A An Illustration of MAT

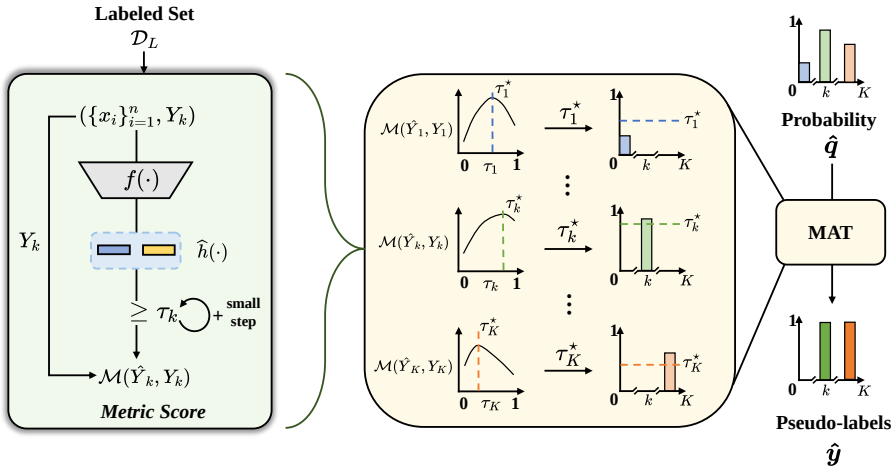


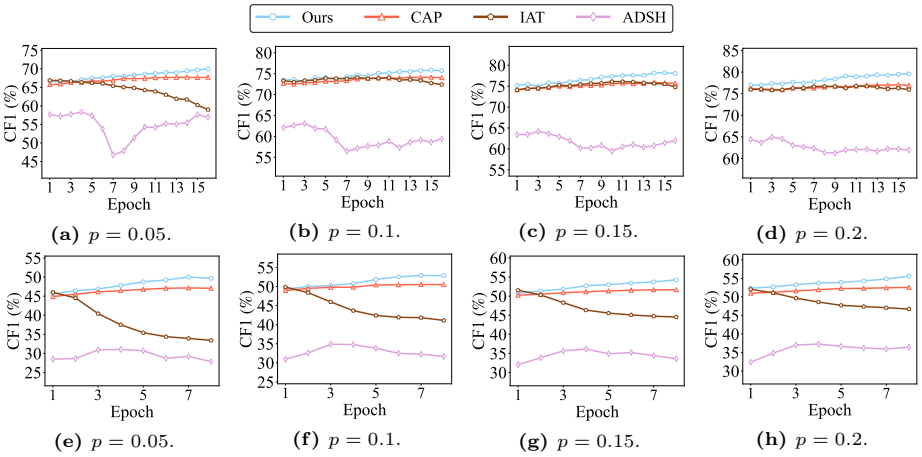
Fig. 5: An illustration of MAT. By feeding instances into the model $f(\cdot) \circ \widehat{h}(\cdot)$, we obtain the predictions. By adjusting τ_k, we can achieve the optimal pseudo-labeling performance $\mathcal{M}(\hat{Y}_k, Y_k)$.

Table 3: The detailed characteristics of three benchmark datasets.

Dataset	# Class.	# Train.	# Val.	# Avg.
VOC	20	5,717	5,823	1.5
COCO	80	82,081	40,137	2.9
NUS	81	150,000	60,260	1.9

B Details of Datasets

The detail characteristics of three benchmark datasets, including PASCAL VOC 2012 (VOC) [14], MS-COCO 2014 (COCO) [30] and NUS-WIDE (NUS), [7] are reported in Table 3. Specifically, VOC contains 5,717 training images and 5,823 validation images for 20 classes. The average number of labels per image in VOC is 1.5. For COCO, there are 82,081 training images and 40,137 validation images for 80 classes, and the average number of labels per image is 2.9. Following [44], we split NUS as 150,000 training images and 60,260 validation images,



(a) $p = 0.05$. **(b)** $p = 0.1$. **(c)** $p = 0.15$. **(d)** $p = 0.2$.

(e) $p = 0.05$. **(f)** $p = 0.1$. **(g)** $p = 0.15$. **(h)** $p = 0.2$.

Fig. 6: The performance of pseudo-labeling on VOC (a-d) and NUS (e-h).

containing 81 classes, where the average number of labels per image is 1.9. In our experiments, we report results on the above validation sets of three datasets.

C More Results of Pseudo-Labeling Performance

In Figure 6, we also perform experiments to examine the quality of pseudo-labels generated by different methods on VOC and NUS in terms of CF1 score. We can observe phenomena similar to those described in the main text on COCO. For the relatively simpler VOC dataset among the three datasets, the first three methods perform comparably across most proportions, and our method outperforms the other two. Even on the challenging NUS dataset, our method exhibits commendable performance compared to the SOTA.

D Reproducibility and Resource Consumption

To verify the reproducibility of our method, we conduct five runs using different random seeds (seed $= \{1, 2, 3, 4, 5\}$) and record the mean and standard deviation of our method's performance. In addition, we compare the results of the second-best method, CAP [44], after the same five runs. Table 4 presents a comparison between our method and CAP in terms of the mAP metric (with mean and standard deviation) across three datasets, showing that our method is not only reproducible but also superior to CAP. Furthermore, we compare the resource consumption of our method and CAP (also depicted in Table 4). Although our method slightly exceeds CAP in terms of memory and time usage, this additional resource investment is acceptable given the significant performance improvement.

Table 4: Mean and standard deviation of mAP(%) in CAP and our method, on three datasets, along with the time/memory comparison. 'Time' is the training time per epoch, including the process of threshold updating, 'GPU' is the max memory allocated during training phase.

Methods	CAP			Ours		
Datasets	VOC	COCO	NUS	VOC	COCO	NUS
$p = 0.05$	77.15±0.58	63.11±0.35	45.30±0.30	81.45±1.50	70.15±0.48	47.42±1.00
$p = 0.10$	82.54±0.20	67.96±0.32	48.89±0.37	85.65±0.92	73.65±0.34	51.01±0.43
$p = 0.15$	83.95±0.24	69.92±0.41	50.53±0.54	87.02±0.67	75.18±0.31	52.15±0.46
$p = 0.20$	85.04±0.32	71.23±0.42	51.82±0.43	87.83±0.38	76.21±0.30	53.37±0.43
Time	0.9min	10.3min	12.2min	1.7min	21.8min	38.9min
GPU	11.1G			14.2G		

Table 5: Mean average precision (mAP %) of the baseline incorporated with different components, on the dataset NUS. The baseline here indicates the method CAP.

MAT	D2L		NUS			
	CDD	GUD	p=0.05	p=0.10	p=0.15	p=0.20
			44.82	48.24	49.90	51.06
✓			45.26	48.88	50.23	51.20
✓	✓		45.74	49.30	50.83	52.04
✓	✓	✓	46.86	50.25	51.61	52.64

E More Ablation Studies

The Study on D2L and MAT. In Table 5, we report the results of ablation experiments on NUS, where the effectiveness of each component in our method is separately validated. Based on the baseline, we gradually introduce these components: metric-adaptive thresholding (MAT, in Section 3.3), correlative/discriminative features decoupling (CDD, in Section 3.2) and generation/utilization of pseudo-labels decoupling (GUD, in Section 3.2). At four different labeled proportions, each component exhibits positive effects.

The Study on Metric Function. Figures 7 and 8 present the analyses of parameters, including the metric function $\mathcal{M}(\cdot, \cdot)$ and the value β in metric F_β, across three datasets. For VOC, the three metric functions perform comparably across the four labeled proportions and the insensitivity of β in F_β remains consistent with COCO. For NUS, choosing F_β appears to be a more suitable metric. However, in cases of low labeled proportions, a higher value of β needs to be selected as performance tends to increase with the increase in β.

Table 6: Average per-class F1 (CF1 %) score of each compared method. Bold represents the highest CF1. LL-* and Top-* select the best-performing method from their respective categories. The detailed method descriptions can be found in 4.1.

CF1 score (%) on VOC.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	19.29	36.42	37.78	40.62	36.93	38.91	**45.60**	44.20	36.15	44.54	41.15
$p = 0.05$	54.00	59.76	62.00	62.20	63.33	60.18	60.80	61.18	52.99	**69.86**	68.50
$p = 0.10$	62.86	62.70	65.75	66.81	67.24	65.18	64.58	65.93	62.41	75.63	**75.94**
$p = 0.15$	64.14	66.17	67.40	67.37	67.68	66.04	66.38	66.72	63.10	77.09	**77.14**
$p = 0.20$	63.96	64.47	67.45	66.80	67.71	66.97	66.34	67.56	63.35	77.88	**79.37**

CF1 score (%) on COCO.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	41.19	41.08	42.80	44.36	47.19	42.03	44.42	43.28	33.53	52.70	**55.32**
$p = 0.05$	51.52	51.05	53.38	53.20	51.69	51.67	53.20	51.49	46.81	60.66	**65.65**
$p = 0.10$	54.11	53.46	56.59	55.92	54.97	55.40	56.17	54.17	48.71	64.11	**68.70**
$p = 0.15$	55.48	55.01	57.75	57.99	56.60	56.55	57.65	55.66	49.89	65.40	**69.83**
$p = 0.20$	56.44	55.78	58.72	59.07	57.63	57.51	58.40	56.72	51.08	66.30	**70.74**

CF1 score (%) on NUS.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	23.68	23.22	20.30	23.42	22.39	22.78	30.19	26.80	16.36	28.81	**42.17**
$p = 0.05$	33.57	32.83	31.70	31.75	34.57	32.10	36.29	33.24	25.48	47.14	**47.47**
$p = 0.10$	36.75	34.35	35.17	34.96	37.24	34.84	38.20	36.15	28.05	49.94	**50.77**
$p = 0.15$	38.33	35.38	35.81	36.61	38.55	35.94	37.79	37.34	28.95	51.14	**51.35**
$p = 0.20$	39.59	36.47	37.21	38.27	39.70	37.16	38.68	38.65	30.31	**52.37**	52.31

Parameter Sensitivity Analyses. In Figures 9 and 10, we demonstrate the performance variation with the parameters n and α within the range $\{2 \times 2, 3 \times 3, 4 \times 4\}$ and $\{0.1, 0.5, 1.0, 1.5, 2.0\}$, respectively. For the sake of presentation, we include figures from the main text where $p = 0.05$ alongside figures with $p = \{0.1, 0.15, 0.2\}$ that were not previously displayed. For parameter n, considering all three datasets, we recommend using $n = 2 \times 2$ for cropping since it not only saves some computational costs but also achieves decent performance. For parameter α, our method is generally insensitive to it. So, we use $\alpha = 1.0$ in all experiments for simplicity.

F More Results of Additional Evaluation Metrics

In Tables 6 and 7, we present additional comparative experimental results that were not reported in the main text. This includes the results of two newly introduced metrics, average per-class F1 score (CF1) and overall F1 score (OF1),

Table 7: Overall F1 (OF1 %) score of each compared method. Bold represents the highest OF1. LL-* and Top-* select the best-performing method from their respective categories. The detailed method descriptions can be found in 4.1.

OF1 score (%) on VOC.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	31.55	43.63	42.33	41.93	43.21	45.48	**52.79**	49.93	46.12	33.57	35.63
$p = 0.05$	60.63	63.47	65.17	64.46	65.16	63.95	64.69	64.99	56.40	**73.98**	72.23
$p = 0.10$	65.36	66.11	68.11	67.73	68.00	67.63	67.54	67.92	61.01	78.39	**79.38**
$p = 0.15$	66.34	66.83	69.07	68.95	68.94	68.55	68.84	68.77	62.25	79.83	**80.54**
$p = 0.20$	66.95	67.25	69.37	69.04	69.22	69.12	69.13	69.42	63.55	80.80	**82.44**

OF1 score (%) on COCO.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	49.72	50.28	51.20	51.91	53.64	51.55	51.76	52.00	45.76	59.99	**62.67**
$p = 0.05$	57.47	57.37	58.96	58.46	57.68	58.37	58.16	57.88	52.88	66.09	**70.88**
$p = 0.10$	59.68	59.67	60.87	60.82	60.12	60.90	60.47	60.14	54.59	68.85	**73.50**
$p = 0.15$	60.76	60.98	61.84	62.22	61.33	62.12	61.64	61.33	55.66	69.94	**74.33**
$p = 0.20$	61.48	61.58	62.51	63.01	62.17	62.73	62.32	62.01	55.89	70.71	**75.10**

OF1 score (%) on NUS.

Method	BCE	ASL	LL-*	PLC	Top-*	IAT	ADSH	FM	DRML	CAP	Ours
$p = 0.01$	47.30	46.89	36.14	48.24	39.03	44.95	47.84	**48.47**	42.38	35.26	40.65
$p = 0.05$	50.05	50.45	50.50	51.28	50.77	50.46	50.94	50.72	46.93	**66.92**	66.23
$p = 0.10$	50.99	51.36	51.48	52.17	51.75	51.45	51.86	51.68	48.07	68.09	**68.50**
$p = 0.15$	51.58	51.95	51.99	52.59	52.17	52.01	52.37	52.06	48.72	68.62	**68.74**
$p = 0.20$	51.72	52.22	52.30	52.91	52.37	52.36	52.67	52.40	49.06	**69.23**	69.15

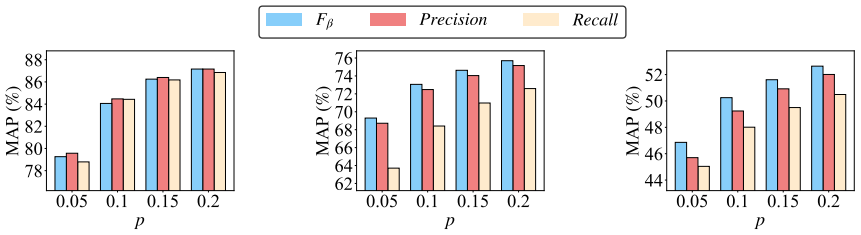
across three datasets and five annotation ratios. Specifically, these two metrics can be computed as follow:

$$\mathrm{CF1} = \frac{2 \times \mathrm{CP} \times \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}}, \quad \mathrm{OF1} = \frac{2 \times \mathrm{OP} \times \mathrm{OR}}{\mathrm{OP} + \mathrm{OR}},$$

and

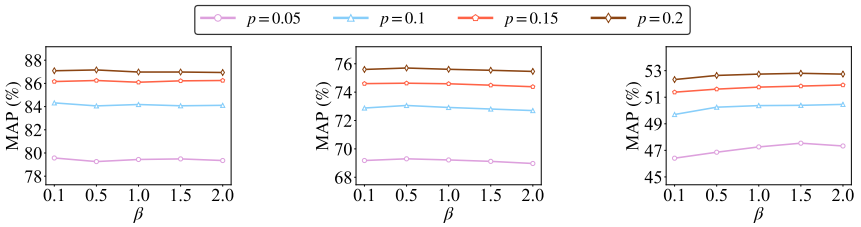
$$\mathrm{CP} = \frac{1}{K} \sum_k \frac{N_k^{TP}}{N_k^{TP} + N_k^{FP}}, \quad \mathrm{OP} = \frac{\sum_k N_k^{TP}}{\sum_k (N_k^{TP} + N_k^{FP})},$$
$$\mathrm{CR} = \frac{1}{K} \sum_k \frac{N_k^{TP}}{N_k^{TP} + N_k^{FN}}, \quad \mathrm{OR} = \frac{\sum_k N_k^{TP}}{\sum_k (N_k^{TP} + N_k^{FN})},$$

where CP, CR are average per-class precision, recall, and OP, OR are overall precision, recall. According to the confusion matrix, $\{N_k^{TP}, N_k^{FP}, N_k^{TN}, N_k^{FN}\}$ indicate the number of true positive, false positive, true negative, false negative for the k-th class. The superiority of our approach is validated by these experimental results.



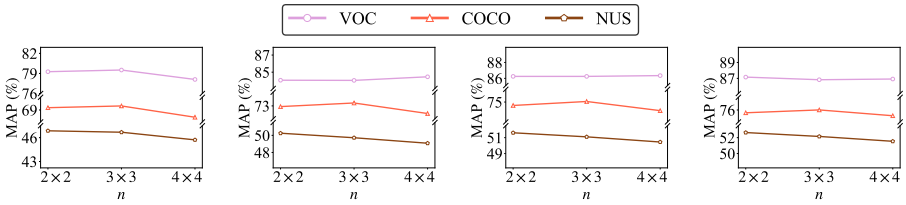
(a) VOC. **(b)** COCO. **(c)** NUS.

Fig. 7: The analyses of metric function $\mathcal{M}(\cdot,\cdot)$ in MAT on three datasets.



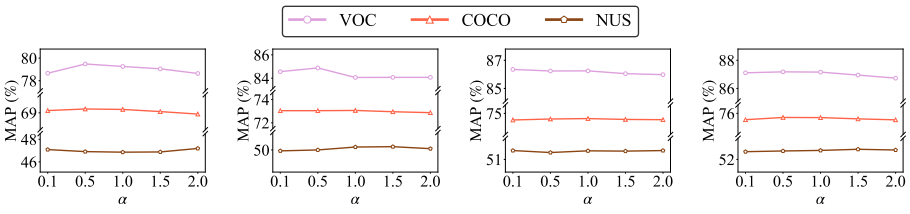
(a) VOC. **(b)** COCO. **(c)** NUS.

Fig. 8: The analyses of value β in metric function F_β on three datasets.



(a) $p = 0.05$. **(b)** $p = 0.1$. **(c)** $p = 0.15$. **(d)** $p = 0.2$.

Fig. 9: The analyses of number of patches n on three datasets.



(a) $p = 0.05$. **(b)** $p = 0.1$. **(c)** $p = 0.15$. **(d)** $p = 0.2$.

Fig. 10: The analyses of temperature α on three datasets.