Assessing the Efficacy of DinoV2-Based Embeddings in Clustering Visual Data from C2C Marketplaces

Michael Okonkwo and Pablo Rivas

Department of Computer Science, Baylor University, Texas, USA
{Michael_Okonkwo1,Pablo_Rivas}@Baylor.edu

Abstract. The rapid expansion of online consumer-to-consumer (C2C) marketplaces has generated a vast amount of image-based data, presenting unique challenges and opportunities for automated analysis. In this work, we propose a robust framework that leverages the DinoV2-base Vision Transformer to convert raw car parts images into high-dimensional embeddings. These embeddings are then effectively reduced via Principal Component Analysis (PCA) and visualized using Uniform Manifold Approximation and Projection (UMAP), revealing intrinsic data structures. By applying k-means clustering and rigorously evaluating cluster quality with the Reduced Silhouette Score, Reduced Calinski-Harabasz Index, and Reduced Davies-Bouldin Index, our experiments demonstrate that a 64-dimensional representation achieves the best balance between intra-cluster cohesion and inter-cluster separation. These promising results underscore the potential of our approach as a scalable and automated solution for monitoring and analyzing visual data in C2C marketplaces.

Keywords: Image Embedding · Clustering · Vision Transformers

1 Introduction

Online consumer-to-consumer marketplaces have experienced rapid expansion in recent years, offering individuals an accessible platform to buy and sell a wide range of products. However, this growth has also led to a substantial increase in the volume of visual data, presenting new challenges in monitoring and analyzing the vast number of image-based listings. Traditional methods that rely on manual review or heuristic approaches are increasingly inadequate given the scale and diversity of the data.

In this study, we propose an AI-assisted framework to evaluate the potential of high-dimensional image embeddings for further analysis using vision-language models (VLMs). Focusing on car parts as a representative domain, our approach leverages state-of-the-art Vision Transformers (ViTs), specifically the DinoV2-base model [24], to transform raw image data into rich numerical representations. These embeddings are designed to capture both local and global semantic features, thereby providing a comprehensive characterization of the visual content.

Following the embedding process, we apply dimensionality reduction techniques such as PCA and UMAP to convert the high-dimensional representations into lower-dimensional spaces suitable for clustering and visualization. We then employ k-means clustering to group similar listings and evaluate the quality of these clusters using metrics including the Reduced Silhouette Score, the Reduced Calinski-Harabasz Index, and the Reduced Davies-Bouldin Index. This evaluation allows us to determine the optimal dimensionality for the embeddings and to assess whether the resulting representation is sufficiently informative for downstream analytical tasks.

It is important to note that the present work does not aim to detect illicit activities within the marketplace directly. Instead, our objective is to rigorously assess the embedding space produced by modern ViTs and to establish a foundation for future work that may integrate these embeddings with VLMs for more complex analyses [10,1]. By doing so, we seek to contribute to developing automated tools capable of analyzing large-scale visual datasets in C2C platforms.

The remainder of the paper is organized as follows. A short review of the literature is presented in Sction 2. Section 3 details the methodology for data collection, image embedding, dimensionality reduction, and clustering evaluation. Section 4 presents the experimental results, and Section 5 concludes with a discussion of the findings and directions for future work.

2 Literature Review

The detection of illegal transactions in online marketplaces is a complex problem that requires a careful integration of advanced computer vision methods, robust unsupervised clustering techniques, and ethical considerations in automated systems. This review synthesizes the state-of-the-art research on these topics, comparing transformer-based models with traditional convolutional neural networks (CNNs), discussing image embedding methods and dimensionality reduction, and examining clustering quality metrics. In addition, we describe prior work that has leveraged multi-modal data to improve detection performance in similar settings.

2.1 Computer Vision Models: Vision Transformers vs. CNNs

Recent research has demonstrated that transformer-based models, particularly ViTs, provide significant advantages over traditional CNNs in certain applications. ViTs use self-attention mechanisms to capture long-range dependencies in an image, whereas CNNs extract features hierarchically through convolutional layers [12]. For example, when large datasets are available, ViTs have been shown to achieve better accuracy by modeling global relationships, while CNNs often offer faster training and lower computational cost for smaller datasets [4,17]. Despite these differences, both architectures remain relevant, and the choice between them depends on the specific constraints of the application.

2.2 Image Embedding and Dimensionality Reduction

Transforming high-dimensional image data into a lower-dimensional space is a crucial step for both visualization and clustering. Techniques such as PCA and UMAP are widely used for dimensionality reduction. PCA captures the directions of maximum variance in the data and can be expressed mathematically as:

$$\mathbf{X}_{\text{reduced}} = \mathbf{X}\mathbf{W}, \tag{1}$$

where $\mathbf{W}$ contains the principal components. However, PCA may not fully capture nonlinear relationships in the data. UMAP, on the other hand, preserves local data structures and can reveal complex patterns that may be critical for detecting illicit activities [16,22].

The quality of clustering on embedded data is often assessed with metrics such as the silhouette score, which for a point i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2}$$

where $a(i)$ is the average distance within the same cluster and $b(i)$ is the average distance to points in the nearest neighboring cluster. Other metrics, such as the Calinski-Harabasz and Davies-Bouldin indices, further help in evaluating the balance between intra-cluster cohesion and inter-cluster separation [23,7].

2.3 Unsupervised Clustering Techniques

Unsupervised clustering methods, including k-means, hierarchical clustering, and DBSCAN, are vital when labeled examples are scarce. k-means clustering, for instance, groups data by minimizing the within-cluster sum of squares. Given a set of points $\{x_1, x_2, \ldots, x_n\}$ and k clusters with centroids $\{c_1, c_2, \ldots, c_k\}$, the objective is to minimize:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - c_j\|^2. \tag{3}$$

Although k-means is computationally efficient, its assumption of spherical clusters can be a limitation in real-world data with irregular shapes [21,18]. Alternative methods such as DBSCAN are better suited for datasets with noise and outliers [15,13].

2.4 Challenges in Processing Large Image Datasets

Law enforcement and regulatory bodies face substantial challenges when processing vast image datasets from online marketplaces. Variability in image format, resolution, and quality complicates the analysis, often overwhelming traditional methods [9,14]. Moreover, the computational requirements to process and analyze these large datasets necessitate the use of efficient algorithms and scalable hardware. Ethical considerations are also paramount, as automated monitoring systems must be designed to ensure fairness and mitigate bias [5,8].

2.5 Multi-Modal Approaches and Prior Work

Integrating multiple data modalities—such as text, images, and audio—can enhance the detection of illicit activities by capturing a fuller picture of online transactions. Multi-modal models can represent semantic information that is not available through single-modality approaches. Our prior work demonstrates this potential. Others such as Hamara [10] and Armijo [1] have worked on similar projects. Hamara and Armijo utilized the Image Bind and ViT-Base models, respectively. Hamara's model was distinct from our own as it identified multimodal patterns. With images serving as an anchor, his approach enabled the representation of semantic meaning across different modalities, even those that are not typically paired together in datasets. In our study, we combine text, image, and audio modalities. While Armijo focused solely on visual data, Hamara's research incorporated both visual and textual data. The ultimate goal of our project is to evaluate the most effective embedding models for detecting online crime via C2C transactions.

2.6 Ethical Considerations

The deployment of automated systems for detecting illegal transactions raises several ethical issues, including privacy, accountability, and the risk of automation bias. As these systems increasingly influence decision-making in sensitive areas, transparency in algorithmic processes becomes essential. Frameworks that promote fairness and mitigate bias must be integrated into the design and implementation of such systems [3,6,2]. Continuous evaluation and adaptation of these ethical standards are necessary as the technology evolves [11,20,19].

We close this section by affirming that the detection of illegal transactions in online marketplaces benefits from the integration of advanced computer vision techniques, effective image embedding, and dimensionality reduction methods, and robust unsupervised clustering. The comparative analysis of Vision Transformers and CNNs, along with the evaluation of clustering quality through established metrics, provides a strong framework for this task. Moreover, multi-modal approaches and ethical considerations are key to ensuring that the developed systems are both effective and responsible.

3 Methodology

This section describes the methodological framework employed to evaluate the representational quality of high-dimensional image embeddings derived from car parts listings. Our approach consists of four main stages: data collection and preprocessing, image embedding, dimensionality reduction, and unsupervised clustering evaluation.

3.1 Data Collection and Preprocessing

The dataset for this study comprises visual data depicting various car parts obtained from online C2C marketplace postings. Raw images were sourced from compressed .tar archives provided by academic collaborators. To ensure data quality and computational efficiency, we developed a preprocessing pipeline that performs image extraction, validation, and embedding. Each image is converted into a Python Imaging Library (PIL) object, assigned a unique identifier, and stored along with its associated metadata. Duplicate and corrupted images are automatically excluded to maintain a consistent dataset.

The overall workflow is summarized in Algorithm 1. This algorithm details the steps involved in extracting images from the tar archive, validating each image, converting it to the PIL format, and generating high-dimensional embeddings using the DinoV2-base model. Including this algorithm in our methodology provides a clear, reproducible, and rigorous description of our data preprocessing procedure.

Algorithm 1 Preprocessing and Embedding Pipeline

```
procedure PreprocessAndEmbed($T$)                ▷ Input: Tar file $T$ with raw images
    $I \leftarrow$ ExtractImages($T$)            ▷ Extract images from the tar archive
    for each image $i \in I$ do
        if IsValid($i$) and NotDuplicate($i$) then
            $img \leftarrow$ ConvertToPIL($i$)   ▷ Convert to PIL image
            $\mathbf{x} \leftarrow$ DinoV2Base($img$)   ▷ Generate ViT embedding
            StoreEmbedding($\mathbf{x}$, GetID($i$))
        end if
    end for
end procedure
```

3.2 Image Embedding

High-dimensional embeddings are generated using the DinoV2-base model [24], a ViT that effectively captures subtle semantic features from images. Batch processing is utilized to efficiently transform large numbers of images into vector representations, ensuring scalability for our dataset.

3.3 Dimensionality Reduction

To facilitate further analysis, the high-dimensional embeddings are reduced using two techniques: PCA and UMAP. PCA is applied to compress the embeddings to 16, 32, 64, and 128 dimensions, while UMAP projects the data into two dimensions for visualization. This step enables the evaluation of how dimensionality impacts the quality of clustering.

3.4 Unsupervised Clustering and Evaluation

The reduced embeddings are clustered using the k-means algorithm with a fixed number of 20 clusters. Clustering performance is quantitatively evaluated using the Reduced Silhouette Score, the Reduced Calinski-Harabasz Index, and the Reduced Davies-Bouldin Index. These established metrics guide the selection of the optimal embedding dimensionality. Additionally, the k-nearest neighbors (KNN) algorithm is applied to the optimal embedding space to retrieve the nearest neighbors for each cluster centroid, providing further insight into the semantic consistency of the clusters.

4 Results

This section presents the evaluation of our embedding space using unsupervised clustering and KNN analysis. Our aim is to assess whether the transformed image embeddings are sufficiently discriminative for further analysis with vision-language models. We first summarize the clustering performance across various dimensionality reductions and then discuss the insights obtained from the KNN evaluation on the optimal 64-dimensional embedding space.

4.1 Clustering Evaluation

Table 1 summarizes the clustering performance metrics obtained from the embeddings reduced via PCA to 16, 32, 64, and 128 dimensions. The metrics include the Reduced Silhouette Score (where higher values indicate better separation), the Reduced Calinski-Harabasz Index (with higher values reflecting improved clustering dispersion), and the Reduced Davies-Bouldin Index (where lower values denote better cluster compactness). As shown, the 64-dimensional embedding achieves the highest Silhouette Score (0.05705) and the lowest Davies-Bouldin Index (3.40935), indicating that this setting provides a more promising representation of the data.

Table 1. Clustering Performance Metrics for Different Dimensionality Reductions

Dims	Silhouette Score	Calinski-Harabasz Idx	Davies-Bouldin Idx
16	0.05394	1526.86	3.58839
32	0.05616	1528.65	3.58213
64	0.05705	1539.84	3.40935
128	0.05610	1544.15	3.50870

Fig. 1–4 display the UMAP visualizations for the embeddings reduced to 16, 32, 64, and 128 dimensions, respectively. These figures visually corroborate the quantitative metrics presented in Table 1, with the 64-dimensional embeddings exhibiting more distinct cluster boundaries.

4.2 KNN Evaluation

To further assess the discriminative power of the 64-dimensional embedding space, we performed a KNN analysis on the clusters obtained from k-means

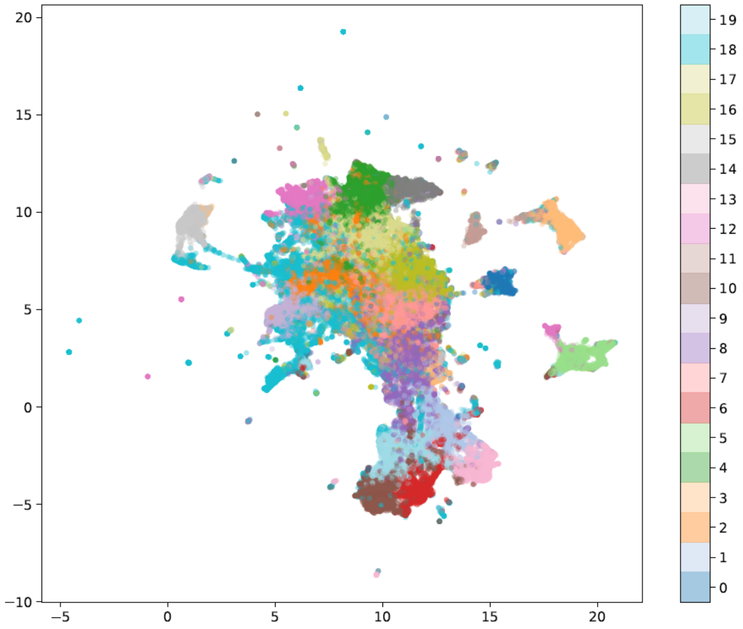

Fig. 1. UMAP visualization of 16-dimensional embeddings.

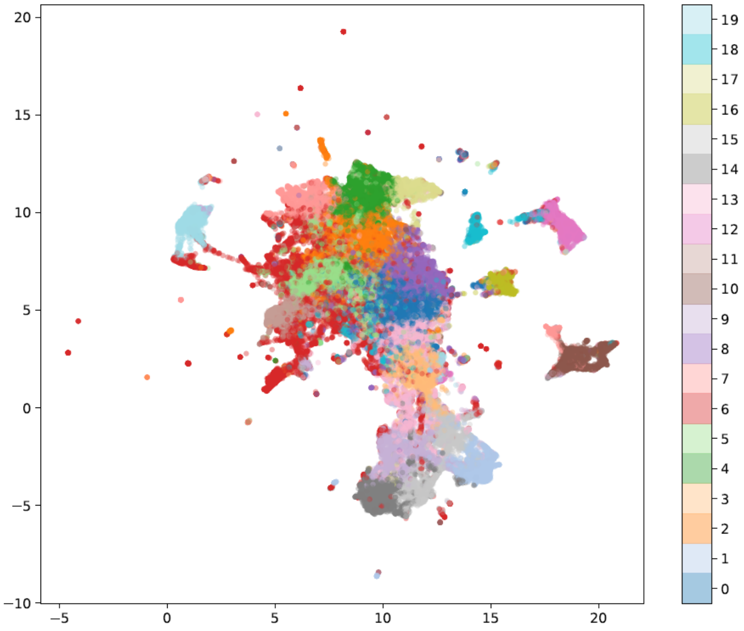

Fig. 2. UMAP visualization of 32-dimensional embeddings.

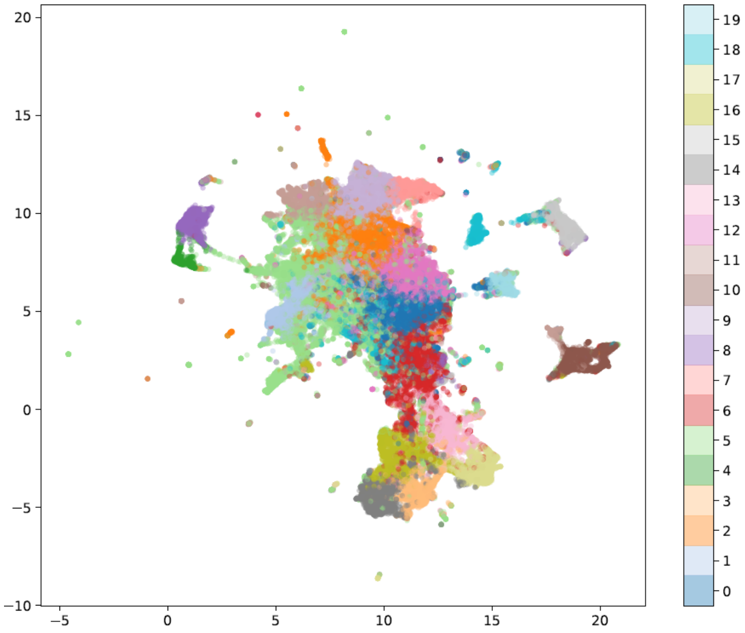

Fig. 3. UMAP visualization of 64-dimensional embeddings.

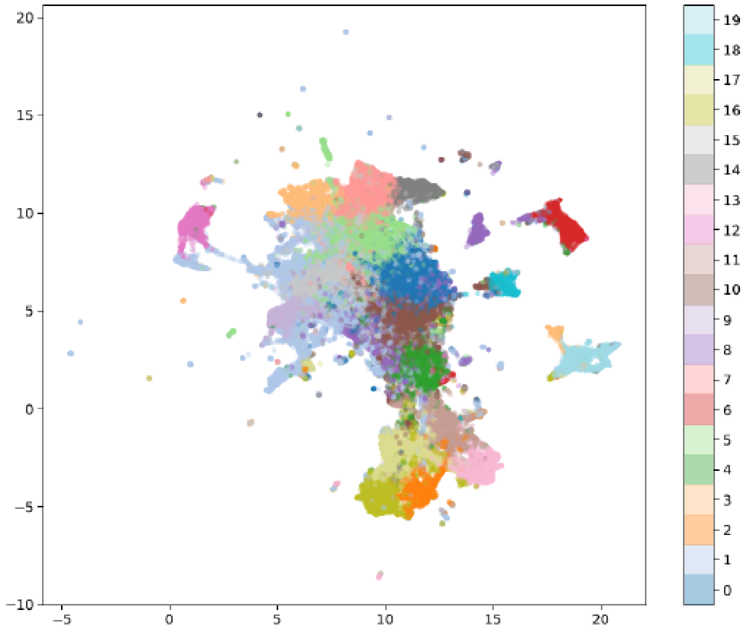

Fig. 4. UMAP visualization of 128-dimensional embeddings.

clustering. For each cluster, the ten nearest neighbor images to the cluster centroid were identified. Fig. 5 presents an illustrative example for a cluster that predominantly comprises images of off-road tires.

Fig. 5 displays multiple photographs of a set of four off-road tires captured from various perspectives. Specifically, the figure includes:

1. A view of the tires leaning against a neutral background, highlighting their rugged tread patterns and multi-spoke wheel design.
2. Close-up shots that emphasize the tire branding, notably the “BIG FOOT A/T” label along with raised white lettering and other specification details.
3. Images that capture the deep tread grooves, showcasing the tire’s texture and design intended for off-road use.
4. An overall lineup of the tires, demonstrating uniformity in appearance and confirming the semantic similarity within the cluster.

These results indicate that the embedding space preserves fine-grained visual details and semantic similarities, thereby supporting its potential for integration with advanced vision-language models in future work. It is important to note that the present study focuses solely on evaluating the embedding space, rather than directly addressing the detection of illicit activity.

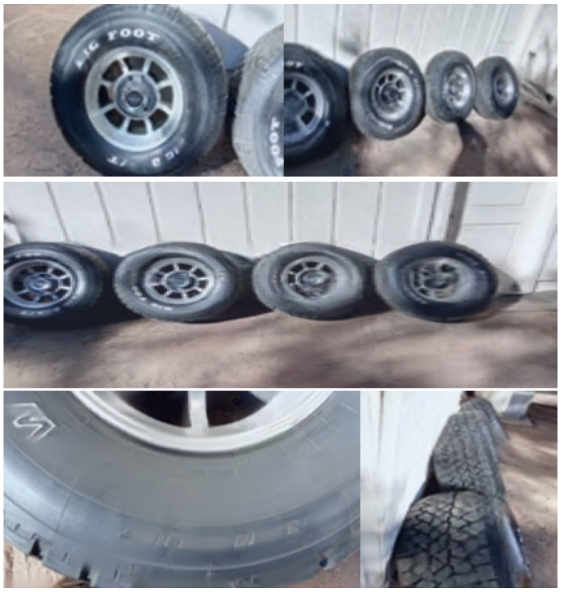

Fig. 5. KNN results for a cluster of off-road tires. The figure shows multiple images of a set of four tires taken from various angles, including close-ups of branding (e.g., “BIG FOOT A/T”) and tread details, as well as an overall lineup, demonstrating the consistency and semantic similarity captured by the embedding space.

5 Conclusion

In this work, we evaluated the potential of high-dimensional image embeddings for subsequent analysis with vision-language models by focusing on unsupervised clustering and KNN evaluation. Our experimental results indicate that reducing the embedding space to 64 dimensions via PCA yielded the best performance in terms of cluster quality, as demonstrated by the highest Reduced Silhouette Score and the lowest Reduced Davies-Bouldin Index. Additionally, the KNN analysis on the 64-dimensional embedding space revealed that the resulting clusters are semantically coherent with respect to the visual characteristics of car parts.

It is important to emphasize that this study does not directly address the detection of illicit activities; rather, it rigorously assesses whether the embedding space is sufficiently informative for more complex downstream tasks. The findings suggest that models such as Dino-V2 effectively capture subtle visual features, providing a solid foundation for future research that may integrate these embeddings with vision-language models to analyze large-scale image data in consumer-to-consumer marketplaces.

Overall, this work contributes to a better understanding of optimal dimensionality settings for clustering high-dimensional embeddings and highlights the promise of transformer-based approaches in representing visual data. Future research will extend these methods to more complex analytical tasks, with careful consideration of ethical implications and data privacy.

Acknowledgments. The authors thank the Rivas.AI Lab (`https://lab.rivas.ai`) for their support and helpful feedback throughout this project. Part of this work was funded by the National Science Foundation under grants CNS-2210091 and CNS-2136961.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armijo, C., Rivas, P.: Exploring visual embedding spaces induced by vision transformers for online auto parts marketplaces. In: AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention Workshop at AAAI 2025 (2025)
2. Bhandari, H.: Automation of databases in oracle. International Journal for Research in Applied Science and Engineering Technology **11**(11), 946–952 (2023). `https://doi.org/10.22214/ijraset.2023.56643`
3. Bonnefon, J., Shariff, A., Rahwan, I.: The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE **107**(3), 502–504 (2019). `https://doi.org/10.1109/jproc.2019.2897447`
4. Calvetti, D., Mêda, P., Sujan, S., Gonçalves, M., Sousa, H.: Challenges of upgrading craft workforce into construction 4.0: framework and agreements. Proceedings of the Institution of Civil Engineers - Management Procurement and Law **173**(4), 158–165 (2020). `https://doi.org/10.1680/jmapl.20.00004`

5. Dyoub, A., Lisi, F.: Logic programming and machine ethics. Electronic Proceedings in Theoretical Computer Science **325**, 6–17 (2020). https://doi.org/10.4204/eptcs.325.6
6. Fossa, F.: Autonomy and automation. the case of connected and automated vehicles pp. 244–248 (2022). https://doi.org/10.33965/ict_wbc_eh2022_202204c030
7. Fusté-Forné, F., Jamal, T.: Co-creating new directions for service robots in hospitality and tourism. Tourism and Hospitality **2**(1), 43–61 (2021). https://doi.org/10.3390/tourhosp2010003
8. Gianfrancesco, M., Tamang, S., Yazdany, J., Schmajuk, G.: Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine **178**(11), 1544 (2018). https://doi.org/10.1001/jamainternmed.2018.3763
9. Goodall, N.: Away from trolley problems and toward risk management. Applied Artificial Intelligence **30**(8), 810–821 (2016). https://doi.org/10.1080/08839514.2016.1229922
10. Hamara, A., Rivas, P.: From latent to engine manifolds: Analyzing imagebind's multimodal embedding space. In: Proceedings of The 26th International Conference on Artificial Intelligence (ICAI'24) (July 2024). https://doi.org/10.48550/arXiv.2409.10528, https://arxiv.org/pdf/2409.10528.pdf, arXiv:2409.10528 [cs.CV]
11. Hiltunen, T.: Reporting and managing ethical issues in intensive care using the critical incident reporting system. Nursing Ethics (2024). https://doi.org/10.1177/09697330241244514
12. Ho, A.: Are we ready for artificial intelligence health monitoring in elder care? BMC Geriatrics **20**(1) (2020). https://doi.org/10.1186/s12877-020-01764-9
13. Ienca, M., Wangmo, T., Jotterand, F., Kressig, R., Elger, B.: Ethical design of intelligent assistive technologies for dementia: a descriptive review. Science and Engineering Ethics **24**(4), 1035–1055 (2017). https://doi.org/10.1007/s11948-017-9976-1
14. Ilyas, Q., Ahmad, M.: An enhanced deep learning model for automatic face mask detection. Intelligent Automation & Soft Computing **31**(1), 241–254 (2022). https://doi.org/10.32604/iasc.2022.018042
15. Javed, H.: Ethical frameworks for machine learning in sensitive healthcare applications. IEEE Access **12**, 16233–16254 (2024). https://doi.org/10.1109/access.2023.3340884
16. Langer, M., König, C.J., Back, C., Hemsing, V.: Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. Journal of Business and Psychology **38**(3), 493–508 (2023). https://doi.org/10.1007/s10869-022-09829-9
17. Nabukenya, S., Okoboi, S., Nakate, V., Twimukye, A., Opio, B., Castelnuovo, B.: Researchers experience of using the regulatory affairs information system (rais) in strengthening research compliance in a large research program: a case study of the infectious diseases institute (idi) in uganda. Jamia Open **5**(3) (2022). https://doi.org/10.1093/jamiaopen/ooac059
18. Onnasch, L., Hösterey, S.: Stages of decision automation: impact on operators' role, awareness and monitoring. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **63**(1), 282–286 (2019). https://doi.org/10.1177/1071181319631126
19. Rotaru, T., Amariei, C.: Ethical issues in research with artificial intelligence systems. In: Radenkovic, M. (ed.) Ethics, chap. 6. IntechOpen, Rijeka (2023). https://doi.org/10.5772/intechopen.1001451

20. Spencer, D.: Automation and well-being: bridging the gap between economics and business ethics. Journal of Business Ethics **187**(2), 271–281 (2022). https://doi.org/10.1007/s10551-022-05258-z
21. Strauß, S.: Deep automation bias: how to tackle a wicked problem of ai? Big Data and Cognitive Computing **5**(2), 18 (2021). https://doi.org/10.3390/bdcc5020018
22. Tiahunova, M.: The automated system of the trolleybus park as part of the sustainable city infrastructure. IOP Conference Series: Earth and Environmental Science **1415**(1), 012023 (2024). https://doi.org/10.1088/1755-1315/1415/1/012023
23. Wen, W., Kuroki, Y., Asama, H.: The sense of agency in driving automation. Frontiers in Psychology **10** (2019). https://doi.org/10.3389/fpsyg.2019.02691
24. Wu, Y., Li, X., Li, J., Yang, K., Zhu, P., Zhang, S.: Dino is also a semantic guider: Exploiting class-aware affinity for weakly supervised semantic segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1389–1397 (2024). https://doi.org/10.1145/3664647.3681710