In2NeCT: Inter-class and Intra-class Neural Collapse Tuning for Semantic Segmentation of Imbalanced Remote Sensing Images

Junao Shen[1, 2], Qiyun Hu[1], Tian Feng[1, 2*], Xinyu Wang[1], Hui Cui[3], Sensen Wu[4], Wei Zhang[1, 5]

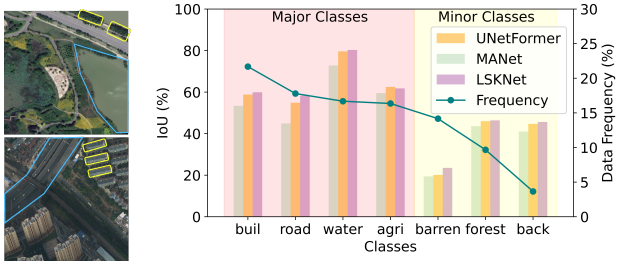
[1]School of Software Technology, Zhejiang University
[2]State Key Lab of CAD&CG, Zhejiang University
[3]Department of Computer Science and Information Technology, La Trobe University, Australia
[4]School of Earth Sciences, Zhejiang University
[5]Innovation Center of Yangtze River Delta, Zhejiang University
{jashen, qiyunhu, t.feng, xinyu.w, t.feng, sensenwu, cstzhangwei}@zju.edu.cn, huicui@latrobe.edu.au

Abstract

Remote sensing images (RSIs) are frequently characterized by multi-scale inter-class objects and inconsistently distributed objects due to scene limitations, which would cause a significant data imbalance challenging the corresponding semantic segmentation. Recent methods have leveraged various deep learning techniques to capture high-quality representations for RSI semantic segmentation, but are hardly capable of addressing the afore-mentioned challenge given their limited explorations toward the mechanisms behind the representations. The recently discovered Neural Collapse (NC) phenomenon in computer vision models suggests the simplex equiangular tight frame (ETF) as the optimal representation structure, which has motivated us to observe that the optimal structure of last-layer representations is disrupted and inter-class representations for minor classes tend to become closer to each other beacuse of data imbalance. To address these issues, we propose Inter-class and Intra-class Neural Collapse Tuning (**In2NeCT**) to optimize the representations that satisfy the simplex ETF, which facilitates the discrimination of inter-class representations and the coherence of intra-class representations. Extensive experiments on three datasets demonstrate that our In2NeCT consistently leads to significant improvements in performance and outperforms the state-of-the-art methods.

Introduction

Semantic segmentation is a fundamental computer vision task and significantly contributes to various analytical applications for remote sensing images (RSIs), including land use classification (Digra, Dhir, and Sharma 2022), agricultural production estimation (Luo et al. 2023), and environmental monitoring (Yuan et al. 2020). As shown in Figure 1 (a), RSIs exhibit distinct characteristics compared to other images: (1) RSIs usually exhibit a multi-scale phenomenon in inter-class objects (blue and yellow); (2) The distribution of objects in each class is inconsistent due to scene limitations (*i.e.*, LoveDA focuses on urban planning with less barren and forest objects in the RSIs). These traits lead to an intrinsic data imbalance and thus would cause a decrease in a method's performance regarding the minor classes with limited samples, as illustrated in Figure 1 (b).



(a) (b)

Figure 1: Challenges in RSI semantic segmentation on LoveDA: (a) Multi-scale inter-class objects, as depicted by the bounding boxes in blue and yellow (b) Performance in Intersection over Union (IoU) decreases significantly regarding the minor classes.

Since the essence of RSI semantic segmentation lies in learning textural and contextual representations for pixel-level classification, conventional methods have adopted techniques on spatial and relational contexts to capture detailed information for segmentation. Nevertheless, these methods are limited in revealing the mechanisms behind the representations, which drives us to raise the following two questions: *(1) What factor(s) cause the decrease in performance regarding the minor classes at the level of representations? (2) How to obtain effective representations for RSI semantic segmentation?*

Recent studies have proposed a phenomenon, namely Neural Collapse (NC) (Papyan, Han, and Donoho 2020; Yang et al. 2022, 2023) in computer vision classification models, which provides useful information on the optimal structure of visual feature representations. Specifically, this phenomenon demonstrates that the last-layer representations within the same class would collapse to their intra-class means near the end of the training phase (*i.e.*, the training loss reaches zero) on a sufficiently large and balanced dataset. Besides, both the intra-class means and their corresponding classifier vectors converge to the vertices of a

*Corresponding author (t.feng@zju.edu.cn).

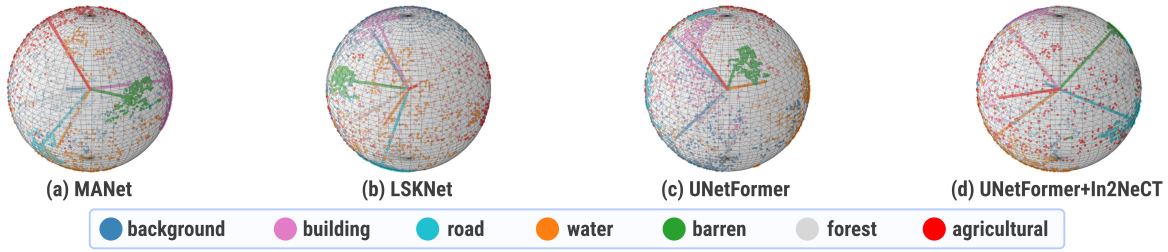


Figure 2: Visualization of the NC phenomenon on the LoveDA. The Arrows indicate classifier vectors, and points represent last-layer representations. (a), (b), and (c) The optimal structure in MANet, LSKNet and UNetFormer is disrupted, resulting in a performance decrease on the imbalanced dataset. (b) Our In2NeCT refines the structure and thus improves the performance.

simplex equiangular tight frame (ETF). The simplex ETF depicts the geometric structure of several vectors that have maximal pairwise angles and equal norms.

The NC phenomenon inspires us to conduct a more thorough examination of the geometry structures of last-layer representations. As shown in Figure 2 (a), (b), and (c), we observe that the optimal structures (*i.e.*, the simplex ETF) in the state-of-the-art MANet (Li et al. 2021), LSKNet (Li and et.al. 2023), and UNetFormer are disrupted. Additionally, inter-class representations for minor classes tend to become closer to each other (*e.g.*, barren and forest), while intra-class representations are more separate. Furthermore, with the results in Figure 1 (b), a negative correlation exists between the strength of NC and performance decrease. Based on these findings, we assume that *a stronger NC of representations associates with a higher performance for RSI semantic segmentation*, which answers to our *question (1)*.

Following our assumption, we propose a novel **In**tra-class and **In**ter-class **Ne**ural **C**ollapse **T**uning (**In2NeCT**) to address the above-discussed challenge of optimal structure disruption for semantic segmentation of imbalanced RSIs. Specifically, to accomplish *question 2*, In2NeCT aims to optimize the representations through two additional branches with explicit constraints to satisfy the simplex ETF. In particular, we introduce a fixed simplex ETF classifier and inter-class regularization loss to optimize the inter-class representations, as well as a balanced intra-class regularization loss to optimize the intra-class representations. As illustrated in Figure 2 (b), our In2NeCT facilitates the representations to evolve toward the simplex ETF for improved performance on minor classes. Meanwhile, we provide a corresponding in-depth theoretical analysis of our In2NeCT. Experiments on three datasets demonstrate that our In2NeCT can outperform the state-of-the-art methods in mean intersection over union (mIoU), average F1 score (AF), and overall accuracy (OA). Our contributions can be summarized as follows:

- For the first time, we extend the NC phenomenon to RSI semantic segmentation and explain the reason behind the limited performances on imbalanced datasets from a perspective of representations;
- We propose an Intra-class and Inter-class Neural Collapse Tuning method for optimal representations to satisfy the simplex ETF;
- Theoretical analysis and extensive experiments demonstrate the effectiveness of the proposed method, which can be integrated with conventional methods and achieve state-of-the-art performances on three datasets.

Related Work

RSI Semantic Segmentation

The fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015) is regarded as the pioneer in the field of semantic segmentation for introducing full convolution and formulating the task as per-pixel classification in an end-to-end manner. Subsequently, numerous advanced methods have been developed to leverage the comprehensive background and spatial information in remote sensing image (RSI) semantic segmentation. Conventional methods have employed techniques in spatial and relational contexts to capture detailed information for segmentation. Specifically, spatial context methods, such as PSPNet (Zhao et al. 2017) and DeepLabv3+ (Chen et al. 2018), employ spatial pyramid pooling (SPP) or atrous spatial pyramid pooling (ASPP) for spatial context integration; Relational context methods, such as MANet (Li et al. 2021) and DANet (Fu et al. 2019), adopt spatial attention and channel attention for selective aggregation. Recently, OCRNet (Yuan, Chen, and Wang 2020) merges class-wise contexts by capturing the global class representations, and UNetFormer (Wang et al. 2022b) and LSKNet (Li and et.al. 2023) refines relational context by CNN-Transformer fusion and large-scale kernels. Nevertheless, these methods are limited in revealing the mechanisms behind the representations.

Neural Collapse

The neural collapse phenomenon was first observed by Papyan et al (Papyan, Han, and Donoho 2020) that the last-layer features of a classification model will collapse into a within-class center at the terminal phase of training on a balanced dataset. This elegant geometric structure is named simplex equiangular tight frame (ETF). Since then, later researchers have endeavored to explain this phenomenon theoretically (Yang et al. 2022, 2023; Weinan and Wojtowytsch 2022). It proved that neural collapse is the global optimality of a classification model with regularization constraint under the CE (Zhang and Sabuncu 2018) and MSE loss functions (Sara, Akter, and Uddin 2019). Recent studies have attempted to induce neural collapse in imbalanced datasets,

and incremental learning. However, these researches into neural collapse are limited in image-level recognition. In this work, we explore the presence of such a solution in RSI semantic segmentation.

Preliminaries

The NC phenomenon (Papyan, Han, and Donoho 2020; Yang et al. 2022) suggests that, at the terminal training phase, last-layer representations would converge to their intra-class means, which would collapse to the vertices of a simplex equiangular tight frame (ETF) together with the linear classifier vectors, leading to the following definitions required in this study:

Simplex ETF A set of vectors $\{m_i \in \mathbb{R}^d | i = 1, \ldots, K, d \geq K-1\}$ is a simplex ETF if satisfying:

$$\mathbf{M} = \sqrt{\frac{K}{K-1}}\mathbf{U}(\mathbf{I}_K - \frac{1}{K}1_K 1_K^\top), \tag{1}$$

where $\mathbf{M} = [m_1, \ldots, m_K] \in \mathbb{R}^{d\times K}$, $\mathbf{U} \in \mathbb{R}^{d\times K}$ allows a rotation and satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_K$, $\mathbf{I}_K$ denotes the identity matrix, and 1_K represents an all-ones vector. All vectors in a simplex ETF have an equal ℓ_2 norm and the same pair-wise angle as follows:

$$m_i^\top m_j = \frac{K}{K-1}\delta_{i,j} - \frac{1}{K-1}, \forall i, j \in [1, K], \tag{2}$$

where $\delta_{i,j}$ equals to 1 when $i = j$ and 0 otherwise. The pair-wise angle $-\frac{1}{K-1}$ is the maximal equiangular separation of K vectors in $\mathbb{R}^d$.

We denote $z_{k,i}$ as the last-layer representation of the i-th sample in the k-th class, $\overline{z}_k = Avg_i\{z_{k,i}\}$ is the mean value of the representations in the k-th class, $z_G = Avg_{k,i}\{k, i\}$ is the global mean of the representations, and w_k is the classifier vector of the k-th class. Based on the description in (Papyan, Han, and Donoho 2020), we summarize three properties by the neural collapse in the vision models as follows:

Intra-Class Representation Collapse ($\mathcal{NC}_1$): The last-layer representations in the k-th class collapse to their prototype (*i.e.*, the k-th intra-class covariance $\Sigma_k := Avg_i\{(z_{k,i} - \overline{z}_k)(z_{k,i} - \overline{z}_k)^\top\} \to 0, \forall k \in [1, K]$).

Inter-Class Representation Collapse ($\mathcal{NC}_2$): The normalized class mean $\overline{z}_k$ converge to a simplex ETF (*i.e.*, $\hat{z}_k = (\overline{z}_k - \overline{z}_G)/\|\overline{z}_k - \overline{z}_G\|, \forall k \in [1, K]$) that satisfies Equation 2.

Classifier Collapse ($\mathcal{NC}_3$): The normalized classifier vectors converge to the same simplex ETF as mean values (*i.e.*, $\hat{w}_k = w_k/\|w_k\| = \overline{z}_k, \forall k \in [1, K]$) that satisfies Equation 2.

Method

Considering the observations illustrated in Figure 1, we propose **In2NeCT**, a novel **In**tra-class and **In**ter-class **Ne**ural **Co**llapse **T**uning method to optimize the process of the RSI semantic segmentation. The proposed method employs two additional regularization branches to address the problems of optimal structure disruption and performance drop related to data imbalance. As shown in Figure 2(b), our In2NeCoT provides the last-layer representations to satisfy the simplex and improves the performance for minor classes as well as the overall performance.

Inter-class and Intra-class Neural Collapse Tuning

As shown in Figure 3, the proposed method comprises three branches on conventional RSI semantic segmentation, inter-class collapse regularization, and intra-class collapse regularization. It is noteworthy that no constraints apply to the exact model used in the RSI semantic segmentation branch. Hence, our In2NeCT has the capability to be inexpensively integrated into any off-the-shelf model for RSI semantic segmentation.

RSI semantic segmentation branch The input image $I \in \mathbb{R}^{H\times W\times 3}$ is processed by a backbone to obtain the representation $\mathbf{Z} \in \mathbb{R}^{H\times W\times d}$. Subsequently, we employ a learnable pixel classifier with an arbitrary form of pixel-wise loss L_{Seg} (*i.e.*, cross-entropy loss) for supervised learning.

Inter-class regularization branch According to the ground truth Y, the representations belonging to the same class $\mathbf{Z_k} = [z_{k,1}, \ldots, z_{k,i}]$ are first gathered to calculate their mean value $\overline{z}_k$ as follows:

$$\overline{z}_k = \frac{1}{n_k}\mathbf{Z_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} z_{k,i}, \tag{3}$$

where n_k is the number of pixels in $\mathbf{Z_k}$. To achieve the properties of $\mathcal{NC}_2$ and $\mathcal{NC}_3$ in Section Preliminaries, we introduce an ETF-structured classifier to this branch. Specifically, the ETF-classifier $\mathbf{W}^* = [\mathbf{w_1}^*, \ldots, \mathbf{w_k}^*]$ is initialized as a simplex ETF following Equation 1. During the training stage, the classifier is fixed (*i.e.*, all parameters are frozen) to enable the maximal equiangular separation (Equation 2) being satisfied as follows:

$$\mathbf{w}_k^{*\top}\mathbf{w}_{k'}^* = \left(\frac{K\delta_{k,k'}}{K-1} - \frac{1}{K-1}\right), \forall k, k' \in \{1, \ldots, K\}, \tag{4}$$

where $\delta_{k,k'} = 1$ when $k = k'$, or otherwise $\delta_{k,k'} = 0$. Afterwards, we adopt a cross entropy-based inter-class regularization loss $\mathcal{L}_{Inter}$ to measure the degree of inter-class collapse on mean values $\overline{\mathbf{Z}} = [\overline{z}_1, \ldots, \overline{z_K}] \in \mathbb{R}^{d\times K}$ as follows:

$$\mathcal{L}_{Inter}\left(\overline{\mathbf{Z}}, \mathbf{W}^*\right) = -\sum_{k=1}^{K}\log\left(\frac{\exp\left(\overline{z}_k^\top \mathbf{w}_k^*\right)}{\sum_{k'=1}^{K}\exp\left(\overline{z}_{k'}^\top \mathbf{w}_{k'}^*\right)}\right). \tag{5}$$

By minimizing the gap between $\overline{z}_k$ with the corresponding fixed w_k^*, it is expected to obtain more discriminative inter-class representations with the simplex ETF.

Intra-class regularization branch Based on the property of $\mathcal{NC}_1$, all intra-class representations would collapse to a single mean value (*i.e.*, $Avg_i\{(z_{k,i} - \overline{z}_k)(z_{k,i} - \overline{z}_k)^\top\} \to 0, \forall k \in [1, K]$). Inspired by contrastive learning (He et al. 2020), we employ the cosine distance with directional information to replace the Euclidean distance as the similarity measure. Considering the data imbalance, we propose a balanced intra-class regularization loss $\mathcal{L}_{Intra}$ via averaging the contribution of different classes (refer Section Theoretical Analysis for details) to measure the degree of intra-class

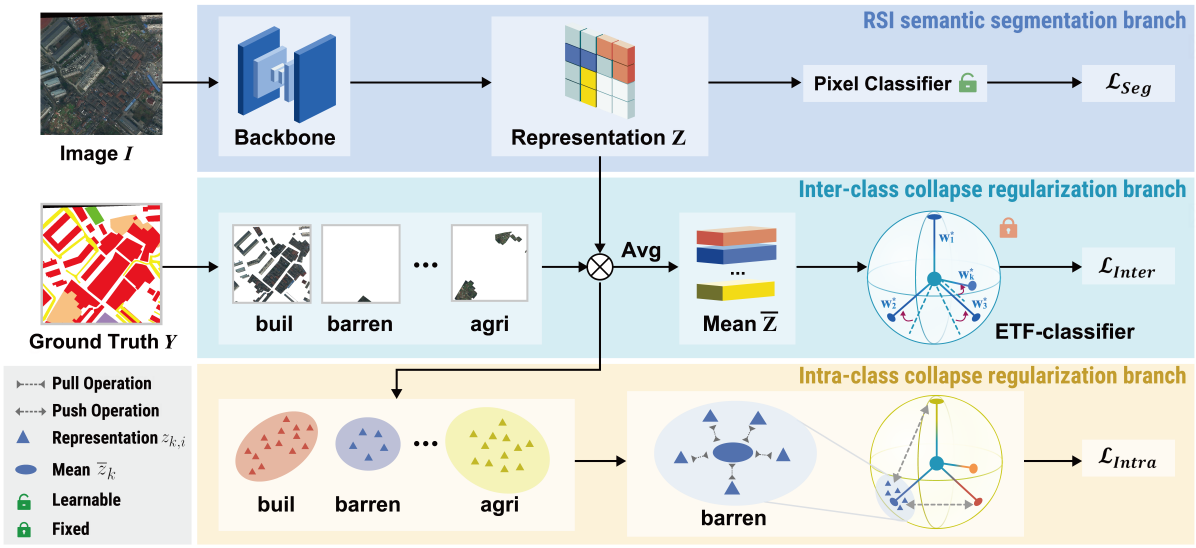


Figure 3: Architecture of the proposed In2NeCT. The RSI semantic segmentation branch integrates the off-the-shelf semantic segmentation model (*i.e.*, UNetFormer). Adopting a fixed ETF-classifier $\mathbf{W}^*$, the inter-class regularization branch employs the loss $\mathcal{L}_{Inter}$ to reach the discrimination of inter-class representations, whereas the intra-class regularization branch adopts the loss $\mathcal{L}_{Intra}$ to ensure the coherence of intra-class representations. The last-layer representations can satisfy the simplex ETF.

collapse upon the representations $\mathbf{Z}$ and their corresponding labels $Y = [1, \ldots, K]$ as follow:

$$\mathcal{L}_{Intra}(\mathbf{Z}) = -\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{1}{n_k - 1} \sum_{i' \neq i}^{n_k} \log\left(\frac{\exp(z_{k,i} z_{k,i'})}{\sum_{k'=1}^{K} \frac{1}{n_{k'}} \sum_{j=1}^{n_{k'}} \exp(z_{k,i} z_{k',j})}\right). \tag{6}$$

Using $\mathcal{L}_{Intra}$ aims to explicitly move the intra-class representations toward each other and the inter-class representations away from each other as much as possible, so as to enable last-layer representations to satisfy the simplex ETF.

Our method incorporates both regularization losses into the loss function of a conventional RSI semantic segmentation method to optimize the segmentation process effectively as follows:

$$\mathcal{L} = \mathcal{L}_{Seg}(\mathbf{Z}, \mathbf{Y}) + w(\mathcal{L}_{Inter}(\overline{\mathbf{Z}}, \mathbf{W}^*) + \mathcal{L}_{Intra}(\mathbf{Z})), \tag{7}$$

where w represents hyperparameters governing the contribution of our In2NeCT. We adopted $w = 0.2$ in all our experiments. It is noteworthy that only the RSI semantic segmentation branch would be preserved in the inference stage, which ensures the efficiency of the proposed method for inference without any additional computation.

Theoretical Analysis

The gradient of $\mathcal{L}_{Inter}$ (Equation 5) *w.r.t* the ETF-classifier $\mathbf{w}_k^*$ and pixel representation $z_{k,i}$ are respectively formulated as follows:

$$\frac{\partial \mathcal{L}_{Inter}}{\partial \mathbf{w}_k^*} = (p(\overline{\mathbf{z}}_k) - 1)\,\overline{z}_k + \sum_{k' \neq k}^{K-1} p(\overline{z}_{k'})\,\overline{z}_{k'},$$
$$= \underbrace{\sum_{i=1}^{n_k} (p(\overline{\mathbf{z}}_k) - 1)\frac{z_i}{n_k}}_{\text{intra-class}} + \underbrace{\sum_{k' \neq k}^{K-1} \sum_{j=1}^{n'_k} p(\overline{z}_{k'})\frac{\mathbf{z}_j}{n_{k'}}}_{\text{inter-class}}, \tag{8}$$

$$\frac{\partial \mathcal{L}_{Inter}}{\partial z_{k,i}} = \frac{\partial \mathcal{L}_{Inter}}{\partial \overline{z}_k}\frac{\partial \overline{z}_k}{\partial z_{k,i}} = \sum_{k=1}^{K} p(\overline{z}_k - 1)(w_{k'}^* - w_k^*)\frac{1}{n_k}, \tag{9}$$

where $p(\overline{z}_k)$ denotes the predict probability of $\overline{z}_k$ to k-th class,

$$p(\overline{z}_k) = \frac{exp(\overline{z}_k^\top w_k^*)}{\sum_{k'=1}^{K} exp(\overline{z}_k^\top w_k^*)}. \tag{10}$$

The gradient *w.r.t* $\mathbf{w}_k^*$ in Equation 8 includes two terms: The *intra-class* term moves $\mathbf{w}_k^*$ toward $\overline{z}_k$ by n_k pixels, and the *inter-class* terms drives $\mathbf{w}_k^*$ against $\overline{z}_k$ by $N - n_k$, where N denotes the number of all pixels. Consequently, the gradients of certain minor classes are overwhelmingly influenced by the intra-class term due to the small n_k and thus the large $N - n_k$. Since the ETF-classifier is fixed, it avoids the imbalanced gradient updating and benefits the performance of minor classes.

The gradient *w.r.t* $z_{k,i}$ in Equation 9 is influenced mostly by $(\mathbf{w}_{k'}^* - \mathbf{w}_k^*)$, which satisfies $\|\mathbf{w}_{k'}^* - \mathbf{w}_k^*\| = 2K/(K-1), \forall k \neq k'$ in our In2NeCT since the ETF-classifier is fixed (Equation 4). It can be interpreted that the fixed ETF-classifier and $\mathcal{L}_{Inter}$ jointly allow the gradi-

Method	LoveDA								Vaihingen			Potsdam		
	back	buil	road	water	barren	forest	agri	mIoU	AF	mIoU	OA	AF	mIoU	OA
MANet (Li et al. 2021)	38.7	51.7	42.6	72.0	15.3	42.1	57.7	45.7	90.41	82.71	90.96	92.90	86.95	91.32
Segmenter (Strudel et al. 2021)	38.0	50.7	48.7	77.4	13.3	43.5	58.2	47.1	88.23	79.44	89.93	92.27	86.48	91.04
LANet (Ding, Tang, and Bruzzone 2021)	40.0	50.6	51.1	78.0	13.0	43.2	56.9	47.6	88.09	79.28	89.83	91.95	85.15	90.84
DeepLabv3+ (Chen et al. 2018)	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6	86.77	77.13	89.12	90.86	84.24	89.18
FarSeg (Zheng et al. 2020)	43.1	51.5	53.9	76.6	9.8	43.3	58.9	48.2	87.88	79.14	89.57	91.21	84.36	89.87
Semantic FPN (Kirillov et al. 2019)	42.9	51.5	53.4	74.7	11.2	44.6	58.7	48.2	87.58	77.94	89.86	91.53	84.57	90.16
PSPNet (Zhao et al. 2017)	44.4	52.1	53.5	76.5	9.7	44.1	57.9	48.3	86.47	76.78	89.36	89.98	81.99	90.14
FLANet (Song et al. 2022)	44.6	51.8	53.0	74.1	15.8	45.8	57.6	49.0	87.44	78.08	89.60	93.12	87.50	91.87
OCRNet (Yuan, Chen, and Wang 2020)	44.2	55.1	53.5	74.3	18.5	43.0	60.5	49.9	89.22	81.71	90.47	92.25	86.14	90.03
SwimUperNet (Liu et al. 2021)	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.0	89.90	81.80	91.00	92.24	86.37	90.98
DANet (Fu et al. 2019)	44.8	55.5	53.0	75.5	17.6	45.1	60.1	50.2	86.88	77.32	89.47	89.60	81.40	89.73
ConvNeXt (Liu et al. 2022)	**46.9**	53.5	56.8	76.1	15.9	47.5	61.8	51.2	90.50	82.87	91.36	93.03	87.17	91.66
ISNet (Jin et al. 2021)	44.4	57.4	58.0	77.5	21.8	43.9	60.6	51.9	90.19	82.36	90.52	92.67	86.58	91.27
BiFormer (Zhu et al. 2023)	43.6	55.3	55.9	79.5	16.9	45.4	61.5	51.2	89.65	81.50	90.63	91.47	84.51	90.17
PoolFormer (Yu et al. 2022)	45.8	57.1	53.3	80.2	19.8	46.1	**64.5**	52.4	89.59	81.35	90.30	92.62	86.45	91.12
MANet (Li et al. 2021)	38.7	51.7	42.6	72.0	15.3	42.1	57.7	45.7	90.41	82.71	90.96	92.90	86.95	91.32
MANet + In2NeCT	41.0	53.4	44.9	72.8	19.4	43.6	59.5	47.8	90.88	83.54	91.36	93.28	88.03	91.72
LSKNet (Li and et.al. 2023)	45.6	59.9	58.3	80.3	23.5	46.4	61.8	53.7	90.77	83.32	91.17	93.12	87.19	92.00
LSKNet + In2NeCT	46.8	**60.8**	**58.8**	**80.7**	**25.5**	47.0	62.5	**54.5**	91.04	**84.61**	**91.92**	**93.44**	**88.10**	**92.53**
UNetFormer (Wang et al. 2022b)	44.7	58.8	54.9	79.6	20.1	46.0	62.5	52.4	90.40	82.70	91.00	92.80	86.80	91.30
UNetFormer + In2NeCT	46.1	**60.8**	57.5	80.5	24.0	**47.7**	63.8	54.3	**91.26**	84.53	91.51	93.21	87.91	91.80
△	+1.4	+2.0	+2.6	+0.9	+3.9	+1.7	+1.3	+1.9	+0.86	+1.83	+0.51	+0.41	+1.11	+0.50

Table 1: Comparison of our In2NeCT and the state-of-the-art methods on LoveDA, ISPRS Vaihingen and ISPRS Potsdam datasets. The best scores are in *bold* and second best underlined. △ denotes the performance improvement over the UNetFormer.

ent transfer and provide the inter-class representations with higher discrimination among minor classes.

The gradient of $\mathcal{L}_{Intra}$ *w.r.t* $z_{i,k}$ is formulated as follows:

$$\frac{\partial \mathcal{L}_{Intra}}{\partial z_{k,i}} = \frac{1}{n_k - 1} \sum_{i' \neq i}^{n_k} (s(z_{k,i}) - 1)(z_{k,i'} - \sum_{k'=1}^{K} \frac{1}{n_{k'}} \sum_{j=1}^{n_{k'}} z_{k',j}), \tag{11}$$

where $s(z_{k,i})$ denotes the similarity of $z_{k,i}$ in the k-th class,

$$s(z_{k,i}) = \frac{\exp\left(z_{k,i} z_{k,i'}\right)}{\sum_{k'=1}^{K} \frac{1}{n_{k'}} \sum_{j=1}^{n_{k'}} \exp\left(z_{k,i} z_{k',j}\right)}. \tag{12}$$

The gradient in Equation 11 is mainly influenced by $z_{k,i'} - \sum_{k'=1}^{K} \frac{1}{n_{k'}} \sum_{j=1}^{n_{k'}} z_{k',j}$. By adopting the class-averaging $(1/n_{k'})$, the contribution of each pixel is balanced by the class frequency, which makes each class equally contribute to the intra-class representation learning in unbalanced data. It suggests that the balanced $\mathcal{L}_{Intra}$ prevents minor classes from being dominated by the pixels of major classes $(\sum_{j=1}^{n_{k'}})$, which allows the balance of each class gradient to reach more coherent intra-class representations.

The above theoretical analysis demonstrates the capability of two regularization branches to make the model conform to the properties of $\mathcal{NC}_1, \mathcal{NC}_2, \mathcal{NC}_3$ in data imbalance, which in turn leads to the representations satisfying the simplex ETF and strengthen NC.

Experiments

Datasets, Metrics, and Implement Details

We conducted the experiments on three RSI datasets: LoveDA (Wang et al. 2022a), ISPRS Vaihingen (Rottensteiner et al. 2020), and ISPRS Potsdam (Rottensteiner et al. 2020) and adopt common metrics to evaluate the performances of our In2NeCT and other methods for comparison.

Datasets **LoveDA (Wang et al. 2022a) dataset** contains 5987 high spatial resolution images (2522, 1669, and 1796 for training, validation, and testing, respectively) of size 1024×1024 in pixels, and involves 7 land classes covering two domains (*i.e.*, urban and rural). LoveDA poses considerable challenges due to the presence of multi-scale objects, and inconsistent class distributions. **Vaihingen (Rottensteiner et al. 2020) dataset** contains 33 TOP image tiles and digital surface models, ranging from 1996×1995 to 3816×2550 in pixels, and involves 6 land classes. We only use the TOP image (14, 1, and 18 for training, validation, and testing, respectively) with near-infrared, red, and green bands. **Potsdam (Rottensteiner et al. 2020) dataset** contains 38 IRRG, RGB, and RGBIR TOP image tiles, of size 6000×6000 in pixels, and involves 6 land classes. We only use the RGB TOP image (23, 1, and 14 for training, validation, and testing, respectively). All images are cropped into 1024×1024 both in Vaihingen and Potsdam.

Implementation details Since our method is orthogonal to the other SOTA methods, we adopt the existing UNetFormer (Wang et al. 2022b), MANet (Li et al. 2021) and LSKNet (Li and et.al. 2023) as our RSI semantic segmentation branch, to demonstrate its flexibility and effectiveness. We implement our In2NeCT on NVIDIA RTX A6000 under the Pytorch framework. For training, the Adam optimizer weight decay of 0.01 is adopted, and the initial learning rate is set to 5e-4 with the cosine annealing learning rate adopted.

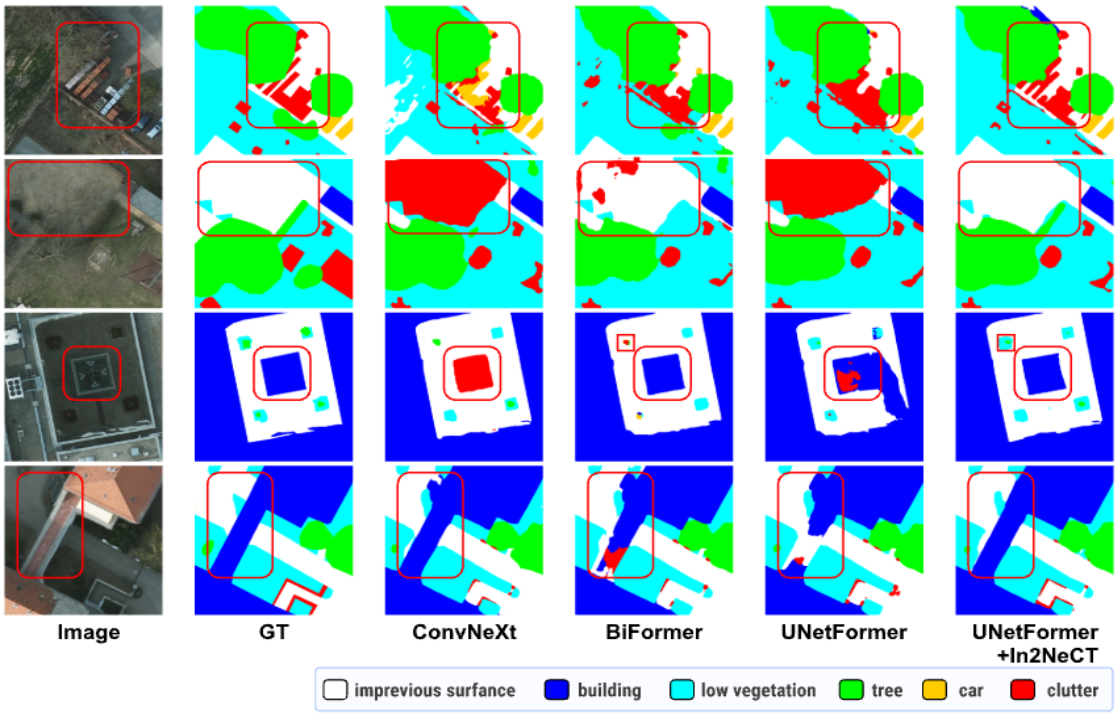


Figure 4: Visualization of semantic segmentation examples on ISPRS Potsdam using In2NeCT and other methods.

The batch size is set to 16, and the epochs are 50, 100, and 100 for LoveDA, Vaihingen, and Potsdam, respectively. For training, we employ random scaling (0.5, 0.75, 1.0, 1.25, 1.5), random vertical flipping, random horizontal flipping, and random rotation data augmentation methods. The augmented images are randomly cropped into patches of size 512×512 in pixels. During the inference, random flipping and multi-scale prediction are used.

Evaluation metrics We adpot *overall accurary* (OA), *average F1 score per class* (AF), and *mean intersection over union* (mIoU) as metrics. Notably, LoveDA is tested online, so we only adopt each class IoU and mIoU as metrics. The mIoU is calculated as follows:

$$IoU_i = \frac{TP_i}{(TP_i + FN_i + FP_i)}, \quad mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU, \tag{13}$$

where TP_i, FN_i, FP_i denote the number of true positives, false negatives, and false positives of class i, respectively. And C denotes the number of classes.

Comparison Results

Quantitative results We compared our In2NeCT with RSI semantic segmentation methods. As shown in Table. 1, integrating In2NeCT significantly outperformed all other methods on three datasets. Specifically, compared to UNetFormer(Wang et al. 2022b), MANet(Li et al. 2021), and LSKNet(Li and et.al. 2023), our In2NeCT obtained an increase by 1.9%, 2.1%, and 0.8% in mIoU on LoveDA. In2NeCT achieved an increase by 1.83%, 0.83%, and 1.29% on ISPRS Vaihingen and by 1.11%, 1.08%, and 0.91% on ISPRS Potsdam in mIoU. The outstanding performances on three datasets demonstrate the effectiveness and consistency of our In2NeCT.

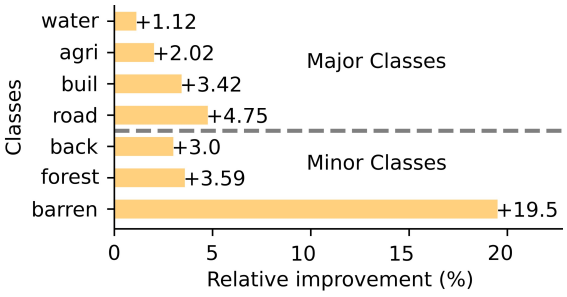


Figure 5: Comparison of In2NeCT compared to UNetFormer about relative IoU improvement of each class on LoveDA.

Furthermore, our In2NeCT achieved greater improvements in minor classes. As shown in Figure 5, compared to UNetFormer, the relative improvement of minor classes has a greater improvement compared to major classes on LoveDA, especially the barren class increased by 19.5%, demonstrating the advantages of our In2NeCT in data imbalance. In addition, we performed a comparison of inter-class collapse degree $\triangle_{Inter} = Avg_k \|\overline{z}_k - \overline{z}_G)/\|\overline{z}_k - \overline{z}_G\| - 1/(K-1)\|$ and intra-class collapse degree $\triangle_{Intra} = Avg_{k,i}\|(z_{k,i} - \overline{z}_k)(z_{k,i} - \overline{z}_k)^\top\|$ in Table 2. The results demonstrated our In2NeCT achieved the improvement in

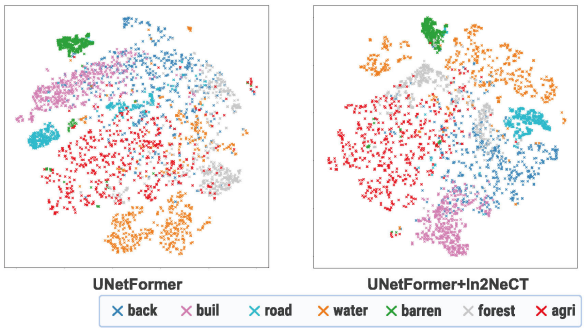


Figure 6: Visualization of the NC degrees of example representation pixels using t-SNE on LoveDA.

Method	$\triangle_{Inter} \downarrow$	$\triangle_{Intra} \downarrow$	mIoU(%) $\uparrow$
UNetFormer	0.047	0.135	52.40
+In2NeCT	0.018	0.113	54.30

Table 2: Comparison of In2NeCT and UNetFormer about inter-class and intra-class collapse degree ($\triangle_{Inter}$, $\triangle_{Intra}$) with mIoU (%) on LoveDA.

mIoU by decreasing $\triangle_{Inter}$ and $\triangle_{Intra}$, which is consistent with our motivation for the representations to satisfy the simplex-ETF and strengthen NC for higher performance.

Qualitative results Figure 4 illustrated the qualitative visualization from ConvNeXt, BiFormer, UNetFormer, and our In2NeCT on ISPRS Potsdam. It is observable that our In2NeCT performs satisfactorily in segmenting the overall framework and edges due to the better representation with the simplex ETF. In addition, the minor classes (*i.e.*, clutter) also performed better with our class-balanced design.

Ablation Studies

To investigate the contributions of ETF-Classifier & $\mathcal{L}_{Inter}$ (Inter-class regularization branch) and $\mathcal{L}_{Intra}$ (Intra-class regularization branch) in our In2NeCT, we performed an ablation study on LoveDA and Vaihingen datasets. As shown in Table 3, each module has a significant positive contribution to the performance of In2NeCT. In particular, ETF-Classifier & $\mathcal{L}_{Inter}$ and $\mathcal{L}_{Intra}$ leads to the increases by 0.5% and 1.2% in mIoU on LoveDA, which suggests that both modules are crucial for the performance improvement.

To intuitively illustrate the influence of our In2NeCT on

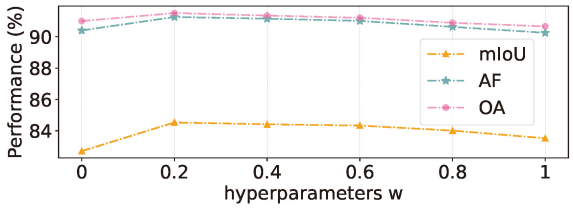


Figure 7: Ablation about the loss weight w and performance on Vaihingen with UNetFormer.

Method	LoveDA	Vaihingen		
	mIoU	AF	mIoU	OA
UNetFormer	52.4	90.40	82.70	91.00
+ ETF & $\mathcal{L}_{Inter}$	52.9	90.81	83.89	91.35
+ $\mathcal{L}_{Intra}$	53.6	90.72	83.54	91.23
+ ETF & $\mathcal{L}_{Inter}$ + $\mathcal{L}_{Intra}$ (Ours)	**54.3**	**91.26**	**84.53**	**91.51**

Table 3: Ablation study about each module of our In2NeCT on LoveDA and ISPRS Vaihingen. Best scores are in *bold*. *ETF* denotes ETF-classifier.

Method	mIoU $\uparrow$	Training (s) $\downarrow$	Testing (s) $\downarrow$
UNetFormer	52.4	1.53	0.48
+ In2NeCT	54.3(+1.9)	1.69(+10.5%)	0.48

Table 4: Comparison of In2NeCT and UNetFormer about performance and the Computational Time Complexity (*i.e.*, second per image) on LoveDA.

the representation space, we visualized the NC degrees using the t-SNE algorithm on randomly sampled 500 representation pixels and corresponding labels for each class on LoveDA, as shown in Figure 6. Compared to UNetFormer, representations from the integration with In2NeCT are more compact (intra-class) and discriminative (inter-class).

We further performed the comparison about the computational complexity. As In2NeCT can be easily integrated into any off-the-shelf RSI semantic segmentation architecture. In the testing of our method, we only preserved the RSI semantic segmentation branch, while the Intra-class and Inter-class regularization branches were discarded. Therefore, the testing of integrating In2NeCT was very efficient. As shown in Table 4, the integration of the additional branches only increase the training time cost by about 10%. All experiments were conducted on the Nvidia GeForce A6000 GPU.

We performed the ablation about the hyperparameter w in loss function (Eq. 7) on the three performance metrics. As shown in Figure 7, the optimal performance was achieved when $w = 0.2$. Thus, we suggest to set $w = 0.2$ for the best performance. Furthermore, please refer to *Technical Appendix* for additional experiments.

Conclusions

In this paper, we explored the last-layer representation of remote sensing image (RSI) semantic segmentation from the perspective of neural collapse. It revealed that the optimal structure of these representations is disrupted in the presence of data imbalance, consequently hurting the performance. To address this challenge, we proposed Intra-class and Inter-class Neural Collapse Tuning (In2NeCT) to optimize the representations to adhere to the simplex equiangular tight frame (ETF). Adopting two additional branches with simplex ETF-classifier & inter-class collapse regularization loss, as well as intra-class collapse regularization loss, In2NeCT effectively enhanced the performance of RSI semantic segmentation. Extensive experiments demonstrated that In2NeCT consistently outperformed the state-of-the-art methods on all three datasets.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62202421 and U23A20311); in part by Zhejiang Provincial Natural Science Foundation of China (Grant No. LTGS23F020001); in part by Ningbo Yongjiang Talent Introduction Program of China (Grant No. 2021A-157-G); in part by the Public Welfare Science and Technology Plan of Ningbo City (Grant No. 2022S125); and in part by the Key Research and Development Program of Ningbo City of China (Grant No. 2023Z130).

References

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Digra, M.; Dhir, R.; and Sharma, N. 2022. Land use land cover classification of remote sensing images based on the deep learning approaches: a statistical analysis and review. *Arabian Journal of Geosciences*, 15(10): 1003.

Ding, L.; Tang, H.; and Bruzzone, L. 2021. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1): 426–435.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. arXiv:1911.05722.

Jin, Z.; Liu, B.; Chu, Q.; and Yu, N. 2021. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7189–7198.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408.

Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; and Atkinson, P. M. 2021. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Li, Y.; and et.al. 2023. Large Selective Kernel Network for Remote Sensing Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16794–16805.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, Z.; Yang, W.; Yuan, Y.; Gou, R.; and Li, X. 2023. Semantic segmentation of agricultural images: a survey. *Information Processing in Agriculture*.

Papyan, V.; Han, X. Y.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Bnitez, S.; and Breitkopf, U. 2020. International society for photogrammetry and remote sensing, 2d semantic labeling contest. *Accessed: Oct*, 29.

Sara, U.; Akter, M.; and Uddin, M. S. 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, 7(3): 8–18.

Song, Q.; Li, J.; Li, C.; Guo, H.; and Huang, R. 2022. Fully attentional network for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2280–2288.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for Semantic Segmentation. arXiv:2105.05633.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2022a. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. arXiv:2110.08733.

Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; and Atkinson, P. M. 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190: 196–214.

Weinan, E.; and Wojtowytsch, S. 2022. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Mathematical and Scientific Machine Learning*, 270–290. PMLR.

Yang, Y.; Xie, L.; Chen, S.; Li, X.; Lin, Z.; and Tao, D. 2022. Do we really need a learnable classifier at the end of deep neural network? *arXiv e-prints*, arXiv–2203.

Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; and Tao, D. 2023. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*.

Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10819–10829.

Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. 2020. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241: 111716.

Yuan, Y.; Chen, X.; and Wang, J. 2020. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 173–190. Springer.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, Z.; Zhong, Y.; Wang, J.; and Ma, A. 2020. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4095–4104.

Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; and Lau, R. W. 2023. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10323–10333.