Enhancing CLIP Embedding Spaces with Contrastive and sliced-Wasserstein Objectives

Siddharth Shah
Department of Computer Science
Vanderbilt University
Nashville, TN
siddharth.p.shah@vanderbilt.edu

Galen Wei
Department of Computer Science
Vanderbilt University
Nashville, TN
galen.wei@vanderbilt.edu

Aditya Shrey
Department of Computer Science
Vanderbilt University
Nashville, TN
aditya.shrey@vanderbilt.edu

Saman Kittani
Department of Computer Science
Vanderbilt University
Nashville, TN
saman.r.kittani@vanderbilt.edu

Abstract

Contrastive Language-Image Pretraining (CLIP) has demonstrated impressive performance across a range of vision-language tasks by learning a shared embedding space for images and text. In this project, we investigate methods to refine and disentangle this shared embedding space by freezing the final later of the CLIP model and retraining the penultimate layers using alterative objective functions. Specifically, we explore variants of contrastive loss and sliced-Wasserstein distance to better align semantic structure and improve embedding separability. Our goal is to enhance representational quality while preserving the zero-shot generalization capabilities of CLIP. We present both qualitative and quantitative evaluations of the modified embedding spaces and discuss the trade-offs of using each loss function.

1 Introduction

Recent advances in multimodal representation learning, exemplified by models such as Contrastive Language-Image Pretraining (CLIP) [11], have significantly enhanced the integration of visual and textual information. CLIP achieves robust zero-shot performance by aligning images and texts in a shared embedding space through a contrastive objective, enabling generalization to previously unseen classes without explicit supervision. However, despite these successes, CLIP's embeddings are not without limitations. Specifically, their learned representations can exhibit semantic entanglement, making fine-grained distinctions among closely related concepts challenging and negatively affecting performance in tasks requiring high semantic precision [16, 15].

Refinement and disentanglement of the learned embedding spaces represent critical challenges to further improve the effectiveness of the multimodal model. Recent literature suggests that embedding clarity and disentanglement directly influence interpretability and downstream performance, particularly in high dimensional or sensitive domains [2, 13]. However, refining embeddings from large pretrained models such as CLIP remains underexplored, particularly with regard to preserving zero-shot capabilities, an essential benefit of these architectures.

In this work, we address these limitations by investigating novel embedding refinement methods for CLIP. Our approach strategically freezes the final layer of the pretrained CLIP model, preserving its well established semantic alignment, and selectively retrains the penultimate embedding layers

Preprint. Under review.

using alternative objective functions designed to improve embedding separability and semantic coherence. Specifically, we experiment with variants of contrastive losses, which directly enhance the discrimination between semantic classes, as well as the sliced-Wasserstein distance, which has shown promise in capturing nuanced geometric structures within high-dimensional spaces. Through this methodological exploration, our study addresses one central question: Can the proposed embedding refinement techniques enhance semantic clarity and disentanglement within CLIP embeddings?

Our primary contributions are as follows: (1) We introduce and evaluate targeted refinement techniques using contrastive loss variants and sliced-Wasserstein distance, demonstrating their effectiveness in improving embedding structure. (2) We provide a qualitative and quantitative evaluation of embedding improvements across multiple vision-language tasks. (3) We offer insights into the inherent trade-offs between the various loss variants presented in this paper.

The remainder of this paper is organized as follows. Section 2 provides background and discusses related research. Section 3 describes our methodological framework in detail, including the choice of alternative objectives and retraining procedures. Section 4 presents empirical evaluations, while Section 5 discusses implications, limitations, and avenues for future research.

2 Related work

2.1 Contrastive representation learning

Contrastive learning has proven foundational in self-supervised learning by structuring the embedding space through similarity based objectives. Frameworks like SimCLR [2], MoCo [4], and InfoNCE [10] have shown that such objectives can yield generalizable visual features by encouraging instance-level discrimination.

While these methods operate in unimodal domains, they inspire the loss structures used in multimodal pretraining regimes such as CLIP. However, their focus is typically on learning representations from scratch. In contrast, our work repurposes contrastive losses for post hoc refinement—retraining intermediate layers to re-shape the embedding geometry while preserving the structure induced by CLIP's pretraining.

2.2 Vision-Language models and CLIP

Multimodal models such as CLIP [11] and ALIGN [5] learn joint embedding spaces by contrasting aligned and misaligned image-text pairs. CLIP, in particular, has demonstrated strong zero-shot performance by training on a vast corpus of noisy internet image-caption pairs, enabling it to generalize without additional supervision.

Despite its scale and effectiveness, CLIP's learned space can be semantically coarse-grained, often collapsing subtle distinctions or allowing entangled concept clusters. Rather than retraining the model end-to-end, we explore whether selectively updating earlier projection layers with structure-aware losses can enhance semantic separability while preserving zero-shot generalization.

2.3 Embedding space refinement

A body of work has investigated how to improve the structure and interpretability of learned embeddings through constraints or post-processing. In deep metric learning, objectives are carefully designed to shape the intra and inter class structure of representations [12]. Other approaches apply regularization to enforce orthogonality or sparsity within learned features [14].

However, such work is typically applied in supervised or unimodal settings and often requires extensive retraining. In contrast, we propose a minimal intervention approach: freezing CLIP's final projection layer and refining only upstream layers using lightweight auxiliary losses. This allows us to isolate and evaluate the effect of loss-driven structure modification on the embedding space.

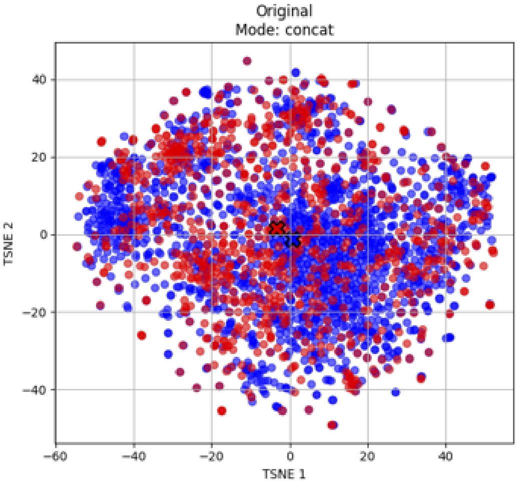

Figure 1: Original CLIP concatenated embedding space on t-SNE.

2.4 Optimal transport and sliced-Wasserstein distance

Optimal transport (OT) offers a principled way to compare and align distributions, with applications in generative modeling, domain adaptation, and structured data analysis. To address the computational burden of classical OT, sliced-Wasserstein distances [7] approximate the transport cost by averaging 1D projections, enabling efficient optimization in high dimensions.

Recent work has used these distances for set-based and distributional representation learning [9]. We extend this paradigm to multimodal representation refinement, using the sliced-Wasserstein distance not as a divergence metric between datasets, but as a training signal to reshape the joint embedding distribution toward greater semantic coherence.

2.5 Zero-Shot Generalization and Disentanglement

Disentangled representations offer interpretability and often enable more reliable generalization, particularly in zero-shot learning contexts. Methods like Semantics Disentangling for GZSL [3] and Disentangled Representation Learning for GZSL [8] propose architectures or supervision regimes to isolate semantic factors for transfer learning.

However, these typically involve designing new model architectures or modifying the training data pipeline. By contrast, we seek to retain CLIP's architecture and zero-shot interface, modifying only internal geometry through auxiliary loss terms. Our work contributes to understanding how such targeted refinements affect the delicate balance between semantic clarity and generalization performance.

3 Methods

3.1 Dataset

We evaluate our methods using the Hateful Memes dataset [6], a benchmark specifically designed to probe multimodal reasoning and nuanced semantic understanding in vision-language models. The dataset poses a particularly challenging task: detecting hateful content that is only discernible through the joint interpretation of both image and text modalities. Many examples are intentionally constructed to be ambiguous or benign in isolation, requiring the model to infer intent and tone through subtle cross-modal cues. This makes the dataset an ideal testbed for assessing the semantic disentanglement and alignment properties of multimodal embeddings. Its fine-grained distinctions also offer a robust setting to evaluate whether refinements to CLIP's embedding space enhance interpretability and discrimination without sacrificing generalization.

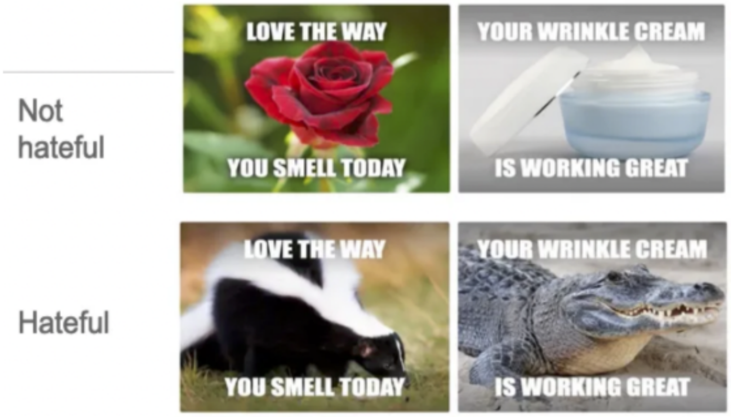

Figure 2: Samples from the dataset demonstrating how an image or text alone can be benign but when together can become hateful.

3.2 Architecture and Fine-Tuning Strategy

We base our model on the pretrained CLIP architecture [11], specifically using the `ViT-B/32` variant. CLIP comprises separate vision and text encoders whose outputs are projected into a shared multimodal embedding space via linear projection heads. To preserve the core semantic alignment capabilities learned during CLIP's large-scale pretraining, we freeze all encoder weights and the final layer of the model. Only the image and text projection layers (i.e., the penultimate linear projection layers) are marked trainable during our refinement phase for contrastive learning.

This selective fine-tuning strategy offers several critical advantages. First, it preserves the robust semantic foundations established during CLIP's extensive pretraining on internet-scale data, which would be prohibitively expensive to replicate. Second, by isolating updates to the projection layers, we maintain the high-level feature extraction capabilities of the encoders while gaining precise control over the geometric structure of the embedding space. This approach strikes an optimal balance between computational efficiency and representational refinement, allowing us to reshape embedding relationships without compromising CLIP's zero-shot generalization capabilities.

3.3 Training Pipeline

We employ the Hateful Memes dataset [6] as our evaluation benchmark, a choice motivated by its inherent multimodal reasoning challenges. During training, each batch consists of paired images and texts that are processed through frozen encoders and trainable projection layers to obtain normalized image and text embeddings. We concatenate these embeddings to form a joint representation matrix of shape $[B, 2D]$, where B is the batch size and D is the embedding dimensionality. This concatenation allows us to treat visual and textual representations as a unified semantic space during optimization.

Each batch is trained using a composite loss function:

$$\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_{\text{CLIP}} + \lambda_s \mathcal{L}_{\text{sep}}$$

where $\mathcal{L}_{\text{CLIP}}$ is the original CLIP contrastive loss, $\mathcal{L}_{\text{sep}}$ is a secondary separation-driven loss (either contrastive or distributional), and λ_c, λ_s are weighting coefficients. This dual-objective approach maintains cross-modal alignment while enforcing improved semantic separation.

3.4 Contrastive Objectives

We explore several loss objectives designed to enforce semantic separability. Let $u, v \in \mathbb{R}^{512}$ denote text and image embeddings, respectively. Furthermore, let $z := u +\!\!+\ v$ be defined as the concatenation of u and v. Furthermore, let S denote the similarity matrix, such that $S_{ij} = \frac{u^{(i)} v^{(j)}}{\|u\| \|v\|}$. Note that (i) indexes a vector i from a batch B, and $S \in [-1, 1]^{|B| \times |B|}$. Conversely, let $Z \in [-1, 1]^{|B| \times |B|}$ denote the similarity matrix, such that $Z_{ij} = \frac{z^{(i)} z^{(j)}}{\|z\| \|z\|}$. Finally, let $L^{(S)}$ and $L^{(Z)}$ be the corresponding label matrices for each similarity matrix.

CLIP-Style Contrastive Loss: We include the standard CLIP contrastive loss as our foundational objective. This loss maximizes cosine similarity between corresponding image-text pairs while minimizing it between mismatched pairs:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\sum_{i=1}^{B}\Big(H(\sigma(S_{i,:}), L^{(S)}_{i,:}) + H(\sigma(S_{:,i}), L^{(S)}_{:,i})\Big)$$

Here, H is cross-entropy and σ is the softmax function. Note that $L^{(S)}$ is a defined as a one-hot probability vector and is defined as:

$$L^{(S)}_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

In other words, the similarity between the text image pairs is maximized. Dissimilarity here is encouraged implicitly by the softmax function; the smaller the logit similarities between non-pairs, the closer the row or column probability vector gets to the one-hot vector. Retaining this loss ensures that we preserve CLIP's cross-modal alignment.

Weighted Supervised Contrastive Loss: Beyond pairwise matching, we use a weighted supervised contrastive objective that explicitly incorporates class-level distinctions across the combined image-text embedding set. Class labels are utilized to create a pull-push dynamic: embeddings with the same class label are pulled together, while those with different labels are pushed apart, determined by the mask matrix M. For clarity, we separate the attraction and repulsion loss terms in this formulation. We define two mask Matrices, $M^{(a)}$ and $M^{(r)}$ to denote the attraction and repulsion masks respectively.

$$M^{(a)}_{ij} = \begin{cases} 1 & h(z^{(i)}) = h(z^{(j)}) \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

$$M^{(r)}_{ij} = \begin{cases} 1 & h(z^{(i)}) \neq h(z^{(j)}) \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

where $h : \mathbb{R}^{1024} \to \mathcal{Y}$ is a function that maps an embedding z to its corresponding class $y \in \mathcal{Y}$. Finally, our weighted supervised contrastive loss is defined as:

$$\mathcal{L}_{\text{SC}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\frac{\|M^{(a)}_{i,:} \odot \log\sigma(Z_{i,:})\|_1}{\|M^{(a)}_{i,:}\|_0} - \gamma\frac{\|M^{(r)}_{i,:} \odot \log\sigma(Z_{i,:})\|_1}{\|M^{(r)}_{i,:}\|_0}\right)$$

where the $\odot$ operation is the Hadamard product and γ is a repulsive parameter for negative pairs. This supervision encourages more semantically coherent clusters. With $\gamma = 0$, supervised contrastive loss is equivalent to cross entropy with an embedding mask. We call this setting "contrastive original". We additionally experiment with values of $\gamma = 1$, which we call "contrastive scaled" and $\gamma = 5$, which we call "contrastive weighted". The hyperparameter γ essentially controls the importance of repulsion during training.

3.5 sliced-Wasserstein Separation Loss

In contrast to contrastive loss, which operates on discrete pairs, we propose a distribution-level separation mechanism using the sliced-Wasserstein Distance (SWD) [7]. For each class label, we treat the embeddings as empirical distributions and seek to maximize the pairwise distance between these respective distributions per sampled training batch:

$$\mathcal{L}_{\text{swd}} = -\frac{1}{K}\sum_{i,j}\text{SWD}(P_i, P_j)$$

where P_i and P_j are the distributions of embeddings from class i and j respectively, and K is the number of unique label pairs.

SWD approximates the optimal transport problem by projecting high-dimensional embeddings onto multiple random 1D subspaces and computing the 1D Wasserstein distance between the projected distributions. Formally, for L projections:

$$\text{SWD}(X, Y) = \frac{1}{L} \sum_{\ell=1}^{L} \|\text{sort}(Xp_\ell) - \text{sort}(Yp_\ell)\|_2^2$$

where p_ℓ are unit vectors sampled uniformly from the unit sphere.

This approach offers several theoretical advantages over traditional contrastive methods. By capturing global distributional differences, it encourages inter-class separation while maintaining intra-class coherence. Additionally, it is less sensitive to individual outliers and can better model the complex, multimodal nature of semantic concepts for nuanced classification tasks.

3.6 Evaluation Metrics During Training

To monitor the refinement process, we use two primary metrics:

Centroid Distance: To assess semantic separability at the class level, we compute the Euclidean distance between class centroids in the joint embedding space:

$$d_{\text{centroid}} = \|\mu_{\text{hateful}} - \mu_{\text{non-hateful}}\|_2$$

A higher centroid distance indicates clearer class-level separation and improved disentanglement. This metric is particularly meaningful in our binary classification context, where maximizing the separation between hateful and non-hateful content directly translates to improved decision boundaries.

Training Loss Trends: We track both $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{sep}}$ across epochs using Weights & Biases to ensure that improvements in separability do not come at the expense of cross-modal alignment. Consistent descent in training loss and validation loss across epochs provides evidence of effective optimization, while centroid distance offers insight into semantic structure evolution. By modifying only embedding projection layers and retaining the geometry, we preserve CLIP's zero-shot utility.

3.7 Implementation Details

We use the Adam optimizer with a learning rate of 5×10^{-4} and batch size of 32, chosen after preliminary experiments indicated this configuration provided stable convergence without overfitting. For contrastive losses, we set the temperature parameter $\tau = 0.07$, which controls the sharpness of the probability distribution and significantly impacts the quality of learned representation (too high and distinctions become muddled, too low and optimization becomes unstable). Each experiment proceeds for 100 epochs.

When using the sliced-Wasserstein loss, we apply 50 random projections per batch, striking a balance between computational efficiency and approximation accuracy. Empirical testing showed that fewer projections led to unstable training, while more projections yielded diminishing returns.

4 Experiments

4.1 Quantitative Evaluation via Centroid Distance

To assess the impact of each refinement method on semantic separability, we compute the average centroid distance between the hateful and non-hateful samples in the joint embedding space. A higher centroid distance indicates better separation between the two classes. As shown in Table 1, the original CLIP model exhibits minimal separation, suggesting considerable overlap in the embedding space. Interestingly, while some variants of contrastive loss improve the distance modestly, the sliced-Wasserstein objective yields the highest centroid distance by a significant margin. This result suggests that the geometric structure imposed by the Wasserstein-based refinement encourages stronger clustering of semantically distinct samples.

Loss Objectives	Centroid Distance
Original	0.0578
Original Contrastive	0.0054044
Scaled Contrastive	0.47258
Weighted Contrastive	0.37795
Supervised Contrastive	0.52027
sliced-Wasserstein	**0.6404**

Table 1: Final centroid distances between hateful and non-hateful sample clusters in the joint embedding space after training convergence, evaluated on the test set. Higher centroid distances indicate greater semantic separation between classes. Refer to Sections 3.4 (Contrastive Loss Variants) and 3.5 (sliced-Wasserstein Distance) for a detailed explanation of each refinement strategy.

4.2 Qualitative Visualization using t-SNE

To better understand how each loss function shapes the embedding space beyond a single scalar measure, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) [1], a widely used dimensionality reduction technique for visualizing high-dimensional data. t-SNE preserves local structure and is suitable for inspecting clusters and neighborhoods in learned representation spaces.

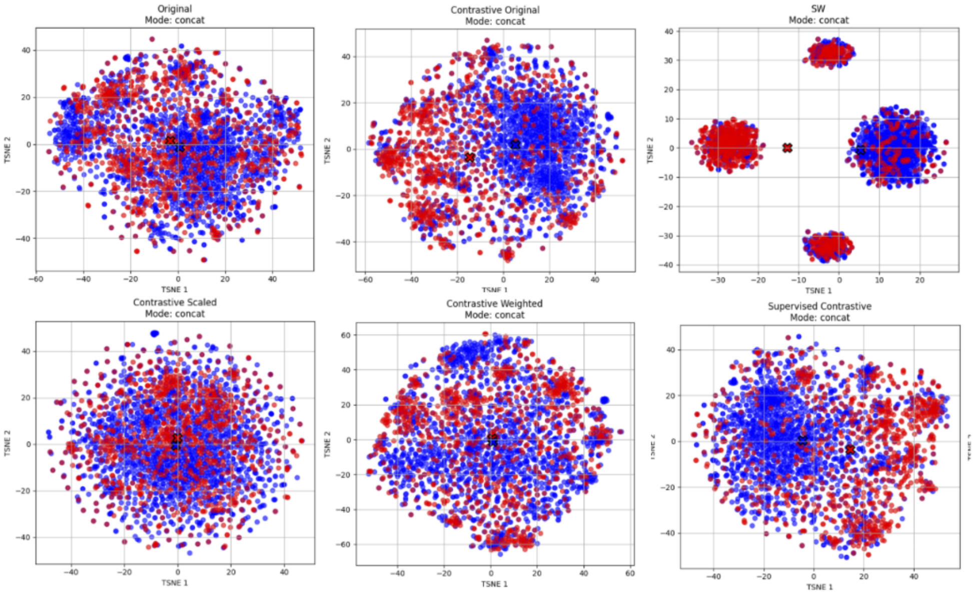

Figure 3: Concatenated embeddings visualized on t-SNE for No Change (upper left), Contrastive Scaled (upper middle), sliced-Wasserstein (upper right), Contrastive Original (bottom left), Constrastive Weighted (bottom middle), and Supervised Contrastive (bottom left).

Across most objectives, including the original and contrastive variants, t-SNE projections of the joint image-text embeddings reveal overlapping clusters, with limited class-level separation between hateful and non-hateful samples. However, the sliced-Wasserstein based refinement produces a notably different structure. As shown in Figure 4, the embeddings exhibit clearer organization, forming three prominent clusters corresponding to hateful content and a distinct fourth cluster for non-hateful samples. This emergent structure suggests a more meaningful semantic geometry.

Further analysis of the visual and textual embeddings in isolation reveals a stark contrast. When visualized separately, each modality forms only two broad groupings aligned with class labels, lacking the nuanced sub-cluster structure found in the joint embedding space. This divergence underscores the importance of cross-modal interaction in generating semantically rich representations and demonstrates how the sliced-Wasserstein loss encourages latent structures that align better with underlying intent.

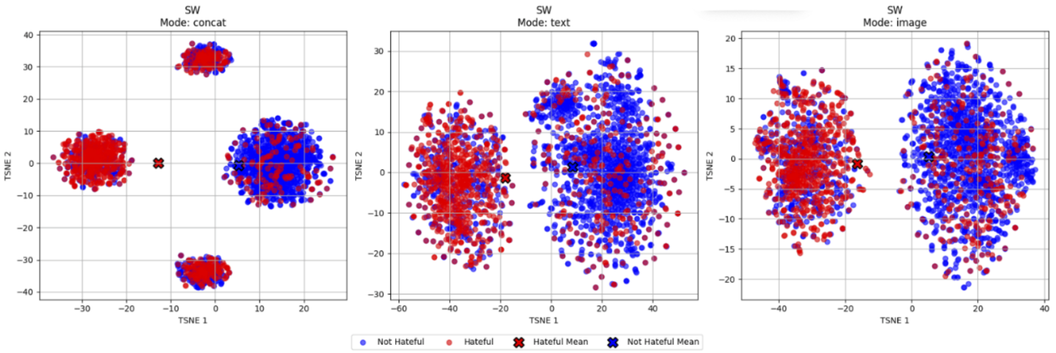

Figure 4: t-SNE visualization of SWCLIP embeddings after refinement with the sliced-Wasserstein loss. From left to right: concatenated image-text embeddings, text-only embeddings, and image-only embeddings. The concatenated space reveals well-separated clusters, including multiple subclusters within the hateful class, indicating improved semantic disentanglement.

5 Conclusion

Our experiments demonstrate that among the tested objective functions, maximizing the sliced-Wasserstein distance yields the most effective separation between hateful and non-hateful samples in the shared CLIP embedding space. While traditional and supervised contrastive losses offer modest improvements over the baseline, their embedding spaces still exhibit overlapping clusters with limited semantic disentanglement. In contrast, the sliced-Wasserstein objective produces distinct, interpretable clusters, often grouping hateful and non-hateful examples into separate, coherent regions. This induces finer substructure in the embedding space, particularly within the hateful class, suggesting a more nuanced capture of semantic variance.

This structural improvement is especially important in the context of multimodal data, where meaning often emerges from subtle interactions between modalities. The t-SNE visualizations support this observation, showing that text-only or image-only embeddings form broader and less informative clusters compared to the fused space. Given the prevalence of sarcasm, indirect language, and visual irony in datasets like Hateful Memes, a loss function that captures higher-order geometric relationships such as sliced-Wasserstein proves more capable of uncovering latent semantics.

For future work, we aim to explore additional loss functions that can further enhance embedding geometry, including those based on mutual information or adversarial contrastive training. Another promising direction is the integration of new data modalities such as audio or user intent metadata to better model context. Finally, moving beyond binary class labels to support multi-class or spectrum-based interpretations of hatefulness could open the door to more realistic, nuanced classification tasks and better alignment with real-world complexity.

References

[1] T. T. Cai and R. Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 2021. Accepted for publication.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[3] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8712–8720, 2021.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[6] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS 2020 Workshop on Datasets and Benchmarks*, 2020.

[7] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

[8] X. Li, Z. Xu, K. Wei, and C. Deng. Generalized zero-shot learning via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1966–1974, 2021.

[9] N. Naderializadeh, S. Kolouri, J. F. Comer, R. W. Andrews, and H. Hoffmann. Set representation learning with generalized sliced-wasserstein embeddings. *arXiv preprint arXiv:2103.03892*, 2021.

[10] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[12] J. D. Seidenschwarz, I. Elezi, and L. Leal-Taixé. Learning intra-batch connections for deep metric learning. In *International conference on machine learning*, pages 9410–9421. PMLR, 2021.

[13] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

[14] C. Wu, S. Zhang, F. Long, Z. Yin, and T. Leng. Towards better orthogonality regularization with disentangled norm in training deep cnns. *arXiv preprint arXiv:2306.09939*, 2023.

[15] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[16] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.