Video-based Person Re-identification without Bells and Whistles

Chih-Ting Liu[1], **Jun-Cheng Chen**[2], **Chu-Song Chen**[3], **Shao-Yi Chien**[1]

[1]Graduate Institute of Electronics Engineering, National Taiwan University
[2]Research Center for Information Technology Innovation, Academia Sinica
[3]Computer Science and Information Engineering, National Taiwan University

CVPR VIRTUAL JUNE 19-25

Motivation

MARS, one of the largest video-based Re-ID dataset, is very noisy.

raw videos

obsolete detection and tracking

training video tracklet

severe spatial and temporal misalignment!

State-of-the-arts !!

[AP3D, ECCV2020] → Normal backbone + Spatial Alignment

Spatial Alignment

Spatial Attention

⋮

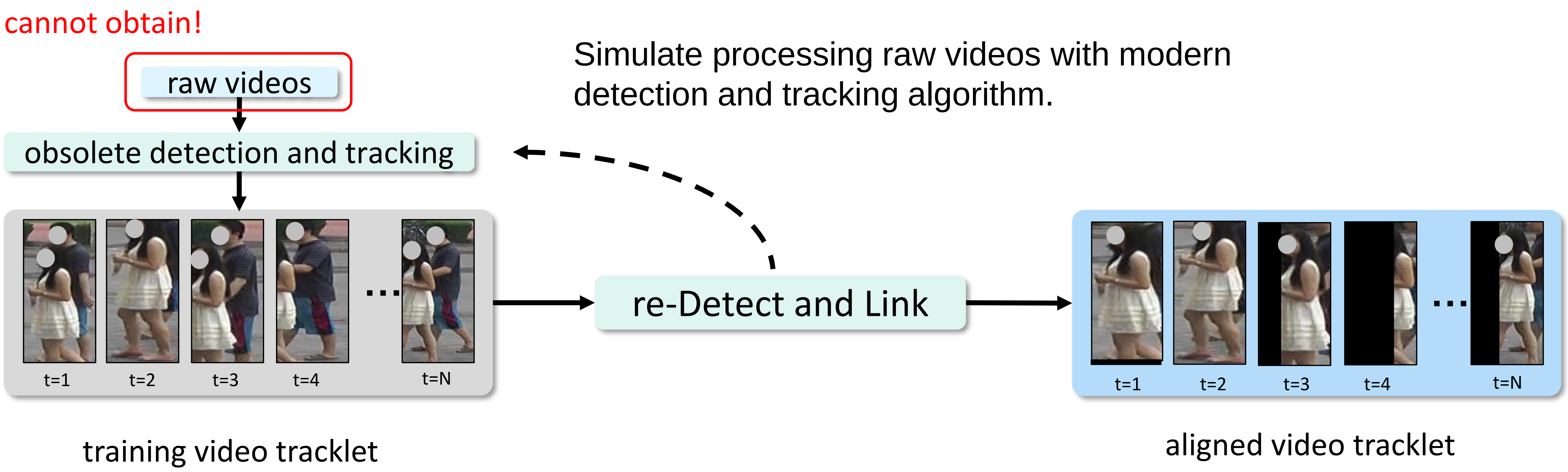
Temporal Attention

various methods

What if all methods train on **already well-aligned** dataset??

re-Detect and Link Module (DL)

Simulate processing raw videos with modern detection and tracking algorithm.

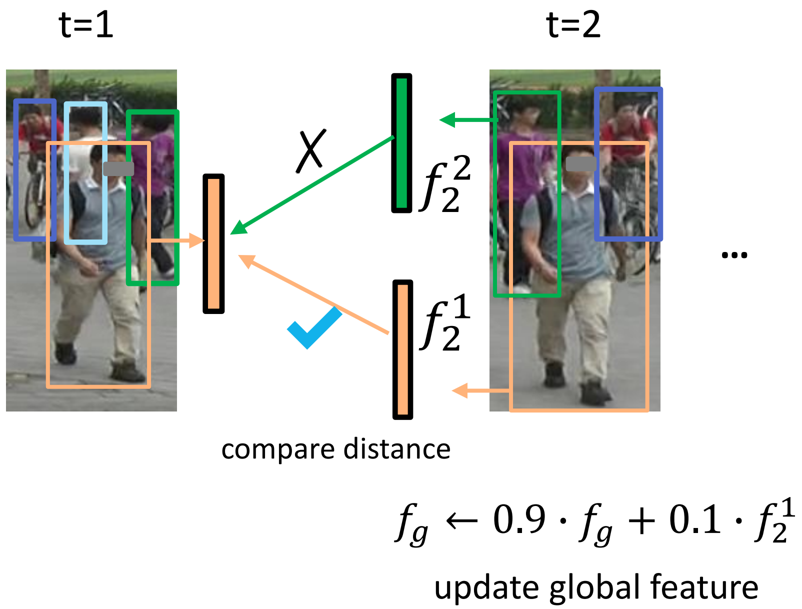


re-Detect and Link Module (DL)

re-Detect

Detect with deep-learning based but efficient object detector.

Link



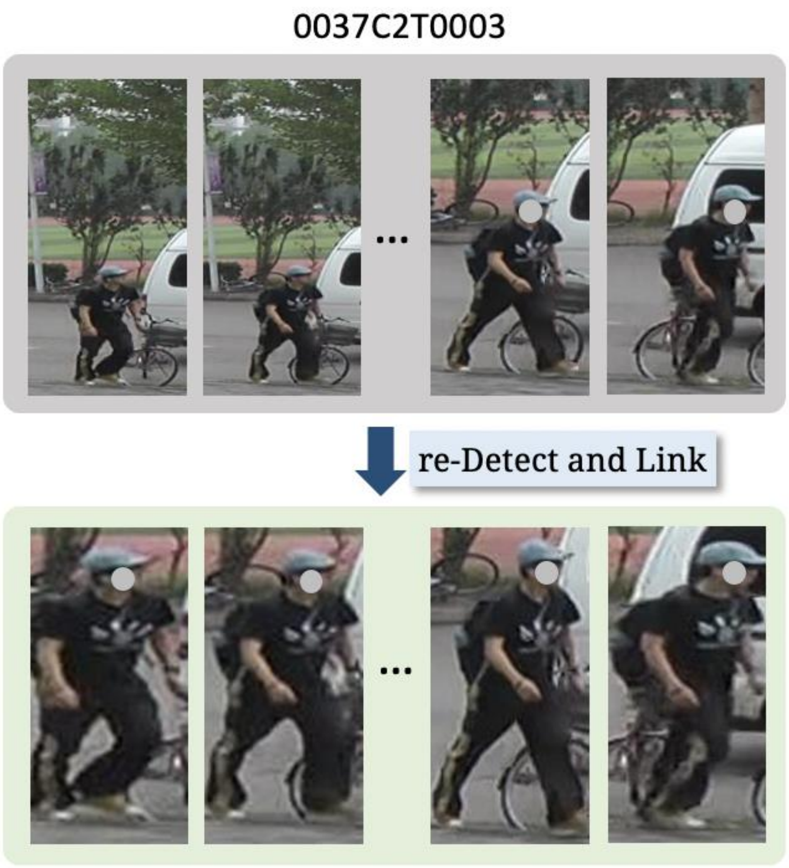
First frame → largest bbox
Latter frames → compare feature distance

Padding

Padding based on aspect ratio and spatial position

Examples of DL

spatial misalignment



multiple identities

Reproduce existing methods on MARS

- We only alter the input tracklet that is processed with our DL module.

Method	Original Results		w/ our DL	
	mAP	rank-1	mAP	rank-1
FT-WFT [30]	82.9	88.6	83.8	90.0
P3D-C [31, 10]	83.1	88.5	85.0	91.0
C2D [10]	83.4	88.9	84.9	91.0
Non-Local [10, 25]	85.0	89.6	**86.2**	**91.4**
TCLNet [14]	85.1	89.8	85.8	90.8
AP3D [10]	**85.1**	**90.1**	85.4	91.0

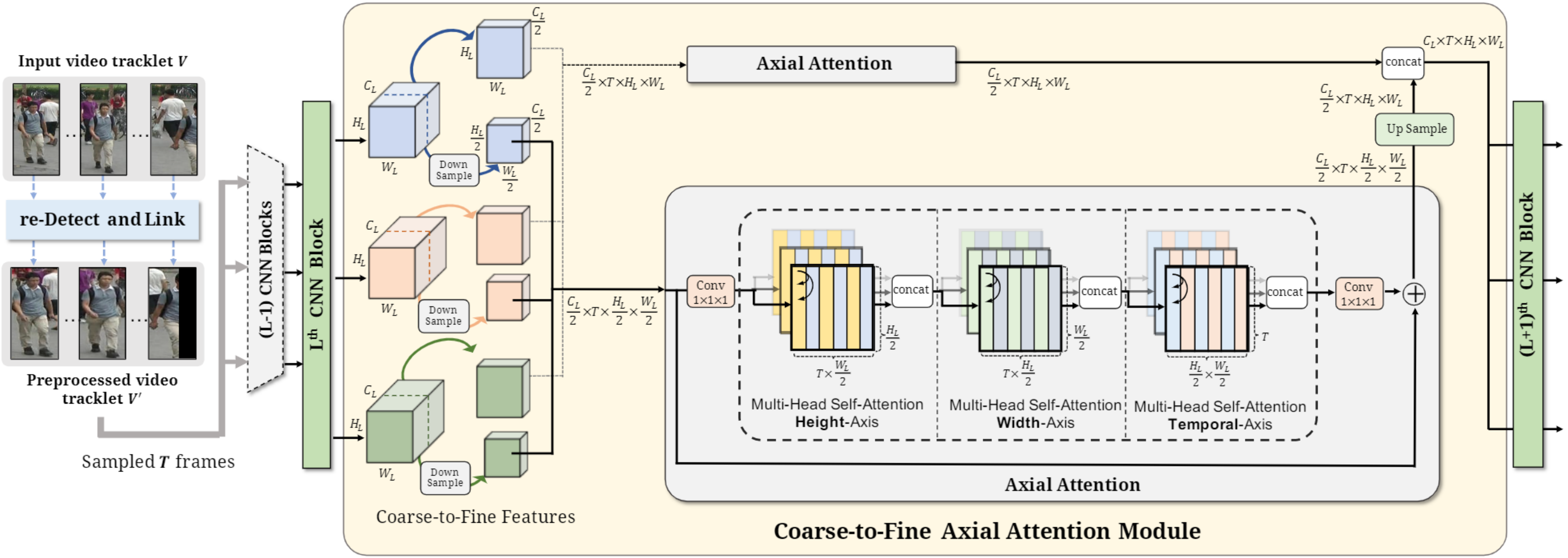
original state-of-the-art (SOTA)

Surprisingly, **without bells and whistles**, a baseline method (C2D) can compete to the SOTA!!!!!

Those methods with **Non-local attention** (spatial and temporal attention) are the new SOTA !!

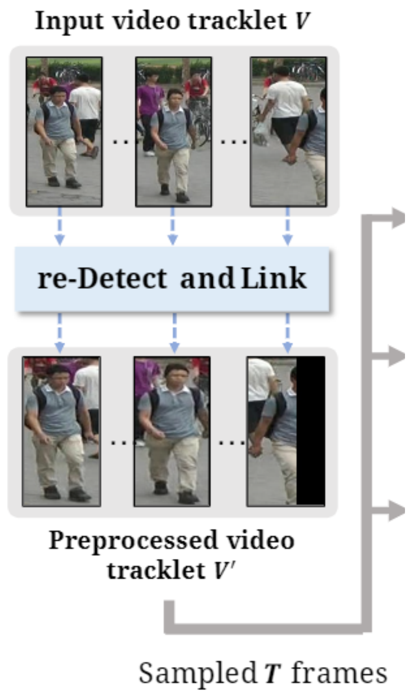
Proposed Video-based Re-ID Architecture

Based on Non-local Network, we proposed **CF-AA Network**.
We replace the Non-local module with our **Coarse-to-Fine Axial Attention module (CF-AA), which is more efficient in computation.**



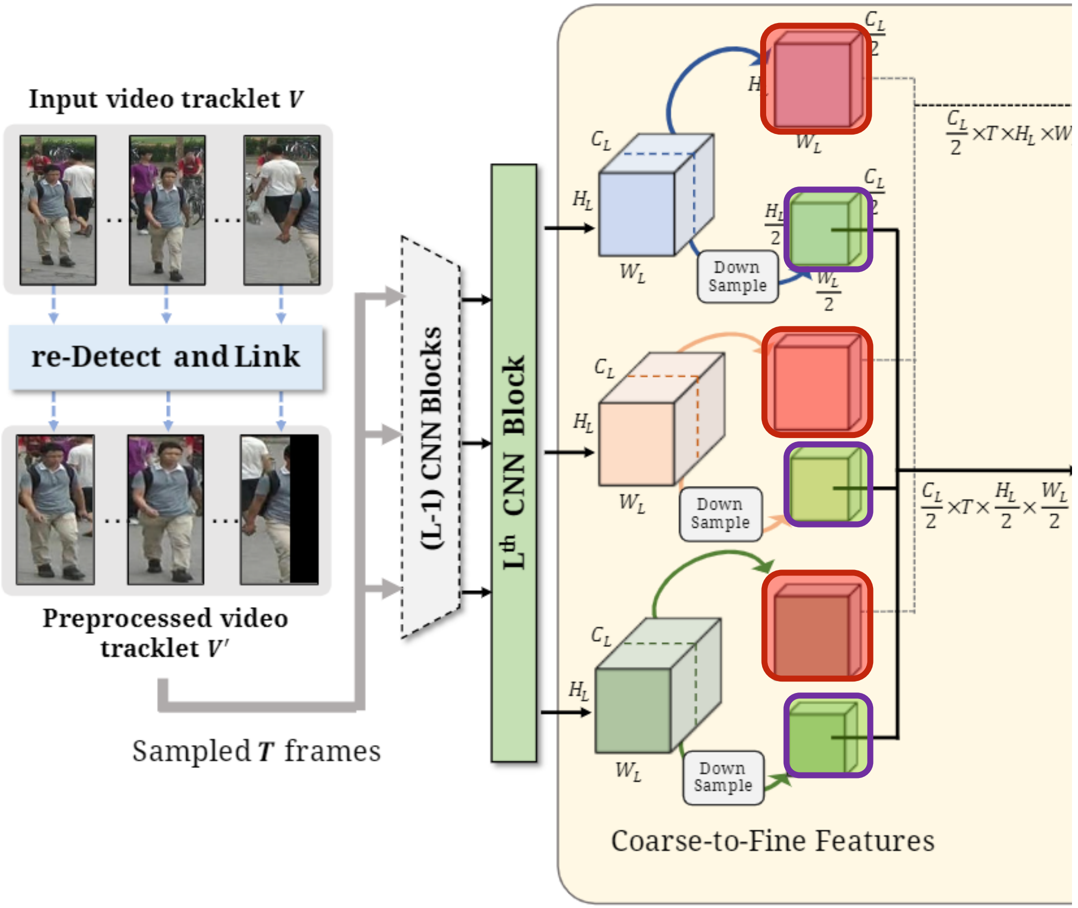
Proposed Video-based Re-ID Architecture

Based on Non-local Network, we proposed **CF-AA Network**.
We replace the Non-local module with our **Coarse-to-Fine Axial Attention module (CF-AA), which is more efficient in computation.**



Proposed Video-based Re-ID Architecture

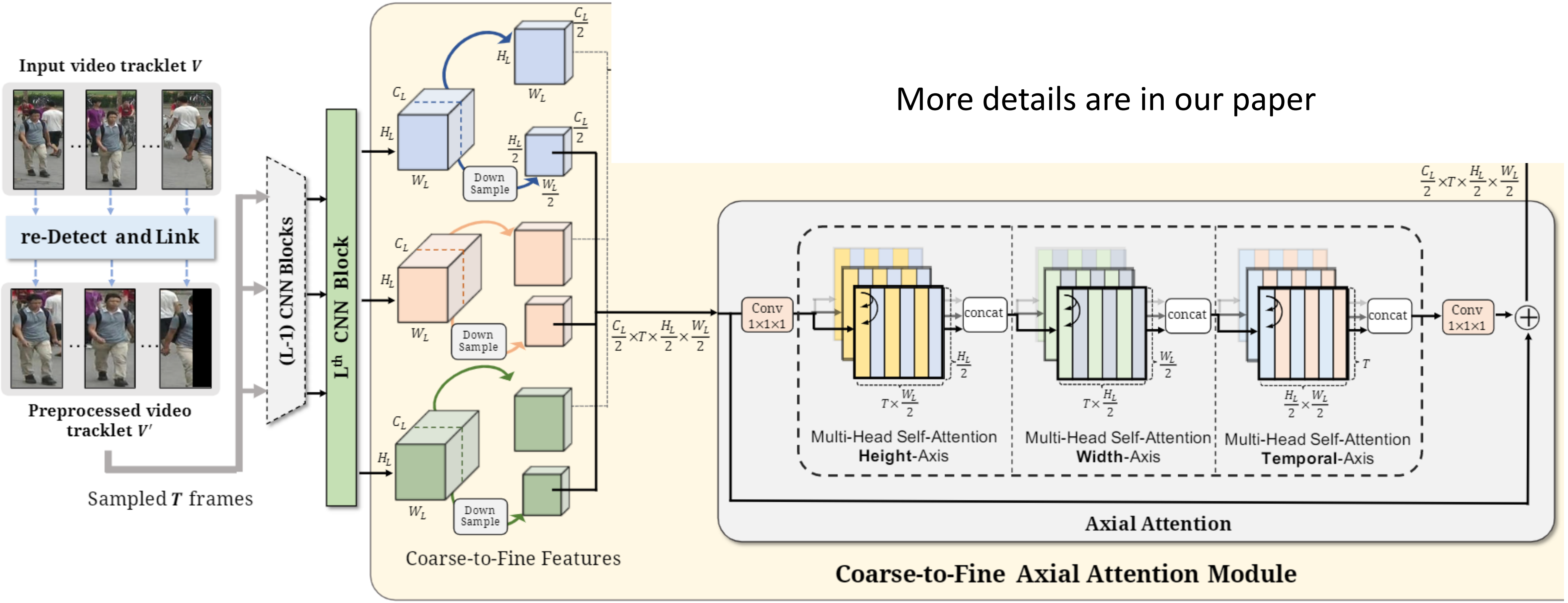
Based on Non-local Network, we proposed **CF-AA Network**.
We replace the Non-local module with our **Coarse-to-Fine Axial Attention module (CF-AA), which is more efficient in computation.**



Proposed Video-based Re-ID Architecture

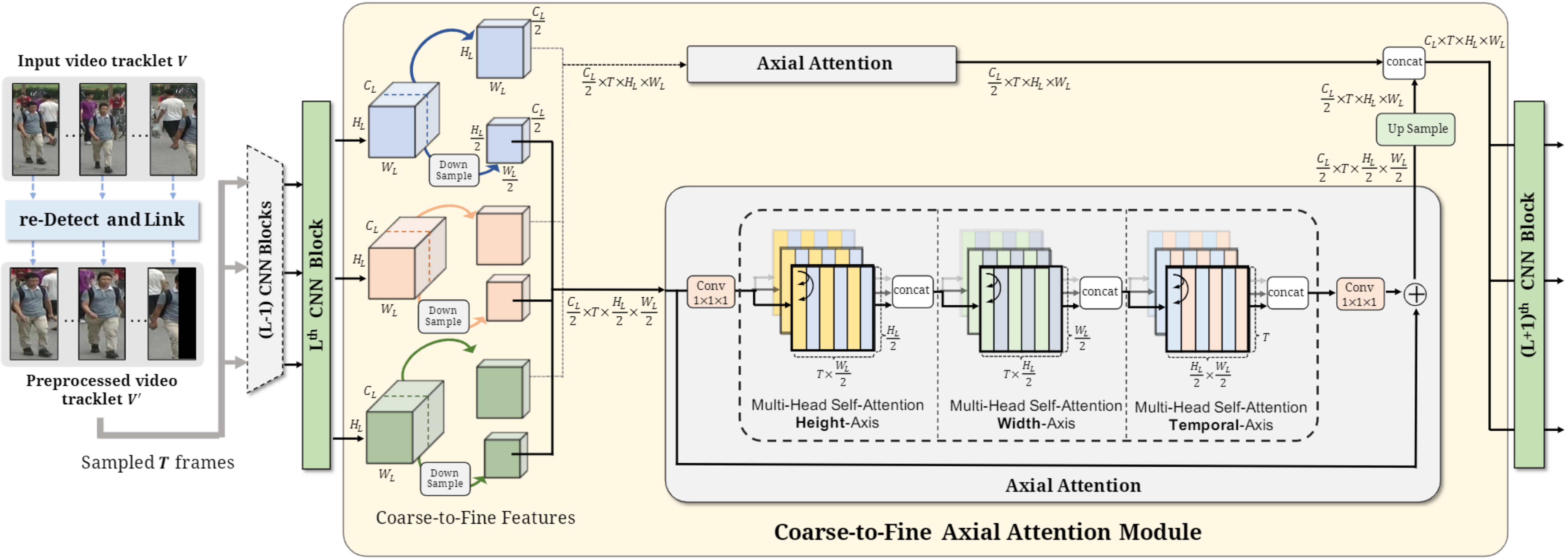
Based on Non-local Network, we proposed **CF-AA Network**.
We replace the Non-local module with our **Coarse-to-Fine Axial Attention module (CF-AA), which is more efficient in computation.**

More details are in our paper



Proposed Video-based Re-ID Architecture

Based on Non-local Network, we proposed **CF-AA Network**.
We replace the Non-local module with our **Coarse-to-Fine Axial Attention module (CF-AA), which is more efficient in computation.**



Experiment Results

Table 2: **The Ablation Study of our DL and CF-AAN.** We compare the effectiveness of our DL and all the components in CF-AAN with the computation cost (GFLOPs) and performance on MARS. Except the baseline itself, all other computation costs are the increase comparing to the baseline method. C_B: the computation cost of the baseline method.

Method	w/ our DL	Self-attention Module				**#GFLOPs**	**MARS**	
		Self-attention	# of heads	Posi. Encoding	# of scales		mAP	R-1
Baseline	✘	✘	✘	✘	✘	24.520 (C_B)	83.4	87.7
	✔	✘	✘	✘	✘		85.1	89.7
Non-local	✔	3D self-attention	1	✘	1	C_B+17.213	86.2	**91.4**
Axial-based	✔	Axial-attention	1	✘	1	C_B+0.361	86.0	91.1
	✔	Axial-attention	8	✘	1	C_B+0.361	86.2	91.2
	✔	Axial-attention	8	Sinusoidal	1	C_B+0.377	86.0	91.1
	✔	Axial-attention	8	Relative	1	C_B+0.424	86.4	91.2
	✔	Axial-attention	8	Relative	2	C_B+0.245	86.4	91.3
	✔	**Axial-attention**	**8**	**Relative**	**4**	C_B**+0.126**	**86.5**	91.3

Our proposed CF-AAN

Compare to SOTA

Table 3: **Comparison with state-of-the-arts (%).** The score with underline is the runner-up.

Method	MARS		DukeV	
	mAP	R-1	mAP	R-1
DRSA (CVPR18)[21]	65.8	82.3	-	-
EUG (CVPR18)[42]	67.4	80.8	78.3	83.6
DuATM (CVPR18)[34]	67.7	81.2	-	-
TKP (ICCV19)[21]	73.3	84.0	91.7	94.0
M3D (AAAI19)[20]	74.1	84.4	-	-
Snippet (CVPR18)[5]	76.1	86.3	-	-
STA (AAAI19)[9]	80.8	86.3	94.9	96.2
VRSTC (CVPR19)[15]	82.3	88.5	93.5	95.0
NVAN (BMVC19)[25]	82.8	90.0	94.9	96.3
FT-WFT (AAAI20)[30]	82.9	88.6	-	-
TCLNet (ECCV20)[14]	85.1	89.8	<u>96.2</u>	**96.9**
AP3D (ECCV20)[10]	85.1	<u>90.1</u>	95.6	96.3
MG-RAFA (CVPR20)[44]	<u>85.9</u>	88.8	-	-
DL+CF-AAN (**Ours**)	**86.5**	**91.3**	**96.2**	<u>96.7</u>

Rectify some errors in MARS testing set

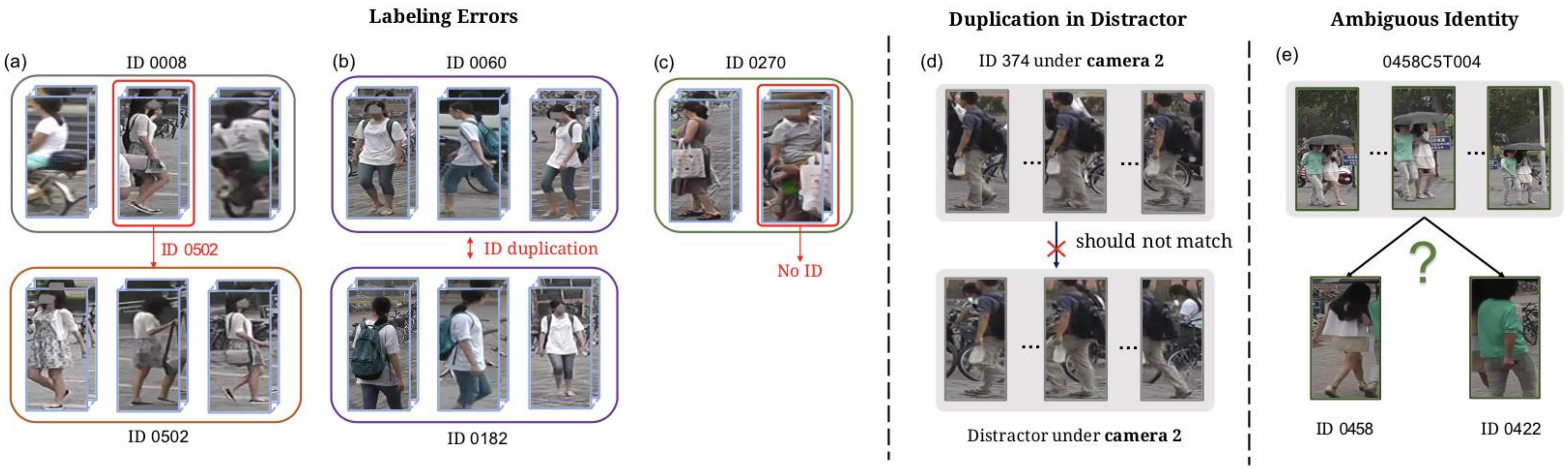


Figure 6: Three kinds of label noises in the MARS testing data.

Conclusion

1. re-Detect and Link module can easily align the noisy input tracklet.
2. With axial-attention, our CF-AAN achieves the state-of-the-arts.
3. We hope the **release of corrected data** can encourage the community for the further development of *invariant representation* on view, pose, illumination, and other variations without the hassle of the spatial and temporal alignment and dataset noise.

link : https://github.com/jackie840129/CF-AAN