STSP: Spatial-Temporal Subspace Projection for Video Class-incremental Learning (Supplementary Material)

Hao Cheng[1,2], Siyuan Yang[1], Chong Wang[1], Joey Tianyi Zhou[3,4], Alex C. Kot[1], and Bihan Wen[1]

[1] Nanyang Technological University, Singapore
{siyuan.yang,wang1711,eackot,bihan.wen}@ntu.edu.sg
[2] Hebei University of Technology, Tianjin, China
2024025@hebut.edu.cn
[3] CFAR, Agency for Science, Technology and Research, Singapore
[4] IHPC, Agency for Science, Technology and Research, Singapore
zhouty@cfar.a-star.edu.sg

In this supplementary material, we first give the proof of Proposition 3 in Sec. A. Then, we provide an additional ablation study on loss functions in our proposed method in Sec. B.

A Proofs of Proposition 3

We first recall the classification metric used in our proposed TSC and Proposition 3 as follows, and then give the proof.

In TSC, for j-th class, we construct a subspace with orthogonal basis $P_j \in \mathbb{R}^{C\times\tau}$ with $P_j^\top P_j = I_\tau, \tau \ll C$, where C and τ indicates the feature dimension and the number of subspace basis for j-th class, respectively. We then adopt current subspaces to predict the class of Z_i based on the ***subspace distance*** $d_j(Z_i)$ between the subspace spanned by temporal video features Z_i and the orthogonal basis of each class as:

$$d_j(Z_i) = -\left\| Z_i Z_i^\top - P_j P_j^\top \right\|_F^2 . \tag{S1}$$

Proposition 3. *Let $f^{TSC}(\,\cdot\,;\mathbf{w}_t)$ denotes the TSC of the network trained on current task $\mathcal{T}_t$ and its corresponding input feature Z_k from any previous task $\mathcal{T}_k$ $(k<t)$. During each training step s, if the TSC parameter update $\Delta\mathbf{w}_{t,s}^{TSC}$ lies in the null space of the subspace spanned by Z_k,* i.e.*,*

$$Z_k Z_k^\top \Delta\mathbf{w}_{t,s}^{TSC} = 0, \tag{S2}$$

similarly we have $f^{TSC}\left(Z_k;\mathbf{w}_t^{TSC}\right) = f^{TSC}\left(Z_k;\mathbf{w}_k^{TSC}\right)$.

Proof. The class weights in TSC contain subspace basis matrices for each class j, i.e.*, $\mathbf{w}_k^{TSC} = \{P_j^k\}_{j=1}^J$, where k and J indicate the current session number and the total number of classes in the current session, respectively.*

The j-th term of the TSC output $f^{TSC}\left(Z_k;\mathbf{w}_k^{TSC}\right)$ for the input Z_k in the k-th session is:

$$\begin{aligned} d_j\left(Z_k;\mathbf{w}_k^{TSC}\right) &= \left\|Z_kZ_k^\top - P_j^kP_j^{k\top}\right\|_F^2 \\ &= tr(Z_k{Z_k}^\top Z_k{Z_k}^\top) + tr(P_j^k{P_j^k}^\top P_j^k{P_j^k}^\top) - 2tr(Z_k{Z_k}^\top P_j^k{P_j^k}^\top) \\ &= tr(Z_k{Z_k}^\top Z_k{Z_k}^\top) + \|I_\tau\|_F^2 - 2tr(Z_k{Z_k}^\top P_j^k{P_j^k}^\top) \\ &= \mathcal{S} - 2tr(Z_k{Z_k}^\top P_j^k{P_j^k}^\top), \end{aligned} \tag{S3}$$

where $\mathcal{S} = tr(Z_k{Z_k}^\top Z_k{Z_k}^\top) + \|I_\tau\|_F^2$ is solely dependent on the input features and remains unaffected by the TSC weight.

After the k-th session, we have the TSC parameters with $\mathbf{w}_k^{TSC}$ as the initialization for the next session $k+1$, i.e., *$\mathbf{w}_{k+1,0}^{TSC} = \mathbf{w}_k^{TSC}$, with $P_{j,0}^{k+1} = P_j^k$. By denoting the learning rate at the training step s for the task $\mathcal{T}_{k+1}$ as α, we have:*

$$\mathbf{w}_{k+1,s}^{TSC} = \mathbf{w}_{k+1,s-1}^{TSC} - \alpha\Delta\mathbf{w}_{k+1,s-1}^{TSC}, \tag{S4}$$

with each term $P_{j,s}^{k+1}$:

$$P_{j,s}^{k+1} = P_{j,s-1}^{k+1} - \alpha\Delta P_{j,s-1}^{k+1}. \tag{S5}$$

Then the TSC output of each class j for the input Z_k of the task $\mathcal{T}_k$ at the training step s in the $(k+1)$-th session can be computed by:

$$d_j\left(Z_k;\mathbf{w}_{k+1,s}^{TSC}\right) = \left\|Z_kZ_k^\top - P_{j,s}^{k+1}P_{j,s}^{k+1\top}\right\|_F^2 = \mathcal{S} - 2tr(Z_k{Z_k}^\top P_{j,s}^{k+1}P_{j,s}^{k+1\top}). \tag{S6}$$

By substituting Eq. (S5) into Eq. (S6), we can obtain:

$$\begin{aligned} d_j\left(Z_k;\mathbf{w}_{k+1,s}^{TSC}\right) &= \mathcal{S} - 2tr(Z_k{Z_k}^\top(P_{j,s-1}^{k+1} - \alpha\Delta P_{j,s-1}^{k+1})(P_{j,s-1}^{k+1\top} - \alpha\Delta P_{j,s-1}^{k+1\top})) \\ &= \mathcal{S} - 2tr(Z_k{Z_k}^\top P_{j,s-1}^{k+1}P_{j,s-1}^{k+1\top}) + 4\alpha tr(Z_k{Z_k}^\top\Delta P_{j,s-1}^{k+1}P_{j,s-1}^{k+1\top}) \\ &\quad - 2\alpha^2 tr(Z_k{Z_k}^\top\Delta P_{j,s-1}^{k+1}\Delta P_{j,s-1}^{k+1\top}). \end{aligned} \tag{S7}$$

Based on Eq. (S2), the last two terms in Eq. (S7) are equal to 0. Hence, we can prove that:

$$d_j\left(Z_k;\mathbf{w}_{k+1,s}^{TSC}\right) = d_j\left(Z_k;\mathbf{w}_{k+1,0}^{TSC}\right) = d_j\left(Z_k;\mathbf{w}_k^{TSC}\right). \tag{S8}$$

Thus, we can extend Eq. (S8) to each novel task $\mathcal{T}_t$ $(t > k)$ as:

$$d_j\left(Z_k;\mathbf{w}_t^{TSC}\right) = d_j\left(Z_k;\mathbf{w}_k^{TSC}\right), \tag{S9}$$

which is equivalent to $f^{TSC}\left(Z_k;\mathbf{w}_t^{TSC}\right) = f^{TSC}\left(Z_k;\mathbf{w}_k^{TSC}\right)$.

B Ablation Study on Loss Functions

We conduct ablation studies on the UCF101 [1] dataset to assess the impact of individual loss components within our proposed method, under both 2×25

Table 1: Ablation studies on loss functions for our proposed STSP method on UCF101 with 5×10 stages and 2×25 stages settings, respectively. Here, L_{OF} and L_{OW} represent the orthogonal loss on video representations and the weights of TSC, respectively.

L_{OF}	L_{OW}	UCF101	
		5×10 stages	2×25 stages
✗	✗	74.73	72.87
✗	✓	76.63	73.75
✓	✗	80.62	74.54
✓	✓	**82.84**	**79.25**

stages and 5×10 stages settings. The results, as detailed in Tab. 1, reveal that each loss component positively affects the model's performance. Notably, the L_{OF} loss term enhances performance by 5.89% at the 10-stage setting, highlighting its role in promoting distinct information capture by individual video frames through orthogonal constraints. This process effectively constructs an appropriate subspace, facilitating subspace-based classification. Additionally, results indicate that the distinct loss components serve complementary functions, further validating our method's capability in effectively modeling the data and classifiers via subspace learning.

References

1. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 2