Irish Interdisciplinary Journal of Science & Research (IIJSR)
Volume 8, Issue 2, Pages 171-185, April-June 2024

Exploiting Diffusion Prior for Out-of-Distribution Detection

Armando Zhu[1], Jiabei Liu[2], Keqin Li[3*], Shuying Dai[4], Bo Hong[5], Peng Zhao[6] & Changsong Wei[7]

[1]Carnegie Mellon University, USA. [2]North Eastern University, USA. [3]AMA University, Philippines. [4]Indian Institute of Technology Guwahati, India. [5]Northern Arizona University, USA. [6]Microsoft, China. [7]Digital Financial Information Technology Co. Ltd., China. Corresponding Author Email: keqin157@gmail.com[*]

DOI: https://doi.org/10.46759/IIJSR.2024.8219

Article Received: 12 April 2024 *Article Accepted: 23 June 2024* *Article Published: 30 June 2024*

ABSTRACT

Out-of-distribution (OOD) detection is crucial for deploying robust machine learning models, especially in areas where security is critical. However, traditional OOD detection methods often fail to capture complex data distributions from large scale date. In this paper, we present a novel approach for OOD detection that leverages the generative ability of diffusion models and the powerful feature extraction capabilities of CLIP. By using these features as conditional inputs to a diffusion model, we can reconstruct the images after encoding them with CLIP. The difference between the original and reconstructed images is used as a signal for OOD identification. The practicality and scalability of our method is increased by the fact that it does not require class-specific labeled ID data, as is the case with many other methods. Extensive experiments on several benchmark datasets demonstrate the robustness and effectiveness of our method, which have significantly improved the detection accuracy.

Keywords: Out-of-distribution detection; Diffusion models; Contrastive language–image pretraining (CLIP); Anomaly detection; Machine learning; Image reconstruction; Feature extraction; Zero-shot classification; Multimodal representations; Safety-critical systems.

1. Introduction

The ability to identify out-of-distribution (OOD) data is a critical component in deploying robust machine learning models in real-world applications [1–9]. OOD detection aims to identify instances that deviate significantly from the training distribution, ensuring the reliability of model predictions and minimizing the risk of erroneous outputs. This capability is particularly crucial in safety-critical domains such as autonomous driving [10–16], healthcare [17–22], and security systems, where the presence of unfamiliar data can lead to catastrophic failures.

Various approaches have been proposed to address the problem of OOD detection, ranging from statistical techniques to deep learning-based methods. Traditional methods often rely on simple feature extraction and anomaly detection algorithms, which may be inadequate for capturing complex data distributions.

Recently, the Contrastive Language–Image Pretraining (CLIP) model has emerged as a powerful backbone for feature extraction. CLIP leverages extensive internet data to learn rich, multimodal representations of images and text, demonstrating impressive zero-shot learning capabilities and effectiveness in various tasks without the need for task-specific fine-tuning.

Similarly, diffusion models have gained attention for their ability to generate high-quality images through a denoising process. These models learn the data distribution by progressively removing noise from a corrupted version of the mage, effectively capturing the underlying data manifold. The prior knowledge embedded in diffusion models can be instrumental in reconstructing images and identifying anomalies. It can be beneficial for 3D vision tasks [23–27] and scene understanding [28–33].

1.1. Study Objectives

The primary objectives of this study are as follows:

(i) To develop a novel OOD detection method that integrates the generative capabilities of diffusion models with the robust feature extraction of CLIP.

(ii) To evaluate the effectiveness of the proposed method in accurately reconstructing images and identifying OOD instances by analyzing the discrepancy between original and reconstructed images.

(iii) To assess the practicality of the proposed method in scenarios without requiring class-specific labeled in-distribution (ID) data.

(iv) To conduct extensive experiments on several benchmark datasets to validate the robustness and efficacy of the proposed method.

(v) To compare the performance of the proposed method with existing OOD detection techniques, highlighting improvements in detection accuracy and scalability.

(vi) To leverage the zero-shot classification capability of CLIP for classification tasks, enabling the use of large in-distribution datasets without the need for labeled OOD data.

In this paper, we propose a novel approach to OOD detection by exploiting the diffusion prior. The core insight of our approach is that a model capable of accurately reconstructing an image indicates that the image is likely part of the distribution the model has learned. Conversely, poor reconstruction suggests that the image is out-of-distribution. Our method involves utilizing the CLIP model to encode the image and using its features as conditional input for the diffusion model. By comparing the discrepancy between the reconstructed image and the original input, we can effectively determine if an image is OOD. This approach is based on the assumption that the model can only accurately reconstruct images of classes it has encountered during training, leveraging both the image input and its feature representation.

Additionally, for classification purposes, we utilize the zero-shot classification capability of CLIP, allowing us to classify images without fine-tuning the model. This is particularly advantageous as it enables the use of large amounts of in-distribution data without requiring labeled OOD data.

Our main contributions can be summarized as follows:

- We propose a novel OOD detection method based on the integration of CLIP and diffusion models.
- We conduct extensive experiments on multiple benchmarks, demonstrating the robustness and efficacy of our method in OOD detection.

2. Related Works

(a) Out-of-Distribution Detection

Several methods have been introduced to address the complex problem of OOD detection [34–41]. A common strategy involves leveraging uncertainty estimation techniques, such as Bayesian modeling [42], to assess prediction uncertainty and identify OOD samples. Prominent techniques in this domain include Maximum Softmax Probabilities (MSP), which uses the maximum softmax output as a confidence measure [43], Mahalanobis distance [44], and Monte Carlo Markov Chain methods that facilitate sampling from high-dimensional distributions [45].

Ensemble models are also widely acknowledged for enhancing the robustness and performance of machine learning systems, including OOD detection [46]. In OOD detection, ensemble methods integrate multiple base models for predictions, fitting into both probabilistic and uncertainty-based frameworks.

Supervised methods have shown some efficacy in reducing the incidence of erroneously high-confidence predictions on OOD inputs [47]; however, they are constrained by the necessity of labeled OOD data for training. Common unsupervised approaches include density estimation techniques [48]. Recent research indicates that augmentation and adversarial perturbation can improve OOD detection performance [49]. A key strength of our proposed OOD detection method is that it does not require specific class labels for training the diffusion model. Instead, it only requires in-distribution samples to learn the distribution, enabling it to determine whether a sample is in-distribution or OOD during testing.

(b) Pre-trained Vision-Language Models

Interpreting the semantic information within images remains a significant challenge in computer vision [50–60]. The emergence of Transformers [61] has made a great impact on not only natural language processing field [62–74], but also vision-related tasks [75], paving the way for the introduction of CLIP [76], a powerful pre-trained vision-language model. By utilizing contrastive learning along with extensive models and datasets [77], CLIP employs image-text pairs for self-supervised training. This strategy has effectively trained the model to align visual and textual representations within a latent space, facilitating robust feature extraction and zero-shot learning capabilities.

(c) Diffusion models

Diffusion denoising probabilistic models, commonly known as diffusion models [78], have gained popularity as a notable class of generative models, recognized for their exceptional synthesis quality and controllability. The fundamental principle of these models involves training a denoising autoencoder to approximate the reverse of a Markovian diffusion process [79]. By leveraging generative training on large-scale datasets with image-text pairs, such as LAION5B [77], diffusion models develop the ability to produce high-quality images featuring diverse content and coherent structures. Recently, a controllable architecture called ControlNet [80] has been introduced, enabling the addition of spatial controls, such as depth maps and human poses, to pre-trained diffusion models, thereby expanding their applicability to controlled image generation.

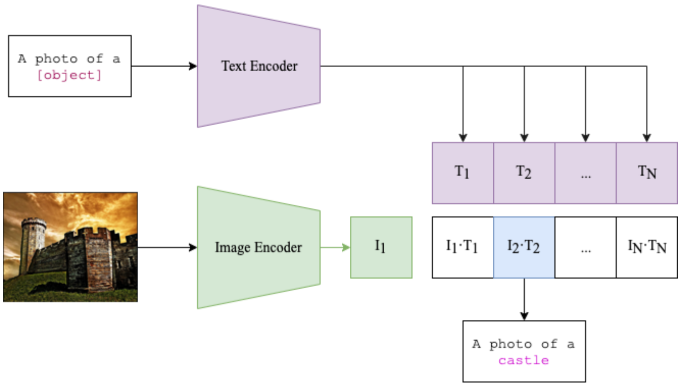

Figure 1. Architecture of CLIP model

3. Method

3.1. CLIP Model

CLIP is a multi-modal vision and language model that has demonstrated impressive results in image-text similarity and zero-shot image classification, leveraging extensive training data and large-scale models. CLIP consists of an image encoder, such as CNN-based or Transformer-like models, and a causal language model to obtain text features. During the pre-training phase, CLIP uses large-scale image-text pairs for self-supervised contrastive learning, aligning images and texts into the same latent space.

As shown in Figure 1, In zero-shot image classification tasks, given M class labels for classification (e.g., "cat", "dog"), CLIP incorporates these class labels into pre-designed hard/unlearnable text prompts, such as "a photo of a [class]", forming a prompt set like "a photo of a cat", "a photo of a dog", etc. These prompts are then fed into the text encoder to obtain M text features T_i, where $i \in \{1, 2, \dots M\}$. The testing image is input into the image encoder to obtain an image feature I_f. The cosine similarity is calculated between the normalized image feature and all text features, formally, $Sim(T_i, I_f) = T_f \cdot I_f$, and the text feature T_i with the highest similarity to I_f is considered the image's category.

3.2. Diffusion U-Net

Diffusion Models are generative models used to generate data similar to the training data. Fundamentally, Diffusion Models work by progressively adding Gaussian noise to training data and then learning to recover the data by reversing this noising process.

Diffusion models achieve high controllability through effective cross-attention layers in the denoising U-Net, facilitating interactions between image features and various conditions. ControlNet, a neural network that enhances image generation in Stable Diffusion by adding extra conditions, allows users to control the images generated more precisely. ControlNet enhances the fine-grained spatial control on latent diffusion models (LDM) by leveraging a trainable copy of the encoding layers in the denoising U-Net as a strong backbone for learning diverse conditional controls.

During the training of the ControlNet framework, images are first projected to latent representations z_0 by a trained VQGAN consisting of the encoder (EEE) and the decoder (DDD). Denoting z_s as the noisy image at the s-th timestep, it is produced by:

$$z_s = \sqrt{\bar{a}_t} z_0 + \sqrt{1 - \bar{a}_t} \qquad (1)$$

where $\bar{a}_t = \prod_{i=1}^{s} a_i$, and $\epsilon \sim N(0, I)$. By utilizing fine-grained conditions, ControlNet achieves controllable human image generation with various conditions based on the semantic information of the input.

3.3. Proposed Out-of-Distribution Detection Method

The features extracted from the CLIP model can be highly beneficial for classifying input images and distinguishing between in-distribution (ID) and out-of-distribution (OOD) samples.

We fine-tune a pre-trained denoising U-Net, guided effectively by the condition injection from features extracted by CLIP. The denoising U-Net is designed to reconstruct input images, and we use the reconstruction error to generate precision-recall curves for OOD detection. To prepare the input images, we convert the cropped image I from pixel space to obtain the latent representation from the image encoder as part of the CLIP model. We then feed the image into the U-Net with guidance extracted from CLIP. The encoder takes grayscale input images of size 128 × 128 × 1 and progressively reduces the spatial dimensions while increasing the number of channels, culminating in a bottleneck layer. The decoder then upscales and reconstructs the original input image through transposed convolutions and activations.

During training, the model is optimized to minimize the Mean Squared Error (MSE) loss between the reconstructed heatmaps and the original input. During inference, the threshold for distinguishing between in-distribution and out-of- distribution samples is set as the maximum reconstruction error of the in-distribution samples. With this approach, any sample with a reconstruction error above the threshold is classified as OOD, while samples below the threshold are considered in-distribution.

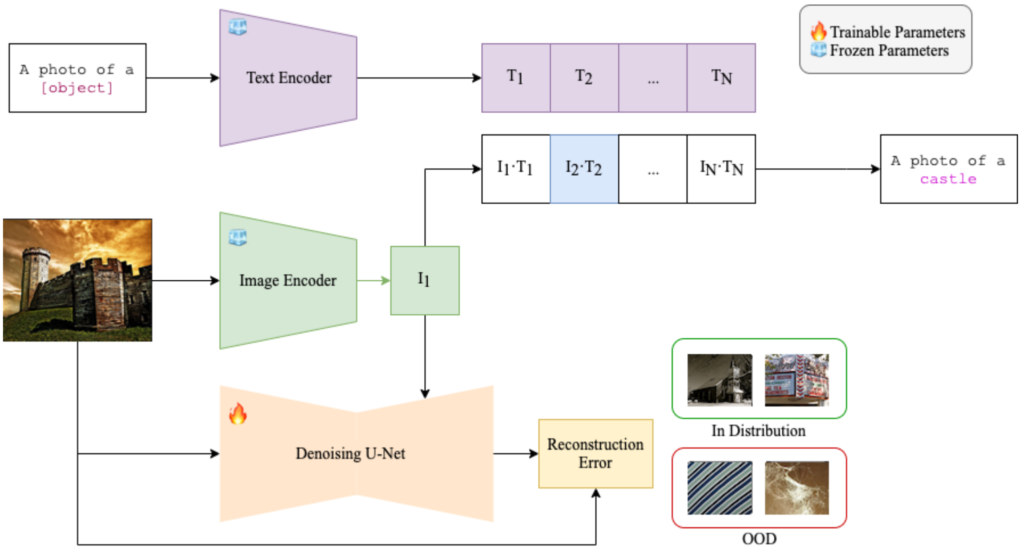

Figure 2. Architecture of the proposed method

4. Experiments

4.1. Experimental Details

4.1.1. Datasets

To evaluate the efficacy of our proposed OOD detection method, we conducted extensive experiments using established benchmarks. We utilized the ImageNet-1K [81] dataset with 1,000 classes as the in-distribution (ID) dataset. For out-of-distribution (OOD) datasets, we selected subsets from Texture [82], iNaturalist [83], Places [84], and SUN [85], ensuring that the concepts in these datasets do not overlap with ImageNet-1K. Specifically, the entire Texture dataset was used for evaluation. Additionally, 110 plant classes not present in ImageNet-1K were selected from iNaturalist, 50 categories not present in ImageNet-1K were selected from Places, and 50 unique nature-related concepts were selected from SUN.

4.1.2. Implementation Details

We employed the CLIP model based on CLIP-B/16, pre-trained from OpenCLIP [86]. For the denoising U-Net, we used Stable Diffusion V1-5 with ControlNet, pre-trained for image generation. The model was fine-tuned using ImageNet-1K samples for 10 epochs. During fine-tuning, the Mean Squared Error (MSE) loss was minimized between the reconstructed images and the original inputs to enhance the model's reconstruction capabilities.

4.2. Comparison with Existing Models

The results of OOD detection on the benchmark datasets are summarized in Table 1. Our proposed method consistently achieves superior or comparable performance across individual OOD datasets and in the averaged results. Compared with zero-shot methods, our approach surpasses the best competing method, CLIPN [87], by approximately 1.5% in FPR95, despite CLIPN requiring an additional large external dataset to train an additional negative text encoder. Although our method is significantly more lightweight than CLIPN in model size, it consistently outperforms CLIPN in both metrics across all OOD datasets. Adapted post-hoc methods generally do not leverage CLIP's capabilities well and thus perform less effectively.

Table 1. Performance metrics across various datasets

Dataset	**iNaturalist**		**SUN**		**Places**		**Texture**		**Average**	
Metrics	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Zero-shot methods										
MCM [88]	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.1	42.82	90.76
GL-MCM [89]	15.18	**96.71**	30.42	93.09	38.85	89.9	57.93	83.63	35.47	90.83
CLIPN [87]	23.94	95.27	26.17	93.92	33.45	**92.28**	**40.83**	90.93	31.1	93.1
CLIP-based posthoc methods										
MSP [43]	74.57	77.74	76.95	73.97	79.12	72.18	73.66	74.84	76.22	74.98
MaxLogit [90]	60.88	88.03	44.83	91.16	55.54	87.45	48.72	88.63	52.49	88.82
ODIN [91]	30.22	94.65	54.04	87.17	55.06	85.54	51.67	87.85	47.75	88.8
ViM [92]	32.19	93.16	54.01	87.19	60.67	83.75	53.94	87.18	50.2	87.82
KNN [93]	29.17	94.52	35.62	92.67	39.61	91.02	64.35	85.67	42.19	90.97
Prompt Learning Methods										
CoOp [94]	29.81	93.77	40.83	93.29	40.11	90.58	45	89.47	51.68	91.78
Proposed Method										
Ours	**15.03**	96.45	**24.95**	**94.59**	**33.17**	90.83	41.85	**91.02**	**28.75**	**93.25**

Our method also substantially surpasses prompt learning-based methods, reducing the FPR95 by about 23%. This indicates that the learned diffusion model provides informed knowledge about OOD data, which is lacking in the competing methods, significantly reducing detection errors.

5. Conclusion

In this paper, we introduced a novel approach to out-of-distribution (OOD) detection by combining the feature extraction capabilities of CLIP with the generative power of diffusion models. Our method involves encoding images with CLIP and using these features as conditional inputs for a diffusion model to reconstruct the images. The discrepancy between the original and reconstructed images serves as a robust indicator for OOD detection.

Our approach offers several advantages over existing methods. Firstly, it does not require labeled OOD data, making it more practical and scalable for real-world applications. By leveraging only in-distribution samples for training, our method effectively discerns between in-distribution and OOD samples during testing. Secondly, the integration of CLIP’s zero-shot classification capability enhances the versatility of our method, allowing for effective image classification without the need for model fine-tuning.

We conducted extensive experiments on multiple benchmark datasets, including ImageNet-1K, Texture, iNaturalist, Places, and SUN. The results demonstrate that our method achieves significant improvements in detection accuracy, with substantial reductions in false positive rates and enhanced detection metrics across diverse datasets. These findings underscore the potential of integrating pre-trained models to enhance the reliability of OOD detection, paving the way for the deployment of more dependable machine learning systems.

Future work could explore further enhancements to our method, such as:

(i) Incorporating additional types of pre-trained models to further enhance detection accuracy.

(ii) Refining the reconstruction process to improve the robustness and accuracy of OOD detection.

(iii) Applying the approach to other domains beyond image data, such as natural language processing or audio data, to broaden its applicability.

(iv) Investigating the impact of different types of noise in the diffusion process to achieve more robust OOD detection.

(v) Integrating the method with real-time systems to evaluate performance in dynamic and unpredictable environments.

(vi) Extending the framework to handle multi-modal data inputs simultaneously, enhancing its capability to detect OOD instances across various data types.

Declarations

Source of Funding

This study did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors declare no competing financial, professional, or personal interests.

Consent for publication

The authors declare that they consented to the publication of this study.

Authors' contributions

All the authors took part in literature review, analysis and manuscript writing equally.

Availability of data and material

All data pertaining to the research is kept in good custody by the authors.

References

[1] Yafeng, Y., Shuyao, H., Zhou, Y., Jiajie, Y., Ziang, L., & Yan, C. (2024). Investigation of customized medical decision algorithms utilizing graph neural networks. ArXiv preprint arXiv: 2405.17460.

[2] Qikai, Y., Panfeng, L., Zhicheng, D., Wenjing, Z., Yi, N., & Xinhe, X. (2024). A comparative study on enhancing prediction in social network advertisement through data augmentation. ArXiv preprint arXiv: 2404. 13812.

[3] Zhenglin, L., Hanyi, Y., Jinxin, X., Jihang, L., & Yuhong, M. (2023). Stock market analysis and prediction using LSTM: A case study on technology stocks. Innovations in Applied Engineering and Technology, Pages 1–6. https://doi.org/10.62836/iaet.v2i1.162.

[4] Jingyu, R., Han, Y., Hao, L., Jiayuan, L., Xiangyue, Z., & Hongli, X. (2023). A bounded near-bottom cruise trajectory planning algorithm for underwater vehicles. Journal of Marine Science and Engineering, 11(1). doi: 10. 3390/jmse11010007.

[5] Xiaosong, W., Yuxin, Q., Jize, X., Zhiming, Z., Ning, Z., Mingyang, F., & Chufeng, J. (2024). Advanced network intrusion detection with tabtransformer. Journal of Theory and Practice of Engineering Science, 4(03): 191–198. https://doi.org/10.53469/jtpes.2024.04(03).18.

[6] Tianrui, L., Qi, C., Changxin, X., Zhanxin, Z., Fanghao, N., Yuxin, Q., & Tsungwei, Y. (2024). Rumor detection with a novel graph neural network approach. ArXiv preprint arXiv: 2403.16206. https://doi.org/10.540 97/farmdr42.

[7] Ao, X., Jingyu, Z., Qin, Y., Liyang, W., & Yu, C. (2024). Research on splicing image detection algorithms based on natural image statistical characteristics. ArXiv preprint arXiv: 2404.16296.

[8] Yu, C., Qin, Y., Liyang, W., Ao, X., & Jingyu, Z. (2024). Research on credit risk early warning model of commercial banks based on neural network algorithm. ArXiv preprint arXiv:2405.10762.

[9] Jingyu, Z., Ao, X., Yu, C., Qin, Y., & Liyang, W. (2024). Research on detection of floating objects in river and lake based on AI intelligent image recognition. ArXiv preprint arXiv: 2404.06883.

[10] Zhicheng, D., Zhixin, L., Siyang, L., Panfeng, L., Qikai, Y., & Edward, W. (2024). Confidence trigger detection: Accelerating real-time tracking-by-detection systems. ArXiv preprint arXiv: 1902.00615.

[11] Chang, Z., Yang, Z., Shaobo, L., Yi, Z., Xingchen, L., & Chiyu, C. (2024). Research on driver facial fatigue detection based on YOLOv8 model. https://doi.org/10.36227/techrxiv.171822194.49730312/v1.

[12] Yucheng, Z., & Guodong, L. (2023). Multimodal event transformer for image-guided story ending generation. ArXiv preprint arXiv: 2301.11357.

[13] Yucheng, Z., & Guodong, L. (2023). Style-aware contrastive learning for multi-style image captioning. ArXiv preprint arXiv: 2301.11367.

[14] Shuyao, H., Yue, Z., Yushan, D., Hao, Q., & Yuhong, M. (2024). Lidar and monocular sensor fusion depth estimation. Applied Science and Engineering Journal for Advanced Research, 3(3): 20–26. https://doi.org/10.5281/zenodo.11347309.

[15] Yuhong, M., Chaoyi, T., Chenghao, W., Hao, Q., & Yushan, D. (2024). Make scale invariant feature transform "fly" with CUDA. International Journal of Engineering and Management Research, 14(3): 38–45. https://doi.org/10.5281/zenodo.11516606.

[16] Shuying, D., Jiajing, D., Yuqiang, Z., Taiyu, Z., & Yuhong, M. (2024). The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. Journal of Computer Technology and Applied Mathematics, 1(1): 6–12. https://doi.org/10.5281/zenodo.10866092.

[17] Jiajie, Y., Linxiao, W., Yulu, G., Zhou, Y., Ziang, L., & Shuyao, H. (2024). Research on intelligent aided diagnosis system of medical image based on computer deep learning. ArXiv preprint arXiv: 2404.18419.

[18] Sheng, C., Xinghui, F., Yulu, W., Lu, D., & Mingxiu, S. (2024). Deep learning-based lung medical image recognition. International Journal of Innovative Research in Computer Science & Technology, 12(3): 100–105. https://doi.org/10.55524/ijircst.2024.12.3.16.

[19] Ziyan, Y., Fei, L., Sheng, C., Weijie, H., Lu, D., & Xinghui, F. (2024). Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis. ArXiv preprint arXiv: 2405.17459.

[20] Jason, S., Ge, S., & Ian, D. (2022). Deep learning in neuroimaging: Overcoming challenges with emerging approaches. Frontiers in Psychiatry, 13: 912600. https://doi.org/10.3389/fpsyt.2022.912600.

[21] Shaojie, L., Yuhong, M., & Zhenglin, L. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. Innovations in Applied Engineering and Technology, Pages 1–6. https://doi.org/10.62836/iaet.vli1.002.

[22] Yuhong, M., Shaojie, L., Yushan, D., Ziyi, Z., & Zhenglin, L. (2024). Password complexity prediction based on RoBERTa algorithm. Applied Science and Engineering Journal for Advanced Research, 3(3): 1–5. https://doi.org/10.5281/zenodo.11180356.

[23] Zhimin, C., Longlong, J., Liang, Y., Yingwei, L., & Bing, L. (2023). Class-level confidence based 3D semi-supervised learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Pages 633–642. https://doi.org/10.1109/wacv56688.2023.00070.

[24] Zhimin, C., Longlong, J., Yang, L., Ying Li, T., & Bing, L. (2021). Multimodal semi-supervised learning for 3D objects. ArXiv preprint arXiv: 2110.11601.

[25] Zhimin, C., Longlong, J., Yingwei, L., & Bing, L. (2024). Bridging the domain gap: Self-supervised 3D scene understanding with foundation models. Advances in Neural Information Processing Systems, 36. https://dl.acm.org/doi/10.5555/3666122.3669589.

[26] Zhimin, C., Yingwei, L., Longlong, J., Liang, Y., & Bing, L. (2023). Point cloud self-supervised learning via 3D to multi-view masked autoencoder. ArXiv preprint arXiv: 2311.10887.

[27] Yijie, W., & Jianhao, W. (2024). Fortifying the global data fortress: a multidimensional examination of cyber security indexes and data protection measures across 193 nations. International Journal of Frontiers in Engineering Technology, 6(2). doi: 10.25236/ijfet.2024.060206.

[28] Yi, X., Junlong, D., Qiang, W., Zhiwen, L., & Ke, Y. (2024). VMT-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, Volume 38, Pages 16085–16093. https://doi.org/10.1609/aaai.v38i14.29541.

[29] Tianchen, D., Yaohui, C., Leyan, Z., Jianfei, Y., Shenghai, Y., Danwei, W., & Weidong, C. (2024). Compact 3D Gaussian splatting for dense visual SLAM. ArXiv preprint arXiv: 2403.11247.

[30] Tianchen, D., Guole, S., Tong, Q., Jianyu, W., Wentao, Z., Jingchuan, W., Danwei, W., & Weidong, C. (2023). PLGSLAM: Progressive neural scene representation with local to global bundle adjustment. ArXiv preprint arXiv: 2312.09866.

[31] Yi, S., Hao, L., Xinxin, L., Wenjing, Z., Chang, Z., & Yizhou, C. (2024). Localization through particle filter powered neural network estimated monocular camera poses. ArXiv preprint arXiv: 2404.17685.

[32] Hao, L., Yi, S., Wenjing, Z., Yuelin, Z., Chang, Z., & Shuyao, H. (2024). Adaptive speed planning for unmanned vehicle based on deep reinforcement learning. ArXiv preprint arXiv: 2404.17379.

[33] Tianchen, D., Hongle, X., Jingchuan, W., & Weidong, C. (2023). Long-term visual simultaneous localization and mapping: Using a Bayesian persistence filter-based global map prediction. IEEE Robotics & Automation Magazine, 30(1): 36–49. https://doi.org/10.1109/mra.2022.3228492.

[34] Jingkang, Y., Kaiyang, Z., Yixuan, L., & Ziwei, L. (2021). Generalized out-of-distribution detection: A survey. ArXiv preprint arXiv: 2110.11334.

[35] Chang, Z., Yang, Z., Jin, C., Yi, S., Xiaoling, C., & Chiyu, C. (2024). Optimizing search advertising strategies: Integrating reinforcement learning with generalized second-price auctions for enhanced ad ranking and bidding. https://ui.adsabs.harvard.edu/link_gateway/2024arXiv240513381Z/doi:10.48550/arXiv.2405.13381.

[36] Chang, Z., Yang, Z., Yuelin, Z., Jin, C., Wenhan, F., Yi, Z., & Cheng, C. (2024). Predict click-through rates with deep interest network model in e-commerce advertising. https://doi.org/10.36227/techrxiv.171822205.50007015/v1.

[37] Ge, S., Jason, S., & Ian, D. (2022). Deep learning for prognosis using task-fMRI: A novel architecture and training scheme. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Pages 1589–1597. https://dl.acm.org/doi/10.1145/3534678.3539362.

[38] Wendong, Y., Yun, J., Yulin, C., Zhengjia, X., & Wenbin, W. (2024). Long-term network structure evolution investigation for sustainability improvement: An empirical analysis on global top full-service carriers. Aerospace, 11(2): 128. https://doi.org/10.3390/aerospace11020128.

[39] Zhengjia, X., Ivan, P., Antonios, T., Raphael, G., Pekka, P., & Smita, T. (2023). Combination and selection of machine learning algorithms in GNSS architecture design for concurrent executions with HIL testing. In 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), Pages 1–9, IEEE. https://doi.org/10.1109/dasc58513.2023.10311160.

[40] Yuelyu, J., Yuhe, G., Runxue, B., Qi, L., Disheng, L., Yiming, S., & Ye, Y. (2023). Prediction of COVID-19 patients' emergency room revisit using multi-source transfer learning. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), Pages 138–144, IEEE. https://doi.org/10.1109/ichi57859.2023.00028.

[41] Yuelyu, J., Zhuochun, L., Rui, M., Sonish, S., Yanshan, W., Zeshui, Y., Hui, J., Yushui, H., Hanyu, Z., & Daqing, H. (2024). Rag-rlrc-laysum at BioLaySumm: Integrating retrieval-augmented generation and readability control for layman summarization of biomedical texts. ArXiv preprint arXiv: 2405.13179.

[42] David, J.C.M. (1992). A practical Bayesian framework for backpropagation networks. Neural computation, 4(3): 448–472. https://doi.org/10.1162/neco.1992.4.3.448.

[43] Dan, H., & Kevin, G. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. ArXiv preprint arXiv: 1610.02136.

[44] Kimin, L., Kibok, L., Honglak, L., & Jinwoo, S. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems, 31.

[45] Christophe, A., Nando, D.F., Arnaud, D., & Michael, I.J. (2003). An introduction to MCMC for machine learning. Machine Learning, 50: 5–43. https://doi.org/10.1007/978-0-387-30164-8_522.

[46] Balaji, L., Alexander, P., & Charles, B. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30.

[47] Terrance, D., & Graham, W.T. (2018). Learning confidence for out-of-distribution detection in neural networks. ArXiv preprint arXiv: 1802.04865.

[48] Durk, P.K., & Prafulla, D. (2018). Glow: Generative flow with invertible 1x1 convolutions. Advances in Neural Information Processing Systems, 31.

[49] Sungik, C., & Sae-Young, C. (2019). Novelty detection via blurring. ArXiv preprint arXiv: 1911.11943.

[50] Yi, X., Siqi, L., Haodi, Z., Junlong, D., Xiaohong, L., Yue, F., Qing, L., & Yuntao, D. (2024). Parameter-efficient fine-tuning for pre-trained vision models: A survey. ArXiv preprint arXiv: 2402.02242.

[51] Yi, X., Junlong, D., Qiang, W., Ke, Y., & Shouhong, D. (2024). Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Volume 38, Pages 16076–16084. https://doi.org/10.1609/aaai.v38i14.29540.

[52] Yi, X., Siqi, L., Pengsheng, J., Yuntao, D., & Chongjun, W. (2023). Self-training with label-feature-consistency for domain adaptation. In International Conference on Database Systems for Advanced Applications, Pages 84–99, Springer. https://doi.org/10.1007/978-3-031-30678-5_7.

[53] Tianchen, D., Nailin, W., Chongdi, W., Shenghai, Y., Jingchuan, W., Danwei, W., & Weidong, C. (2024). Incremental joint learning of depth, pose and implicit scene representation on monocular camera in large-scale scenes. ArXiv preprint arXiv: 2404.06050.

[54] Huajun, Z., Su, D., Yining, Y., Jiachen, Z., & Yafeng, Y. (2024). Multi-scale image recognition strategy based on convolutional neural network. Journal of Computing and Electronic Information Management, 12(3): 107–113. https://doi.org/10.54097/ro4puyx5.

[55] Zijun, G., Qi, W., Taiyuan, M., Xiaohan, C., et al. (2024). An enhanced encoder-decoder network architecture for reducing information loss in image semantic segmentation. ArXiv preprint arXiv: 2406.01605.

[56] Qishi, Z., Yuhan, M., Erdi, G., Dan, S., & Haowei, Y. (2024). Innovations in time related expression recognition using LSTM networks. International Journal of Innovative Research in Computer Science & Technology, 12(3): 120–125. https://doi.org/10.55524/ijircst.2024.12.3.19.

[57] Yutian, Y., Zexi, C., Yafeng, Y., Muqing, L., & Tana, G. (2024). A new method of image recognition based on deep learning generates adversarial networks and integrates traditional algorithms. Journal of Computing and Electronic Information Management, 13(1): 57–61. https://doi.org/10.54097/qnq8g0tj.

[58] Panfeng, L., Youzuo, L., & Emily, S.F. (2019). Contextual hourglass network for semantic segmentation of high resolution aerial imagery. ArXiv preprint arXiv: 1810.12813.

[59] Zhicheng, D., Panfeng, L., Qikai, Y., Siyang, L., & Qingtian, G. (2024). Regional style and color transfer. ArXiv preprint arXiv: 2404.13880.

[60] Yuelyu, J., Yuheng, S., Wei, W., Ruoyi, X., Zhongqian, X., & Huiyun, L. (2023). Improving emotional expression and cohesion in image-based playlist description and music topics: A continuous parameterization approach. ArXiv preprint arXiv: 2310.01248.

[61] Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N.G., Łukasz, K., & Illia, P. (2017). Attention is all you need. Advances in neural Information Processing Systems, 30.

[62] Yuelyu, J., Zeshui, Y., & Yanshan, W. (2024). Assertion detection large language model in-context learning Lora fine-tuning. ArXiv preprint arXiv: 2401.17602.

[63] Taiyuan, M., Yun, Z., Xiaohan, C., Zijun, G., Qi, W., & Haowei, Y. (2024). Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks. ArXiv preprint arXiv: 2405.11704.

[64] Lingxi, X., Muqing, L., Yinqiu, F., Meiqi, W., Ziyi, Z., & Zexi, C. (2024). Exploration of attention mechanism-enhanced deep learning models in the mining of medical textual data. ArXiv preprint arXiv: 2406. 00016.

[65] Ke, X., Yu, C., Shiqing, L., Junjie, G., Jue, X., & Mengfang, S. (2024). Advancing financial risk prediction through optimized LSTM model performance and comparative analysis. ArXiv preprint arXiv: 2405.20603.

[66] Panfeng, L., Qikai, Y., Xieming, G., Wenjing, Z., Zhicheng, D., & Yi, N. (2024). Exploring diverse methods in visual question answering. ArXiv preprint arXiv: 2404.13565.

[67] Panfeng, L., Mohamed, A., & Rada, M. (2023). Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks. ArXiv preprint arXiv: 2311.10944.

[68] Zhenglin, L., Yangchen, H., Mengran, Z., Jingyu, Z., Jinghao, C., & Houze, L. (2024). Feature manipulation for DDPM based change detection. ArXiv preprint arXiv: 2403.15943.

[69] Yucheng, Z., Tao, S., Xiubo, G., Chongyang, T., Can, X., Guodong, L., Binxing, J., & Daxin, J. (2022). Towards robust ranker for text retrieval. ArXiv preprint arXiv: 2206.08063.

[70] Yucheng, Z., Xiubo, G., Tao, S., Chongyang, T., Guodong, L., Jian-Guang, L., & Jianbing, S. (2023). Thread of thought unraveling chaotic contexts. ArXiv preprint arXiv: 2311.08734.

[71] Yucheng, Z., Tao, S., Xiubo, G., Chongyang, T., Jianbing, S., Guodong, L., Can, X., & Daxin, J. (2024). Fine-grained distillation for long document retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, Volume 38, Pages 19732–19740. https://doi.org/10.1609/aaai.v38i17.29947.

[72] Yuhong, M., Hao, Q., Yushan, D., Ziyi, Z., & Zhenglin, L. (2024). Large language model (LLM) AI text generation detection based on transformer deep learning algorithm. ArXiv preprint arXiv: 2405.06652.

[73] Yucheng Z., & Guodong, L. (2023). Improving cross-modal alignment for text-guided image in painting. ArXiv preprint arXiv: 2301.11362.

[74] Tianrui, L., Shaojie, L., Yushan, D., Yuhong, M., & Shuyao, H. (2024). Spam detection and classification based on DistilBERT deep learning algorithm. Applied Science and Engineering Journal for Advanced Research, 3(3): 6–10. https://doi.org/10.5281/zenodo.11180574.

[75] Alexey, D., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., Mostafa, D., Matthias, M., George, H., & Sylvain, G. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv preprint arXiv: 2010.11929.

[76] Alec, R., Jong, W.K., Chris, H., Aditya, R., Gabriel, G., Sandhini, A., Girish, S., Amanda, A., Pamela, M., Jack, C., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, Pages 8748–8763, PMLR.

[77] Christoph, S., Romain, B., Richard, V., Cade, G., Ross, W., Mehdi, C., Theo, C., Aarush, K., Clayton, M., Mitchell, W., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35: 25278–25294.

[78] Jonathan, H., Ajay, J., & Pieter, A. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840–6851.

[79] Jascha, S.D., Eric, W., Niru, M., & Surya, G. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, Pages 2256–2265, PMLR.

[80] Lvmin, Z., Anyi, R., & Maneesh, A. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Pages 3836–3847. https://doi.org/10.1109/iccv51070.2023.00355.

[81] Jia, D., Wei, D., Richard, S., Li-Jia, L., Kai, L., & Li, F.F. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Pages 248–255, IEEE. https://doi.org/10.1109/cvpr.2009.5206848.

[82] Mircea, C., Subhransu, M., Iasonas, K., Sammy, M., & Andrea, V. (2014). Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Pages 3606–3613. https://doi.org/10.1109/cvpr.2014.461.

[83] Grant, V.H., Oisin, M.A., Yang, S., Yin, C., Chen, S., Alex, S., Hartwig, A., Pietro, P., & Serge, B. (2018). The iNaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Pages 8769–8778. https://doi.org/10.1109/cvpr.2018.00914.

[84] Bolei, Z., Agata, L., Aditya, K., Aude, O., & Antonio, T. (2017). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1452–1464. https://doi.org/10.1109/cvpr.2018.00914.

[85] Jianxiong, X., James, H., Krista, A.E., Aude, O., & Antonio, T. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Pages 3485–3492, IEEE. https://doi.org/10.1109/cvpr.2010.5539970.

[86] Mehdi, C., Romain, B., Ross, W., Mitchell, W., Gabriel, I., Cade, G., Christoph, S., Ludwig, S., & Jenia, J. (2023). Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Pages 2818–2829. https://doi.org/10.1109/cvpr52729.2023.00276.

[87] Hualiang, W., Yi, L., Huifeng, Y., & Xiaomeng, L. (2023). CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Pages 1802–1812. https://doi.org/10.1109/iccv51070.2023.00173.

[88] Yifei, M., Ziyang, C., Jiuxiang, G., Yiyou, S., Wei, L., & Yixuan, L. (2022). Delving into out-of-distribution detection with vision-language representations. Advances in Neural Information Processing Systems, 35: 35087–35102.

[89] Atsuyuki, M., Qing, Y., Go, I., & Kiyoharu, A. (2023). Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. ArXiv preprint arXiv: 2304.04521.

[90] Dan, H., Steven, B., Mantas, M., Andy, Z., Joe, K., Mohammadreza, M., Jacob, S., & Dawn, S. (2019). Scaling out-of-distribution detection for real-world settings. ArXiv preprint arXiv: 1911.11132.

[91] Shiyu, L., Yixuan, L., & Rayadurgam, S. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. ArXiv preprint arXiv: 1706.02690.

[92] Haoqi, W., Zhizhong, L., Litong, F., & Wayne, Z. (2022). VIM: Out-of-distribution with virtual-logit matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Pages 4921–4930. https://doi.org/10.1109/cvpr52688.2022.00487.

[93] Yiyou, S., Yifei, M., Xiaojin, Z., & Yixuan, L. (2022). Out-of-distribution detection with deep nearest neighbors. In International Conference on Machine Learning, Pages 20827–20840, PMLR.

[94] Kaiyang, Z., Jingkang, Y., Chen, C.L., & Ziwei, L. (2022). Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Pages 16816–16825. https://doi.org/10.1109/cvpr52688.2022.01631.