3D filtering function is fully-differentiable. In order to make the 3D band-pass filtering compatible with CNNs, we approximate it with two sequential channel-wise[1] convolutional layers [33], as shown in Figure 1. We name the discrete approximation Motion Band-Pass Module (MBPM) which can be expressed in an engineering form as follows:

$$
\mathbf{\Gamma} \approx \text{MBPM}(\mathbf{I}) = H^{3\times1\times1}_{s\times1\times1}(LoG^{1\times k\times k}_\sigma(\mathbf{I})), \tag{3}
$$

where $LoG^{1\times k\times k}_\sigma$ is referred to as a spatial channel-wise convolutional layer [33] with a $k \times k$ kernel, each channel

---

[1]Also referred to as "depth-wise". We use the term "channel-wise" to avoid confusions with the network depth.

of which is initialized with a Laplacian of Gaussian distribution with scale σ. The sum of kernel values is normalized to 1. Meanwhile, $H^{3\times1\times1}_{s\times1\times1}$ is referred to as a temporal channel-wise convolutional layer with a temporal stride s. In each channel, the kernel value of $H^{3\times1\times1}_{s\times1\times1}$ is initialized with $\left[-\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}\right]$, which is a high-pass filter. In order to let the MBPM kernel parameters fine-tune on video streams, we embed the MBPM within a CNN for end-to-end training, optimized with the video classification loss.

4. Busy-Quiet Net (BQN)

As illustrated in Figure 2, the BQN architecture contains two pathways, Busy and Quiet, operating in parallel on two distinct video data components, which are separated by the MBPM. The Busy and Quiet pathways are bridged by multiple Band-Pass Lateral Connections (see Section 4.2). These lateral connections enable information fusion between the two processing pathways.

4.1. Busy and Quiet pathways

Busy pathway. The Busy pathway is designed to learn fine-grained movement features. It takes as input the information filtered by the MBPM, which contains critical motion information located at the boundaries of objects or regions that have significant temporal change. The stride of $H^{3\times1\times1}_{s\times1\times1}$ from equation (3) is set to $s = 3$, which means that for every three consecutive RGB frames, the MBPM generates one-frame output. The MBPM output preserves the temporal order of the video frames while significantly reducing the redundant temporal information. We intend to utilize larger spatial input sizes for the Busy pathway to

extract more distinct textures.

Quiet pathway. The Quiet pathway focuses on processing quiet information, representing the characteristics of large regions of movement, such as the movement happening in the plain-textured background regions. The input to the Quiet pathway is considered to be the complementary of the MBPM output:

$$2\text{D-DownSamp}(\text{Avg}_{3\times1\times1}^{3\times1\times1}(\boldsymbol{I}) - \boldsymbol{\Gamma}), \tag{4}$$

where $\text{Avg}_{3\times1\times1}^{3\times1\times1}$ is a temporal average pooling with a stride of 3. In the spatial dimensions, we perform bilinear down-sampling (*i.e.* 2D-DownSamp) to reduce the redundant spatial information shared by neighboring locations. In Section 5.3, we explore the effect on performance when varying the input size of the Quiet pathway.

4.2. Band-Pass Lateral Connection (BPLC)

In the proposed BQN, we include a novel Band-Pass Lateral Connection (BPLC) module which has an MBPM embedded. The BPLCs established between the two pathways, Busy and Quiet, provide a mechanism for information exchange, enabling an optimal fusion of video information characterized by different frequency bands. Different from the lateral connections in [12, 13, 14, 30], the BPLC, enabled by MBPM, performs feature fusion and feature selection simultaneously, which shows higher performance than other lateral connection designs, according to the experimental results. We denote the two inputs of BPLC from the i-th residual blocks in the Busy and Quiet pathways, as $\mathbf{x}_f^i$ and $\mathbf{x}_c^i$, respectively. For simplifying the notation, we assume that $\mathbf{x}_f^i$ and $\mathbf{x}_c^i$ are of the same size. When their sizes are different, we adopt bilinear interpolation to match them in size. The output $\mathbf{y}_f^i$ and $\mathbf{y}_c^i$ for the Busy and Quiet, respectively, is given by

$$\mathbf{y}_f^i = \begin{cases} \text{BN}(\text{MBPM}(\mathbf{x}_c^i)) + \mathbf{x}_f^i & \text{if } \mod(i,2) = 0, \\ \mathbf{x}_f^i & \text{otherwise,} \end{cases}$$

$$\mathbf{y}_c^i = \begin{cases} \mathbf{x}_c^i & \text{if } \mod(i,2) = 0, \\ \text{BN}(\phi(\mathbf{x}_f^i)) + \mathbf{x}_c^i & \text{otherwise,} \end{cases}$$

$$i = 1, 2, \ldots, B \tag{5}$$

where B denotes the number of residual blocks in the backbone network (considered as the network with residual block designs in the experiments). $\phi(\cdot)$ is the linear transformation that can be implemented as a $1 \times 1 \times 1$ convolution, or alternatively, when the channel number is very large, as a bottleneck MLP for reducing computation. BN indicates Batch Normalization [23], the weights of which are initialized to zero. For the MBPM in BPLC, the convolutional stride of $H_{s\times1\times1}^{3\times1\times1}$ from equation (3) is set to $s = 1$, maintaining the same temporal size.

The fusion direction of BPLC reverses back and forth, as indicated in Figure 2, providing better communication
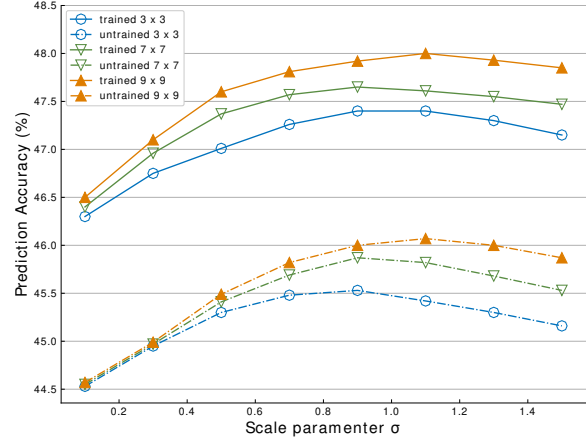

Figure 3: Results on SS V1 when varying the scale σ and kernel size $k \times k$ of the spatial channel-wise convolution in MBPM. The results represent averages of multiple runs.

for the two pathways than the unidirectional lateral connections in [12, 30] whose information fusion direction is fixed always fusing the information from a certain pathway to the other. By default, we place a BPLC between the two pathways right after each pair of residual blocks. The MBPM embedded in BPLC acts as a soft feature selection gate, which allows only for the busy information from the Quiet pathway to flow to the Busy pathway during the information fusion process. The exploration of various lateral connection designs is provided in Section 5.3 and this setting is shown to give the best performance in our experiments.

5. Experiments

In this section, we provide the results of the experiments for the proposed MBPM and BQN. Then, we compare these results with the state-of-the-art. Unless otherwise stated, we use ResNet50 (R50) with TSM [29], as the backbone of our model.

5.1. Datasets and Implementation Details

Datasets. We evaluate our approach on Something-Something V1 [15], Kinetics400 [3], UCF101 [36] and HMDB51 [26]. Most of the videos in Kinetics400 (K400), UCF101 and HMDB51 can be accurately classified by only considering their background scene information, while the temporal relation between frames is not very important. In Something-Something (SS) V1, many action categories are symmetrical (*e.g.* "Pulling something from left to right" and "Pulling something from right to left"). Discriminating these symmetric actions requires models with strong temporal modeling ability.

Training & Testing. Aside from X3D-M [11], the backbone networks are pretrained on ImageNet [6]. For training, we utilize the dense sampling strategy [46] for Kinetics400.
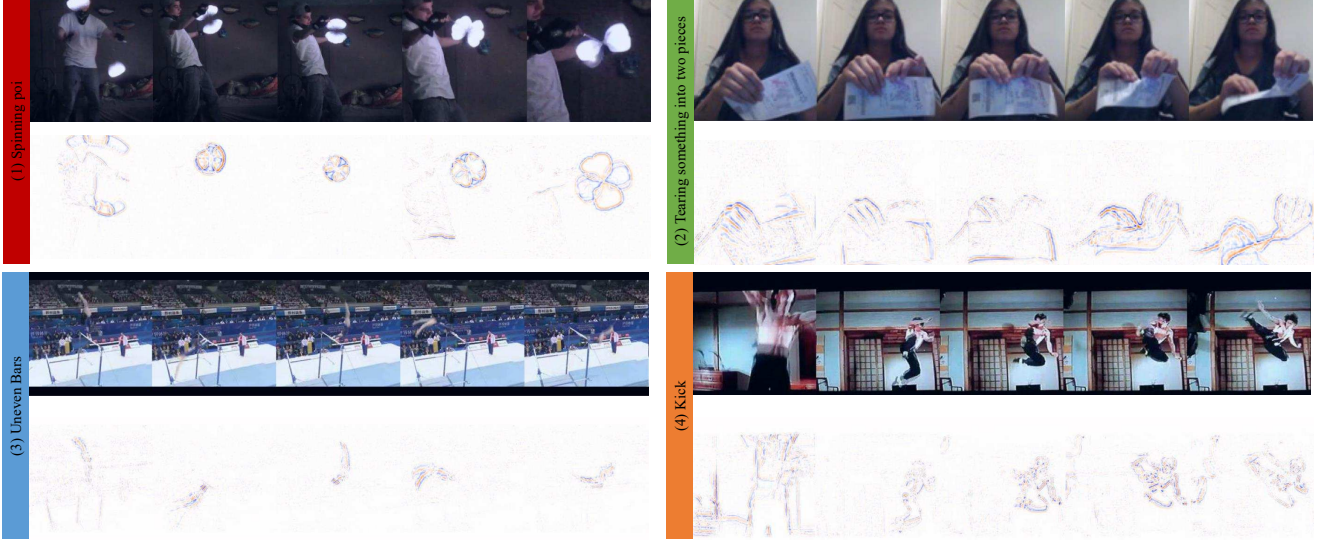
Figure 4: Videos and their MBPM outputs. The video clips (1)-(4) are from Kinetics, SS V1, UCF and HMDB, respectively.

Table 1: MBPM *vs.* other motion representation methods. † denotes our reimplementation. The additional parameters and the required computation (FLOPs) are reported.

Rep. Method	Efficiency Metrics		UCF101	SS V1	K400
	FLOPs	#Param.			
RGB (baseline)	-	-	87.1	46.5	71.2
RGB Diff [45]	-	-	87.0	46.6	71.4
TV-L1 Flow [50]	-	-	88.5	37.4	55.7
DI† [1]	-	-	86.2	43.4	68.3
FlowNetC† [22]	444G	39.2M	87.3	26.3	-
FlowNetS† [22]	356G	38.7M	86.8	23.4	-
TVNet† [9]	3.3G	0.2K	88.6	45.2	58.5
PA [52]	2.8G	1.1K	89.5	45.1	57.3
MBPM	0.3G	0.2K	**90.3**	**48.0**	**72.3**

Table 2: Using different CNNs as backbones on SS V1. ResNet50 and MobileNetV2 have TSM [29] embedded.

CNN backbone	Modality	pretrain	Seg. (N)	Acc.
ResNet50 [18]	RGB	ImageNet	8	46.5
	RGB+Flow			49.8
	MBPM			48.0
	RGB+MBPM			50.3
MobileNetV2 [33]	RGB	ImageNet	8	38.7
	MBPM			39.8
X3D-M [11]	RGB	None	16	45.5
	MBPM			46.9

As for the other datasets, we utilize the uniform sampling strategy as shown in Figure 2, where a video is equally di-

vided into N segments, and 3 consecutive frames in each segment are randomly sampled to constitute a video clip of length $T = 3N$. Unless specified otherwise, a default video clip is composed of $N = 8$ segments with a spatial size of 224^2. During the tests, we sample a single clip per video with center cropping for efficient inference [29], which is used in our ablation studies. When pursuing high accuracy, we consider sampling multiple clips&crops from the video and averaging the prediction scores of multiple space-time "views" (spatial crops×temporal clips) used in [12]. More training details can be found in Appendix A.

5.2. Ablation Studies for MBPM

Instantiations and Settings. In the MBPM, the scale σ from equation (2) and the kernel size of the spatial channel-wise convolution $LoG_\sigma^{1 \times k \times k}$ have a significant impact on the performance. We vary the scale σ and the kernel size to search for the optimal settings. Meanwhile, in order to highlight the importance of the training for the MBPM, we compare the performance when using trained MBPM with that of untrained MBPM whose kernel weights are not op-timized with the classification loss. The results on SS V1 are shown in Figure 3. We summarize two facts: 1) the optimal value of σ for the MBPM changes when the ker-nel size changes, and the MBPM with $\sigma = 1.1$ and a spa-tial kernel size 9×9 gives the best performance within the searching range. 2) optimizing the parameters of MBPM with the video classification loss generally produces higher prediction accuracy. In our preliminary work, we have ver-ified that different datasets share the same optimal settings of MBPM. More results on other datasets are provided in Appendix B. In the following experiments, we set MBPM in the Busy pathway as trainable with the scale $\sigma = 1.1$ and the kernel size of 9×9, unless specified otherwise.

Table 3: Ablation Studies for BQN on Something-Something V1. We show top-1 and top-5 prediction accuracy (%), as well as computational complexity measured in GFLOPs for a single crop & single clip.

Model	Top-1	Top-5	GFLOPs
Quiet	46.5	75.3	32.8
Busy	48.0	76.8	32.8
Quiet+Busy	50.3	79.0	65.7
BQN	**51.6**	**80.5**	65.9

# BPLC	Top-1	Top-5	GFLOPs
0	49.6	78.9	65.7
4	50.2	79.2	65.8
8	50.7	79.7	65.8
16	**51.6**	**80.5**	65.9

Fusion Method	Position	Top-1	Top-5
Average	before fc	50.9	79.8
Average	after fc	**51.6**	**80.5**
Max	after fc	50.1	78.7
Concatenation	before fc	51.3	80.2

(a) **Complementarity of Quiet and Busy.** "Quiet" and "Busy" refer to that the Quiet and Busy pathways are trained separately.

(b) **Effect of BPLCs.** The BQN model with more BPLCs has higher accuracy.

(c) **Fusion Strategies.** The fully-connected layers of the two pathways share the parameters.

Input size for Quiet	Input size for Busy	Top-1	Top-5	GFLOPs
$224^2 \times 8$	$224^2 \times 8$	51.6	80.5	65.9
$160^2 \times 8$	$224^2 \times 8$	51.3	80.1	50.5
$160^2 \times 8$	**$256^2 \times 8$**	**51.7**	**80.5**	**60.7**
$160^2 \times 6$	$256^2 \times 8$	49.6	78.3	55.5

Design	Top-1	Top-5
LC-A	50.9	79.8
LC-B	50.9	79.7
LC-C	51.5	80.2
BPLC	51.6	80.5
LC-E	51.3	79.9

Stages	# BPLC	Top-1	Top-5
res2	1	49.8	79.1
res2,res3	2	50.1	78.7
res2,res3,res4	3	50.2	79.0
res2,res3,res4,res5	4	50.2	79.2

(d) **Spatio-temporal input size.** The input size is formatted with (width$^2\times$time).

(e) **Various LC designs.** 16 LCs are set in the BQN.

(f) **Stage for adding BPLCs.** In each stage, we set one BPLC after its first residual block.
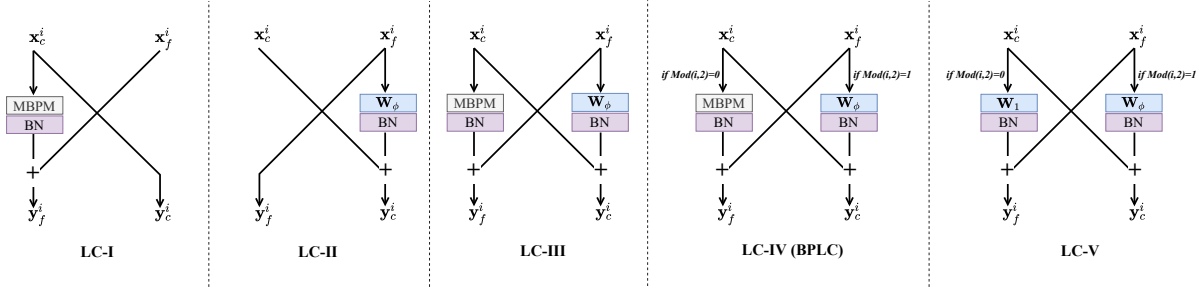

Figure 5: Diagrams of various lateral connection (LC) designs. Bilinear interpolation is used for resizing the feature maps when $\mathbf{x}_c^i$ and $\mathbf{x}_f^i$ do not have the same spatial size. i refers to the index of the residual block. $\mathbf{W}_\phi$ and $\mathbf{W}_1$ denote the weights of the linear transformation.

Efficiency and Effectiveness of the MBPM. We draw an apple-to-apple comparison between the proposed MBPM and other motion representation methods [1, 9, 22, 45, 50, 52]. The comparison results are shown in Table 1. The motion representations produced by these methods are used as inputs to the backbone network. The prediction scores are obtained by the average consensus of eight temporal segments [45]. More details about the implementations can be found in Appendix C. The proposed MBPM outperforms the other motion representation methods by big margins, while its computation is nearly negligible, which strongly demonstrates the high efficiency and effectiveness of the MBPM. Moreover, the two-stream fusion of "RGB+MBPM" has higher accuracy than the fusion of "RGB+Flow", according to the results in Table 2.

Generalization to different CNNs. The proposed MBPM is a generic plug-and-play unit. The performance of existing video models could be boosted by simply placing an MBPM after their input layers. Table 2 show that MobileNetV2 [33] and X3D-M [11] have steady performance improvement after being equipped with our MBPM.

Visualization Analysis. We provide the visualization results of four videos and their corresponding MBPM outputs in the top and bottom rows from Figures 4 (1)-(4). From these results it can be observed that the extracted representations are stable when jittering and other camera movements are present. Also the results from Figures 4 (1)-(4) show that MBPM not only suppresses the stationary information and the background movement, but also highlights the boundaries of moving objects, which are of vital importance for action discrimination. For example, in the "spinning poi" video, from Figure 4-(1), MBPM highlights the poi's movement rather than the movement of the background or that of the performer. More visualization results and visual comparison with other motion representation methods are provided in Appendix D.

Table 4: Results on Something V1. "N/A" indicates the numbers are not available. † denotes our reimplementation.

Method	Pretrain	Backbone	Frames×Crops×Clips	FLOPs	#Param.	Top-1 (%)	Top-5 (%)
NL I3D GCN [47]		3D R50	32×3×2	303G×3×2	62.2M	46.1	76.8
TRN$_{RGB+Flow}$ [53]		BNInception	(8+48)×1×1	N/A	36.6M	42.0	-
TSM$_{RGB+Flow}$ [29]	ImageNet	R50	(16+96)×1×1	N/A	48.6M	49.7	78.5
TEA [28]		R50	16×3×10	70G×3×10	24.4M	52.3	81.9
bLVNet-TAM [10]		bLR50	8×1×2	12G×1×2	25M	46.4	76.6
PAN$_{Full}$ [52]		TSM R50	40×1×2	67.7G×1×2	-	50.5	79.2
ir-CSN [42]	None	3D R152	32×1×10	74G×1×10	29.7M	49.3	-
TSM R50 [29]		R50	16×1×1	65G×1×1	24.3M	47.2	77.1
BQN		TSM R50	24×1×1	60G×1×1	47.4M	51.7	80.5
BQN	ImageNet	TSM R50	24×3×2	60G×3×2	47.4M	53.3	82.0
BQN		TSM R50	48×3×2	121G×3×2	47.4M	54.3	82.0
BQN		TSM R101	48×3×2	231G×3×2	85.4M	54.9	81.7
X3D-M† [11]	None	-	16×3×2	6.4G×3×2	3.3M	46.7	75.5
BQN	None	X3D-M	48×3×2	9.7G×3×2	6.6M	50.6	79.2
BQN	K400	X3D-M	48×3×2	9.7G×3×2	6.6M	53.7	81.8
BQN$_{En}$	ImageNet + K400	TSM R101 +X3D-M	(48+48)×3×2	241G×3×2	92M	**57.1**	**84.2**

5.3. Ablation Studies for BQN

BQN *vs*. Quiet+Busy. In order to evaluate the effectiveness of the proposed BQN architecture, we compare BQN with the simple fusion (Quiet+Busy), which mimics the two-stream model [34] by averaging the predictions of two pathways trained separately. Table 3a shows that the simple fusion of two individual pathways (Quiet+Busy) generates higher top-1 accuracy (50.3%) than the individual pathways, which indicates that the features learned by the Quiet pathway and by the Busy pathway are complementary. Surprisingly, BQN has 51.6% top-1 accuracy, which is 1.3% better than the Quiet+Busy fusion. The high-performance gain strongly demonstrates the advantages of the proposed BQN architecture.

Fusion strategies applied at the end of the Busy and Quiet pathways also influence the performance of BQN. Table 3c shows the results of different fusions. We observe that the average fusion gives the best result among the listed methods, while the concatenation fusion is second only to the averaging. Besides, placing the average fusion layer after the fully-connected layer is better than placing it before.

Effectiveness of the BPLC. We can set a maximum of up to 16 BPLCs in the BQN architecture when using TSM R50 [29] as the backbone[2]. For the BPLCs in stage res2, res3 and res4, we set the spatial kernel size of MBPM as 7×7, and the scale $\sigma = 0.9$. As for the stage res5, whose feature size is relatively small, the kernel size is therefore set to 3×3. Table 3f, illustrates that adding BPLCs to all processing stages is helpful for improving performance. From Table 3b, we can observe that the model performance

improves gradually as the number of BPLCs increases. The substantial performance gains demonstrate the importance of using BPLCs for BQN.

Lateral Connection (LC) Designs. In order to illustrate the rationality of the proposed BPLC design, we compare it with other LC designs. The diagrams of different LC designs are illustrated in Figure 5, where LC-I and LC-II are unidirectional, and LC-III is bidirectional. The results from Table 3e show that the bidirectional design LC-III has higher accuracy than the unidirectional designs LC-I and LC-II. Among the listed designs, the proposed BPLC, which reverses the information fusion direction back and forth, provides the highest accuracy. We also compare the BPLC with LC-V that does not contain an MBPM. As a result, LC-V shows lower accuracy than the BPLC, which demonstrates the importance of MBPM for the BPLC.

Spatial-temporal input size. In BQN, the Busy pathway takes as input the MBPM output, which has the same spatial size as the raw video clip, while the temporal size is one-third of the raw video clip length. Meanwhile, the Quiet pathway takes as input the complementary of the MBPM output, given by equation (4). Table 3d shows that with the same temporal size of 8 for the inputs, the spatial size combination of 160^2 and 256^2 for the Quiet and Busy, respectively, has slightly better top-1 accuracy (+0.1%) than the combination of 224^2 and 224^2 but saves 5.2 GFLOPs in computational cost. We also attempt to reduce the temporal input size of the Quiet pathway. However, this would result in a performance drop. One possible explanation is that due to the temporal average pooling in the Quiet pathway, the input's temporal size is already reduced to one-third of the raw video clip. An even smaller temporal size could fail to

---

[2]ResNet50 contains four stages, named res2, res3, res4, res5, respectively. These stages are composed of 3, 4, 6, 3 residual blocks, respectively.

Table 5: Comparison results on Kinetics400. We report the inference cost of multiple "views" (spatial crops × temporal clips). † denotes our reimplementation.

Method	Backbone	Frames × views	FLOPs	Top-1 (%)	Top-5 (%)
bLVNet-TAM [10]	bLR50	16×9	561G	72.0	90.6
TSM [29]	R50	16×30	2580G	74.7	-
STM [24]	R50	16×30	2010G	73.7	91.6
X3D-M† [11]	-	16×30	186G	75.1	92.2
SlowFast$_{4\times16}$ [12]	3D R50	32×30	1083G	75.6	92.1
ip-CSN [42]	3D R101	32×30	2490G	76.8	92.5
SmallBigNet [27]	R101	32×12	6552G	77.4	93.3
PAN$_{Full}$	TSM R50	40×2	176G	74.4	91.6
I3D$_{RGB}$ [3]	Inc. V1	64×N/A	N/A	71.1	89.3
Oct-I3D [5]	-	N/A×N/A	N/A	74.6	-
NL I3D [46]	3D R101	128×30	10770G	77.7	93.3
BQN	TSM R50	48×10	1210G	76.8	92.4
BQN	TSM R50	72×10	1820G	77.3	93.2
BQN	X3D-M	48×30	291G	77.1	92.5

preserve the correct temporal order of the video, and therefore harms the temporal relation modeling.

5.4. Comparisons with the State-of-the-Art

We compare BQN with current state-of-the-art methods on the four datasets. In BQN, the Quiet and Busy pathways' spatial input size is set to 160^2 and 256^2, respectively.

Results on Something-Something V1. Table 4 summarizes the comprehensive comparison, including the inference protocols, corresponding computational costs (FLOPs) and the prediction accuracy. Our method surpasses all other methods by good margins. For example, the multi-clip accuracy of BQN$_{24f}$ with TSM R50 is 7.2% higher than NL I3D GCN$_{32f}$ [47] while requiring 5× fewer FLOPs. Among the models based on ResNet50, BQN$_{48f}$ has the highest top-1 accuracy (54.3%), which surpasses the second-best, TEA$_{16f}$ [28], by a margin of +2%. Furthermore, our signal-clip BQN$_{24f}$ has higher accuracy (51.7%) than most other multi-clip models, requiring only 60 GFLOPs. By adopting a deeper backbone (TSM R101), BQN$_{48f}$ has 54.9% top-1 accuracy, higher than any single model. When using X3D-M as the backbone, BQN achieves the ultimate efficiency, possessing very low redundancy in both feature channel and spatio-temporal dimensions. BQN with X3D-M processes 4× more video frames than vanilla X3D-M, with only 50% additional FLOPs. Compared with TSM R50$_{16f}$, BQN with X3D-M trained from scratch produces 3.4% higher top-1 accuracy with the computation complexity of 14% of TSM R50$_{16f}$. The ensemble version **BQN**$_{En}$ achieves the state-of-the-art top-1/5 accuracy (57.1%/84.2%).

Table 6: Results on HMDB51 and UCF101. We report the mean class accuracy over the three official splits.

Method	Backbone	HMDB51	UCF101
StNet [17]	R50	-	93.5
TSM [29]	R50	73.5	95.9
STM [24]	R50	72.2	96.2
TEA [28]	R50	73.3	96.9
DI Four-Stream [1]	ResNeXt101	72.5	95.5
TVNet [9]	BNInception	71.0	94.5
TSN$_{RGB+Flow}$ [45]	BNInception	68.5	94.0
I3D$_{RGB+Flow}$ [3]	3D Inception	**80.7**	**98.0**
PAN$_{Full}$ [52]	TSM R50	77.0	96.5
BQN	**TSM R50**	**77.6**	**97.6**

Results on Kinetics400, UCF101 and HMDB51. Table 5 shows the comparison results on Kinetics400. For fair comparison, we only list the models with the spatial input size of 256^2. BQN$_{72f}$ with TSM R50 achieves 77.3%/93.2% top-1/5 accuracy, which is better than the 3D CNN-based architecture, I3D [3], by a big margin of +6.2%/3.9%. When BQN uses TSM R50 or X3D-M as the backbone, it consistently shows higher accuracy than SlowFast$_{4\times16}$. Particularly, BQN with X3D-M has 1.5% higher top-1 accuracy than SlowFast$_{4\times16}$, while requiring 3.7× fewer FLOPs. Meanwhile, BQN$_{72f}$ with TSM R50 is 2.7% better than Oct-I3D [5] for top-1 accuracy. The results on two smaller datasets, UCF101 and HMDB51, are shown in Table 6, where we report the mean class accuracy over the three official splits. We pretrain our model on Kinetics400 to avoid overfitting. The accuracy of our method is obtained by the inference protocol (3 crops×2 clips). BQN with TSM R50 outperforms most other methods except for I3D$_{RGB+Flow}$, which uses additional optical flow input modality.

6. Conclusion

This paper develops a novel video representation learning mechanism called Motion Band-Pass Module (MBPM). The MBPM can distill important motion cues corresponding to a set of band-pass spatio-temporal frequencies. We design a spatio-temporal architecture called Bus-Quiet Net (BQN), enabled by MBPM, to separately process busy and quiet video data information. The busy-quiet disentangling enables efficient video processing by allocating additional resources to the Busy stream and less to the Quiet. Our methods can also be used for video processing and analysis in various applications.

References

[1] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi. Action recognition with dynamic image networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(12):2799–2813, 2018. 2, 5, 6, 8

[2] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. Gcnet: Nonlocal networks meet squeeze-excitation networks and beyond. In *Proc. IEEE Int. Conference Computer Vision Workshops (ICCV-w)*, 2019. 2

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 4724–4733, 2017. 1, 2, 4, 8

[4] C. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. *arXiv preprint arXiv:1807.03848*, 2018. 2

[5] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 3435–3444, 2019. 2, 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 248–255, 2009. 4

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Longterm recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 2625–2634, 2015. 2

[8] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[9] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang. End-to-end learning of motion representation for video understanding. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 6016–6025, 2018. 2, 5, 6, 8

[10] Q. Fan, C. Chen, H. Kuehne, M. Pistoia, and D. Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *Advances Neural Information Process. Systems (NIPS)*, pages 2264–2273, 2019. 2, 7, 8

[11] C Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 203–213, 2020. 1, 2, 4, 5, 6, 7, 8

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proc. IEEE Conf. Computer Vision Pattern Recog. (CVPR)*, pages 6202–6211, 2019. 2, 4, 5, 8

[13] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3468–3476, 2016. 4

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 1933–1941, 2016. 2, 4

[15] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, volume 1, pages 5842–5850, 2017. 2, 4

[16] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 6546–6555, 2018. 1, 2

[17] D. He, Z. Zhou, C. Gan, F. Li, X. Liu, Y. Li, L. Wang, and S. Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proc. AAAI Conference on Artif. Intel.*, volume 33, pages 8401–8408, 2019. 8

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 770–778, 2016. 2, 5

[19] G. Huang and A. G. Bors. Learning spatio-temporal representations with temporal squeeze pooling. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2103–2107, 2020. 2

[20] G. Huang and A. G. Bors. Region-based non-local operation for video classification. In *Proc. Int. Conference on Pattern Recognition (ICPR)*, pages 10010–10017, 2021. 2

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 4700–4708, 2017. 2

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, volume 2, pages 2462–2470, 2017. 2, 5, 6

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conference Mach. Learn. (ICML), vol. PMLR 37*, page 448–456, 2015. 4

[24] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 2000–2009, 2019. 8

[25] A. Krizhevsky, I. Sutskever, and G Hinton. Imagenet classification with deep convolutional neural networks. In *Advances Neural Information Process. Systems (NIPS)*, pages 1097–1105, 2012. 2

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 2556–2563, 2011. 2, 4

[27] X. Li, Y. Wang, Z. Zhou, and Y. Qiao. Smallbignet: Integrating core and contextual views for video classification.

In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 1092–1101, 2020. 8

[28] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 909–918, 2020. 7, 8

[29] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 7083–7093, 2019. 1, 2, 4, 5, 7, 8

[30] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 2117–2125, 2017. 4

[31] A. Piergiovanni and S. Ryoo. Representation flow for action recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 9945–9953, 2019. 2

[32] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 5533–5541, 2017. 1, 2

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 4510–4520, 2018. 3, 5, 6

[34] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances Neural Information Process. Systems (NIPS)*, pages 568–576, 2014. 2, 7

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conference Learn. Representations (ICLR), arXiv preprint arXiv:1409.1556*, 2015. 2

[36] K. Soomro, Amir R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 4

[37] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 1390–1399, 2018. 2

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 1–9, 2015. 2

[39] M. Tan and Q. V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. Int. Conference Mach. Learn. (ICML), vol. PMLR 97*, pages 6105–6114, 2019. 2

[40] G. W Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *Proc. European Conference Computer Vision (ECCV)*, pages 140–153, 2010. 1, 2

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 4489–4497, 2015. 1, 2

[42] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Video classification with channel-separated convolutional networks. In *Proc. IEEE Int. Conference Computer Vision (ICCV)*, pages 5552–5561, 2019. 1, 2, 7, 8

[43] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 6450–6459, 2018. 1, 2

[44] J. Wang, A. Cherian, F. Porikli, and S. Gould. Video representation learning using discriminative pooling. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 1149–1158, 2018. 2

[45] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. European Conference Computer Vision (ECCV), vol LNCS 9912*, pages 20–36, 2016. 2, 5, 6, 8

[46] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 7794–7803, 2018. 2, 4, 8

[47] X. Wang and A. Gupta. Videos as space-time region graphs. In *Proc. European Conference Computer Vision (ECCV)*, pages 399–417, 2018. 7, 8

[48] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy tradeoffs in video classification. In *Proc. European Conference Computer Vision (ECCV), vol. LNCS 11219*, pages 305–321, 2018. 1, 2

[49] J. Yue-Hei Ng, M. Hauseknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proc. IEEE Conference Computer Vision Pattern Recog. (CVPR)*, pages 4694–4702, 2015. 2

[50] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Proc. Joint Pattern Recog. Symp., vol. LNCS 4713*, pages 214–223, 2007. 2, 5, 6

[51] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018. 2

[52] C. Zhang, Y. Zou, G. Chen, and L. Gan. Pan: Persistent appearance network with an efficient motion cue for fast action recognition. In *Proc. ACM Int. Conference Multimedia*, pages 500–509, 2019. 2, 5, 6, 7, 8

[53] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proc. European Conference Computer Vision (ECCV), vol LNCS 11205*, pages 803–818, 2018. 7