DEAR: Depth-Estimated Action Recognition

Sadegh Rahmani, Filip Rybansky, Quoc Vuong, Frank Guerin, Andrew Gilbert

Scan to read

- **Contributions:**
  - ❑ Using estimated depth as a complementary modality instead of using depth camera/sensors
  - ❑ Designing a new network architecture to encode and fuse RGB frames and paired Depth maps
  - ❑ Improving action recognition using estimated depth maps on SSv2 compared to SOTA

Cross Attention (Middle Fusion)

$$CrossAttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$CrossAttention(Q_i, K_{ds}, V_{ds})$
$CrossAttention(Q_{ds}, K_i, V_i)$
$CrossAttention(Q_i, K_{dv}, V_{dv})$
$CrossAttention(Q_{dv}, K_i, V_i)$
i, ds, dv are image, depth_s4v, depth_videomamba

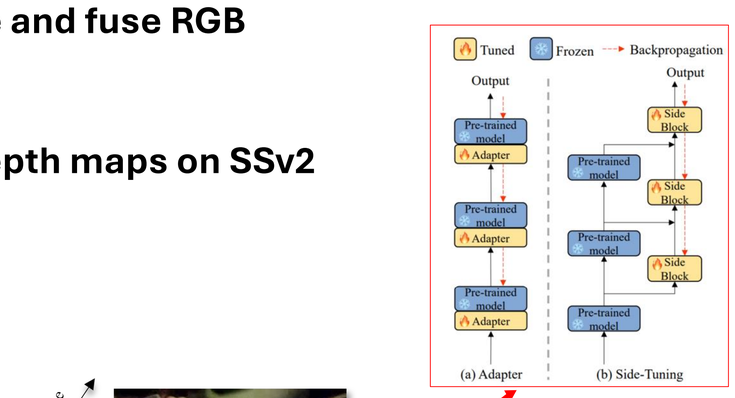
Branch Classification

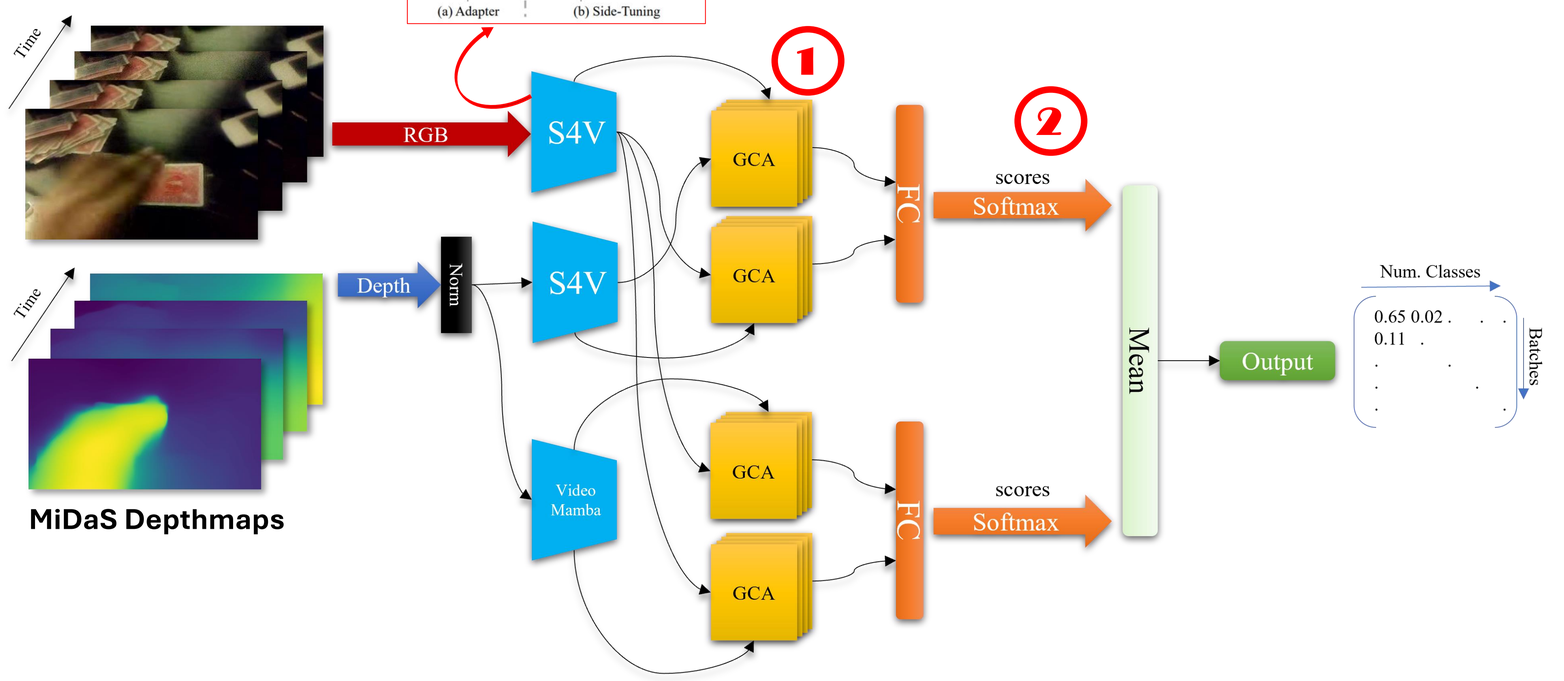
$$prob_{ids} = \sigma(FC(Concat(CA_{ids}, CA_{dsi})))$$
$$prob_{idv} = \sigma(FC(Concat(CA_{idv}, CA_{dvi})))$$

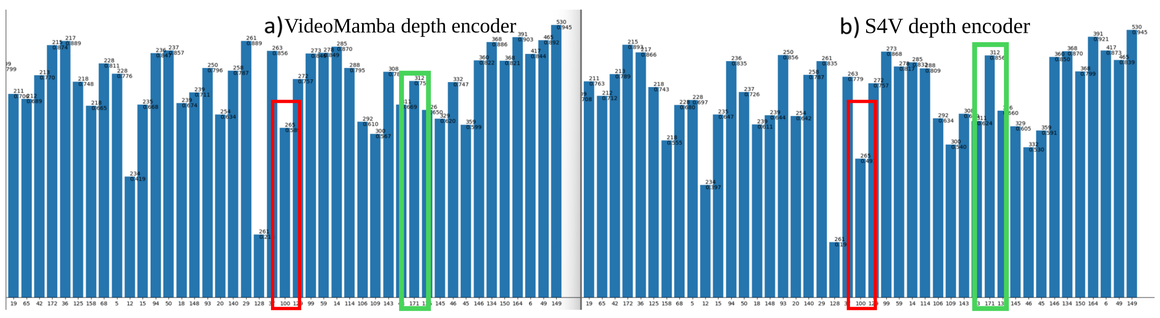
Softmax average (Late Fusion)

$$\hat{Y} = \frac{prob_{ids} + prob_{idv}}{2}$$

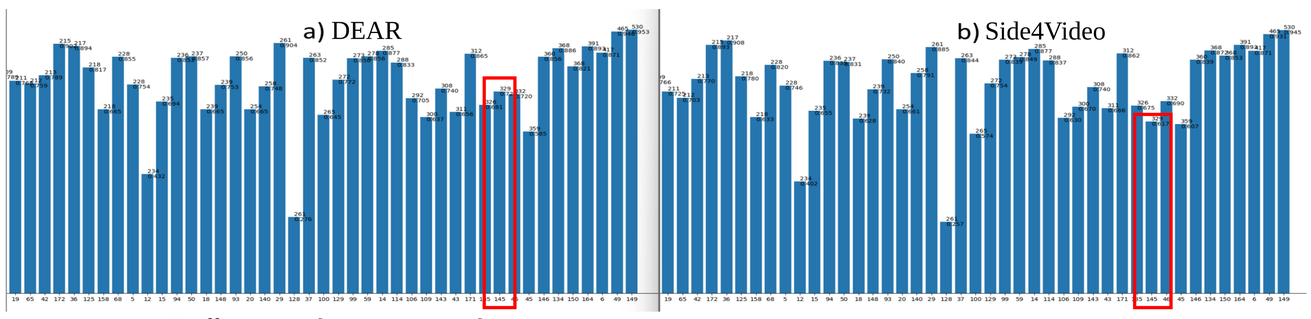
Method	Modality	Backbone	pre-train	Top-1
SlowFast [70]	RGB	ResNet101	K400	63.1
TSM [71]	RGB	ResNet50	K400	63.4
TimeSformer [72]	RGB	ViT-L	IN-21K	62.4
Mformer [73]	RGB	ViT-L	IN-21K+K400	68.1
ViViT FE [74]	RGB	ViT-L	IN-21k+K400	65.9
VIMPAC [75]	RGB	ViT-L	HowTo100M	68.1
VideoMamba-S [25]	RGB	-	IN-1K	66.6
VideoMamba-Tiny [25]	RGB	-	IN-1K	65.1
VideoMamba-Tiny [25]	Depth map	-	IN-1K	52.0
S4V* [24]	RGB	ViT-B	Clip-400M	70.2
S4V* [24]	Depth map	ViT-B	Clip-400M	56.0
Proposed	RGB+Depth(VideoMamba)	ViT-B	Clip-400M	70.0
Proposed	RGB+Depth	ViT-B	Clip-400M	70.3
Proposed	RGB+Depth+ Depth(VideoMamba)	ViT-B	Clip-400M	**71.0**







Pushing something so that it slightly moves *Uncovering something*

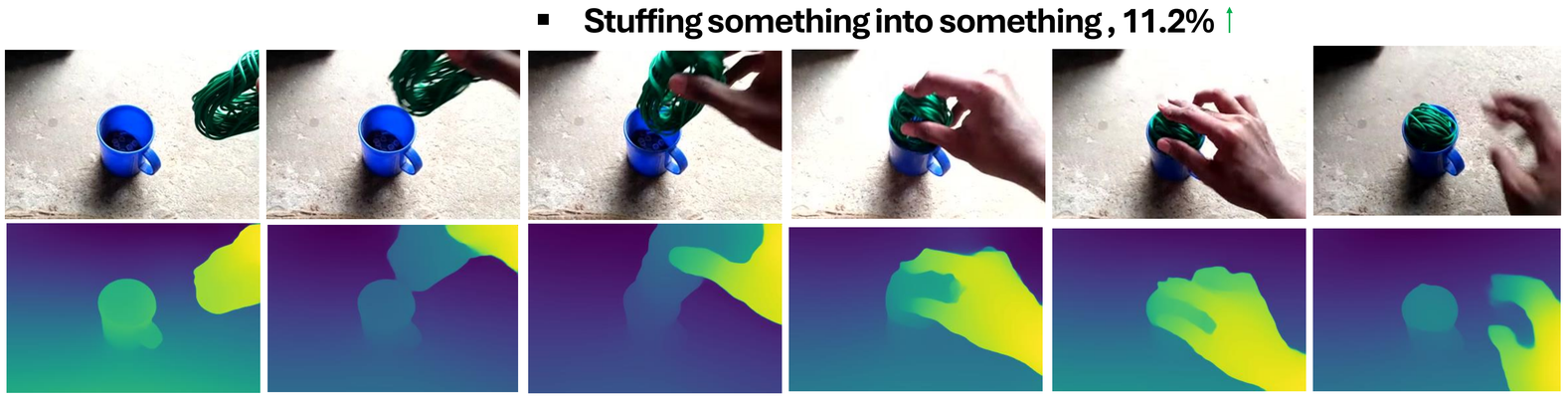


Stuffing something into something

- **Why two depth encoder branches:**
  - ❑ Different understanding of the input
  - ❑ Various distribution over classes accuracy
  - ❑ Combining models, leads to higher performance

- **Classes get affected the most by depthmaps:**
  - ❑ Stuffing something into something , 11.2% ↑
  - ❑ Putting something into something , 7.5% ↑
  - ❑ Pushing something so that it slightly moves , 7.1% ↑
  - ❑ Tearing something just a little bit , 3.2% ↓

▪ Stuffing something into something , 11.2% ↑



- **Visual reasons for positive effect of depth maps**
  - ❑ Similar colour temperature between mug and rope (and gradient colour on the mug) shows the equal distance of those objects from the camera which show the rope is not beside or on the mug.

▪ Tearing something just a little bit , 3.2 ↓

- **Visual reasons for negative effect of depth maps**
  - ❑ Due to the tiny size of the ripping area and the fact that hands filled a large portion of the receptive field with the greatest colour temperature, no specific feature is achievable on torn paper.

Funded By LEVERHULME TRUST