Final Project Review

Adaptive Multi-Modal Learning for Efficient Video Recognition

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING,
KANCHEEPURAM

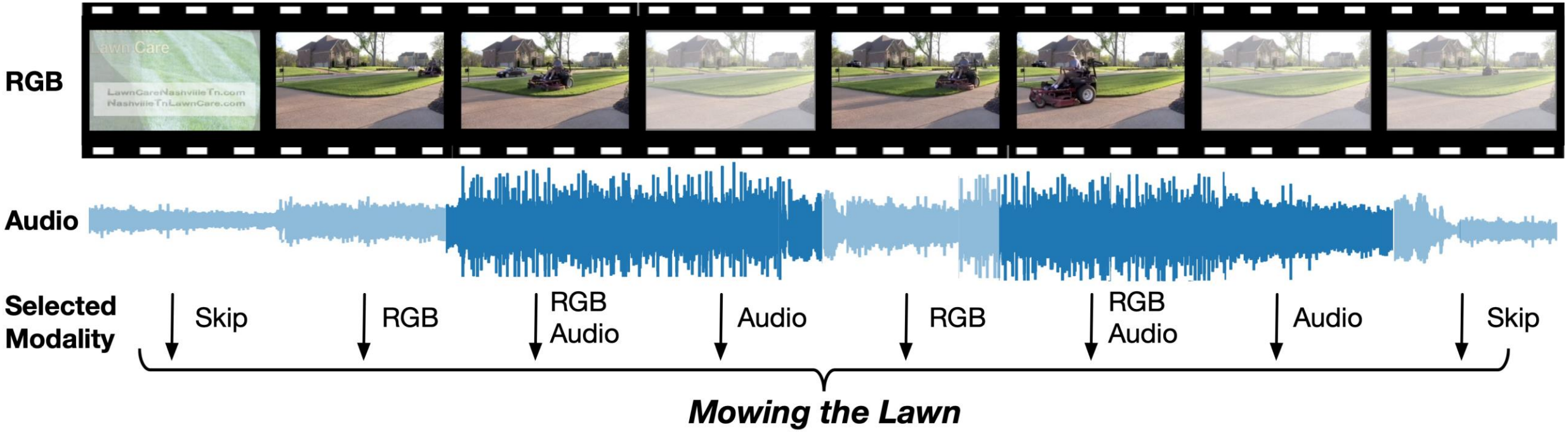
Krishna Kumar Sutar
(CED17I003)

Table of Contents

- Motivation
- Key Idea
- Framework
- Dataset and Architecture
- Results and Visualizations
- Future Scopes

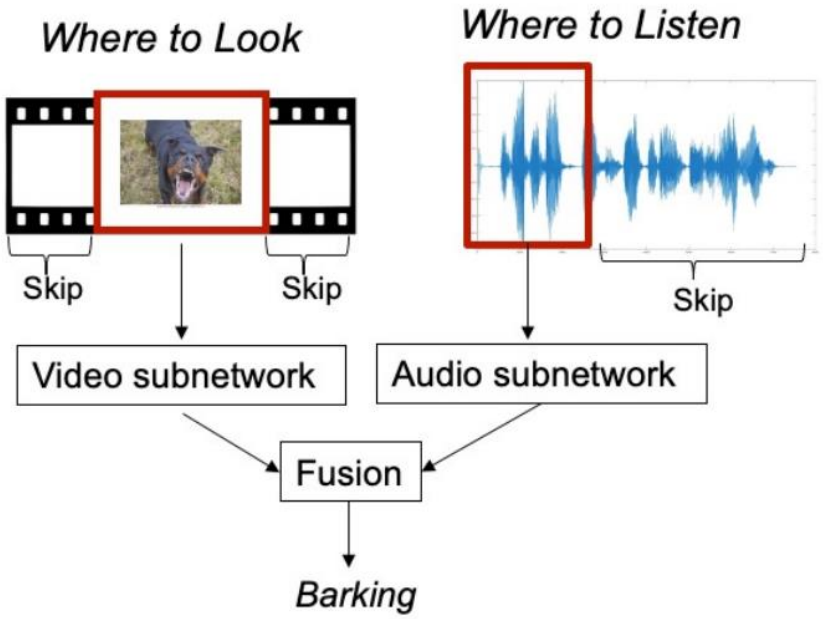
Motivation

- Most of the multi-modal video recognition methods are computationally expensive .
- Utilizing information from all input modalities may be counterproductive.
- Selection of cheap modality with good performance becomes essential.



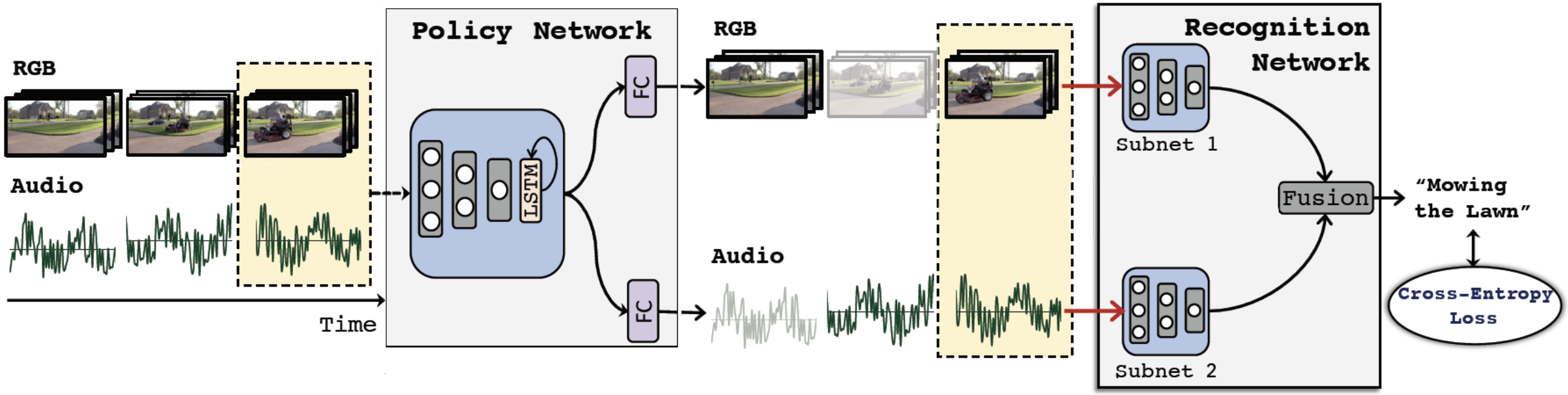
- The lawn mower is moving with relevant audio only in the 3rd and 6th segment.
- In contrast to the commonly used one-size-fits-all scheme for multi-modal learning, we would like these decisions to be made individually per input segment

Key Idea



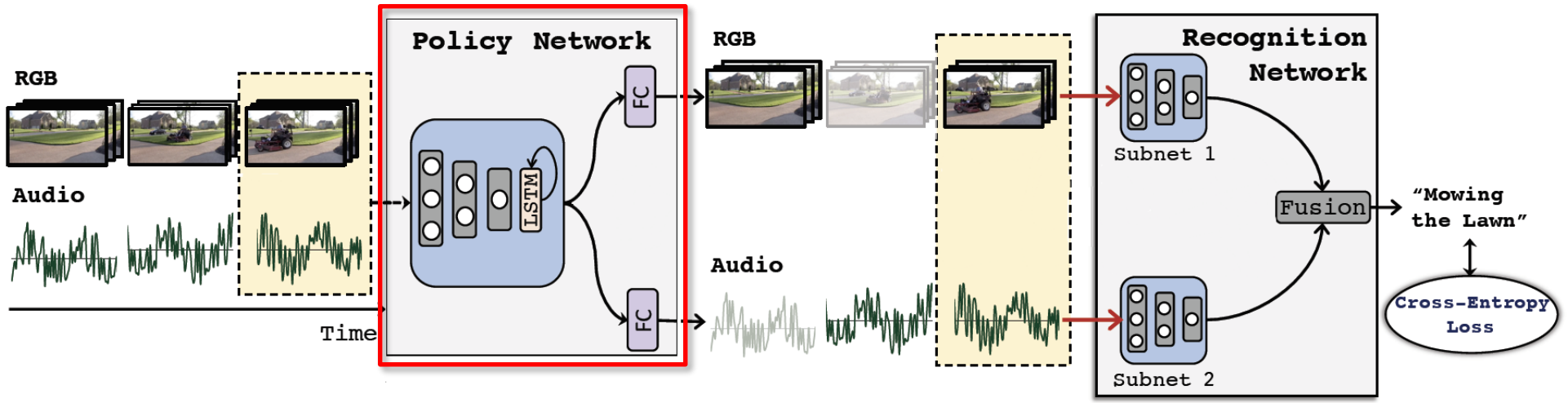
Adaptive Multi-Modal Learning : A novel approach to learn a decision policy that selects optimal modalities conditioned on the inputs for efficient video recognition.

Framework



Our Framework consists of a policy network and a Recognition network composed of different subnetworks that are trained jointly for recognizing the videos.

Multi-Modal Policy Network

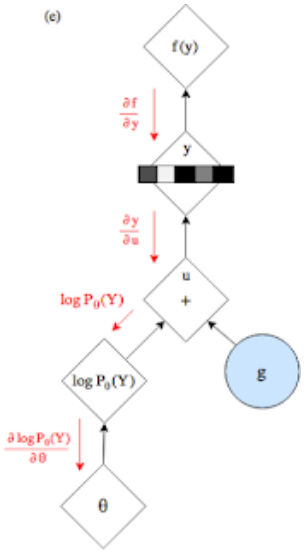


$$h_t, o_t = \text{LSTM}(f_t, h_{t-1}, o_{t-1})$$

- Given the hidden state, the policy network estimates a policy distribution for each modality and then sample binary decisions on whether to select the modality or not at that time step using Gumbel-Softmax operation.

$$\hat{P}_k = \arg\max_{i \in \{0,1\}} (\log z_{i,k} + G_{i,k}), \qquad k \in [1, ..., K]$$

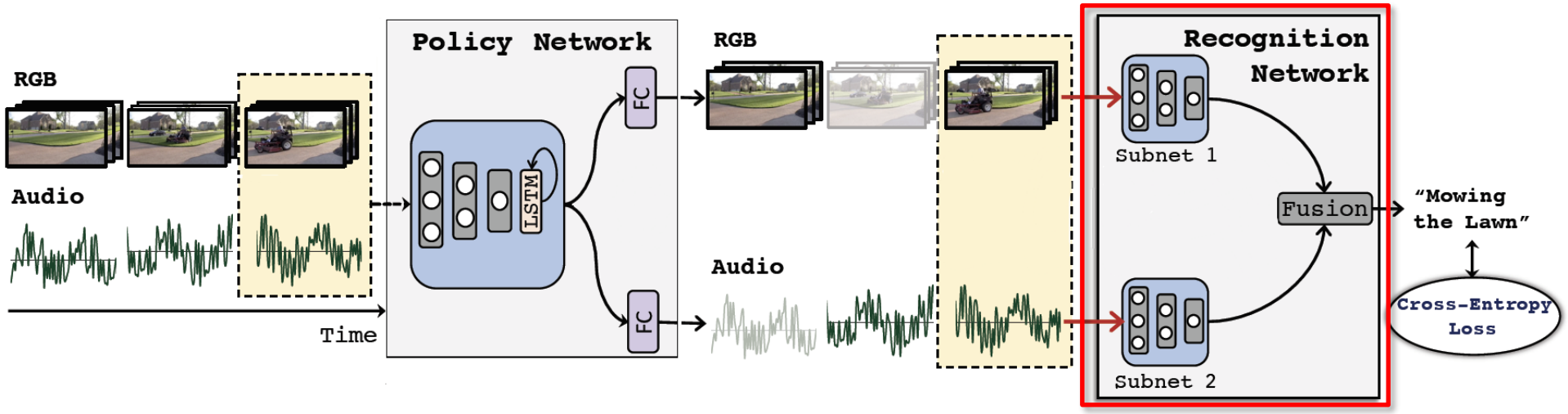
$$P_{i,k} = \frac{\exp((\log z_{i,k} + G_{i,k})/\tau)}{\sum_{j \in \{0,1\}} \exp((\log z_{j,k} + G_{j,k})/\tau)}$$



Training using Gumbel-Softmax Sampling

- An effective way to replace non differentiable sample from corresponding differentiable sample from Gumbel-Softmax distribution

Recognition Network



The selected modalities are passed to the recognition network to generate segment level predictions. Finally, the network averages all segment level predictions to obtain video level prediction.

$$\mathbb{E}_{(V,y)\sim\mathcal{D}_{train}}\left[-y\log(\mathcal{P}(V;\Theta))+\sum_{k=1}^{K}\lambda_k\mathcal{C}_k\right]$$

$$\mathcal{C}_k=\begin{cases}(\frac{|U_k|_0}{C})^2 & \text{if correct}\\ \gamma & \text{otherwise}\end{cases}$$

Loss Function

- First part is standard cross entropy loss used to measure classification quality.
- Second part drives the network to learn a policy that favors selection of modality that is computationally efficient for recognizing.

Datasets and Architecture

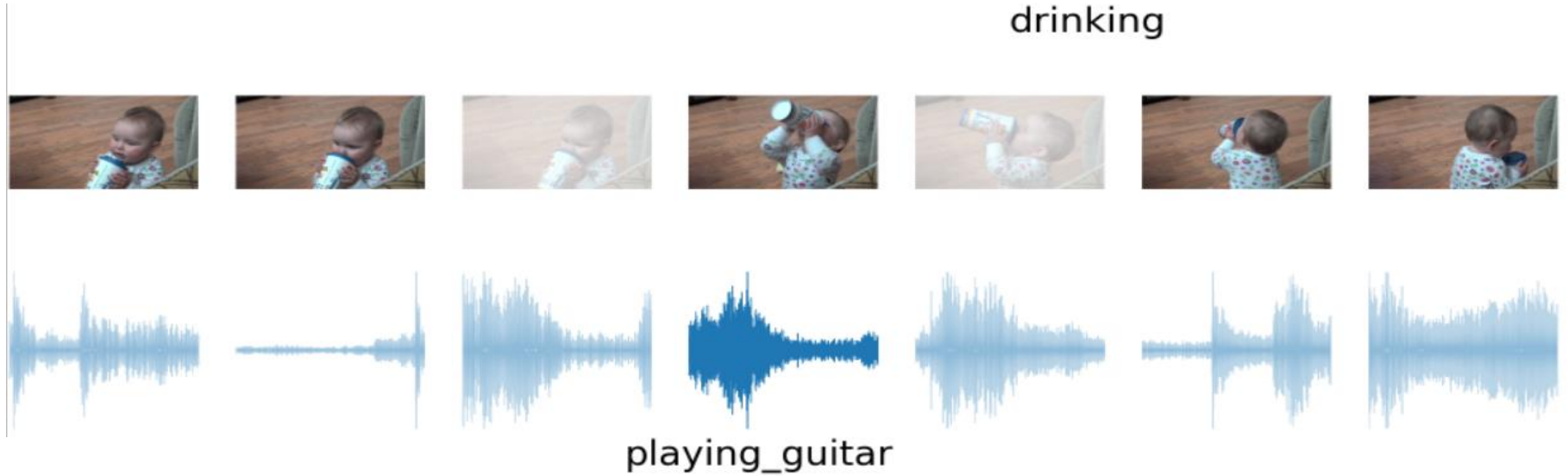
- Datasets :
  - Kinetics400 : Training: 2,521 videos ; Testing: 1,462 videos ; Classes: 400
- Tasks:
  - RGB + Audio
- Model Architectures:
  - Policy Network: MobileNetV2 (For feature extraction)
  - Recognition Network:
    - RGB frames Recognition - ResNet-50
    - Audio Recognition – MobileNetV2
- Evaluation Metrics – Accuracy, Selection rate, FLOPs

Results

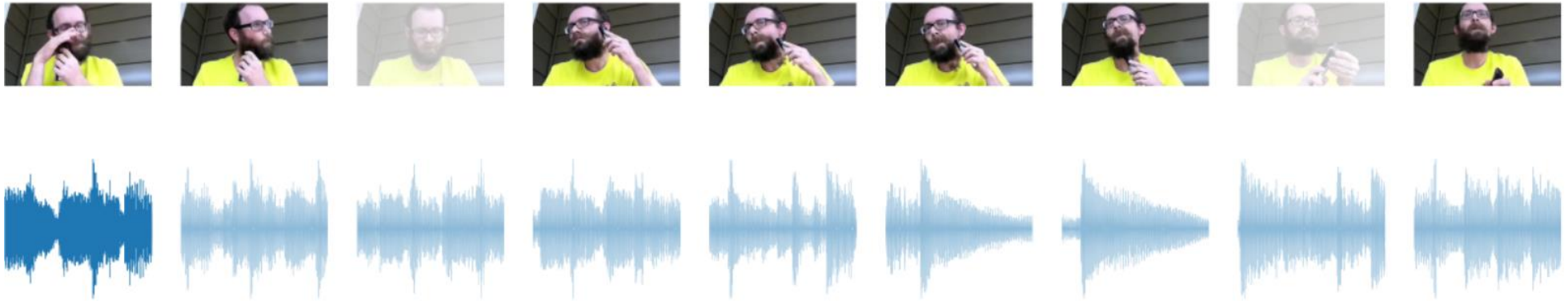
Dataset	Kinetics400			
Method	Accuracy (%)	Selection Rate (%)		GFLOPs
		RGB	Audio	
RGB	74.3	100	-	201.45
AUDIO	49.2	-	100	4.38
Weighted Fusion	76.18	100	100	210.17
Our Model	76.47	89.05	10.32	134.47 (-36.01%)

Method	Kinetics400	
	Accuracy (%)	GFLOPs
LiteEval	70.02	165.06
Our Model	76.47	134.47 (-18.53%)

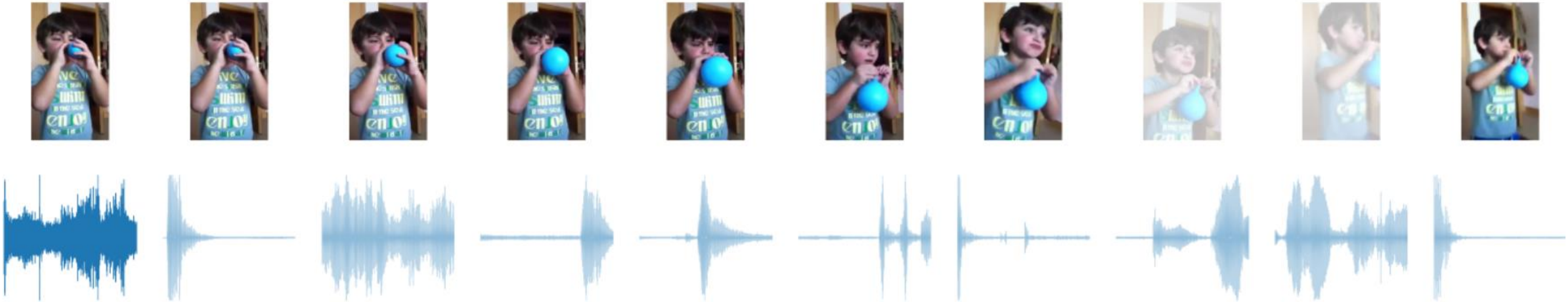
Visualizations



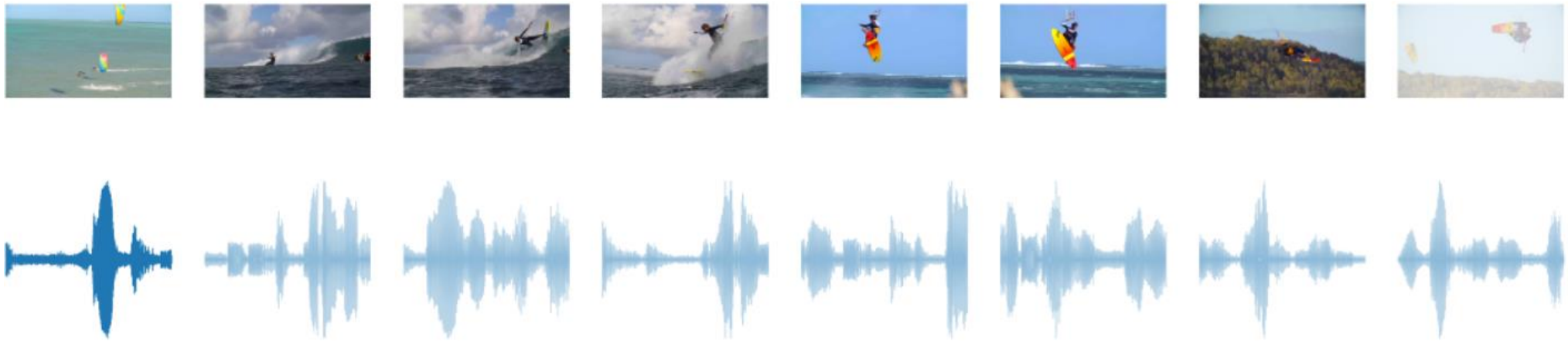
trimming_or_shaving_beard



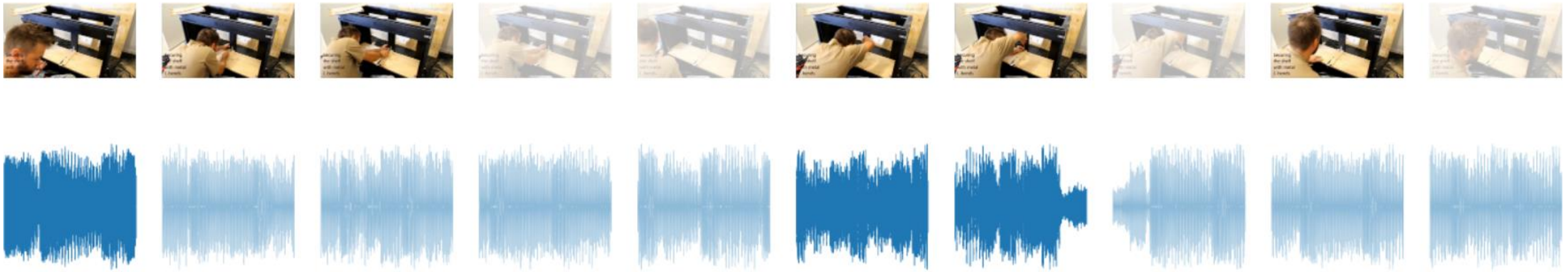
balloon_blowing



kitesurfing



building_cabinet



Future Scopes

- Understand modality clues to decide whether other modalities are necessary or not.
  - E.g. audio modality clues are the cheapest of all the modality, so they should be leveraged such that after utilizing information from audio modality clues, decision must be taken on whether RGB modality is necessary or not.

Thank You