Appendix

This appendix provides further details as referenced in the main paper: Section A contains detailed description of proposed STS conv. Section B contains further results ablations on Kinetics-400.

A Formula of STS Conv

We give the formal definition of STS convolution as following. Given the input $\mathbf{x} \in \mathbb{R}^{C\times T\times H\times W}$ and a 3D Conv with weights $\theta \in \mathbb{R}^{C_{in}\times C_{out}\times K_t\times K_h\times K_w}$ (for simplicity, $C_{in} = C_{out} = C, K_t = 3$), We first decomposes the θ along the channel dimension into two groups: $(\alpha, \beta) \in \mathbb{R}^{C\times C_{1/2}\times 3\times K_h\times K_w}$. α is for the static appearance modeling so we can split it along temporal dimension into $(\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^{C\times C_{1/2}\times K_h\times K_w}$. β is for dynamic motion modeling. To preserve α_1's appearance modeling ability, we initialize the α_0 and α_2 with zeros. Then we aim at leveraging the *untouched* α_0 and α_2 to enlarge spatial receptive field. Specifically, we reshape each frame x_t into x_t^{row} with size of $(C, W\times H)$ and $x_t^{col.}$ with size of $(C, H\times W)$. Similarly, α_0 and α_2 should be reshaped. Finally, we gather the feature

$$\mathbf{y}_t = \boldsymbol{concat}(\underbrace{Conv1D(x_t^{row};\alpha_0) + Conv2D(x_t;\alpha_1) + Conv1D(x_t^{col.};\alpha_2)}_{Static} \quad (1)$$

$$, \underbrace{Conv3D(\mathbf{x};\beta)}_{Dynamic}). \quad (2)$$

B Additional Ablation Study

B.1 Case Study of Slowfast

We believe that a proper initialization method and training schedule are the two keys to boosting 3D CNNs' performance. First, we pre-train the two branches together while SlowFast only initializes the slow branch due to its structural changes. As shown in Table 1, pre-training both branches with STS improves SlowFast pipeline by **0.3%** on the same amount of budget. Second, further increasing the pre-training budget to 300 epochs readily outperforms the from-scratch result by **1.3%** with only $\times$**0.8** computation.

Model	Pre-train Branch	Pre-train	Fine-tune	Total Budgets	K400
SlowFast (from scratch)	-	-	256	$\times$1	75.6
SlowFast (previous pipeline)	slow	90	100	$\times$0.5	75.4
STS-SlowFast (our pipeline)	slow+fast	90	100	$\times$**0.5**	**75.7**
STS-SlowFast (our pipeline)	slow+fast	300	100	$\times$**0.8**	**76.9**

Table 1: Investigation of pre-training in SlowFast 4×16 .

B.2 Dilated Conv v.s. STS Conv

During fine-tuning, reshaping the *untouched* kernels in spatial space can enlarge the receptive field to boost performance. Two reshaped 1D Convs can obtain larger receptive filed than two same-directional dilated Convs. We ablate dilated Conv and have two observations. 1) Reshaping 1D Conv achieves better results than dilated Conv on SSV2 (61.4% *vs*. 61.1%) and K400 (74.7% *vs*. 74.5%). 2) Dilated Conv is better than the baseline (61.1% *vs*. 60.4% on SSV2, 74.5% *vs*. 74.3% on K400), suggesting the effectiveness of enlarging receptive field in the static channel.

ResNet50-3x3x3	dilated rate	Effective Receptive field	K400	SS-V2
Baseline	-	3×3	74.3	60.4
w/ dilated conv	2	$3 \times 3 + 5 \times 5$	74.5	61.2
w/ dilated conv	3	$3 \times 3 + 7 \times 7$	74.4	61.1
w/ two orthogonal 1D convs	-	$1 \times 9 + 3 \times 3 + 9 \times 1$	**74.7**	**61.4**

Table 2: Dilated Conv v.s. STS Conv.