Fast Video Action Recognition from Compressed Domain Motion Information
基于压缩域信息及运动估计的快速视频动作识别

冯君逸
21821114
计算机科学与技术学院
fengjunyi@zju.edu.cn

摘要

近年来，运动 *(Motion)* 信息被广泛应用在各类视频理解的视觉任务中，在视频动作识别分类任务中，常见的算法是融合传统像素域的 *RGB* 网络特征与运动特征。在各类算法中，运动信往往特指光流 *(Optical Flow)*，但进行运动估计 *(Motion Estimation)* 所需的时间成本相对较高。因此，为了减少计算量，最近一些工作开始直接使用视频压缩域中的运动向量 *(Motion Vectors)* 以及残差 *(Residuals)* 信息作为模型输入。尽管提升了速度，但是由于运动向量的低分辨率以及高噪声，这些方法损失了一定模型精度。为了减少噪声的影响，本文引入一个轻量的类光流运动场生成网络 *(Flow-Like Motion Generator, FLMG)*，以运动向量为输入，生成精度更接近光流场的运动估计图。在公开数据集 *HMDB-51* 的实验结果表明，我们提出的网络在保证较高预测精度的同时，相比传统利用光流的多路网络极大提升了模型的运行速度。*(*本文代码已公开至 *https://github.com/Sixkplus/cv-project-action-recognition)*

1. 背景介绍

视频中包含丰富的视觉信息。不同于图像，视频中不但有每帧图片的像素域特征，还包含连续帧之间的时域运动信息。近年来的工作表明在视频分析问题 (如动作识别 [18, 34, 28]，动作定位 [1, 7, 39, 43]，以及时空异常检测 [21, 25] 等问题) 的相关算法中将运动信息引入模型对于性能提升具有很明显的效果。目前，这些领域的最优算法基本都遵循 Two-stream [28, 5, 33] 的结构，即包含两个 CNN 网络——一个以 RGB 帧作为输入，另一个以光流为输入，二者提取的特征在底层进行融合。此外，还有使用 3D-CNN 对视频进行时空特征提取的算法 [33, 32]。

一个困扰研究人员的问题是提取运动特征的光流估计算法以及 3D 卷积需要消耗较多的时间。因此，一方面，更快的基于 CNN 的光流估计网络 [9, 15] 被提出；另一方面，最近的工作 [41, 38, 42] 开始避免计算光流，取而代之地，它们使用视频压缩域中包含的运动信息去直接代替光流。具体地，它们使用 MPEG-4 [2] , H.264 [20] 等较新的视频编码标准中定义的运动向量 (Motion Vectors, MV) 和残差 (Residual, Res) 信息。最近，Wu 等人提出的 CoViAR [38] 模型充分利用了压缩域的多模态特征 (RGB, MV, Res) 设计了三个 CNN 网络，在避免额外解码操作的同时，充分利用了压缩域的运动特征；在该模型中，每个独立 CNN 网络分别对视频压缩域中三种模态的输入进行处理，并进行端到端的预测，最后，将每个独立网络的预测结果进行融合给出最终预测。 CoViAR 模型很好地利用了压缩域信息，可以达到很快的运行速度，但受限于运动向量的粗糙性，其精度不高，为了提升准确性，Wu 等人在其基础模型的基础上又加入了精细光流，但运行速度大大减慢。

基于 MV 的多分支模型精度不高的主要原因在于 MV 的粗糙性。一方面，MV 描述的是 16×16 大小的像素区域的运动，因此其真实分辨率较低；另一方

面，在视频压缩算法中，尽管 MV 的粗糙性不可避免，但有残差信息 Res 作为补偿；然而，基于 MV 的深度学习算法缺少对残差信息的有效使用。基于以上观察，参考 Shou 等人提出的 DMC-Net [27]，本文使用一种更准确、更全面地利用视频压缩域信息的视频动作视频算法。具体地，观察到视频压缩域中多模态运动特征 MV 与 Res 的特点与其互相之间的关联，我们引入一个轻量级的类光流运动场生成网络 (Flow-Like Motion Generator, FLMG)，以压缩域中原始运动向量 MV 作为主要输入，辅以残差信息 Res 进行细节补充，经过一个轻量级的 CNN 生成质量高于 MV 的类光流场的运动特征，并在此基础上设计动作识别网络。我们的工作及其效果总结如下：

- 复现 CoViAR 模型，加入并对其模型框架进行改进，使得整个模型不再依赖于光流提取算法，同时能达到非常快的速度。
- 我们提出了一个轻量级的类光流运动场生成网络 (Flow-Like Motion Generator, FLMG)，用来生成适用于动作识别任务的较高质量的类光流场，同时保证该模块比现有光流提取算法快数十倍至数百倍。
- 我们在 HMDB-51 [17] 数据集上测试了我们的算法，结果表明，我们的模型在保证快速运行的同时缩短了与基于精细光流模型之间的精度差距。

2. 相关工作

2.1. 视频动作识别

受到卷积神经网络 CNN 在图片分类任务中成功应用的启发，目前动作识别的前沿方法主要也使用 CNN 作为模型基础。著名的 Two-Stream 网络 [28] 采用两个独立的 2D-CNN 分别利用光流及 RGB 图片帧进行预测，并将两路结果进行融合。但区别于图像，视频本身具有特定的时域结构并且包含运动信息，这些信息对于人分辨动作起到关键作用。基于这些观察，以下工作对时域运动信息进行了更有效的建模，包括 3D-CNN [32, 5], Temporal Segment Network [36] 以及 Non-Local Network [37]。此外，动作识别领域的公认的一点是，即使模型本身已经包含了运动信息 (如 3D-CNN)，在算法中融合光流特征 [5, 33] 总会得到一定性能提升。

2.2. 压缩域信息与动作识别

由于光流信息需要额外的操作进行提取，因此，一系列工作开始使用视频压缩域中已有的运动信息替代光流。Zhang 等人[41, 42, 6]提出用 Motion Vector Stream 替代 Optical Flow Stream，但这些方法仍需完整解码出 RGB 图片，并且忽略了压缩域的其他模态特征。最近提出的 CoViAR [38] 模型充分利用了视频压缩域中的各种模态，大幅提升了识别速度；但该方法的性能与 Two-Stream 的方法相比仍有差距，这主要是由于 Wu 等人未对低分辨率、多噪声的 MV 进行处理便直接使用；此外，CoViAR 模型尽管利用了残差信息，但使用方法并未完全发挥该模态信息的特点。我们希望更好地利用压缩域的特征，对运动 MV 进行精炼，使得基于 MV 的轻量级模型能够达到接近 Two-Stream 模型的性能。

2.3. 运动表示与光流估计

传统的光流提取方法致力于对连续帧之间的像素位置差距进行建模 [3, 14, 40, 30]。传统光流算法受限于无法并行，运行效率较低，近年来，出现了一系列基于 CNN 的光流估计算法，包括 FlowNet系列 [9, 15]、SpyNet [24] 以及 PWC-Net [31]，这些基于 CNN 的光流算法在公开数据集 MPI Sintel [4]、KITTI [22] 上取得了更准确的结果，并且其中一些版本可以在高性能 GPU 上实时运行。

对于视频动作识别这一特定任务，光流这一运动信息可以提升模型的预测准确性，但最近 Sevilla 等人的工作 [26] 显示光流的精细程度与识别准确率相关性不是非常强。因此，开始有工作把重点放在生成更适合于动作识别任务的运动特征问题上。比如， Fan 等人 [10] 设计一个类似 TV-L1 运动估计的可微子网络，可以和下游的视频理解任务共同训练，得到更适合具体任务的运动特征；Ng 等人 [22] 在模型中加入了一个全卷积残差模块来预测光流。

与上述做法使用 RGB 图像帧作为输入不同，我们希望基于视频压缩域的运动向量以及残差信息，设计一个类光流的运动特征生成模块，目的是得到比 MV 更精细的适用于动作识别任务的运动特征；该模块可以与下游任务共同训练，达到比基于 MV 的 Two-Stream 方法更高的准确度。

3. 算法描述

在本章节，我们将对我们的模型进行详细描述，模型结构如图 3 所示。首先，我们将介绍视频压缩域特征的基本知识；其次，我们介绍模型的 Baseline, CoViAR [38]；接下来，我们介绍模型的核心模块——类光流运动场生成网络 (Flow-Like Motion Generator, FLMG)；最后，我们将给出整个模型在训练过程中的目标函数。

3.1. 视频压缩与编解码

在接下来的模型描述中，我们以 MPEG-4 Part2 [19] 为默认的视频压缩标准。在现有常用的视频压缩算法中，压缩域 (非码流) 主要包含三种模态的信息：I 帧的 RGB 信息、P 帧参考相应 I 帧的运动向量 MV，以及 P 帧对应的残差信息 Res。其中，I 帧是视频中的关键帧，压缩域中存储完整的信息；运动向量一般是基于 16×16 大小的像素块 (也称为宏块) 的像素域运动信息，描述某一个块相对于最近的 I 帧在图像坐标系中的移动量；残差信息是在像素域 P 帧的原始帧与经过 MV 运动补偿后相应图像之间的 RGB 之差。在以下实验中，我们设定 $GOP = 12$，即相邻两个 I 帧之间有 11 个 P 帧。假设原视频的分辨率为 $W \times H$，则对应的 MV 大小为 $W \times H \times 2$，Res 的大小为 $W \times H \times 3$。另外需注意，由于 MV 描述的是宏块的运动，因此其实际有效分辨率是原大小的 $1/16$。视频解压缩得到三种模态信息的过程如 图 1 所示。

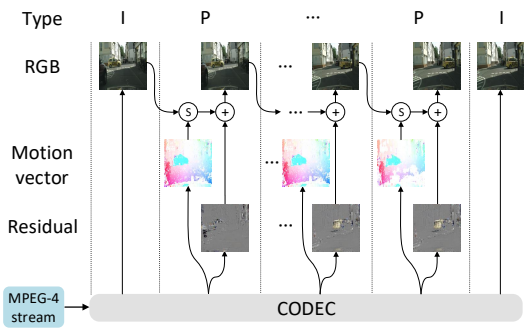

图 1. 压缩视频解码得到三种模态数据的示意图。

3.2. 基本模型

我 们 以 Compressed Video Action Recognition (CoViAR) [38] 为基础，简单修改后实现我们的模型。该模型示意图见 图 2。其主要构成为三个 CNN Stream，分别以视频压缩域中的三种不同模态的特征作为输入，并作出预测，并最终进行结果融合。我们选择 ResNet [13] 为网络的主体结构；根据观察，MV 的分辨率较低、Res 的信息集中在物体边缘处，因此，对于这两种模态的输入，我们使用 ResNet-50 进行处理，而对于蕴含信息量更多的 RGB 图像，我们使用 ResNet-152 作为主体结构。

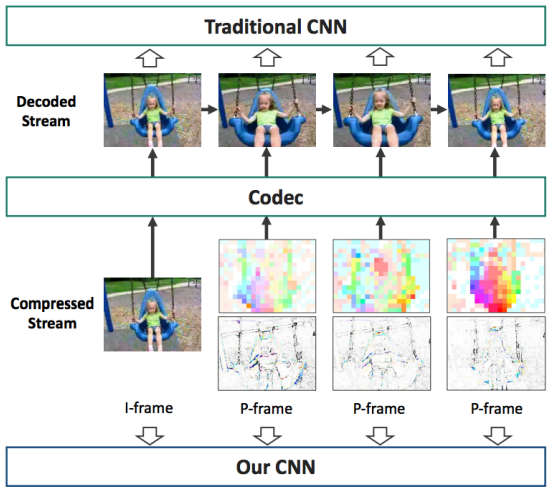

图 2. Coviar [38] 模型示意图，该图取自其论文。其中，'Our CNN' 包含三个并行的 CNN，分别以压缩域中的三种模态信息作为输入。

3.3. 类光流运动场生成网络

为了提升 MV 的精确性，使其能对最终识别准确率起到更好的促进作用，我们在 CoViAR 模型的基础上引入了一个新的轻量级的类光流运动场生成网络 (Flow-Like Motion Generator, FLMG)。FLMG 的示意图对应 图 3 中生成 New Mv 的部分，该模块由6层简单的 $Conv - ReLu$ 模块构成。注意到压缩域中的另一种特征 Res 主要描述视频中运动物体的边界，为了使生成的运动特征更加精细，我们把 Res 与 MV 合并后作为 FLMG 的输入，用 TV-L1 [40] 的输出结果作为 FLMG 的 Ground Truth 进行监督，以常用的 L2 MSE 作为损失函数。直观地，我们的目标是该模块能够对 MV 进行细节增强及平滑，因此在 FLMG 中，我们也使用残差模块对每层网络的输入输出进行相加，希望学习到增强的部分。

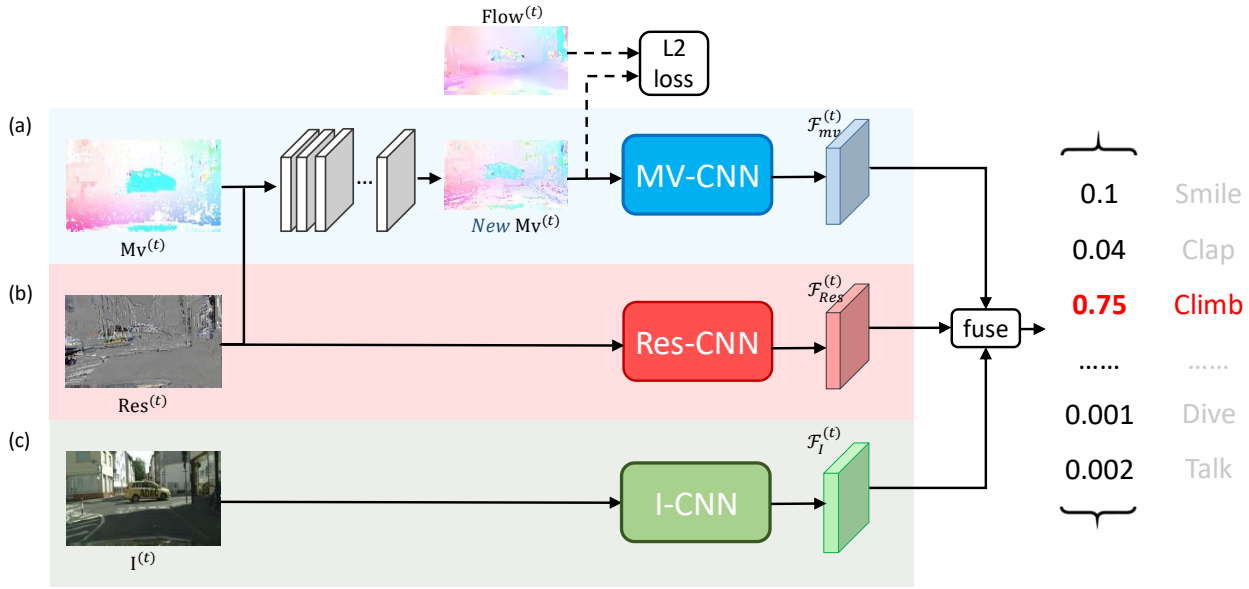

图 3. 改进后的模型示意图。三种模态的信息对应三个 CNN。其中，FLMG 模块对应 (a) 分支中的前半部分。

3.4. 损失函数与训练目标

模型的损失函数主要由两部分组成，即 FLMG 生成类光流部分的损失以及最终的分类损失。接下来将详细介绍这两部分损失函数。以下描述中假设 x 为特定帧，x_{MV} 为对应帧的运动向量，x_{Res} 为相应的残差信息，C 为运动种类数目。

3.5. 生成函数损失

假设我们的生成网络 FLMG 对应的函数为 $\phi_{FLMG}(x_{MV}, x_{Res})$，光流生成函数为 $\phi_{FlOW}(x)$，则我们对生成模型的损失定义为：

$$\mathcal{L}_{FLMG} = ||\phi_{FLMG}(x_{MV}, x_{Res}) - \phi_{FlOW}(x)||_2^2 \quad (1)$$

3.6. 分类函数损失

对于三个独立的 CNN 网络，其各自输出的 One-hot 形式的特征为$\mathcal{F}_I, \mathcal{F}_{MV}, \mathcal{F}_{Res} \in \mathbb{R}^C$ 以及最终融合后的预测结果 $\mathcal{F}_{fuse}$，我们对这几个预测结果均进行损失计算，损失的计算均使用 Softmax Cross Entropy 作为损失函数，假设样本 x 对应的真实标签为 c_g，则分类部分的损失函数为：

$$\mathcal{L}_I = -\sum_x \log \frac{e^{\mathcal{F}_I(c_g)}}{\sum_{c=1}^{C} e^{\mathcal{F}_I(c)}} \quad (2)$$

$$\mathcal{L}_{MV} = -\sum_x \log \frac{e^{\mathcal{F}_{MV}(c_g)}}{\sum_{c=1}^{C} e^{\mathcal{F}_{MV}(c)}} \quad (3)$$

$$\mathcal{L}_{Res} = -\sum_x \log \frac{e^{\mathcal{F}_{Res}(c_g)}}{\sum_{c=1}^{C} e^{\mathcal{F}_{Res}(c)}} \quad (4)$$

$$\mathcal{L}_{fuse} = -\sum_x \log \frac{e^{\mathcal{F}_{fuse}(c_g)}}{\sum_{c=1}^{C} e^{\mathcal{F}_{fuse}(c)}} \quad (5)$$

加权的分类损失函数汇总为：

$$\mathcal{L}_{cls} = \mathcal{L}_{fuse} + \alpha_1 \cdot \mathcal{L}_I + \alpha_2 \cdot \mathcal{L}_{MV} + \alpha_3 \cdot \mathcal{L}_{Res} \quad (6)$$

最终，整个模型的待优化目标函数为：

$$\mathcal{L}_{tot} = \mathcal{L}_{cls} + \gamma \cdot \mathcal{L}_{FLMG} \quad (7)$$

4. 实验

在本章节，我们将给出实验的细节、给出定性定量的结果，并与最新的动作识别算法进行对比。

4.1. 数据集和衡量指标

接下来的实验部分，本文使用 HMDB-51 [17] 为实验数据集。该数据集包含6766个视频，共有51个动作分类。HMDB 公开了三个数据的训练/测试划分，考虑到时间受限，本文的实验采用其中的 HMDB-Split2 为划分标准进行准确度测试。在视频识别领域，常用的数据集还包括 UCF-101 [29] 以及 Kinetics [4]。

HMDB 中的所有视频均为单标签数据，因此我们使用该领域通用的 top-1 accuracy 作为分类效果的衡量指标。

4.2. 实现细节

4.2.1 训练细节

我们对于基础模型 CoViAR [38] 的实现与原论文稍有不同，为了充分利用预训练模型，我们对 I, MV, Res 分别使用在 ILSVRC 2012-CLS 分类数据集 [8] 上预训练好的 ResNet152, ResNet50, ResNet50 作为分类网络骨架（原论文中对于 MV, Res 采用的是 ResNet18 模型）；以 OpenCV 中内置的 TV-L1 光流算法 [40] 监督 FLMG 的训练；最终结果融合的过程，我们使用各支 CNN 预测结果 One hot 向量的加权平均。为了统一视频格式，我们设置视频编解码标准为 MPEG-4，Group of Pictures (GOP) 为12，把所有视频分辨率缩放至 340×256，随机选取 224×224 的区域并且做随机翻转以进行数据增强。训练过程中，我们使用 Adam 优化方法 [16]，对于预训练部分的网络参数，设置学习率为 $1e-5$，其余部分参数学习率为 $1e-3$，至损失稳定后统一学习率。

4.2.2 测试细节

测试阶段，我们遵循 CoViAR [38] 中的设置：在待测视频中均匀地选取25帧；每帧选取5个 224×224 的区域，并进行翻转；所有250个预测分数 $(25 \times 5 \times 2)$ 取平均，作为最终的预测结果。

4.3. 结果分析

4.3.1 训练过程分析

使用 Tensorboard 记录训练过程中多个分类目标函数的下降趋势，如图 4 所示:

4.3.2 FLMG

我们对比加入 FLMG 前后模型在 HMDB-51 [17] 上的准确率，结果如见表 1 ：

Model	Top-1 Accuracy
I-CNN	49.12
MV-CNN	43.38
Res-CNN	47.38
I-CNN + MV-CNN + Res-CNN	55.67
Ours(I-CNN + MV-CNN + Res-CNN + FLMG)	**61.19**

表 1. 各个模型在 HMDB-51 数据集的准确率结果对比。

实验结果表明，FLMG 模块的加入使得生成的运动特征能够更好地适应动作识别任务，帮助产生更有判别力的特征。

4.3.3 动作识别效果

接下来，我们将我们的模型与主流动作识别算法进行对比，结果见表 2 。

Model	Top-1 Accuracy
基于解码后 RGB 视频帧的算法	
ResNet-152 [11]	46.7
ResNet-50 [11]	48.9
I3D RGB [5]	49.8
C3D [32]	51.6
Res3D [32]	54.9
ActionFlowNet [23]	56.4
ArtNet [35]	70.9
TVNet [10]	71.0
基于 Two-stream 的算法	
Two Stream [12]	65.4
I3D [5]	66.4
TSN [36]	69.4
CoViAR + Optical Flow [38]	70.2
基于视频压缩域特征的算法	
CoViAR [38]	59.1
DMC-Net [27]	62.8
Ours(I-CNN + MV-CNN + Res-CNN)	55.67
Ours(I-CNN + MV-CNN + Res-CNN + FLMG)	61.19

表 2. 各个模型在 HMDB-51 数据集的准确率结果对比。

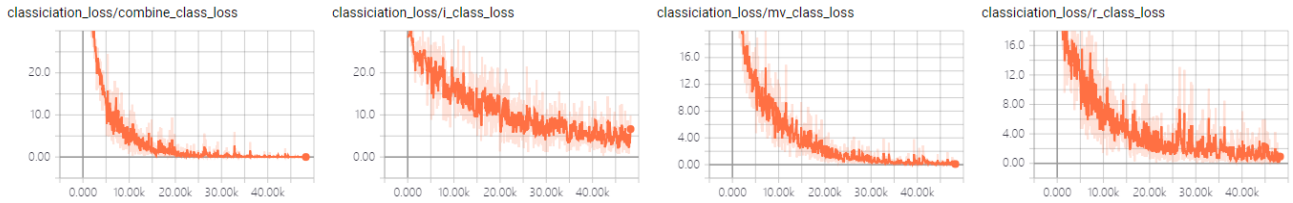

图 4. 分类损失随训练迭代数目增长的下降趋势。其中，combine, mv 及 r 模态的损失下降较快，关键帧 RGB 图像 i 对应的损失下降较慢，原因是该 CNN stream 为 ResNet152，层数较多，收敛速度更慢；且静态图像中包含的运动信息不多。

5. 总结

本文提出了一种更有效地利用视频压缩域多模态特征进行视频动作识别的算法。具体地，我们引入了一个轻量级的类光流运动特征生成模块。实验结果表明，我们加入的模块在基本不影响运行速度的条件下实现了准确度的提升，使得整个模型学习得到了更有判别力的运动特征。受计算资源及时间限制，本文的实验还不够完备，未来将探索模型在更多数据集上的性能，并尝试在生成模型部分加入对抗损失以取得更好的效果。

参考文献

[1] Y. Bai, H. Xu, K. Saenko, and B. Ghanem. Contextual multi-scale region convolutional 3d network for activity detection. *arXiv preprint arXiv:1801.09184*, 2018.

[2] F. Bellard, M. Niedermayer, et al. Ffmpeg: A complete, cross-platform solution to record, convert and stream audio and video, 2010.

[3] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61(3):211–231, 2005.

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] A. Chadha, A. Abbas, and Y. Andreopoulos. Compressed-domain video classification with deep neural networks: "there's way too much information to decode the matrix". In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1832–1836. IEEE, 2017.

[7] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[10] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang. End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6016–6025, 2018.

[11] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[15] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[18] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008-IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8. IEEE Computer Society, 2008.

[19] D. Le Gall. Mpeg: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–59, 1991.

[20] D. LEGALL. A video compression standard for multimedia applications. *Commun. ACM*, 34:226–252, 1993.

[21] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[22] M. Menze, C. Heipke, and A. Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 2015.

[23] J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018.

[24] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017.

[25] D. Ryan, Denman, C. Fookes, and S. Sridharan. Textures of optical flow for real-time anomaly detection in crowds. In *2011 8th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 230–235. IEEE, 2011.

[26] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition*, pages 281–297. Springer, 2018.

[27] Z. Shou, Z. Yan, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, X. Lin, and S.-F. Chang. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. *arXiv preprint arXiv:1901.03460*, 2019.

[28] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[29] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[30] D. Sun, S. Roth, J. Lewis, and M. J. Black. Learning optical flow. In *European Conference on Computer Vision*, pages 83–97. Springer, 2008.

[31] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[34] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[35] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.

[36] 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 *European conference on computer vision*, pages 20–36. Springer, 2016.

[37] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[38] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Compressed video action recognition.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6026–6035, 2018.

[39] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.

[40] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.

[41] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2016.

[42] 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.

[43] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.