Supplementary Materials

Anonymous CVPR 2021 submission

Paper ID 11846

Abstract

This material includes additional experimental results and architecture details of the Semi-Supervised Knowledge Transfer (SSKT).

1. SSKT with Multiple Tasks and Problem Domains

Prerequisite Learning based Knowledge Transfer (SSKT) supports a training procedure that enables transfer learning in a variety of scenarios using deep neural networks. SSKT has a structure that transfers pretrained knoweldge naturally, without compromising the training information of the pretrained network or requiring additional supervision in the target task training process. We achieved this goal using the soft label-based knowledge transfer techniques with auxiliary task learning through self-supervision, for the various domain of image recognition variants. Final formulation of SSKT as follows:

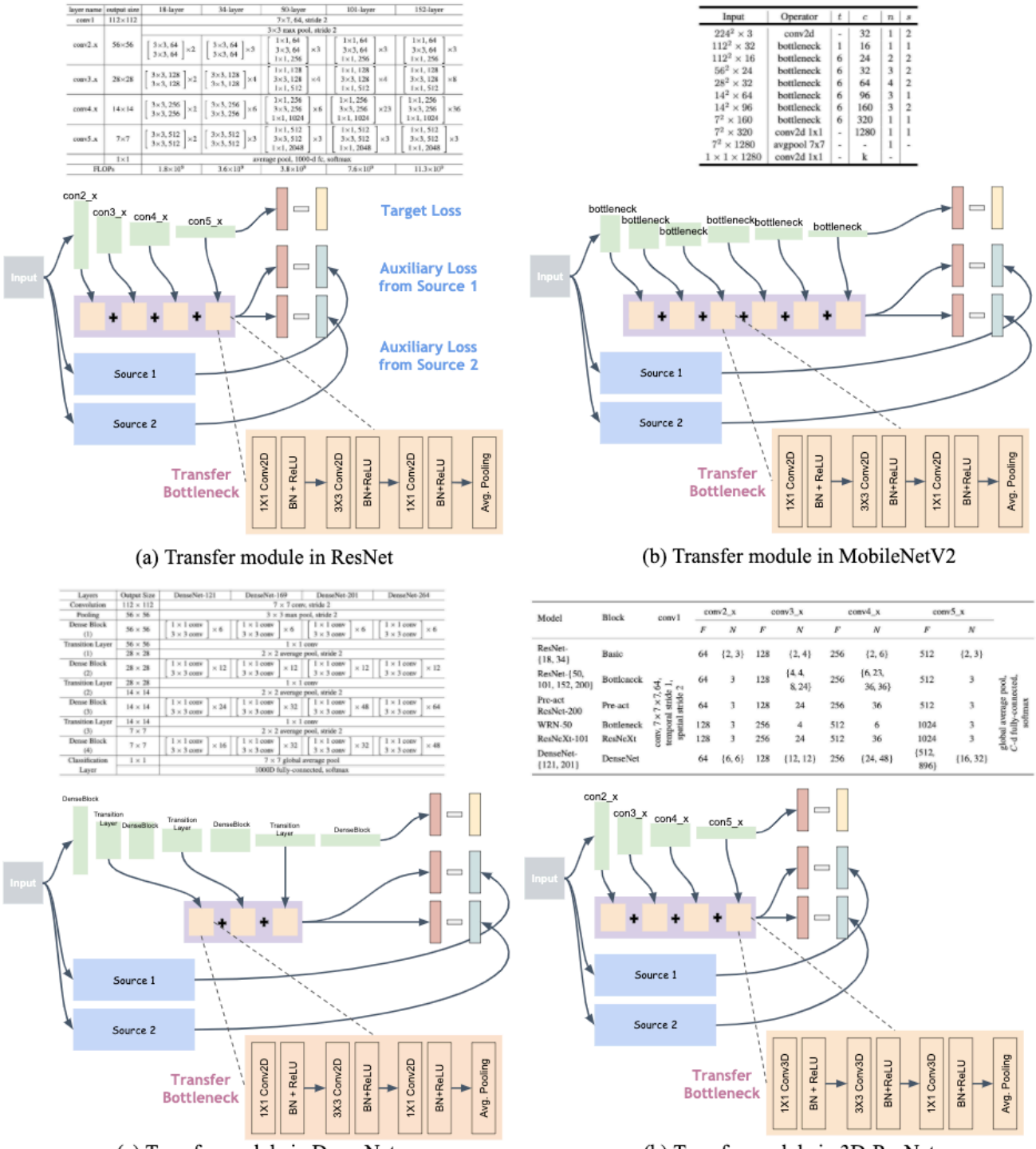
$$\begin{aligned}\underset{\theta_t}{\operatorname{argmin}} \Big(& L(h_t^{prim}(x_i;\theta_t.D_t,T_t),y_{t,i}^{prim}) \\ & +\alpha(L(h_t^{aux}(f_{s_1}(x_i);\theta_t.D_t,T_{s_1}),y_{s_1,i}^{aux}) \\ & +L(h_t^{aux}(f_{s_2}(x_i);\theta_t.D_t,T_{s_2}),y_{s_2,i}^{aux})+\dots \\ & +L(h_t^{aux}(f_{s_M}(x_i);\theta_t.D_t,T_{s_M}),y_{s_M,i}^{aux}))\Big). \end{aligned} \quad (1)$$

We define a multi-tast network $h_t(x;\theta_t,D_t,T_t)$, where x is the input, θ_t is a parameter of the target network, D_t is a target dataset, and T_t is the task to be trained. θ_t is updated simultaneously through target loss and auxiliary loss during training to solve the primary task. $h_s(x;\theta_s,D_s,T_s)$ describes a source network that receives the input x and delivers knowledge to the target network. θ_s denotes a parameter trained by the source task T_s for the source data set D_s. θ_s is not updated during the target task training. i is the i^{th} batch of the training data, α is balanced parameter for total loss, and $y_{s,i}^{aux}=h_s(x_i;\theta_s,D_s,T_s)$ is the softmax output from the pretrained source network and conveys the dark knowledge of the pretrained dataset by soft labels. The data transformation function f_s converts the data type to match the source task to infer the recognition information to the task of the source domain. For example, if T_t is an action recognition problem using 3D-CNN, the input $x^{w\times h\times d}\in D_t$ is defined as a three-dimensional tensor. In this case, if a pretrained network for transfer learning is obtained through the image recognition problem T_s using 2D-CNN, $f_s: x^{w\times h\times d}\rightarrow \hat{x}^{w\times h}$ should be defined as a function that maps a three-dimensional tensor to a two-dimensional matrix into which h_s can be input. Up to M number of different type of transformation functions could be defined.

Transfer Modules Depending on CNN Architecture. To encourage predicting $y_{s,i}^{aux}$ by h_t, we design bottleneck structure based transfer module supporting auxiliary task using feature output from each convolutional block. Figure 1 shows configuration of transfer modules depending on each CNN architecture for its problem domain. We applied the transfer module to four different CNN architectures such as ResNet (He et al. 2016), DenseNet (Huang et al. 2017), MobileNet (Sandler et al. 2018), and 3D-ResNet (Hara, Kataoka, and Satoh 2018) for each problem domain.

2. Experiments Results

We provide performance for all experimental conditions for each dataset, in addition to the results contained in the paper. For fair comparison of SSKT, the experimental conditions consist of a combination of the type of source and target network, the presence or absence of a transfer module, and a loss function. For model architecture and hyperparameters configuration for training (See Table 1 of the paper). Same as the experiment results of the paper, the datasets of the source task are ImageNet (Deng et al. 2009) and Places365 (Zhou et al. 2018), and the datasets of the target task are CIFAR10/100 (Krizhevsky 2009), STL10 (Coates, Lee, and Ng 2011), ImageNet, Places365, PASCAL VOC (Everingham et al. 2015), UCF101 (Soomro, Zamir, and Shah 2012), and HMDB51 (Kuehne et al. 2011). Tables 1 to 6 provide performance according to the experimental conditions of each dataset. At the same time, we



(a) Transfer module in ResNet

(b) Transfer module in MobileNetV2

(a) Transfer module in DenseNet

(b) Transfer module in 3D-ResNet

Figure 1. **Schematic of the transfer modules for efficient SSKT with different CNN architectures.** The transfer module used in the SSKT consists of summation of feature output of bottleneck layers from each convolutional block. Schematic shows and example of the transfer module with different CNN architectures for SSKT using multiple sources.

included all experimental results for hyperparameter optimization. Figure 2 shows the performance changes of the hyperparameters for loss in the STL10 and PASCAL VOC datasets, the structure of the source and target network, and

Table 1. SSKT-based transfer learning performance change for CIFAR10 dataset compared to the training from scratch. All experiments evaluated test performance 3 times from the same random seed for the model. TM stands for Transfer Module and R[depth] stands for ResNet structure. The best performance of each network architecture highlighted in **bold**.

T_s	T_t	Model	Method	TM	Loss	acc.
-	C10	R20	scratch	-	CE	92.19±0.09
P		R20	SSKT	x	CE+CE	92.21±0.06
		R20	SSKT	x	CE+KD	92.24±0.14
		R20	SSKT	o	CE+CE	92.23±0.04
		R20	SSKT	o	CE+KD	92.25±0.04
I		R20	SSKT	x	CE+CE	92.28±0.07
		R20	SSKT	x	CE+KD	92.34±0.07
		R20	SSKT	o	CE+CE	92.44±0.05
		R20	SSKT	o	CE+KD	92.29±0.0
P+I		R20	SSKT	x	CE+CE	91.9±0.1
		R20	SSKT	x	CE+KD	**92.46±0.15**
		R20	SSKT	o	CE+CE	92.42±0.07
		R20	SSKT	o	CE+KD	92.22±0.17
-		R32	scratch	-	CE	93.21±0.09
P		R32	SSKT	x	CE+CE	92.77±0.14
		R32	SSKT	x	CE+KD	92.87±0.31
		R32	SSKT	o	CE+CE	92.65±0.26
		R32	SSKT	o	CE+KD	92.59±0.22
I		R32	SSKT	x	CE+CE	93.26±0.08
		R32	SSKT	x	CE+KD	92.78±0.2
		R32	SSKT	o	CE+CE	93.25±0.12
		R32	SSKT	o	CE+KD	92.88±0.07
P+I		R32	SSKT	x	CE+CE	92.88±0.15
		R32	SSKT	x	CE+KD	93.07±0.09
		R32	SSKT	o	CE+CE	**93.38±0.02**
		R32	SSKT	o	CE+KD	93.1±0.22

the presence or absence of a transfer module. The abbreviations for the datasets and model architectures listed in all experimental tables are as follows:

Datasets: ImageNet (I), Places365 (P), CIFAR10 (C10), CIFAR100 (C100), STL10 (S10), PASCAL VOC (VOC), UCF101 (U101), and HMDB51 (H51).

Model architectures: ResNet (R), DenseNet (D), MobileNetV2 (MV2), and 3D-ResNet (3DR).

Finally, we included the results according to the experimental configuration for further experiments for SSKT analysis. Table 7 shows the experimental results for DenseNet121 and MoblieNetV2, and Table 8 shows the evaluation results for each experimental setting with fine-tuning scenario.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *In Proc. of ICML*, 2020. 3

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models

Table 2. SSKT-based transfer learning performance change for CIFAR100 dataset compared to the training from scratch.

T_s	T_t	Model	Method	TM	Loss	acc.
-	C100	R20	scratch	-	CE	68.26±0.36
P		R20	SSKT	x	CE+CE	67.65±0.21
		R20	SSKT	x	CE+KD	68.01±0.42
		R20	SSKT	o	CE+CE	67.96±0.27
		R20	SSKT	o	CE+KD	67.83±0.34
I		R20	SSKT	x	CE+CE	68.3±0.17
		R20	SSKT	x	CE+KD	68.37±0.23
		R20	SSKT	o	CE+CE	**68.63±0.12**
		R20	SSKT	o	CE+KD	68.35±0.1
P+I		R20	SSKT	x	CE+CE	67.87±0.17
		R20	SSKT	x	CE+KD	68.13±0.05
		R20	SSKT	o	CE+CE	68.56±0.23
		R20	SSKT	o	CE+KD	67.84±0.28
-		R32	scratch	-	CE	70.33±0.19
P		R32	SSKT	x	CE+CE	69.97±0.16
		R32	SSKT	x	CE+KD	69.93±0.21
		R32	SSKT	o	CE+CE	69.69±0.19
		R32	SSKT	o	CE+KD	69.92±0.31
I		R32	SSKT	x	CE+CE	70.6±0.05
		R32	SSKT	x	CE+KD	70.17±0.14
		R32	SSKT	o	CE+CE	70.75±0.06
		R32	SSKT	o	CE+KD	70.0±0.11
P+I		R32	SSKT	x	CE+CE	69.25±0.58
		R32	SSKT	x	CE+KD	69.22±0.43
		R32	SSKT	o	CE+CE	**70.94±0.36**
		R32	SSKT	o	CE+KD	69.44±0.01

Table 3. SSKT-based transfer learning performance change for STL10 dataset compared to the training from scratch.

T_s	T_t	Model	Method	TM	Loss	acc.
-	STL10	R20	scratch	-	CE	81.15±0.34
P		R20	SSKT	x	CE+CE	81.56±0.32
		R20	SSKT	x	CE+KD	80.88±0.19
		R20	SSKT	o	CE+CE	82.76±0.05
		R20	SSKT	o	CE+KD	81.06±0.2
I		R20	SSKT	x	CE+CE	82.2±0.17
		R20	SSKT	x	CE+KD	80.82±0.14
		R20	SSKT	o	CE+CE	83.45±0.07
		R20	SSKT	o	CE+KD	81.3±0.39
P+I		R20	SSKT	x	CE+CE	82.46±0.24
		R20	SSKT	x	CE+KD	81.47±0.22
		R20	SSKT	o	CE+CE	**84.56±0.35**
		R20	SSKT	o	CE+KD	81.33±0.11
-		R32	scratch	-	CE	81.19±0.17
P		R32	SSKT	x	CE+CE	82.1±0.14
		R32	SSKT	x	CE+KD	81.29±0.22
		R32	SSKT	o	CE+CE	83.06±0.27
		R32	SSKT	o	CE+KD	81.19±0.12
I		R32	SSKT	x	CE+CE	82.88±0.33
		R32	SSKT	x	CE+KD	81.4±0.23
		R32	SSKT	o	CE+CE	83.68±0.28
		R32	SSKT	o	CE+KD	81.76±0.18
P+I		R32	SSKT	x	CE+CE	82.39±0.15
		R32	SSKT	x	CE+KD	79.8±0.47
		R32	SSKT	o	CE+CE	**83.4±0.2**
		R32	SSKT	o	CE+KD	80.05±1.06

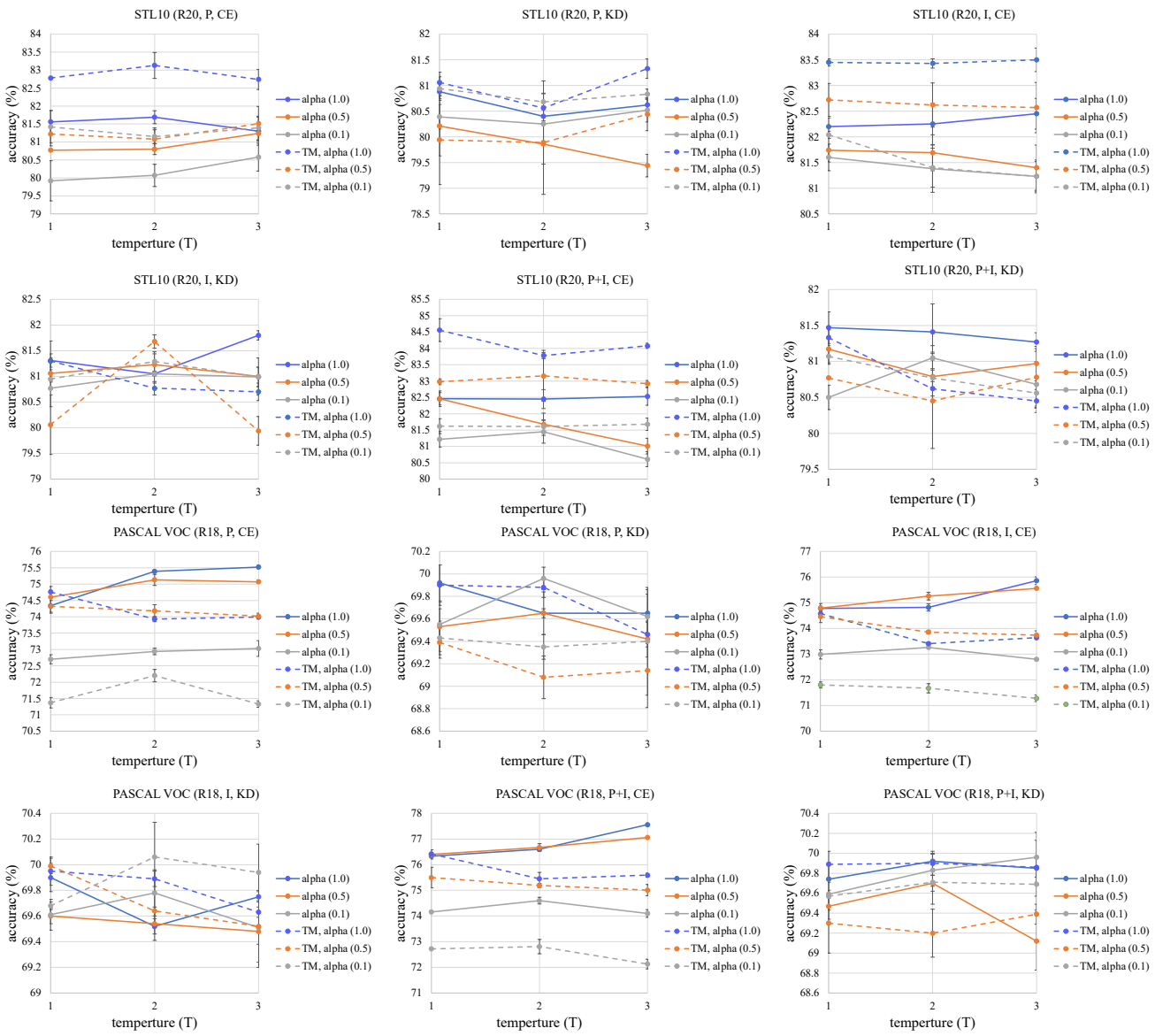


Figure 2. **Graph visualization for parameter optimization of SSKT.** The title of each graph is composed of D_t (target model, T_s, auxiliary loss). T is the temparture parameter of each auxiliary loss, and α is the balance parameter of the total loss.

are strong semi-supervised learners. *In Proc. of NeurIPS*, 2020. 3

[3] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. *In Proc. of CVPR*, 2020. 2

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. *In Proc. of AISTAT*, 2011. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *In Proc. of CVPR*, 2009. 5

[6] Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *In Proc. of NeurIPSW*, 2019. 3, 8

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge - a retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 6

[8] Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *In Proc. of ICML*, 2018. 2

[9] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *In Proc. of ICCV*, 2019. 3

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can

Table 4. SSKT-based transfer learning performance change for ImageNet and Places365 compared to the training from scratch.

T_s	T_t	Model	Method	TM	Loss	acc.
-	P	R18	scratch	-	CE	50.92
P		R18	SSKT	x	CE+CE	54.41
		R18	SSKT	x	CE+KD	53.42
		R18	SSKT	o	CE+CE	54.5
		R18	SSKT	o	CE+KD	54.11
I		R18	SSKT	x	CE+CE	53.47
		R18	SSKT	x	CE+KD	53.51
		R18	SSKT	o	CE+CE	53.67
		R18	SSKT	o	CE+KD	53.44
P+I		R18	SSKT	x	CE+CE	**54.78**
		R18	SSKT	x	CE+KD	54.5
		R18	SSKT	o	CE+CE	54.62
		R18	SSKT	o	CE+KD	54.5
-	I	R18	scratch	-	CE	64.14
P		R18	SSKT	x	CE+CE	64.18
		R18	SSKT	x	CE+KD	64.21
		R18	SSKT	o	CE+CE	64.99
		R18	SSKT	o	CE+KD	63.53
I		R18	SSKT	x	CE+CE	67.79
		R18	SSKT	x	CE+KD	66.0
		R18	SSKT	o	CE+CE	67.46
		R18	SSKT	o	CE+KD	65.65
P+I		R18	SSKT	x	CE+CE	**70.57**
		R18	SSKT	x	CE+KD	67.42
		R18	SSKT	o	CE+CE	67.64
		R18	SSKT	o	CE+KD	66.81

Table 5. SSKT-based transfer learning performance change for PASCAL VOC compared to the training from scratch.

T_s	T_t	Model	Method	TM	Loss	acc.
-	VOC	R18	scratch	-	BCE	67.28±0.25
P		R18	SSKT	x	BCE+CE	74.34±0.23
		R18	SSKT	x	BCE+KD	69.92±0.16
		R18	SSKT	o	BCE+CE	74.76±0.17
		R18	SSKT	o	BCE+KD	69.9±0.18
I		R18	SSKT	x	BCE+CE	74.78±0.09
		R18	SSKT	x	BCE+KD	69.9±0.35
		R18	SSKT	o	BCE+CE	74.58±0.11
		R18	SSKT	o	BCE+KD	69.95±0.19
P+I		R18	SSKT	x	BCE+CE	76.33±0.0
		R18	SSKT	x	BCE+KD	69.74±0.17
		R18	SSKT	o	BCE+CE	**76.42±0.06**
		R18	SSKT	o	BCE+KD	69.89±0.13
-		R34	scratch	-	BCE	66.0±0.49
P		R34	SSKT	x	BCE+CE	73.83±0.38
		R34	SSKT	x	BCE+KD	69.93±0.03
		R34	SSKT	o	BCE+CE	75.65±0.12
		R34	SSKT	o	BCE+KD	69.51±0.13
I		R34	SSKT	x	BCE+CE	74.25±0.12
		R34	SSKT	x	BCE+KD	70.05±0.14
		R34	SSKT	o	BCE+CE	75.14±0.14
		R34	SSKT	o	BCE+KD	70.18±0.11
P+I		R34	SSKT	x	BCE+CE	75.88±0.1
		R34	SSKT	x	BCE+KD	70.15±0.09
		R34	SSKT	o	BCE+CE	**77.02±0.02**
		R34	SSKT	o	BCE+KD	70.58±0.35
-		R50	scratch	-	BCE	61.16±0.34
P		R50	SSKT	x	BCE+CE	63.29±1.43
		R50	SSKT	x	BCE+KD	65.5±0.2
		R50	SSKT	o	BCE+CE	74.44±0.06
		R50	SSKT	o	BCE+KD	65.94±0.09
I		R50	SSKT	x	BCE+CE	63.96±2.74
		R50	SSKT	x	BCE+KD	66.11±0.32
		R50	SSKT	o	BCE+CE	74.24±0.05
		R50	SSKT	o	BCE+KD	65.77±0.13
P+I		R50	SSKT	x	BCE+CE	69.27±0.21
		R50	SSKT	x	BCE+KD	66.0±0.29
		R50	SSKT	o	BCE+CE	**77.1±0.14**
		R50	SSKT	o	BCE+KD	66.22±0.23

spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *In Proc. of CVPR*, 2018. 7

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *In Proc. of CVPR*, 2020. 3

[12] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *In Proc. of ICCV*, 2019. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In Proc. of CVPR*, 2016. 2

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *In Proc. of NIPSW*, 2014. 1, 2, 4, 8

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *In Proc. of CVPR*, 2017. 7

[16] Inseop Chung Nojun Kwak Jangho Kim, Minsung Hyun. Feature fusion for online mutual knowledge distillation. *In Proc. of ICPR*, 2020. 2

[17] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetful learning for domain expansion in deep neural networks. *In Proc. of AAAI*, 2018. 3

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR abs/1705.06950*, 2017. 7

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 6

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *In Proc. of NIPS*, 2012. 2

[21] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso A. Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. *In Proc. of ICCV*, 2011. 7

[22] Zhizhong Li and Derek Hoiem. Learning without forgetting. *In Proc. of ECCV*, 2016. 3

[23] Shikun Liu, Andrew J. Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. *In Proc. of NeurIPS*, 2019. 3

Table 6. SSKT-based transfer learning performance change for UCF101 and HMDB51 compared to the training from scratch.

T_s	T_t	Model	Method	TM	Loss	acc.
-	U101	3DR18	scratch	-	CE	43.28
P		3DR18	SSKT	x	CE+CE	44.1
		3DR18	SSKT	x	CE+KD	44.79
		3DR18	SSKT	o	CE+CE	45.35
		3DR18	SSKT	o	CE+KD	43.95
I		3DR18	SSKT	x	CE+CE	46.62
		3DR18	SSKT	x	CE+KD	40.35
		3DR18	SSKT	o	CE+CE	44.26
		3DR18	SSKT	o	CE+KD	38.95
P+I		3DR18	SSKT	x	CE+CE	**52.19**
		3DR18	SSKT	x	CE+KD	43.68
		3DR18	SSKT	o	CE+CE	47.09
		3DR18	SSKT	o	CE+KD	45.0
-	H51	3DR18	scratch	-	CE	17.14
P		3DR18	SSKT	x	CE+CE	18.18
		3DR18	SSKT	x	CE+KD	17.33
		3DR18	SSKT	o	CE+CE	18.77
		3DR18	SSKT	o	CE+KD	17.59
I		3DR18	SSKT	x	CE+CE	18.64
		3DR18	SSKT	x	CE+KD	18.12
		3DR18	SSKT	o	CE+CE	18.38
		3DR18	SSKT	o	CE+KD	18.77
P+I		3DR18	SSKT	x	CE+CE	19.75
		3DR18	SSKT	x	CE+KD	18.38
		3DR18	SSKT	o	CE+CE	**20.54**
		3DR18	SSKT	o	CE+KD	17.99

Table 7. SSKT-based transfer learning performance change for UCF101 and HMDB51 compared to the training from scratch with MobileNet V2 (MV2) and DenseNet121 (D121).

T_s	T_t	Model	Method	TM	Loss	acc.
-	S10	MV2	scratch	-	CE	72.26±0.83
P		MV2	SSKT	x	CE+CE	75.79±0.19
		MV2	SSKT	x	CE+KD	74.0±0.35
		MV2	SSKT	o	CE+CE	75.28±0.49
		MV2	SSKT	o	CE+KD	73.37±1.8
I		MV2	SSKT	x	CE+CE	76.08±0.63
		MV2	SSKT	x	CE+KD	74.39±0.82
		MV2	SSKT	o	CE+CE	75.35±0.61
		MV2	SSKT	o	CE+KD	72.6±0.67
P+I		MV2	SSKT	x	CE+CE	76.69±0.18
		MV2	SSKT	x	CE+KD	73.29±0.89
		MV2	SSKT	o	CE+CE	**76.96±0.39**
		MV2	SSKT	o	CE+KD	73.35±0.99
-		D121	scratch	-	CE	72.02±0.48
P		D121	SSKT	x	CE+CE	76.17±0.35
		D121	SSKT	x	CE+KD	74.83±0.59
		D121	SSKT	o	CE+CE	73.46±0.62
		D121	SSKT	o	CE+KD	72.55±0.43
I		D121	SSKT	x	CE+CE	76.0±0.33
		D121	SSKT	x	CE+KD	73.7±0.22
		D121	SSKT	o	CE+CE	74.35±0.3
		D121	SSKT	o	CE+KD	71.13±0.59
P+I		D121	SSKT	x	CE+CE	**77.03±0.17**
		D121	SSKT	x	CE+KD	73.76±0.84
		D121	SSKT	o	CE+CE	76.09±0.26
		D121	SSKT	o	CE+KD	70.94±1.14

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *In Proc. of ECCV*, 2016. 1, 2

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *In Proc. of CVPR*, 2015. 1, 2

[26] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. Subclass distillation. *CoRR abs/2002.03936*, 2020. 2

[27] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Imagenet classification with deep convolutional neural networks. *In Proc. of CVPR*, 2019. 2

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *In Proc. of NeurIPS*, 2019. 5, 6

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *In Proc. of NIPS*, 2015. 1, 2

[30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *In Proc. of ICMR*, 2015. 2

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *In Proc. of CVPR*, 2018. 7

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In Proc. of ICLR*, 2015. 2

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR abs/1212.0402*, 2012. 7

[34] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *In Proc. of ICANN*, 2018. 1

[35] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. *In Proc. of CVPR*, 2020. 2

[36] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *In Proc. of NIPS*, 2014. 1

[37] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. *In Proc. of CVPR*, 2018. 1

[38] Amir Zamir, Alexander Sax, Teresa Yeo, Oğuzhan Kar, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas Guibas. Robust learning through cross-task consistency. *In Proc. of CVPR*, 2020. 1

Table 8. SSKT results using pretrained weights. *ft* stands for fine-tuning and K stands for Kinetics-400 dataset (Kay et al. 2017).

T_s	T_t	Model	Method	TM	Loss	acc.
-	VOC	R18	*ft* (I)	-	CE	90.52±0.11
P		R18	SSKT	x	CE+CE	89.3±0.04
		R18	SSKT	x	CE+KD	**92.28±0.06**
		R18	SSKT	o	CE+CE	90.83±0.04
		R18	SSKT	o	CE+KD	92.21±0.05
I		R18	SSKT	x	CE+CE	91.29±0.03
		R18	SSKT	x	CE+KD	92.26±0.07
		R18	SSKT	o	CE+CE	91.58±0.15
		R18	SSKT	o	CE+KD	92.19±0.09
P+I		R18	SSKT	x	CE+CE	91.28±0.05
		R18	SSKT	x	CE+KD	92.19±0.07
		R18	SSKT	o	CE+CE	91.25±0.08
		R18	SSKT	o	CE+KD	92.25±0.07
-	U101	3DR18	*ft* (K)	-	CE	83.95
P		3DR18	SSKT	x	CE+CE	84.53
		3DR18	SSKT	x	CE+KD	**84.58**
		3DR18	SSKT	o	CE+CE	83.87
		3DR18	SSKT	o	CE+KD	83.98
I		3DR18	SSKT	x	CE+CE	81.99
		3DR18	SSKT	x	CE+KD	83.42
		3DR18	SSKT	o	CE+CE	84.29
		3DR18	SSKT	o	CE+KD	84.37
P+I		3DR18	SSKT	x	CE+CE	78.56
		3DR18	SSKT	x	CE+KD	84.14
		3DR18	SSKT	o	CE+CE	82.81
		3DR18	SSKT	o	CE+KD	84.19
-	H51	3DR18	*ft* (K)	-	CE	56.64
P		3DR18	SSKT	x	CE+CE	56.77
		3DR18	SSKT	x	CE+KD	56.77
		3DR18	SSKT	o	CE+CE	57.75
		3DR18	SSKT	o	CE+KD	**57.82**
I		3DR18	SSKT	x	CE+CE	56.18
		3DR18	SSKT	x	CE+KD	56.9
		3DR18	SSKT	o	CE+CE	53.3
		3DR18	SSKT	o	CE+KD	57.75
P+I		3DR18	SSKT	x	CE+CE	54.48
		3DR18	SSKT	x	CE+KD	56.05
		3DR18	SSKT	o	CE+CE	57.29
		3DR18	SSKT	o	CE+KD	57.1

[39] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. *In Proc. of CVPR*, 2018. 2, 8

[40] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, , and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 5