Acknowledging the Unknown for Multi-label Learning with Single Positive Labels (Supplementary Material)

Donghao Zhou[1,2], Pengfei Chen[3], Qiong Wang[1], Guangyong Chen[4(✉)], and Pheng-Ann Heng[1,5]

[1] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
dh.zhou@siat.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Tencent Technology, Shenzhen, China
[4] Zhejiang Lab, Hangzhou, China
gychen@zhejianglab.com
[5] The Chinese University of Hong Kong, Hong Kong, China

A Derivation of the Gradient Equations

In this section, we provide detailed derivation of the gradients of AN and EM loss (i.e. Eq. 2 and Eq. 5). following the notations in the main paper, the gradients of $\mathcal{L}_+$, $\mathcal{L}_-$ and $\mathcal{L}_\varnothing$ for the logit g are

$$\begin{aligned}\frac{\partial \mathcal{L}_+}{\partial g} &= \frac{\partial \mathcal{L}_+}{\partial p}\frac{\partial p}{\partial g} = -\frac{1}{p}\cdot p(1-p) = p-1 \\ &= \frac{1}{1+e^{-g}} - 1 = \frac{-e^{-g}}{1+e^{-g}},\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_-}{\partial g} &= \frac{\partial \mathcal{L}_-}{\partial p}\frac{\partial p}{\partial g} = \frac{1}{1-p}\cdot p(1-p) \\ &= p = \frac{1}{1+e^{-g}},\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_\varnothing}{\partial g} &= \frac{\partial \mathcal{L}_\varnothing}{\partial p}\frac{\partial p}{\partial g} = \alpha[\log p - \log(1-p)]\cdot p(1-p) \\ &= \alpha \log\frac{p}{1-p}\cdot p(1-p) \\ &= -\alpha \log e^{-g}\cdot \frac{1}{1+e^{-g}}\frac{e^{-g}}{1+e^{-g}} \\ &= \frac{-\alpha e^{-g}\log e^{-g}}{(1+e^{-g})^2}.\end{aligned} \tag{11}$$

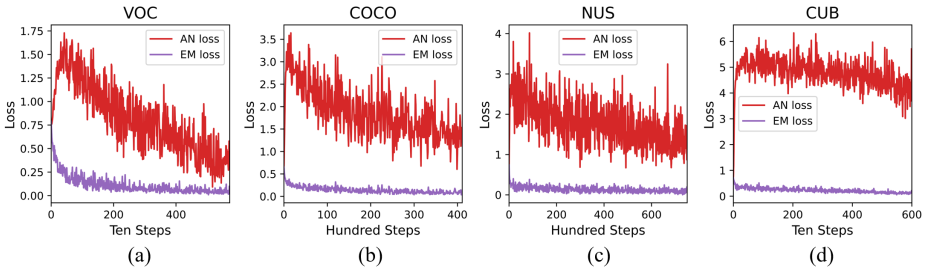

Fig. 7. Training losses of annotated positive labels (i.e. $\mathcal{L}_+$) on all four datasets from the models trained with AN and EM loss, where α of EM loss is set to the corresponding value shown in Table 5

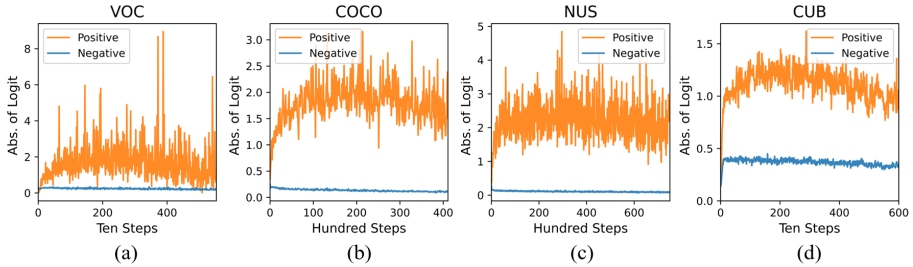

Fig. 8. Absolute values of the logits for unannotated positive and negative labels, produced by the model trained with EM loss on all four datasets

B Additional Empirical Evidence for EM Loss

In Sec. 3.2, we have claimed that the gradient regime of EM loss leads to three behaviours beneficial to model training: 1) Learning from annotated labels preferentially. 2) Mitigating the effect of label noise. 3) Maintaining confident positive predictions. In this section, we provide more empirical evidence to further verify these three advantages respectively, and thus further demonstrate the effectiveness of EM loss. As done in the main paper, for empirical analysis, we consider AN loss as the baseline of SPML, and adopt AN and EM loss in model training respectively. Note that the experimental setup and the hyperparameters of each method are the same as those in benchmark experiments.

B.1 Learning from Annotated Labels Preferentially

We have claimed that the model training would be dominated by assumed negative labels when adopting AN loss, whereas EM loss can lead the model to preferentially learn from annotated positive labels. To verify this, we present the training losses of annotated positive labels (i.e. $\mathcal{L}_+$) on VOC in the main paper, where α of EM loss is set to 1. We are curious about if this improvement of EM loss can perform well on all four datasets when α is set to a more proper

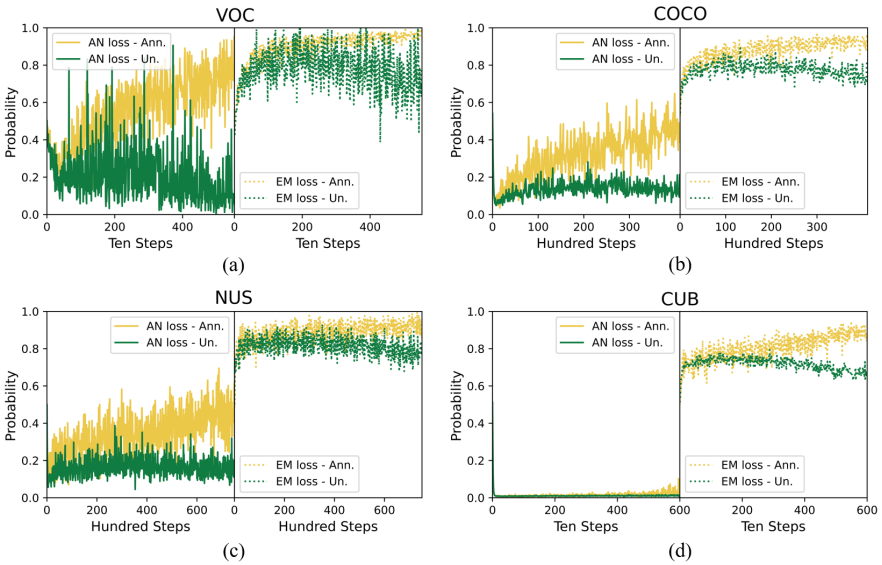

Fig. 9. Predicted probabilities for annotated and unannotated positive labels, produced by the models trained with AN (*left*) and EM (*right*) loss on all four datasets

value. Thus, we provide more visualization of $\mathcal{L}_+$ for comparison, where α is set to the corresponding value shown in Table 5. As shown in Fig. 7, on all four datasets, $\mathcal{L}_+$ of AN loss would increase in early training, since the model trained with AN loss preferentially focuses on fitting the numerous assumed negative labels. However, $\mathcal{L}_+$ of EM loss can gradually decrease and is more stable, since the gradients of EM loss for unannotated labels would be relatively low in early training (see Fig. 2(a)). Moreover, we have also claimed that EM loss tends to keep the predictions of unannotated labels ambiguous. To empirically observe this, we visualize the absolute values of the logits produced by the model trained with EM loss. In Fig. 8, it can be observed that EM loss would keep near-zero logits for numerous unannotated negative labels, which results in small gradients for them throughout training. As for the relatively large results of unannotated positive labels, we would discuss them in Sec. B.3.

B.2 Mitigating the Effect of Label Noise

In Sec. 3.2, we have also claimed that the model trained with AN loss would suffer from the false negative labels (i.e. unannotated positive labels which are assumed as negative ones), whereas EM loss can mitigate the effect of label noise. Trained with AN loss, the model would be confused by annotated and unannotated positive labels, which results in unconfident and even incorrect predictions for positive labels. To demonstrate this, we present the predicted probabilities for annotated and unannotated positive labels, which are produced by the models

trained with AN and EM loss. As shown in Fig. 9, the model trained with AN loss would produce low predicted probabilities for positive labels, especially for the unannotated ones. However, the model trained with EM loss can produce relatively confident positive predictions for both unannotated and annotated positive labels, since EM loss does not introduce any false negative labels and is capable of encouraging the model to learn from annotated positive ones.

B.3 Maintaining Confident Positive Predictions

Different from AN loss, EM loss can maintain confident positive predictions for unannotated labels due to its special gradient regime. This claim can be verified by visualizing the logits for unannotated positive labels, since confident positive predictions are always associated with large logits for unannotated positive labels. For instance, a logit of 1 (resp. 2, 3) corresponds to a predicted probability of 0.73 (resp. 0.88, 0.95). As shown in Fig. 8, compared to the logits of annotated negative labels, the model trained with EM loss would produce relatively large logits for unannotated positive labels, since EM loss would not over-suppress confident positive predictions by providing large gradients for them. This difference between the logits of unannotated positive and negative labels shows that EM loss can indeed maintain confident positive predictions, instead of keeping near-zero logits for all unannotated labels without distinction.

C Details of the Experimental Setup

In this section, we provide more details of the experimental setup, including dataset descriptions and hyperparameter tuning and selection, to ensure the fairness and reproducibility of our experiments.

C.1 Datasets

The following large-scale multi-label datasets are used in our experiments: PASCAL VOC 2012 (VOC) [3], MS-COCO 2014 (COCO) [6], NUS-WIDE (NUS) [1], and CUB-200-2011 (CUB) [8]. VOC consists of 5,717 training images and 20 classes, and we report test results on its official validation set with 5,823 images. COCO contains 82,081 training images and 80 classes, and we also report test results on its official validation set with 40,137 images. NUS consists of 81 classes and contains 150,000 training images and 60,260 testing images collected from Flickr. Instead of re-crawling the NUS images as done in [2], we use the *official version* of NUS in our experiments, which has less manual intervention and thus is fairer. CUB is divided into 5,994 training images and 5,794 test images, consisting of 312 classes (i.e. binary attributes of birds). For reference purposes, we summarize the statistics of the datasets in Table 4, which shows the diversity of these four popular multi-label datasets.

Table 4. Statistics of the datasets, including the number of classes, the number of images on the split datasets, and the number of ground-truth positive and negative labels per image on the training sets

Statistics		**VOC**	**COCO**	**NUS**	**CUB**
# Classes		20	80	81	312
# Images	Training	4574	65665	120000	4795
	Validation	1143	16416	30000	1199
	Test	5823	40137	60260	5794
# Labels Per Training Image	Positive	1.46	2.94	1.89	31.4
	Negative	18.54	77.06	79.11	280.6

C.2 Comparing Methods

We compare our method with the following methods: 1) AN loss (Eq. 1): The widely recognized baseline of SPML, which assumes all unannotated labels are negative. 2) EntMin [4]: A widely adopted method of semi-supervised learning, i.e. entropy minimization regularization, which aims to minimize the entropy of predicted probabilities for unannotated labels. 3) Focal loss [5]: An efficient method to handle label imbalance. 4) ASL [7]: One of the state-of-the-art methods of multi-label classification, which can mitigate the effect of mislabeled samples. 5) ROLE [2]: the state-of-the-art method of SPML, which adopts a label estimator and exploits the average number of positive labels to perform regularization. 6) ROLE+LI [2]: ROLE is combined with the "LinearInit" training fashion, i.e. firstly training the model with the backbone being frozen before end-to-end training. Note that unannotated labels are also assumed as negative ones in Focal loss and ASL. Besides, we also compare our method to the baseline of SPML with the following improvement: 1) DW: Down-weighting $\mathcal{L}_-$ of Eq. 1. 2) L1R/L2R: adopting l_1/l_2 regularization. 3) LS: Label smoothing for all labels. 4) N-LS: Label smoothing for only assumed negative labels.

C.3 Hyperparameters

For each method, method-specific hyperparameters are tuned on all four datasets respectively, and the hyperparameters with the best mAP on validation sets are selected for the final evaluation. The detailed hyperparameter tuning and selection of our experiments are as follows:

1. **DW**: A hyperparameter tuned in $\{0.01, 0.02, 0.1, 0.2, 0.4, 0.9\}$ is used to down-weight $\mathcal{L}_-$ of Eq. 1. Finally, 0.1 is selected for VOC, COCO, and NUS, and 0.02 is selected for CUB.
2. **L1R/L2R**: A hyperparameter tuned in $\{1e-9, 1e-8, 1e-7, 1e-6, 1e-5\}$ is used to control the strength of l_1/l_2 regularization. In L1R, we select $1e-6$ (resp. $1e-7$, $1e-7$, $1e-9$) for VOC (resp. COCO, NUS, CUB). In L2R, we select $1e-6$ (resp. $1e-7$, $1e-6$, $1e-8$) for VOC (resp. COCO, NUS, CUB).

Table 5. Hyperparameters of our method on all four datasets in our experiments

hyperparameters	**VOC**	**COCO**	**NUS**	**CUB**
Batch Size	8	16	16	8
Learning Rate	$1e-5$	$1e-5$	$1e-5$	$1e-4$
α	0.2	0.1	0.1	0.01
β	0.02	0.9	0.2	0.4
$\theta\%$	90%	90%	90%	90%
T_{w}	5	5	4	3

3. **LS/N-LS**: Label smoothing coefficient is tuned in $\{0.1, 0.2, 0.3\}$. In LS, we select 0.1 for all four datasets. In N-LS, we select 0.3 for VOC, and select 0.1 for the other datasets.
4. **EntMin** [4]: A hyperparameter tuned in $\{0.01, 0.02, 0.1, 0.2, 0.4, 0.9\}$ is used to control the strength of entropy minimization regularization. Finally, we select 0.01 (resp. 0.9, 0.4, 0.4) for VOC (resp. COCO, NUS, CUB).
5. **Focal loss** [5]: There are a focusing parameter γ and a balance parameter α in Focal loss. As recommended in [7], we set $\gamma = 2$, and tune $\alpha \in \{0.25, 0.5, 0.75\}$. Finally, we select $\alpha = 0.75$ for all four datasets.
6. **ASL** [7]: There are two hyperparameters (i.e. γ_+ and γ_-) used to control the focusing levels of positive and negative labels respectively, and a hyperparameter (i.e. m) used to act as the proposed probability margin in ASL. As done in [7], we set $\gamma_+ = 0$, and tune $\gamma_- \in \{1, 2\}$ and $m \in \{0, 0.05, 0.2\}$ with a grid search. Finally, we select $\gamma_- = 2$, $m = 0.2$ for VOC, COCO, and NUS, and select $\gamma_- = 1$, $m = 0$ for CUB.
7. **ROLE/ROLE+LI** [2]: The experimental results are reproduced by reimplementing the methods exactly following the hyperparameters in [2].
8. **EM loss/APL**: For our method, we tune $\alpha, \beta \in \{0.01, 0.02, 0.1, 0.2, 0.4, 0.9\}$. Moreover, we set $\theta\% = 90\%$ for all datasets and empirically select T_{w} for each dataset. For convenience, the final hyperparameters of our method are shown in Table 5, including the selected batch sizes and learning rates.

D Detailed Analysis for APL

As an extension to the ablation study of APL in Sec. 4.3, we provide detailed analysis for it in this section to further demonstrate the contribution of the components adopted in APL. We focus on answering the following key questions:

Question 1: Is a high sample proportion necessary for generating negative pseudo-labels?
Answer 1: In Table 3, it can be observed that generating negative pseudo-labeling with a low sample proportion just leads to a tiny mAP increment. As shown in Fig. 3(b), generating negative pseudo-labeling with a low sample proportion gradually reduces performance as pseudo-labeling goes on, since

Table 6. Precision (averaged in 3 runs) of positive pseudo-labels generated by the similar positive pseudo-labeling on four multi-label datasets. Note that the sample proportion for positive pseudo-labeling is set to 10%

	VOC	**COCO**	**NUS**	**CUB**
Precision	15.66%	15.29%	9.18%	19.34%

the model may be overfitting to few negative pseudo-labels. Thus, adopting a high-tolerance strategy is necessary for generating negative pseudo-labels.
Question 2: Is assigning hard labels also able to boost performance?
Answer 2: As shown in Table 3, when assigning hard labels instead of soft ones, pseudo-labeling would not significantly boost performance, since label noise may be contained in the generated negative pseudo-labels. As a solution, soft labels can mitigate this damaging impact and make negative pseudo-labels participate in model training in a more appropriate way, which is beneficial to better performance (see Table 3).
Question 3: Does down-weighting contribute to performance improvement?
Answer 3: As shown in Fig. 3(b), pseudo-labeling without down-weighting (i.e. β of Eq. 8 is set to 1) can still achieve stable training, whereas properly performing down-weighting for the loss of pseudo-labels can lead to further performance improvement (see Table 3).
Question 4: What is the effect of performing positive pseudo-labeling?
Answer 4: As shown in Table 3, it is worth noting that performing similar positive pseudo-labeling would cause a performance drop. Since positive labels are the tiny minority of multi-label annotations, positive pseudo-labeling would introduce a large amount of label noise, even with a small sample proportion. To empirically observe this, we present the precision of positive pseudo-labels in Table 6, which shows that performing positive pseudo-labeling can only generate positive pseudo-labels with very low precision due to the positive-negative label imbalance of unannotated labels. Therefore, considering this imbalance, we choose to adopt an extreme low-tolerance strategy for positive pseudo-labels, aiming to avoid introducing any noisy positive pseudo-labels for more stable training (see Fig. 3(b)).

E Evaluation with Other Metrics

In multi-label learning, mAP is the primary metric to evaluate model performance. To further verify the effectiveness of our method, we perform an additional evaluation with two metrics (i.e. micro-F1 and macro-F1). Specifically, we compare EM loss with AN loss in the SPML setting, and report the performance of BCE loss on the fully labeled datasets. As commonly done, the thresholds for micro-F1 and macro-F1 are set to 0.5 for BCE and AN loss. Since EM loss tends to keep ambiguous predictions for unannotated labels, the predicted probabilities for negative labels produced by the model trained with EM loss would be

Table 7. Experimental results of BCE loss, AN loss, and EM loss on four SPML benchmarks with two additional metrics (i.e. micro-F1 and macro-F1). Note that the model trained with BCE loss adopts full annotations for training and the best performance of the methods in the SPML setting is marked in bold

Datasets	VOC		COCO		NUS		CUB	
Methods	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BCE loss	85.18	82.84	76.01	71.84	68.97	50.59	47.54	22.15
AN loss	73.24	71.66	38.03	41.85	28.24	21.36	0	0
EM loss	**85.36**	**82.78**	**71.58**	**66.74**	**66.83**	**45.47**	**43.85**	**20.38**

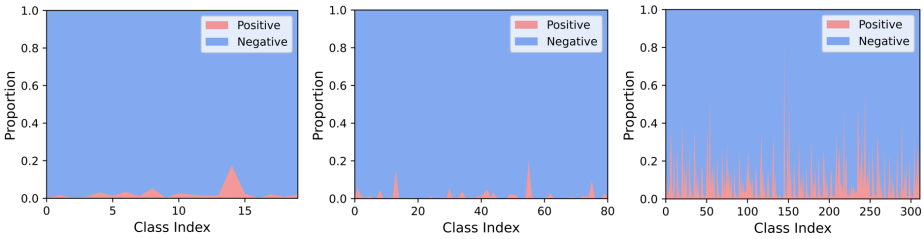

Fig. 10. Proportions of unannotated positive and negative labels of each class on the other datasets

near 0.5 (see Fig. 11). Therefore, for fair comparison, we set the thresholds to 0.75 for EM loss. As shown in Table 7, AN loss achieves poor performance on all four datasets, especially on CUB (both micro-F1 and macro-F1 are 0). However, our EM loss can still perform well in the evaluation with these two metrics. For instance, EM loss can achieve 85.36% micro-F1 on VOC, which even exceeds the results of being trained with full annotations.

F More Illustrative Examples

In this section, we provide more illustrative examples to support our observations in the main paper, including the positive-negative label imbalance of unannotated labels and the distinguishability of model predictions.

F.1 Positive-Negative Label Imbalance of Unannotated Labels

As an extension to Fig. 2(c), we present the proportions of unannotated positive and negative labels of each class on the other datasets in Fig. 10, which shows that the positive-negative labels imbalance of unannotated labels is an inherent property of SPML. As shown in Fig. 10, it is worth noting that the proportions of unannotated negative labels on some classes are lower than the predefined sample proportion for negative pseudo-labeling (i.e. 90%). Fortunately, with a

self-paced procedure, it is not often that APL would generate negative pseudo-labels with a sample proportion of 90% before early stopping, which does not damage the high precision of pseudo-labels generated by APL (see Table 2).

F.2 Distinguishability of Model Predictions

In Fig. 11, we visualize the predicted probabilities for positive and negative labels on more classes of COCO, aiming to further compare the effect of AN and EM loss on the distinguishability of model predictions. It can be observed that the model trained with EM loss can produce more distinguishable predictions for positive and negative labels. Moreover, we also present the class name and the percentage of mAP increment in the caption of each subfigure in Fig. 11, which shows that distinguishability improvement indeed contributes to model performance as we expect. Especially, as shown in Fig. 11(g), the model trained with EM loss can produce more distinguishable predictions on the "knife" class, even though they are rare and small objects in the images of COCO.

References

1. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009)
2. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
3. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep **8**, 5 (2011)
4. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004)
5. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
7. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
8. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., California Institute of Technology (2011)

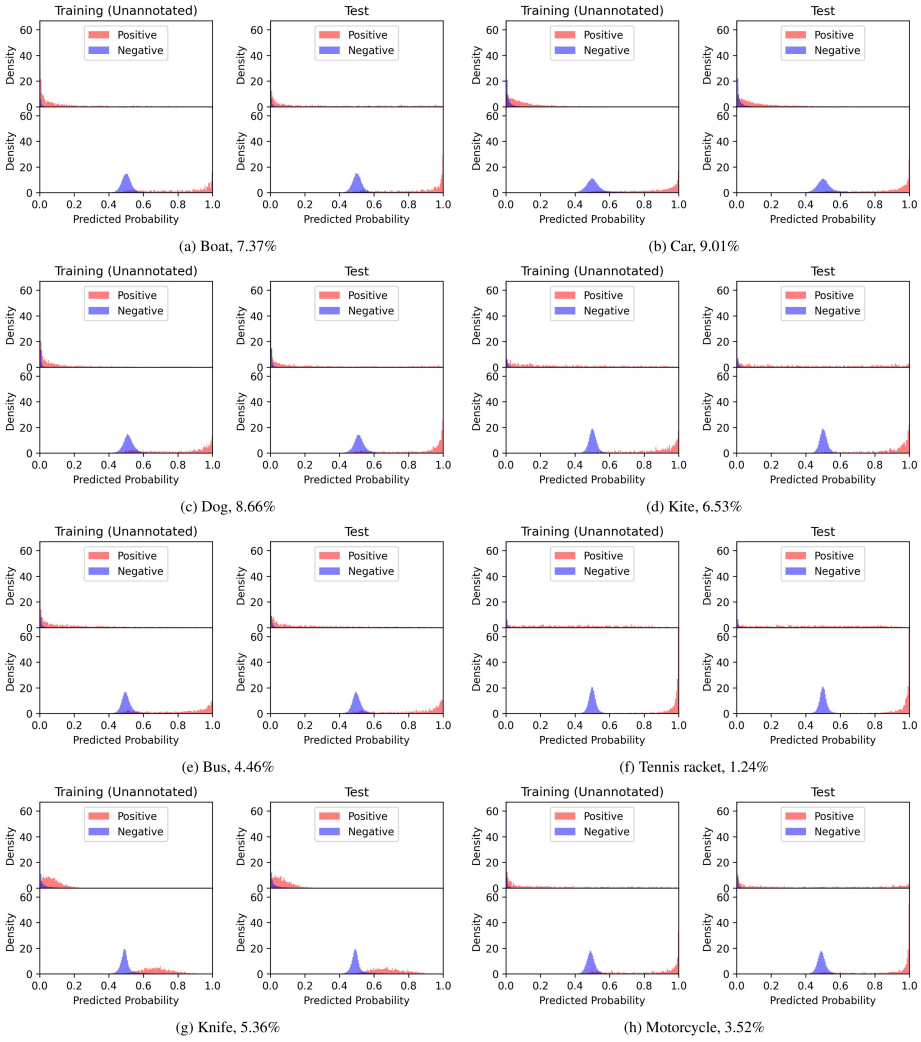

Fig. 11. Densities of the predicted probabilities for training and test images on more classes of COCO, produced by the models trained with AN (*top*) and EM (*bottom*) loss. Note that we only visualize the *unannotated* labels of training images. The caption of each subfigure contains the class name and the percentage of mAP increment. For clear comparison, we limit the y-axis to the same scale as Fig. 4(b).