# Learning Motion in Feature Space: Locally-Consistent Deformable Convolution Networks for Fine-Grained Action Detection

Khoi-Nguyen C. Mac [1] Dhiraj Joshi [2] Raymond A. Yeh [1] Jinjun Xiong [2] Rogerio S. Feris [2] Minh N. Do [1]

[1] University of Illinois at Urbana-Champaign [2] IBM Research AI

ICCV 2019 Seoul, Korea

## Motivation

***Task:*** Predict segments of fine-grained action from a given video

- Fine-grained actions are difficult to visually distinguish based on individual frames
- Two-stream approaches are computationally expensive in general (optical flow + multi-stream inference)

***Contributions:***

- Modeling motion in feature space using changes in adaptive receptive fields over time
- Introducing local coherency constraint to enforce consistency in motion
- Constructing a backbone single-stream network to jointly learn spatio-temporal features

## Problem Formulation

***Background:***

- Standard convolution: $\mathbf{y}[n] = \sum_k \mathbf{w}[-k]\mathbf{x}[n+k]$
- Deformable convolution: for $\ddot{\Delta} \in \mathbb{R}^{N\times K}$,
  $\mathbf{y}[n] = \sum_k \mathbf{w}[-k]\mathbf{x}\left(n+k+\ddot{\Delta}_{n,k}\right), \quad \left(\ddot{\Delta}_{n,k} = (\mathbf{h}_k * \mathbf{x})[n]\right)$
- Adaptive receptive field at time $t$:
  $$\ddot{\mathbf{F}}^{(t)} \in \mathbb{R}^{N\times K} \quad \text{where} \quad \ddot{\mathbf{F}}^{(t)}_{n,k} = n+k+\ddot{\Delta}^{(t)}_{n,k}$$

***Temporal modeling:*** $\ddot{\mathbf{r}}^{(t)} = \ddot{\mathbf{F}}^{(t)} - \ddot{\mathbf{F}}^{(t-1)} = \ddot{\Delta}^{(t)} - \ddot{\Delta}^{(t-1)}$,
$\ddot{\mathbf{r}}^{(t)} \neq 0$ only for deformable convolutions

No guarantee of local consistency in receptive fields:

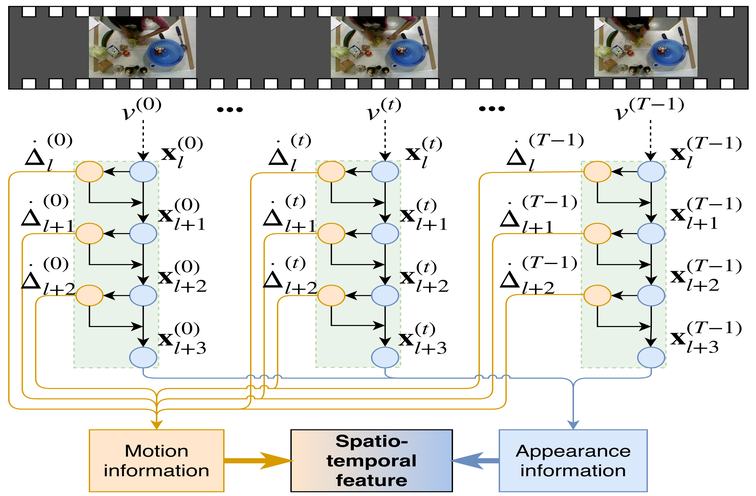
- $\ddot{\Delta}_{n,k}$ corresponds to $\mathbf{x}[n+k] = \mathbf{x}[m]$, where $m = n+k$.
- Multiple ways to decompose: $m = (n-l)+(k+l)$, for any $l$

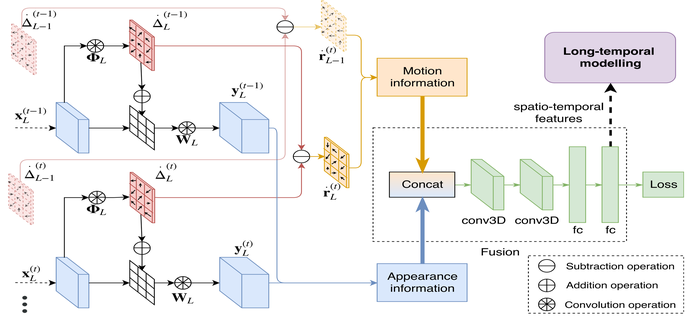
***Locally-consistent deformable convolution (LCDC)***:

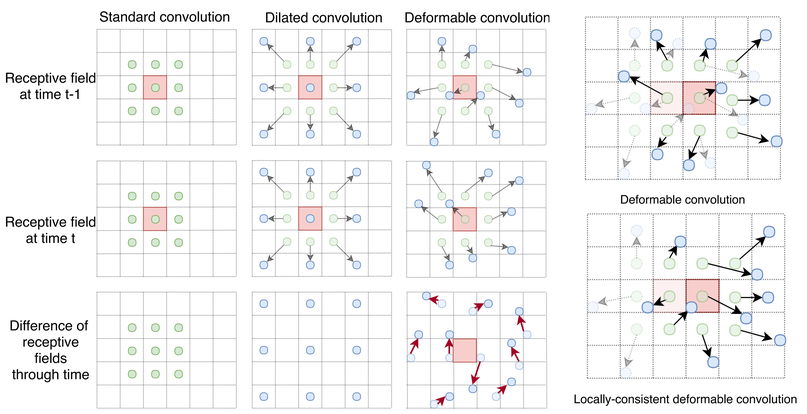
$$\mathbf{y}[n] = \sum_k \mathbf{w}[-k]\mathbf{x}\left(n+k+\dot{\Delta}_{n+k}\right), \quad \text{where} \quad \dot{\Delta} \in \mathbb{R}^N$$

- Local coherency constraint: LCDC is a special case of deformable convolution where $\ddot{\Delta}_{n,k} = \dot{\Delta}_{n+k}, \quad \forall n,k.$
- *Appearance information:* from the last layer.
- *Motion information:* from $\dot{\Delta}$ of multiple layers in feature space

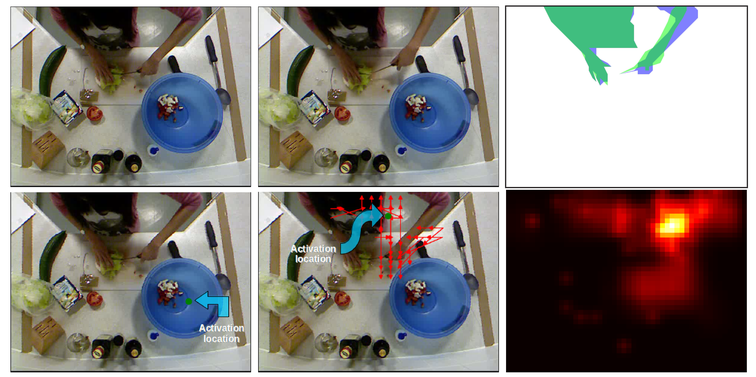
## Our Approach




## Results




***50 Salads dataset:*** 50 salad making videos, each 5-10 minutes. We report 2 granularity levels: *mid* (17 classes) and *eval* (9 classes)

| | Model | Spatial comp | Temporal comp (short) | Long-temporal | F1@10 | Edit | Acc |
|---|---|---|---|---|---|---|---|
| Mid | SpatialCNN [16] | RGB | MHI | - | 32.3 | 24.8 | 54.9 |
| | (SpatialCNN) + ST-CNN [16] | RGB | MHI | 1D-Conv | 55.9 | 45.9 | 59.4 |
| | (SpatialCNN) + DilatedTCN [15] | RGB | MHI | DilatedTCN | 52.2 | 43.1 | 59.3 |
| | (SpatialCNN) + ED-TCN [15] | RGB | MHI | ED-TCN | 68.0 | 59.8 | 64.7 |
| | (SpatialCNN) + TDRN [17] | RGB | MHI | TDRN | (72.9) | (66.0) | (68.1) |
| | LCDC | RGB | Learned deformation | - | 43.99 | 33.38 | 67.27 |
| | LCDC + ST-CNN | RGB | Learned deformation | 1D-Conv | 60.01±0.42 | 51.35±0.12 | 68.45±0.15 |
| | LCDC + DilatedTCN | RGB | Learned deformation | DilatedTCN | 58.21±0.59 | 48.54±0.52 | 69.28±0.25 |
| | LCDC + ED-TCN | RGB | Learned deformation | ED-TCN | **73.75±0.54** | **66.94±1.33** | **72.12±0.41** |
| Eval | Spatial CNN [16] | RGB | MHI | - | 35.0 | 25.5 | 68.0 |
| | (SpatialCNN) + ST-CNN [16] | RGB | MHI | 1D-Conv | 61.7 | 52.8 | 71.3 |
| | (SpatialCNN) + DilatedTCN [15] | RGB | MHI | DilatedTCN | 55.8 | 46.9 | 71.1 |
| | (SpatialCNN) + ED-TCN [15] | RGB | MHI | ED-TCN | 76.5 | 72.2 | 73.4 |
| | LCDC | RGB | Learned deformation | - | 56.56 | 45.77 | 77.59 |
| | LCDC + ST-CNN | RGB | Learned deformation | 1D-Conv | 70.46±0.41 | 62.71±0.46 | 77.84±0.26 |
| | LCDC + DilatedTCN | RGB | Learned deformation | DilatedTCN | 67.59±0.42 | 58.97±0.55 | 78.29±0.29 |
| | LCDC + ED-TCN | RGB | Learned deformation | ED-TCN | **80.22±0.21** | **74.56±0.70** | **78.90±0.25** |

***GTEA dataset:*** 28 kitchen activity videos, each around 1 minute. On average, a video has 19 segments from 7 classes

| Model | Spatial comp | Temporal comp (short) | Long-temporal | F1@10 | Edit | Acc |
|---|---|---|---|---|---|---|
| SpatialCNN [16] | RGB | MHI | - | 41.8 | - | 54.1 |
| (SpatialCNN) + ST-CNN [16] | RGB | MHI | 1D-Conv | 58.7 | - | 60.6 |
| (SpatialCNN) + DilatedTCN [15] | RGB | MHI | DilatedTCN | 58.8 | - | 58.3 |
| (SpatialCNN) + ED-TCN [15] | RGB | MHI | ED-TCN | 72.2 | - | 64.0 |
| (SpatialCNN) + TDRN [17] | RGB | MHI | TDRN | (79.2) | (74.1) | (70.1) |
| LCDC | RGB | Learned deformation | - | 52.42 | 45.38 | 55.32 |
| LCDC + ST-CNN | RGB | Learned deformation | 1D-Conv | 62.23±0.69 | 55.75±0.94 | 58.36±0.45 |
| LCDC + DilatedTCN | RGB | Learned deformation | DilatedTCN | 62.08±0.85 | 55.13±0.79 | 58.07±0.30 |
| LCDC + ED-TCN | RGB | Learned deformation | ED-TCN | **75.39±1.33** | **72.84±0.84** | **65.34±0.54** |

***Ablation Study:*** on split 1 of 50 Salads dataset

| Model | Spatial comp | Temporal comp (short) | Fusion scheme | Acc | Total params | Deform params |
|---|---|---|---|---|---|---|
| SpatialCNN | RGB (single) | MHI (multi) | Stacked inputs | 60.99 | - | - |
| NaiveAppear | RGB (single) | - | - | 68.45 | 38.9M | - |
| NaiveTempAppear | RGB (multi) | Avg feat frames (multi) | - | 71.52 | 38.9M | - |
| OptFlowMotion | - | OptFlow (multi) | - | 25.67 | 134.1M | - |
| TwoStreamNet | RGB (multi) | OptFlow (multi) | Avg scores | 71.82 | 173.0M | - |
| DC | RGB (multi) | Learned deformation (w/o local coherency) (multi) | 3D-Conv | 72.25 | 45.7M | 995.5K |
| LCDC | RGB (multi) | Learned deformation (multi) | 3D-Conv | **73.77** | **42.7M** | **27.7K** |