Stony Brook University
UNIVERSITY OF NORTH CAROLINA CHARLOTTE

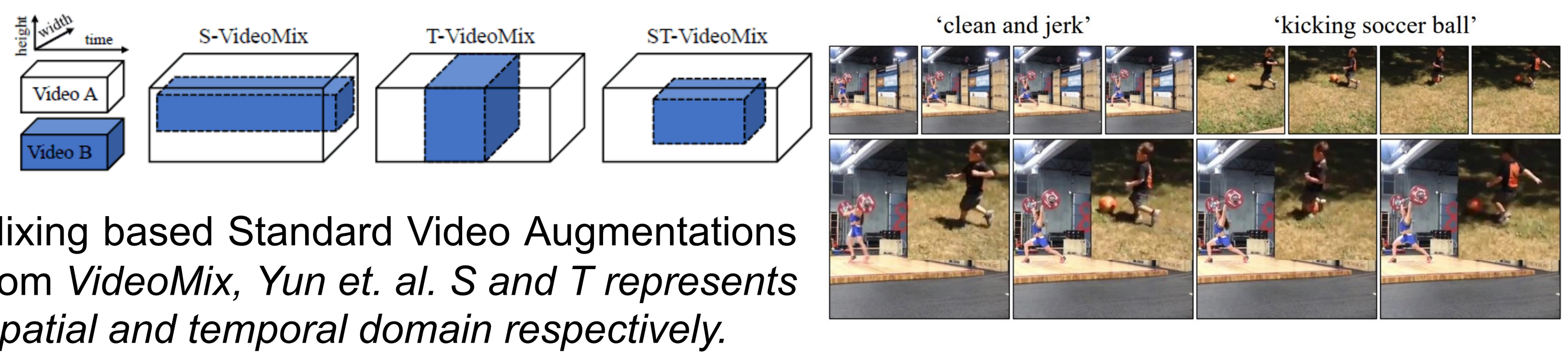
Cross-modal Manifold Cutmix for Self-supervised Video Representation Learning

MVA2023

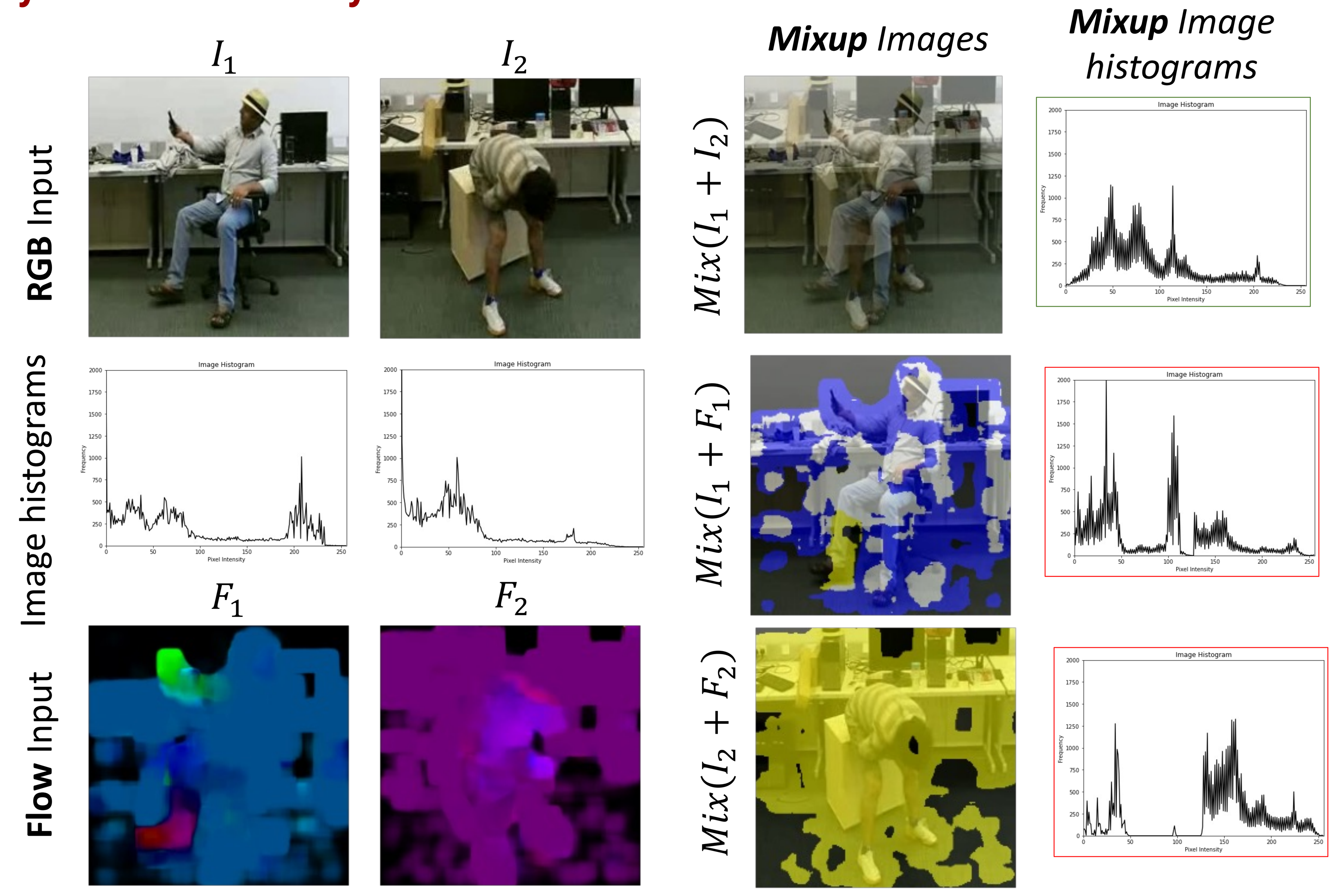
Srijan Das[1] and Michael S. Ryoo[2]
[1]UNC Charlotte [2]Stony Brook University

Motivation

- **Self-supervised Video representation learning** require large-scale video datasets, which is impractical in limited data domains.
- **Mixing strategies** have not been explored exhaustively in video domain.
- How do we leverage **multiple modalities** in video representation learning?



Mixing based Standard Video Augmentations from *VideoMix, Yun et. al. S and T represents Spatial and temporal domain respectively.*



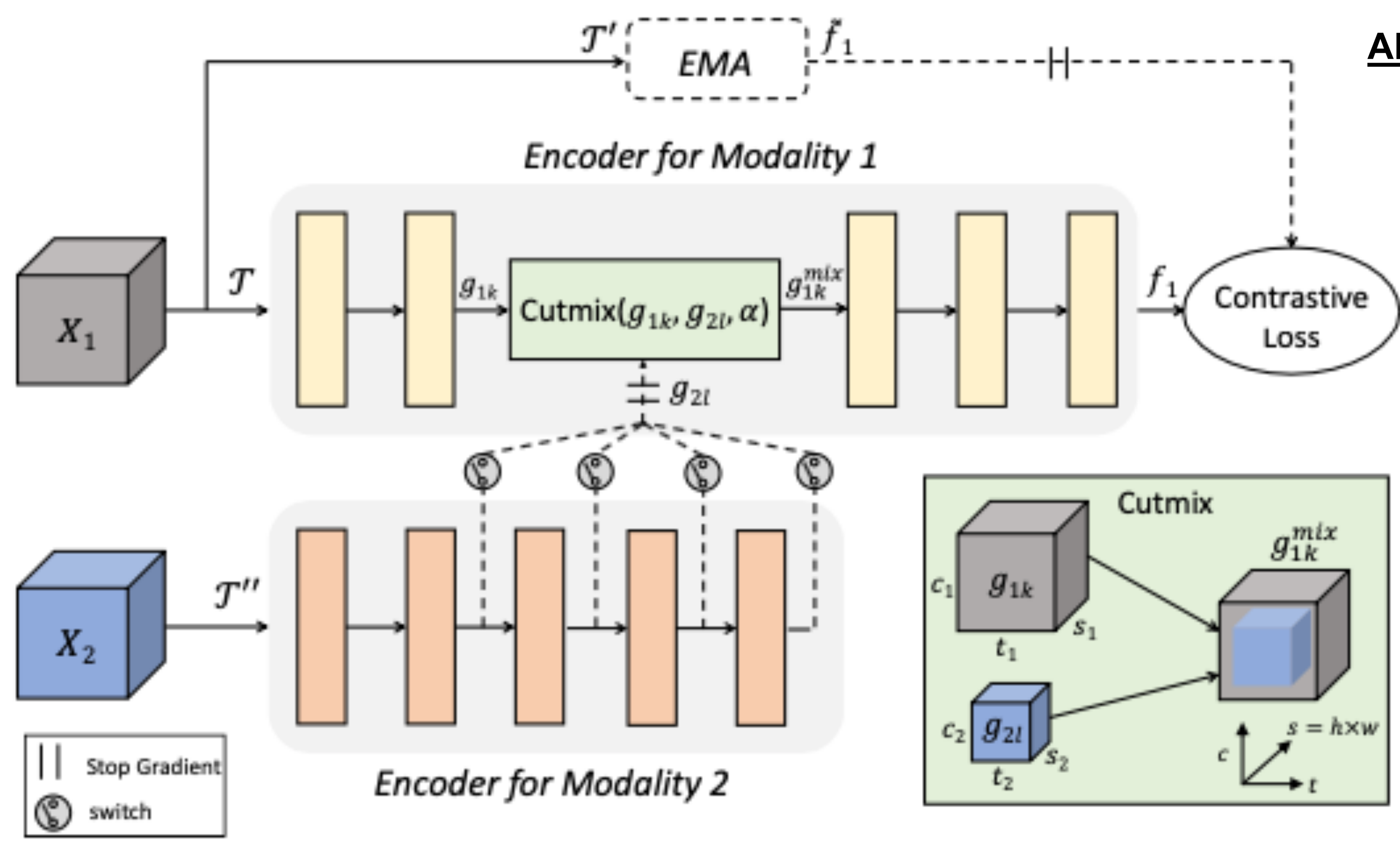
★ Input distribution changes when mixing different modalities!

Proposed Video Augmentation for Self-supervised Video Representation Learning

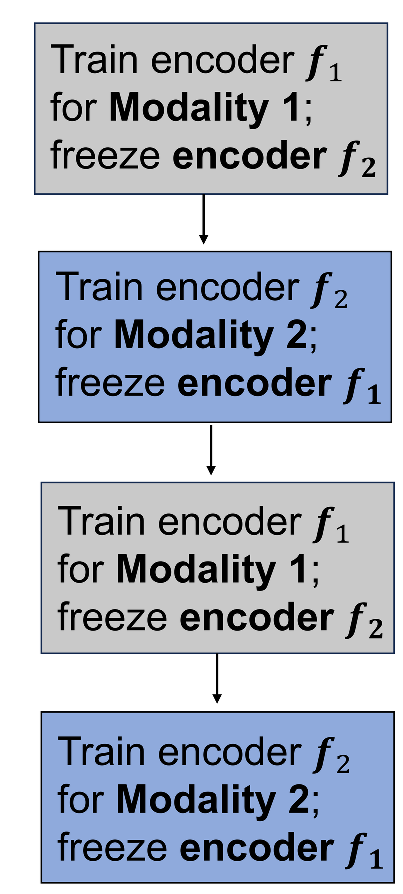
- **CMMC** (*Cross-modal Manifold Cutmix*) –
  - Performs data mixing across modalities of a video in their **hidden intermediate representations**.
  - CMMC enables video models in learning **self-supervised representations** with **Limited data**.
  - First video augmentation that performs data mixing across **channels**.

Algorithm: *Pytorch-style Pseudocode of CMMC for Modality 1:*

```
Input:    X1, X2
Output:  loss

alpha, k = 1.0, rand(1, N)                 # N is the layers in the encoder
l = rand(k, N)
# T and T'' are the augmentations
x1, x̃1 = T(X1)                              # Two views of the modality 1
x2 = T''(X2)                                # modality 2 data

#f1 and f2 are the video encoders for modality 1 and 2
g1k = f1.partial_forward(x1, 0, k)
g2l = f2.partial_forward(x2, 0, l)
g1k^mix, labels_new, lam = CutMix(g1k , g2l , alpha)

z1 = normalize(f1.partial_forward(g1k^mix, l, N))
z2 = normalize(f̃1.forward(x̃1))
z2, g2l = z2.detach(), g2l .detach()        # no gradient flow

# compute loss
labels = zeros(len(x1))
logits = matmul(z1, z2.T) / t              # t is the temperature
loss = lam * CE(logits, labels) + (1 - lam) * CE(logits, labels_new)
```



Alternative Training strategy



CutMix operation across modalities -

$$g_{1k}^{mix} = M \odot g_{1k} + (1\text{-}M) \odot g_{2l}$$

g_{1k}: intermediate representation of **modality 1** at k^{th} layer; g_{2l}: intermediate representation of **modality 2** at l^{th} layer

Randomly sample a center coordinate (M_{cc}, M_{cc}, M_{cc}, M_{cc}) from $U(0, c_1)$, $U(0, t_1)$, $U(0, h_1)$, and $U(0, w_1)$.

Binary Mask M is obtained from -

$$M_{c1}, M_{c2} = M_{cc} - {c^2}/{2}, M_{cc} + {c^2}/{2}$$
$$M_{t1}, M_{t2} = M_{tc} - {t^2}/{2}, M_{tc} + {t^2}/{2}$$
$$M_{h1}, M_{h2} = M_{hc} - {h^2}/{2}, M_{hc} + {h^2}/{2}$$
$$M_{w1}, M_{w2} = M_{wc} - {w^2}/{2}, M_{wc} + {w^2}/{2}$$

Mixing coefficient returned by cutmix ⟶ $\lambda = 1 - \sum_{c,t,w,h} M_{c,t,w,h} / vol(g_{1k})$

Experiments

Impact of different strategies in CMMC for downstream tasks like Action classification and video retrieval.

	Cross-Modal Manifold Mixing strategies	Action cls. Linear probe UCF	Action cls. Linear probe HMDB	Retrieval R@1 UCF	Retrieval R@1 HMDB
RGB	MoCo (Baseline)	46.8	23.1	33.1	15.2
	+ mixup	52.8	24.4	37.6	17.6
	+ CM mixup	53.9	25.1	40.3	17.6
	+ CM cutmix	**55.8**	**28.3**	**42.8**	**19.1**
OF	MoCo (Baseline)	66.8	30.3	45.2	20.8
	+ mixup	68.6	33.1	48.7	19.5
	+ CM mixup	70.4	33.1	51.2	21.0
	+ CM cutmix	**72.4**	**34.9**	**53.9**	**23.1**
2stream	MoCo (Baseline)	68.1	33.1	49.8	21.9
	+ mixup	71.3	36.3	53.8	24.5
	+ CM mixup	72.2	35.9	56.1	25.3
	+ CM cutmix	**74.0**	**38.1**	**58.1**	27.1

State-of-the-art comparison of CMMC with 3D Poses for **Action classification** on NTU-60 dataset.

Method	$\mathcal{M}$	NTU-60 CS	NTU-60 CV
LongTGAN (Wang et al., ICCV 17)	J	39.1	48.1
MS^{2}L (Lin et al., MM 20)	J	52.6	-
AS-CAL (Nie et al., ECCV 20)	J	58.5	64.8
P&C (Rao et al., CVPR 20)	J	50.7	76.3
SeBiReNet (Li et al., CVPR 21)	J	-	79.7
SkeletonCLR* (Baseline)	J + M	70.1	77.2
CMMC (Skeleton)	J + M	72.5	79.1
2s-CrosSCLR (Li et al., CVPR 21)	J + M	74.5	82.1
CMMC (2s-Skeleton)	J + M	**75.2**	**83.1**

State-of-the-art comparison of CMMC for **Nearest Neighbor video retrieval** on *UCF101*. Testing set clips are used to retrieve training set videos and *R@k* is reported for k ∈ {1, 5, 10, 20}.

Method	Dataset	UCF101 R@1	R@5	R@10	R@20
Jigsaw (Noroozi et al., ECCV 16)	UCF	19.7	28.5	33.5	40.0
OPN (Lee et al., ICCV 17)	UCF	19.9	28.7	34.0	40.6
RL-method (Benaim et al., CVPR 20)	UCF	25.7	36.2	42.2	49.2
VCOP (Xu et al., CVPR 19)	UCF	14.1	30.3	40.4	51.1
VCP (Luo et al., AAAI 20)	UCF	18.6	33.6	42.5	53.5
MemDPC (Han et al., ECCV 20)	UCF	20.2	40.4	52.4	64.7
SpeedNet (Benaim et al., CVPR 20)	K400	13.0	28.1	37.5	49.5
CoCLR (Han et al., NeurIPS 20)	UCF	55.9	70.8	76.9	82.5
CMMC	UCF	**58.1**	**76.5**	**83.4**	**88.7**

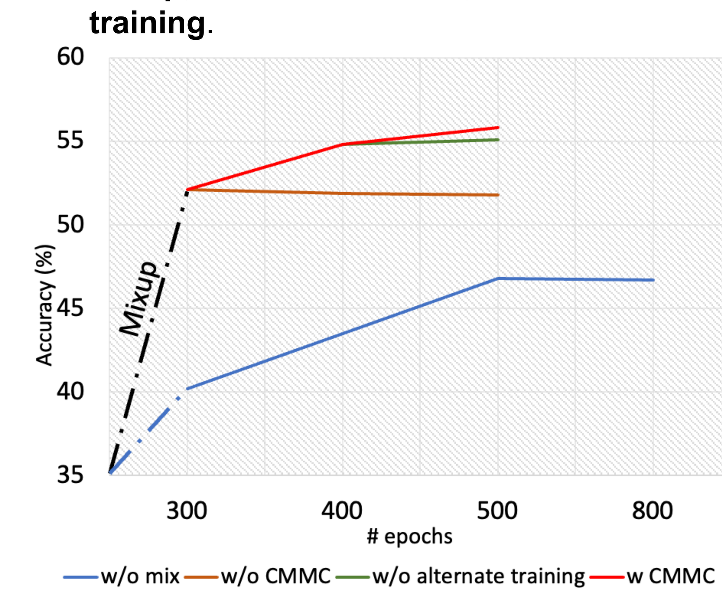
State-of-the-art comparison of CMMC with existing SSL based video models on **Action classification** dataset, UCF-101 and HMDB.

Method	GFLOPs	$\mathcal{M}$	Linear Probe UCF	Linear Probe HMDB	Fine-tune UCF	Fine-tune HMDB
OPN (Lee et al., ICCV 17)	16	V	-	-	59.6	23.8
VCOP (Xu et al., CVPR 19)	12.5	V	-	-	72.4	30.9
CoCLR-RGB (Han et al., NeurIPS 20)	11	V	70.2	39.1	81.4	52.1
ρBYOL† (Feichtenhofer et al., CVPR 21)	22	V+F	70.2	37.8	84.9	57.6
CoCLR (Han et al., NeurIPS 20)	22	V+F	72.1	40.2	87.3	58.7
CMMC	22	V+F	74.0	38.1	87.5	59.1
CoCLR+CMMC	11	V	71.3	39.4	82.5	53.2
CoCLR+CMMC	22	V+F	**74.7**	**40.8**	**87.9**	**59.0**

State-of-the-art comparison of CMMC on NTU-60 using RGB (R) and Pose (P) modalities.

	Method	$\mathcal{M}$	Extra Data	Pre-train Dataset	NTU-60 CS	NTU-60 CV
Supervised	I3D (Carreira and Zisserman, CVPR 17)	R	✓	K400	85.5	87.3
	NPL (Piergiovanni and Ryoo, CVPR 21)	R	✓	K400	-	93.7
	STA (Das et al., ICCV 19)	R+P	✓	K400	92.2	94.6
	VPN (Das et al., ECCV 20)	R+P	✓	K400	**93.5**	**96.2**
SSL	MoCo (S3D)	R	×	NTU-60	87.5	91.3
	CMMC	R	×	NTU-60	88.1	92.0
	CMMC	R+P	×	NTU-60	91.4	95.1

Accuracy vs #epochs graph illustrating the improvements in models trained with **CMMC** compared to **baselines without using mixup, manifold mix,** and **alternate training**.



Conclusion

- CMMC is a video augmentation that can be enable video SSLs in Limited domains.
- CMMC can be used with any SSL.
- CMMC can be applied across various modalities, including RGB, Optical Flow, and Poses.
- CMMC improves performance of downstream tasks like video retrieval and action classification.