

Dissected 3D CNNs: Temporal Skip Connections for Efficient Online Video Processing

Okan Köpüklü¹, Stefan Hörmann¹, Fabian Herzog¹, Hakan Cevikalp², Gerhard Rigoll¹

¹ Technical University of Munich

² Eskisehir Osmangazi University

Abstract

Convolutional Neural Networks with 3D kernels (3D CNNs) currently achieve state-of-the-art results in video recognition tasks due to their supremacy in extracting spatiotemporal features within video frames. There have been many successful 3D CNN architectures surpassing the state-of-the-art results successively. However, nearly all of them are designed to operate offline creating several serious handicaps during online operation. Firstly, conventional 3D CNNs are not dynamic since their output features represent the complete input clip instead of the most recent frame in the clip. Secondly, they are not temporal resolution-preserving due to their inherent temporal downsampling. Lastly, 3D CNNs are constrained to be used with fixed temporal input size limiting their flexibility. In order to address these drawbacks, we propose dissected 3D CNNs, where the intermediate volumes of the network are dissected and propagated over depth (time) dimension for future calculations, substantially reducing the number of computations at online operation. For action classification, the dissected version of ResNet models performs 74-90% fewer computations at online operation while achieving $\sim 5\%$ better classification accuracy on the Kinetics-600 dataset than conventional 3D ResNet models. Moreover, the advantages of dissected 3D CNNs are demonstrated by deploying our approach onto several vision tasks, which consistently improved the performance.

1. Introduction

Convolutional Neural Networks (CNNs) have dominated the majority of computer vision tasks ever since AlexNet [25] won the ImageNet Challenge (ILSVRC 2012 [34]). In order to harness a similar performance as 2-dimensional (2D) CNNs achieved on image-based tasks, 3-dimensional (3D) CNNs have been proposed by adding an additional depth dimension to convolutional and pooling layers. However, 3D CNNs have significantly more parameters and

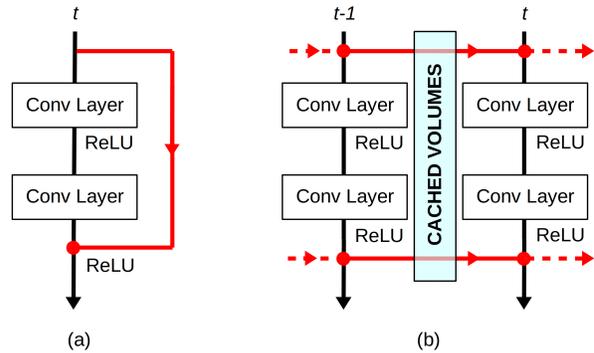


Figure 1: Comparison of spatial skip connections (a) first proposed in [12] and temporal skip connections (b) proposed in this work. At every iteration, only the computations for the most recent frame performed. Afterwards, intermediate volumes from the skip connections are updated to be used for the next iteration. This way, recomputation of previous frames is saved. Skip connections are denoted with red lines.

computations at inference time than their 2D counterparts making them more challenging to train and prone to overfitting. The overfitting problem is resolved with the availability of large scale video datasets, such as Kinetics [3], Sports-1M [19]. Nevertheless, computational cost remains as the biggest drawback of 3D CNNs.

Currently, the primary trend in video recognition tasks is to increase network performance by building deeper and wider 3D CNN architectures [11, 9, 3]. However, these architectures are typically designed to operate offline, ignoring the requirements of online operation. Firstly, most of the 3D CNNs deploy temporal downsampling to reduce the computational cost at the later stages of the network and provide translation invariance (in the time dimension) to the internal representation. This causes the network to become non-dynamic (i.e. the final decision of the network might be due to any frame in the input clip, not due to the latest in-

roduced frame), which is of utmost importance for online operation. Moreover, the resulting network is not temporal resolution-preserving. Secondly, 3D CNNs are typically built to work with a fixed number of input frames. Therefore, online operating frameworks usually use 3D CNNs in a sliding window, either with small temporal stride [20, 22] or larger stride [32]. In the former case, there is a severe resource waste due to reprocessing frames in the overlapping regions, which are already processed in the previous timestamps. In the latter case, there is an information loss since relations between some of the frames are not exploited. These issues make most of the 3D CNNs unsuitable for online operation.

In order to address the limitations mentioned above, we propose a novel 3D CNN architecture, Dissected 3D CNNs (D3D), by incorporating temporal skip connections. Skip connections are first proposed in ResNets [12] to overcome the issue of vanishing/exploding gradients. Spatial skip connections, which are depicted in Fig. 1(a), can be in the form of summation [12] or concatenation [15, 30]. As opposed to spatial skip connections, we propose temporal skip connections to create a network for efficient online operation. The general idea of the proposed architecture is depicted in Fig. 1(b). Intermediate volumes are always stored in a cache, and only the computations for the new available frame are performed at each iteration. After the computations, the previous cached volumes are replaced with the most recent intermediate feature volumes coming from the skip connections. This way, the volumes in dissected 3D CNN architecture are propagated without calculating them repeatedly. We incorporate 3D convolutions since we apply concatenation operation in the depth dimension at the skip connections. Although summation is also possible at temporal skip connections, we will show in our ablation study that temporal information is lost with the summation operation, which leads to inferior results. Moreover, spatial skip connections are still applicable on top of temporal skip connections.

To obtain the networks final decision, dissected 3D CNN architecture still needs a spatiotemporal modeling mechanism at the end. Although the conventional way of using a fully connected layer is a valid option, a Recurrent Neural Network (RNN) block can also be applied. The RNN block makes the D3D architecture independent of the number of input frames and performs better, as shown in our ablation study. Moreover, any 3D CNN architecture can be converted to its dissected version. Overall, Dissected 3D CNNs bring the following advantages:

- D3Ds provide frame-level features. Hence they are dynamic.
- D3Ds operate at any number of input frames.
- A large number of computations are saved at online operation. D3D versions of ResNet-18,50,101 perform

74-90% less computation at online operation while achieving $\sim 5\%$ better classification accuracy compared to conventional ResNet models on Kinetics-600 dataset.

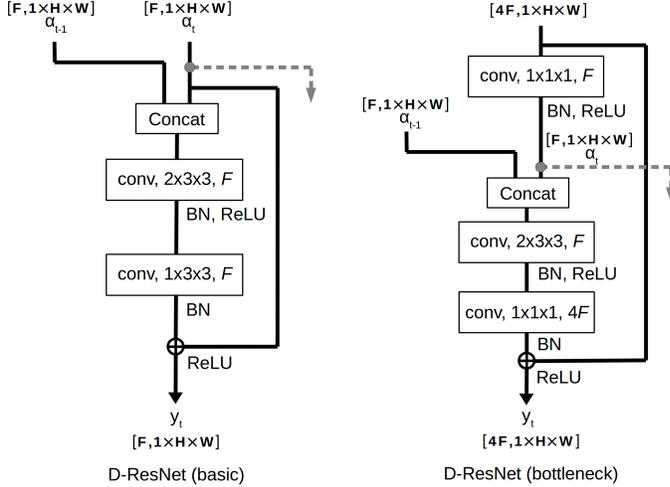
- Any frame-level task can leverage from D3D architecture if the frames are obtained from continuous video streams.

The remaining part of the paper is organized as follows: In Section 2, the related work is presented. Section 3 introduces the D3D architecture and elaborates training and evaluation processes. Section 4 presents experiments and results on various video-based computer vision tasks. Finally, Section 5 concludes the paper.

2. Related Work

Video-based computer vision tasks. The proposed D3D is designed to propagate spatiotemporal information with frame-level correspondence. Therefore, any task requiring to process continuous video streams can benefit from D3D including action/activity recognition with datasets UCF-101 [38], HMDB [26], Kinetics [3]; video object detection task such as ImageNet VID [34]; spatiotemporal action localization task such as Atomic Visual Actions (AVA) dataset [10]; video object tracking (VOT) task such as [24, 45]; multi-object tracking (MOT) such as [7]; video person re-identification task such as MARS (Motion Analysis and Re-identification Set) dataset [52]; gait recognition task such as Casia-B dataset [50]; video face recognition such as YouTube Faces [43] and many other tasks. Currently, state-of-the-art architectures either use only spatial content by processing the input frame-by-frame ignoring the temporal content [4, 42, 2] or utilize offline-trained 3D CNN architectures in a non-dynamic way [20, 1]. By utilizing our proposed D3D architecture, all video-based computer vision tasks can incorporate temporal information.

3D CNN architectures. Ji *et al.* propose a 3D CNN architectures for the first time in [17]. Ever since then, there have been plenty of 3D CNN architectures to achieve better accuracies at video classification tasks such as C3D [39], I3D [3], R(2+1)D [40], P3D [33], SlowFast [9], etc. The effect of dataset size is investigated in [11] together with the performance of widely-used architectures such as ResNet [12], DenseNet [15], ResNext [46]. In [21], 3D versions of popular resource-efficient architectures are investigated for video classification tasks. However, all these architectures are designed for offline operation and do not meet the requirements for online operation, as they operate with a fixed number of input frames. Moreover, the number of floating point operations (FLOPs) is in the order of 10s-100s GFLOPs at inference time, which is too costly for online operation.



Layer	ResNet-18	ResNet-{50,101}
block	basic	bottleneck
conv1	conv(3×7×7), stride (1, 2, 2) F:64	
pool	MaxPool(1×3×3), stride (1, 2, 2)	
conv2_x	N:2, F:64	N:3, F:64
conv3_x	N:2, F:128	N:4, F:128
conv4_x	N:2, F:256	N:{6, 23}, F:256
conv5_x	N:2, F:512	N:3, F:512
conv_last	—	conv(1×1×1), stride (1, 1, 1), F:512
	global average pooling, spatiotemporal modeling	

Figure 2: Basic and bottleneck blocks used in ResNet architecture. F , BN , and $ReLU$ denote the number of feature maps (i.e. channels), batch normalization [16], and rectified linear unit, respectively. $Concat$ denotes concatenation at depth dimension while \oplus denotes to element-wise addition.

Table 1: Dissected ResNet architectures. F is the number of feature channels corresponding in Fig. 2, and N refers to the number of blocks in each layer.

Online video processing architectures. For gesture recognition, Molchanov *et al.* propose to use 3D CNN to extract features followed by an LSTM for online recognition [32]. However, this approach is *near-dynamic* since the 3D CNN processes non-overlapping 8-frame clips. Köpüklü *et al.* propose to use a two-level hierarchical framework for online gesture recognition [20]. This architecture is again *near-dynamic* since the detector also takes 8-frame clips with a sliding window. For spatiotemporal action localization task, [36, 18] propose to use a detector to obtain frame-level detections and create action tubes with further post-processing. However, these methods make use of optical flow modality in order to incorporate motion information, which requires a substantial amount of computation. In [23], the YOWO (you only watch once) architecture is proposed, where spatiotemporal and fine-spatial features are concurrently extracted via 3D and 2D CNNs, and actions are detected on the key-frame. YOWO is a *dynamic* architecture in this regard. However, YOWO is not watching once since it operates using a sliding window for continuous videos, and 15 frames of a 16-frame clip have already been processed (*watched*) in the previous step. So there is a serious amount of repetitive computation at online operation, which can be avoided. The closest work to ours is [35], in which Singh *et al.* propose to decompose a 3D convolutional block into a 2D spatial convolution followed by a recurrent unit for temporal modeling. However, in this work, convolutions are performed in 2D (i.e. depth dimension of the convolutional kernels are always 1) and temporal information is captured only with recurrent units. Moreover, putting a recurrent unit at each layer of the network is

too costly in terms of computation complexity. To the best of our knowledge, D3D is the first architecture proposing to propagate intermediate volumes of the complete 3D CNN architecture to reduce the computational complexity during online operation.

3. Methodology

In this section, we first elaborate on the D3D architecture details, which reduces the computational complexity substantially during online operation. Secondly, we mention possible options for spatiotemporal modeling. Finally, training details are described.

3.1. Dissected 3D CNN Architecture

In order to demonstrate the advantages of the proposed D3D architecture, we have created the dissected version of the ResNet family (named as D-ResNet) and compared its performance with the conventional ResNet family as in [21]. The details of the proposed D-ResNet models are shown in Table 1 and Fig. 2. Spatial downsampling is performed at *conv1*, *pool*, *conv3_1*, *conv4_1*, and *conv5_1* with a stride of 2. No temporal downsampling is employed. Unlike the ResNet architecture, we reduced the depth dimension of the initial convolutional layer of the basic block and the middle convolutional layer of the bottleneck block to 2 since we cache only previous intermediate volumes. We also modify the second convolutional layer of the basic block and set its depth dimension to 1. These modifications lead to parameter reduction of $\sim 50\%$ on D-ResNet-18 and $\sim 23\%$ on D-ResNet-50,101 compared to conventional 3D ResNet architectures.

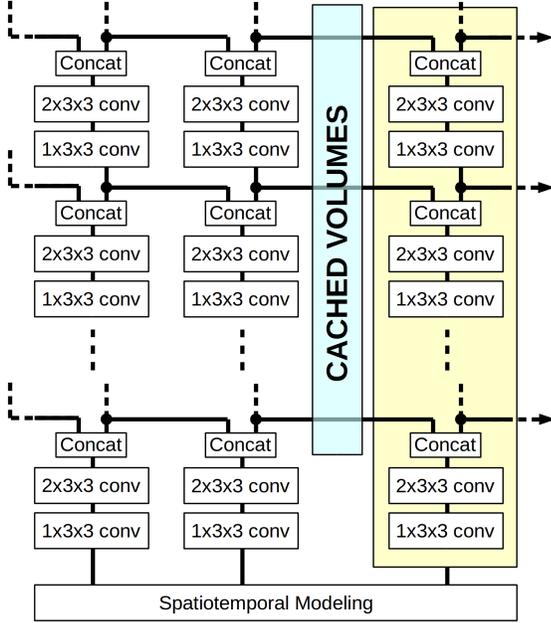


Figure 3: Proposed Dissected 3D CNN architecture using basic D-ResNet block. Spatial skip connections are excluded for the sake of simplicity.

An illustration of Dissected 3D CNN architecture with basic D-ResNet block is shown in Fig. 3. The primary motivation to create such an architecture is to dispense with the recomputation of already processed frames of the video stream during online operation. For that, intermediate volumes of the architecture are stored in a cache (blue region in Fig. 3) and used at inference. Throughout the inference, previous intermediate volumes are replaced with the current ones to be used in the next iteration. Therefore, only the computations within the yellow region in Fig. 3 are performed at online operation. Moreover, the designed D3D architecture does not employ (i) temporal downsampling and (ii) padding from right to ensure dynamic online operation.

At the network’s input, the current frame together with the previous two frames are passed to the network in order to capture pixel-wise motion information, which is critical for motion intensive datasets such as Jester dataset [31]. At the first iteration, the initial frame is replicated since there are no previous frames. Same padding is also applied at *concat* operations for the first iteration as the cache for the intermediate volumes is empty. For D-ResNet-50,101 architectures, an additional *conv_last* block is used in order to reduce the output feature dimension from 2048 to 512. So, all D-ResNet architectures produce a 512-dimensional feature vector for every frame. After obtaining frame-level features, a spatiotemporal modeling mechanism is required to produce class-conditional probabilities, which is explained in the next section.

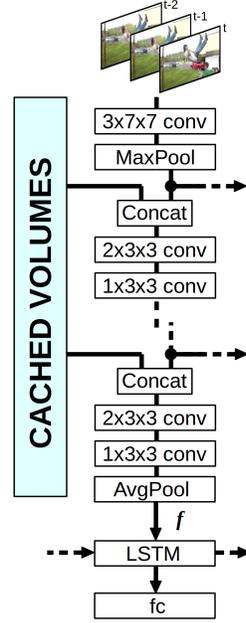


Figure 4: Purely dynamic D-ResNet-18 architecture. Spatial skip connections are excluded for the sake of simplicity.

3.2. Spatiotemporal Modeling Mechanism

The typical approach for spatiotemporal modeling is to conclude the network with a fully connected (fc) layer. This approach is also how we trained our architectures from scratch. However, the fc layer at the end of the network requires a fixed number of frames as input. Moreover, the dynamicity condition of the architecture is not met since the decision is made with all output features.

In order to achieve a purely dynamic system, we have considered two popular RNN blocks: Long Short-Term Memory (LSTM) [14] and gated recurrent unit (GRU) [5]. However, joint end-to-end training of the feature extraction and RNN blocks is not feasible due to the computational and memory complexity of back-propagating through the long video, as described in [44]. To this end, we have extracted the output features f (before the fully connected layer - see Fig. 4) of all video frames for the training and test set and trained the recurrent blocks separately. For example, each video in the Kinetics dataset lasts around 10 seconds, which makes 250 frames if the video is recorded with 25 fps. After applying the recurrent block, an fc layer is used at the last output of the recurrent block to map the hidden feature map to the number of classes. We have named the resulting network as purely dynamic D-ResNet-18 architecture since the network produces a decision using the most recent frame at every iteration. Purely dynamic D-ResNet-18 architecture is shown in Fig. 4. In the experiments section, we will validate the advantages of recurrent spatiotemporal modeling techniques.

Model	Skip Connection	Params	MFLOPs	St-Modeling	Accuracy (%)
D-ResNet-18	None	11.02M	438	fc	58.74
D-ResNet-18	Summation	11.02M	438	fc	58.40
D-ResNet-18	Concatenation	15.74M	602	fc	61.41

Table 2: Performance Comparison for different temporal skip connections at online operation on the Kinetics-600 validation set. The number of FLOPs and parameters are calculated excluding the spatiotemporal modeling mechanism.

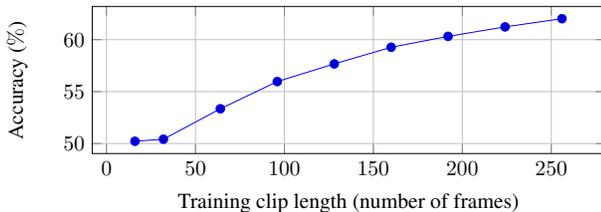


Figure 5: Influence of using different clip lengths at training on the accuracy of the Kinetics-600 validation set when training a D-ResNet-18-lstm.

Layer	1-layer	2-layer	3-layer
GRU	61.08	61.43	61.38
LSTM	61.10	62.02	60.83
fc	61.41		

Table 3: Accuracy on the Kinetics-600 validation set for different spatiotemporal modeling mechanisms using D-ResNet-18 architecture.

3.3. Implementation Details

Learning: We initially train our D3D architectures with fc layer at the end. 19 frames are fed to the network, but only the last 16 output features are used for loss computation. Moreover, the initial frames are solely utilized to properly initialize cached intermediate volumes. Stochastic Gradient Descent (SGD) is applied with standard categorical cross-entropy loss as an optimizer. The largest fitting batch size is selected for mini-batch size, which is typically in the order of 128 clips. The networks are trained from scratch with a learning rate initialized with 0.1 and reduced 3 times with a factor of 10^{-1} when the validation loss converges. For temporal augmentation, clips are selected from a random position in the video. For spatial augmentation, clips are selected from a random spatial position with a randomly selected scale from $\{1, \frac{1}{2^{1/4}}, \frac{1}{2^{3/4}}, \frac{1}{2}\}$ in order to perform multi-scale cropping as in [11].

For the training of the RNN blocks, we again use SGD with identical learning rates. However, we apply different augmentation schemes. First, the number of input features is selected randomly between [16, 'number of frames in the video'] and padded with zero to obtain a fixed size of input size for all videos. In this way, the RNN blocks can learn all *short*-, *medium*- and *long*-range dependencies. Moreover, videos are down-sampled by 2, 3 and 4 with probabilities of 30%, 14% and 11%, respectively. We also replaced random parts of the input features with noise to enable RNN blocks to ignore unrelated parts of the input. In order to increase regularization, we also leverage Gaussian noise with zero mean and 0.005 variance at the input features and 0.3 dropout at the hidden layers of RNN blocks. For the hidden

layers of RNN blocks, dimension is set to 1024.

Recognition: Kinetics-600 clips are selected by a sliding window with stride of 1 for fc spatiotemporal modeling. Afterwards, class scores are averaged for all the clips. For RNN blocks, the complete input is fed to the network and the last output of the RNN block is used for the final prediction.

Implementation: Network architectures are implemented in PyTorch. Our code and pretrained models will be made publicly available¹.

4. Experiments

4.1. Video Classification Task

Comparison of different temporal skip connection operations: We first compare the performance of different temporal skip connection operations. Table 2 shows the comparison of applying summation, concatenation and no temporal skip connections on D-ResNet-18 architecture. For the sake of fairness, at each iteration all networks receive the current frame together with the two previous frames as input and apply a 3D convolution layer as the first operation. For summation and no temporal skip connection, 2D convolution layer is applied afterward, whereas for concatenation temporal skip connection, a 3D convolution layer is used since volumes are concatenated along the depth dimension. Although using a 3D convolution layer increases the number of parameters and floating point operations, concatenation achieves the best performance with a margin of $\sim 2.7\%$.

¹<https://github.com/okankop/Dissected-3D-CNNs>

Model	Params	MFLOPs	Speed (ms)	St-Modeling	Accuracy (%)
3D ResNet-18 [21]	32.97M	5556	3.00	fc	57.65
3D ResNet-50 [21]	43.01M	6780	5.46	fc	63.00
3D ResNet-101 [21]	82.06M	10610	7.04	fc	64.18
D-ResNet-18	15.74M	602	0.33	fc	61.41 +3.76
D-ResNet-50	33.12M	1337	0.75	fc	67.35 +4.35
D-ResNet-101	62.12M	2760	1.46	fc	68.78 +4.60
D-ResNet-18	15.74M	602	0.33	LSTM	62.02 +4.37
D-ResNet-50	33.12M	1337	0.75	LSTM	68.22 +5.22
D-ResNet-101	62.12M	2760	1.46	LSTM	69.17 +4.99

Table 4: Comparison of D-ResNet architecture with conventional ResNet architecture over offline classification accuracy, number of parameters, computation complexity (FLOPs) at online operation on the Kinetics-600 validation set. The number of FLOPs and parameters are calculated excluding the spatiotemporal modeling mechanism. For each architecture, the speed refers to single inference time measured using NVIDIA Titan XP GPU for a batch size of 8.

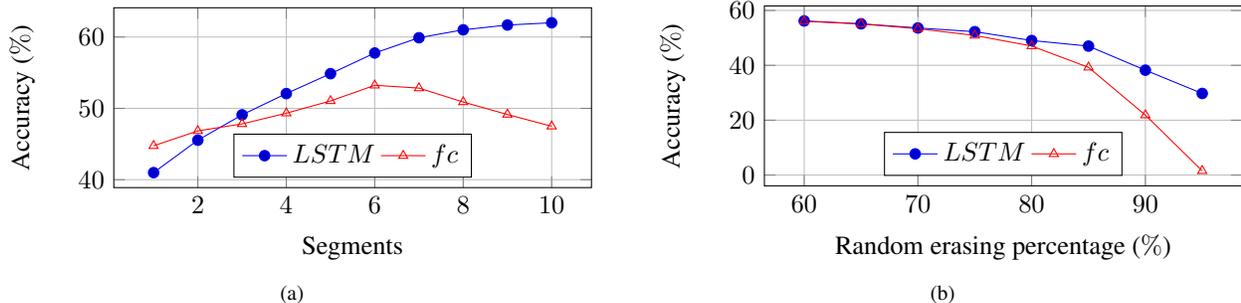


Figure 6: Causality analysis of deployed spatiotemporal modeling mechanisms. In (a), videos are separated into ten equal segments and network outputs at each segment are averaged for *fc* and *LSTM*. In (b), the network outputs at the middle parts of the videos are replaced with the Gaussian noise.

We would like to note that summation does not bring any performance gain and even performs slightly worse than no temporal skip connection. We infer that this is due to the loss of temporal information after the summation operation.

It is also interesting to see that D-ResNet-18 with no skip connection achieves even better than conventional 3D ResNet-18 architecture in Table 4. This contradicts the findings of [40], where f-R2D achieves 1.3% worse accuracy than R3D. Our only difference from f-R2D in [40] is that we apply a 3D convolution layer at the first convolution operation, which was enough to capture necessary motion information to outperform R3D. Besides, we can conclude that preserving temporal-resolution in the network (i.e. not applying temporal downsampling) increases classification performance, although this also increases the computation and memory load at inference time.

Analysis of different spatiotemporal modeling mechanisms: We investigate the performance of applying fc, LSTM or GRU as spatiotemporal modeling mechanism. Ta-

ble 3 shows the comparison of fc with LSTM and GRU for different number of hidden layers. Both recurrent blocks perform better with 2 hidden layers while LSTM achieves the best performance. Hence, from this point onwards, we always use two layers for LSTM.

Effect of different clip lengths on training recurrent blocks: At the training of RNN blocks, clip length plays an important role on the final classification performance. We have investigated the effect of different clip lengths at training time on the classification performance for D-ResNet-18-lstm, as illustrated in Fig. 5. The results clearly show that a longer clip leads to higher classification accuracy. The reason is that LSTMs can learn the important/unimportant features and store/remove them in their cell state easier when they observe longer clips.

Performance comparison of D-ResNet architectures with different depths: Comparative results are shown in Table 4. As usual, increasing network depth yields

Model	St-Modeling	Accuracy (%)
3D ResNet-18 [21]	fc	93.34
D-ResNet-18	fc	94.58 +1.24

Table 5: Comparison of dissected and conventional ResNet-18 architectures on the Jester validation set. Both architectures take 16-frames input (downsampling of 2 is applied) with 112×112 spatial resolution.

higher accuracies. Moreover, D-ResNet performs 74-90% less computation at online operation while achieving $\sim 5\%$ better classification accuracy compared to conventional ResNet models on Kinetics-600. This is since D3D uses the previous computations efficiently by caching the intermediate volumes of the network. We must also note that D-ResNet architectures also have fewer parameters compared to conventional ones.

Causality analysis of D-ResNet-18 architecture: The essential property of an online system is that the architecture should be causal. To validate the causality of the proposed D3D, we designed two tests. Firstly, we make a segment-level classification test, where we have divided input videos into ten equal parts and outputs are averaged within each segment. Fig. 6(a) shows the comparison of fc and LSTM. Since fc treats each clip independently and the middle parts of the videos are typically more informative, a bowed curve is achieved. On the other hand, LSTM stores the relevant features in its cell state over time, leading to increased accuracy with rising segment numbers. Keeping in mind that an entirely causal system should improve monotonically, D3D satisfies this criterion. Secondly, we have replaced the middle parts of the videos with the Gaussian noise. As we increase the erased percentage from the middle part of the videos, accuracy drops linearly for both fc and LSTM till 60% erasure. If we keep increasing the erasure percentage, it becomes more and more important to use the information at the beginning and end of the videos jointly. Therefore, LSTM outperforms fc more and more as the erasure percentage increases. Specifically, with an erasure percentage of 95%, fc achieves 1.53% accuracy, whereas LSTM achieves 29.74% accuracy.

4.2. Gesture Recognition Task

Gesture recognition can be viewed as a very similar task to action recognition task. In action recognition, although it is still necessary to capture motion patterns, the network especially needs to capture spatial patterns. For example, In the Kinetics-600 dataset, there are nine different “eating something” classes where “something” is one of “burger, cake, carrot, chips, doughnut, hotdog, ice cream, spaghetti, watermelon”. For the correct classification, the network

Model	Accuracy (%)	mAP
2D ResNet-50	80.8	69.0
D-ResNet-50	81.3 +0.5	69.1 +0.1

Table 6: Comparison of our D-ResNet-50 architecture with 2D ResNet-50 on the validation set of the MARS dataset.

must recognize the objects in the videos correctly. On the other hand, the spatial content in gesture videos are similar: A person in front of a camera performing a hand gesture. For the correct classification, the motion of the hand must be captured by the network.

To inspect the D3D architectures ability to capture motion patterns, we have experimented with the Jester dataset [31], which is the largest available hand gesture dataset currently. Training details are kept exactly the same as previous settings. In Table 5, D-ResNet-18 achieves 1.24% more classification accuracy than conventional 3D ResNet-18.

4.3. Video Person Re-Identification (ReID) Task

Person Re-identification aims to match a queried data with its true owner in the gallery set. In video person ReID, both the query and gallery are person tracklets, which usually consist of a varying number of frames. Most state-of-the-art approaches leverage 2D CNN architectures for video ReID [28, 27, 37]. However, 2D CNN architectures process individual frames independently, hence they cannot incorporate temporal information between frames. In this section, we demonstrate that our proposed D3D architecture can increase the performance over 2D CNNs.

Our person ReID architecture is as follows. Given input video clips, a backbone network extracts features for each frame and these features are averaged to get final feature representing the given input clip. We utilized classification loss and triplet loss in order to train the network. For the classification loss, we consider person identities as category-level annotations and train a linear layer followed by a softmax operation to get class-conditional probabilities. Then, our classification loss \mathcal{L}_C is the cross entropy error between the predicted classes and the ground truth classes. For the triplet loss \mathcal{L}_T , our data loader randomly selects N video clips for each person, which is used for hard sample mining [13]. The final loss is $\mathcal{L} = \mathcal{L}_C + \mathcal{L}_T$. The architecture is trained end-to-end using the final loss \mathcal{L} . In our experiments, we used $N = 4$ and each clip contains 4 frames at training time. At test time, we have loaded all frames in person videos to get final video features.

In our experiments, we have used the MARS dataset [52] for performance evaluation. For the backbone network in the architecture described above, we have compared the conventional 2D ResNet-50 with our D-ResNet-50 architecture. Both models are inflated from ImageNet

pretrained model. We trained the networks for 150 epochs using Adam optimizer with an initial learning rate 0.0003, which is divided by 10 every 60 epochs. In the MARS dataset, all person detections are already cropped, hence there is no pixel-wise correspondence at consecutive frames in tracklets. Therefore, we used single frames at the input of the D-ResNet-50 architecture. The comparative results are shown in Table 6. D-ResNet-50 architecture manages to capture discriminative motion information of identities, possibly gait-related information, which slightly increases the performance.

4.4. Video Face Recognition Task

In the domain of video face recognition, typical approaches [29, 41, 8] leverage the features obtained by training on big datasets containing still images followed by simple average pooling of the features without emphasis on the quality of every frame. More sophisticated approaches combine the feature extraction network to aggregate the features based on their importance with a feature aggregation network [49, 48, 47, 53]. However, temporal information is discarded as frames are treated as an unordered set of faces. Compared to these approaches, our D3D architecture can cope with this task while only consisting of one single network.

Before training the network, we preprocess the VoxCeleb2 dataset [6] by extracting 3 frames per clip, which are aligned using facial landmarks extracted using the MTCNN [51] and cropped to 112×112 pixels. First, we pretrain a 2D ResNet-18 with a 256-dimensional bottleneck layer on single image recognition on the VoxCeleb2 dataset using cross entropy loss with Adam optimizer, 50% dropout, an initial learning rate of 0.05 and a batch size of 100 for 50 epochs. We decided against pretraining on a bigger dataset containing still images, as otherwise, the adaption to D-ResNet-18 gets overshadowed by the dataset change. For training the D-ResNet-18, we inflate the weights of the 2D ResNet-18 and finetune using 5 frames per sample and a frame at the input with a lower learning rate of 0.01 and additional motion blur data augmentation for 1 epoch. Apart from these changes, parameters are identical to the pretraining. Our experiments showed that motion blur data augmentation did not improve the accuracy of the 2D ResNet-18, whereas it improves accuracy while finetuning the D-ResNet-18.

We evaluated our approach on the YouTube Faces dataset [43] following the standard verification protocol. We computed the Euclidean distance after l2-normalization and taking the average of the features. The preprocessing is done similarly to the VoxCeleb2 dataset. However, we resample the videos to obtain a fixed number of frames to show the dependency of the accuracy on the number of frames as shown in Table 7. In contrast to the 2D ResNet-18, our D-ResNet-18 continues to improve with increasing number

Model	Fusion	Accuracy (%)			
		5	25	50	100
2D ResNet-18	avg all	93.18	93.66	93.64	93.66
D-ResNet-18	avg all	92.40	93.90	93.74	93.74
D-ResNet-18	avg 5:end	93.12	93.80	93.92	94.10

Table 7: Evaluation on YouTube Faces dataset resampled to different number of frames.

of frames. We also evaluated discarding the first four features in the average due to cache initialization (denoted by avg 5:end), which resulted in another minor improvement. Note that for 5 frames *avg 5:end* is equal to taking only the last feature, which is substantially higher than the accuracy of the 2D-ResNet-18 for a single frame per video (88.78%). This demonstrate that our network is capable of propagating useful information through time.

5. Conclusion

In this work, we have addressed the computational complexity drawback of 3D CNNs and proposed a novel Dissected 3D CNN (D3D) architecture. The D3D architecture caches the intermediate volumes of the network and propagates them for future calculation, which reduces the computation around 74-90% during online operation for D-ResNet family. Besides reducing complexity during online operation, D-ResNet family achieves $\sim 5\%$ higher classification accuracy compared to classical ResNet family on Kinetics-600 dataset. We believe that this performance improvement arises since D3D networks are temporal resolution preserving and produce frame level fine-grained features. In this work, only ResNet family is converted to its dissected version and evaluated. However, any CNN architecture can be converted to its dissected version for efficient online video processing. The proposed D3D architecture successfully models temporal information and can be employed at any video based computer vision task. In our experiments, we have successfully validated the effectiveness of D3D architecture on four different vision tasks: activity/action recognition, gesture recognition, video person re-identification and video face recognition. For all these tasks, D3D consistently improved the performance. We believe that the D3D architecture will be actively used in many other video based tasks by the vision community.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation GPUs used in this study.

References

- [1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. *arXiv preprint arXiv:2003.08429*, 2020.
- [2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 941–951, 2019.
- [3] Joao Carreira and Andrew 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.
- [4] Hakan Cevikalp, Hasan Saribas, Burak Benligiray, and Sinem Kahvecioglu. Visual object tracking by using ranking loss. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.
- [7] Patrick Dendorfer, Hamid Rezaatofghi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Cvpr19 tracking and detection challenge: How crowded can it get? *arXiv preprint arXiv:1906.04567*, 2019.
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.
- [10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.
- [11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.
- [18] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.
- [19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [20] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.
- [21] Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [22] Okan Kopuklu, Yao Rong, and Gerhard Rigoll. Talking with your hands: Scaling hand gestures and recognition with cnns. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [23] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.
- [24] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebhay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [26] H. Kuhne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [27] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsuper-vised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [28] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.
- [29] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [30] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.
- [31] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland 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.
- [32] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.
- [33] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.
- [34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [35] Gurkirt Singh and Fabio Cuzzolin. Recurrent convolutions for causal 3d cnns. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [36] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017.
- [37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.
- [38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [40] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar 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.
- [41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.
- [42] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1328–1338, 2019.
- [43] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.
- [44] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross 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.
- [45] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [47] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 782–797, 2018.
- [48] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018.
- [49] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017.
- [50] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006.
- [51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
- [52] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*. Springer, 2016.
- [53] Yujie Zhong, Relja Arandjelović, and Andrew Zisserman. Ghostvlad for set-based face recognition. In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.