Protecting Patient Privacy During Action Recognition Using Extreme Low-Resolution

Samuel Kwong
Stanford University
samkwong@stanford.edu

Boxiao Pan
Stanford University
bxpan@stanford.edu

Abstract

Human behavior analysis is of growing importance in clinical environments. In order for a system to automatically analyze the behavior of humans, it must first be able to recognize their actions. With recent deep learning advances in computer vision models, 3D convolutional neural networks have been successful in human action recognition from video input. In healthcare settings, however, patient privacy must also be protected in the process of analyzing behavior. In this work, we take an approach of only feeding the model extreme low-resolution video input to anonymize the humans present. The biggest challenge is to effectively capture the action cues within the drastically downsampled video. We also experiment using a novel approach of cross-resolution knowledge distillation, incorporating discriminative cues present in high-resolution videos into their low-resolution counterpart, while only taking in low-resolution videos at test time. Through experiments run on UCF-101 and Jester, we show the limitations of this approach and address the difficulty of deducing high-resolution information from low-resolution videos. Code for this work is publicly available at https://github.com/samjkwong/ELR-Action-Recognition.

1 Introduction

In recent years, there have been many advances in machine learning approaches for clinician and patient behavior analysis in clinical settings. One such approach is vision-based, which benefits from new innovations in deep Convolutional Neural Network (CNN) architectures. Automatic human behavior analysis via computer vision is growing in importance in healthcare, especially in recognizing and analyzing ICU patient mobility, patient care tasks and clinicians' procedures [6, 20, 21, 22]. It is of vital necessity, however, to protect the privacy of patients when deploying these computer vision models.

This poses a problem, since state-of-the-art video action recognition models typically require high-resolution videos such that the subject is identifiable. Previous works have used depth videos in order to preserve the temporal motions of the subjects while hiding their identities [6, 20]. Depth videos, however, require significant overhead and setting up multiple depth sensors in patient room settings. RGB video input, in contrast, provides color information to help with classification and is typically a lower cost solution compared to depth sensors, since only one camera is required and a single RGB camera for this task would cost less than a single depth sensor. A remaining challenge would be to anonymize the video input taken by an RGB camera.

A popular approach for video anonymization for privacy protection in action recognition is to only use extreme low-resolution input videos at test time, such that the subject is unidentifiable. Prior works [1, 11, 12, 19] use a resolution of 12×16 on popular video datasets such as UCF-101 [14], a common dataset used for action recognition benchmarking.

Since low-resolution (LR) videos contain very limited amounts of information compared to their high-resolution (HR) counterparts, action prediction from only LR input tends to drastically underperform when compared to that on HR input. It is possible that a key to improving performance is to leverage discriminative queues learned from HR input during training while still only evaluating on LR input at test time. Various approaches such as a fully-coupled two-stream spatio-temporal network [19] and spatio-temporal attention transfer [1] have been used to meet this goal. We run experiments on a novel approach of cross-resolution knowledge distillation.

In this work, we establish a means of testing extreme low-resolution action recognition, experimenting on two video datasets and testing if our proposal of cross-resolution knowledge distillation is able to further improve performance. Our contributions are as follows:

- We run baselines on UCF-101 [14] and Jester [10] using I3D [2], and achieve 92.3% and 95.2%, respectively.
- We create two extreme low-resolution action recognition datasets, UCF-101-ELR and Jester-ELR, by downsampling each frame of videos from UCF-101 and Jester to 12×16. Experiments on UCF-101-ELR and Jester-ELR result in 69.2% and 92.3%, respectively, demonstrating a drop in performance when training and testing only on LR input.
- We test whether it may be possible to use knowledge distillation between HR and LR branches and propose a new architecture to this end. We achieve 67.9% on UCF-101-ELR and 92.4% on Jester-ELR. Based on our experiments and ablation study, results show that it is in fact difficult to distill information from different video resolutions and that, in some scenarios, it hurts performance. To the best of our knowledge, no other work has experimented with this approach. Thus, we show that cross-resolution knowledge distillation does not seem to improve extreme low-resolution action recognition tasks.

2 Related Work

General Video Classification. Spatio-temporal reasoning is one of the main topics for video understanding. With the success of CNNs on image recognition [8], many deep neural architectures are proposed correspondingly in the space-time domain. C3D [15] and I3D [2] construct hierarchical spatio-temporal understanding by performing 3D convolutions. A two-stream network [3] receives additional motion information by fusing an extra optical flow branch. A Temporal Segment Network (TSN) [16], on the other hand, takes advantage of the fact that huge redundancy exists between adjacent video frames via sparse frame sampling. While arguing previous methods fail to capture long-term temporal dependency, several recent works [23, 17, 4, 18] make attempts to extend their horizons to a wider range. Specifically, Temporal Relational Reasoning (TRN) [23] extends TSN by considering multi-level sampling frequency. Non-local network (NL) [17] explicitly creates long-term spatio-temporal links among features. SlowFast network [4] exploits multiple time scale by creating two pathways with different temporal resolution. Alternatively, Long-term feature bank (LFB) [18] directly stores long-term features and later correlates them with short-term features.

Extreme Low-Resolution Action Recognition. Because of the importance of extreme low-resolution action recognition in far-view surveillance and privacy-preserving behavior analysis, it has gained increasing attention and various approaches have been proposed to solve the problem. Ryoo *et al.*[11] proposes a data augmentation approach by applying different transformations to the same low-resolution video, and learns a tighter decision boundary with the augmented samples. A follow-up work [12] further leverages positive / negative sampling and minimizes an extra contrastive loss. Xu *et al.*[19] incorporates optical flow to better leverage the temporal cues. Very recently, Bai *et al.*[1] proposes a spatio-temporal attention transfer pipeline to help the model attend to spatio-temporally important regions.

Knowledge Distillation. Knowledge distillation was first proposed in [7], where the distillation is performed from a large model to a small one by minimizing the KL divergence between their logits distributions. Later, Lopez-Paz *et al.* [9] generalizes distillation to incorporate privileged information, which is some additional information that is available during training but not accessible during testing. One application of this approach is by treating the extra modality as the privileged information [5]. In our case, we view the high-resolution videos as the privileged information available during training stage, and perform knowledge distillation from them to the the low-resolution video branch.

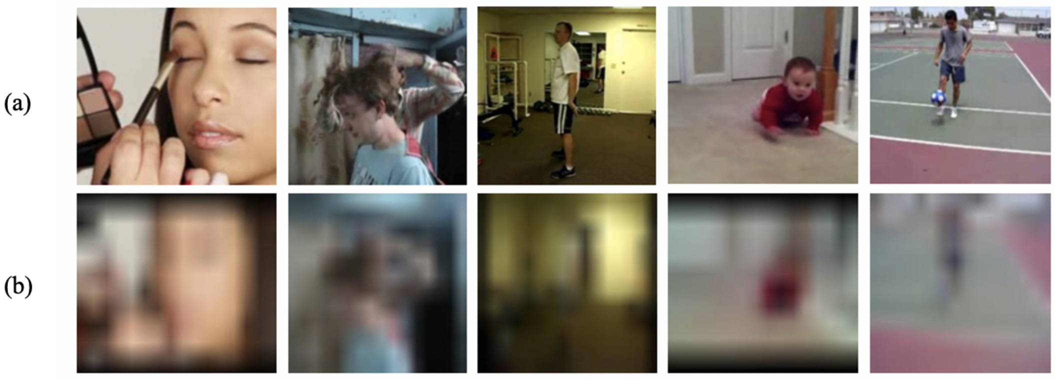

Figure 1: Data samples from UCF-101 (a) and UCF-101-ELR (b). Note that the original frame sizes vary (which also leads to the black rim in some samples in the bottom row) and for presentation purpose, we crop the center area of frames from UCF-101.

3 Data

UCF-101 Dataset. UCF-101 [13] is a popular video action recognition dataset that has 101 action classes and 13320 video samples, split into around 9.5K for training and 3.7K for testing. The dataset videos are sampled as 25 FPS RGBs.

We use Split 1 of UCF-101 to build our extreme low-resolution action recognition dataset. Specifically, we first downsample each frame to 12×16 and then upsample it back to 224×224 with bilinear interpolation, effectively retaining the anonymization from 12×16 resolution. We name this dataset UCF-101-ELR and use it throughout our experiments. Fig. 1 shows examples from UCF-101 and UCF-101-ELR, and it can be observed that subjects are anonymized in UCF-101-ELR.

Jester Dataset. Jester [10] is another video action recognition dataset that has 27 action classes for hamd gestures performed in front of a webcam and 148K video samples, split into 118.5K for training and 14.7K for validation. The dataset videos are sampled as 12 FPS RGBs. As with UCF-101, we also build our extreme low-resolution action recognition dataset, Jester-ELR, and use it throughout our experiments.

4 Approach

4.1 Overview

The framework of our proposed method for cross-resolution knowledge distillation is shown in Fig. 2. During the training process, our goal is to distill the discriminative cues learned from HR videos into the LR pathway. To this end, we leverage both the HR and LR videos and train a separate feature extractor, which is an I3D network [2], as well as a classifier, for each pathway. Then we distill the knowledge from the HR branch to the LR branch by minimizing a KL divergence [7] loss between the logits of the two branches.

4.2 Feature Extraction

Given a sequence of RGB frames $\{x_1, x_2, \ldots, x_T\}$, we extract 3D features out of them with I3D [2] as $F_{3D} = \{v_1, v_2, \ldots, v_L\}$ with $v_l \in \mathbb{R}^{d_{3D}}$, where $d_{3D} = 1024$.

Model	Dataset	Test Acc.	
		Stride 1	Stride 4
RGB-I3D	UCF-101	92.3	91.8
RGB-I3D	UCF-101-ELR	68.5	69.2
Cross-Resolution KD	ECF-101-ELR	66.6	67.9

Table 1: Action classification accuracy (%) on UCF-101 and UCF-101-ELR test sets.

4.3 Kullback-Leilber Divergence

The Kullback-Leilber (KL) divergence measurement defines how much one probability distribution differs from another. For probability distributions P and Q, the KL divergence from Q to P is as follows:

$$D_{KL} = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{1}$$

KL divergence is especially helpful in attaining cross-network knowledge distillation, between a teacher and student network. In our method, we assign our teacher network as the HR branch and our student network as the LR branch.

4.4 Cross-Resolution Knowledge Distillation

We make an important observation that when viewed at the same time period, motion in an LR video is to be less ambiguous due to the reference in the HR video counterpart. Hence, we argue that there is shared information in the representation between high and low-resolution videos, and the former can aid the recognition of the latter by instilling knowledge. We therefore propose distilling information learned by the logits in the HR branch of our network.

Our approach uses late fusion on the classification logits of both HR and LR branches by minimizing the KL divergence on the two probability distributions. Formally, let $P^{high}(x)$ be the probability distribution (pre-Softmax logits) across the class set C from the HR branch, and $P^{low}(x)$ be the distribution from the LR branch. This method minimizes the KL divergence loss:

$$L_{logit} = - \sum_{x \in C} P^{high}(x) \log \left(\frac{P^{low}(x)}{P^{high}(x)} \right) \tag{2}$$

5 Experiments on UCF-101-ELR and Jester-ELR

We perform experiments on UCF-101-ELR. The task, as in UCF-101, is action recognition. We implement our LR approach that serves as our baseline for reference. Then we train an HR branch that exhibits the performance gap when using original resolution videos and extreme low-resolution videos from UCF-101. We also then use the HR branch to implement and train a cross-resolution distilled model from our method, to see if the resulting model learns from both HR and LR branches.

5.1 Implementation Details

Our first experiment is training I3D directly on UCF-101-ELR without using any knowledge distillation. Specifically, we use I3D pretrained on ImageNet & Kinetics as in [2] and then finetune it on UCF-101-ELR. The preliminary results are shown in Table 1. We sample at the center of the video. If the input video is too short, we perform padding by replicating the last frame. We also report the result of I3D on the original high-resolution UCF-101 dataset in Table 1.

From the results we can see that the baseline I3D performs much worse on UCF-101-ELR than on the original UCF-101, with a performance gap of -23.8% when using a temporal stride of 1. For our next step, we implement the knowledge transfer procedure to test if knowledge distillation has an effect on the action recognition.

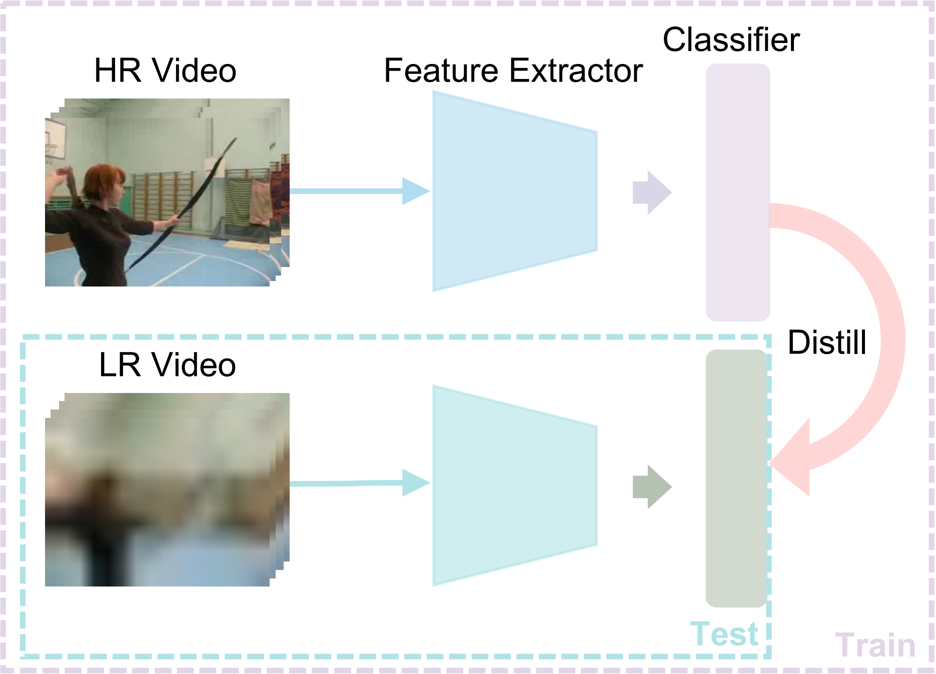

Figure 2: Overview of the proposed framework for cross-resolution knowledge distillation. During training, it takes both a high-resolution (HR) and a low-resolution (LR) video, and passes them subsequently into a feature extractor and a classifier. Knowledge distillation is then performed from the HR branch to the LR branch. At test time, only the LR branch is used and classifies the action in the video correspondingly.

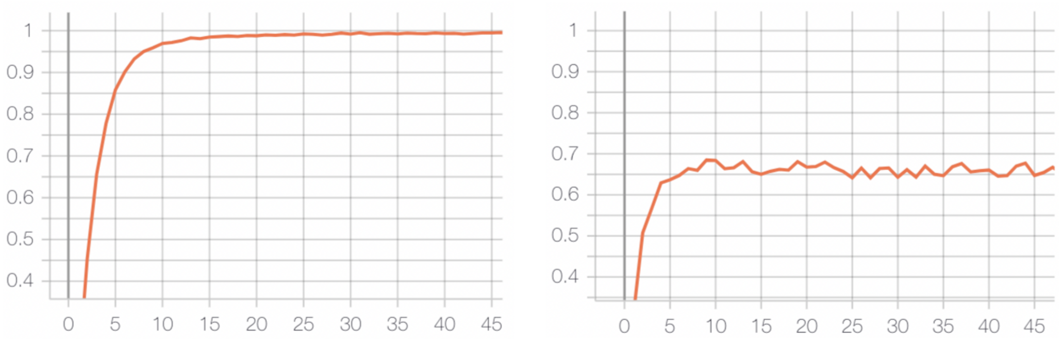

Figure 3: Accuracy curves for LR training on UCF-101-ELR for train (left) and test (right).

Training. For both datasets, we use Adam optimizer with an initial learning rate of 0.0001, batch size of 16 for optimization and early stopping. For UCF-101-ELR, we sample at 25 FPS and with a temporal stride of 1, with a fixed clip size of 64 frames. For Jester-ELR, we sample at 12 FPS and with a temporal stride of 1, with a fixed clip size of 32. For both datasets, to prevent overfitting we use temporal augmentation and sample each video starting from a random offset at least 64 frames from the end of the video. If the video is shorter than 64 frames for UCF-101-ELR or 32 frames for Jester-ELR, we perform padding by replicating the last frame.

Inference. At test time, we maintain the same configuration as during training except we temporally sample from the center of a video. Sampling from a fixed temporal range enables consistent testing throughout experiments, and we choose the center since the performed action is more likely to occur in the middle of the video sequence as opposed to the very beginning or very end.

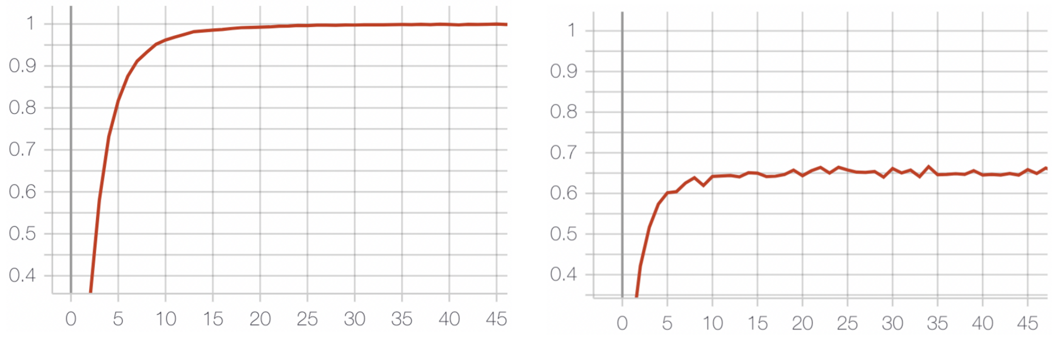

Figure 4: Accuracy curves for Cross-Resolution KD training on UCF-101-ELR for train (left) and test (right).

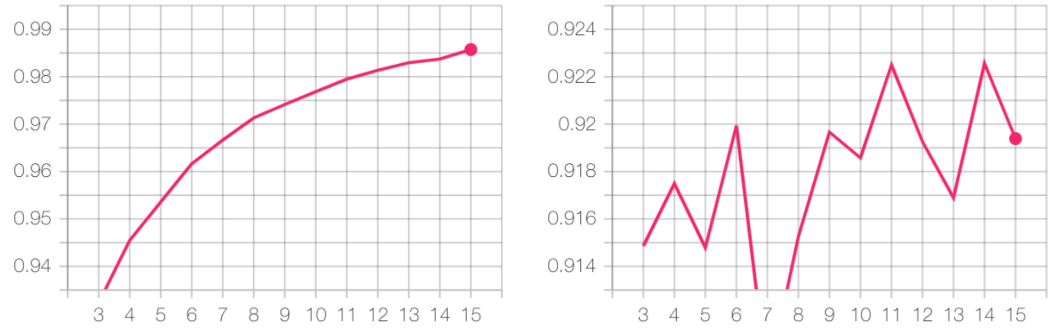

Figure 5: Accuracy curves for LR training on Jester-ELR for train (left) and val (right).

5.2 Ablation Experiment

Temporal Sampling. We run experiments on two different configurations in terms of temporal sampling, which is how large of a stride we use when sampling frames from each video. Results in Table 1 show that a shorter stride of 1 is more successful than a longer stride of 4 for HR input. For LR input, on the other hand, a longer stride of 4 is more successful than a shorter stride of 1. This could be because as the resolution decreases in the video input, a larger temporal window can aid the model in differentiating certain actions from others. Thus, a larger stride is more desirable in the LR configuration. At a normal resolution, a larger temporal window may not outweigh the cost of lost motion information by skipping intermediate frames. Thus, a shorter stride is more desirable in the LR configuration.

6 Conclusion

In clinical environments, in order for a system to automatically analyze the behavior of humans, it must first be able to recognize their actions while protecting patients' privacy. In this work, we take an approach of only feeding the model extreme low-resolution video input to anonymize the humans present. We establish a means of testing extreme low-resolution action recognition and experimenting on two video datasets. Additionally, we show that cross-resolution knowledge distillation is not able to further improve performance on extreme low-resolution action recognition. Although knowledge distillation has been shown to work on various tasks, it seems to not be successful for cross-resolution data transfer. Future work can further analyze why cross-resolution is a difficult task to distill information from one resolution branch to another.

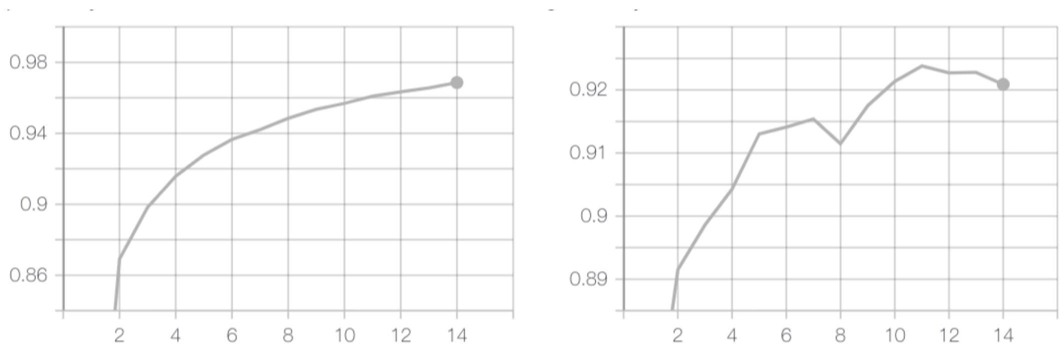

Figure 6: Accuracy curves for Cross-Resolution KD training on Jester-ELR for train (left) and val (right).

Model	Dataset	Val Acc
RGB-I3D	Jester	95.2
RGB-I3D	Jester-ELR	92.3
Cross-Resolution KD	Jester-ELR	92.4

Table 2: Action classification accuracy (%) on Jester and Jester-ELR test sets.

References

[1] Y. Bai, G. Dai, and L. Chen. Extreme low resolution activity recognition with spatial-temporal attention transfer. *arXiv preprint arXiv:1909.03580*, 2019.

[2] 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.

[3] 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.

[4] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

[5] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[6] A. Haque, M. Guo, A. Alahi, S. Yeung, Z. Luo, A. Rege, J. Jopling, L. Downing, W. Beninati, A. Singh, T. Platchek, A. Milstein, and L. Fei-Fei. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. 2017.

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[10] J. Materzynska, G. Berger, I. Bax, and R. Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[12] M. S. Ryoo, K. Kim, and H. J. Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[19] M. Xu, A. Sharghi, X. Chen, and D. J. Crandall. Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1607–1615. IEEE, 2018.

[20] R. F. J. J. e. a. Yeung, S. A computer vision system for deep learning-based detection of patient mobilization activities in the icu. npj Dgit. Med., 2019.

[21] S. Yeung, A. Alahi, A. Haque, B. Peng, Z. Luo, A. Singh, T. Platchek, A. Milstein, and F.-F. Li. Vision-based hand hygiene monitoring in hospitals. In *AMIA*, 2016.

[22] S. Yeung, N. L. Downing, F.-F. Li, and A. Milstein. Bedside computer vision - moving artificial intelligence from driver assistance to patient safety. New England Journal of Medicine, 2018.

[23] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.