# Coherent Temporal Synthesis for Incremental Action Segmentation

CVPR2024

Guodong Ding, Hans Golong and Angela Yao

National University of Singapore

# Incremental tasks are ...

## Extensively studied in image domain

- Image classification,
- Object detection,
- Semantic segmentation, etc.

# Incremental tasks are ...

## Extensively studied in image domain

- Image classification,
- Object detection,
- Semantic segmentation, etc.

## Underexplored in video domain

- Action recognition,
- ... (more to come)

## Fundamental Categories of IL algorithms

- Replay/Rehersal

  replay a few of the data samples previously seen tasks (exemplar, generative)
- Regularization [1]

  consolidates the past knowledge, controlling the network weights updates
- Architectural

  dynamically changes the model's architecture, isolating task-specific parameters

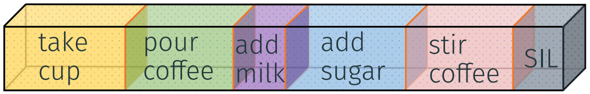
## Procedural Videos

Series of actions performed in some **constrained but non-unique order** to achieve some intended high-level goal.

## Procedural Videos

Series of actions performed in some **constrained but non-unique order** to achieve some intended high-level goal.

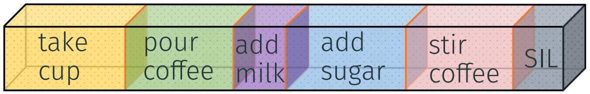
Make coffee



## Procedural Videos

Series of actions performed in some **constrained but non-unique order** to achieve some intended high-level goal.

Make coffee



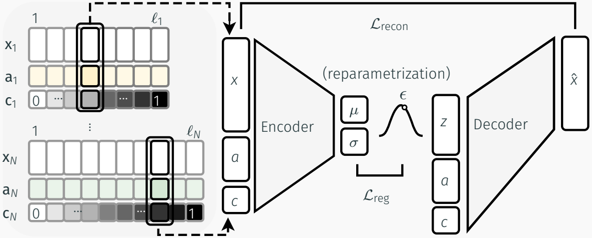

## Video Replay

- (Symbolic) Action sequence
  - take cup - pour coffee - add milk – add sugar – stir coffee – SIL
- Action duration
  - 180 - 150 - 90 - 140 - 160 - 100
- (Segmental) Action features

## Action Modeling via Conditional VAE

The Encoder takes as input

- $x$ - frame feature
- $a$ - action label
- $c$ - coherence variable
  - relative temporal progression of a frame within the action [0-1]



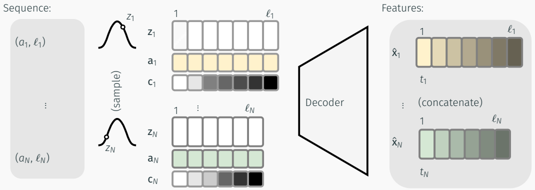

The Decoder

- samples a latent variable
- outputs the reconstruction of the original feature

## Action Synthesis with Decoder

Frames in the same segment have

- consistent action label
- identical sampled latent variable
- varying coherence variable
  - in accordance to their temporal location




Generated segments are concatenated in time to form the replay video.

# Incremental Training

Whenever new task data comes

## Action Segmentation

- construct replay data with generators from previous tasks
- learn segmentation with both incoming data and replay data

# Incremental Training

## Whenever new task data comes

### Action Segmentation

- construct replay data with generators from previous tasks
- learn segmentation with both incoming data and replay data

### Video Replay

- train new generator with incoming data
- cache generator in task stask

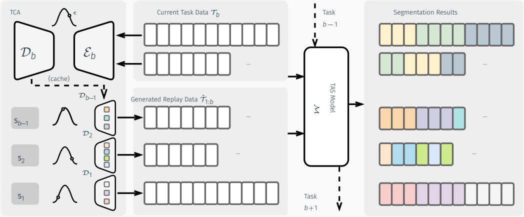
## Whenever new task data comes

### Action Segmentation

- construct replay data with generators from previous tasks
- learn segmentation with both incoming data and replay data

### Video Replay

- train new generator with incoming data
- cache generator in task stask




Iterate between Action Segmentation and Video Replay.

# Effectiveness on two benchmarks with two backbones

## Improvement

- significant improvements over standard finetune approach without data replay
- improvements compared to exemplar-saving counterpart

## Gaps

- large performance gap compared to using original frame features

| # Tasks | | MSTCN | | | | | ASFormer | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Edit | F1 @ {10, 25, 50} | | | Acc | Edit | F1 @ {10, 25, 50} | | |
| Breakfast | | | | | | | | | | | |
| 10 | Finetune | 7.4 | 7.2 | 7.5 | 7.0 | 5.4 | 9.9 | 9.8 | 10.3 | 9.4 | 7.5 |
| | Exemplar | 16.1 | 13.3 | 13.8 | 12.5 | 9.5 | 12.4 | 11.2 | 11.7 | 10.7 | 8.5 |
| | Ours | **29.4** | **25.9** | **26.3** | **23.5** | **17.7** | **34.2** | **32.4** | **33.1** | **30.1** | **23.4** |
| | Original | 43.1 | 41.1 | 41.2 | 37.6 | 29.5 | 48.1 | 45.2 | 45.9 | 42.4 | 34.2 |
| 5 | Finetune | 15.4 | 15.8 | 16.6 | 15.8 | 12.7 | 15.7 | 16.1 | 16.9 | 15.8 | 13.2 |
| | Exemplar | 32.5 | 28.9 | 30.8 | 28.5 | 22.9 | 29.5 | 27.5 | 28.7 | 26.7 | 22.0 |
| | Ours | **54.5** | **49.4** | **51.1** | **46.9** | **37.7** | **57.2** | **56.8** | **58.3** | **54.0** | **43.6** |
| | Original | 60.4 | 59.1 | 60.3 | 56.1 | 46.0 | 65.1 | 64.2 | 65.6 | 61.5 | 51.0 |
| YouTube Instructional | | | | | | | | | | | |
| 5 | Finetune | 13.6 | 2.8 | 3.6 | 2.7 | 0.6 | 13.9 | 11.5 | 11.1 | 9.8 | 6.3 |
| | Exemplar | **30.8** | 19.7 | 19.8 | 16.0 | 9.3 | 22.1 | 18.9 | 17.7 | 15.3 | 10.0 |
| | Ours | 30.2 | **25.0** | **21.9** | **18.5** | **11.1** | **25.2** | **20.9** | **20.1** | **17.5** | **11.4** |
| | Original | 55.9 | 39.4 | 38.1 | 32.2 | 19.1 | 59.2 | 51.1 | 45.4 | 39.1 | 25.5 |

# Ablations

## Temporal Coherence

| | SD | FD | TC | Acc | Edit | F1 @ {10, 25, 50} | | |
|---|---|---|---|---|---|---|---|---|
| Exemplar | ✓ | ✗ | ✗ | 27.8 | 35.6 | 36.1 | 31.7 | 24.3 |
| $\text{Ours}_{\text{random}}$ | ✓ | ✓ | ✗ | 32.9 | 38.9 | 40.0 | 35.6 | 27.2 |
| $\text{Ours}_{\text{static}}$ | ✓ | ✗ | ✗ | 37.9 | 42.9 | 43.8 | 38.9 | 29.0 |
| Ours | ✓ | ✓ | ✓ | **41.8** | **45.0** | **47.0** | **41.5** | **32.0** |

SD - segment-level diversity FD - frame-level diversity TC - temporal coherence

- Without temporal coherence, static segmnet works better than random
- All factors considered together achieves the best performance

## Replay Size

| $M$ | Acc | Edit | F1 @ {10, 25, 50} | | |
|---|---|---|---|---|---|
| 30 | 34.0 | 39.6 | 41.0 | 34.8 | 24.7 |
| 60 | 35.4 | 41.2 | 42.3 | 36.0 | 25.6 |
| 90 | 36.2 | **42.3** | 43.9 | **37.3** | **26.8** |
| 120 | **38.0** | **42.3** | **44.0** | 37.1 | 26.2 |

- A larger replay size leads to better performance
- saturates and no further gain with replay size

## TCA Training data

| | $\mathcal{T}(\%)$ | Acc | Edit | F1 @ {10, 25, 50} | | |
|---|---|---|---|---|---|---|
| Exemplar | - | 22.6 | 34.8 | 36.0 | 32.4 | 25.2 |
| Ours | 25 | 41.7 | 43.2 | 46.1 | 40.9 | 31.5 |
| | 50 | 42.1 | 43.3 | 45.1 | 40.5 | 31.5 |
| | 75 | 45.3 | 45.9 | 47.8 | **43.7** | **34.7** |
| | 100 | **47.4** | **46.9** | **48.2** | 42.8 | 33.4 |

- Access to more real data helps build the generative ability

# Conclusion

## Take aways

- Generative replay approaches are better desired for procedural videos
- Temporal coherence is essential for video replay
- This is an underexplored area full of research possibilities

*Thank you!*