MICCAI2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

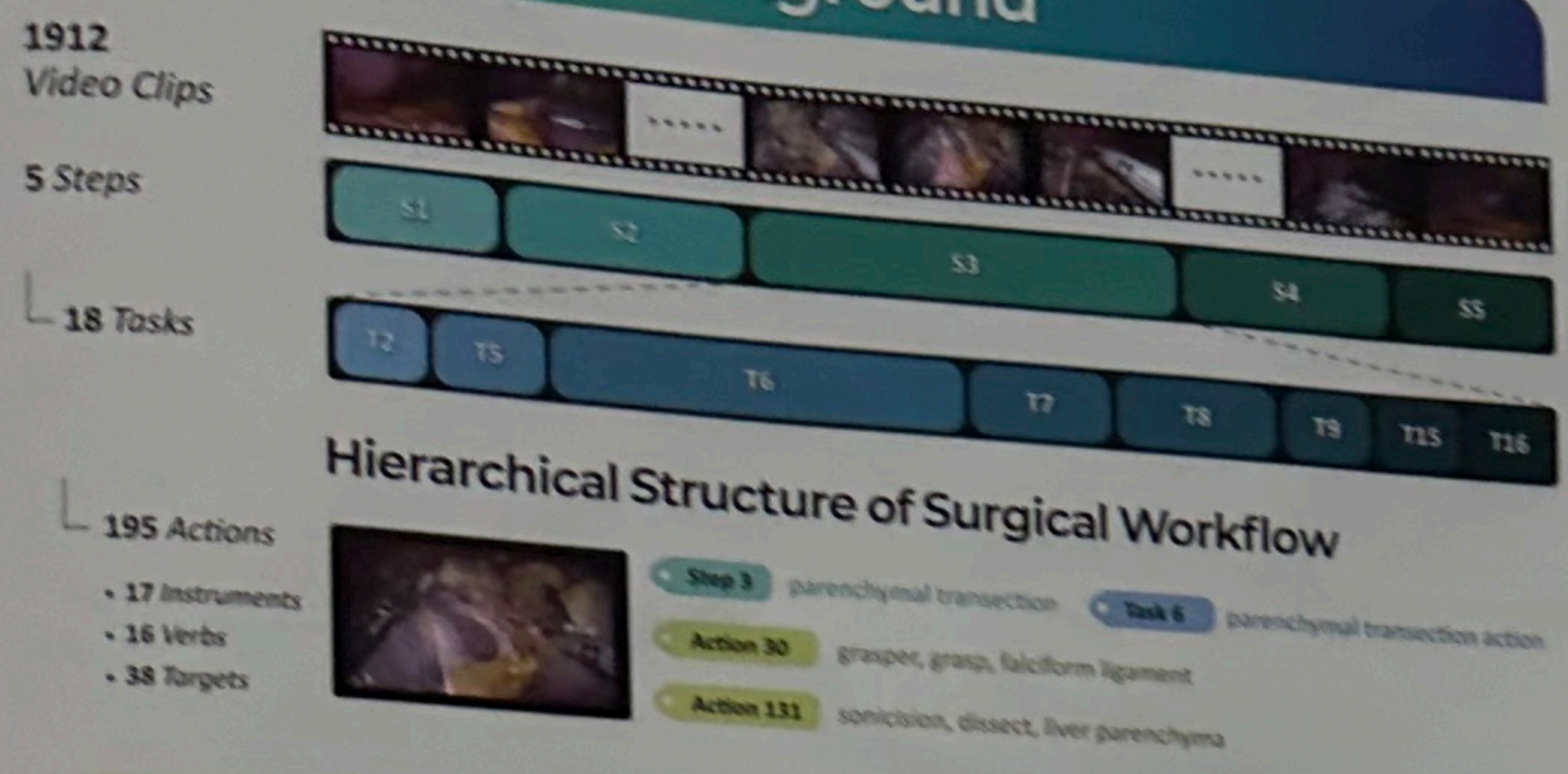
# CurConMix: A Curriculum Contrastive Learning Framework for Enhancing Surgical Action Triplet Recognition

Yongjun Jeon[1,†], Jongmin Shin[2,†], Seonmin Park[2], Bogeun Kim[2], Kanggil Park[2], Namkee Oh[2,‡], Kyu-Hwan Jung[1,2,‡]

[1]Sungkyunkwan University [2]Samsung Medical Center

MI DAS LAB
SMC SAMSUNG MEDICAL CENTER
SUNGKYUNKWAN UNIVERSITY

## Background



Hierarchical Structure of Surgical Workflow

## Challenges

- Complex Interdependencies
- Fine-grained Distinctions
- Severe Class Imbalance & Limited Training Data

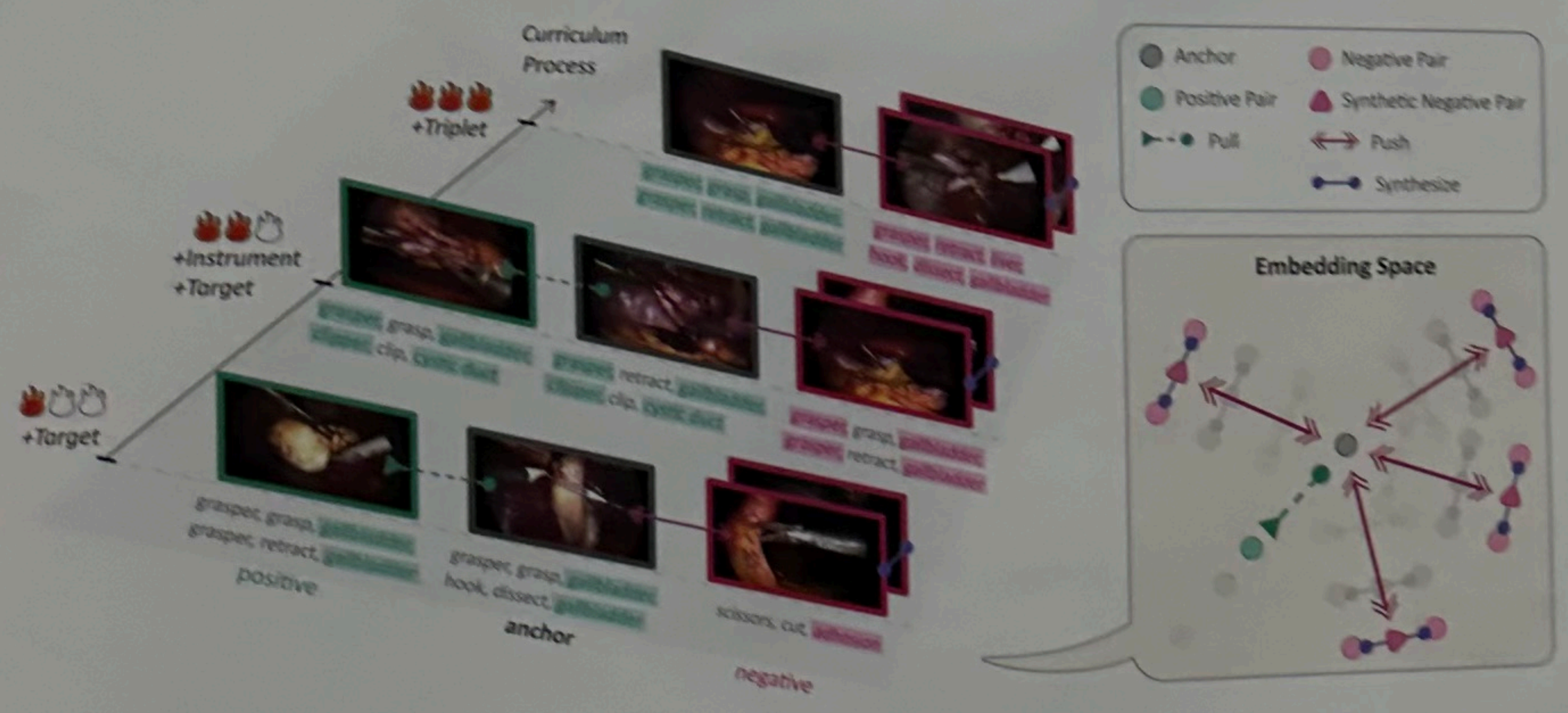
## Contributions

- Curriculum Contrastive Learning with Feature Mixup (CurConMix)
- Hard Pair Sampling & Synthetic Negatives to increase diversity and robustness
- State-of-the-Art Performance

## Methods

### Pre-Training Stage : Curriculum Contrastive Learning

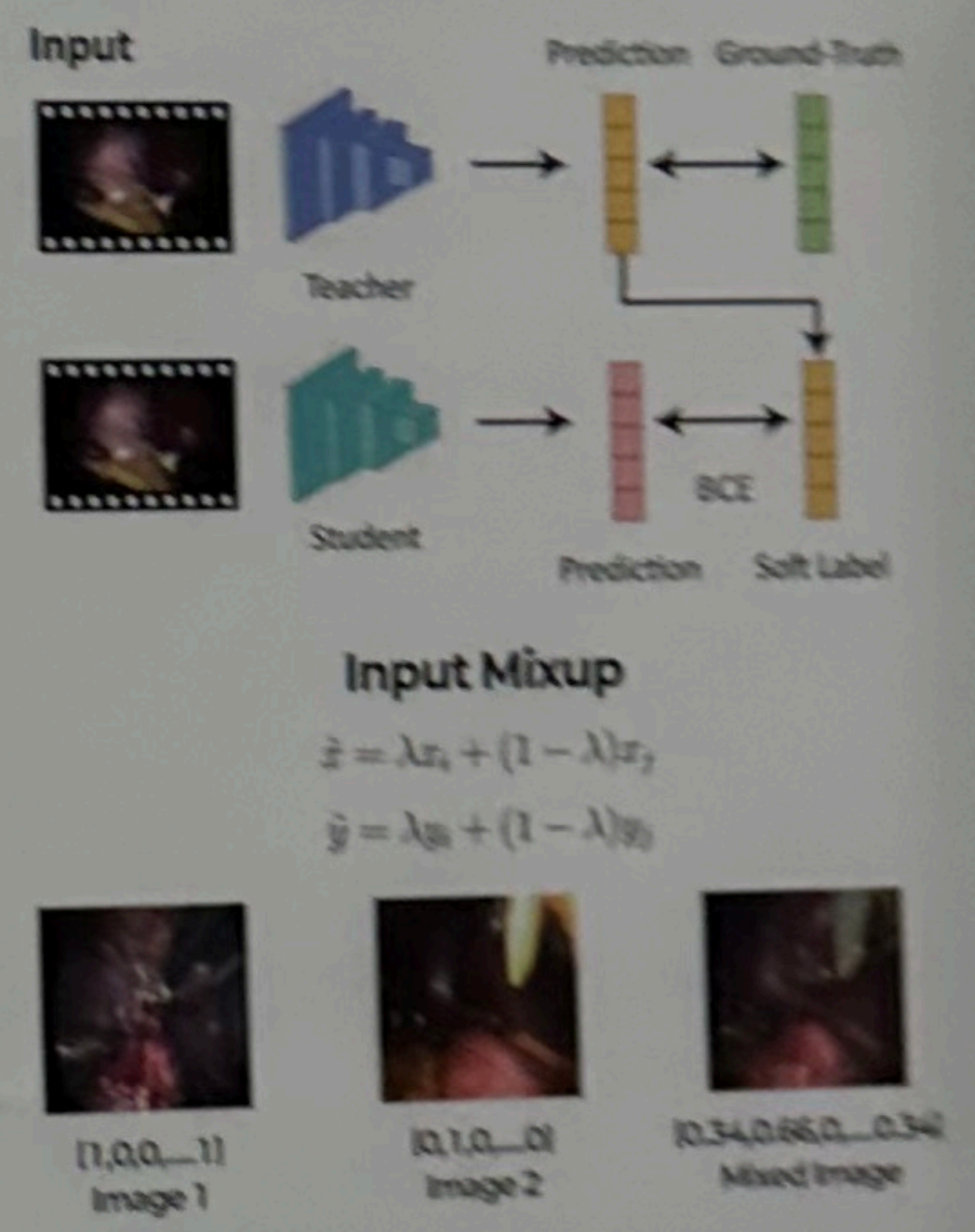
1. Pair Generation Rule – Hard Negative Sampling
2. Pair Generation Rule – Hard Positive Sampling
3. Synthetic Hard Negatives
4. Supervised Contrastive Learning


Curriculum
Process
+Triplet
+Instrument
+Target
+Target
grasper, grasp, gallbladder
grasper, retract, gallbladder
positive
grasper, grasp, gallbladder
hook, dissect, gallbladder
anchor
scissors, cut, adhesion
negative
grasper, grasp, gallbladder
clipper, clip, cystic duct
grasper, retract, gallbladder
clipper, clip, cystic duct
grasper, grasp, gallbladder
grasper, retract, gallbladder
grasper, grasp, gallbladder
grasper, retract, gallbladder
grasper, retract, liver
hook, dissect, gallbladder
Anchor
Negative Pair
Positive Pair
Synthetic Negative Pair
Pull
Push
Synthesize
Embedding Space

(a) Curriculum Contrastive Learning

(b) Contrastive Learning

### Fine-Tuning Stage : Self-Distillation for Action-Triplet Recognition



## Results

Table 1. Comparison of single models from different approaches on the provided 5-fold validation split of the CholecT45 dataset. **Bold** font indicates the best performance within comparable models. Results marked with † were reproduced using the official code. TERL-B(384) was reproduced with a batch size of 12 due to hardware constraints.

| Method | Backbone | $AP_I$ | $AP_V$ | $AP_T$ | $AP_{IV}$ | $AP_{IT}$ | $AP_{IVT}$ |
|---|---|---|---|---|---|---|---|
| RDV [15] | Res18 | 89.3±2.1 | 62.0±1.3 | 40.0±1.4 | 34.0±3.3 | 30.8±2.1 | 29.4±2.8 |
| RiT [16] | Res18 | 88.6±2.6 | 64.0±2.5 | 43.4±1.4 | 38.3±3.5 | 36.9±1.0 | 29.7±2.6 |
| TDN [2] | Res50 | 91.2±1.9 | 65.3±2.8 | 43.7±1.6 | - | - | 33.8±2.5 |
| MT4MTL-KD [6] | SwinL(384) | 93.1±2.1 | 71.8±3.4 | 48.8±3.8 | 44.9±2.4 | 43.1±2.0 | 37.1±0.5 |
| SelfD [21] | SwinB(224)† | 90.3±2.3 | 67.4±1.5 | 47.9±1.8 | 43.7±4.1 | 42.9±1.6 | 37.1±1.9 |
| TERL-T [5] | SwinT(224)† | 93.5±1.5 | 71.4±2.2 | 47.2±2.6 | 44.7±3.8 | 42.0±2.4 | 35.7±1.6 |
| TERL-B [5] | SwinB(224)† | 93.9±2.0 | 70.8±2.3 | 49.4±4.7 | 43.9±3.4 | 43.6±2.6 | 35.6±1.4 |
| TERL-B [5] | SwinB(384)† | 94.1±2.3 | 73.0±1.4 | 51.1±3.8 | 46.5±4.9 | 44.9±1.8 | 37.7±1.5 |
| TERL-Ens [5] | Ensemble† | 94.6±1.9 | 73.5±1.9 | 50.8±3.3 | 47.3±4.1 | 45.3±1.9 | 38.5±1.1 |
| CurConMix-T | SwinT(224) | 90.4±2.1 | 67.8±1.8 | 48.3±3.4 | 43.3±2.9 | 43.3±1.8 | **37.7±2.1** |
| CurConMix-B | SwinB(224) | 90.4±3.0 | 68.2±1.5 | 49.7±2.5 | 44.8±5.4 | 45.3±2.4 | **38.8±2.8** |
| CurConMix-B | SwinB(384) | 90.9±2.0 | 68.3±1.3 | 49.8±3.2 | 45.2±4.2 | 45.1±1.1 | **39.1±2.0** |
| CurConMix-Ens | Ensemble | 91.7±2.2 | 69.5±0.4 | 51.3±2.9 | 46.3±5.0 | 47.1±1.6 | **40.7±2.1** |

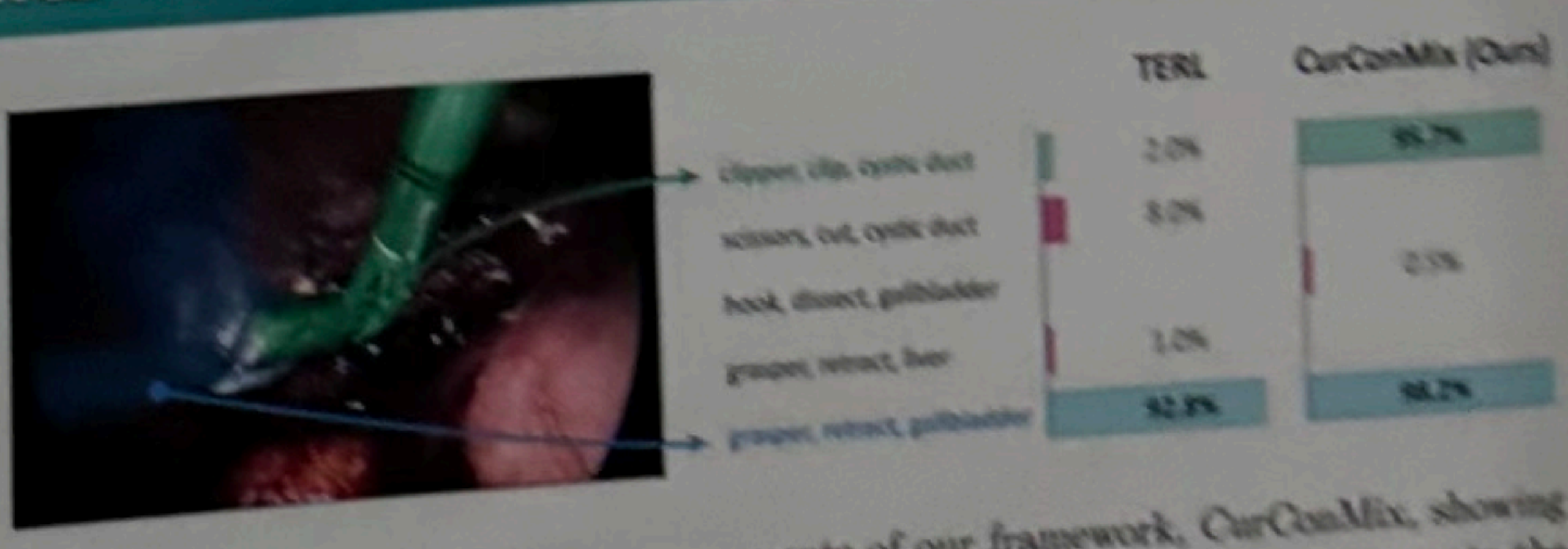


Table 2. Ablation study on the components of our framework, CurConMix, showing performance improvement as each component is added. The first row represents the baseline model, marked with an *.

| Contrastive | Curriculum | Input Mixup | Feature Mixup | $AP_{IVT}$ |
|---|---|---|---|---|
| | | | | 37.1* |
| ✓ | | | | 37.8 |
| ✓ | ✓ | | | 38.1 |
| ✓ | ✓ | ✓ | | 38.3 |
| ✓ | ✓ | ✓ | ✓ | 38.8 |

Paper ID #0871