

A Survey of Authentic-Image Out-of-Context Mismatch Detection Cross-Modal Alignment, Decision Routes, and Evaluation Comparability

Xingya Tian

*School of Mathematics, Shandong University, Jinan, China
3287484903@qq.com*

Abstract. Authentic-image out-of-context (OOC) mismatch detection concerns cases in which an unmanipulated image is paired with text or metadata that reassigns its original event context. Unlike deepfake detection or image-tampering forensics, the core task is to determine whether the image and the accompanying claim refer to the same underlying event context, especially when topical similarity masks contextual inconsistency, rather than to assess visual authenticity. This survey reviews existing research on this problem from the perspective of cross-modal alignment and consistency judgment. Relevant studies remain scattered across evidence construction, training-signal design, and different decision routes, while reported results are often difficult to compare directly. Accordingly, the survey synthesizes the literature through a chain of evidence construction, decision routes, and evaluation/failure analysis, covering global, region-level, and entity-level alignment together with direct decision, contrastive-learning-enhanced decision, and retrieve-evidence-verify. We further summarize major comparability issues, typical failure modes, and their corresponding priorities for future research. The survey aims to provide a concise and judgment-oriented overview of authentic-image out-of-context detection.

Keywords: authentic-image out-of-context mismatch, cross-modal alignment, consistency judgment, evaluation comparability

1. Introduction

With the widespread adoption of multimodal content in social media and news communication, misinformation increasingly appears as image-text packages. Among them, out-of-context (OOC) mismatch with authentic images is especially deceptive: the image itself is real, yet the accompanying text or metadata shifts its original time, place, or event narrative. Unlike deepfakes or image-tampering forensics, the key challenge here is not visual forgery, but whether the image and the accompanying text or metadata point to the same underlying event context [1-4]. The central difficulty, therefore, lies in distinguishing topical similarity from context consistency.

This survey is restricted to authentic-image OOC mismatch detection, in which the image remains genuine while the accompanying text or metadata reassigns its original event context. Unlike studies on deepfake detection, image-manipulation forensics, generic multimodal fake news

detection, or social-context modeling, the present review focuses on whether an image-text package refers to the same underlying event context. Although broad surveys of multimodal fact-checking and multimodal fake news detection already exist [5, 6], research on this task remains fragmented across cross-modal alignment, learning-signal design, and evidence-based or classification-oriented decision strategies. Reported results are also difficult to compare because mismatch type, context granularity, negative-sample difficulty, and evaluation settings vary substantially, with hard mismatches remaining a major source of error [7, 8]. Recent work further extends this line from within-package matching to context recovery and evidence-aware verification [9-13]. This survey therefore provides a unified framework for integrating existing studies on authentic-image OOC mismatch detection.

On this basis, we organize the literature along the chain of cross-modal alignment - from alignment to decision - evaluation logic, comparability and failure analysis. Section 2 reviews evidence construction at the global, region-level, and entity-level. Section 3 compares three representative decision routes: direct decision, contrastive-learning-enhanced decision, and retrieve-evidence-verify. Section 4 discusses evaluation logic, comparability, and typical failure modes.

The main contributions of this survey are as follows. (1) We provide a task-specific review framework for authentic-image OOC mismatch detection, organized around evidence construction, decision routes, and evaluation/failure analysis. (2) Based on this framework, we systematically review cross-modal alignment methods and representative consistency-judgment routes, and assess them by two unified criteria: whether they distinguish topical similarity from context consistency, and whether they adapt to hard mismatches. (3) We summarize key comparability issues, route-aware reporting considerations, and typical failure modes together with their optimization implications, with the aim of supporting more consistent evaluation and future research on authentic-image OOC detection.

Figure 1 presents the overall organization of the survey.

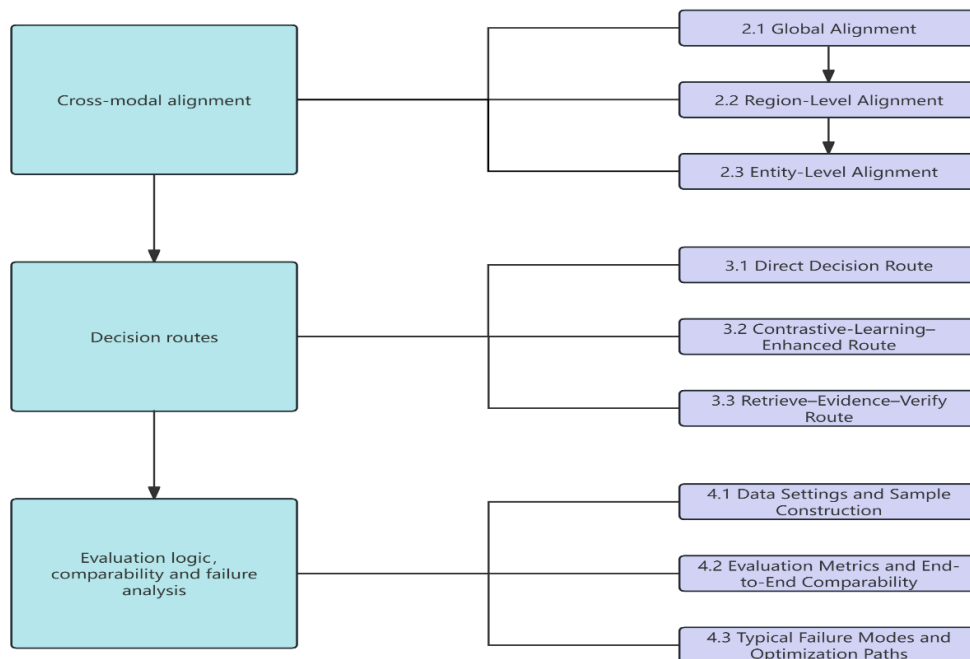


Figure 1. Overall analytical framework of this survey

2. Cross-modal alignment: representation and interaction

2.1. Global alignment

Global alignment encodes the whole image and the whole text/caption into global representations and derives a package-level consistency signal [1, 2].

One representative line treats global alignment as semantic integrity assessment. Multimedia Semantic Integrity Assessment Using Joint Embedding of Images and Text learn a joint image-text embedding, computes an image-text consistency score (ICCS), and judges a query package by its inlier or outlier in the learned reference distribution [2]. This design provides a stable overall consistency signal, but its evidence remains coarse-grained and does not directly identify which contextual anchors cause the mismatch.

For image repurposing, where surface semantics may remain coherent while key contextual anchors shift, Deep Multimodal Image-Repurposing Detection strengthens global judgment through a reference-set–retrieve–compare strategy rather than relying only on within-package matching [1].

Overall, global alignment is efficient for package-level consistency estimation, but it is limited in separating topical similarity from true context consistency, especially under hard mismatches.

2.2. Region-level alignment

Region-level alignment establishes correspondences between image regions and textual words or phrases, thus providing finer evidence units than global alignment for consistency judgment. In image-text contextual mismatch detection, its value lies in revealing local inconsistencies even when the overall semantics still appear similar [14, 15].

One representative route performs explicit region-word matching. SCAN formulates image-text matching as latent semantic alignments between regions and words, and computes similarity through Stacked Cross Attention [14]. Its key idea is to make region-word correspondence part of the matching process itself, so that the model can learn which words should align with which image regions. This route is more sensitive than global alignment to local conflicts, but it still depends on whether critical contextual cues are visually salient and textually alignable.

Another representative route internalizes such fine-grained correspondence into pretrained cross-modal Transformers. LXMERT learns correspondences between textual elements and image regions through cross-modal attention [16]. On this basis, Oscar further introduces detected object tags as alignment anchors, making interaction more directly organized around object-level semantic units [15]. For contextual mismatch detection, this route provides more structured local evidence than overall similarity alone, but its effectiveness remains limited when decisive anchors are weak, implicit, or only indirectly expressed.

2.3. Entity-level alignment

Entity-level alignment focuses on whether the explicitly mentioned entities in the text are consistent with entity evidence that can be located or recognized in the image. Its goal is to ground contextual mismatch in named and verifiable anchors, such as locations, persons, organizations, and brands [1, 2].

Its typical procedure includes three steps: text-side entity extraction and representation learning [17], image-side entity-cue extraction [18-20], and entity-level consistency verification. On the text side, the system identifies entities such as person names, place names, and organization names in the

caption and builds semantic representations for matching. On the image side, it extracts evidence for locations, textual entities, and organization or brand entities through place recognition, scene text detection, and logo detection, respectively [18-20]. Compared with region-level alignment, this line further compresses local evidence into explicit contextual anchors and is therefore better suited to cases in which locations, persons, or organizations are reassigned [1, 2].

At the decision level, entity-level alignment can be modeled in two ways. In the implicit setting, the overall consistency problem is decomposed into whether entities such as locations, persons, organizations, and brands match [1, 2]. In the explicit setting, cross-modal entity consistency is treated as an independent verification task: the system extracts persons, places, and events from the text through entity linking, introduces reference images of these entities as visual evidence, and then computes consistency measures for different entity types [21].

Compared with overall semantic matching, entity-level alignment is more effective for authentic-image OOC mismatch because it targets the anchors most likely to be reassigned while surface semantics remain coherent. This makes it more capable of separating topical similarity from context consistency, especially under hard mismatches. Its limitation, however, is that performance depends on whether critical anchors can be reliably extracted, recognized, and aligned across modalities.

3. From alignment to decision: training and inference strategies

3.1. Direct decision route

The direct decision route maps the alignment evidence constructed in section 2 directly to a consistent / inconsistent output, without introducing external evidence. Its core logic is to make a judgment from the evidence available inside the image-text package itself. Existing methods mainly follow two forms: score-based decision and classification-based decision.

Score-based decision makes judgments from an image-text consistency score. MAIM is a representative example: it casts semantic integrity assessment as the detection of abnormal consistency scores. This route is simple and easy to deploy, but its output still mainly reflects the strength of within-package matching [2].

Classification-based decision instead feeds the aligned joint representation into a classification head and directly outputs the probability of match / mismatch. Vision-language pretraining models such as VisualBERT and LXMERT follow this idea by learning image-text representations with matching classification as an objective [16, 22]. For OOC mismatch, recent work also often adopts a direct pipeline of feature extraction, fusion, and classification, and uses synthetic data to alleviate the lack of labeled samples [23].

Overall, this route is most suitable when sufficient within-package evidence is available. Its limitation is that it remains closer to semantic matching than to event-context verification, and is therefore less effective under hard mismatches or evidence-insufficient settings [1, 2, 23].

3.2. Contrastive-learning-enhanced route

The contrastive-learning-enhanced route improves image-text contextual mismatch detection by enlarging the distance between matched and mismatched samples in the representation space [24, 25]. Its core value lies in strengthening discrimination during training, so that the model is less likely to confuse topical similarity with context consistency.

This route mainly addresses two problems. The first is hard-sample discrimination: negative pairs may remain highly similar to true pairs at the topical level while critical contextual anchors have

already shifted [3, 4]. The second is robustness to pairing noise: in large-scale weakly supervised image-text data, noisy pairs can easily be learned as if they were true alignments [26, 27].

Existing work mainly strengthens this route through harder negative construction and more stable training signals. In the image-text setting, ALBEF combines image-text contrastive learning, in-batch hard negative mining, and Momentum Distillation, thereby improving discrimination under both hard negatives and noisy supervision [26]. In addition, combining discriminative matching objectives with region-word alignment or object-tag anchors can further sharpen cross-modal discrimination under weak supervision [15, 28].

Overall, this route improves sensitivity to hard mismatches through stronger training signals, but it still mainly enhances discriminative boundaries rather than performing explicit event-context verification [26, 29, 30].

3.3. Retrieve–evidence–verify route

The retrieve–evidence–verify route is motivated by cases in which within-package alignment signals are insufficient for stable judgment. In image–text contextual mismatch and image repurposing, mismatched samples may remain semantically coherent on the surface while key contextual anchors such as locations, persons, or organizations have already shifted, and image-side cues may also be weak or missing [1, 2, 23]. This route therefore extends within-package alignment with external evidence retrieval and verification [1, 31].

The first step is to construct retrievable cues. Queries usually center on textual entities and visible anchors in the image, including location appearance, scene structure, readable text, and brand logos [1]. Evidence retrieval may take either an image-to-image or a text-to-text form. The former recovers contextual clues through place recognition or image retrieval, with representative methods such as NetVLAD and DELF; the latter retrieves candidate evidence from text corpora through methods such as DPR, REALM, and RAG [23, 31-35]. For textual and logo cues in images, EAST and LogoDet-3K provide basic support for text entry points and organization-anchor filtering, respectively [19, 20].

At the verification stage, candidate evidence must be filtered, reranked, and judged for consistency. This process can be viewed as an extension of FEVER-style evidence verification to the task discussed in this paper [1, 31]. Recent studies further move this route from direct evidence use toward evidence selection and adaptation: CMIE improves evidence relevance and explainability [9], while COVE and E2LVLM show that external evidence can further support context recovery and evidence-enhanced final judgment [10, 11].

Overall, this route is most suitable when within-package evidence is insufficient and external references are available. Its main limitation is that performance depends heavily on retrieval quality and evidence coverage, so a more systematic cross-route comparison is deferred to section 4.

To summarize the main trade-offs across the three decision routes, Table 1 compares them from the perspectives of evidence source, training signal, interpretability, and system cost.

Table 1. Cross-route comparison of three decision routes in authentic-image OOC detection

Dimension	Direct decision [1, 2, 16, 22, 23]	Contrastive-enhanced [15, 24-30]	Retrieve-evidence-verify [1, 9-13, 18-20, 23, 31-35]
Evidence source	Within-package evidence only.	Within-package evidence with stronger discriminative representation learning.	Within-package evidence plus external retrieved evidence.

Table 1. (continued)

Training signal	Consistency scores or classification heads over aligned representations.	Hard negatives, contrastive objectives, and noise-robust training signals.	Retrieval + verification; effectiveness depends on retrieved evidence quality.
Interpretability	Usually outputs a score or label; limited explanation.	Improves discrimination, but gives little explicit evidence support.	Can return external supporting evidence and stronger verifiable grounding.
System cost	Usually the least costly route.	Mainly increases training complexity.	Highest cost due to retrieval, evidence handling, and multi-stage inference.

4. Evaluation logic, comparability, and failure analysis

4.1. Data settings and sample construction

4.1.1. Datasets and data resources

Evaluation in image-text contextual mismatch detection first depends on how datasets define the task. Existing resources can be grouped into three categories.

The first includes task-oriented datasets that directly target contextual mismatch or image repurposing, such as MAIM, MEIR, NewsCLIPPings, and COSMOS. These resources all relate to authentic-image OOC settings, but differ in whether they emphasize package-level semantic integrity, contextual-anchor replacement, or larger-scale contextual shifts [1-4].

The second includes broader misinformation resources, which mainly provide supplementary corpora or auxiliary evaluation settings rather than the core task definition [36-39]. The third includes foundational visual or image-text resources that mainly support weakly supervised pretraining or evidence extraction [27, 40]. Benchmark-oriented analyses further show that resource design strongly affects what systems appear to learn and how comparable reported gains really are [7, 8].

4.1.2. Negative-sample construction and task difficulty

Negative-sample construction can be roughly grouped into four types. (I) random mismatch, which directly shuffles image-text pairs. It is simple, but usually creates only weak difficulty and is closer to general relevance judgment. (II) entity replacement or context reassignment. This type keeps the overall narrative superficially coherent while replacing key contextual anchors such as locations, persons, organizations, or events [1, 2]. It is therefore closer to the real form of image repurposing or OOC mismatch. (III) retrieval-based hard negatives. These are built by retrieving semantically similar instances from large corpora or image repositories, so that negative samples remain close to true samples in topic, scene, or objects, while their key contexts are still inconsistent [3, 4]. Such negatives are closer to the difficult mismatches encountered in real deployment. (IV) manually constructed or in-the-wild samples. These are often more realistic, but they are also harder to scale, annotate, and control consistently.

Overall, negative-sample construction directly determines whether evaluation mainly reflects coarse relevance or the harder ability to distinguish topical similarity from true context consistency.

4.2. Evaluation metrics and end-to-end comparability

Common classification metrics such as Accuracy, Precision, Recall, and F1 are widely used in matched / mismatched classification, but they are not equally informative for authentic-image OOC detection. In particular, Accuracy is not recommended as a primary metric, because this task is naturally affected by class imbalance and hard mismatches [41, 42], and a single overall score can easily obscure the real difficulty of context-consistency judgement [7, 8].

To improve cross-study comparability, this survey proposes a compact evaluation protocol with three mandatory basic measures: Macro-F1, which reduces the dominance of majority classes; AUC-PR, which is more informative under imbalance [41]; and hard-mismatch F1, which separately reports performance on the most difficult cases where topical similarity is preserved but contextual anchors have shifted [7, 8]. Together, these measures better reflect whether a method can distinguish topical similarity from context consistency.

In addition, different decision routes should report route-specific supplementary measures. Direct decision methods should report the stability of consistency-score calibration; contrastive-learning-enhanced methods should report hard-negative separability in representation space; and retrieve-evidence-verify methods should report Top-5 retrieval precision and evidence support effectiveness. Therefore, methods from different routes should not be compared mechanically by a single F1 value alone, because their outputs are shaped by different evaluation logics and system dependencies [7, 8, 31-34].

4.3. Typical failure modes and research priorities

Based on the discussion above, typical failures in image-text contextual mismatch detection can be grouped into three broad categories.

The first type is failure caused by insufficient within-package evidence. In such cases, the sample may remain semantically coherent at the surface level while key contextual anchors such as locations, persons, organizations, or events have already shifted, leading the model to judge "related" rather than "consistent" [1, 23]. This problem becomes more severe when crucial contextual cues are weak, missing, or visually inaccessible [14, 18-20].

The second type is failure caused by insufficient or distorted training signals. If negative samples are too easy, the model tends to learn only coarse relevance rather than boundaries that are truly useful for contextual mismatch detection; pairing noise in large-scale weakly supervised image-text data can further weaken or distort the learned correspondences [3, 4, 26, 27].

The third type is failure caused by instability in the external evidence chain. In the retrieve-evidence-verify route, errors may arise when query cues are incomplete, correct evidence is not retrieved, candidate evidence is insufficient or conflicting, or the final verification fails to produce a reliable judgment [31, 32, 35].

Overall, these three failure types suggest three corresponding priorities for future work. First, evidence construction should move toward stronger fine-grained alignment and implicit-anchor extraction, so that systems can rely less on coarse package-level similarity and more on explicit contextual cues. Second, training design should adopt more standardized hard-negative construction and more robust objectives under noisy supervision, in order to improve discrimination on topically similar but contextually inconsistent cases. Third, retrieval-based systems should improve the coordination between retrieval and verification, together with multi-evidence fusion, so that external evidence becomes not only available but also relevant, sufficient, and verifiable.

5. Conclusion

This survey reviewed authentic-image OOC mismatch detection, focusing on cases where the image is genuine but its event context is reassigned by text or metadata. We organized existing studies through a unified framework of evidence construction, decision routes, and evaluation/failure analysis, and clarified how data settings, negative-sample difficulty, and route heterogeneity affect cross-study comparison. On this basis, the main bottlenecks can be summarized as insufficient within-package evidence, distorted training signals, and unstable external evidence chains. Future progress therefore depends on better contextual-anchor grounding, stronger hard-negative design, and more reliable retrieval-verification mechanisms.

References

- [1] Sabir, E., AbdAlmageed, W., Wu, Y., et al. (2018). Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 1337–1345).
- [2] Jaiswal, A., Sabir, E., AbdAlmageed, W., et al. (2017). Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1465–1471).
- [3] Luo, G., Darrell, T., & Rohrbach, A. (2021). Newscippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6801–6817).
- [4] Aneja, S., Bregler, C., & Nießner, M. (2021). Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.
- [5] Akhtar, M., Schlichtkrull, M., Guo, Z., et al. (2023). Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5430–5448).
- [6] Li, X., Qiao, J., Yin, S., et al. (2025). A survey of multimodal fake news detection: A cross-modal interaction perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence*.
- [7] Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., et al. (2024). Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1), 4.
- [8] Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., et al. (2025). Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5041–5050.
- [9] Li, F., Wu, J., He, C., et al. (2025). CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. *Findings of the Association for Computational Linguistics: ACL 2025*, 9342–9354.
- [10] Tonglet, J., Thiem, G., & Gurevych, I. (2025). COVE: COntext and VERacity prediction for out-of-context images. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2029–2049.
- [11] Wu, J., Fu, Y., Yu, N., et al. (2025). E2lvm: Evidence-enhanced large vision-language model for multimodal out-of-context misinformation detection. *arXiv preprint arXiv:2502.10455*.
- [12] Qi, P., Yan, Z., Hsu, W., et al. (2024). Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062.
- [13] Tonglet, J., Moens, M. F., & Gurevych, I. (2024). “Image, tell me your story!” Predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 7845–7864).
- [14] Lee, K. H., Chen, X., Hua, G., et al. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 201–216).
- [15] Li, X., Yin, X., Li, C., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121–137). Springer International Publishing.
- [16] Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5100–5111).

- [17] Lample, G., Ballesteros, M., Subramanian, S., et al. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 260–270).
- [18] Arandjelovic, R., Gronat, P., Torii, A., et al. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5297–5307).
- [19] Zhou, X., Yao, C., Wen, H., et al. (2017). East: An efficient and accurate scene text detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5551–5560).
- [20] Wang, J., Min, W., Hou, S., et al. (2022). Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1), 1–19.
- [21] Müller-Budack, E., Theiner, J., Diering, S., et al. (2020). Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (pp. 16–25).
- [22] Li, L. H., Yatskar, M., Yin, D., et al. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv: 1908.03557*.
- [23] Shalabi, F., Nguyen, H. H., Felouat, H., et al. (2023). Image-text out-of-context detection using synthetic multimodal misinformation. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 605–612). IEEE.
- [24] Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv: 1807.03748*.
- [25] Chen, T., Kornblith, S., Norouzi, M., et al. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.
- [26] Li, J., Selvaraju, R., Gotmare, A., et al. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 9694–9705.
- [27] Sharma, P., Ding, N., Goodman, S., et al. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.
- [28] Chen, Y. C., Li, L., Yu, L., et al. (2020). Uniter: Universal image-text representation learning. *European Conference on Computer Vision*, 104–120.
- [29] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.
- [30] Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *Proceedings of the 38th International Conference on Machine Learning*, 4904–4916.
- [31] Thorne, J., Vlachos, A., Christodoulopoulos, C., et al. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819).
- [32] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781).
- [33] Guu, K., Lee, K., Tung, Z., et al. (2020). Retrieval augmented language model pre-training. In *International Conference on Machine Learning* (pp. 3929–3938). PMLR.
- [34] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- [35] Noh, H., Araujo, A., Sim, J., et al. (2017). Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3456–3465).
- [36] Liu, F., Wang, Y., Wang, T., et al. (2021). Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6761–6771).
- [37] Biamby, G., Luo, G., Darrell, T., et al. (2022). Twitter-COMMS: Detecting climate, COVID, and military multimodal misinformation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1530–1549).
- [38] Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 6149–6157).
- [39] Shu, K., Mahudeswaran, D., Wang, S., et al. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), 171–188.
- [40] Lin, T. Y., Maire, M., Belongie, S., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer International Publishing, Cham.

- [41] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).
- [42] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.