A Instantiations Architecture

In this work, our TEAM is instantiated with ResNet architecture to formulate TEAM-Net. I-frame pathway utilizes ResNet-50 as the backbone while MV and residual use ResNet-18 as the backbone. Details of the TEAM-Net architecture and feature dimension at each stage can be referred in Figure 1 and Table 2. As mentioned in the main paper, the input is sampled based on the GOP. In Figure 1, T GOPs are sampled from a video, which results in T I-frames and T P-frames (MVs and residuals) to be fed to the network.

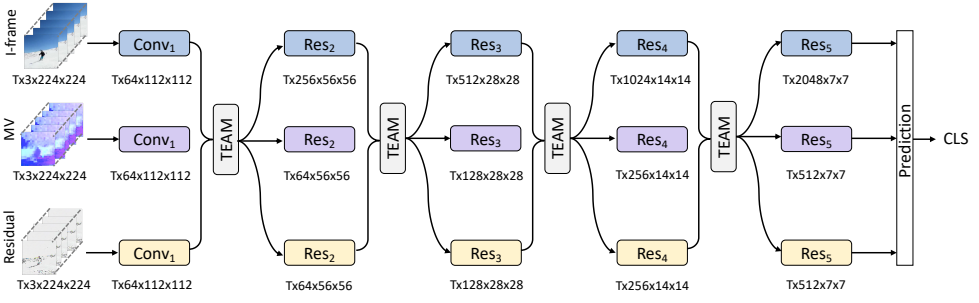


Figure 1: An example instantiation of the TEAM-Net. The dimensions of features at each stage are denoted as $\{T \times C \times H \times W\}$. The I-frame path ultilizes ResNet-50 while MV and residual paths use ResNet-18.

B Instantiated with TSM

Table 1 shows the TEAM-Net using TSM as the backbone. It can be seen that TEAM is able to improve the TSM baseline consistently. Together with TSN experiment demonstrated in the main paper, these results show that TEAM-Net is well-generalized to different backbones and datasets.

Table 1: TEAM-Net consistently improves the TSM baseline for three datasets. All methods use 8 frames for fair comparison.

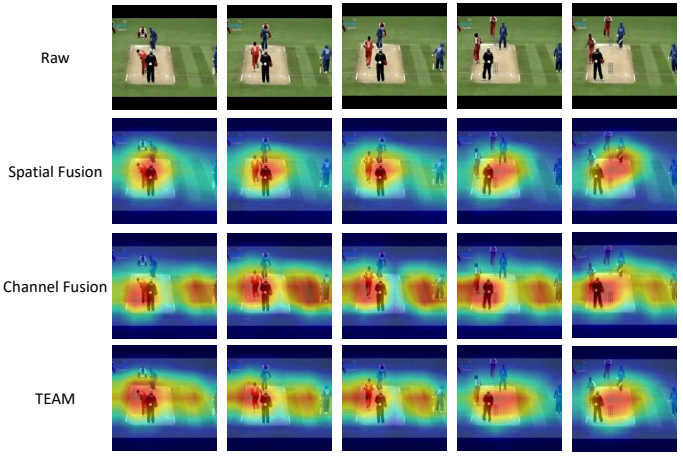
Model	Kinetics-400	UCF-101	HMDB-51
TSM	71.8	94.5	70.7
TEAM-Net	**73.0** (↑1.2)	**95.2** (↑0.7)	**74.3** (↑3.0)

Table 2: TEAM-Net is instantiated with the ResNet architecture. I-frame pathway utilizes ResNet-50 as the backbone while MV and residual (R) both utilize ResNet-18 as the backbone.

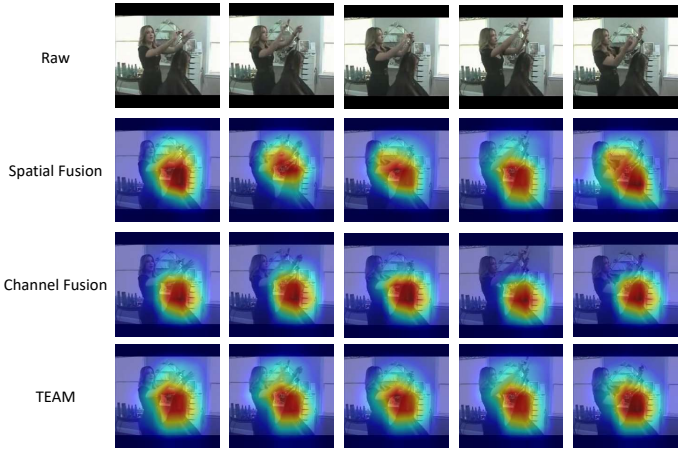
Stage	I-frame	MV/Residual	Output size
Input	——		I-frame: $T \times 224 \times 224$ MV/R: $T \times 224 \times 224$
conv_1	$1 \times 7 \times 7$, 64, stride 1, 2, 2		I-frame: $T \times 112 \times 112$ MV/R: $T \times 112 \times 112$
pool_1	$1 \times 3 \times 3$, max, stride 1, 2, 2		I-frame: $T \times 56 \times 56$ MV/R: $T \times 56 \times 56$
TEAM			I-frame: $T \times 56 \times 56$ MV/R: $T \times 56 \times 56$
res_2	$\begin{bmatrix} 1\times1\times1,64 \\ 1\times3\times3,64 \\ 1\times1\times1,256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times3\times3,64 \\ 1\times3\times3,64 \end{bmatrix} \times 2$	I-frame: $T \times 56 \times 56$ MV/R: $T \times 56 \times 56$
TEAM			I-frame: $T \times 56 \times 56$ MV/R: $T \times 56 \times 56$
res_3	$\begin{bmatrix} 1\times1\times1,128 \\ 1\times3\times3,128 \\ 1\times1\times1,512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times3\times3,128 \\ 1\times3\times3,128 \end{bmatrix} \times 2$	I-frame: $T \times 28 \times 28$ MV/R: $T \times 28 \times 28$
TEAM			I-frame: $T \times 28 \times 28$ MV/R: $T \times 28 \times 28$
res_4	$\begin{bmatrix} 1\times1\times1,256 \\ 1\times3\times3,256 \\ 1\times1\times1,512 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times3\times3,256 \\ 1\times3\times3,256 \end{bmatrix} \times 2$	I-frame: $T \times 14 \times 14$ MV/R: $T \times 14 \times 14$
TEAM			I-frame: $T \times 14 \times 14$ MV/R: $T \times 28 \times 28$
res_5	$\begin{bmatrix} 1\times1\times1,512 \\ 1\times3\times3,512 \\ 1\times1\times1,2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times3\times3,512 \\ 1\times3\times3,512 \end{bmatrix} \times 2$	I-frame: $T \times 7 \times 7$ MV/R: $T \times 7 \times 7$
global average pool, FC			$T \times CLS$
temporal average			CLS

C Class-specific Visualization

As mentioned in the main paper, the channel fusion focuses on *'what'* by taking global spatial information into account while the spatial fusion focuses on *'where'* by looking at the local spatial content. Figure 2 shows that the channel fusion fails in local action reasoning. Figure 3 shows that the spatial fusion is sensitive to unrelated local movements.

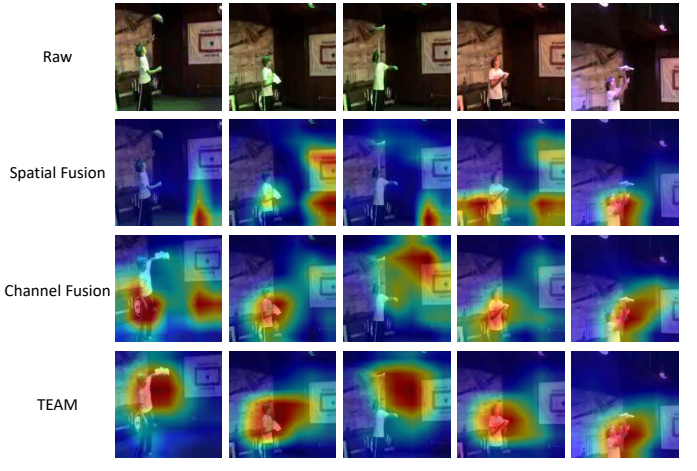


(a) `Cricket Shot`

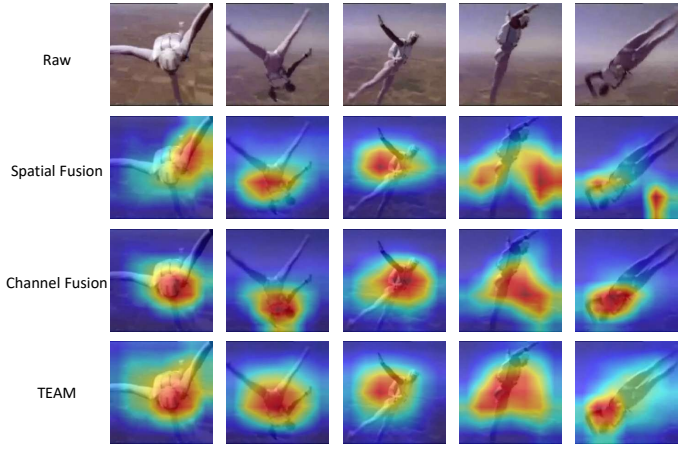


(b) `Hair Cut`

Figure 2: Class-specific visualization for channel fusion *only*, spatial fusion *only* and our proposed TEAM module. *The channel fusion only reasoning is inaccurate in this case.*



(a) Pizza Tossing



(b) Sky Diving

Figure 3: Class-specific visualization for channel fusion *only*, spatial fusion *only* and our proposed TEAM module. *The spatial fusion only is sensitive to unrelated movements.*