Multi-Label Knowledge Distillation Supplementary Materials

Penghui Yang[1,2*], Ming-Kun Xie[1,2*], Chen-Chen Zong[1,2], Lei Feng[3],
Gang Niu[4], Masashi Sugiyama[4,5], Sheng-Jun Huang[1,2†]
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[4]RIKEN Center for Advanced Intelligence Project, [5]The University of Tokyo, Tokyo, Japan
phyang.cs@gmail.com, mkxie@nuaa.edu.cn, chencz@nuaa.edu.cn, lfengqaq@gmail.com,
gang.niu.ml@gmail.com, sugi@k.u-tokyo.ac.jp, huangsj@nuaa.edu.cn

1. More Details of Implementation

In order to validate the proposed method with diverse architectures, we employ some commonly used models, including ResNet [4], Wide ResNet (WRN) [12], RepVGG [2], Swin Transformer [6], and MobileNet v2 [8]. For all the backbones, we utilize their pre-trained version on the ImageNet [1] as our base model.

For all experiments, similar to the previous work [7], we employ the label-wise embedding encoder consisting of a cross-attention module and a feed-forward fully-connected layer [10]. The cross-attention module takes full queries and feature maps as the input. We assign a query per class to ensure that each query corresponds to a single semantic. The multi-label classifier $\boldsymbol{h}(.)$ is a fully-connected layer for each class, which outputs a predicted logit for a class label based on the input label-wise embedding.

2. More Results on NUS-WIDE

Table 1 reports comparison results on NUS-WIDE with the same and different architectures of student and teacher models. For distillation between the same architectures, we choose a ResNet-101 [4] as the teacher and a ResNet-34 as the student. For distillation between different architectures, we choose a Swin-T [6] as the teacher and a MobileNet v2 [8] as the student. From the tables, it can be observed that the proposed L2D significantly outperforms all comparing methods, which convincingly validates the effectiveness of the proposed label-wise embeddings distillation.

Table 1. Results on NUS-WIDE validation.

Teacher	ResNet-101			Swin-T		
Student	ResNet-34			MobileNet v2		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	55.32	75.56	61.31	59.73	77.30	65.44
Student	53.41	75.10	60.08	54.49	75.72	61.74
RKD	53.62	75.20	59.91	54.76	75.69	61.74
PKT	53.55	75.08	60.35	54.59	75.69	61.74
ReviewKD	53.52	75.23	60.44	54.85	75.84	61.75
MSE	53.52	75.13	59.94	54.86	75.80	61.69
PS	54.14	75.43	60.79	55.18	75.91	62.35
MLD	54.44	75.36	60.73	55.36	76.00	62.52
L2D	**55.31**	**76.17**	**62.79**	**56.91**	**76.92**	**63.89**

Table 2. Results of reversed knowledge distillation on MS-COCO validation, where the backbones of teacher and student model are respectively ResNet-34 and ResNet-101. The numbers in the brackets indicate the performance gaps between the student and the teacher.

Metrics	mAP	OF1	CF1
Teacher	70.19	72.30	66.50
Student	73.98 (+3.79)	75.01 (+2.71)	70.12 (+3.62)
RKD	74.03 (+3.84)	74.96 (+2.66)	70.01 (+3.51)
PKT	73.95 (+3.76)	74.94 (+2.64)	69.98 (+3.48)
ReviewKD	74.02 (+3.83)	74.96 (+2.66)	70.07 (+3.57)
MSE	74.21 (+4.02)	75.12 (+2.82)	70.18 (+3.68)
PS	74.70 (+4.51)	75.78 (+3.48)	71.08 (+4.58)
MLD	74.64 (+4.45)	75.78 (+3.48)	71.10 (+4.60)
L2D	**75.51(+5.32)**	**76.25(+3.95)**	**71.75(+5.25)**

3. More Results on Pascal VOC 2007

Table 3 and Table 4 report comparison results on Pascal VOC 2007 with the same and different architectures of student and teacher models. From the tables, it can be observed that the proposed L2D significantly outperforms all comparing methods, which convincingly validates the effectiveness of the proposed label-wise embeddings distillation. Compared with the results on MS-COCO, the performance gap between L2D and comparing methods seems to

*Both authors contributed equally to this research.
†Correspondence to: Sheng-Jun Huang (huangsj@nuaa.edu.cn).

become smaller. One possible reason is that VOC only contains about 1.5 labels per image, which leads conventional KD methods to obtain a better performance.

4. Reversed Knowledge Distillation

In practice, teacher networks are always pretrained without any knowledge of the student's architecture, which makes it possible that the student is more complicated than the teacher. Previous study [11] proved that the superior network can also be enhanced by learning from a weak network. To explore the performance of our method in this setting, we further conduct experiments on MS-COCO. In detail, we use a ResNet-34 as the teacher and a ResNet-101 as the student, which makes the vanilla student model outperforms the teacher. From Table 2, we can find that our method still outperforms all the other methods. This implies that our method is also effective for the reversed KD setting.

5. Parameter Sensitivity Analysis

In this section, we study the influence of balancing parameters λ_{MLD}, λ_{CD} and λ_{ID} on the performance of L2D. A commonly used setting of the hyperparameter in vanilla KD that balances KL divergence against CE is 0.9 [9], which means the balancing parameter for CE is 0.1 and the one for KL divergence is 0.9. So we choose 10 for the balancing parameter for MLD, which is closest to the setting of vanilla KD. We set the balancing parameter for ID loss larger than CD loss considering that the ID loss may carry less information because there are only less than 3 labels for a instance on average, though it seems unnecessary since parameter sensitivity experiments in Figure 1 show that the performance of L2D are not sensitive to all of our balancing parameters.

6. Visualization of Attention Maps

To further show the effectiveness of our proposed method L2D, we visualize some attention maps of the penultimate layer in the visual backbones using LayerCAM [5] implemented by François-Guillaume Fernandez [3]. We compare attention maps of the student model trained by L2D with some other methods in Figure 2 3 4. We compare L2D with: 1)Vanilla: student trained without distillation; 2)ReviewKD: a classical feature-based method which has the state-of-the-art performance among all conventional KD methods. In each figure, the first column shows the raw picture and the other columns show class activation maps overlaying on the raw picture. Each row represents a certain class. From these figures, we can find that L2D can locate the specified object more precisely than the other methods, which means it can not only pay attention to target objects, but also resist interference from similar but unrelated objects. All these comparisons show that L2D outperforms all comparing methods. It validates the effectiveness of our proposed label-wise embeddings distillation and shows great potential in MLKD.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009. 1

[2] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021. 1

[3] François-Guillaume Fernandez. Torchcam: class activation explorer. https://github.com/frgfm/torch-cam, 2020. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[5] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, pages 5875–5888, 2021. 2

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[7] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. *arXiv preprint arXiv:2111.12933*, 2021. 1

[8] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 1

[9] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1

[11] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

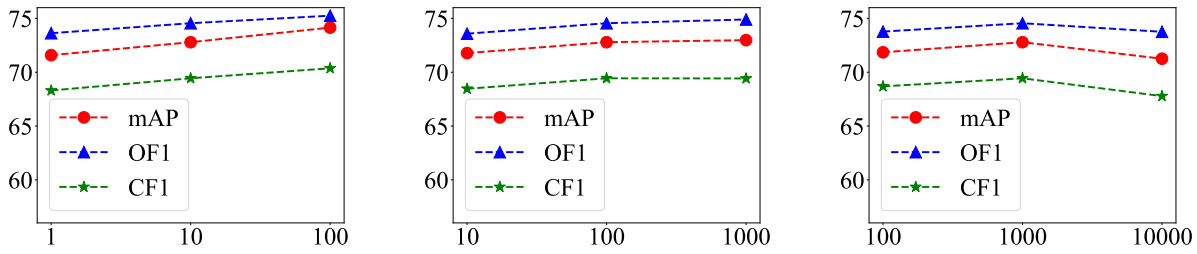
[12] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016. 1

Table 3. Results on Pascal VOC 2007 validation teacher and student models are in the **same** architectures.

Teacher	RepVGG-A2			ResNet-50			WRN-101			Swin-S		
Student	RepVGG-A0			ResNet-18			WRN-50			Swin-T		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	86.20	85.63	82.62	86.73	84.92	81.21	88.00	87.03	83.72	92.75	91.05	88.82
Student	83.79	83.36	79.83	84.01	83.60	79.42	88.52	87.21	84.08	91.31	89.98	88.00
RKD	83.75	83.41	79.85	84.48	83.54	79.83	88.21	87.33	84.55	91.52	90.44	88.51
PKT	83.63	83.53	80.04	84.12	83.10	79.31	87.69	87.07	84.14	91.28	90.17	88.03
ReviewKD	83.87	83.98	80.54	83.71	83.01	79.25	88.23	87.13	84.20	91.45	90.17	88.06
MSE	84.02	83.67	79.94	84.23	83.16	79.29	88.04	86.49	83.57	91.06	89.99	87.66
PS	83.77	83.74	80.28	84.44	83.78	79.95	88.30	86.92	83.91	91.21	90.25	88.12
MLD	83.65	83.66	80.02	84.48	84.07	80.29	88.29	87.16	84.25	91.43	90.72	88.81
L2D	**84.56**	**84.37**	**80.82**	**85.71**	**85.70**	**82.11**	**89.52**	**88.25**	**85.69**	**91.92**	**91.34**	**89.58**

Table 4. Results on Pascal VOC 2007 validation where teacher and student models are in the **different** architectures.

Teacher	ResNet-50			Swin-T			ResNet-50			Swin-T		
Student	RepVGG-A0			ResNet-18			MobileNet v2			MobileNet v2		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	86.73	84.92	81.21	91.43	89.81	87.63	86.73	84.92	81.21	91.43	89.81	87.63
Student	83.79	83.36	79.83	84.01	83.60	79.42	86.12	85.01	81.76	86.12	85.01	81.76
RKD	84.26	84.29	80.70	83.27	83.05	79.55	86.22	84.97	81.76	85.68	85.31	81.57
PKT	83.93	83.79	80.03	83.45	83.25	79.64	86.10	84.84	81.66	85.67	85.22	81.68
ReviewKD	84.07	83.62	80.34	83.37	83.08	78.93	85.87	85.04	81.73	85.69	85.10	81.56
MSE	84.01	84.05	80.52	83.60	83.06	79.46	86.20	84.94	81.84	85.80	85.51	81.98
PS	84.80	84.46	81.13	83.97	83.75	79.86	86.26	85.47	82.06	86.07	85.73	82.39
MLD	85.07	84.91	81.55	84.61	84.26	80.78	86.38	85.67	82.43	86.11	85.98	82.55
L2D	**86.26**	**85.85**	**82.55**	**85.87**	**85.67**	**82.17**	**87.32**	**86.48**	**83.26**	**87.37**	**86.88**	**83.68**



(a) λ_{MLD} (b) λ_{CD} (c) λ_{ID}

Figure 1. Student models' performance comparisons with different values of λ_{MLD}, λ_{CD} and λ_{ID} respectively on MS-COCO with a ResNet-101 as the teacher and a ResNet-34 as the student.

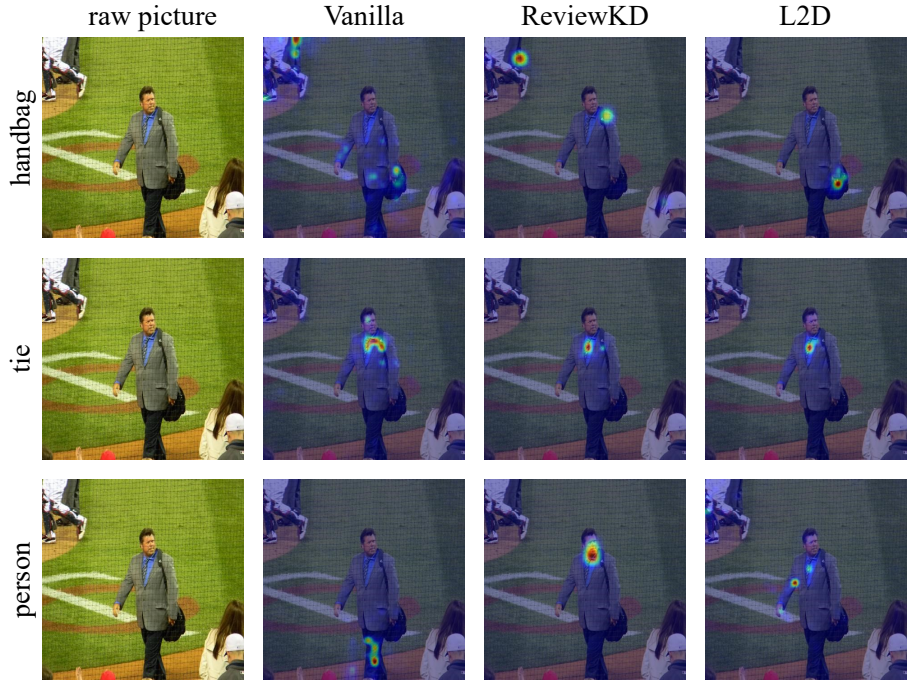


Figure 2. An example of visualization of attention maps. We can find that on attention head for class *handbag*, both Vanilla and ReviewKD are interfered by some other objects and do not pay all attention to the handbag, but our L2D resists such interference successfully.

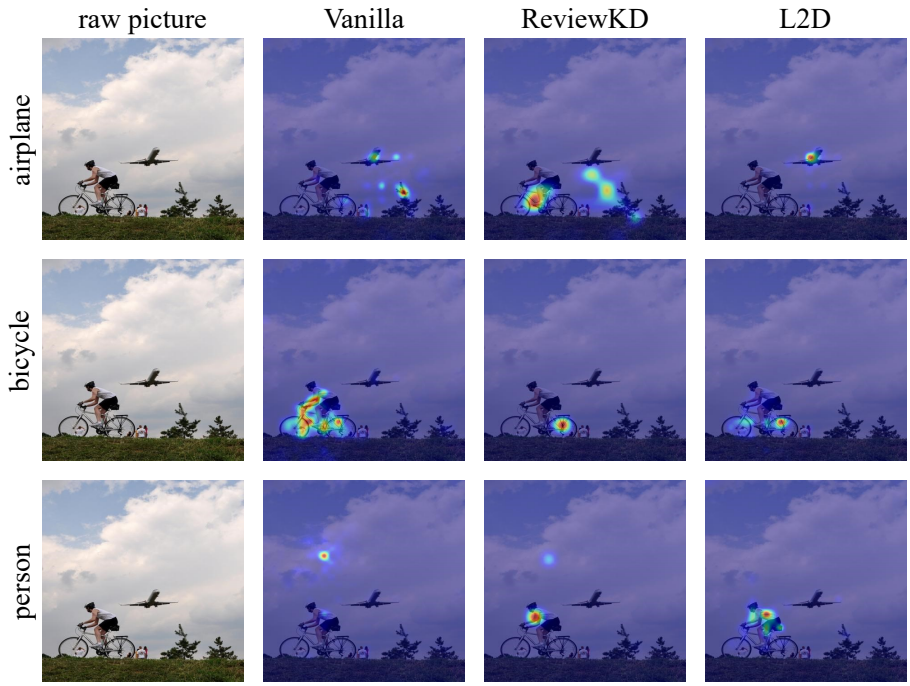


Figure 3. An example of visualization of attention maps. We can find that on attention head for class *airplane*, both Vanilla and ReviewKD do not pay all attention to the airplane: Vanilla is interfered by the plant and ReviewKD is interfered by the boy. But our L2D resists such interference successfully. On attention head for class *person*, both Vanilla and ReviewKD are interfered by the shades on the cloud, but L2D is not.

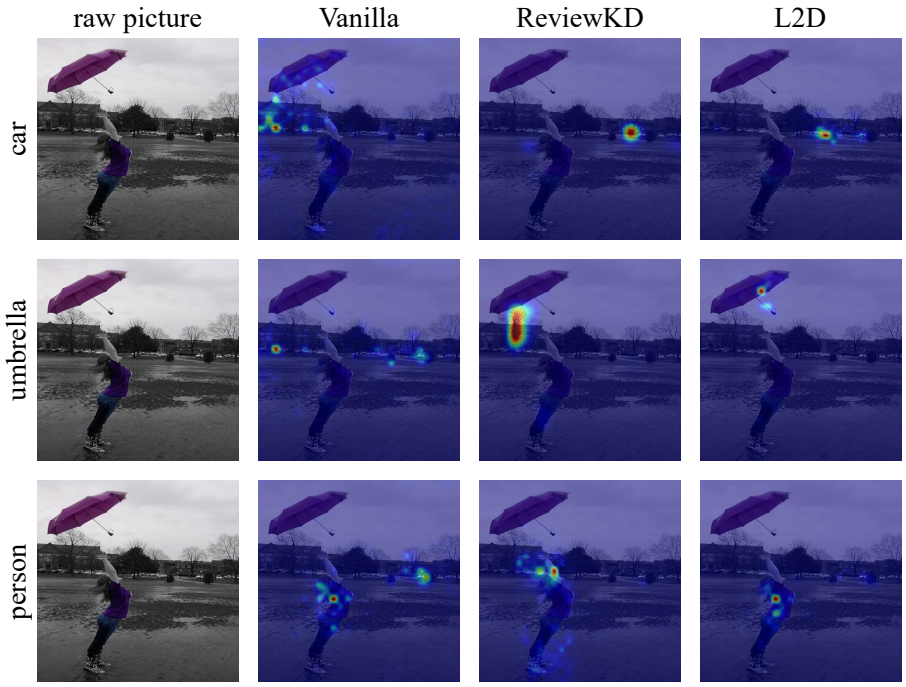


Figure 4. An example of visualization of attention maps. We can observe that on attention heads for class *car* and class *person*, Vanilla pays some of its attention to unrelated objects. On attention head for class *umbrella*, ReviewKD pays some of its attention to the house. Only L2D can concentrate on these three targets precisely.