Vision-Language Hash Learning for Remote Sensing Scene Retrieval Based on Asymmetric Semantic Representation Mining

Ganav Hishra
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
ganavwork@ku.edu

Kasper Kennedy
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
kennedykasper@uab.edu

Bennett A. Carpenter
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
bacarpenter@missouri.edu

Kasper Burton
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
burtonkasper@oregonstate.edu

Abstract

The exponential growth of remote sensing imagery archives demands scalable, semantically precise retrieval mechanisms capable of bridging the gap between high-dimensional visual data and human-expressed queries. Vision-language hash learning has emerged as a compelling paradigm, encoding cross-modal semantic correspondences into compact binary codes that support fast approximate nearest neighbor search in large-scale repositories. This paper presents a systems-oriented examination of vision-language hash learning for remote sensing scene retrieval, founded upon asymmetric semantic representation mining. Conventional symmetric alignment strategies often fail to account for the inherent information density imbalance between satellite imagery and textual descriptions, where visual scenes contain rich spectral and spatial detail that is only partially captured in short query statements. We argue that intentionally asymmetric modalities of representation, in which the visual and language encoders learn complementary rather than strictly matched embeddings, unlock superior retrieval fidelity when combined with sophisticated hash coding. The paper foregrounds architectural trade-offs, infrastructure requirements, deployment models, and governance frameworks that shape the real-world viability of such systems. We discuss how decisions regarding model complexity, hash code length, training data composition, and inference distribution carry profound implications for sustainability, fairness, and robustness. Cross-domain comparisons with multimedia and medical image retrieval highlight unique challenges in the remote sensing domain, including geospatial bias, temporal variability, and the coexistence of heterogeneous sensor modalities. Policy considerations

around data sovereignty, dual-use governance, and the environmental footprint of large-scale multi-modal training are integrated into a holistic assessment. The paper concludes by identifying open research frontiers at the intersection of asymmetric learning, hash-based indexing, and socio-technical infrastructure design.

Keywords

remote sensing scene retrieval, vision-language hashing, asymmetric semantic mining, deep learning, cross-modal retrieval, large-scale systems, robustness, fairness, governance.

1. Introduction

The proliferation of Earth observation satellites and airborne sensors has generated an unprecedented volume of remotely sensed imagery, transforming fields as diverse as precision agriculture, disaster response, urban planning, and climate monitoring. Efficiently searching such repositories for scenes that match a conceptual query, such as "coastal wetlands with sparse urban encroachment," demands retrieval systems that transcend traditional metadata-based filtering. Cross-modal retrieval, particularly the alignment of visual data with natural language, has attracted substantial attention following breakthroughs in deep convolutional [1, 2] and transformer-based architectures [3]. The vision-language pre-training paradigm, epitomized by models that learn joint embedding spaces from paired image-text corpora [4, 5], provides a foundation for bridging the semantic gulf between pixel arrays and human-interpretable concepts. However, deploying such models directly for remote sensing scene retrieval introduces distinctive scalability and semantic fidelity challenges. High-dimensional dense embeddings incur prohibitive storage and search costs when datasets exceed petabyte scales, motivating the adoption of hash learning techniques that compress representations into compact binary codes amenable to Hamming distance ranking.

Hashing for large-scale retrieval has a rich history, spanning from data-independent locality-sensitive hashing to learning-based methods that optimize binary codes for specific similarity metrics [6, 7]. Deep hashing, wherein a neural network jointly learns feature representations and their quantization into binary codes, has demonstrated state-of-the-art performance across multiple retrieval benchmarks [8]. When extended to the cross-modal setting, deep hashing must reconcile two fundamentally different data distributions, visual and textual, within a shared binary space. Early efforts often employed symmetric alignment losses that enforce strict semantic parity between the two modalities. Yet remote sensing scenes present an informational asymmetry that resists such symmetry. A panchromatic or multispectral image capturing complex land cover and object interactions encodes a breadth of detail that a concise textual description cannot fully articulate; conversely, language queries may express abstract or functional relationships (e.g., "vulnerable informal settlements") that lack explicit visual signatures. Asymmetric representation mining, wherein the visual and language encoders are deliberately configured to learn complementary rather than identical semantics, offers a principled strategy to exploit this natural divergence while sharpening retrieval discrimination.

This paper provides an interdisciplinary, systems-level analysis of vision-language hash learning for remote sensing scene retrieval, with asymmetric semantic representation mining as its conceptual backbone. Unlike purely algorithmic contributions, we emphasize the structural trade-offs, infrastructure demands, and socio-technical implications that determine whether such systems transition from laboratory benchmarks to operational deployments. The discussion encompasses encoder architectures, hash coding schemes, training data governance,

distributed inference topologies, fairness across geographic and socioeconomic strata, robustness to distributional shifts, and the environmental impacts of large-scale model training. Throughout, we draw upon cross-domain insights from content-based image retrieval in the multimedia domain, medical image analysis, and industrial search systems to situate the unique challenges of remote sensing. The remainder of the paper is organized as follows. Section 2 surveys related foundational work in deep hashing, cross-modal retrieval, and remote sensing scene understanding. Section 3 develops the conceptual framework of asymmetric semantic representation mining. Section 4 presents a system architecture analysis and discusses deployment models. Section 5 addresses governance, fairness, and policy dimensions. Section 6 examines robustness and sustainability. Section 7 concludes with forward-looking perspectives.

2. Related Foundational Work

The intellectual lineage of the proposed system draws from multiple intersecting research streams. Foundational studies in supervised and unsupervised hashing established the viability of binary codes for large-scale nearest neighbor search, with early spectral and iterative quantization methods [6] proving that compact hashes could preserve semantic neighborhoods. Supervised hashing with kernels extended this principle by incorporating label information to enhance discrimination [7], laying the groundwork for deep hashing architectures that unify feature extraction and hash coding within end-to-end trainable networks [8]. In the remote sensing domain, deep learning has revolutionized scene classification and retrieval [9], with specialized deep hashing approaches introduced to handle the high intra-class variance and geographic diversity of overhead imagery [10]. These works primarily operate in a single-modal setting, either visual-to-visual or text-to-text, and do not address the cross-modal alignment necessary for vision-language retrieval.

Cross-modal retrieval has been extensively investigated in the multimedia community. Early probabilistic models for automatic image annotation learned correlations between visual segments and keyword vocabularies [11]. The advent of deep learning introduced correspondence autoencoders and adversarial training schemes that map visual and textual features into a common latent space [12]. Vision-language pre-training models, such as CLIP and its successors, popularized contrastive alignment of large-scale image-text pairs collected from the web [4]. However, the generic nature of such corpora, dominated by everyday objects and scenes, limits their direct transferability to remote sensing, where the lexicon includes specialized geospatial semantics and the visual statistics differ markedly from natural images. Simultaneously, asymmetric learning paradigms have been explored to address modality imbalance. Multi-task architectures that assign different roles to visual and linguistic branches have been shown to outperform symmetric fusion strategies in visual search and visual question answering [13]. The notion of asymmetric semantic excavation, as advanced in recent self-supervised hashing methods, specifically targets the uneven distribution of discriminative cues between representations [16], providing a direct conceptual anchor for the asymmetric mining we extend to the cross-modal case.

Additional relevant threads include weakly supervised localization methods for remote sensing that exploit partial annotations to guide attention toward semantically salient regions [14], and adaptive vision-language alignment frameworks that modulate the contribution of each modality based on input characteristics [15]. Transformer-based architectures, initially proposed for machine translation and subsequently adapted to vision, have become the de facto backbone for multi-modal fusion [22]. Their self-attention mechanisms naturally

accommodate cross-modal interactions but introduce quadratic computational complexity, an important systems consideration for large-scale hashing. Distributed deep learning frameworks and model quantization techniques, while not retrieval-specific, form the operational substrate for training and deploying billion-parameter cross-modal hash models. The convergence of these streams motivates our holistic examination of vision-language hash learning grounded in asymmetric semantic representation mining, which reinterprets algorithmic components through the lens of systemic viability.

3. Conceptual Foundations of Asymmetric Semantic Mining

Symmetric alignment, the predominant paradigm in cross-modal representation learning, posits that visual and textual embeddings of a given scene should map to identical or highly similar points in a shared semantic space. This approach assumes rough informational parity between the two modalities, a condition violated in remote sensing. A satellite image captures continous spectral reflectance values across a spatial grid, revealing texture, context, and object arrangements that are often irrelevant to a high-level textual query; inversely, a query such as “economically marginalized agricultural region” encodes sociological knowledge unobservable from strict visual patterns. Forcing perfect alignment compels each modality to discard information that cannot be reconciled, thereby flattening the representation and reducing its discriminative power for retrieval tasks where the query and target modality carry complementary expectations.

Asymmetric semantic representation mining deliberately designs encoder architectures and loss landscapes to preserve and exploit modality-specific richness while still enabling meaningful cross-modal comparison. In practice, this can entail equipping the vision encoder with higher capacity or a different inductive bias than the language encoder, or configuring the hash layer such that the binary codes of the two modalities reside in distinct subspaces of the Hamming space whose distance metric is still well-defined. The learning objective may combine instance-level contrastive terms, which pull associated image-text pairs closer, with set-level distribution alignment terms that prevent mode collapse without imposing pointwise identity. The self-supervised asymmetric excavation concept, originally formulated for single-modal deep hashing [16], demonstrated that allowing the target binary codes and the continuous embeddings to evolve with different degrees of constraint can yield superior intra-modal retrieval. In the cross-modal remote sensing context, this logic extends to giving the vision pipeline greater representational latitude to model spectral nuance while the language pipeline specializes in compositional semantics, with the hash projection acting as an interface that translates between the two worlds.

The mining of asymmetric semantics also requires careful treatment of negative sampling and hard example selection. Symmetric models typically treat all mismatched image-text pairs as negatives, but in remote sensing this can produce false negatives when two distinct queries describe complementary facets of the same scene (e.g., “forested watershed” and “high runoff zone”). A mining strategy aware of the partial overlap between modalities can adjust the margin of contrastive loss based on estimated semantic distance, reducing the penalty for pairs that are plausibly related. Such margin-scalable constraints, as explored in hashing literature [16], align with the broader asymmetric philosophy by introducing flexibility into the alignment criterion. The resulting hash codes preserve fine-grained visual distinctions that would be erased under a strict one-to-one mapping, while still enabling effective text-based retrieval. The systems implication is that hash code length can often be reduced without

degrading recall, because the binary dimensions are used more efficiently when modalities are not forced to replicate the same information.

4. System Architecture and Deployment Trade-offs

Designing a production-grade vision-language hash retrieval system for remote sensing demands navigating a multidimensional space of architectural choices, each with infrastructure and cost consequences. At the core, the architecture comprises a visual encoder, a language encoder, cross-modal interaction layers, a hash projection head, and an index serving layer. The visual encoder can be based on convolutional backbones such as ResNet [21] or vision transformers; the language encoder is typically a transformer model pre-trained on large text corpora and fine-tuned with domain-specific geospatial vocabularies using subword tokenization strategies [24]. A critical decision is whether cross-modal fusion occurs late, after independent encoding, or early through cross-attention blocks that intermix visual and language tokens. Late fusion simplifies distributed deployment because the two encoders can be hosted on separate hardware and scaled independently, whereas early fusion yields richer interactions at the expense of tightly coupled inference pipelines and higher memory footprints. For remote sensing, where batch retrieval of millions of images per query is common, late fusion architectures that pre-compute database-side binary codes and keep the query encoder lightweight are preferred for sustainability and latency reasons.

The hash projection layer, typically a fully connected layer followed by a sign activation function or a straight-through estimator during training, translates the fused embedding into b-bit binary codes. Code length b is a primary system dial: shorter codes accelerate Hamming distance computation and index memory, while longer codes preserve more semantic fidelity. Empirical evidence from large-scale multimedia benchmarks indicates diminishing returns beyond 64 to 128 bits for many tasks, but the optimal length for asymmetric cross-modal remote sensing settings, where fine-grained land cover distinctions matter, must be empirically determined through rigorous validation on geographically and temporally diverse datasets. The training process itself is a significant engineering undertaking; it requires curating millions of image-text pairs, a non-trivial task given the predominance of noisy labels and unstructured metadata in Earth observation archives. Weakly supervised techniques [14] and automated captioning pipelines can augment human-annotated data, but introduce governance risks concerning label quality and potential bias propagation.

Infrastructure considerations extend to the index serving layer. Once all database images are encoded into binary codes, an index structure such as a multi-index hashing table or a graph-based approximate nearest neighbor index must support sub-linear query time. In cloud-native deployments, the index can be partitioned across multiple nodes with replication for fault tolerance. Edge deployments, envisioned for field-based rapid response scenarios where connectivity is intermittent, require compressed indexes that fit within the memory constraints of embedded systems. The asymmetry between offline database indexing and online query encoding creates favorable scalability characteristics: the expensive vision encoding can be performed incrementally as new imagery arrives, while the lightweight language query encoder runs on a thin client. However, the system must account for temporal drift in land cover, requiring a lifecycle management strategy for re-indexing cycles, versioning of hash models, and backward compatibility of binary codes.

Data governance in such a pipeline intersects with international frameworks regulating satellite data sharing and privacy. High-resolution imagery may inadvertently capture identifiable individuals or sensitive infrastructure, and the retrieval system’s ability to

aggregate scenes matching specific patterns raises dual-use concerns. A governance layer must enforce fine-grained access control lists that restrict certain query templates or return only coarsened hash results without raw image content. Additionally, the provenance and licensing of training data must be tracked to comply with the differing data policies of public agencies and commercial satellite operators. Such requirements demand that the system architecture be instrumentation-ready from the outset, with logging, audit trails, and model cards documenting performance characteristics and limitations.

5. Governance, Fairness, and Policy Implications

The deployment of vision-language hash retrieval in high-stakes domains such as disaster response and environmental regulation carries profound fairness and governance obligations. Training data drawn disproportionately from certain geographic regions, climate zones, or levels of economic development can embed representational biases that manifest as differential retrieval accuracy. A query for "drought-affected farmland," for example, may over-return scenes from well-documented regions in North America or Europe while under-returning relevant scenes from the Global South, where labeled drought examples are scarcer. Such disparities can skew resource allocation decisions, amplify existing inequities, and erode trust in AI-assisted geospatial analysis. Mitigation requires deliberate data curation strategies that oversample underrepresented ecoregions, combined with algorithmic fairness constraints incorporated during hash learning, such as equalized odds criteria across geographic strata. The asymmetric framework offers a partial avenue for fairer representation, as it prevents the language encoder from imposing Anglophone-centric semantic structures onto visual representations that may encode different functional land-use patterns.

Policy frameworks are still nascent for cross-modal AI systems that simultaneously process remotely sensed data and natural language. The European Union's AI Act classifies certain applications of remote sensing, particularly those involving biometric inference or critical infrastructure monitoring, as high-risk, triggering requirements for transparency, human oversight, and robustness. A vision-language hash retrieval system deployed in this context must provide explanations for its retrieval results, a challenging task for binary codes that are inherently non-interpretable. One emerging mitigation is to couple the hash index with a post hoc explanation module that highlights the image regions and query terms most influential in the similarity judgment, leveraging attention maps or occlusion-based sensitivity analysis. Furthermore, dual-use concerns demand that developers and deployers implement gating mechanisms that prevent the system from being used for indiscriminate surveillance or targeting of vulnerable populations. Institutional review processes, comparable to those employed in the biomedical domain, could be adapted to assess the ethical permissibility of specific retrieval use cases before granting API access.

The cross-border nature of satellite data introduces jurisdiction-specific regulatory friction. An Earth observation hash index hosted in one country but queried from another may violate local data sovereignty laws if the binary codes, despite their obfuscated form, can be reverse-engineered to reveal sensitive landscape features. Although binary codes are designed for efficiency rather than human interpretability, concatenating hash bits with auxiliary metadata could potentially enable reconstruction attacks under certain conditions. Therefore, security assessments should treat the hash index as personal or quasi-sensitive data under the strictest applicable regime and apply appropriate encryption at rest and in transit. Standardization efforts within the geospatial community, such as the Open Geospatial Consortium's APIs, can facilitate interoperability while embedding these privacy and security controls by default.

6. Robustness and Sustainability

Robustness of vision-language hash retrieval to distributional shifts is paramount given the dynamic nature of Earth's surface and the diversity of sensor platforms. A model trained on cloud-free, nadir-view summer imagery may fail catastrophically when confronted with off-nadir winter scenes exhibiting snow cover and low sun angles. Asymmetric mining provides a buffer against such shifts by allowing the vision encoder to retain redundant sensory features that become discriminative under domain shift, even if those features were not strongly aligned with training-time language. Nevertheless, systematic robustness engineering requires adversarial training against naturally occurring corruptions, continuous monitoring of query-drift, and periodic retraining triggered by statistical divergence detectors. The deployment infrastructure must support A/B testing of new hash models against live traffic, with fallback mechanisms to earlier model versions if retrieval quality degrades. Comparisons with content-based image retrieval systems in the medical domain, where robustness to scanner variability is a well-studied challenge, offer transferable practices including domain-adversarial feature normalization and calibrated confidence scoring.

Sustainability considerations encompass both the carbon footprint of training large vision-language models and the energy efficiency of the resulting hash retrieval services. Training a transformer-based cross-modal architecture on millions of high-resolution remote sensing patches may consume megawatt-hours of electricity, a non-trivial environmental externality that must be weighed against the societal benefits of improved retrieval. The asymmetric design can mitigate this cost by permitting separate and potentially asynchronous training of the visual and language encoders, allowing the re-use of pre-trained checkpoints that amortize their training emissions over many downstream tasks. Hash code length directly influences the energy budget of large-scale inference: shorter codes reduce the power consumed by Hamming distance computations and memory accesses in the serving infrastructure, aligning with green AI principles. Model compression techniques, including knowledge distillation from a large asymmetric teacher to a compact student model, can further shrink the operational footprint without severely compromising accuracy. Life-cycle assessment methodologies adapted from data center engineering can help operators track and optimize the carbon intensity of their retrieval services over time.

The interplay between robustness and sustainability reveals a systemic tension. Enhancing robustness often requires larger and more diverse training sets, which in turn escalate data storage emissions and training compute. Addressing this tension demands that system architects adopt a multi-objective design process, where retrieval accuracy, fairness, robustness, latency, and environmental impact are co-optimized rather than sequentially traded off. This multi-objective perspective aligns with emerging standards for responsible AI systems engineering and points toward the need for declarative optimization frameworks that allow operators to specify constraints and automatically derive Pareto-optimal configurations for hash model training and deployment.

7. Conclusion

This paper has presented a comprehensive systems analysis of vision-language hash learning for remote sensing scene retrieval, organized around the principle of asymmetric semantic representation mining. By intentionally designing visual and language encoders to capture complementary rather than strictly matched semantics, the approach addresses the inherent informational imbalance between overhead imagery and textual descriptions, yielding hash codes that are more discriminative and compact. The architectural dissection revealed how

choices in encoder type, fusion point, hash length, and indexing topology propagate into infrastructure demands, deployment flexibility, and operational cost. Beyond algorithmic performance, we underscored the governance, fairness, and policy dimensions that are inseparable from real-world deployment. Representational biases in geospatial training data, dual-use risks, and legal constraints on cross-border data flows demand that the system be designed with ethical safeguards and regulatory compliance as first-class requirements. Robustness to environmental and sensor-induced distribution shifts and the environmental sustainability of training and inference were examined as deeply intertwined concerns that call for multi-objective engineering frameworks.

Looking forward, several research frontiers beckon. The integration of asymmetric mining with emerging foundation models for Earth observation, capable of handling multiple modalities beyond visible-spectrum and text, could unlock retrieval across radar, hyperspectral, and temporal query formats. The development of standardized benchmarks that evaluate not only retrieval precision but also fairness across demographic and ecological strata would accelerate progress toward equitable geospatial AI. Cross-disciplinary collaboration among remote sensing scientists, systems engineers, legal scholars, and policy makers will be essential to navigate the evolving landscape of regulation and public expectation. Vision-language hash learning, grounded in asymmetric semantic representation mining, stands as a powerful enabling technology for making the planet's vast and growing archives of remotely sensed knowledge truly searchable, provided that its systemic implications are confronted with intellectual rigor and ethical foresight.

References

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

2. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763.

5. Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020). Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 11336–11344.

6. Gong, Y., Lazebnik, S., Gordo, A., & Perronnin, F. (2013). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2916–2929.

7. Liu, W., Wang, J., Ji, R., Jiang, Y.-G., & Chang, S.-F. (2012). Supervised hashing with kernels. IEEE Conference on Computer Vision and Pattern Recognition, 2074–2081.

8. Zhu, H., Long, M., Wang, J., & Cao, Y. (2016). Deep hashing network for efficient similarity retrieval. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 2415–2421.

9. Lu, X., Zhang, L., & Li, Z. (2020). Learning discriminative deep features for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 58(8), 5666–5681.

10. Li, Y., Zhang, Y., Huang, X., & Han, J. (2021). Remote sensing image retrieval using deep hashing with weighted triplet loss. IEEE Geoscience and Remote Sensing Letters, 19, 1–5.

11. Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. Proceedings of the 26th Annual International ACM SIGIR Conference, 119–126.

12. Feng, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. Proceedings of the 22nd ACM International Conference on Multimedia, 7–16.

13. Wang, Z., Li, Q., & Tao, D. (2016). Asymmetric multi-task learning for visual search. IEEE Transactions on Image Processing, 25(8), 3869–3882.

14. Zhang, D., Han, J., Zhao, L., & Meng, D. (2019). Leveraging prior-knowledge for weakly supervised object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 57(9), 6962–6975.

15. Peng, Y., Qi, J., & Zhuo, Y. (2020). MAVA: Multi-level adaptive visual-textual alignment by cross-media bidirectional matching. Proceedings of the 28th ACM International Conference on Multimedia, 1728–1736.

16. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

17. Ma, L., Liu, Y., & Liu, X. (2022). Learning to hash for big data: Current status and future trends. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 7015–7036.

18. Norouzi, M., Fleet, D. J., & Salakhutdinov, R. (2012). Hamming distance metric learning. Advances in Neural Information Processing Systems, 25, 1061–1069.

19. Weiss, Y., Torralba, A., & Fergus, R. (2008). Spectral hashing. Advances in Neural Information Processing Systems, 21, 1753–1760.

20. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.

21. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

23. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

24. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.

25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.