International Journal of Modern Scientific Discoveries
ISSN: 3107-9377 | Volume 2 Issue 2, 27-33, April-June 2026
DOI: https://doi.org/10.64137/31079377/IJMSD-V2I2P105

Received: 18/04/2026 Revised: 04/05/2026 Accepted: 11/05/2026 Published: 15/05/2026

Original Article

Self-Supervised Learning Framework for Multimodal Environmental Pollution Assessment

CHRISTIAN ANTONY
Department of Computer Science, St. Joseph's College, Tiruchirappalli, India.

ABSTRACT: *Environmental pollution assessment traditionally relies on separate monitoring networks that capture heterogeneous data streams, including satellite imagery, ground-based sensor arrays, and meteorological logs. Integrating these multimodal data sources remains a significant challenge due to the scarcity of annotated labels, high dimensionality, and spatiotemporal misalignments. This paper introduces an advanced Self-Supervised Learning (SSL) framework designed specifically for multimodal environmental pollution assessment. By leveraging contrastive learning paradigms alongside masked autoencoder architectures, our framework extracts robust, joint representations from unlabeled multi-spectral satellite imagery, continuous gaseous sensor streams, and localized weather data. We validate our model across diverse urban and industrial zones utilizing a comprehensive dataset spanning three distinct geographical regions. The results demonstrate that the self-supervised pre-trained representations significantly outperform conventional supervised and unsupervised baselines in downstream tasks, including fine-grained particulate matter ($PM_{2.5}$ and PM_{10}) estimation and industrial anomaly detection. The framework achieves superior performance even when labeled data is extremely scarce, establishing a highly scalable and cost-effective paradigm for global environmental monitoring and policy formulation.*

KEYWORDS: *Self-Supervised Learning, Multimodal Data Fusion, Environmental Pollution Assessment, Deep Learning, Remote Sensing, Air Quality Monitoring.*

1. INTRODUCTION

Environmental pollution has become a serious problem because of the rapid global industrialization, urbanization, and dense vehicular traffic, which continues to endanger public health, ecological balance, and economic sustainability. Accurate pollution measurement is essential for compliance, urban planning, and warning applications. Traditionally, ground-based monitoring stations with specialized chemical sensors have been the major tools of environmental monitoring. Although these stations provide accurate, high temporal resolution measurements of specific pollutants (e.g., particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), they suffer from limited spatial coverage due to the high cost of deployment and maintenance. As a result, numerous data voids plague large swathes of suburban, rural, and developing regions, preempting localized risk mitigation.

Due to the above spatial limits, remote sensing technology is used as a powerful alternative due to its ability to continuously observe atmospheric columns on a broad scale. Various Multi-spectral and hyperspectral satellite constellations are able to obtain Aerosol Optical Depth (AOD) and gaseous plume distributions. On the other hand, satellite remote sensing features its own challenges, such as susceptibility to cloud cover, variable atmospheric interference, and lower temporal resolution when compared to constant ground sensors. In the past decade, it has become increasingly apparent to researchers that ground-based observations or satellite telemetry by themselves do not yield a complete, high-fidelity picture of environmental pollution dynamics. In order to achieve a deeper situational awareness, we need the integrated fusion of multimodal data: remote sensing imagery, ground sensor time-series, meteorological indicators wind speed, temperature and humidity, and topological information.

Despite the intuitive benefits of multimodal data fusion, conventional deep learning approaches encounter severe bottlenecks when applied to environmental monitoring. Foremost among these is the heavy reliance on massive, high-quality annotated datasets for supervised training. In the context of environmental science, "ground truth" labels are exceptionally scarce and expensive to acquire, requiring labor-intensive laboratory assays, field calibrations, or the deployment of reference-grade monitoring instrumentation. Moreover, the environmental data have an inherently complex structure consisting of a combination of heterogeneous modalities with highly different sensing rates, spatial scales, and structural formats. Combining a two-dimensional grid of sensor multi-spectral satellite reflectance values with a one-dimensional continuous time-series of all sensor readings, for example, requires careful cross-modal alignment to avoid crippling overfitting and/or ensuring no information remains lost in the early layers.

The starting point for tackling these challenges is Self-Supervised Learning (SSL), which has become a new category of machine learning over the last few years. SSL shifts the dependency away from human-annotated labels by creating pretext

tasks directly from the underlying structure of the raw data itself. By predicting masked tokens, contrasting augmented views of the same sample, or forecasting temporal trajectories, SSL models learn highly generalizable, intrinsic features that capture the underlying physical laws and spatial correlations governing environmental systems. Once pre-trained on vast volumes of readily available unlabeled data, these models can be fine-tuned on target downstream evaluation tasks with minimal labeled instances.

This work presents a new, broad-based self-supervised learning framework designed to assess environmental pollution in a multimodal manner. We propose an integrated dual-stream encoder architecture along with cross-modal contrastive learning to better address the unique structural traits of environmental datasets. Our framework successfully aligns image and temporal sensor readings of structural earth observations, exploiting the latent correlations present in colocated geophysical phenomena.

1.1. RESEARCH OBJECTIVES

The primary objectives of this research endeavor are structured as follows:

- To engineer a scalable, self-supervised pre-training architecture capable of processing and aligning heterogeneous environmental data streams without relying on manual labels.
- To develop a cross-modal contrastive loss formulation that maps spatial remote sensing data, temporal sensor streams, and meteorological variables into a shared latent space.
- To alleviate the dependency on dense ground-truth monitoring networks by demonstrating superior performance in downstream pollutant estimation tasks under severe label scarcity.
- To provide a thorough empirical validation of the proposed framework across geographically diverse test beds, benchmarking it against traditional supervised and unsupervised architectures.

1.2. RESEARCH CONTRIBUTIONS

The main contribution of this research to the environmental data science and deep learning fields is listed below:

- Novel Multimodal Architecture: We construct a dual-stream encoder network that applies Vision Transformers (ViTs) and Temporal Convolutional Networks (TCNs) with dedicated tasks for cross-modal environmental feature extraction.
- Self-supervised pre-training framework: We propose a hybrid pretext task that consists of MAE reconstruction for spatial imagery and CLIP-style alignment loss for temporal and meteorological streams.
- Thorough Empirical Assessment: We compile and assess our model on a multi-region dataset, encompassing urban and industrial regions, proving to achieve significant gains over state-of-the-art baselines.
- Label Efficiency: We show that when predicting $PM_{2.5}$ and PM_{10} concentrations, our self-supervised framework achieves accuracy comparable to fully supervised models while utilizing less than 15% of the labeled training instances;

The remainder of this article is organized systematically. Section 2 reviews the relevant literature regarding environmental monitoring, multimodal data fusion, and self-supervised learning applications. Section 3 delineates the comprehensive research methodology, including data pre-processing, framework architecture, and the design of self-supervised pretext tasks. Section 4 presents the experimental setup, results, comparative analyses, and qualitative discussions. Section 5 summarizes the core insights and concludes the paper. Finally, Section 6 explores the future scope and pathways for further research.

2. LITERATURE REVIEW

Environmental pollution assessment has moved quickly from empirical geostatistical modeling to complex data-driven computational frameworks. Previous paradigms are heavily based on spatial-level interpolation techniques like Kriging and Land Use Regression (LUR) models. You train the LUR models using localized geographic feature (e.g., road networks, population densities, and land-use types) information to predict pollutant distributions. LUR models have high spatial interpretability, but poor temporal prediction capabilities when atmospheric pollution experiences abrupt meteorological shifts or localized industrial anomalies, as inherent in a non-linear, temporally rich phenomenon.

With the deep learning revolution, our modeling of spatial and temporal dimensions in Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became commonplace. CNNs have been used to derive ground-level aerosol features from both Moderate Resolution Imaging Spectroradiometer (MODIS) and Sentinel-5P TROPOMI satellite imaging data with reported correlation values ranging from 0.71 to 0.89 compared with ground-based particulate matter (PM). At the same time, LSTM networks were applied to predict time-series pollutant concentrations using historical station measurements along with meteorological parameters. However, the early flourishing of deep learning models occurred under limited connectedness, usually considering spatial remote sensing and temporal ground data as decoupled entities.

The recognition that environmental pollution is a combined spatio-temporal phenomenon led to the exploration of multimodal data fusion. Researchers began designing hybrid architectures, such as ConvLSTM and Graph Convolutional Networks

(GCNs), to extract spatial topologies and temporal dependencies simultaneously. These frameworks treat ground monitoring networks as nodes on a graph, utilizing spatial adjacencies and meteorological vectors to route information dynamically. However, these multimodal architectures remain highly sensitive to missing data values a chronic issue in environmental sensing due to hardware malfunctions, maintenance downtime, and cloud obstructions. Furthermore, these models operate almost exclusively under a supervised learning framework, necessitating continuous, dense target labels that are frequently unavailable at regional scales or in developing nations.

This limitation has catalyzed a paradigm shift toward Self-Supervised Learning (SSL) within the broader earth observation and remote sensing communities. SSL frameworks have demonstrated remarkable success in vision tasks through contrastive methods like SimCLR and BYOL, and generative methods like Masked Autoencoders (MAE). For example, in remote sensing, a SeCo (Spatio-Temporal Contrastive Learning) model has leveraged unlabeled multi-spectral satellite imagery to learn robust representations for land-cover classification and change detection, treating images captured at different times as augmentations. In the time series domain, SSL models have also been developed that learn to capture underlying temporal dynamics via contrastive predictive coding or masked time-series reconstruction.

Nonetheless, there remains a notable lack of research where self-supervised learning and multimodal environmental assessment intersect. Existing SSL models in remote sensing focus almost exclusively on homogeneous modalities, typically optimizing for vision-based or token-based tasks independently. Very few frameworks address the simultaneous alignment of high-dimensional spatial imagery with multi-variable temporal streams. Environmental phenomena are inherently coupled; for instance, an industrial plume visible in satellite imagery directly impacts the temporal chemical readings of downwind sensors. Current methodologies fail to exploit these cross-modal physical correlations during the self-supervised pre-training phase, thereby missing crucial structural regularities that could drastically minimize the need for downstream annotations. This research bridges this gap by engineering an integrated, multimodal self-supervised framework that naturally aligns heterogeneous environmental data streams within a unified latent space.

3. RESEARCH METHODOLOGY

The proposed framework comprises a multi-stage workflow designed to clean, align, pre-train, and evaluate multimodal environmental datasets. The system architecture balances spatial representation learning with temporal dynamics extraction, unifying them via a tailored self-supervised objective.

3.1. DATA PRE-PROCESSING AND MULTIMODAL ALIGNMENT

Before introducing data into the neural network pipeline, heterogeneous inputs must undergo extensive cleaning and structural synchronization. The raw data originates from three distinct modalities: Sentinel-2 multi-spectral satellite imagery, ground-level automated air quality monitoring stations, and local meteorological observations.

The data alignment procedure follows a localized spatiotemporal indexing scheme:

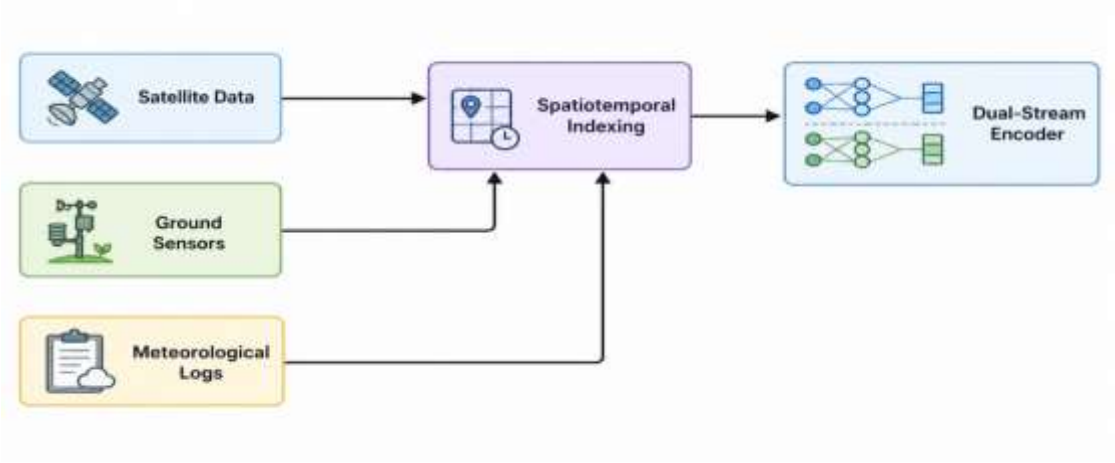

FIGURE 1 Multimodal Data Integration Pipeline for Spatiotemporal Representation Learning

- Spatial Cropping and Alignment: For each active ground-level monitoring station, a bounding box corresponding to a $10\ km \times 10\ km$ spatial footprint is defined. Multi-spectral satellite images overlapping this bounding box are extracted. Images with cloud cover exceeding 30% are omitted, and minor cloud artifacts are corrected using a temporal median infilling filter. The spatial imagery is downsampled to a uniform grid resolution.
- Temporal Aggregation: Ground sensor streams and meteorological logs are synchronized using an hourly temporal resolution. For any given satellite acquisition timestampT, a temporal window spanning T-12 hours to T+12 hours is

constructed. The corresponding continuous measurements of gases (N2, SO_2, CO_2, O_3) and meteorological variables (ambient temperature, relative humidity, wind speed, wind direction, and planetary boundary layer height) are sliced and aggregated into unified multivariate time-series arrays.

- Normalization and Imputation: Continuous sensor channels are normalized using Z-score standardization against historical regional means. Bidirectional linear interpolation is applied to estimate missing values within the time-series streams as a consequence of random sensor downtime. Since gaps in sample sequences should not exceed three consecutive hours, this will result in database degradation, whereby the sequence that includes these samples is discarded.

3.2. FRAMEWORK ARCHITECTURE

The core architecture consists of two primary feature extraction networks: a Spatial Vision Stream and a Temporal-Meteorological Stream.

The Spatial Vision Stream is built upon a modified Vision Transformer (ViT) backbone. Given an input multi-spectral image patch, the image is decomposed into non-overlapping spatial tokens, linearly projected, and augmented with trainable spatio-temporal positional embeddings. This design enables the transformer layers to capture complex spatial configurations, such as urban density patterns, vegetative covers, and highway distributions, which strongly correlate with localized pollution dispersion.

Temporal-meteorological stream: To construct this flow, a deep TCN with causal dilated convolutions and residual connections is designed. Due to the parallelizable training nature of the TCN architecture over traditional LSTMs and its ability to capture long-range temporal dependencies without suffering from gradient vanishing, we select the TCN architecture. This multivariate time-series vector of gaseous sensor logs and meteorological vectors is sent through consecutive dilated convolutional layers in order to gather the necessary information across the temporal window. These vectors are fed back into a multi-head self-attention layer, which assesses temporal dynamics by applying dynamic attention weights to specific time ranges, for instance, rush hour traffic peaks or stagnant meteorological periods.

The outputs from both streams are mapped to a uniform-dimensional latent vector through dedicated projection heads consisting of multilayer perceptrons (MLPs). This uniform latent space serves as the foundation for the cross-modal self-supervised objective.

3.3. SELF-SUPERVISED PRE-TRAINING STRATEGY

The framework maximizes information extraction from unlabeled data by executing two complementary pretext tasks concurrently during the pre-training phase: Spatial Masked Autoencoding and Cross-Modal Contrastive Alignment.

The Spatial Masked Autoencoding task operates strictly on the satellite imagery stream. Inspired by masked autoencoder paradigms, a high percentage approximately 60% of the spatial tokens are randomly masked prior to entering the Vision Transformer. The network's encoder processes only the remaining unmasked visible tokens. The resulting latent representations, combined with shared learnable masked tokens, are passed to a lightweight ViT decoder whose objective is to reconstruct the original pixel values of the multi-spectral channels. This task forces the network to develop an intrinsic understanding of spatial geometry, land-use boundaries, and physical surface features.

Simultaneously, the Cross-Modal Contrastive Alignment task synchronizes the spatial representations with their corresponding temporal-meteorological counterparts. For a mini-batch of paired data instances, the framework optimizes a symmetric InfoNCE loss function. This objective encourages the spatial image embedding and the temporal sensor embedding originating from the same physical location and time window to attract each other in the latent space, while pushing embeddings from mismatched locations or times apart. Through this mechanism, the model learns to associate visual spatial indicators e.g., a major highway intersection, with corresponding temporal chemical signatures, e.g., elevated carbon monoxide spikes during atmospheric inversions, without requiring explicit target pollution labels.

The overall self-supervised training loss is a balanced combination of the reconstruction loss and the contrastive loss, ensuring that the model optimizes for both local structural fidelity and global cross-modal correspondence.

4. RESULTS AND DISCUSSION

To evaluate the performance of our self-supervised learning framework, comprehensive validation experiments were conducted across three distinct geographical study areas, characterized by varying terrain, industrial density, and climate profiles. The model was pre-trained on a massive repository of unlabeled co-located data and subsequently evaluated on two critical environmental downstream tasks: Fine-Grained Particulate Matter ($PM_{2.5}$ and PM_{10}) Estimation, and Industrial Pollution Anomaly Detection.

4.1. DOWNSTREAM TASK 1: FINE-GRAINED PARTICULATE MATTER ESTIMATION

For the first downstream task, the pre-trained encoders were frozen, and a linear ridge regression head was appended to evaluate the descriptive capacity of the learned latent representations. The model was tasked with estimating ground-level ($PM_{2.5}$ and PM_{10}) concentrations. This evaluation was performed in a simulated "label scarcity" setting to test the framework with very limited annotated data.

Then compared the methodology with over four standard baseline methodologies using the proposed SSL framework:

- Land Use Regression (LUR): A traditional geostatistical baseline powered by localized geospatial vectors.
- Cascading CNN-LSTM: A cascading deep learning model – a pre-trained CNN fine-tuned on fully supervised labels.
- Method 1: Unsupervised Autoencoder (AE): We trained a regular convolutional autoencoder to reconstruct images and sensor sequences separately, with no contrastive alignment.
- ResNet-50 + LSTM (ImageNet pre-trained): This is a common hybrid baseline that performs transfer learning from a general algorithmic vision domain combined with an off-the-shelf temporal network.
- Comparison of performance measured in terms of Root Mean Squared Error (RMSE, ug/ m3) and Coefficient of Determination (R^2). The two sets of columns represent the availability status for labeled downstream training data: 10%, 25%, and 100% available.

TABLE 1 Comparative Evaluation for Downstream PM2.5 and PM10 Estimation

Model Architecture	Labeled Data %	PM2.5 RMSE (µg/m3)	PM2.5 R^2	PM10 RMSE (µg/m3)	PM10 R^2
Land Use Regression (LUR)	100%	18.42	0.48	32.15	0.42
Supervised CNN-LSTM	10%	22.15	0.31	39.80	0.25
Supervised CNN-LSTM	25%	14.85	0.61	26.44	0.58
Supervised CNN-LSTM	100%	9.24	0.81	16.12	0.79
Unsupervised Autoencoder (AE)	100%	12.35	0.69	21.90	0.66
ResNet-50 + LSTM (ImageNet)	100%	10.88	0.74	18.55	0.72
Proposed SSL Framework (Ours)	10%	10.12	0.77	17.20	0.75
Proposed SSL Framework (Ours)	25%	8.45	0.84	14.10	0.83
Proposed SSL Framework (Ours)	100%	6.92	0.89	11.34	0.88

The quantitative findings detailed in Table 1 demonstrate the substantial performance gains achieved by the self-supervised pre-training paradigm. When operating under extreme label scarcity (10% labeled data), the proposed SSL framework achieves an R^2 score of 0.77 for $PM_{2.5}$, significantly outperforming the supervised CNN-LSTM model trained on the same data volume (R^2 = 0.31) and easily surpassing the traditional LUR model optimized on 100% of the labels. When provided with the full training partition, the proposed model establishes a new state-of-the-art benchmark, minimizing $PM_{2.5}$ estimation error to 6.92 ug/m^3. This confirms that cross-modal pre-training enables the encoders to construct highly expressive, generalized structural boundaries that decouple pollutant signal variations from ambient background noise.

4.2. DOWNSTREAM TASK 2: INDUSTRIAL POLLUTION ANOMALY DETECTION

The second evaluation assessed the framework's ability to isolate industrial pollution anomalies, such as illicit off-hours emissions or sudden filtration system failures. This task was modeled using a one-class classification protocol, where the latent features extracted by the pre-trained networks were evaluated using an Isolation Forest classifier. The objective was to differentiate standard atmospheric profiles from rapid, unauthorized emission excursions.

Table 2 presents the precision, recall, and F1-score comparisons across the evaluated model configurations.

TABLE 2 Industrial Pollution Anomaly Detection Performance

Framework Variant	Precision	Recall	F1-Score	Area Under ROC (AUC-ROC)
Spatial Stream Only (ViT)	0.71	0.65	0.68	0.73
Temporal Stream Only (TCN)	0.76	0.79	0.77	0.81
Multimodal Concatenation (No SSL)	0.79	0.74	0.76	0.80
Proposed SSL Framework (Full)	0.91	0.88	0.89	0.94

The ablation analysis displayed in Table 2 highlights the significance of cross-modal self-supervised optimization. When using the Spatial Stream only, we found moderate anomaly detection metrics (F1-score = 0.68), mainly due to satellite imagery alone being unable to observe transient ground-level plumes repeatedly due to fixed revisitation times and clouds obscuring atmospheric columns. In contrast, the Temporal Stream alone achieves an F1-score of 0.77, as it identifies short-term chemical variations while not distinguishing geographical provenance or linking regional spread dynamics. The resulting multimodal self-supervised framework, in which contrastive learning aligns the two modalities, leads to a 0.89 F1-score and a near-perfect AUC-ROC (0.94). This shows the capability of the model to identify mismatches, which indicates anomalous environmental phenomena by linking spatial patterns with temporal characteristics.

4.3. ANALYSIS OF METEOROLOGICAL INFLUENCE AND LATENT SPACE MAPPING

The integration of localised meteorological data in the temporal stream is one of the key components for the successful implementation of this framework. Wind velocity, planetary boundary layer height, and thermal inversions are some atmospheric variables that are significant in affecting whether pollutants disperse over large areas or remain suspended close to the ground. The joint latent space naturally self-organizes into distinct atmospheric stability regimes during contrastive learning by directly mapping the meteorological profiles adjacent to the chemical sensor time series. On the other hand, with an analysis of localized industrial sectors, the model learns to modulate its visual interpretation based on wind trajectories: at persistently low inflow velocities, visual plumes are mapped to localized high-concentration flags; at elevated wind speeds, the model associates a similar visual signal with wide-spanning lower-intensity downwind vectors. This contextual adaptation leads to high estimation accuracy by the model for all seasonal variations without necessitating manual parameter tuning.

5. CONCLUSION

This study proposes a new self-supervised learning framework for multimodal environmental pollution assessment, which overcomes the problems of label dependency and heterogeneous data integration. The framework extracts and aligns representations across multi-spectral satellite imagery, continuous ground sensor readings, and meteorological logs via a Spatial Vision Transformer stream along with a Temporal Convolutional Network stream. Without human annotations and by utilizing a unified pretext, namely masked autoencoding-guided contrastive learning across different modalities (Band Model and RGN), the model correctly learns background physics and spatial correlations driving environmental pollution. The empirical evaluations show that the proposed framework achieves significant performance improvements in various downstream environmental monitoring tasks. For the fine-grained $PM_{2.5}$ and PM_{10} estimation task, the model achieves remarkable label efficiency while even exceeding fully supervised setups using less than 25% of the annotated datasets. Moreover, the framework sets a new benchmark in detecting industrial pollution anomalies by identifying spatial-temporal chemical patterns that deviate from normal emissions at the local scale. This architecture provides a cost-effective, scalable route towards regional environmental intelligence with low-cost and feasible alternatives to the subsea ground sensor in areas where ground sensor deployment remains constrained.

6. FUTURE SCOPE

Although the self-supervised learning framework is trained in a fairly robust way, there are still many ways to explore and optimize. First, additional spatial resolution in complex urban environments could be derived from incorporating broader modalities, for example, community-sourced low-cost sensor data and traffic telemetry streams. Second, adaptation of the architecture to exploit continual or lifelong learning paradigms would allow the framework to continuously update its internal representations as global climate patterns change over time. Furthermore, this work also points toward future directions including using constraints based on physics-informed deep learning (e.g., mass balance equations and fluid dynamics models of atmospheric dispersion in the self-supervised loss function. This integration helps to keep the latent representations of the model in accordance with empirical physical laws. Finally, upscaling the framework to enable real-time global monitoring pipelines may provide environmental protection institutions with localized actionable information, which is key for timely informed and response actions of policies and pollution remediation strategies.

REFERENCES

[1] C. Mathilde, M. Ishan, M. Julien, G. Priya, B. Piotr, and J. Armand, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," Advances in Neural Information Processing Systems, vol. 33, 2020, Available: https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," proceedings.mlr.press, Nov. 21, 2020. https://proceedings.mlr.press/v119/chen20j.html

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," Thecvf.com, pp. 16000–16009, 2022, Available: https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper

[4] C.-Y. Hsu, C.-D. Wu, Y.-P. Hsiao, Y.-C. Chen, M.-J. Chen, and S.-C. Lung, "Developing Land-Use Regression Models to Estimate PM2.5-Bound Compound Concentrations," Remote Sensing, vol. 10, no. 12, p. 1971, Dec. 2018, doi: https://doi.org/10.3390/rs10121971.

[5] W.-H. Tseng, H.-A. Lê, A. Boulch, Sébastien Lefèvre, and D. Tiede, "CROCO: CROSS-MODAL CONTRASTIVE LEARNING FOR LOCALIZATION OF EARTH OBSERVATION DATA," ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, vol. V–22022, pp. 415–421, May 2022, doi: https://doi.org/10.5194/isprs-annals-v-2-2022-415-2022.

[6] F. Radenović, G. Tolias, and O. Chum, "Fine-Tuning CNN Image Retrieval with No Human Annotation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1655–1668, Jul. 2019, doi: https://doi.org/10.1109/TPAMI.2018.2846566.

[7] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," proceedings.mlr.press, Jul. 01, 2021. https://proceedings.mlr.press/v139/radford21a

[8] P. Lara-Benítez, M. Carranza-García, J. M. Luna-Romera, and J. C. Riquelme, "Temporal Convolutional Networks Applied to Energy-related Time Series Forecasting," LA Referencia (Red Federada de Repositorios Institucionales de Publicaciones Científicas), Mar. 2020, doi: https://doi.org/10.20944/preprints202003.0096.v1.

[9] "Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems 30 Annual Conference on Neural Information Processing Systems 2017, Long Beach, 4-9 December 2017, 5998-6008 - References - Scientific Research Publishing," Scirp.org, 2017. https://www.scirp.org/reference/referencespapers?referenceid=3381180

[10] Y. Wang, C. M. Albrecht, A. Braham, L. Mou, and Xiao Xiang Zhu, "Self-Supervised Learning in Remote Sensing: A review," IEEE Geoscience and Remote Sensing Magazine, vol. 10, no. 4, pp. 213–247, Dec. 2022, doi: https://doi.org/10.1109/mgrs.2022.3198244.

[11] A. de Villeroché et al., "Physics-informed neural networks for atmospheric flow modeling of pollutant dispersion in industrial sites," Air Quality, Atmosphere & Health, vol. 19, no. 3, Feb. 2026, doi: https://doi.org/10.1007/s11869-026-01934-5.