# Skeleton Aware Multi-modal Sign Language Recognition

## SAM-SLR

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li and Yun Fu
Department of Electrical and Computer Engineering
Northeastern University, Boston MA, USA

Speaker: Songyao Jiang

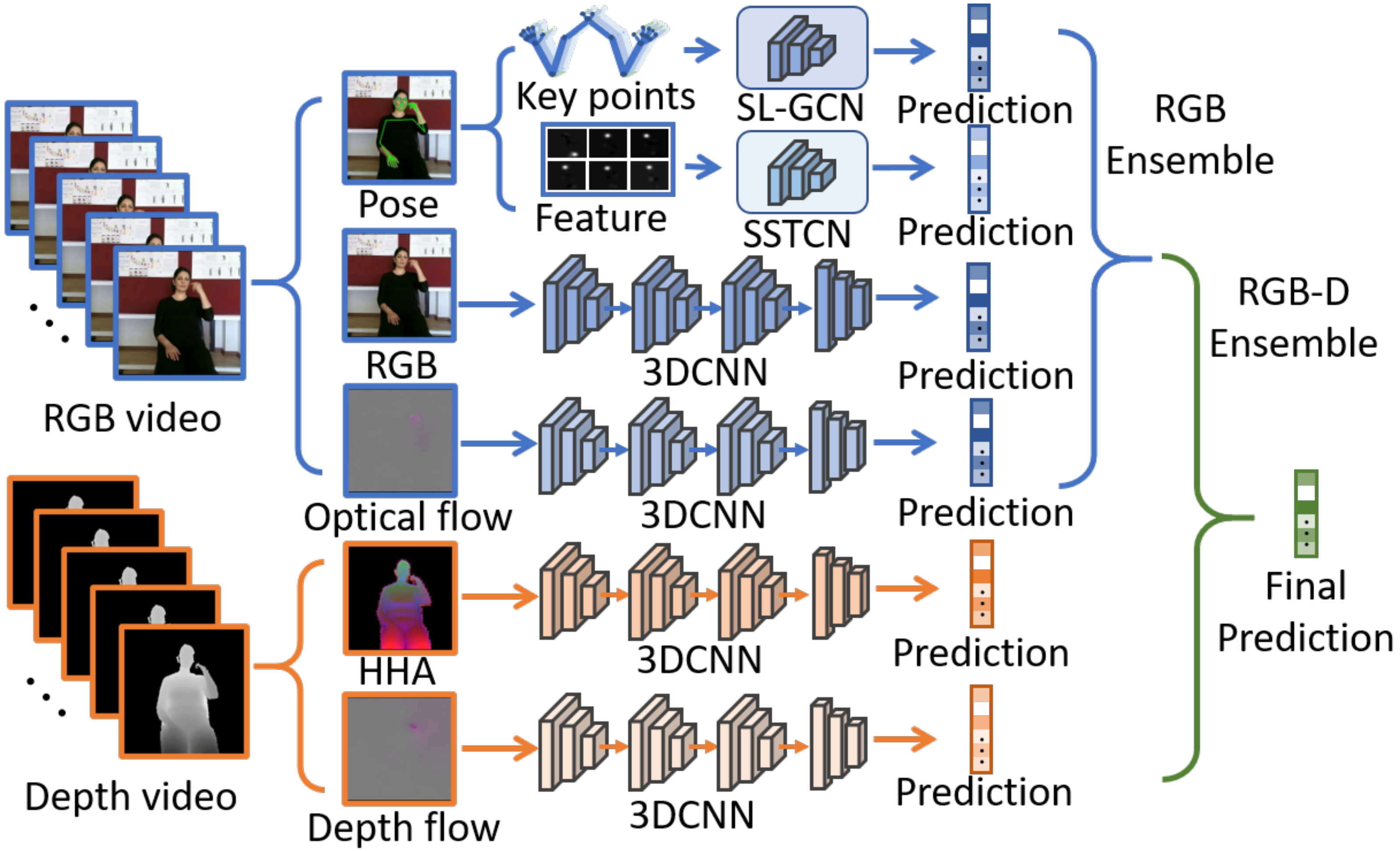
# Agenda

- Introduction and motivation
- Pipelines of SAM-SLR
  - Preprocessing of modalities
  - SL-GCN
  - 3DCNN
  - SSTCN
- Experimental results
- Conclusion

# Introduction and Motivation

- Sign Language Recognition (SLR) is a more challenging problem:
  - Sign language requires both global body motion and delicate arm/hand gestures to distinctly and accurately express its meaning.
  - Similar gestures can even impose various meanings depending on the number of repetitions.
  - Different signers may perform sign language differently (e.g., speed, localism, left-handers, right-handers, body shape)
- Preliminary findings and our assumptions:
  - Skeleton based methods become popular in action recognition.
  - Skeleton based methods act as strong complements to RGB / RGB-D based methods.
  - Different modalities contain different valuable information. Their ensembles always improve the overall performance. (e.g. RGB + Optical flow)
- Problem: no ground-truth keypoints provided.
- Basic ideas:
  - Use whole-body pose estimator to provide whole-body skeleton keypoints.
  - Use as many modalities as we can to improve the overall accuracy.

# Pipelines of SAM-SLR Framework




Three types of models
- SL-GCN
- SSTCN
- 3DCNN

RGB Track:
- Skeleton
- Pose Feature
- RGB
- Optical Flow

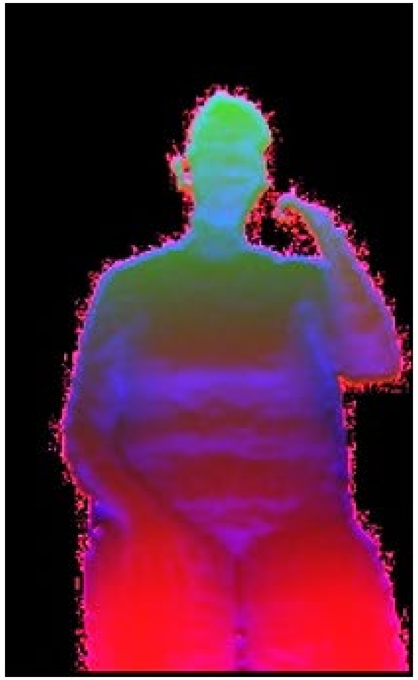
RGB-D Track:
- Skeleton
- Pose Feature
- RGB
- Optical Flow
- HHA
- Depth Flow

Code available on GitHub: https://github.com/jackyjsy/CVPR21Chal-SLR

# Visualization of Modalities Used

Whole-body Pose

Depth



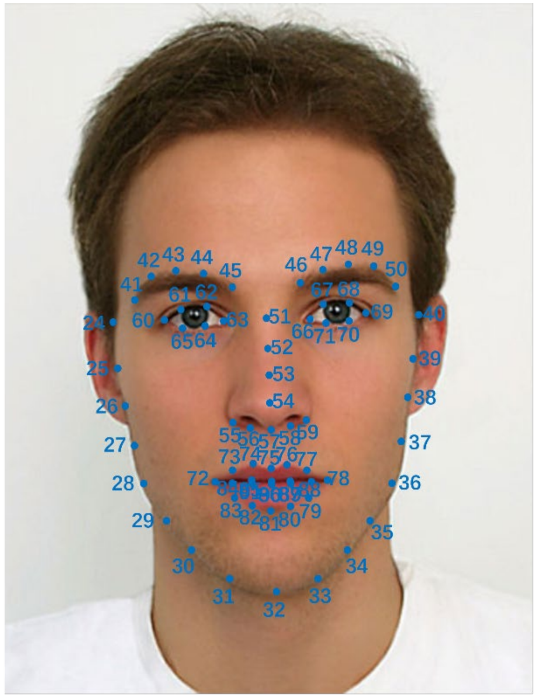
HHA

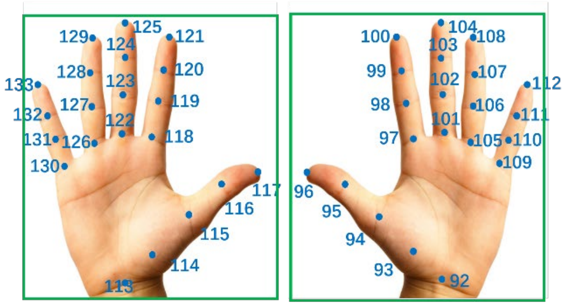
RGB Optical Flow

Depth Optical Flow

# Whole-body Pose Estimation

- Traditional 2D human pose estimation:
  - 16 points or 17 points only
  - Does not include hand keypoints
- Problems using separate hand pose model:
  - Hand pose estimator cannot work without detector.
  - Hand detector fails due to motion blur / low resolution.
- 133-point whole-body keypoints[1]:
  - Face: 68 points
  - Body: 17 points
  - Hands: 34 points
  - Feet: 6 points
- Advantages of whole-body keypoints estimator:
  - Consistent and faithful estimation of hand keypoints
  - Resistant to motion blurs

[1] Jin et al., Whole-body human pose estimation in the wild. In ECCV, 2020.

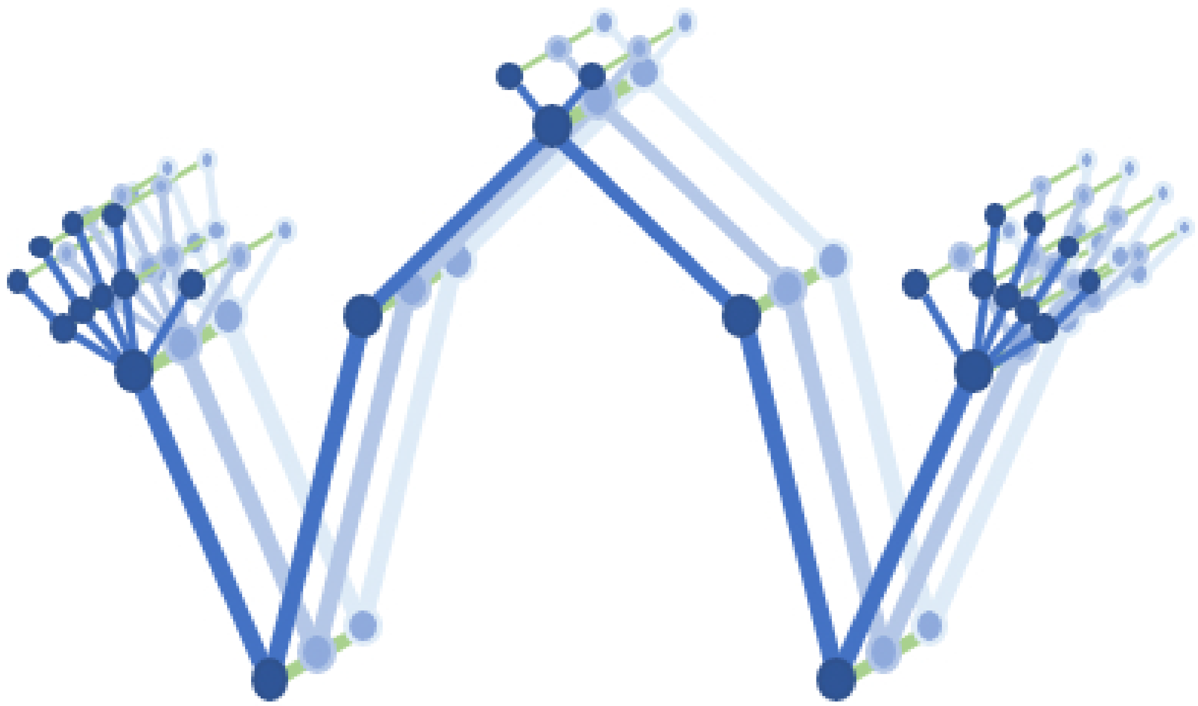
# Sign Language Skeleton Graph Construction

- Spatial and Temporal Graph

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \text{if } d(v_i, v_j) = 1 \\ 0 & \text{else} \end{cases}$$

where $d(v_i, v_j)$ calculate the minimum distance between skeleton node $v_i$ and $v_j$ .

- Graph Reduction
  - Motivation: too many nodes introduce extra noise into the spatio-temporal graph.
  - 133 nodes are trimmed to 27 nodes.
  - The remaining graph contains 10 nodes for each hand and 7 nodes for the upper body.

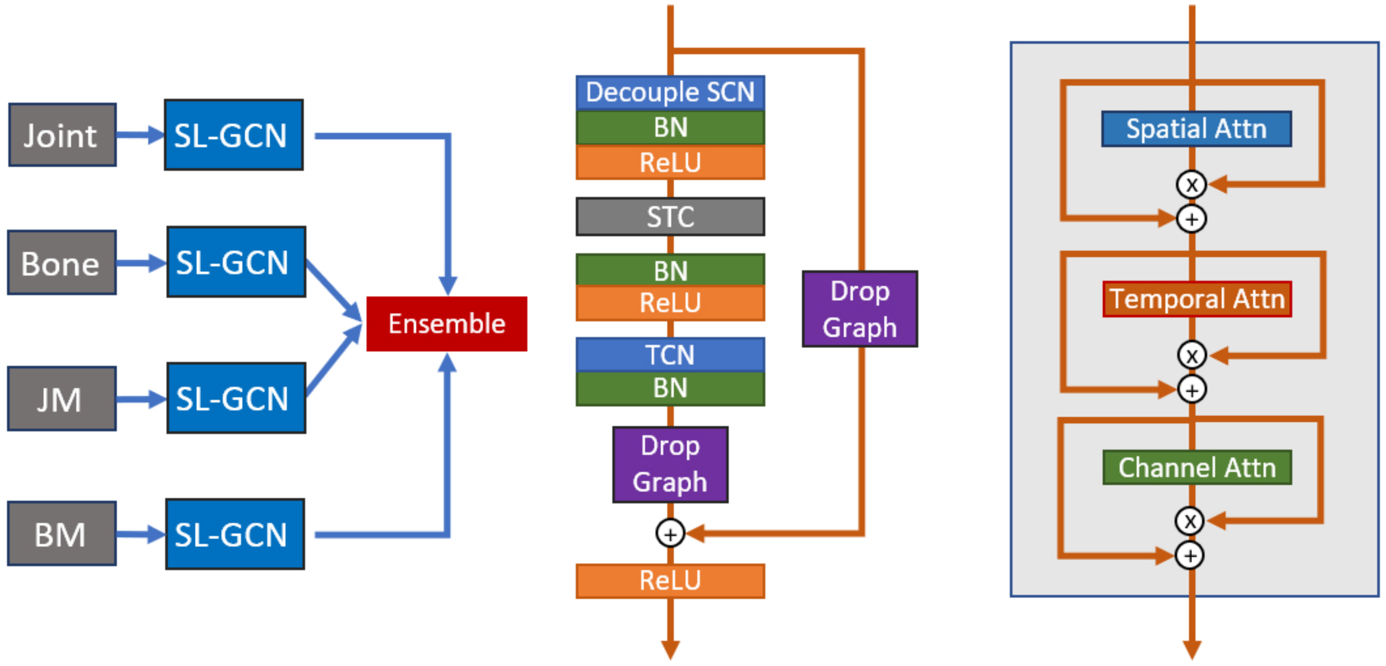


# Sign Language GCN Framework (SL-GCN)

Basic SL-GCN Block:
- ST-GCN Modules [2]
- Decouple SCN [3]
- Drop Graph Module [3]

STC Attention Module [4]:
- Spatial Attention
- Temporal Attention
- Channel Attention

Multi-stream Workflow:
- Joint
- Bone
- Joint Motion
- Bone Motion




(c) Multi-stream SL-GCN (d) SL-GCN Block (e) STC Module


[2] Yan et al., Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI, 2018.
[3] Cheng et al., Decoupling GCN with DropGraph module for skeleton-based action recognition. In ECCV, 2020.
[4] Shi et al., Skeleton-based action recognition with directed graph neural networks. In CVPR, 2019.

# Performance of SL-GCN

Recognition rate on AUTSL validation set is shown below:

Multi-stream performance on val set

| Streams | Top-1 | Top-5 |
|---|---|---|
| Joint | 95.02 | 99.21 |
| Bone | 94.70 | 99.14 |
| Joint Motion | 93.01 | 98.85 |
| Bone Motion | 92.49 | 98.78 |
| Multi-stream | **95.45** | **99.25** |

Ablation studies of SL-GCN

| Variations | Top-1 |
|---|---|
| SL-GCN (Joint) | **95.02** |
| w/o Graph Reduction | 63.69 |
| w/o Decouple GCN | 94.66 |
| w/o Drop Graph | 94.81 |
| w/o Keypoints Augmentation | 90.16 |
| w/o STC Attention | 93.53 |

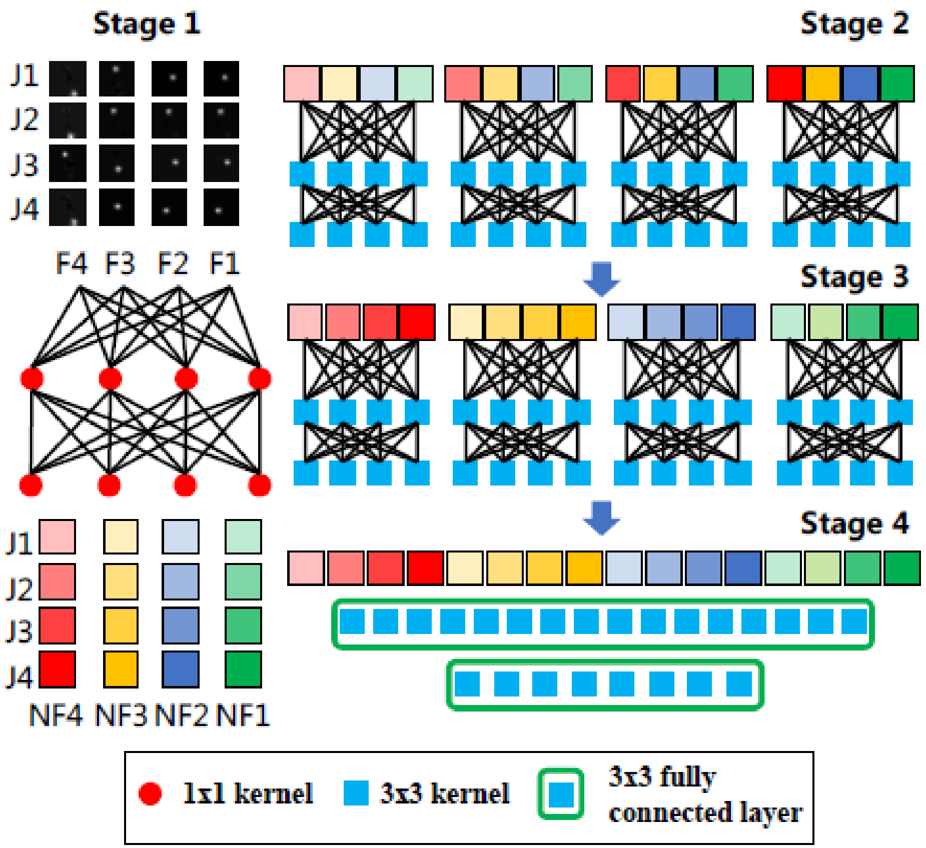
# Separable Spatial-Temporal Convolution Network (SSTCN)

Motivation:

- Factorization of 3D convolution will be easier to optimize compared to full 3D filters where appearance and dynamics are jointly intertwined.

Basic SSTCN block contains 4 stages:

- Stage 1: Extract features along temporal dimension
- Stage 2: Extract features along temporal dimension grouped by joints
- Stage 3: Extract features along joints dimension grouped by frames.
- Stage 4: Extract features through fully convolutional layers



# Pros and Cons of Skeleton-based SLR

Pros:

- Accuracy is high.
- No interference of background.
- Signer-invariant.
- Light-weight network, easy to train.

Cons:

- Finger keypoints estimation may not be accurate.

Solution:

- Those inaccurate keypoints may be corrected by other modalities in ensemble models.

Example: that figure was not captured

# 3D Convolutions Neural Networks (3DCNN)

- Popular 3D CNNs in video classification:
  - I3D, ResNet3D, SlowFast,
- Baseline: ResNet2+1D-18
- Swish Activation: $f(x) = x \cdot \mathrm{Sigmoid}(x).$
- Label smoothing: $q'(k|x) = (1-\epsilon)\delta_{k,y} + \epsilon u(k),$
- Corresponding Cross-entropy

$$H(q', p) = -\sum_{k=1}^{K} \log p(k) q'(k) = (1-\epsilon) H(q, p) + \epsilon H(u, p),$$

# Training 3DCNNs (RGB Modality)

- Pretraining 3DCNNs:
  - Import weights trained on Kinectic-300 action recognition datasets.
  - Pretrain on Chinese Sign Language (CSL) dataset

- Ablation studies:
  - w/o label smoothing
  - w/o swish activation
  - w/o pretraining on CSL
  - Use ResNet3D-18 backbone instead

| 3D CNN Variations | Top-1 |
|---|---|
| Ours (RGB Frame) | **94.77** |
| w/o Label Smoothing | 93.75 |
| w/o Swish Activation | 92.88 |
| w/o Pretraining on CSL | 93.41 |
| w/ ResNet3D-18 Backbone | 93.10 |

Table 5. Ablation studies on 3D CNN using RGB frames.

# Multi-modal Ensemble

- Simple ensemble by adding up class scores with weights:

$$q_{\text{RGB}} = \alpha_1 q_{\text{skel}} + \alpha_2 q_{\text{RGB}} + \alpha_3 q_{\text{flow}} + \alpha_4 q_{\text{feat}}, \qquad (6)$$

$$\begin{aligned} q_{\text{RGB-D}} = & \alpha_1 q_{\text{skel}} + \alpha_2 q_{\text{RGB}} + \alpha_3 q_{\text{flow}} + \alpha_4 q_{\text{feat}} \\ & + \alpha_5 q_{\text{HHA}} + \alpha_6 q_{\text{depthflow}}, \end{aligned} \qquad (7)$$

- Other ensemble methods tried:
  - Using fully-connected layers before or after class scores.
  - Problem: introduces too many parameters.

- Hyper-parameter tuning:
  - Hyper parameters are tuned on validation set.
  - Rule of thumb: higher-accuracy model is given larger weights.
  - Introduce new model one by one while keeping the existing weights fixed.

# Overall Performance: Multi-modal and Ensembles

| Modality | Top-1 | Top-5 |
|---|---|---|
| Baseline RGB | 42.58 | - |
| Baseline RGB-D | 63.22 | - |
| Keypoints | **95.45** | **99.25** |
| Features | 94.32 | 98.84 |
| RGB Frames | 94.77 | 99.48 |
| RGB Flow | 91.65 | 98.76 |
| Depth HHA | 95.13 | 99.25 |
| Depth Flow | 92.69 | 98.87 |

Table 6. Results of single modalities on AUTSL validation set.

| Ensemble | K | F | R | O | H | D | Top-1 | Top-5 |
|---|---|---|---|---|---|---|---|---|
| Skeleton | ✓ | ✓ | | | | | 96.11 | 99.43 |
| RGB+Flow | | | ✓ | ✓ | | | 95.77 | 99.52 |
| RGB All | ✓ | ✓ | ✓ | ✓ | | | 96.96 | 99.68 |
| Depth | | | | | ✓ | ✓ | 95.76 | 99.41 |
| RGB+D | | | ✓ | ✓ | ✓ | ✓ | 96.27 | 99.66 |
| RGBD All | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **97.10** | **99.73** |

Performance of multi-modal ensembles on val set

# Overall Performance: Test Phase

- During test phase, we finetune our models using the validation set.
- The finetuned results further improve the recognition rate.
  - RGB: 98.42%
  - RGB-D: 98.53%
- The above results won the 1st rank in both RGB and RGB-D tracks.

| | Finetune | Track | Top-1 |
|---|---|---|---|
| Baseline | - | RGB | 49.23 |
| Baseline | - | RGB-D | 62.03 |
| Ensemble | No | RGB | 97.51 |
| Ensemble | No | RGB-D | 97.68 |
| Ensemble | w/ Val | RGB | **98.42** |
| Ensemble | w/ Val | RGB-D | **98.53** |

Performance of submissions in the challenge test set

# Conclusions

- We proposed a novel Skeleton Aware Multimodal SLR framework (SAM-SLR) to take advantage of multi-modal information towards effective SLR.
- Our frameworks includes:
  - SL-GCN for skeleton keypoints modality.
  - SSTCN for skeleton features modality.
  - 3DCNN baselines for RGB, Optical Flow and Depth modalities.
- Our multi-modal ensemble results achieves the state-of-the-art performance and won the challenge in both RGB and RGB-D tracks.

# Thank you!