

REVIEW ARTICLE

Scene Classification Using Deep Learning: A Critical Review of Representations, Benchmarks, and Evaluation

Sandeep Kumar Aazad ^{a,*}, Kritika Chaudhary ^b, Taniya Saini ^c, Ashok Ajad ^c, L N Sastry Varanasi ^d

^aB. Tech (2021-25) in Electrical and Electronics Engineering, National Institute of Technology Andhra Pradesh, Tadepalligudem, Andhra Pradesh 534 101, India.

^bB. Tech (2025-29) in Electronics and Communication Engineering, College of Engineering Roorkee, Uttarakhand 247 667, India.

^cM. Tech (2016-18) in Computer Science Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand 826 004, India.

^dFaculty of Electrical Engineering Department, National Institute of Technology Andhra Pradesh, Tadepalligudem, Andhra Pradesh 534 101, India.

Abstract

Scene classification assigns semantic categories to images by integrating information about objects, spatial layout, texture, and environmental context. Deep learning has shifted this task from manually designed descriptors toward transferable representations, but reported improvements remain difficult to interpret when datasets, supervision, and evaluation protocols differ. This narrative review examines deep learning for static visual scene classification across indoor, outdoor, and aerial imagery. It organizes the literature along three complementary axes: learning signal, representation architecture, and prediction task. The discussion connects convolutional feature extraction, local and global feature aggregation, semantic and graph-based reasoning, and generative representation learning with selected developments in vision transformers, self-supervised pretraining, and vision-language models. Benchmark characteristics are examined alongside the distinction between single-label, multi-label, and zero-shot evaluation. A source-checked numerical example illustrates why domain-specific pretraining can improve some benchmarks while leaving others unchanged or worse. The synthesis identifies dataset bias, incomplete split reporting, spatial leakage, semantic ambiguity, and deployment cost as persistent obstacles to reliable comparison. It also establishes a reporting framework covering class-wise performance, uncertainty, computational resources, and distribution shift. The review concludes that useful progress should be assessed through reproducible gains under matched protocols and credible transfer to new environments, rather than isolated accuracy values.

Keywords: Scene classification, Convolutional neural networks, Vision transformers, Remote sensing, Representation learning, Vision-language models, Evaluation protocols.

Article information:

ISSN: 3107-9466 (Online)

Published by: **Krrish Scientific Publications Pvt. Ltd.**

DOI: <https://doi.org/10.71426/jcdt.v2.i2.pp182-192>

Received: 21 Jul. 2026 | Revised: 09 Oct. 2026 | Accepted: 15 Oct. 2026 | Published: 26 Oct. 2026

Copyright ©2026 Author(s).

This is an open-access article distributed under the terms of the [Creative Commons Attribution-NonCommercial 4.0 International License \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

A scene category describes an environment through the configuration and interaction of its visual components. A kitchen and a dining room may contain similar furniture,

yet differ in object arrangement, functional equipment, and the relative prominence of particular regions. Likewise, aerial images of industrial and residential areas may share roofs and roads while differing in spatial density and land-use structure. Scene classification therefore requires more than recognizing a dominant object. It combines local evidence with spatial and semantic context, as reflected in scene-centric benchmarks such as SUN and Places [1], [2].

The task is usually formulated as image-level prediction over a predefined label set. Its applications include image retrieval, indoor robotic perception, road-environment in-

*Corresponding author

Email addresses: sandeepkumaraazad622@gmail.com (Sandeep Kumar Aazad), ckritika49@gmail.com (Kritika Chaudhary), tania.gagiyan1@gmail.com (Taniya Saini), cse.aa9@gmail.com (Ashok Ajad), sastry.ramyasri@gmail.com (L N Sastry Varanasi).

terpretation, and remote sensing. These applications do not impose identical requirements. A retrieval system may prioritize semantic similarity; a mobile robot may require calibrated confidence and low latency; a land-use model must cope with geographic variation and changing acquisition conditions. Evaluation should reflect these differences rather than treating a single benchmark score as evidence of universal usefulness.

Early scene-recognition pipelines relied on engineered descriptors and classifiers. Deep learning changed the representation stage by allowing features to be learned from data and transferred between tasks. Convolutional network features can encode both object appearance and scene structure, but their usefulness depends on the pretraining source, spatial scale, and downstream protocol [3], [4]. Deep architectures do not remove preprocessing or model-selection requirements, and additional layers do not guarantee higher accuracy. The effective learning problem still depends on annotation quality, sample diversity, optimization, and the relationship between training and deployment distributions.

An enduring challenge is to reconcile global scene structure with local discriminative content. Distinctive parts can identify a scene when whole-image appearance is ambiguous [5]. Conversely, an isolated chair or vehicle rarely determines a unique environment. Approaches based on region aggregation, semantic segmentation, objectness, and layout graphs address different aspects of this tension [6]–[9]. Their architectural differences should be interpreted in terms of the information each method preserves or discards, not simply as successive entries in an accuracy ranking.

Remote sensing introduces additional variation in ground sampling distance, orientation, geographic location, and sensor characteristics. The NWPU-RESISC45 benchmark and a subsequent deep-learning survey document the importance of intra-class diversity and inter-class similarity in this domain [10], [11]. Small, repeatedly reused datasets can support controlled experiments but provide limited evidence of geographic generalization. Furthermore, a tile assigned one land-use label may contain several land-cover components; single-label and multi-label formulations consequently answer different questions.

More recent representation-learning methods broaden the available supervision. Vision transformers learn from image patches, self-supervised models exploit unlabeled images, and vision-language models associate visual representations with textual descriptions [12], [14]. These developments change the comparison problem: a frozen pretrained encoder, a fully fine-tuned classifier, and a zero-shot text-matching system have different access to target labels and different computational costs. They should be compared within clearly identified evaluation regimes.

This review develops a critical account of these relationships. Its contributions are a taxonomy that separates supervision from architecture; a synthesis of mechanisms for combining global, local, and contextual information; a benchmark and metric framework that exposes common comparison errors; and a research agenda centered on transfer, uncertainty, and deployment. The emphasis is on static visual scenes. Acoustic scene classification, wearable-sensor activity recognition, medical-image diagnosis, and text localization are distinct tasks and are not pooled with visual scene results. Video and temporal remote sensing appear

only when they clarify transfer or pretraining considerations.

2. Review scope and evidence handling

2.1. Narrative design and coverage

This article is a critical narrative review rather than a systematic review or a meta-analysis. Its historical core concerns scene-classification methods represented in the 2012–2021 literature, supplemented by earlier benchmark papers and selected developments from 2021–2025. The later material provides conceptual coverage of transformers, self-supervision, and vision-language transfer; it does not constitute an exhaustive census of recent publications.

The starting corpus was the bibliography and study descriptions in the supplied manuscript. That manuscript named several search services but did not provide reproducible database exports, exact query histories, screening decisions, or a deduplication record. Accordingly, the claimed total of 73 screened studies and the associated annual-publication chart are not used as evidence. No retrospective screening counts or PRISMA-style flow totals are inferred. This distinction matters because a bibliography assembled for narrative discussion cannot establish publication prevalence or systematic coverage.

Studies were retained for discussion when they addressed image-level visual scene categories, transferable scene representations, or a benchmark directly relevant to that task. Adjacent methods are discussed only to explain a specific mechanism, such as semantic segmentation supporting scene recognition. A targeted update inspected primary papers and official benchmark descriptions using combinations of scene classification, remote sensing, vision transformer, self-supervised learning, and vision-language model terms. Foundational general-purpose models were included to explain their relevance to scene representation; their success on other tasks is not treated as a measured scene-classification gain.

2.2. Comparison unit and strength of evidence

The appropriate comparison unit is a method evaluated under a specified dataset, split, supervision regime, and metric. Architectural summaries and quantitative evidence serve different purposes. A methodological description can explain a design choice even when the supplied account lacks enough protocol detail for numerical ranking. Quantitative claims require a traceable source and a clear interpretation of the reported measure. For this reason, the numerical example in Section 7 is drawn from a single primary study, while the broader literature table synthesizes mechanisms and limitations.

The review distinguishes published observations from its own analysis. It does not report newly trained models, reproduced experiments, statistical significance, or a pooled effect size. Where older study descriptions contained ambiguous dataset names, displaced reference numbers, or incomplete training ratios, their accuracy values were excluded from the comparative narrative. A separate editorial audit records the treatment of the original bibliography so that omission does not erase the provenance of the supplied material.

Table 1: Evidence fields needed to interpret a scene-classification result.

Field	Why it affects interpretation
Task and label space	Single-label categories, multiple labels, and unknown-class rejection have different prediction and scoring rules.
Dataset version and split	Sample counts alone do not identify the images available for training, selection, and testing.
Pretraining and target labels	Zero-shot, linear-probe, few-shot, and fine-tuned results involve different supervision budgets.
Input and inference policy	Resolution, crops, region proposals, and model ensembles can materially change both accuracy and cost.
Metric and uncertainty	Overall accuracy, mean-class accuracy, and macro-F1 weight errors differently; variation across runs must be stated.
Compute and transfer	Hardware, latency, memory, and held-out domains determine whether a result supports the intended use.

3. Problem formulation and taxonomy

3.1. Single-label, multi-label, and open-set prediction

Let x_i denote an image and $\mathcal{Y} = \{1, \dots, C\}$ a set of scene categories. A single-label model produces class probabilities $p_\theta(c | x_i)$ and predicts

$$\hat{y}_i = \arg \max_{c \in \mathcal{Y}} p_\theta(c | x_i). \quad (1)$$

A common supervised objective is cross-entropy (3):

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \log p_\theta(y_i | x_i). \quad (2)$$

This formulation assumes that each evaluated image has one relevant label in the predefined vocabulary. The assumption is convenient for benchmarking, but it can obscure mixed land use or overlapping functional environments.

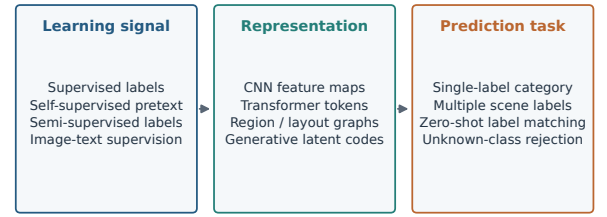
For multi-label prediction, $\mathbf{y}_i \in \{0, 1\}^C$ allows several labels to be active. Independent sigmoid outputs q_{ic} can be trained using binary cross-entropy (3):

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C [y_{ic} \log q_{ic} + (1 - y_{ic}) \log (1 - q_{ic})]. \quad (3)$$

Label thresholds must be selected using validation data. Multi-label prediction is not equivalent to assigning one category from each of several attribute groups, such as road location, weather, and time of day; the latter can be implemented as a multi-task model with separate output heads [15], [16].

Open-set recognition permits rejection when a sample does not belong to the known categories. This differs from zero-shot classification, where labels can be specified through text without fitting a target-specific classifier. A zero-shot system may still force every image into one supplied label and therefore need not detect unknown environments. Open-world evaluation additionally considers adaptation over time and retention of previously learned categories. These distinctions are increasingly relevant to remote sensing, as illustrated by OpenEarthSensing [17].

Three complementary axes for organizing scene classifiers



Example: self-supervised ViT features + supervised linear scene classifier

Figure 1: A multi-axis taxonomy of visual scene classification.

3.2. Architecture and supervision are separate axes

CNNs, recurrent networks, transformers, graph networks, and autoencoders describe architectural or modeling choices. Supervised, self-supervised, semi-supervised, and image-text learning describe the training signal. These categories intersect: a CNN can be trained with labels or a self-supervised objective, and an autoencoder representation can support a supervised classifier. An RNN is not inherently an unsupervised or generative method; likewise, a deep neural network is an umbrella term rather than a separate hybrid category.

Figure 1 uses this separation to avoid a misleading hierarchy. Generative modeling characterizes a learning objective or probabilistic formulation, while transfer learning describes reuse across data or tasks. A useful method description therefore identifies the encoder, training signal, feature aggregation, prediction head, and adaptation protocol together.

4. Deep representations for scene classification

4.1. Global CNN features and transfer learning

A global CNN pipeline maps an image to a feature vector and learns a classifier over that representation. The encoder may remain frozen, be partially adapted, or be fine-tuned end to end. Early scene-classification studies using AlexNet and related architectures demonstrated the practical value of transferring learned features [18], [19], [3]. The key distinction is between feature transfer and full retraining: a lightweight classifier on a frozen representation requires fewer target-specific parameters, whereas

fine-tuning can better adapt features when sufficient representative data are available.

Pretraining on objects and pretraining on scenes supply different inductive information. Object-centric features can emphasize localized entities, while scene-centric features can better encode the environmental context surrounding them. Herranz et al. [4] show that scale interacts with this source bias. A multiscale model should therefore explain why a particular representation is appropriate at each scale. Concatenating features from several networks without controlling capacity and pretraining does not isolate the contribution of scale modeling.

For remote sensing, transfer learning also has to accommodate differences between ground-level photographs and overhead imagery. Li et al. [20] investigate pretrained deep CNNs for high-spatial-resolution scenes. The relevant question is not whether transfer is universally preferable, but which parameters and training strategy are appropriate for the available target data. If only a small labeled set is available, a frozen or partially adapted encoder offers a useful baseline against which more elaborate methods should be assessed.

Several variants alter the computational or training representation. Gradient-boosting random convolutional networks combine learned components within an ensemble framework [21], while compressed-domain classification examines how remote-sensing representations can be extracted within a JPEG 2000 workflow [22]. These choices should be evaluated jointly with preprocessing and inference cost. Compression may change the input information itself, so accuracy and efficiency should be measured under specified compression settings rather than treated as independent properties.

4.2. Local regions, pooling, and multiscale aggregation

Global pooling offers a compact representation but may suppress a small, diagnostic object. Local methods retain evidence from regions or patches, then aggregate it into an image-level descriptor. Distinctive-part approaches and discriminative meta-object discovery address this limitation by selecting informative components [5], [6]. Their benefit depends on whether region selection retains meaningful context and rejects repetitive or irrelevant patches.

Pooling determines which relationships survive aggregation. Orderless representations offer tolerance to rearrangement but can lose information about layout; spatial partitions retain coarse location at the cost of sensitivity to alignment. Semantic Fisher vectors aggregate local semantic descriptors [23]. Related Fisher-kernel work investigates dependencies among local observations [24]. Shared-codebook patch representations offer another route to combining local feature information [25]. These mechanisms are relevant because two scenes can contain similar objects but differ in how those objects are organized.

Multiscale pooling and context modeling attempt to preserve both local detail and broader structure [26], [27]. Their evaluation should distinguish the effect of additional information from the effect of additional computation. A model that evaluates many crops or region proposals receives a larger inference budget than a single-pass baseline. Reporting only the final accuracy conceals this difference.

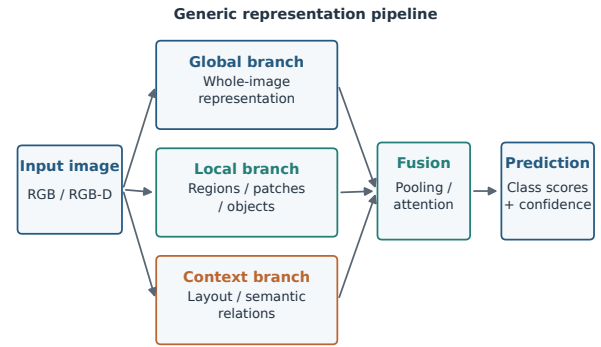


Figure 2: Generic pathways for combining whole-image, local, and contextual evidence.

A fair ablation keeps the backbone and pretraining source fixed while changing the aggregation mechanism, then reports both classification and resource measures.

Saliency-guided fusion provides an explicit mechanism for emphasizing selected image regions [28]. However, saliency does not necessarily identify the evidence required by the scene label. A visually prominent object may be shared across categories, while weak background structure may be decisive. Thus, qualitative saliency maps should be accompanied by controlled occlusion or region-removal experiments before they are interpreted as explanations of model behavior.

4.3. Semantic context, attributes, and layout graphs

Semantic representations convert low-level appearance into objects, regions, or attributes. Semantic-aware recognition combines RGB features with information from semantic segmentation [7]; attribute learning emphasizes properties complementary to object and scene features [29]. These approaches make the representation more structured, but introduce dependence on the quality and vocabulary of upstream predictors. A segmentation model trained in one domain may omit categories or boundaries that are relevant in another.

Objectness-based recognition filters local evidence according to its usefulness for scene discrimination [8]. Layout graphs take a further step by modeling relations among selected regions. Prototype-agnostic scene layout constructs image-specific structure and uses graph-based processing to integrate spatial and semantic relations [9]. This provides a principled alternative to assuming a fixed arrangement for every example of a class. Its limitations include the computational cost of region construction and the possibility that incorrect region relationships propagate through the graph.

Joint object detection, scene classification, and segmentation can exploit complementary predictions [30]. Nevertheless, task coupling does not guarantee better scene recognition: a dominant auxiliary objective can divert capacity away from the target task. Evaluations should compare the joint system with a matched scene-only baseline and disclose the additional annotations it uses. Figure 2 summarizes the representation pathways without claiming that every reviewed method contains all branches.

4.4. Generative and semi-supervised representations

Generative representation learning is motivated in part by the expense of scene annotation. Deep belief networks have been used to construct and select features for remote-sensing classification [31]. Attention GANs extend unsupervised feature learning with attention mechanisms for aerial scenes [32]. These methods should be described in terms of their representation objective and downstream classifier, rather than assuming that an unsupervised pre-training stage eliminates label requirements throughout the pipeline.

A central issue is the alignment between the learning objective and the target semantics. Reconstruction can preserve texture and illumination without necessarily separating scene categories. Adversarial learning can encourage distributional realism without establishing reliable class boundaries. Therefore, evaluation should examine whether the learned features improve a fixed downstream classifier and whether the gains persist when pretraining data, backbone capacity, and label budgets are controlled.

Semi-supervised learning combines labeled and unlabeled target data. Han et al. [33] investigate a generative framework with deep features for high-resolution remote sensing. Such methods can expand the effective training set, but inferred labels may amplify early mistakes. Confidence thresholds, class balance, and the source of unlabeled images become part of the method. Unlabeled access to the eventual test distribution must be disclosed, since transductive adaptation is a different setting from inductive evaluation on unseen environments.

4.5. Hybrid architectures and feature supervision

The term hybrid covers several distinct mechanisms. Locally supervised deep hybrid models inject supervisory information into intermediate features [34]; knowledge-guided multiresolution models address ambiguity across resolutions [35]; multi-instance learning combines evidence from several regions of an image [36]. Attention pooling similarly changes how spatial evidence contributes to the final representation [37]. These methods should be compared according to the problem they address, rather than grouped merely because they combine components.

CNN-recurrent combinations can model a sequence of spatial features, as in Inception-LSTM scene classification [38]. For static images, however, the ordering of that sequence is a modeling choice, not an observed temporal process. Sensitivity to traversal order or patch arrangement should therefore be tested. An RNN using hand-engineered descriptors [39] also differs fundamentally from an end-to-end recurrent visual representation, even if both use recurrent units.

Architecture optimization introduces a further evaluation burden. Multi-objective search for UAV scene classification [40] is motivated by competing accuracy and computational demands. Search time, tuning resources, and the final deployment configuration must all be reported. A model selected through extensive optimization should not be compared with an untuned baseline without acknowledging the unequal search budget.

5. Transformers, self-supervision, and vision-language transfer

5.1. Patch-based visual representations

Vision transformers process image patches as tokens and use attention to exchange information across them [12]. This offers a different mechanism for modeling context from convolutional feature hierarchies. For scene recognition, patch relationships can represent long-range structure, but the usefulness of that structure depends on token resolution and pretraining. Small diagnostic objects may be poorly represented when patches are too coarse.

The practical comparison is consequently broader than CNN versus transformer. Models should be matched or clearly characterized by input resolution, pretraining data, parameter count, and inference policy. A transformer pre-trained on a much larger corpus cannot establish an architectural advantage over a CNN trained only on the target benchmark. Conversely, a well-adapted convolutional model remains a relevant baseline when latency or memory constrains deployment. Hybrid or hierarchical architectures should be assessed through the same resource-aware criteria.

5.2. Self-supervised learning and domain structure

DINOv2 demonstrates that self-supervised pretraining on curated images can yield reusable visual features [13]. In scene-classification studies, this motivates comparing frozen-feature evaluation with end-to-end adaptation. A linear probe measures the accessibility of target-label information in the representation, while fine-tuning measures the combined effect of representation quality and adaptation. Reporting only the stronger result obscures this distinction.

Remote-sensing pretraining can exploit information absent from conventional RGB photographs. SatMAE incorporates temporal and multispectral structure through masked pretraining [41]. Its relevance to static scene classification is primarily as a source of transferable features; temporal or spectral inputs must not be silently treated as equivalent to one RGB image. A comparison should state which modalities are used during pretraining and which are required at inference. This is especially important when a deployment sensor cannot supply all of the information available during model development.

The research question is whether pretraining learns features that remain useful under changes in geography, season, or sensor. In-domain pretraining is not sufficient evidence of this property. A stronger design reserves independent regions or acquisition groups for evaluation, tests the sensitivity to the number of labeled examples, and reports the performance of rare classes. These are review recommendations, not claims that all self-supervised methods already satisfy them.

5.3. Language-aligned scene representations

CLIP trains image and text encoders to align paired representations and enables label descriptions to define a zero-shot classifier [14]. For scene analysis, this changes the label interface: categories can be expressed as descriptions rather than as only numerical identifiers. Prompt wording

and the candidate label set become part of the evaluation protocol. A label description selected using test performance constitutes adaptation and should not be presented as an untouched zero-shot setting.

For normalized image and text embeddings $\mathbf{v}(x)$ and $\mathbf{t}_c$, a typical score is expressed as in (4).

$$s_c(x) = \frac{\mathbf{v}(x)^\top \mathbf{t}_c}{\tau}, \quad \hat{y} = \arg \max_c s_c(x), \quad (4)$$

where $\tau > 0$ is a temperature. Softmax scores over these similarities depend on the supplied candidate set; they are not automatically calibrated probabilities of correctness. Open-set rejection therefore requires an explicit decision rule and a separate evaluation.

RemoteCLIP adapts image-text learning to remote sensing by converting heterogeneous annotations into caption-based supervision [42]. This supports discussion of domain adaptation, but it does not establish uniformly better performance across datasets. The within-study example in Section 7 makes that limitation visible. Language alignment also introduces annotation and vocabulary biases, so tests should include prompt stability and semantically neighboring classes.

6. Benchmarks and application requirements

6.1. Scene-centric datasets

Table 2 summarizes selected benchmarks and distinguishes image collection size from evaluation size. MIT Indoor67 contains 15,620 images, while its official experiment uses 80 training and 20 testing images per category [43]. SUN397 contains 397 categories and 108,754 images according to the benchmark source [1]. These datasets probe complementary aspects of scene recognition: object-rich indoor environments and a broader set of scene categories.

The Places work describes a larger scene-centric resource [2]. Its overall collection should not be conflated with a specific Places365 release, training subset, or validation split. Dataset names alone are insufficient to establish reproducibility. Authors should identify the release, class vocabulary, sample lists, and whether any filtering was performed.

6.2. Remote sensing, multiple labels, and modality

AID contains 30 aerial scene categories and 10,000 images [44]. NWPU-RESISC45 expands category coverage and image diversity [10]. These collections support standardized experiments, but random image splits alone do not measure transfer to a new geographic region. Tiles may share acquisition conditions or nearby structures, and repeated sampling from a familiar distribution can overstate deployment performance. Geographic holdouts are a complementary test, not a substitute silently imposed on published benchmark protocols.

MLRSNet addresses multi-label semantic understanding in high-spatial-resolution remote sensing [45]. Its motivation illustrates why land-use scene labels and land-cover labels should not be merged indiscriminately. An image may depict an airport as a scene while containing roads,

buildings, and vegetation as component labels. Data augmentation for multi-label land-cover categorization also requires transformations that preserve the validity of all labels [46]. The report must specify the label ontology and the way overlapping labels are scored.

Depth information can help separate appearance from geometry in indoor settings. RGB-D scene methods [47] must nonetheless handle missing or noisy depth and distinguish measured depth from derived encodings. Dataset graphics should not imply that every RGB benchmark includes depth or geometric annotations. Modality-specific preprocessing, alignment, and missing-data policies are necessary parts of a reproducible comparison.

6.3. Application-specific interpretation

For road environments, scene-level predictions describe context and can complement more localized perception outputs [15], [16]. They do not replace lane detection, object detection, or motion estimation. A whole-image class can be correct while a safety-relevant local feature is missed. Evaluation should therefore separate the scene-classification endpoint from downstream system performance.

For aerial emergency response, category recognition may assist image triage [48], but practical usefulness also depends on image quality, latency, and coverage of unfamiliar conditions. For forensic image retrieval, semantic categorization can support organization of visual records [49], [50]; it should not be interpreted as independent determination of a criminal event. These applications demonstrate why task boundaries matter: classification performance must be tied to the decision it actually supports.

7. Critical comparison of the evidence

7.1. What the methodological literature supports

Table 3 compares representative mechanisms. It does not rank methods across incompatible experiments. The literature supports several recurring design principles: retain diagnostic local evidence, preserve enough context to disambiguate shared objects, adapt the representation to the image domain, and use the available supervision explicitly. These principles can be implemented through different network families, and architectural complexity alone is not evidence of a better solution.

A recurring confound is the mixture of several changes in one proposed pipeline. For example, a new pooling mechanism may be introduced together with a different backbone, higher input resolution, extra pretraining data, and an ensemble classifier. The resulting improvement cannot be attributed uniquely to pooling. A useful ablation changes one component at a time where feasible and reports interactions when components are jointly necessary. The same reasoning applies to learned prompts, auxiliary segmentation branches, and graph-based context models.

7.2. A traceable numerical example

Table 4 and Figure 3 reproduce selected zero-shot accuracies from Table III of the RemoteCLIP paper, using the same ViT-L/14 backbone [42]. AID improves by 18.02

Table 2: Selected visual scene benchmarks. (Note: Collection sizes are not necessarily the number of images used by a reported experiment.)

Dataset	Categories	Images	Interpretation and protocol requirement
MIT Indoor67 [43]	67	15,620	Indoor RGB; official 80/20 train/test images per category.
SUN397 [1]	397	108,754	Diverse RGB scenes; specify the benchmark split files and aggregation across splits.
AID [44]	30	10,000	Aerial RGB; state the training fraction and repeated partitions.
NWPU-RESISC45 [10]	45	31,500	Aerial RGB, 700 images per class; disclose split ratio and random seeds.
Places [2]	Release-specific	About 10 million overall	Scene-centric collection; identify the exact release and its train/validation/test subsets.

Table 3: Mechanism-oriented comparison of representative scene-classification approaches. (Note: Limitations are the review’s analytical assessment, not measured rankings.)

Approach	Information retained or learned	Main comparison requirement
CNN transfer [3], [20]	Global visual features reused from a source task	Match pretraining, adaptation, and input resolution.
Local regions and pooling [6], [23], [27]	Diagnostic parts and aggregated patch evidence	Account for region count, crops, and lost spatial order.
Semantic and attribute features [7], [29]	Object, region, and attribute context	Disclose upstream labels and robustness to semantic errors.
Layout graphs [9]	Image-specific spatial and semantic relations	Isolate graph benefit from region selection and encoder capacity.
Generative / semi-supervised learning [32], [33]	Features learned with reduced direct labeling	Identify unlabeled data access and pseudo-label selection.
Hybrid and multi-instance models [34], [36]	Intermediate supervision and weighted local evidence	Use matched component ablations and inference budgets.
Self-supervised transformers [13], [41]	Transferable patch representations	Separate frozen probes from adaptation and modality effects.
Vision-language transfer [14], [42]	Alignment of images with category descriptions	Record prompts, label candidates, and target-label access.

Table 4: Published zero-shot accuracy (%) with ViT-L/14; source: Liu et al. [42], Table III. Differences are percentage points.

Dataset	CLIP	RemoteCLIP	Difference
EuroSAT	60.21	59.94	−0.27
RESISC45	69.23	79.84	+10.61
AID	69.88	87.90	+18.02

percentage points and RESISC45 by 10.61, whereas EuroSAT decreases by 0.27. These observations illustrate benchmark-dependent transfer, not a universal ordering of models. No new experiments or significance tests are reported here.

7.3. Why cross-paper accuracy rankings are unreliable

The numerical example is intentionally narrower than an aggregate leaderboard. Across papers, identical dataset names can conceal different training fractions, data filtering, and test-time processing. A supervised classifier with extensive target labels is not directly comparable to a zero-shot model. A mean-class score is also not interchangeable with overall accuracy when category frequencies differ.

An evidence table should therefore record the protocol before presenting a score. If the necessary details are un-

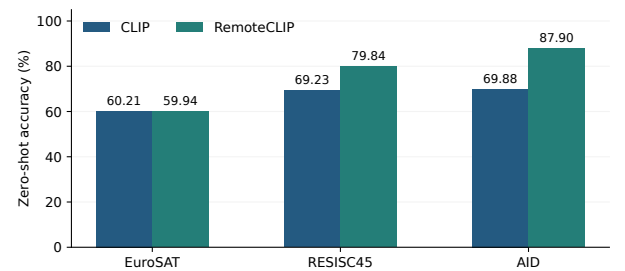


Figure 3: Selected published zero-shot results [42], redrawn from the values in Table 4.

available, the study can still contribute to methodological discussion, but its value should not determine a claimed winner. Missing uncertainty estimates further restrict interpretation: a difference of a few tenths of a percentage point may reflect partition or optimization variability. The review does not infer statistical significance from such point estimates.

8. Evaluation metrics and reproducibility

8.1. Class-wise and aggregate measures

For a single-label problem, define a confusion matrix M with true classes in rows and predicted classes in columns.

For class c , $TP_c = M_{cc}$, $FN_c = \sum_{j \neq c} M_{cj}$, and $FP_c = \sum_{i \neq c} M_{ic}$. A false negative is an example that belongs to class c but is predicted as another class. True negatives are all examples that belong to neither the true nor predicted positive set for the one-versus-rest problem. Figure 4 fixes the axis convention explicitly.

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad (5)$$

$$R_c = \frac{TP_c}{TP_c + FN_c}, \quad (6)$$

$$F_{1,c} = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \quad (7)$$

$$\text{Specificity}_c = \frac{TN_c}{TN_c + FP_c}. \quad (8)$$

The policy for zero denominators should be stated. Overall accuracy and macro-F1 are given as in 9.

$$\text{Accuracy} = \frac{\sum_c M_{cc}}{N}, \quad \text{Macro-F1} = \frac{1}{C} \sum_{c=1}^C F_{1,c}. \quad (9)$$

Mean-class accuracy is the average class recall, $C^{-1} \sum_c R_c$. It weights categories equally, while overall accuracy weights individual images equally. For imbalanced data, reporting both reveals whether improvements are concentrated in common categories. Macro-F1 is the mean of class-level F1 scores, not generally the harmonic mean of macro-precision and macro-recall.

Multi-label classification requires label-wise rather than mutually exclusive class counts. Macro- and micro-averaged F1, label-wise average precision, and exact-match accuracy answer different questions. Exact match requires every label of an image to be correct and can therefore be much more demanding than per-label accuracy. The aggregation convention and threshold policy must accompany the numerical result.

8.2. Ranking, calibration, and uncertainty

A receiver operating characteristic (ROC) curve varies the decision threshold and plots true-positive rate against false-positive rate. For multiclass problems, the one-versus-rest or one-versus-one construction and averaging rule must be specified. ROC-AUC uses confidence scores, not just hard predictions. In settings with rare positive labels, precision–recall analysis provides complementary information about the precision attainable at a given recall. The schematic ROC in Figure 4 illustrates the geometry only and contains no experimental data.

Confidence and correctness should be evaluated separately. A classifier can achieve high accuracy while assigning excessive confidence to its remaining errors. Calibration assessment should specify the scoring rule or reliability-bin construction and use held-out calibration data. Thresholds for abstention, unknown-scene rejection, or human review should be chosen before the final test and accompanied by the fraction of examples retained.

Repeated runs and independent partitions address different sources of variability. Multiple random seeds on one fixed split estimate optimization variability; repeated splits

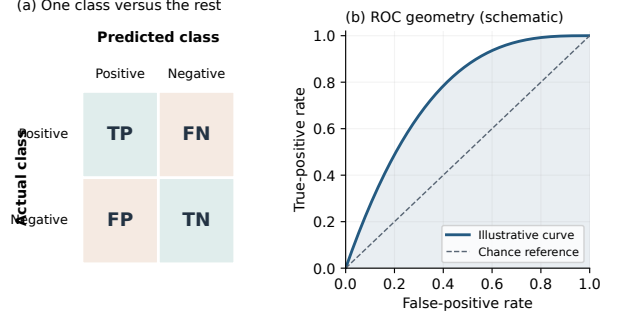


Figure 4: Evaluation conventions: (a) a binary confusion matrix with actual classes in rows; (b) an illustrative ROC curve.

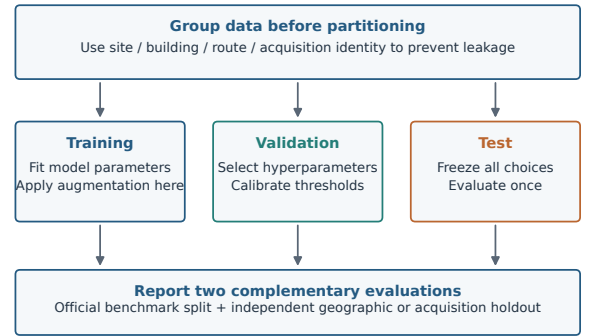


Figure 5: Recommended evaluation design for reproducible scene classification.

also expose sampling variability. Neither substitutes for an independent geographic or acquisition holdout. Confidence intervals should respect the sampling unit: resampling highly correlated tiles as if they were independent can produce overly narrow intervals. Group-level resampling is more appropriate when the deployment claim concerns unseen sites or scenes.

8.3. Preventing leakage and reporting resource use

Data should be grouped and partitioned before augmentation, patch extraction, or learned normalization whenever those operations could share information across subsets. Near-duplicate images, views of the same room, and neighboring aerial tiles should be examined explicitly. Pretrained models introduce an additional provenance question: public evaluation images may overlap with large training corpora. If overlap cannot be established or excluded, that limitation should be stated rather than interpreted as proof of contamination.

Figure 5 summarizes a recommended evaluation design. Official benchmark splits preserve comparability with prior work; independent group holdouts test a stronger transfer claim. Both are useful when their purposes remain distinct. Hyperparameter selection, early stopping, prompt selection, and threshold calibration belong to the training/validation process, leaving the test set for final evaluation.

Deployment reporting should include the hardware, software environment, precision, input resolution, batch size, and latency distribution. Parameter count and floating-point operation estimates do not uniquely determine la-

tency. Region generation, preprocessing, data movement, and auxiliary semantic models may dominate total execution time. Energy and memory measurements require stated boundaries, such as encoder-only versus complete-pipeline inference. Model compression is useful only if the resulting accuracy, calibration, and latency satisfy the intended operating conditions.

9. Research priorities and limitations

9.1. Generalization beyond familiar benchmark distributions

A priority is to evaluate transfer across buildings, cities, seasons, and sensors with enough metadata to explain failures. Domain adaptation and simulation-to-reality transfer [51] are relevant mechanisms, but their value depends on the target information available during adaptation. A model allowed to use unlabeled deployment data should be distinguished from a model evaluated without any target access. Source-only and adaptation-assisted results should be reported separately.

Open-world benchmarks such as OpenEarthSensing [17] motivate evaluation of semantic novelty and distribution shift alongside retention of known categories. A convincing system must both recognize familiar environments and handle uncertainty when its label vocabulary is incomplete. For practical assessment, rejection quality, adaptation cost, and catastrophic forgetting deserve attention alongside conventional closed-set accuracy.

9.2. Label efficiency and semantic ambiguity

Self-supervised pretraining, few-shot adaptation, and language supervision offer different ways to reduce manual labeling. They should be compared under explicit annotation budgets and with independent trials of the labeled subset. Selecting a particularly favorable few-shot sample can overstate reliability. Active learning should also account for the cost of obtaining labels, including cases in which land-use ambiguity requires expert interpretation.

Scene labels may encode visual appearance, human function, or administrative definitions. These meanings do not always agree. Hierarchical and multi-label annotations can expose that ambiguity, while a flat label space can conceal it as apparent model error. Future datasets should document category definitions, ambiguous cases, and annotation disagreement. New models can then be assessed on whether they represent this structure usefully rather than merely fitting a single adjudicated label.

9.3. Efficient context and reliable explanations

Global, local, and semantic branches increase the information available to a classifier but also increase the number of possible failure points. Future work should examine which branches are needed under particular operating conditions and whether conditional computation can reduce cost. The relevant target is a measured trade-off among recognition, confidence, and resource use, rather than the minimum parameter count alone.

Explanations should be tested for faithfulness. Attention maps, class-activation maps, and saliency overlays

can visualize internal responses but do not by themselves establish causal dependence. Region ablation, controlled background replacement, and counterfactual object edits can test whether a model relies on the intended evidence. Such interventions require care because unrealistic edited images may introduce artifacts unrelated to the semantic hypothesis.

9.4. Limits of this review

The review is selective and narrative. It does not estimate the prevalence of method families, provide an exhaustive current leaderboard, or reproduce the reviewed models. Much of the historical corpus emphasizes CNN-based scene recognition, while the later additions provide targeted coverage of new representation paradigms. The source-checked numerical comparison is illustrative and cannot establish a pooled advantage across the full literature. Older methodological references were reconciled by identity and topic, but not all underlying experiments were independently re-extracted. These limits should be considered when using the review to select baselines or frame new empirical studies.

10. Conclusion

Deep learning for scene classification has developed from transferable convolutional features toward structured context models and broader pretrained representations. The most useful distinction is not a single hierarchy of network names, but the interaction among learning signal, representation, label space, and evaluation protocol. Local aggregation, semantic context, and spatial relations address complementary weaknesses of whole-image recognition; self-supervised and language-aligned models alter the dependence on target annotations. None of these mechanisms removes dataset bias or guarantees reliable deployment. Reproducible progress requires matched comparisons, explicit treatment of label and split differences, and evaluation under realistic distribution changes. Future studies should connect classification quality with uncertainty, annotation cost, and computational resources so that benchmark improvements support clearly defined application claims.

Declarations and ethical statements

Conflict of interest: The author declares that there is no conflict of interest.

Funding statement: The author declares that no specific funding was received for this research.

Artificial Intelligence usage statement: During the preparation of this manuscript, the authors utilized ChatGPT solely for language refinement and grammatical corrections. The authors carefully reviewed and revised the generated content and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Availability of data and materials: No new image dataset or trained classifier was produced for this review. The accompanying source package contains editable figure-generation code, the numerical values used in the published-result illustration, and an editorial reference audit. Benchmark access points are provided in the cited dataset sources.

CRediT authorship contribution statement

Sandeep Kumar Aazad: Conceptualization, Methodology, Visualization, Writing- Original draft & editing. **Kritika Chaudhary:** Conceptualization, Methodology, Data-collection & curation. **Taniya Saini:** Conceptualization, Methodology, Data-collection & curation. **Ashok Ajad:** Conceptualization, Visualization, Data-collection & curation. **L N Sastry Varanasi:** Supervision, Formal analysis & Writing- editing.

Publisher's note

Krrish Scientific Publications Pvt. Ltd. and the *Journal of Computing and Data Technology* remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- [1] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010. p. 3485-3492. Available from: <https://doi.org/10.1109/CVPR.2010.5539970>
- [2] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018;40(6):1452-1464. Available from: <https://doi.org/10.1109/TPAMI.2017.2723009>
- [3] Koskela M, Laaksonen J. Convolutional Network Features for Scene Recognition. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014. p. 1169-1172. Available from: <https://doi.org/10.1145/2647868.2655024>
- [4] Herranz L, Jiang S, Li X. Scene Recognition with CNNs: Objects, Scales and Dataset Bias. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 571-579. Available from: <https://doi.org/10.1109/CVPR.2016.68>
- [5] Juneja M, Vedaldi A, Jawahar CV, Zisserman A. Blocks That Shout: Distinctive Parts for Scene Classification. In: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013. p. 923-930. Available from: <https://doi.org/10.1109/CVPR.2013.124>
- [6] Wu R, Wang B, Wang W, Yu Y. Harvesting Discriminative Meta Objects with Deep CNN Features for Scene Classification. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision*. 2015. p. 1287-1295. Available from: <https://doi.org/10.1109/ICCV.2015.152>
- [7] López-Cifuentes A, Escudero-Viñolo M, Bescós J, García-Martín Á. Semantic-aware scene recognition. *Pattern Recognition*. 2020;102:107256. Available from: <https://doi.org/10.1016/j.patcog.2020.107256>
- [8] Cheng X, Lu J, Feng J, Yuan B, Zhou J. Scene recognition with objectness. *Pattern Recognition*. 2018;74:474-487. Available from: <https://doi.org/10.1016/j.patcog.2017.09.025>
- [9] Chen G, Song X, Zeng H, Jiang S. Scene Recognition With Prototype-Agnostic Scene Layout. *IEEE Transactions on Image Processing*. 2020;29:5877-5888. Available from: <https://doi.org/10.1109/TIP.2020.2986599>
- [10] Cheng G, Han J, Lu X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*. 2017;105(10):1865-1883. Available from: <https://doi.org/10.1109/JPROC.2017.2675998>
- [11] Cheng G, Xie X, Han J, Guo L, Xia GS. 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*. 2020;13:3735-3756. Available from: <https://doi.org/10.1109/JSTARS.2020.3005403>
- [12] Dosovitskiy A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *Proceedings of the International Conference on Learning Representations*. 2021. Available from: <https://doi.org/10.48550/arXiv.2010.11929>
- [13] Oquab M, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*. 2024. Available from: <https://openreview.net/pdf?id=a68SUt6zFt>
- [14] Radford A, et al. Learning Transferable Visual Models From Natural Language Supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. 2021;139:8748-8763. Available from: <https://proceedings.mlr.press/v139/radford21a.html>
- [15] Duong TT, Nguyen TP, Jeon JW. Combined Classifier for Multi-label Network in Road Scene Classification. In: *Proceedings of the 2020 IEEE International Conference on Consumer Electronics - Asia*. 2020. p. 1-4. Available from: <https://doi.org/10.1109/ICCE-Asia49877.2020.9276981>
- [16] Chen L, Zhan W, Tian W, He Y, Zou Q. Deep Integration: A Multi-Label Architecture for Road Scene Recognition. *IEEE Transactions on Image Processing*. 2019;28(10):4883-4898. Available from: <https://doi.org/10.1109/TIP.2019.2913079>
- [17] Xiang X, et al. OpenEarthSensing: Large-Scale Fine-Grained Benchmark for Open-World Remote Sensing. *arXiv [Preprint]*. 2025;arXiv:2502.20668. Available from: <https://doi.org/10.48550/arXiv.2502.20668>
- [18] Sun J, Cai X, Sun F, Zhang J. Scene Image Classification Method Based on Alex-Net Model. In: *Proceedings of the 2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems*. 2016. p. 363-367. Available from: <https://doi.org/10.1109/ICCS.2016.7586482>
- [19] Sharma S, Juneja A, Sharma N. Using Deep Convolutional Neural Network in Computer Vision for Real-World Scene Classification. In: *Proceedings of the 2018 IEEE 8th International Advance Computing Conference*. 2018. p. 284-289. Available from: <https://doi.org/10.1109/IADCC.2018.8692121>
- [20] Li W, et al. Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2020;13:1986–1995. Available from: <https://doi.org/10.1109/JSTARS.2020.2988477>
- [21] Zhang F, Du B, Zhang L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Transactions on Geoscience and Remote Sensing*. 2016;54(3):1793–1802. Available from: <https://doi.org/10.1109/TGRS.2015.2488681>
- [22] Preethy Byju A, Sumbul G, Demir B, Bruzzone L. Remote-Sensing Image Scene Classification With Deep Neural Networks in JPEG 2000 Compressed Domain. *IEEE Transactions on Geoscience and Remote Sensing*. 2021;59(4):3458–3472. Available from: <https://doi.org/10.1109/TGRS.2020.3007523>
- [23] Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N. Scene Classification With Semantic Fisher Vectors. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. p. 2974–2983. Available from: <https://doi.org/10.1109/CVPR.2015.7298916>
- [24] Cinbis RG, Verbeek J, Schmid C. Approximate Fisher Kernels of Non-iid Image Models for Image Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016;38(6):1084–1098. Available from: <https://doi.org/10.1109/TPAMI.2015.2484342>
- [25] Jiang S, Chen G, Song X, Liu L. Deep Patch Representations with Shared Codebook for Scene Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*. 2019;15(1s):5:1–5:17. Available from: <https://doi.org/10.1145/3231738>
- [26] Song X, Jiang S, Herranz L. Multi-Scale Multi-Feature Con-

- text Modeling for Scene Recognition in the Semantic Manifold. *IEEE Transactions on Image Processing*. 2017;26(6):2721–2735. Available from: <https://doi.org/10.1109/TIP.2017.2686017>
- [27] Yoo D, Park S, Lee JY, Kweon IS. Multi-Scale Pyramid Pooling for Deep Convolutional Representation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2015. p. 71–80. Available from: <https://doi.org/10.1109/CVPRW.2015.7301274>
- [28] Shawky OA, Hagag A, El-Dahshan ESA, Ismail MA. A very high-resolution scene classification model using transfer deep CNNs based on saliency features. *Signal, Image and Video Processing*. 2021;15:817–825. Available from: <https://doi.org/10.1007/s11760-020-01801-5>
- [29] Zeng H, Song X, Chen G, Jiang S. Learning Scene Attribute for Scene Recognition. *IEEE Transactions on Multimedia*. 2020;22(6):1519–1530. Available from: <https://doi.org/10.1109/TMM.2019.2944241>
- [30] Yao J, Fidler S, Urtasun R. Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012. p. 702–709. Available from: <https://doi.org/10.1109/CVPR.2012.6247739>
- [31] Zou Q, Ni L, Zhang T, Wang Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters*. 2015;12(11):2321–2325. Available from: <https://doi.org/10.1109/LGRS.2015.2475299>
- [32] Yu Y, Li X, Liu F. Attention GANs: Unsupervised Deep Feature Learning for Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2020;58(1):519–531. Available from: <https://doi.org/10.1109/TGRS.2019.2937830>
- [33] Han W, Feng R, Wang L, Cheng Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2018;145:23–43. Available from: <https://doi.org/10.1016/j.isprsjprs.2017.11.004>
- [34] Guo S, Huang W, Wang L, Qiao Y. Locally Supervised Deep Hybrid Model for Scene Recognition. *IEEE Transactions on Image Processing*. 2017;26(2):808–820. Available from: <https://doi.org/10.1109/TIP.2016.2629443>
- [35] Wang L, Guo S, Huang W, Xiong Y, Qiao Y. Knowledge Guided Disambiguation for Large-Scale Scene Classification With Multi-Resolution CNNs. *IEEE Transactions on Image Processing*. 2017;26(4):2055–2068. Available from: <https://doi.org/10.1109/TIP.2017.2675339>
- [36] Ajjaji DA, Alsaed MA, Alsawayed AS, Alhichri HS. Multi-Instance Neural Network Architecture for Scene Classification in Remote Sensing. In: *Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS)*. 2019. p. 1–5. Available from: <https://doi.org/10.1109/ICCISCI.2019.8716411>
- [37] Bi Q, Qin K, Zhang H, Xie J, Li Z, Xu K. APDC-Net: Attention Pooling-Based Convolutional Network for Aerial Scene Classification. *IEEE Geoscience and Remote Sensing Letters*. 2020;17(9):1603–1607. Available from: <https://doi.org/10.1109/LGRS.2019.2949930>
- [38] Dong Y, Zhang Q. A Combined Deep Learning Model for the Scene Classification of High-Resolution Remote Sensing Image. *IEEE Geoscience and Remote Sensing Letters*. 2019;16(10):1540–1544. Available from: <https://doi.org/10.1109/LGRS.2019.2902675>
- [39] Verma A, Sharma S, Gupta P. RNN-LSTM Based Indoor Scene Classification with HoG Features. In: Luhach AK, Singh D, Hsiung PA, Hawari KB, Lingras P, Singh PK, editors. *Advanced Informatics for Computing Research*. Communications in Computer and Information Science. Vol. 955. Singapore: Springer; 2019. p. 149–159. Available from: https://doi.org/10.1007/978-981-13-3140-4_14
- [40] Rajagopal A, et al. A Deep Learning Model Based on Multi-Objective Particle Swarm Optimization for Scene Classification in Unmanned Aerial Vehicles. *IEEE Access*. 2020;8:135383–135393. Available from: <https://doi.org/10.1109/ACCESS.2020.3011502>
- [41] Cong Y, Khanna S, Meng C, Liu P, Rozi E, He Y, Burke M, Lobell DB, Ermon S. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. *Advances in Neural Information Processing Systems*. 2022;35. Available from: <https://doi.org/10.52202/068431-0015>. Preprint: <https://doi.org/10.48550/arXiv.2207.08051>
- [42] Liu F, Chen D, Guan Z, Zhou X, Zhu J, Ye Q, Fu L, Zhou J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*. 2024;62:5622216. Available from: <https://doi.org/10.1109/TGRS.2024.3390838>. Author manuscript: <https://arxiv.org/abs/2306.11029>
- [43] Quattoni A, Torralba A. Recognizing Indoor Scenes. In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. p. 413–420. Available from: <https://doi.org/10.1109/CVPR.2009.5206537>
- [44] Xia GS, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L, Lu X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2017;55(7):3965–3981. Available from: <https://doi.org/10.1109/TGRS.2017.2685945>
- [45] Qi X, Zhu P, Wang Y, Zhang L, Peng J, Wu M, Chen J, Zhao X, Zang N, Mathiopoulos PT. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2020;169:337–350. Available from: <https://doi.org/10.1016/j.isprsjprs.2020.09.020>
- [46] Stivaktakis R, Tsagkatakis G, Tsakalides P. Deep Learning for Multilabel Land Cover Scene Categorization Using Data Augmentation. *IEEE Geoscience and Remote Sensing Letters*. 2019;16(7):1031–1035. Available from: <https://doi.org/10.1109/LGRS.2019.2893306>
- [47] Pujar K, Chickerur S, Patil MS. Combining RGB and Depth Images for Indoor Scene Classification Using Deep Learning. In: *Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. 2017. p. 1–8. Available from: <https://doi.org/10.1109/ICCIC.2017.8524231>
- [48] Kyrkou C, Theodorides T. Deep-Learning-Based Aerial Image Classification for Emergency Response Applications Using Unmanned Aerial Vehicles. In: *Proceedings of the CVPR International Workshop on Computer Vision for UAVs*. 2019. Available from: <https://doi.org/10.48550/arXiv.1906.08716>
- [49] Abu Hana RO, Freitas COA, Oliveira LS, Bortolozzi F. Crime scene classification. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*. 2008. p. 419–423. Available from: <https://doi.org/10.1145/1363686.1363789>
- [50] Liu Y, Wang F, Hu D, Fan J. Multi-feature fusion with SVM classification for crime scene investigation image retrieval. In: *Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. 2017. p. 160–165. Available from: <https://doi.org/10.1109/SIPROCESS.2017.8124525>
- [51] Bird JJ, Faria DR, Ekart A, Ayrosa PPS. From Simulation to Reality: CNN Transfer Learning for Scene Classification. In: *Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS)*. 2020. p. 619–625. Available from: <https://doi.org/10.1109/IS48319.2020.9199968>