Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Label-Graph Guided Semantic Alignment for Multi-Class Remote Sensing Image Recognition

KUN ZHOU
Zhejiang Business Technology Institute
LIWEI ZHU
Zhejiang Shangong Intelligent Technology
YI ZHANG
Zhejiang Business Technology Institute
CUNCUN WEI
weicuncun@163.com

Zhejiang Business Technology Institute

Research Article

Keywords: Remote sensing image classification, Label graph, Semantic alignment, Inter-class dependency, Confusion-aware learning

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9190415/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Label-Graph Guided Semantic Alignment for Multi-Class Remote Sensing Image Recognition

Kun Zhou[1,2], Liwei Zhu[3], Yi Zhang[1], Cuncun Wei[1*]

[1*]School of Artificial Intelligence, Zhejiang Business Technology Institute, Ningbo, 315012, China.
[2]School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia.
[3]Zhejiang Shangong Intelligent Technology Co., Ltd., Hangzhou, 315000, China.

*Corresponding author(s). E-mail(s): weicuncun@163.com;
Contributing authors: zhoukun@student.usm.my; zhuliwei@zeiet.com;
rhettzhang@live.com;

Abstract

In multi-class remote sensing image classification, semantic confusion among similar classes often limits performance, as conventional one-hot labels cannot express inter-class relationships. To address this issue, we propose a Label-Graph-Guided Semantic Alignment method that leverages a data-driven label graph to enhance both the training supervision and the model predictions. Specifically, we construct a label graph from the confusion matrix to capture class similarities, and introduce two novel mechanisms: Graph-Soft at the supervision level and Graph-Logit at the prediction level. Graph-Soft utilizes the label graph to refine one-hot labels into soft label vectors. Meanwhile, Graph-Logit encourages the model's logits to respect label similarities, aligning the semantic space of predictions with that of labels. The proposed approach is lightweight and architecture-agnostic, introducing negligible computational overhead and requiring no additional annotations. Experiments on multiple remote sensing scene classification benchmarks (EuroSAT, MLRSNet and RESISC45) with different backbones (ResNet18, ResNet50) demonstrate consistent performance gains. Our method outperforms standard one-hot training across all tested datasets, improving classification accuracy, AUC, and F1-score. Notably, it achieves up to a 2-3% increase in accuracy on challenging benchmarks compared to the baseline. These results validate that incorporating label graph knowledge effectively reduces semantic confusion and enhances multi-class classification performance. The

source code is publicly available at https://github.com/zhoukuniyc/GLSGLR/tree/master.

Keywords: Remote sensing image classification, Label graph , Semantic alignment , Inter-class dependency, Confusion-aware learning

1 Introduction

Remote sensing image classification is a fundamental task in Earth observation, supporting a wide range of applications such as land-use and land-cover mapping, environmental monitoring, and urban planning [1]. With the rapid growth of high-resolution satellite imagery and large-scale benchmarks, deep learning based methods have become the dominant paradigm for scene recognition in remote sensing [2]. Despite this progress, multi-class remote sensing image recognition remains challenging due to the inherent complexity of scenes, including large intra-class variability and strong inter-class visual similarity [3]. Visually distinct semantic categories often share similar spatial structures and textures, leading to frequent confusion among related classes.

Most existing approaches formulate remote sensing scene classification as a standard multi-class problem and adopt one-hot label supervision with cross-entropy loss [4]. This formulation implicitly assumes that class labels are mutually exclusive and semantically independent. However, such an assumption is often unrealistic in remote sensing scenarios, where semantic relationships naturally exist among categories. For example, classes such as *residential*, *commercial*, and *industrial* areas share common visual components, while different vegetation-related scenes may differ only subtly in appearance. Treating these categories as equally distinct can result in over-confident predictions and poor discrimination for visually similar classes [5]. As a consequence, conventional one-hot supervision may limit the ability of deep models to handle semantic ambiguity in complex remote sensing scenes.

To alleviate this issue, recent studies have explored incorporating additional semantic information into the learning process. Some works focus on confusion-aware or hard-sample oriented learning strategies to emphasize easily confused categories [6]. Others investigate hierarchical semantic structures to better reflect category relations and reduce ambiguity [7]. In parallel, graph-based representations have attracted increasing attention as a flexible way to model dependencies among labels or semantic concepts in remote sensing analysis. However, many existing methods rely on complex graph neural networks, external knowledge bases, or additional annotations, which may increase model complexity and limit their practical applicability.

Motivated by these observations, this work investigates how structured semantic relationships among class labels can be exploited in a lightweight and data-driven manner to improve multi-class remote sensing image recognition. Instead of treating labels as isolated entities, we aim to leverage inter-class dependencies to guide the learning process, while preserving the simplicity and generality of standard classification frameworks.

The main contributions of this work are summarized as follows:

- We identify the limitation of one-hot supervision in multi-class remote sensing image recognition and highlight the importance of explicitly modeling inter-class semantic relationships.
- We propose a label-graph guided semantic alignment framework that leverages structured label dependencies to enhance both training supervision and prediction behavior, without modifying backbone architectures.
- We develop a lightweight and data-driven approach to construct and utilize label graphs, requiring no additional annotations or external knowledge sources.
- Extensive experiments on multiple remote sensing benchmark datasets and different CNN backbones demonstrate the effectiveness and generality of the proposed method.

The remainder of this paper is organized as follows. Section II reviews related work on remote sensing image classification and semantic label modeling. Section III presents the proposed label-graph guided semantic alignment framework. Section IV describes the experimental setup and reports comparative results on benchmark datasets. Finally, Section V concludes this paper and discusses potential directions for future work.

2 Related Work

2.1 Remote Sensing Image Classification

Remote sensing image scene classification has been transformed by deep learning, achieving remarkable performance improvements in recent years. Early approaches relied on hand-crafted features or bag-of-visual-words models, but modern methods predominantly use deep neural networks for feature learning and classification [8]. Comprehensive surveys (e.g., Cheng et al. 2020) have documented how convolutional neural networks (CNNs) became the state-of-the-art, addressing challenges like high intra-class diversity and low inter-class separability in aerial scenes. Current research focuses on improving network architectures and learning strategies to better capture discriminative information from limited and complex remote sensing data. For example, many works introduce attention mechanisms and multi-scale feature fusion into CNN models to enhance representation of both global context and local details. Tang et al. (2021) propose an attention-consistent network (ACNet) that uses dual-branch Siamese CNNs with spatial and channel attention modules, achieving notable accuracy gains on several scene benchmarks [9]. Similarly, vision transformers and hybrid CNN-Transformer models have been explored to exploit long-range dependencies. Yang et al. (2023) design a multi-scale transformer for scene classification, demonstrating that transformer-based frameworks can rival or surpass CNNs in remote sensing tasks by capturing global context across high-resolution images [10]. Beyond architecture innovations, researchers have also leveraged techniques like transfer learning from ImageNet, data augmentation, and ensemble learning to boost classification performance on remote sensing scene datasets. As a result of these advances, today's best models exceed 95% accuracy on challenging benchmarks (e.g., RESISC45), though issues like

limited labeled data and class imbalance continue to motivate ongoing improvements in remote sensing image classification.

2.2 Label Dependency Modeling and Graph-Based Supervision

Standard image classification training treats class labels as independent one-hot vectors, which neglects semantic relationships among classes and can lead to overconfident predictions. Recent work aims to model label dependencies during training to overcome the limitations of one-hot supervision. Label smoothing is a widely used strategy that softens the hard one-hot targets by assigning a small probability to incorrect classes. This regularization mitigates overconfidence and can improve generalization [11]. However, conventional label smoothing typically uses a uniform distribution for the off-target classes, which is independent of the data and does not reflect actual class similarities. To address this, researchers have proposed adaptive or informed smoothing approaches. For instance, some methods replace the uniform label prior with confusion-aware soft labels learned from the model's confusion matrix or a teacher network. This allows the training targets to allocate higher probability to classes that are semantically or visually similar to the true class, thereby encoding inter-class correlation information. Zhang et al. (2022) introduced a two-branch "FocusNet" architecture to explicitly handle confusing classes: a primary network identifies the most confusable classes for each sample, and an auxiliary branch is trained to distinguish the true label from these confusing classes, with a focus-picking loss that emphasizes hard samples. Their results demonstrated that accounting for class confusion during training can yield better generalization than naive one-hot or uniformly smoothed labels [11].

Another line of research uses graph-based supervision to model label dependencies. In multi-label remote sensing tasks, where an image can have multiple scene tags, graph neural networks have been employed to capture co-occurrence and semantic relations among labels. For example, Ma et al. (2024) construct a label correlation graph and integrate it with a Graph Convolutional Network (GCN) to propagate information between labels, improving multi-label classification performance by exploiting label co-occurrence patterns [12]. Even for strictly single-label (multi-class) scene classification, a label graph can encode similarities,such as hierarchical relationships or spectral confusion,between different land-use classes. Incorporating such a label graph into the training objective is an emerging idea to relax the overly rigid one-hot supervision. Some recent approaches extend the idea of knowledge graphs or GCNs to multi-class settings by treating each class as a node in a graph and encouraging the model's predictions to respect the similarity structure among classes. This graph-based label smoothing can be seen as a generalization of confusion-aware learning, where the affinity between classes (derived from domain knowledge or data-driven metrics) dictates the soft label distribution. In summary, prior works on label dependency modeling,from label smoothing variants to confusion-focused losses and graph-based label correlations,show that enriching the supervision signal beyond one-hot labels can improve classifier robustness and accuracy. These methods inform our approach, which further innovates by introducing a graph-guided label supervision framework. Unlike previous methods, our framework explicitly encodes a prior label relationship graph into the training process (both at the soft label and logit levels), enabling the

model to learn class decision boundaries that align with known inter-class relationships. This offers a novel way to leverage domain knowledge of label semantics, setting our work apart from existing one-hot or purely data-driven smoothing techniques.

3 Proposed Method

3.1 Framework

We address the multi-class remote sensing image recognition problem by proposing a label-graph guided semantic alignment framework. Given an input image x, a backbone network $f_\theta(\cdot)$ produces a logit vector $\mathbf{z} \in \mathbb{R}^C$ over C scene classes, which is converted into a class probability distribution via the softmax function.

Unlike conventional training schemes that rely on one-hot supervision and assume label independence, our framework explicitly models inter-class semantic relationships through a label graph [13]. The label graph serves as a fixed semantic prior and is injected into the learning process through two complementary components: (i) *Graph-Soft supervision*, which aligns the training targets with structured label semantics, and (ii) *Graph-Logit regularization*, which enforces semantic consistency at the prediction level. Both components operate purely in the label space and do not modify the backbone architecture, making the framework lightweight and model-agnostic.

An overview of the proposed framework is illustrated in Fig. 1, where the label graph is constructed once and shared across all training samples, and its influence is injected through supervision- and prediction-level alignment.

3.2 Label Graph Construction

To model semantic relationships among scene categories, we define a label graph $G = (\mathcal{C}, \mathbf{A})$, where $\mathcal{C} = \{1, \ldots, C\}$ denotes the set of classes and $\mathbf{A} \in \mathbb{R}^{C \times C}$ is an affinity matrix. Each entry $\mathbf{A}_{ij}$ represents the semantic relatedness between class i and class j.

In this work, the label graph is constructed in a data-driven manner from class confusion statistics. Specifically, a baseline classifier is first trained using standard one-hot supervision. We then compute its normalized confusion matrix on a held-out split, which reflects the frequency with which samples from one class are misclassified as another. Classes that are frequently confused are considered to have higher semantic affinity. Based on these statistics, we derive the affinity matrix $\mathbf{A}$ by assigning larger weights to such class pairs.

To ensure numerical stability and interpretability, we enforce $\mathbf{A}_{ii} = 1$ for all classes and normalize each row of $\mathbf{A}$ such that $\sum_j \mathbf{A}_{ij} = 1$. Once constructed, the label graph is fixed throughout training and shared across all samples. This design avoids introducing additional learnable parameters and allows the graph to function as a lightweight semantic prior.

3.3 Graph-Soft Supervision

Standard cross-entropy loss employs one-hot targets, implicitly assuming that all non-ground-truth classes are equally unrelated [14]. Such rigid supervision ignores inter-class semantic similarity and may hinder learning when visually similar classes are

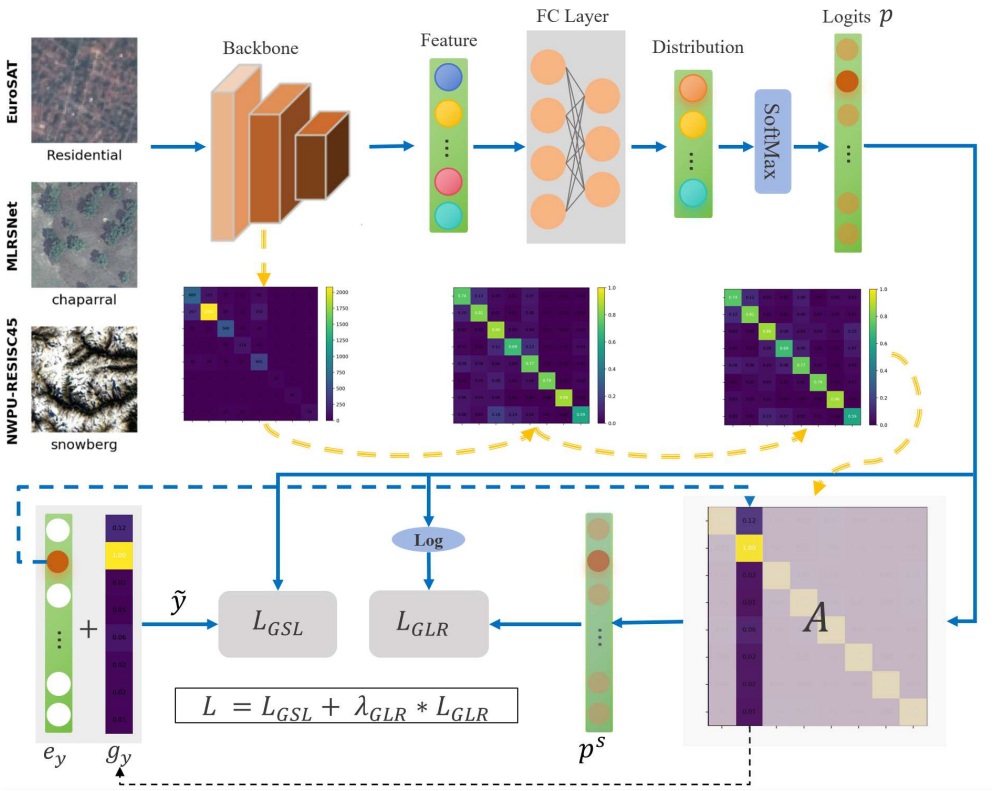


Fig. 1 Framework of the proposed label-graph guided semantic alignment approach. A label graph encoding inter-class semantic relationships is constructed in advance and injected into training through Graph-Soft supervision and Graph-Logit regularization.

present. To address this issue, we introduce *Graph-Soft supervision*, which replaces hard targets with graph-guided soft labels.

For a training sample with ground-truth label $y \in \{1, \ldots, C\}$, we denote the one-hot target by $\mathbf{e}_y \in \mathbb{R}^C$ and retrieve the y-th row of the label graph, denoted by $\mathbf{A}_{y,:}$. The soft target distribution is defined as

$$\tilde{\mathbf{y}} = (1 - \alpha)\mathbf{e}_y + \alpha\, \mathbf{A}_{y,:}, \tag{1}$$

where $\alpha \in [0, 1]$ controls the contribution of the label graph. This formulation preserves the ground-truth label while allocating nonzero probability mass to semantically related classes.

Given the predicted distribution $p_\theta(c \mid x) = \mathrm{softmax}(\mathbf{z})_c$, the Graph-Soft loss is computed as the cross-entropy between $\tilde{\mathbf{y}}$ and $p_\theta(c \mid x)$.

Formally, the Graph-Soft supervision loss is defined as

$$\mathcal{L}_{\mathrm{GSL}} = -\sum_{c=1}^{C} \tilde{y}_c \log p_\theta(c \mid x), \tag{2}$$

where $\tilde{y}_c$ denotes the c-th element of the graph-guided soft target $\tilde{\mathbf{y}}$. Compared with the standard cross-entropy loss, this formulation relaxes the one-hot constraint and encourages the model to allocate probability mass to semantically related classes according to the label graph.

3.4 Graph-Logit Regularization

While Graph-Soft supervision injects semantic structure into the training targets, it does not explicitly constrain the model predictions. To further enforce semantic alignment, we introduce *Graph-Logit regularization*, which operates directly on the predicted class distribution.

Let $\mathbf{p} = \mathrm{softmax}(\mathbf{z}/\tau)$ denote the temperature-scaled prediction, where $\tau > 0$ controls distribution smoothness. Using the label graph, we compute a graph-smoothed distribution

$$\mathbf{p}^s = \mathrm{Normalize}\big(\mathbf{pA}\big), \tag{3}$$

where $\mathrm{Normalize}(\cdot)$ denotes row-wise normalization. This operation propagates confidence mass to semantically related classes according to the label graph.

We define the Graph-Logit regularization term as the Kullback-Leibler divergence

$$\mathcal{L}_{\mathrm{GLR}} = D_{\mathrm{KL}}(\mathbf{p}^s \,\|\, \mathbf{p}) = \sum_{c=1}^{C} p_c^s \left(\log p_c^s - \log p_c\right), \tag{4}$$

which encourages the predicted distribution to be consistent with its graph-smoothed counterpart.

3.5 Training Objective

The final training objective combines standard classification loss with the proposed graph-guided components. When Graph-Soft supervision is used, the conventional one-hot cross-entropy loss is replaced by the Graph-Soft loss. Graph-Logit regularization is added as an auxiliary term. The objective is expressed as

$$\mathcal{L} = \mathcal{L}_{\mathrm{GSL}} + \lambda_{\mathrm{GLR}}\, \mathcal{L}_{\mathrm{GLR}}, \tag{5}$$

where λ_{GLR} controls the contribution of Graph-Logit regularization. This design allows the two components to be flexibly combined while keeping the training procedure simple and efficient.

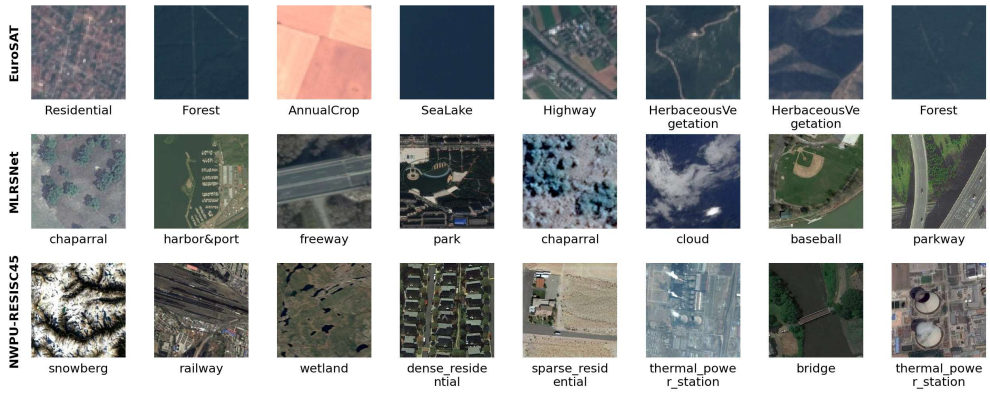


Fig. 2 Dataset overview with randomly sampled images from three remote sensing benchmarks. Each row corresponds to one dataset (EuroSAT, MLRSNet_scene46, and RESISC45), and each column shows an example scene category, highlighting the large intra-dataset diversity and inter-dataset differences.

Table 1 Datasets and data splits used in our experiments.

Dataset	#Classes	Train	Val	Test
EuroSAT (2750)	10	18,900	4,050	4,050
RESISC45	45	18,900	6,300	6,300
MLRSNet_scene46	46	76,402	16,364	16,398

4 Experiments

4.1 Datasets and Splits

We evaluate our method on three multi-class remote sensing benchmarks, covering both scene classification and land-cover recognition. Fig. 2 provides a visual overview of the datasets, illustrating the diverse scene appearances and semantic variations across different benchmarks.

For RESISC45 [15], we directly use the official train/validation/test split provided by the HuggingFace dataset interface (`timm/resisc45`). For MLRSNet_scene46 [16], we follow the predefined split files (`train.csv/val.csv/test.csv`) under `splits_scene46` configurations, where each image is associated with a single scene label derived from its filename. For EuroSAT [17], we construct a fixed stratified split using `val_ratio` $= 0.15$ and `test_ratio` $= 0.15$ with `split_seed` $= 42$, and keep this split identical across all training seeds to ensure fair and reproducible comparisons.

Table 1 summarizes the dataset statistics and split protocols used throughout our experiments.

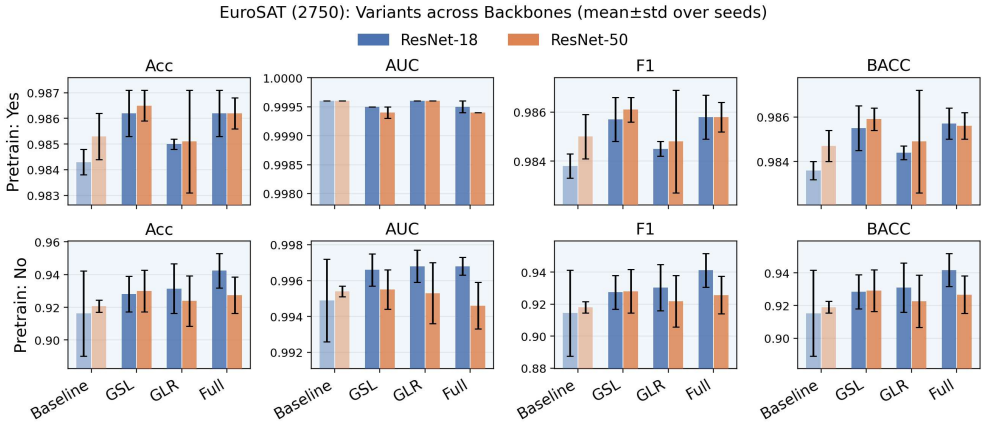


Fig. 3 Performance comparison on EuroSAT under different backbones, pretraining settings, and graph-guided variants. Results are reported as mean±std over multiple seeds.

4.2 Implementation Details

All experiments are implemented in PyTorch with mixed-precision training [18] . We adopt ResNet-18 and ResNet-50 backbones [19] and set the input resolution to 224×224. Training uses AdamW with learning rate 1×10^{-4} and weight decay 1×10^{-4}, batch size 512, and a maximum of 500 epochs [20]. We apply early stopping with patience 6 based on a validation score defined as the harmonic mean of macro-F1 and balanced accuracy (BACC), and report test performance of the best validation checkpoint.

Data augmentation follows the training pipeline in our code: random horizontal flip ($p = 0.5$), vertical flip ($p = 0.2$), and random rotation (limit $15°$, $p = 0.3$), followed by normalization. Unless otherwise specified, results are averaged over three random seeds $\{42, 1337, 2025\}$ and reported as mean±std.

4.3 Compared Variants

We compare four training variants under identical backbones and hyperparameters: **(i) Baseline** (standard cross-entropy with one-hot labels), **(ii) GSL** (Graph-Soft supervision only), **(iii) GLR** (Graph-Logit regularization only), and **(iv) Full** (GSL+GLR). For all graph-based variants, we first run a warmup stage of 5 epochs using standard cross-entropy, then build the label graph on the validation split using class-conditional mean predicted probabilities (`use_prob_graph=1`) and apply symmetrization (`sym_graph=1`). The default hyperparameters are $\alpha = 0.3$ for GSL, $\lambda = 0.1$ and $\tau = 1.0$ for GLR.

4.4 Evaluation Metrics

We report accuracy (Acc), macro-AUC (OvR), macro-F1, and balanced accuracy (BACC) following the evaluation function in our code [21]. All results are computed on the test split and summarized as mean±std across seeds.

Table 2 Test performance (mean±std over seeds) on EuroSAT (2750). Best results in each (Backbone, Pretrain) group are bolded per metric.

Backbone	Pretrain	Variant	Acc	AUC	F1	BACC
resnet18	Yes	baseline	0.9843 ± 0.0005	$\mathbf{0.9996 \pm 0.0000}$	0.9838 ± 0.0005	0.9836 ± 0.0004
resnet18	Yes	gsl	$\mathbf{0.9862 \pm 0.0009}$	0.9995 ± 0.0000	0.9857 ± 0.0009	0.9855 ± 0.0010
resnet18	Yes	glr	0.9850 ± 0.0002	$\mathbf{0.9996 \pm 0.0000}$	0.9845 ± 0.0003	0.9844 ± 0.0003
resnet18	Yes	full	$\mathbf{0.9862 \pm 0.0009}$	0.9995 ± 0.0001	$\mathbf{0.9858 \pm 0.0009}$	$\mathbf{0.9857 \pm 0.0007}$
resnet18	No	baseline	0.9162 ± 0.0261	0.9949 ± 0.0023	0.9144 ± 0.0269	0.9153 ± 0.0263
resnet18	No	gsl	0.9282 ± 0.0109	0.9966 ± 0.0009	0.9274 ± 0.0105	0.9285 ± 0.0105
resnet18	No	glr	0.9314 ± 0.0152	$\mathbf{0.9968 \pm 0.0009}$	0.9303 ± 0.0145	0.9310 ± 0.0151
resnet18	No	full	$\mathbf{0.9424 \pm 0.0105}$	$\mathbf{0.9968 \pm 0.0005}$	$\mathbf{0.9411 \pm 0.0105}$	$\mathbf{0.9418 \pm 0.0101}$
resnet50	Yes	baseline	0.9853 ± 0.0009	$\mathbf{0.9996 \pm 0.0000}$	0.9850 ± 0.0009	0.9847 ± 0.0007
resnet50	Yes	gsl	$\mathbf{0.9865 \pm 0.0006}$	0.9994 ± 0.0001	$\mathbf{0.9861 \pm 0.0005}$	$\mathbf{0.9859 \pm 0.0005}$
resnet50	Yes	glr	0.9851 ± 0.0020	$\mathbf{0.9996 \pm 0.0000}$	0.9848 ± 0.0021	0.9849 ± 0.0023
resnet50	Yes	full	0.9862 ± 0.0006	0.9994 ± 0.0000	0.9858 ± 0.0006	0.9856 ± 0.0006
resnet50	No	baseline	0.9207 ± 0.0038	0.9954 ± 0.0003	0.9180 ± 0.0036	0.9190 ± 0.0036
resnet50	No	gsl	$\mathbf{0.9299 \pm 0.0127}$	$\mathbf{0.9955 \pm 0.0011}$	$\mathbf{0.9280 \pm 0.0136}$	$\mathbf{0.9292 \pm 0.0127}$
resnet50	No	glr	0.9239 ± 0.0154	0.9953 ± 0.0017	0.9218 ± 0.0160	0.9227 ± 0.0160
resnet50	No	full	0.9275 ± 0.0111	0.9946 ± 0.0013	0.9256 ± 0.0117	0.9267 ± 0.0115

Table 3 Test performance (mean±std over seeds) on MLRSNet_scene46. Best results in each (Backbone, Pretrain) group are bolded per metric.

Backbone	Pretrain	Variant	Acc	AUC	F1	BACC
resnet18	Yes	baseline	0.9707 ± 0.0011	$\mathbf{0.9997 \pm 0.0000}$	0.9705 ± 0.0013	0.9708 ± 0.0010
resnet18	Yes	gsl	0.9729 ± 0.0007	$\mathbf{0.9997 \pm 0.0000}$	0.9727 ± 0.0008	0.9729 ± 0.0007
resnet18	Yes	glr	$\mathbf{0.9736 \pm 0.0006}$	$\mathbf{0.9997 \pm 0.0000}$	$\mathbf{0.9734 \pm 0.0006}$	$\mathbf{0.9736 \pm 0.0006}$
resnet18	Yes	full	0.9732 ± 0.0007	$\mathbf{0.9997 \pm 0.0000}$	0.9731 ± 0.0008	0.9734 ± 0.0007
resnet18	No	baseline	0.9265 ± 0.0089	$\mathbf{0.9990 \pm 0.0001}$	0.9264 ± 0.0089	0.9272 ± 0.0086
resnet18	No	gsl	0.9328 ± 0.0073	$\mathbf{0.9990 \pm 0.0001}$	0.9330 ± 0.0072	0.9336 ± 0.0072
resnet18	No	glr	0.9337 ± 0.0075	$\mathbf{0.9990 \pm 0.0001}$	0.9338 ± 0.0077	0.9342 ± 0.0074
resnet18	No	full	$\mathbf{0.9348 \pm 0.0067}$	$\mathbf{0.9990 \pm 0.0001}$	$\mathbf{0.9353 \pm 0.0069}$	$\mathbf{0.9353 \pm 0.0066}$
resnet50	Yes	baseline	$\mathbf{0.9754 \pm 0.0008}$	$\mathbf{0.9998 \pm 0.0000}$	$\mathbf{0.9754 \pm 0.0008}$	$\mathbf{0.9755 \pm 0.0008}$
resnet50	Yes	gsl	0.9732 ± 0.0008	$\mathbf{0.9998 \pm 0.0000}$	0.9733 ± 0.0008	0.9734 ± 0.0008
resnet50	Yes	glr	0.9740 ± 0.0007	$\mathbf{0.9998 \pm 0.0000}$	0.9740 ± 0.0007	0.9741 ± 0.0007
resnet50	Yes	full	0.9734 ± 0.0008	$\mathbf{0.9998 \pm 0.0000}$	0.9735 ± 0.0008	0.9736 ± 0.0008
resnet50	No	baseline	0.9409 ± 0.0055	$\mathbf{0.9992 \pm 0.0001}$	0.9409 ± 0.0058	0.9414 ± 0.0054
resnet50	No	gsl	0.9443 ± 0.0078	$\mathbf{0.9992 \pm 0.0001}$	0.9445 ± 0.0077	0.9448 ± 0.0078
resnet50	No	glr	0.9452 ± 0.0074	$\mathbf{0.9992 \pm 0.0001}$	0.9454 ± 0.0072	0.9459 ± 0.0071
resnet50	No	full	$\mathbf{0.9460 \pm 0.0075}$	$\mathbf{0.9992 \pm 0.0001}$	$\mathbf{0.9461 \pm 0.0074}$	$\mathbf{0.9465 \pm 0.0073}$

Table 4 Test performance (mean±std over seeds) on RESISC45 (`timm/resisc45`). Best results in each (Backbone, Pretrain) group are bolded per metric.

Backbone	Pretrain	Variant	Acc	AUC	F1	BACC
resnet18	Yes	baseline	0.9428 ± 0.0006	$\mathbf{0.9992 \pm 0.0000}$	0.9428 ± 0.0006	0.9428 ± 0.0006
resnet18	Yes	gsl	0.9506 ± 0.0010	$\mathbf{0.9992 \pm 0.0000}$	0.9506 ± 0.0010	0.9506 ± 0.0010
resnet18	Yes	glr	0.9513 ± 0.0010	$\mathbf{0.9992 \pm 0.0000}$	0.9513 ± 0.0010	0.9513 ± 0.0010
resnet18	Yes	full	$\mathbf{0.9521 \pm 0.0010}$	$\mathbf{0.9992 \pm 0.0000}$	$\mathbf{0.9520 \pm 0.0010}$	$\mathbf{0.9521 \pm 0.0010}$
resnet18	No	baseline	0.9133 ± 0.0013	0.9987 ± 0.0001	0.9132 ± 0.0013	0.9133 ± 0.0013
resnet18	No	gsl	0.9231 ± 0.0011	$\mathbf{0.9988 \pm 0.0001}$	0.9230 ± 0.0011	0.9231 ± 0.0011
resnet18	No	glr	0.9237 ± 0.0012	$\mathbf{0.9988 \pm 0.0001}$	0.9236 ± 0.0012	0.9236 ± 0.0012
resnet18	No	full	$\mathbf{0.9253 \pm 0.0013}$	$\mathbf{0.9988 \pm 0.0001}$	$\mathbf{0.9250 \pm 0.0013}$	$\mathbf{0.9251 \pm 0.0013}$
resnet50	Yes	baseline	0.9581 ± 0.0005	$\mathbf{0.9994 \pm 0.0000}$	0.9581 ± 0.0005	0.9581 ± 0.0005
resnet50	Yes	gsl	0.9599 ± 0.0005	$\mathbf{0.9994 \pm 0.0000}$	0.9599 ± 0.0005	0.9599 ± 0.0005
resnet50	Yes	glr	0.9602 ± 0.0004	$\mathbf{0.9994 \pm 0.0000}$	0.9602 ± 0.0004	0.9602 ± 0.0004
resnet50	Yes	full	$\mathbf{0.9606 \pm 0.0004}$	$\mathbf{0.9994 \pm 0.0000}$	$\mathbf{0.9605 \pm 0.0004}$	$\mathbf{0.9605 \pm 0.0004}$
resnet50	No	baseline	0.9292 ± 0.0012	0.9989 ± 0.0000	0.9291 ± 0.0012	0.9291 ± 0.0012
resnet50	No	gsl	$\mathbf{0.9351 \pm 0.0009}$	$\mathbf{0.9990 \pm 0.0000}$	$\mathbf{0.9350 \pm 0.0009}$	$\mathbf{0.9350 \pm 0.0009}$
resnet50	No	glr	0.9350 ± 0.0010	$\mathbf{0.9990 \pm 0.0000}$	$\mathbf{0.9350 \pm 0.0010}$	$\mathbf{0.9350 \pm 0.0010}$
resnet50	No	full	0.9317 ± 0.0012	0.9989 ± 0.0000	0.9315 ± 0.0012	0.9315 ± 0.0012

4.5 Main Results

Tables 2, 3, and 4 summarize the quantitative results across three remote sensing benchmarks. For clarity, Fig. 3, Fig. 4, and Fig. 5 further visualize these results by grouping variants and backbones under different pretraining settings.

Incorporating label-graph guidance leads to consistent improvements over the baseline in most configurations. As illustrated in the figures, the gains are particularly pronounced when training from scratch (i.e., without ImageNet pretraining), where the label graph acts as an additional semantic prior that compensates for weaker initialization. In these settings, graph-guided variants consistently outperform the baseline across accuracy, F1-score, and balanced accuracy.

On EuroSAT and RESISC45, the **Full** model achieves the strongest or tied-best performance for nearly all backbone and pretraining combinations, a trend that is clearly reflected by the higher bars across metrics in Fig. 3 and Fig. 5. On MLRSNet_scene46, most graph-guided variants still improve upon the baseline; however, a slight degradation is observed for ResNet-50 with ImageNet pretraining (Fig. 4). This suggests that the effectiveness of label-graph guidance can depend on the backbone capacity and the strength of initialization, with diminishing returns when strong pretrained representations are already available.

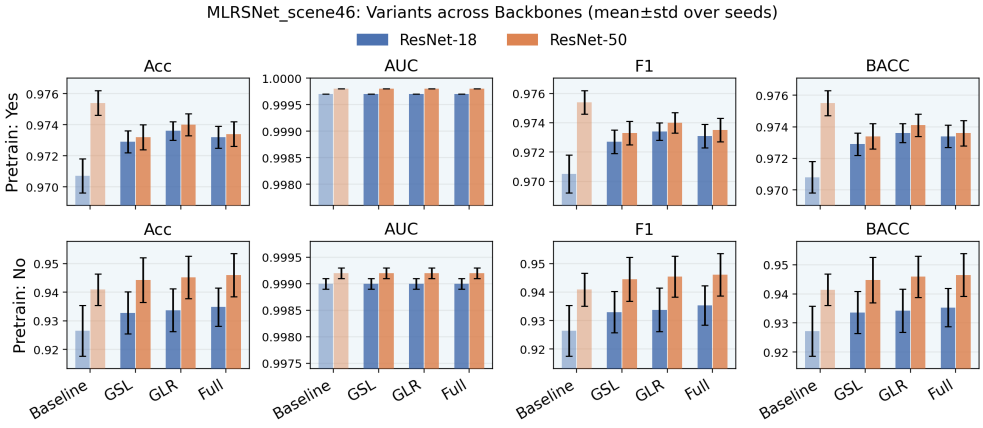


Fig. 4 Performance comparison on MLRSNet_scene46. Graph-guided variants are compared with the baseline across different backbones and pretraining configurations.

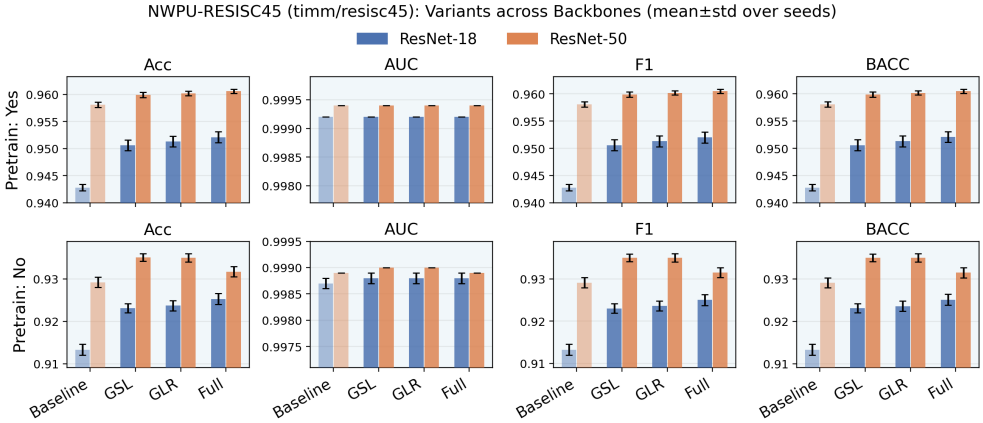


Fig. 5 Performance comparison on RESISC45 (`timm/resisc45`) across different backbones, pretraining settings, and model variants.

4.6 Ablation and Analysis

The ablation results further clarify the roles of individual components. Across all datasets, both **GSL** and **GLR** yield consistent gains when applied in isolation, as visually evidenced by their systematic improvements over the baseline in Fig. 3–5. Specifically, GSL enhances training supervision by replacing rigid one-hot targets with graph-guided soft labels, while GLR regularizes the prediction space by encouraging consistency between the original and graph-smoothed outputs.

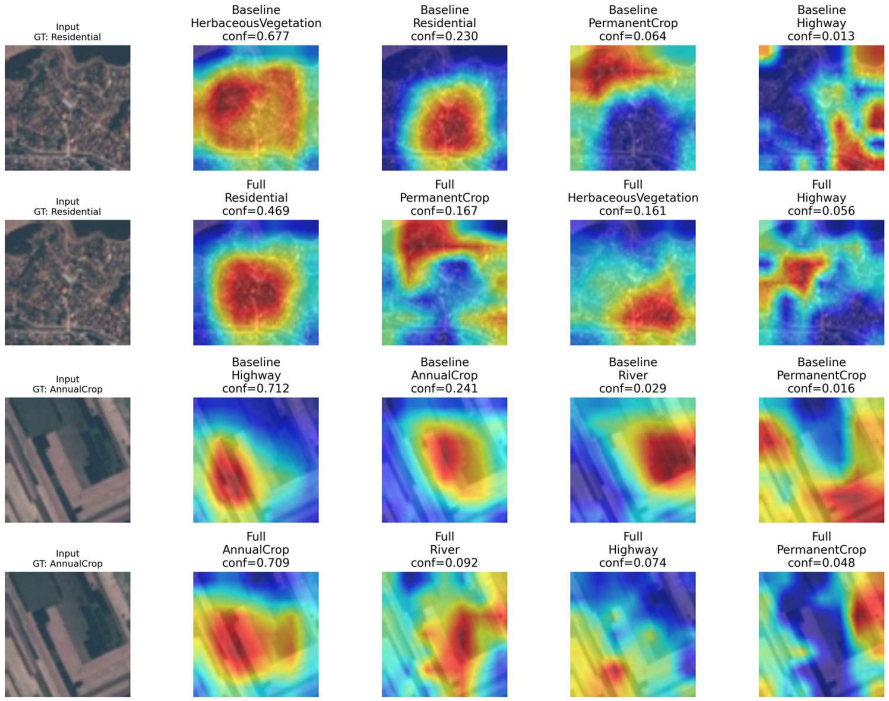


Fig. 6 Qualitative comparison of class activation maps (CAMs) on the EuroSAT dataset. For each image, the first column shows the input image and its ground-truth label. The top row visualizes the top-4 predictions of the baseline model, while the bottom row shows the top-4 predictions of the proposed full model, with classes sorted by confidence. All visualized samples are randomly selected from images misclassified by the baseline model.

When combined, the **Full** model generally achieves the most stable and strongest performance across metrics and backbones. The complementary effects of supervision-level (GSL) and prediction-level (GLR) semantic alignment are reflected by the consistently higher performance bars of the Full variant in the visual comparisons. These observations confirm that jointly exploiting label-graph information at both stages leads to more robust and transferable representations.

4.7 Qualitative Analysis

To further understand the effect of label-graph guidance beyond quantitative metrics, we provide qualitative comparisons between the baseline model and the proposed full model on the EuroSAT dataset. Fig. 6 visualizes representative examples that are *misclassified by the baseline model*, where we show the top-ranked predictions of each model together with their corresponding class activation maps (CAMs), sorted by confidence.

As shown in Fig. 6, when the baseline model fails, its high-confidence predictions often rely on spatially diffuse activation patterns that cover broad regions of the image, suggesting a stronger dependence on coarse textures or local visual cues.

In contrast, the full model consistently produces more compact and structure-aware activation maps, even when conditioned on non-ground-truth classes. Notably, for the ground-truth category, the full model tends to focus on regions that better correspond to meaningful scene layouts, while suppressing spurious responses to visually correlated but semantically distinct classes. These qualitative observations complement the quantitative results and indicate that the proposed label-graph guidance improves semantic disentanglement and leads to more interpretable and discriminative decision evidence, particularly in challenging cases where the baseline model fails.

5 Conclusion

This paper investigated the role of inter-class semantic relationships in multi-class remote sensing image recognition and presented a label-graph guided semantic alignment framework. Different with conventional approaches that rely solely on one-hot supervision and assume label independence, the proposed method explicitly incorporates a label graph as a lightweight semantic prior to guide model learning.

The proposed framework introduces two complementary components. Graph-Soft supervision replaces rigid one-hot targets with graph-guided soft labels, enabling the model to account for semantic similarity among classes during training. Graph-Logit regularization further enforces prediction-level consistency by aligning the model output distribution with its graph-smoothed counterpart. Both components operate purely in the label space and are fully compatible with standard CNN backbones, requiring no architectural modification or additional annotations.

Extensive experiments on multiple remote sensing benchmarks demonstrate that explicitly modeling label dependencies leads to consistent performance improvements in most settings. In particular, the proposed approach is especially effective when training from scratch or under challenging semantic ambiguity, while maintaining stable behavior across different backbone architectures. We also observe that the effectiveness of label-graph guidance may vary with model capacity and initialization strength, highlighting the importance of careful graph construction and integration.

Future work will explore adaptive and sample-aware label graph construction strategies, as well as extending the proposed framework to multi-label and hierarchical remote sensing recognition tasks. We believe that incorporating structured semantic priors into supervision and prediction remains a promising direction for improving robust remote sensing image understanding.

Declarations

Funding

This work was supported by the Zhejiang Office of Philosophy and Social Science under Project No. 23NDJC399YBM.

Code availability

Name of the code/library: GLSGLR (Graph-based Label Softening and Logit Regularization)

Contact: E-mail: `zhoukun@student.usm.my`

Hardware requirements: A CUDA-enabled GPU is recommended for efficient training. The code has been tested on a single NVIDIA GPU with at least 8 GB of memory. CPU-only execution is possible but significantly slower.

Program language: Python ($\geq$ 3.8)

Software required: PyTorch, torchvision, numpy, scikit-learn, albumentations, tqdm, opencv-python, and the `datasets` library (required for the HuggingFace dataset backend).

Program size: Approximately a few megabytes, excluding datasets and trained model checkpoints.

Availability: The source code supporting the findings of this study is publicly available at: https://github.com/zhoukuniyc/GLSGLR/tree/master

References

[1] Thapa, A., Horanont, T., Neupane, B., Aryal, J.: Deep learning for remote sensing image scene classification: A review and meta-analysis. Remote Sensing **15**(19), 4804 (2023)

[2] Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE geoscience and remote sensing magazine **5**(4), 8–36 (2017)

[3] Zhang, L., Zhang, L., Du, B.: Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and remote sensing magazine **4**(2), 22–40 (2016)

[4] Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)

[5] Koller, C., Kauermann, G., Zhu, X.X.: Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in earth observation? IEEE Transactions on Geoscience and Remote Sensing **62**, 1–11 (2023)

[6] Xiong, Y., Zhang, P., Dou, Y., Xu, K., Niu, X.: Focus on hard categories and hard examples: Remote sensing image scene classification via expert model and hard example mining. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2021)

[7] Gu, X., Zhang, C., Shen, Q., Han, J., Angelov, P.P., Atkinson, P.M.: A self-training hierarchical prototype-based ensemble framework for remote sensing scene classification. Information Fusion **80**, 179–204 (2022)

[8] Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S.: Remote sensing image scene

classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 3735–3756 (2020)

[9] Tang, X., Ma, Q., Zhang, X., Liu, F., Ma, J., Jiao, L.: Attention consistent network for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 2030–2045 (2021)

[10] Yang, Y., Jiao, L., Liu, F., Liu, X., Li, L., Chen, P., Yang, S.: An explainable spatial–frequency multiscale transformer for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)

[11] Zhang, X., Sheng, Z., Shen, H.-L.: Focusnet: Classifying better by focusing on confusing classes. Pattern Recognition **129**, 108709 (2022)

[12] Ma, B., Wu, F., Hu, T., Fathollahi, L., Sui, X., Liu, Y., Gantumur, B.: Label-driven graph convolutional network for multilabel remote sensing image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 2245–2255 (2023)

[13] Sun, S., Dustdar, S., Ranjan, R., Morgan, G., Dong, Y., Wang, L.: Remote sensing image interpretation with semantic graph-based methods: A survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 4544–4558 (2022)

[14] Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems **32** (2019)

[15] Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)

[16] Qi, X., Zhu, P., Wang, Y., Zhang, L., Peng, J., Wu, M., Chen, J., Zhao, X., Zang, N., Mathiopoulos, P.T.: Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISPRS Journal of Photogrammetry and Remote Sensing **169**, 337–350 (2020)

[17] Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[19] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[20] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[21] Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)