A Supplementary: Introduction

In this supplementary material, we provide more details regarding implementation details in Appendix B, more analysis of ERDA in Appendix C, full experimental results in Appendix D, and studies on parameters in Appendix E.

B Supplementary: Implementation and Training Details

For the RandLA-Net [34] and CloserLook3D [51] baselines, we follow the instructions in their released code for training and evaluation, which are here (RandLA-Net) and here (CloserLook3D), respectively. Especially, in CloserLook3D[51], there are several local aggregation operations and we use the "Pseudo Grid" (KPConv-like) one, which provides a neat re-implementation of the popular KPConv [68] network (rigid version). For point transformer (PT) [97], we follow their paper and the instructions in the code base that claims to have the official code (here). For FixMatch [64], we use the publicly available implementation here.

Our code and pre-trained models will be released.

C Supplementary: Delving into ERDA with More Analysis

Following the discussion in Sec. 3, we study the impact of entropy regularization as well as different distance measurements from the perspective of gradient updates.

In particular, we study the gradient on the score of the i-th class *i.e.,* s_i, and denote it as $g_i = \frac{\partial L_p}{\partial s_i}$. Given that $\frac{\partial p_j}{\partial s_i} = \mathbb{1}_{(i=j)} p_i - p_i p_j$, we have $g_i = p_i \sum_j p_j (\frac{\partial L_p}{\partial p_i} - \frac{\partial L_p}{\partial p_j})$. As shown in Tab. 8, we demonstrate the gradient update $\Delta = -g_i$ under different situations.

In addition to the analysis in Sec. 3.2, we find that, when $\mathbf{q}$ is certain, *i.e.,* $\mathbf{q}$ approaching a one-hot vector, the update of our choice $KL(\mathbf{p}||\mathbf{q})$ would approach the infinity. We note that this could be hardly encountered since $\mathbf{q}$ is typically also the output of a softmax function. Instead, we would rather regard it as a benefit because it would generate effective supervision on those model predictions with high certainty, and the problem of gradient explosion could also be prevented by common operations such as gradient clipping.

In Fig. 4, we provide visualization for a more intuitive understanding on the impact of different formulations for L_{DA} as well as their combination with L_{ER}. Specifically, we consider a simplified case of binary classification and visualize their gradient updates when λ takes different values. We also visualize the gradient updates of L_{ER}. By comparing the gradient updates, we observe that only $KL(\mathbf{p}||\mathbf{q})$ with $\lambda = 1$ can achieve small updates when $\mathbf{q}$ is close to uniform ($q = \frac{1}{2}$ under the binary case), and that there is a close-0 plateau as indicated by the sparse contour lines.

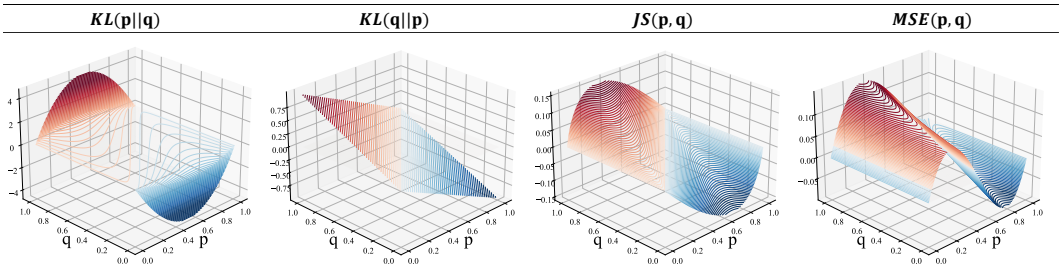
Additionally, we also find that, when increasing the λ, all distances, except the $KL(\mathbf{p}||\mathbf{q})$, are modulated to be similar to the updates of having L_{ER} alone; whereas $KL(\mathbf{p}||\mathbf{q})$ can still produce effective updates, which may indicate that $KL(\mathbf{p}||\mathbf{q})$ is more robust to the λ.

D Supplementary: Full Results

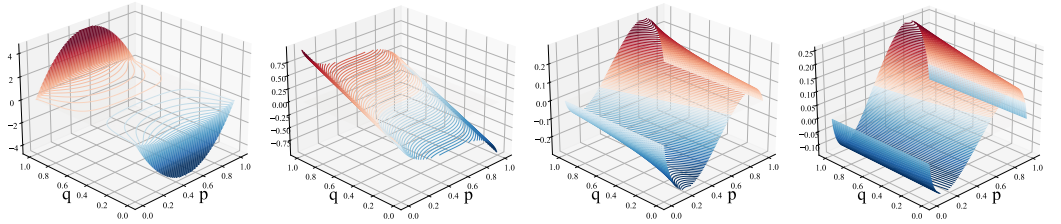
We provide full results for the experiments reported in the main paper. For S3DIS [2], we provide the full results of S3DIS with 6-fold cross-validation in Tab. 10. For ScanNet [16] and SensatUrban [33], we evaluate on their online test servers, which are here and here, and provide the full results in Tab. 11 and Tab. 12, respectively.

L_{DA}	$KL(\mathbf{p}\|\|\mathbf{q})$	$KL(\mathbf{q}\|\|\mathbf{p})$	$JS(\mathbf{p},\mathbf{q})$	$MSE(\mathbf{p},\mathbf{q})$
L_p	$H(\mathbf{p},\mathbf{q}) - (1-\lambda)H(\mathbf{p})$	$H(\mathbf{q},\mathbf{p}) - H(\mathbf{q}) + \lambda H(\mathbf{p})$	$H(\frac{\mathbf{p}+\mathbf{q}}{2}) - (\frac{1}{2}-\lambda)H(\mathbf{p}) - \frac{1}{2}H(\mathbf{q})$	$\frac{1}{2}\sum_i (p_i - q_i)^2 + \lambda H(\mathbf{p})$
g_i	$p_i \sum_j p_j(-\log\frac{q_i}{q_j} + (1-\lambda)\log\frac{p_i}{p_j})$	$p_i - q_i - \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$	$p_i \sum_j p_j(\frac{-1}{2}\log\frac{p_i+q_i}{p_j+q_j} + (\frac{1}{2}-\lambda)\log\frac{p_i}{p_j})$	$p_i(p_i - q_i) - p_i \sum_j p_j(p_j - q_j) - \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$
Situations	$\Delta = -g_i$			
$p_k \to 1$	0	$q_i - \mathbb{1}_{k=i}$	0	0
$q_1 = ... = q_K$	$(\lambda-1)p_i \sum_j p_j \log\frac{p_i}{p_j}$	$\frac{1}{K} - p_i + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$	$p_i \sum_{j\neq i} p_j(\frac{1}{2}\log\frac{Kp_i+1}{Kp_j+1} + (\lambda-\frac{1}{2})\log\frac{p_i}{p_j})$	$-p_i^2 + p_i \sum_j p_j^2 + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$
$q_i \to 1$	$+\inf$	$1 - p_i + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$	$p_i \sum_{j\neq i} p_j(\frac{1}{2}\log\frac{p_i+1}{p_j} + (\lambda-\frac{1}{2})\log\frac{p_i}{p_j})$	$-p_i^2 + p_i(1-p_i) + p_i \sum_j p_j^2 + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$
$q_{k\neq i} \to 1$	$-\inf$	$-p_i + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$	$p_i \sum_{j\neq i} p_j(\frac{1}{2}\log\frac{p_i}{p_j+\mathbb{1}_{j=k}} + (\lambda-\frac{1}{2})\log\frac{p_i}{p_j})$	$-p_i^2 - p_i p_k + p_i \sum_j p_j^2 + \lambda p_i \sum_j p_j \log\frac{p_i}{p_j}$

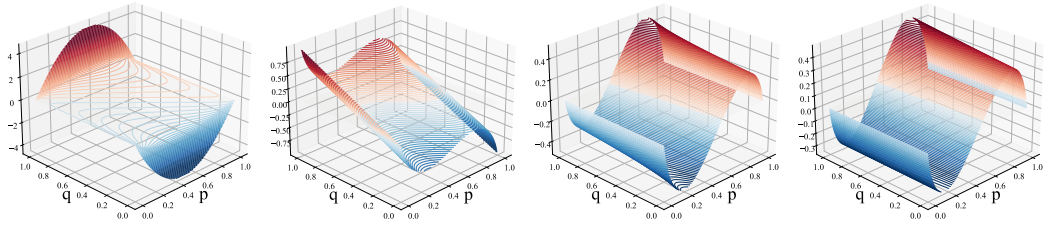
Table 8. The formulation of L_p using different functions to formulate L_{DA}. We present the gradients $g_i = \frac{\partial L_p}{\partial s_i}$, and the corresponding update $\Delta = -g_i$ under different situations. Analysis can be found in the Sec. 3.2 and Appendix C.



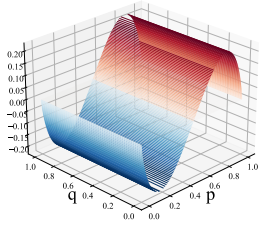
(a) $\lambda = 0$



(b) $\lambda = 1$



(c) $\lambda = 2$



(d) Gradient updates of L_{ER}

Figure 4. Contour visualization of the gradient update with binary classes for better understanding. For a clearer view, we use red for positive updates and blue for negative updates, the darker indicates larger absolute values and the lighter indicates smaller absolute values.

m	mIoU
0.9	66.19
0.99	66.80
0.999	67.18
0.9999	66.22

(a) **Momentum update**.

projection	mIoU
-	65.90
linear	66.55
2-layer MLPs	67.18
3-layer MLPs	66.31

(b) **Projection network**.

α	mIoU
0.001	65.25
0.01	66.01
0.1	67.18
1	65.95

(c) **Loss weight**.

Table 9. Parameter study on ERDA. If not specified, the model is RandLA-Net with ERDA trained with loss weight $\alpha = 0.1$, momentum $m = 0.999$, and 2-layer MLPs as projection networks under 1% setting on S3DIS. Default settings are marked in gray.

E Supplementary: Ablations and Parameter Study

We further study the hyper-parameters involved in the implementation of ERDA with the prototypical pseudo-label generation, including loss weight α, momentum coefficient m, and the use of projection network. As shown in Tab. 9, the proposed method acquires decent performance (mIoUs are all > 65 and mostly > 66) in a wide range of different hyper-parameter settings, compared with its fully-supervised baseline (64.7 mIoU) and previous state-of-the-art performance (65.3 mIoU by HybridCR [47]).

Additionally, we suggest that the projection network could be effective in facilitating the ERDA learning, which can be learned to specialize in the pseudo-label generation task. This could also be related to the advances in contrastive learning. Many works [10, 11, 25] suggest that a further projection on feature representation can largely boost the performance because such projection decouples the learned features from the pretext task. We share a similar motivation in decoupling the features for ERDA learning on the pseudo-label generation task from the features for the segmentation task.

settings	methods	mIoU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
Fully	RandLA-Net + ERDA	71.0	94.0	96.1	83.7	59.2	48.3	62.7	73.6	65.6	78.6	71.5	66.8	65.4	57.9
	CloserLook3D + ERDA	73.7	94.1	93.6	85.8	65.5	50.2	58.7	79.2	71.8	79.6	74.8	73.0	72.0	59.5
	PT + ERDA	76.3	94.9	97.8	86.2	65.4	55.2	64.1	80.9	84.8	79.3	74.0	74.0	69.3	66.2
1%	RandLA-Net + ERDA	69.4	93.8	92.5	81.7	60.9	43.0	60.6	70.8	65.1	76.4	71.1	65.3	65.3	55.0
	CloserLook3D + ERDA	72.3	94.2	97.5	84.1	62.9	46.2	59.2	73.0	71.5	77.0	73.6	71.0	67.7	61.2
	PT + ERDA	73.5	94.9	97.7	85.3	66.7	53.2	60.9	80.8	69.2	78.4	73.3	67.7	65.9	62.1

Table 10. The full results of ERDA with different baselines on S3DIS 6-fold cross-validation.

settings	methods	mIoU	bathtub	bed	books.	cabinet	chair	counter	curtain	desk	door	floor	other	pic	fridge	shower	sink	sofa	table	toilet	wall	wndw
Fully	CloserLook3D + ERDA	70.4	75.9	76.2	77.0	68.2	84.3	48.1	81.3	62.1	61.4	94.7	52.7	19.9	57.1	88.0	75.9	79.9	64.7	89.2	84.2	66.6
20pts	CloserLook3D + ERDA	57.0	75.1	62.5	63.1	46.0	77.7	30.0	64.9	46.1	43.6	93.3	36.0	15.4	38.0	73.6	51.6	69.5	47.2	83.2	74.5	47.8
0.1%	RandLA-Net + ERDA	62.0	75.7	72.4	67.9	56.9	79.0	31.8	73.0	58.1	47.3	94.1	47.1	15.2	46.3	69.2	51.8	72.8	56.5	83.2	79.2	62.0
1%	RandLA-Net + ERDA	63.0	63.2	73.1	66.5	60.5	80.4	40.9	72.9	58.5	42.4	94.3	50.0	35.0	53.0	57.0	60.4	75.6	61.9	78.8	73.8	62.6

Table 11. The full results of ERDA with different baselines on ScanNet [16] test set, obtained from its online benchmark site by the time of submission.

settings	methods	mIoU	OA	Ground	Vegetation	Buildings	Walls	Bridge	Parking	Rail	Roads	Street Furniture	Cars	Footpath	Bikes	Water
Fully	RandLA-Net + ERDA	64.7	93.1	86.1	98.1	95.2	64.7	66.9	59.6	49.2	62.5	46.5	85.8	45.1	0.0	81.5
0.1%	RandLA-Net + ERDA	56.4	91.1	82.0	97.4	93.2	56.4	57.1	53.1	5.2	60.0	33.6	81.2	39.9	0.0	74.2

Table 12. The full results of ERDA with different baselines on SensatUrban [33] test set, obtained from its online benchmark site by the time of submission.