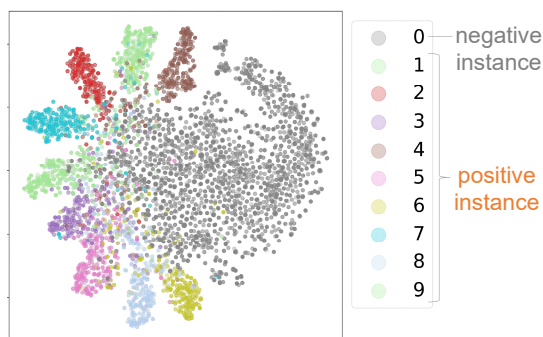
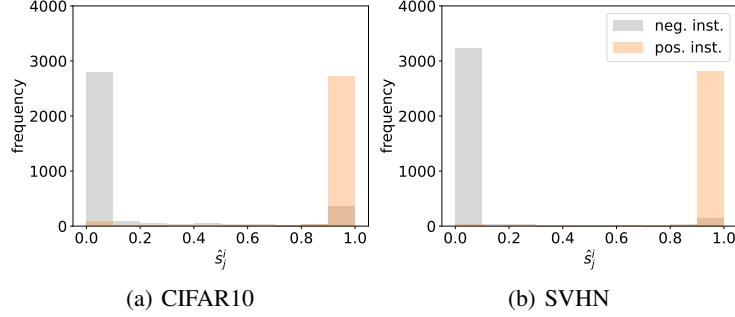


**Table A.** Experimental results on CIFAR10 and SVHN which have widely been used in previous papers on LLP. Despite the challenging LPLP scenario, our method achieved a competitive performance with the oracles (SL and LLP) results, which used more informative labels as training data. Our method outperformed the baseline methods (PPL and Two-stage) in the same scenario (LPLP). The performance of PPL was the worst. We consider that the extension of the conventional proportion loss is insufficient for LPLP. The comparison with Two-stage shows the effectiveness of our joint learning of MIL and LLP.

Setting	Given label	Method	CIFAR10		SVHN	
			Acc.[%] $\uparrow$	mIoU[%] $\uparrow$	Acc.[%] $\uparrow$	mIoU[%] $\uparrow$
SL	Instance label $y$	CE	80.44 $\pm$ 0.84	53.31 $\pm$ 1.32	88.91 $\pm$ 2.03	70.42 $\pm$ 4.24
LLP	(Comp.) label proportion	PL	78.78 $\pm$ 1.30	49.04 $\pm$ 3.17	89.32 $\pm$ 1.27	69.48 $\pm$ 2.97
LPLP	Partial label proportion	PPL	70.72 $\pm$ 0.91	34.44 $\pm$ 1.58	80.66 $\pm$ 4.37	53.51 $\pm$ 9.23
		Two-stage	76.76 $\pm$ 1.51	46.54 $\pm$ 2.71	85.25 $\pm$ 2.94	61.40 $\pm$ 6.66
		Ours	<b>77.04<math>\pm</math>0.64</b>	<b>49.18<math>\pm</math>1.02</b>	<b>88.05<math>\pm</math>0.83</b>	<b>68.82<math>\pm</math>2.02</b>



**Fig. A.** Feature distributions of the test samples from SVHN by t-SNE. Each color indicates each class, where gray shows the negative samples. The negative instances (gray) were successively separated from the positive instances (non-gray). This indicates that our method successfully trained the MIL classifier. Furthermore, the feature distributions of sub-classes of the positive class are separated successfully. It indicates the LLP module after MIL works well.



**Fig. B.** Histogram of the estimated positive score  $\hat{s}_i^j$  by MIL in CIFAR10 and SVHN. In the results, almost all instances are successfully classified. In addition, the estimated score tends to take 0 or 1 after training, although the score can take a value from 0 to 1  $\hat{s}_j^i \in [0, 1]$  to give weight for each instance during training. It indicates that the soft mask can work as a hard mask in the end of training.

**Table B.** Experimental results when changing the bag size (32, 64, 128) while the number of bags was fixed. Our method was the best in all bag sizes. It shows the robustness of our method for the bag size.

Bag size		32		64		128	
Setting	Method	Acc.[%] $\uparrow$	mIoU[%] $\uparrow$	Acc.[%] $\uparrow$	mIoU[%] $\uparrow$	Acc.[%] $\uparrow$	mIoU[%] $\uparrow$
SL	CE	80.44 $\pm$ 0.84	53.31 $\pm$ 1.32	80.29 $\pm$ 0.51	54.92 $\pm$ 1.16	81.17 $\pm$ 0.64	57.69 $\pm$ 0.87
LLP	PL	78.78 $\pm$ 1.30	49.04 $\pm$ 3.17	79.39 $\pm$ 0.97	50.23 $\pm$ 1.60	81.08 $\pm$ 0.73	53.61 $\pm$ 1.72
LPLP	PPL	70.72 $\pm$ 0.91	34.44 $\pm$ 1.58	74.12 $\pm$ 0.83	41.56 $\pm$ 1.38	75.21 $\pm$ 1.10	44.36 $\pm$ 2.05
	Two-stage	76.76 $\pm$ 1.51	46.54 $\pm$ 2.71	78.19 $\pm$ 1.05	49.34 $\pm$ 1.86	79.00 $\pm$ 0.81	52.56 $\pm$ 1.20
	Ours	<b>77.04<math>\pm</math>0.64</b>	<b>49.18<math>\pm</math>1.02</b>	<b>78.95<math>\pm</math>1.21</b>	<b>52.45<math>\pm</math>1.29</b>	<b>80.47<math>\pm</math>0.72</b>	<b>55.13<math>\pm</math>0.76</b>

**Table C.** Experimental results when using a different LLP backbone, LLP-VAT, instead of PL on CIFAR10 to show the extension ability of our method. In this experiment, LLP-VAT is an oracle that uses the complete label proportion. The backbone method for LLP was changed to LLP-VAT in two baseline methods and our method. Our method outperformed the comparative baseline methods and achieved comparative performances with the oracle. Any LLP methods can be applied to our method.

Method	Acc.[%] $\uparrow$	mIoU[%] $\uparrow$
LLP-VAT	78.81 $\pm$ 1.53	48.73 $\pm$ 3.24
PPL (w/ LLP-VAT)	71.08 $\pm$ 0.72	34.92 $\pm$ 1.56
Two-stage (w/ LLP-VAT)	77.39 $\pm$ 1.02	46.50 $\pm$ 2.06
Ours (w/ LLP-VAT)	<b>77.68<math>\pm</math>0.87</b>	<b>49.21<math>\pm</math>0.78</b>