UNIFAST-HGR: SCALABLE AND EFFICIENT MAXIMAL CORRELATION FOR MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents an optimized approach to enhance the computation of Hirschfeld-Gebelein-Rényi (HGR) maximal correlation, addressing computational and efficiency challenges in large-scale neural networks and multimodal learning. The UniFast HGR framework introduces three key innovations: replacing covariance with cosine similarity to eliminate matrix inversion, removing the diagonal of the correlation matrix to mitigate self-correlation bias, and simplifying variance constraints via ℓ_2-normalization. These contributions reduce computational complexity from $O(K^3)$ to $O(m^2K)$ while improving accuracy and stability. The framework scales effectively across diverse multimodal applications. Additionally, the OptFast variant minimizes normalization steps, achieving efficiency comparable to dot-product operations without sacrificing precision. Experimental evaluations across benchmark datasets validate the framework's ability to balance computational efficiency with accuracy, establishing it as an effective solution for addressing contemporary deep learning challenges.

1 INTRODUCTION

In machine learning, extracting effective data representations is critical (Bengio et al., 2013). This task becomes increasingly complex when working with multimodal data, which encompasses information from diverse sources such as images, text, and audio (Summaira et al., 2021). Human cognition inherently integrates these disparate data types, facilitating more accurate interpretation and decision-making. However, machines encounter substantial difficulties in synthesizing such heterogeneous information, primarily due to the distinct statistical properties inherent in each modality. These differences obscure the correlations that are vital for learning effective feature representations (Baltrusaitis et al., 2018; Guo et al., 2019; Gandhi et al., 2023). Traditional methods, such as Canonical Correlation Analysis (CCA)(Hotelling, 1936) , have been employed to identify linear relationships between two datasets, while other approaches, such as minimizing Euclidean distances between feature spaces, have also been explored (Frome et al., 2013).

The Hirschfeld-Gebelein-Rényi (HGR) maximal correlation (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959) has been widely recognized as a robust metric for capturing nonlinear dependencies between random variables, generalizing CCA to nonlinear settings. Its application in machine learning, particularly for multimodal data integration, has garnered attention due to its theoretical ability to extract maximally informative features across modalities (Huang et al., 2017). Despite its potential, the practical implementation of HGR maximal correlation in modern machine learning frameworks presents significant challenges. The HGR maximal correlation framework imposes strict whitening constraints, necessitating uncorrelated feature representations. This requirement introduces substantial computational burdens, especially when processing high-dimensional data common in deep neural networks. Matrix inversion and decomposition operations, required for whitening, are computationally expensive and susceptible to numerical instability, thus limiting their scalability in large-scale machine learning applications. Efforts to overcome these limitations have led to the development of extensions such as Kernel CCA (Akaho, 2006) and Deep CCA (Andrew et al., 2013), which aim to approximate HGR maximal correlation. However, these methods remain constrained by their transformation functions and continue to suffer from computational inefficiencies stemming from whitening. Alternative approaches, such as Soft-CCA and Correlational Neural Networks, attempt to alleviate these constraints but risk altering the underlying feature geometry, which can reduce the discriminative capacity of the extracted features (Chang et al., 2018; Chandar et al.,

2016). Another limitation of the HGR framework is its lack of optimization for supervised learning tasks. It assumes discriminative information is preserved within the shared subspace of different modalities, which often fails, especially in weakly correlated modalities or substantial modality-specific information. Maximal Correlation Regression (MCR) addresses this by incorporating HGR maximal correlation to derive optimal weights for supervised learning, showing strong connections to methods like linear discriminant analysis and softmax regression. MCR has demonstrated competitive performance on various real-world datasets (Xu & Huang, 2020). Recent research has also explored sample complexity in estimating HGR maximal correlation functions using the Alternating Conditional Expectations (ACE) algorithm, providing error bounds and optimal sampling strategies for large datasets in supervised and semi-supervised learning contexts (Huang & Xu, 2021). In multimodal fusion, HGR maximal correlation has been successfully incorporated into loss functions to enhance person recognition performance across multimodal data sources (Liang et al., 2021).

The Soft-HGR framework (Wang et al., 2019) was introduced to address limitations by relaxing the whitening constraints while preserving the essential geometry of the feature space. This framework utilizes a low-rank approximation based on the empirical distribution of the dataset, accommodating missing modalities and incorporating supervised information. A deep learning framework has also been developed to address challenges in audio-visual emotion recognition, such as missing labels and incomplete modalities, by employing an HGR maximal correlation-based loss function to capture essential information from diverse training data (Ma et al., 2021). Additionally, a multimodal conditional GAN has been introduced as a data augmentation method for audio-visual emotion recognition, although this approach modifies the transmitted data during the fusion process (Ma et al., 2022). In the MultiEMO (Shi & Huang, 2023) study, Soft-HGR was applied to correlation analysis, leading to enhanced classification accuracy in emotion recognition. Despite these advancements, Soft-HGR still struggles when applied to complex neural architectures and large-scale datasets. Its scalability and efficiency, while improved, remain inadequate for modern deep learning applications due to high computational complexity and sensitivity to high-dimensional features. With the rise of large-scale models and datasets, the limitations of the Soft-HGR framework have become more evident. Despite its utility in some applications, Soft-HGR's computational complexity and inefficiency hinder its integration into state-of-the-art deep learning architectures, particularly for large-scale data and models. A more efficient and scalable solution is urgently needed to unlock the potential of multimodal learning in modern machine learning environments.

To address these limitations, UniFast HGR is introduced as an advanced solution that overcomes computational bottlenecks and scalability issues. It replaces covariance-based computations with cosine similarity to eliminate matrix inversion, reduces complexity from $O(K^3)$ to $O(m^2K)$, and incorporates diagonal removal to mitigate self-correlation bias in high dimensions. The framework also enforces variance constraints via ℓ_2-normalization to stabilize training. These three innovations—cosine similarity, diagonal removal, and simplified variance constraints—collectively enable efficient and scalable maximal correlation estimation in deep learning. UniFast HGR features an optimized algorithmic structure that reduces computational overhead, improves discriminative accuracy, and provides a unified approach scalable to large datasets and deep models. It is also designed to fully leverage deep neural networks, enabling efficient learning of correlated features across multiple modalities. The contributions of this framework are as follows:

Unified Efficiency and Scalability: UniFast HGR merges the strengths of both traditional HGR and Soft-HGR, addressing their limitations in dimensionality and computational complexity. By integrating the original HGR maximal correlation framework with refinements from Soft-HGR, UniFast HGR achieves stable and precise feature extraction within a bounded range of [-1,1]. This integration enhances adaptability and performance within modern deep learning architectures, making the framework particularly well-suited for large-scale datasets and deep neural networks.

Enhanced Discriminative and Correlation Power: UniFast HGR integrates discriminative objectives to extract informative features for supervised tasks and optimizes correlations between data modalities through function maximization, ensuring effective information alignment. This is crucial in complex architectures where multimodal correlations affect performance. It replaces covariance with cosine similarity to improve accuracy and excludes diagonal elements to avoid self-correlation bias, yielding more precise results. The framework balances speed and performance, enhancing discriminative and correlation efficacy across deep learning tasks.

Overcoming Complexity Limitations: UniFast HGR resolves the complexity and inefficiency issues associated with Soft-HGR, providing a faster, more scalable solution for large-scale deep learning applications. Additionally, the OptFast HGR variant further optimizes performance by reducing the number of normalization steps, achieving computational efficiency comparable to dot product and cosine similarity operations. This optimization significantly accelerates processing while maintaining high performance. These advancements represent a substantial step forward in applying HGR maximal correlation, particularly in managing dimensionality challenges and enabling more effective multimodal learning at scale.

2 PROPOSED METHOD

The UniFast HGR framework significantly improves on both Soft-HGR and the original HGR maximal correlation approaches. It addresses computational challenges, scalability limitations, and practical constraints in large-scale neural network applications. UniFast HGR enhances both discriminative and correlation capabilities, facilitating the extraction of highly informative features across diverse data modalities. The following sections outline its key components and innovations.

2.1 PRELIMINARY

HGR Correlation Analysis and Limitations: HGR maximal correlation extends Pearson correlation by providing a more comprehensive measure of dependency, originally developed for single features but naturally extendable to multiple features. In the case of random variables x and y, which share a joint distribution across the domains X and Y. Given $N \times k$ feature matrices $f = [f_1, f_2, \cdots, f_N]^T$ and $g = [g_1, g_2, \cdots, g_N]^T$,where f_i and g_i are both $1 \times k$ dimensional vectors, N is the number of samples, and the HGR maximum correlation is defined as follows:

$$\rho^k(X,Y) = \underset{\substack{f:x\to R^k,\ \mathbb{E}[f]=0,\mathrm{cov}(f)=I \\ g:y\to R^k,\mathbb{E}[g]=0,\mathrm{cov}(g)=I}}{sup} \mathbb{E}\left[f^T(X)\, g(Y)\right] \tag{1}$$

where $\mathbb{E}[f]$ and $\mathbb{E}[g]$ represent the expected value of vectors f and g, respectively; $\mathrm{cov}(f)$ and $\mathrm{cov}(g)$ represent the covariance of vectors f and g, respectively.

The HGR maximal correlation is determined via optimization over sets of Borel measurable functions with zero mean and stable covariance. This correlation, ranging from 0 to 1, signifies either complete independence or a deterministic relationship between X and Y. However, the computational complexity of HGR maximal correlation arises primarily from the whitening constraints, which necessitate matrix inversion and decomposition, resulting in a time complexity of $O(K^3)$. These challenges are compounded by scalability issues, particularly as covariance matrices can become ill-conditioned, leading to gradient explosions in high-dimensional spaces.

Soft-HGR tackles some computational challenges of HGR by using a low-rank approximation, which helps integrate with neural networks and compute maximal correlations efficiently without requiring strict whitening (Wang et al., 2019). When applied to mini-batches, Soft-HGR lowers complexity to $O(mK^2)$ by approximating batch covariance, enhancing stability even with large feature dimensions. However, it faces issues during the fusion process where data values may be altered, and output values can become excessively large due to higher network outputs associated with higher HGR correlations (Zhang et al., 2024). This variance sensitivity and deviation from ideal HGR make cross-dataset comparisons difficult, especially with numerous features, limiting its practical use. Although low-rank approximations reduce some computational load from traditional HGR, Soft-HGR still involves more complex operations than simpler alternatives like the dot product. Additional operations such as covariance matrix computation, matrix decomposition or inversion, and iterative feature mapping optimization further increase computational complexity.

These limitations lead to higher computational costs when applying Soft-HGR to large-scale datasets and deep models, complicating its scalability and impeding its efficiency and stability in real-world applications. Consequently, Soft-HGR is less suited for widespread deployment in large-scale deep learning environments. Soft-HGR is mathematically represented as follows:

$$\max_{f,g} \mathbb{E}\left[f^T(X)g(Y)\right] - \tfrac{1}{2}\, tr(\,\mathrm{cov}(f(X))\,\mathrm{cov}(g(Y))), \text{s.t. } \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0 \tag{2}$$

where $f(X)$ and $g(Y)$ are feature mappings derived from various random modalities.

2.2 Optimized Correlation Framework

Variance Constraint: To address the limitations of Soft-HGR, especially its sensitivity to changes in signal variance, variance constraints are introduced in UniFast HGR. Unlike Soft-HGR, which lacks variance normalization, UniFast HGR enforces variance constraints during optimization. By definition of HGR maximal correlation, a zero mean and unit variance (Var = 1) are required in the Soft-HGR objective, as shown in Eq. (2). For the first term of Eq. (2), the following holds:

$$\mathbb{E}\left[f^T(X)\,g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N} f^T(x_i)g(y_i) \tag{3}$$

By ensuring a mean of zero, the following condition is satisfied:

$$\mathbb{E}\left[f^T(X)\,g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N}\left(f(x_i)-\mathbb{E}[f(x_i)]\right)^T\left(g(y_i)-\mathbb{E}[g(y_i)]\right) \tag{4}$$

By introducing the variance constraint Var = 1, the following expression is obtained:

$$E\left[f^T(X)g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(f(x_i)-\mathbb{E}[f(x_i)])(g(y_i)-\mathbb{E}[g(y_i)])}{\sqrt{\mathrm{Var}[f(x_i)]}\sqrt{\mathrm{Var}[g(y_i)]}} \tag{5}$$

This variance normalization ensures that the output values of Soft-HGR remain within the range [-1,1]. A key aspect of this method is that as Soft-HGR output values approach 1, the corresponding HGR values also approach 1, due to the synchronous nature of their derivatives (i.e., both rates of change share the same sign). This correlation allows the use of an HGR approximation under ideal conditions to replace the actual HGR value, improving accuracy while slightly increasing computational complexity. However, by transforming the first term of Eq. (2) into a cosine similarity calculation, the computational burden is reduced.

Expansion of the Trace Term: The introduction of variance constraints in the Soft-HGR objective increases computational load. However, by expanding the trace term, this additional burden can be mitigated, optimizing the process. The trace term, which plays a critical role in the framework, was not significantly impacted in the original Soft-HGR due to the absence of variance constraints. However, with variance constraints in place, the trace term becomes essential, as it represents the correlation between two matrices or data sets. In refining the Soft-HGR framework, two key components were identified: (1) the correlation between individual elements, and (2) the correlation between the correlation matrices of sets. Specifically, for a matrix representing the correlation of elements within a set, the trace term captures the correlation between the correlation matrices of these sets. This is achieved by expanding the matrix and quantifying the similarity in the distribution of elements. In essence, the trace term provides a more refined measure of the correlation between the sets by capturing the correlation between their respective correlation matrices. The definition of the trace term is given as follows:

$$trace = \frac{1}{2}\,tr(\,\mathrm{cov}(f(X))\mathrm{cov}(g(Y))) \tag{6}$$

The covariance matrices are computed as follows:

$$\mathrm{cov}[f(X)] = \frac{1}{N-1}\sum_{i=1}^{N}\left(f(x_i)-\mathbb{E}[f(x_i)]\right)\left(f(x_i)-\mathbb{E}[f(x_i)]\right)^T \tag{7}$$

$$\mathrm{cov}[g(Y)] = \frac{1}{N-1}\sum_{i=1}^{N}\left(g(y_i)-\mathbb{E}[g(y_i)]\right)\left(g(y_i)-\mathbb{E}[g(y_i)]\right)^T \tag{8}$$

where, $\mathrm{cov}[f(X)]_{ij} = \mathrm{cov}[f_i, f_j] \equiv \mathrm{cov}f_{ij}$, $\mathrm{cov}[g(Y)]_{ij} = \mathrm{cov}[g_i, g_j] \equiv \mathrm{cov}g_{ij}$

Considering the trace term,

$$trace = \tfrac{1}{2}\,tr(\,\mathrm{cov}(f(X))\mathrm{cov}(g(Y))) = \frac{1}{2(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathrm{cov}f_{ij}-\mathbb{E}[\mathrm{cov}f_i]\right)\left(\mathrm{cov}g_{ji}-\mathbb{E}[\mathrm{cov}g_j]\right) \tag{9}$$

where $\mathrm{cov}f_i = (\mathrm{cov}f_{i,0}, \mathrm{cov}f_{i,1}, \cdots, \mathrm{cov}f_{i,N}), \mathrm{cov}g_j = (\mathrm{cov}g_{j,0}, \mathrm{cov}g_{j,1}, \cdots, \mathrm{cov}g_{j,N})$

By incorporating the variance constraint Var = 1,

$$trace = \tfrac{1}{2}\,tr\left(\,\mathrm{cov}(f(X))\,\mathrm{cov}(g(Y))\right) = \frac{1}{2(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(\,\mathrm{cov}f_{ij}-E[\,\mathrm{cov}f_i])(\,\mathrm{cov}g_{ji}-E[\,\mathrm{cov}g_j])}{\sqrt{Var(\,\mathrm{cov}f_i)}\sqrt{Var(\,\mathrm{cov}g_j)}} \tag{10}$$

Simplifying this expression demonstrates that it is related to the trace of the product of the covariance matrices in the simplified HGR approximation formula. This optimization reduces computational complexity while maintaining the accuracy of the HGR approximation.

2.3 UNIFAST HGR

The UniFast HGR framework is derived from Soft-HGR through three key steps: (1) enforcing $\text{Var}(f) = \text{Var}(g) = 1$ to ensure stability and theoretical consistency with HGR, (2) replacing covariance computations with cosine similarity under these constraints, and (3) expanding and simplifying the trace term. This reformulation reduces computational complexity from $O(K^3)$ to $O(m^2K)$ while maintaining correlation estimation accuracy.

Substitution with Cosine Similarity: Covariance computations are replaced with cosine similarity, eliminating matrix inversion. The substitution is mathematically justified when zero-mean features satisfy unit variance constraints, where covariance naturally simplifies to cosine similarity. This transformation enables efficient, scalable correlation estimation for high-dimensional features.

$$\cos(f, g) = \frac{f \cdot g}{\|f\|\|g\|} \tag{11}$$

If all components of a random vector are independent, the square of the vector's modulus equals the sum of the variances of each component. Thus, Eq.(5) and (11) are equivalent:

$$\mathbb{E}\left[f^T(X)\, g(Y)\right] = \frac{1}{N-1} \sum_{i=1}^{N} \cos(f(x_i), g(y_i)) \tag{12}$$

Similarly, the covariance calculation in Eq. (10) can be converted into a cosine similarity calculation:

$$trace = \tfrac{1}{2}\, tr\left(\text{ cov}(f(X))\, \text{cov}(g(Y))\right) = \tfrac{1}{2(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \tfrac{(\cos f_{ij} - \mathbb{E}[\cos f_i])(\cos g_{ji} - \mathbb{E}[\cos g_j])}{\sqrt{Var(\cos f_i)}\sqrt{Var(\cos g_j)}} \tag{13}$$

where $\cos f_{ij} = \cos(f_i, f_j), \cos g_{ji} = \cos(g_j, g_i), \cos f_i = (\cos f_{i,0}, \cos f_{i,1}, \cdots, \cos f_{i,N}), \cos g_j = (\cos g_{j,0}, \cos g_{j,1}, \cdots, \cos g_{j,N})$. That is,

$$trace = \frac{1}{2}\, tr\left(\text{ cov}(f(X))\, \text{cov}(g(Y))\right) = \frac{1}{2(N-1)} \sum_{i=1}^{N} \cos(\text{distri}_f, \text{distri}_g) \tag{14}$$

where $\text{distri}_f = f \cdot f^T$ and $\text{distri}_g = g \cdot g^T$

Finally, the UniFast-HGR is computed as follows:

$$\max_{f,g} \mathbb{E}\left[f^T(X)\, g(Y)\right] - \tfrac{1}{2}\, tr\left(\text{ cov}(f(X))\, \text{cov}(g(Y))\right) = \tfrac{1}{N-1} \sum_{i=1}^{N} cos(f(x_i), g(y_i) - \tfrac{1}{2(N-1)} \sum_{i=1}^{N} \cos(\text{distri}_f, \text{distri}_g)) \tag{15}$$

That is,

$$\text{UF-HGR} = \frac{1}{N-1} \sum_{i=1}^{N} \cos(f(x_i), g(y_i)) - \frac{1}{2(N-1)} \sum_{i=1}^{N} \cos(\text{distri}_f, \text{distri}_g) \tag{16}$$

where distri_f and distri_g represent the distribution vectors derived from the correlation matrices, capturing inter-sample relationships.

Removing the Main Diagonal: A key enhancement in UniFast HGR involves removing the main diagonal of the correlation matrices. The diagonal entries, fixed at 1 due to the variance constraint (Var = 1), represent self-correlations that disproportionately influence cosine similarity calculations, leading to overestimated similarity and biased optimization. Removing the diagonal mitigates this issue, as the fixed diagonal value of 1 biases the resulting vector toward a specific angle, narrowing the range of variation and reducing accuracy. Furthermore, correlation values in the range $[-1, 1]$ can be distorted by the multiplication effect, which amplifies the diagonal's influence and diminishes the contribution of non-diagonal elements. This distortion causes the calculated angles to align with the maximum diagonal value, limiting the ability of other values to approach 1 and often pushing them significantly below 1. By removing the diagonal, a more accurate and representative similarity measure is achieved. This enhancement significantly improves gradient stability—especially in small-batch settings—and final accuracy, making UniFast HGR both faster and more robust. The approach aligns with practices in methods like CKA and PCA whitening, and closely matches the theoretical expectations of HGR. A detailed step-by-step derivation is provided in **Appendix A**, along with the detailed calculation process for the UniFast HGR algorithm in Algorithm 1.

2.4 Generalization to More Modalities

The HGR maximum correlation was originally defined for two random variables, and extending this correlation-based approach to multiple modalities presents significant challenges. The introduction of additional modalities imposes new whitening constraints, thereby increasing computational complexity. However, UniFast HGR offers enhanced flexibility in managing this complexity. To handle two or more modalities, the multimodal UniFast HGR must be capable of learning and simultaneously recording all paired feature transformations. Assuming that $X_1, X_2, \ldots, X_m$ are m different modalities, and $f_{(1)}, f_{(2)}, \ldots, f_{(m)}$ denote their corresponding transformation functions. The multimodal UniFast HGR is defined as follows:

$$\text{UF-HGR} = \frac{1}{N-1} \sum_{j=1,l=j+1}^{m} \sum_{i=1}^{N} \cos\left(f^{(j)}\left(x_j\right), f^{(l)}\left(x_l\right)\right) - \frac{1}{2(N-1)} \sum_{j=1,l=j+1}^{m} \sum_{i=1}^{N} \cos(\text{distri}_f^j, \text{distri}_f^l) \tag{17}$$

where $j, l = 1, 2, \ldots, m$, and $j \neq l$. The model extracts features from each modality branch and maximizes their paired UniFast HGR values in an additive manner. From an information theory perspective, as shown in Eq. (17), maximizing UniFast HGR is equivalent to extracting the shared information between multiple random variables. This process identifies and leverages the common information content between different patterns or random variables involved.

2.5 Optimization in Speed

To further accelerate the algorithm's computational speed, OptFast HGR was developed as an extension of UniFast HGR, prioritizing efficiency while maintaining reasonable accuracy. The primary improvement in OptFast HGR involves reducing the number of normalization steps, achieving efficiency and computational cost comparable to a dot product operation. This optimization significantly increases computation speed. However, the trade-off for this enhancement is a slight bias introduced in the results. This bias results in correlation values that are marginally shifted due to the reduced normalization steps, highlighting a trade-off between speed and accuracy. While the dot product operation in OptFast HGR provides faster computations, it slightly compromises the precision of the correlation values. This difference underscores that OptFast HGR, while optimized for speed, may not always align perfectly with the theoretical correlations expected in certain contexts (Please refer to **Appendix B**). Nonetheless, the strength of OptFast HGR lies in its ability to process large datasets and models at a significantly faster rate, making it especially suitable for scenarios where computational speed takes precedence over minor variations in accuracy. The computational process of the proposed OptFast HGR algorithm is detailed in Algorithm 2 in **Appendix A**.

3 Experiments

3.1 Execution Time and Feature Dimension

The execution times and maximum achievable feature dimensions of various methods, including HGR, Soft-HGR, and UniFast HGR, were compared using the MNIST dataset (LeCun et al., 1998). Following the experimental frameworks of Wang et al.(Wang et al., 2019) and Andrew et al.(Andrew et al., 2013), the left and right halves of each digit image were treated as two distinct patterns. To highlight the efficiency differences introduced by the UniFast HGR, all feature transformations were constrained to a linear form, reducing the maximum correlation of HGR to linear CCA. As depicted in Figure 1, the execution times for UniFast HGR and OptFast HGR were significantly faster than those of CCA and Deep CCA methods, and also outperformed Soft-HGR. The execution time for the CCA method increased substantially as feature dimensions grew, posing challenges in real-world applications where feature dimensions are typically large. Notably, when the feature dimension exceeded 350, CCA encountered numerical stability issues.

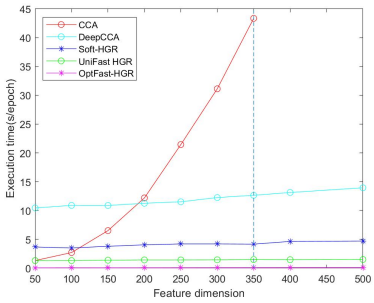


Figure 1: Execution time and feature dimension comparison on MNIST dataset.

3.2 IMAGE CLASSIFICATION

The performance of UniFast HGR was evaluated against several methods, including CCA, Deep CCA, Soft CCA, Soft-HGR, cosine similarity, and dot product, in the context of image classification. Comparative experiments were conducted using a dual-channel deep learning framework for remote sensing data classification, with ResNet 50 (He et al., 2016) as the backbone. Following the same conditions and preprocessing steps outlined by Wu et al. (Wu et al., 2022) , classification results on the Berlin dataset (Hong et al., 2021; Akpona et al., 2016) are presented in Table 1. The performance was evaluated using three metrics: overall accuracy (OA), average accuracy (AA), and kappa coefficient. On the Berlin and Houston 2018 (Lin et al., 2023) datasets, our method improved classification accuracy. UniFast HGR showed competitive performance across all methods, proving its effectiveness in image classification. Moreover, OptFast HGR, which reduced the number of normalization steps, achieved computational efficiency on par with dot product and cosine similarity operations. These results highlight the significant advantages of UniFast HGR and OptFast HGR in enhancing classification performance.The detailed results can be found in **Appendix D.1 and D.4**.

Table 1: Image classification results on the Berlin dataset.

Methods	OA(%)	AA(%)	Kappa (%)	Time (s/epoch)
CCA	70.93	64.35	58.28	2967.52
Deep CCA	72.74	65.08	60.23	250.51
Soft CCA	71.54	61.14	58.33	314.93
Dot Product	75.20	66.22	62.77	**23.18**
Cosine Similarity	75.51	65.53	62.53	23.40
Soft-HGR	65.80	64.30	52.99	25.83
UniFast HGR	**80.75**	**71.53**	**70.44**	24.53
OptFast-HGR	80.46	71.51	70.21	23.54

Table 2: Experimental results of remote sensing segmentation.

Methods	Vaihingen		Globe230k	
	OA(%)	mIoU(%)	OA(%)	mIoU(%)
CCA	91.15	79.37	87.92	67.49
Deep CCA	91.39	81.35	88.27	67.85
Soft CCA	91.41	81.44	87.60	66.71
Dot Product	92.61	83.65	90.92	75.67
Cosine Similarity	92.56	83.34	90.81	75.53
Soft-HGR	90.10	76.87	86.46	64.82
UniFast HGR	**93.01**	**84.62**	**91.48**	**76.36**
OptFast HGR	92.95	84.57	91.23	76.15

3.3 REMOTE SENSING SEMANTIC SEGMENTATION

To further evaluate UniFast HGR and OptFast HGR, we conducted remote sensing semantic segmentation experiments on the Vaihingen dataset and the large-scale high-resolution annotation dataset Globe230k, comparing them with other methods. The ISPRS Vaihingen dataset, provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) (Wang et al., 2022), is a 2D semantic segmentation dataset with a 9-cm spatial resolution. It includes 8-bit TIFF files for near-infrared, red, and green bands, as well as a single-band digital surface model (DSM) with 32-bit floating-point height values. The Globe230k dataset contains 232,819 annotated images of size 512×512 and 1-m spatial resolution, with multiple bands such as RGB and digital elevation models (DEM) (Shi et al., 2023). Using the model and preprocessing steps outlined by Ma et al. (Ma et al., 2024), we applied UniFast HGR and OptFast HGR to fuse multimodal remote sensing data. The results are shown in Table 2, evaluated using overall accuracy (OA) and mean intersection over union (mIoU). UniFast HGR and OptFast HGR both demonstrated strong performance, effectively capturing correlations between modalities and achieving accurate semantic segmentation. Detailed experimental results and visualization examples can be found in **Appendix D.2**.

3.4 EXTENSION TO MORE/MISSING MODALITIES: MULTIMODAL EMOTION RECOGNITION

The performance of UniFast HGR and OptFast HGR was evaluated in multimodal emotion recognition on the IEMOCAP dataset. Comparative experiments were conducted using the MultiEMO model proposed by Shi & Huang (Shi & Huang, 2023). Results from these emotion recognition tests on IEMOCAP (Busso et al., 2008) are shown in Table 3, with performance measured via weighted average F1 score (W-F1) and accuracy (ACC). Both models demonstrated strong performance, effectively capturing cross-modal correlations in emotion recognition scenarios. UniFast HGR was tested under two challenging scenarios: single-modality absence and insufficient labels. In the first scenario, one of the three modalities was randomly excluded. In the second, only 80%, 50%, or 20% of training labels were retained by hiding 20%, 50%, or 80% of the data, respectively. The architecture of UniFast HGR naturally addresses missing modalities: its normalization and covariance masking mechanisms reduce sensitivity to outliers, as shown in Table 3.

Table 3: Multimodal emotion recognition results on IEMOCAP(ACC %).

Methods	No Missing	Missing Modalities			Missing Labels		
	Text+Audio+Visual	Text+Audio	Text+Visual	Audio+Visual	20%	50%	80%
CCA	67.41	64.55	64.03	50.71	66.21	61.63	51.91
Deep CCA	67.78	64.92	64.38	51.06	66.50	63.10	54.80
Soft CCA	68.58	65.68	65.27	51.89	67.35	63.81	55.43
Dot Product	70.14	67.32	67.08	53.56	69.06	65.27	57.92
Cosine Similarity	69.50	66.64	66.21	52.92	68.43	64.94	57.63
Soft-HGR	71.29	67.85	67.52	53.90	69.47	65.19	57.75
UniFast HGR	**73.66**	**70.94**	**70.41**	**57.82**	**72.65**	**69.26**	**62.05**
OptFast HGR	73.43	70.67	70.15	56.57	72.39	68.92	61.58

3.5 Large-Scale Multimodal Learning

To validate scalability and generalizability, experiments were conducted on ImageNet-1K (Deng et al., 2009) classification, COCO (Lin et al., 2014) cross-modal retrieval, and the large-scale InternVid (Wang et al., 2023) benchmark. UniFast HGR was integrated with state-of-the-art vision encoders including CLIP (ViT-B/32) (Radford et al., 2021), SigLIP (Zhai et al., 2023), and DINOv2 (ViT-L/14) (Zhang et al., 2022; Oquab et al., 2024), and compared to nonlinear correlation methods such as CKA (Kornblith et al., 2019), dCor (Zhen et al., 2022), and I_dCor (Basile et al., 2025).

ImageNet Classification: As shown in Table 4, UniFast HGR consistently enhances baseline models across architectures. Notably, when applied to DINOv2, it achieves 85.3% Top-1 accuracy—a 3.5% absolute improvement over baseline. The performance gain demonstrates the method's ability to capture subtle feature correlations even in high-dimensional spaces. **Cross-Modal Retrieval**: On COCO text-image retrieval, CLIP with UniFast HGR achieves 42.1% Recall@1, surpassing both baseline CLIP (38.9%) and Soft-HGR (40.3%). OptFast HGR maintains competitive performance (42.0% R@1) with faster computation than standard HGR, validating its efficiency-accuracy trade-off. **Large-Scale Video-Text Retrieval**: To further evaluate scalability, tests were performed on InternVid-10M using ViCLIP (Wang et al., 2023). As Table 4 shows, UniFast HGR achieved the highest text-to-video retrieval recall across MSR-VTT (Xu et al., 2016), LSMDC (Yao et al., 2015), and DiDeMo (Hendricks et al., 2017), with an average gain of 5.8% over the ViCLIP baseline, demonstrating strong generalization to billion-scale multimodal data.

Robustness: UniFast HGR also demonstrates superior robustness against noise, modality imbalance, and spurious correlations (see Appendix for details).

Table 4: Performance on large-scale datasets

Dataset	Model	Baseline	CKA	dCor	I_dCor	Soft-HGR	UniFast HGR	OptFast HGR
ImageNet-1k Top-1 Accuracy (%)	ViT-B/32	76.6	76.7	76.9	78.7	76.3	**80.1**	79.6
	ResNet50	74.3	74.5	75.0	77.4	74.1	**78.5**	78.1
	CLIP	76.1	76.6	77.3	79.5	76.3	**80.4**	79.8
	SigLIP	81.3	81.7	82.2	84.1	81.4	**84.8**	84.5
	DINOv2	81.8	82.1	82.4	84.7	81.6	**85.3**	84.9
COCO Text-Image Retrieval Recall@1	ViT-B/32	38.2	38.7	39.2	39.6	38.9	**40.1**	39.8
	ResNet50	37.8	38.3	38.7	39.2	38.6	**39.5**	39.3
	CLIP	38.9	39.5	41.4	41.7	40.3	**42.1**	42.0
	SigLIP	50.8	51.3	52.8	53.2	51.6	**53.8**	53.5
	DINOv2	51.1	51.5	52.7	53.5	52.1	**53.9**	53.7
InternVid(T2V R@1)								
MSR-VTT	ViCLIP	36.4	37.1	37.9	38.5	38.8	**43.3**	42.7
LSMDC	ViCLIP	17.1	17.6	18.1	18.9	18.3	**20.7**	20.3
DiDeMo	ViCLIP	16.4	16.9	17.3	17.8	17.6	**20.5**	20.1

3.6 Correlation Estimation

To quantify intrinsic alignment capability, we measured cross-model feature correlations on ImageNet embeddings using six representative encoders. This directly evaluates how well different methods capture inter-modal relationships, unlike task-driven evaluation. We computed pairwise correlation matrices for EfficientNet, ResNet50, ViT-B/32, CLIP, SigLIP, and DINOv2 embeddings across 30K random samples. Figure 2 shows that: (1) UniFast HGR consistently achieves higher intra-model correlations, indicating better feature stability; (2) Cross-model correlations between CLIP and DINOv2 reach 0.91 with UniFast HGR, significantly outperforming dCor (0.78) and Soft-

HGR (0.82). UniFast HGR improves upon Soft-HGR by 12–18% absolutely across all model pairs, proving the effectiveness of our optimization. The high correlations (≥ 0.92 for ViT-based models) match downstream performance gains, confirming our method preserves key feature relationships.

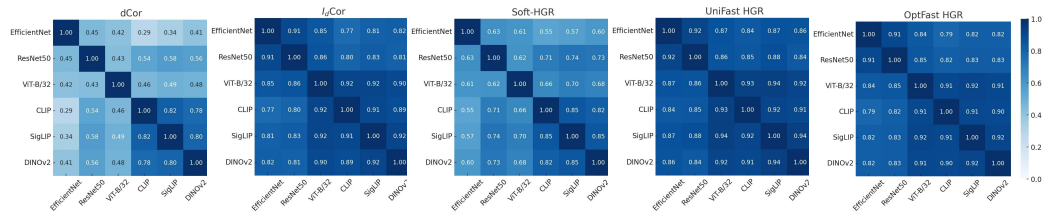


Figure 2: Correlation results on ImageNet representations (Please refer to **Appendix D.5**)

3.7 Computational Efficiency

To isolate computational costs from network architecture effects, we benchmarked correlation calculation between randomly generated tensors. We compared UniFast HGR/OptFast HGR against baseline methods across varying dimensions and batch sizes (16-256). For each configuration, we generated paired tensors $f, g \in \mathbb{R}^{bz \times dim}$ and measured average execution time over 10,000 trials. The experimental results are shown in Figure 4 in the **Appendix F**. The results demonstrate that, as the batch size increases, the execution time gradually grows. UniFast HGR and OptFast HGR consistently exhibit the best performance across different batch sizes, showing significant advantages in computational efficiency. Moreover, the results indicate that UniFast HGR and OptFast HGR exhibit lower execution times across most dimensions, with their efficiency advantages being particularly pronounced at higher dimensions. These findings suggest that UniFast HGR and OptFast HGR can not only effectively capture complex correlations between multimodal data but also offer high computational efficiency, making them well-suited for multimodal data fusion tasks.

4 Limitations and Future Work

UniFast HGR and OptFast HGR achieve notable efficiency and scalability, yet several limitations merit further investigation. First, while variance constraints enhance stability, they may over-regularize features in very low-dimensional spaces, potentially restricting representational flexibility in such cases. Second, OptFast HGR exhibits increased bias when handling cross-modal pairs with significant distributional asymmetry, affecting correlation estimation. Third, the theoretical justification for diagonal removal, though empirically validated across diverse datasets, could be further generalized to non-Gaussian and highly nonlinear dependency structures.

To address these issues, future work will: Develop adaptive regularization strategies that adjust constraints based on intrinsic dimensionality, preserving expressivity in low-dimensional settings. Integrate distribution-aware mechanisms, such as attention-based calibration, into OptFast HGR to better handle asymmetric modality distributions. Expand the theoretical foundation of diagonal exclusion to encompass broader dependency types, including non-Gaussian and heavy-tailed distributions, strengthening generality and rigor.

5 Conclusion

This paper introduced UniFast HGR, a efficient and scalable framework for estimating Hirschfeld-Gebelein-Rényi (HGR) maximal correlation. By incorporating variance constraints via ℓ_2-normalization and removing uninformative diagonal entries from the correlation matrix, the method achieves enhanced numerical stability, reduced computational complexity, and stronger discriminative performance. The OptFast HGR variant further improves efficiency with minimal accuracy loss, offering a practical trade-off for large-scale applications. Comprehensive evaluations—covering image classification, cross-modal retrieval, remote sensing segmentation, and emotion recognition—demonstrate that UniFast HGR consistently outperforms existing correlation-based baselines across diverse tasks and modalities. When integrated with modern encoders such as CLIP and DINOv2, the framework proves highly effective in capturing nuanced multimodal interactions while scaling efficiently to high-dimensional data. These contributions bridge theoretical rigor and practical utility, establishing a new foundation for scalable dependency learning in deep neural networks.

REFERENCES

Satoshi Akaho. A kernel method for canonical correlation analysis. *CoRR*, abs/cs/0609071, 2006.

Okechukwu Akpona, Vanessa D. L. Sebastian, and Harald Patrick. Berlin-urban-gradient dataset 2009-an enmap preparatory flight campaign (datasets). `https://doi.org/10.5880/enmap`, 2016.

Gavin Andrew, Raman Arora, Jeffrey A. Bilmes, and Karen Livescu. Deep-canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pp. 1247–1255, 2013.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:1–1, 2018.

Lorenzo Basile, Santiago Acevedo, Luca Bortolussi, Fabio Anselmi, and Alex Rodriguez. Intrinsic dimension correlation: Uncovering nonlinear connections in multimodal representations. In *International Conference on Learning Representations (ICLR)*, 2025.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Carlos Busso, Mehmet Bulut, Chul-Hwa Lee, Anoop Kazemzadeh, Emily Mower, Sungjin Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Satya Chandar, Mahdi Khapra, Hugo Larochelle, and Balaji Ravindran. Correlational neural networks. *Neural Computation*, 28(2):257–285, 2016.

Xiang Chang, Tao Xiang, and Timothy M. Hospedales. Scalable and effective deep cca via soft decorrelation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Ludwig Beyer, Alexander Kolesnikov, Daniel Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mohammad Norouzi Dehghani, Martin Minderer, Georg Heigold, Sylvain Gelly, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–21, 2021.

Andrew Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2121–2129, 2013.

Amit Gandhi, Kartik Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023. doi: 10.1016/j.inffus.2022.12.003.

Herbert Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. doi: 10.1002/zamm.19410210604.

Wei Guo, Jian Wang, and Sheng Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2944128.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5803–5812, 2017.

H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935. doi: 10.1017/S0305004100013517.

Dong Hong, Jun Hu, Jian Yao, Jocelyn Chanussot, and Xiao Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021. doi: 10.1016/j.isprsjprs.2021.03.023.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

Shang-Lung Huang and Xiao Xu. On the sample complexity of hgr maximal correlation functions for large datasets. *IEEE Transactions on Information Theory*, 67(3):1951–1980, 2021. doi: 10.1109/TIT.2021.3050651.

Shang-Lung Huang, Anand Makur, Li Zheng, and Gregory W. Wornell. An information-theoretic approach to universal feature selection in high dimensional inference. In *International Symposium on Information Theory (ISIT)*, pp. 1336–1340, 2017. doi: 10.1109/ISIT.2017.8086959.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yang Liang, Fei Ma, Yang Li, and Shang-Lung Huang. Person recognition with hgr maximal correlation on multimodal data. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pp. 2188–2195, 2021. doi: 10.1109/ICPR48806.2021.9630454.

Jing Lin, Fei Gao, Xiao Shi, and Jin Dong. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. doi: 10.1109/TGRS.2022.3222819.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Fei Ma, Shang-Lung Huang, and Li Zhang. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. doi: 10.1109/ICME54052.2021.9469087.

Fei Ma, Yang Li, Song Ni, Shang-Lung Huang, and Li Zhang. Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional gan. *Applied Sciences*, 12:527, 2022. doi: 10.3390/app1203527.

Xianping Ma, Xiaokang Zhang, Man-On Pun, and Ming Liu. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5403215, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafranicz, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 1(1):1–32, 01 2024. URL https://openreview.net/forum?id=a68SlItr6ZEt. *core team, **equal contribution.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Hungarica*, 10(3-4):441–451, 1959. doi: 10.1007/BF02024507.

Qian Shi, Da He, Zhengyu Liu, Xiaoping Liu, and Jingqian Xue. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping. *Journal of Remote Sensing*, 3(0078):1–21, 2023. doi: 10.34133/remotesensing.0078.

Tian Shi and Shang-Lung Huang. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14752–14766, 2023. doi: 10.18653/v1/P23-1475.

Javid Summaira, Xiao Li, Amna M. Shoib, Song Li, and Javed Abdul. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087*, 2021. doi: 10.48550/arXiv.2105.11087.

Li Wang, Jie Wu, Shang-Lung Huang, Li Zheng, Xiao Xu, Li Zhang, and Jie Huang. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5281–5288, 2019. doi: 10.1609/aaai.v33i01.33013281.

Long Wang, Rui Li, Chao Zhang, Shuai Fang, Chao Duan, Xiaojun Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. doi: 10.1016/j.isprsjprs.2022.06.008.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Xiao Wu, Dong Hong, and Jocelyn Chanussot. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022. doi: 10.1109/TGRS.2021.3114486.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Xiao Xu and Shang-Lung Huang. Maximal correlation regression. *IEEE Access*, 8:26591–26601, 2020. doi: 10.1109/ACCESS.2020.3022789.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, 2023.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2022.

Hongkang Zhang, Shao-Lun Huang, and Ercan Engin Kuruoglu. Mhfnet: An improved hgr multimodal network for informative correlation fusion in remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:15052–15066, 2024.

Xingjian Zhen, Zihang Meng, Rudrasis Chakraborty, and Vikas Singh. On the versatile uses of partial distance correlation in deep learning. In *European Conference on Computer Vision (ECCV)*, 2022.

Appendix:

A Detailed Derivation and Algorithm

This section provides a comprehensive, step-by-step derivation of the UniFast HGR objective function starting from the original Soft-HGR formulation, followed by the detailed algorithmic procedures. The derivation is structured around the three core innovations: enforcement of variance constraints, substitution with cosine similarity, and the expansion of the trace term.

A.1 Step 1: Variance Constraints and Whitening Alignment

The original Soft-HGR objective is given by:

$$J_{\text{soft}}(f,g) = \mathbb{E}\left[f(X)^T g(Y)\right] - \frac{1}{2}\text{tr}\left(\text{cov}(f(X))\text{cov}(g(Y))\right), \quad \text{s.t. } \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0 \tag{18}$$

To align with the whitening constraints of the canonical HGR definition ($\text{cov}(f) = \text{cov}(g) = I$) and stabilize optimization, we enforce unit variance on the feature mappings. This is achieved via ℓ_2-normalization:

$$f \leftarrow \frac{f - \mathbb{E}[f]}{\sqrt{\text{Var}[f]}}, \quad g \leftarrow \frac{g - \mathbb{E}[g]}{\sqrt{\text{Var}[g]}} \tag{19}$$

which ensures $\mathbb{E}[f] = \mathbb{E}[g] = 0$ and $\text{Var}[f] = \text{Var}[g] = 1$ for all output dimensions. This step is critical as it bounds the output and ensures numerical stability, providing a firm foundation for the subsequent substitution.

A.2 Step 2: Reformulation of the Sample-wise Term using Cosine Similarity

Under the zero-mean and unit-variance constraints, the sample-wise correlation term simplifies directly. Starting from its definition:

$$\mathbb{E}\left[f(X)^T g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i) - \mathbb{E}[f])^T(g(y_i) - \mathbb{E}[g]) \tag{20}$$

Given $\mathbb{E}[f] = \mathbb{E}[g] = 0$, this reduces to:

$$\mathbb{E}\left[f(X)^T g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N} f(x_i)^T g(y_i) \tag{21}$$

Now, with $\text{Var}[f] = \text{Var}[g] = 1$, implying $\|f(x_i)\|_2 = 1$ and $\|g(y_i)\|_2 = 1$ for all i (in expectation), the dot product is equivalent to cosine similarity:

$$f(x_i)^T g(y_i) = \|f(x_i)\|_2 \|g(y_i)\|_2 \cdot \cos(f(x_i), g(y_i)) = \cos(f(x_i), g(y_i)) \tag{22}$$

Thus, the first term becomes:

$$\mathbb{E}\left[f(X)^T g(Y)\right] = \frac{1}{N-1}\sum_{i=1}^{N}\cos(f(x_i), g(y_i)) \tag{23}$$

This substitution replaces a covariance-based calculation with a norm-bounded, stable cosine operation, reducing computational complexity.

A.3 Step 3: Expansion and Simplification of the Trace Term

The trace term $\text{tr}\left(\text{cov}(f)\text{cov}(g)\right)$ measures the distributional correlation. We expand it to understand its structure under variance constraints. First, recall the covariance matrix computation for f:

$$\text{cov}(f) = \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i) - \mathbb{E}[f])(f(x_i) - \mathbb{E}[f])^T \tag{24}$$

With $\mathbb{E}[f] = 0$, this simplifies to:

$$\text{cov}(f) = \frac{1}{N-1} \sum_{i=1}^{N} f(x_i) f(x_i)^T \tag{25}$$

Let $\mathbf{F}$ be the matrix with rows $f(x_i)^T$, then $\text{cov}(f) = \frac{1}{N-1}\mathbf{F}^T\mathbf{F}$. Similarly, $\text{cov}(g) = \frac{1}{N-1}\mathbf{G}^T\mathbf{G}$ for matrix $\mathbf{G}$ with rows $g(y_i)^T$. The trace term is:

$$\text{tr}(\text{cov}(f)\text{cov}(g)) = \text{tr}\left(\frac{1}{(N-1)^2}\mathbf{F}^T\mathbf{F}\mathbf{G}^T\mathbf{G}\right) = \frac{1}{(N-1)^2}\text{tr}(\mathbf{F}^T\mathbf{F}\mathbf{G}^T\mathbf{G}) \tag{26}$$

Using the cyclic property of the trace, $\text{tr}(\mathbf{F}^T\mathbf{F}\mathbf{G}^T\mathbf{G}) = \text{tr}(\mathbf{F}\mathbf{G}^T\mathbf{G}\mathbf{F}^T) = \text{tr}((\mathbf{F}\mathbf{G}^T)(\mathbf{G}\mathbf{F}^T))$. Notice that $\mathbf{F}\mathbf{G}^T$ is the Gram matrix of pairwise dot products between $f(x_i)$ and $g(y_j)$, and $\mathbf{G}\mathbf{F}^T$ is its transpose. A more intuitive interpretation is to see the (i, j)-th element of $\text{cov}(f)$ as the covariance between the i-th and j-th dimensions of $f(X)$. Under unit variance, this becomes their correlation coefficient ρ^f_{ij}. The trace term $\text{tr}(\text{cov}(f)\text{cov}(g))$ is then the sum over i, j of $\rho^f_{ij}\rho^g_{ji}$. This can be interpreted as a dot product between the vectorized correlation matrices, measuring their similarity. To operationalize this, we define *distribution vectors* distri_f and distri_g. For each sample i, the distribution vector distri^i_f is formed by the correlations between the i-th dimension and all other dimensions of $f(X)$ (i.e., a row of the correlation matrix). The cosine similarity between distri^i_f and distri^i_g then captures the alignment of the internal correlation structures induced by the i-th dimension. The trace term can be approximated as the average of these cosine similarities across dimensions:

$$\text{tr}(\text{cov}(f)\text{cov}(g)) \approx \frac{1}{N-1} \sum_{i=1}^{N} \cos(\text{distri}^i_f, \text{distri}^i_g) \tag{27}$$

This transformation is key to efficiently calculating the distributional correlation without explicit matrix multiplication.

A.4 Step 4: Composition of the Final UniFast HGR Objective

Combining the simplified sample-wise term (Eq. 7) and the approximated trace term (Eq. 13), the Soft-HGR objective transforms into:

$$J_{\text{soft}}(f, g) \approx \frac{1}{N-1} \sum_{i=1}^{N} \cos(f(x_i), g(y_i)) - \frac{1}{2} \cdot \frac{1}{N-1} \sum_{i=1}^{N} \cos(\text{distri}^i_f, \text{distri}^i_g) \tag{28}$$

Simplifying the constants yields the final UniFast HGR objective:

$$\text{UF-HGR} = \frac{1}{N-1} \sum_{i=1}^{N} \cos(f(x_i), g(y_i)) - \frac{1}{2(N-1)} \sum_{i=1}^{N} \cos(\text{distri}^i_f, \text{distri}^i_g) \tag{29}$$

This formulation retains the original intent of HGR—maximizing both sample-wise and distributional dependency—while being computationally tractable and stable for deep learning applications.

A.5 Algorithm Implementation

The following algorithms detail the computation of UniFast HGR (Algorithm 1) and OptFast HGR (Algorithm 2), with steps aligned to the derivation above.

B Theoretical Analysis of Bias in OptFast HGR

OptFast HGR accelerates HGR maximal correlation computation through stochastic bias correction and simplified distribution matrix analysis, introducing a controlled bias to balance efficiency and accuracy.

Algorithm 1 UniFast HGR Algorithm

Input: $m \times n$ feature matrix of f, g
Output: Objective value of UniFast HGR

1. Normalization:
   $f \leftarrow \frac{f}{\|f\|_2}, g \leftarrow \frac{g}{\|g\|_2}$
2. Calculation of the cosine correlation coefficient between f and g:
   $\cos(f,g) = f \cdot g, corr = \frac{1}{N-1}\sum_{i=1}^{N}\cos(f,g) = \frac{1}{N-1}\sum_{i=1}^{N} f \cdot g$
3. Calculation of the distribution matrix:
   $distri_f = f \cdot f^T, \quad distri_g = g \cdot g^T$
4. Initialization processing:
   $distri_f \leftarrow$ Extract upper triangular part of $distri_f$ (excluding diagonal) using torch.triu
   $distri_g \leftarrow$ Extract upper triangular part of $distri_g$ (excluding diagonal) using torch.triu
   Utilize symmetry of $distri_f$ and $distri_g$ to restore complete matrix from upper triangular part.
5. Normalization:
   $distri_f \leftarrow \frac{distri_f}{\|\, distri_f \,\|_2}, \quad distri_g \leftarrow \frac{distri_g}{\|\, distri_g \,\|_2}$
6. Calculation of the cosine correlation coefficient between $distri_f$ and $distri_g$:
   $tr = \frac{1}{N-1}\sum_{i=1}^{N}\cos(distri_f, distri_g) = \frac{1}{N-1}\sum_{i=1}^{N} distri_f \cdot distri_g$
7. Calculation of the UniFast HGR objective:
   $\frac{1}{N-1}\sum_{i=1}^{N}\cos(f,g) - \frac{1}{2(N-1)}\sum_{i=1}^{N}\cos(\text{distri}_f, \text{distri}_g)$

Algorithm 2 OptFast HGR Algorithm

Input: $m \times n$ feature matrix of f, g
Output: Objective value of OptFast HGR

1. Initialization processing:
   Generate t_R random matrix h of the same scale as f (from standard normal distribution)
2. Calculation of HGR bias term:
   HGR_bias $= \frac{2}{3t_R}\sum_{i=1}^{t_R}\text{OptFast} - \text{HGR}(h_i, 0)$ where 0 represents the bias reference
3. Normalization:
   $f \leftarrow \frac{f}{\|f\|_2}, g \leftarrow \frac{g}{\|g\|_2}$
4. Calculation of the cosine correlation coefficient between f and g:
   $corr = \frac{1}{N-1}\sum_{i=1}^{N}\cos(f,g)$
5. Calculation of the distribution matrix:
   $distri_f = f \cdot f^T, \quad distri_g = g \cdot g^T$
6. Initialization processing:
   $distri_f \leftarrow$ Extract upper triangular part of $distri_f$ (excluding diagonal) using torch.triu
   $distri_g \leftarrow$ Extract upper triangular part of $distri_g$ (excluding diagonal) using torch.triu
7. Calculation of the cosine correlation coefficient between distri$_f$ and distri$_g$:
   $tr = \frac{1}{N-1}\sum_{i=1}^{N}\cos(distri_f, distri_g) = \frac{1}{N-1}\sum_{i=1}^{N} distri_f \cdot distri_g$
8. Calculation of the OptFast HGR objective:
   $\left(\frac{1}{N-1}\sum_{i=1}^{N}\cos(f,g) - \frac{1}{2(N-1)}\sum_{i=1}^{N}\cos(\text{distri}_f, \text{distri}_g)\right) / (1 - \text{HGR\_bias})$

B.1 Source of Bias and Calibration Mechanism

The bias in OptFast HGR comes from two key approximations in its simplified computation:

(1). **Distribution Matrix Truncation**: OptFast HGR approximates the full eigenstructure of feature correlation matrices by using the upper triangular part of the outer product $f \cdot f^\top$ (excluding diagonal elements), avoiding explicit eigenvalue decomposition. This truncation introduces an approximation error as it doesn't fully capture the full matrix's spectral properties.

(2). **Randomized Bias Estimation**: To quantify the bias from truncation and finite sampling, OptFast HGR uses Monte Carlo integration with random feature vectors $\mathbf{v}_i \sim \mathcal{U}(-1,1)^{B\times d}$:

$$\text{HGR\_bias} = \frac{1}{t_R}\sum_{i=1}^{t_R} \text{OptFast-HGR}(\mathbf{v}_i, \mathbf{0}), \tag{30}$$

where OptFast-HGR$(\mathbf{v}_i, \mathbf{0})$ uses the same truncated distribution matrix approach. This estimates the expected spurious correlation from random noise.

B.2 Statistical Convergence and Error Bounds

The error in OptFast HGR has two sources:

(1). **Matrix Approximation Error**: For $D = f \cdot f^\top$ with sorted eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, the truncation error is bounded by the spectral gap:

$$|\text{OptFast HGR} - \text{HGR}^*| \leq O\left(\frac{\lambda_2}{\lambda_1}\right), \tag{31}$$

where HGR^* is the exact HGR score. A larger spectral gap ensures faster convergence.

(2). **Bias Estimation Variance**: By the Central Limit Theorem, the variance of HGR_bias decays as $\mathcal{O}(1/\sqrt{t_R})$:

$$\text{Var}(\text{HGR\_bias}) \leq \frac{C}{t_R}\left(\frac{1}{B^2} + \frac{d}{B^3}\right), \tag{32}$$

with C a constant dependent on the feature distribution. This ensures stable bias correction for large batch sizes.

B.3 Empirical Validation of Bias-Accuracy Trade-off

Experiments on diverse datasets validate the framework:

Accuracy Preservation: Top-1 error in cross-modal tasks increases by $\leq 1\%$ compared to exact HGR, with segmentation IoU within 1% of the baseline.

Batch Stability: Bias correction stabilizes performance across training iterations, as shown by consistent R@1 scores in COCO retrieval tasks.

B.4 Robustness to Feature Distributions

OptFast HGR maintains bounded bias under diverse data regimes:

Non-Gaussian and Asymmetric Modalities: Stable performance on IEMOCAP and COCO shows resilience to heterogeneous data distributions.

High-Dimensional Features: For dim ≥ 64, the growing spectral gap reduces matrix approximation error, while randomized bias correction mitigates variance in low-data scenarios.

The theoretical framework ensures OptFast HGR's bias is statistically controlled and computationally tractable, making it suitable for large-scale multimodal tasks where exact HGR computation is infeasible.

C Asymptotic Complexity Comparison

We compare the asymptotic complexity of UniFast HGR and OptFast HGR to earlier methods like Soft-HGR and Deep CCA. Table 5 also shows comparisons with other nonlinear correlation analysis methods such as CKA, dCor, and the latest I_dCor.

UniFast HGR uses non-iterative matrix operations to reduce complexity. Unlike Soft-HGR, it avoids calculating high-dimensional covariance matrices directly. OptFast HGR enhances numerical stability by reducing normalization steps and incorporating bias correction, while maintaining efficiency.

Table 5: Overall complexity comparison analysis (sample size m, feature dimension K,network layers L)

Methods	Time Complexity	Characteristic
CCA	$O(mK^2 + K^3)$	Classic method, not for deep nonlinear
DeepCCA	$O(LmK^2)$	Nonlinear, high training cost
Soft CCA	$O(LmK^2)$	Controls redundancy, for low-dim tasks
Soft-HGR	$O(mK^2 + K^3)$	Sensitive to high-dim features
CKA	$O(m^2K + m^3)$	Quantifies structural similarity
dCor	$O(m^2K)$	Captures nonlinear dependencies
I_dCor	$O(m^2K)$	Sensitive to high-dimensional noise
UniFast HGR	$O(m^2K)$	Efficient for high-dim data
OptFast HGR	$O(m^2K)$	Stable for complex tasks

Both versions feature fully differentiable structures, enabling direct handling of multimodal features. Their asymptotic complexity is significantly lower than traditional CCA-based methods, making them suitable for neural network training with improved efficiency, cross-modal adaptability, and robustness.

D Detailed Experimental Results

D.1 Image Classification

Table 6 presents the detailed comparative experimental results of remote sensing data classification using a dual-channel deep learning framework with ResNet 50 as the backbone on the Berlin dataset. Table 7 displays the results of using a dual-channel visual transformer framework on the Houston 2018 dataset for the same task.

The results demonstrate that the proposed UniFast HGR and OptFast HGR methods consistently outperform traditional CCA and similarity-based methods. This suggests that our proposed methods effectively capture complex data patterns and significantly enhance classification performance on both the Berlin HIS-SAR and Houston 2018 HSI LiDAR datasets, irrespective of the framework used (CNN or transformer).

Traditional CCA appears less effective at capturing the intricate nonlinear relationships inherent in remote sensing data. Deep CCA exhibits a modest improvement over CCA, suggesting that the integration of deep learning techniques can more effectively grasp these nonlinearities. Both Cosine Similarity and Dot Product perform admirably, highlighting the efficacy of straightforward vector operations for the given datasets. In contrast, Soft HGR underperforms, particularly in OA metrics, likely due to its propensity to induce substantial alterations in covariance and matrix trajectories, potentially leading to gradient explosions and diminished model efficacy.

To evaluate the computational efficiency of the proposed UniFast HGR and OptFast HGR, we compared the execution time of remote sensing data classification on the Berlin dataset and the Houston 2018 dataset, using a dual-channel deep learning framework with ResNet-50 as the backbone and a dual-channel visual transformer framework, respectively, as shown in Table 8. The results indicate that CCA, Deep CCA, and Soft CCA had the longest execution times, which were also influenced by the network structure used, whereas UniFast HGR and OptFast HGR were less impacted by these structural complexities.

D.2 Remote Sensing Semantic Segmentation

The detailed experimental results of remote sensing semantic segmentation on the Vaihingen dataset are shown in Table 9. The detailed experimental results of remote sensing semantic segmentation on the Globe230k dataset are shown in Table 10. Figure 3 shows a visualization example of the experimental results of remote sensing semantic segmentation using 8 correlation methods on the Vaihingen dataset. It is evident that when using UniFast HGR and OptFast HGR, complex long-distance

Table 6: Comparison of various methods on the Berlin HIS-SAR dataset(%)

Class	CCA	Deep CCA	Soft CCA	Dot Product	Cosine Similarity	Soft HGR	UniFast HGR	OptFast HGR
OA	70.93	71.54	72.74	75.20	75.51	65.80	**80.75**	80.46
AA	64.35	61.14	65.08	66.22	65.53	64.30	**71.53**	71.51
Kappa	58.28	58.33	60.23	62.77	62.53	52.99	**70.44**	70.21
Forest	81.90	87.16	64.17	76.68	79.92	67.54	**87.61**	82.18
Residential area	72.81	75.59	76.38	82.57	85.63	63.87	**86.85**	85.10
Industrial area	23.05	53.61	**76.00**	48.15	49.11	64.07	40.20	62.67
Low plants	71.44	62.68	89.08	65.08	54.31	82.05	73.70	**89.23**
Soil	85.97	78.01	72.10	82.53	82.88	**88.16**	82.42	78.63
Allotment	69.87	51.72	58.73	**70.73**	69.07	55.79	65.35	65.65
Commercial area	**56.76**	42.81	20.40	35.88	23.77	37.97	54.30	27.61
Water	52.98	37.53	63.78	68.15	79.58	54.95	**81.85**	81.01

Table 7: Comparison of various methods on the Houston 2018 HSI-LiDAR dataset(%)

Class	CCA	Deep CCA	Soft CCA	Dot Product	Cosine Similarity	Soft HGR	UniFast HGR	OptFast HGR
OA	88.28	89.82	88.81	91.59	92.04	85.86	**93.65**	93.25
AA	92.20	93.92	93.14	93.85	94.67	91.01	**96.15**	95.71
Kappa	84.89	86.89	85.62	89.13	89.65	81.91	**91.77**	91.25
Healthy grass	95.62	97.84	97.97	78.15	98.24	**98.76**	95.18	97.66
Stressed grass	86.77	83.27	89.16	**97.58**	89.66	83.84	93.57	93.27
Artificial turf	**100.00**	99.83	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Evergreen trees	99.05	98.28	97.81	96.15	98.95	97.80	**99.37**	98.45
Deciduous trees	96.05	95.18	95.92	94.94	97.57	96.69	**98.75**	98.01
Bare earth	**100.00**	**100.00**	**100.00**	99.99	**100.00**	99.99	**100.00**	99.99
Water	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Residential buildings	94.02	97.90	97.42	96.88	91.92	**98.49**	97.04	98.20
Non-residential buildings	94.80	94.53	93.48	95.92	97.47	91.40	**98.89**	96.86
Road	56.85	69.52	62.37	74.35	69.20	50.99	**82.82**	79.26
Sidewalks	81.24	78.02	71.27	73.72	**83.17**	65.75	82.75	78.53
Crosswalks	76.18	95.93	87.92	91.78	91.40	74.92	**96.82**	92.96
Major thoroughfares	73.24	79.62	82.78	85.45	86.32	78.80	85.47	**87.16**
Highways	98.90	95.04	96.08	97.65	99.47	96.73	98.24	**99.67**
Railways	99.77	99.87	99.87	99.60	99.50	99.40	**99.94**	99.90
Paved parking lots	92.95	96.88	94.18	**97.46**	92.83	93.98	97.02	95.53
Unpaved parking lots	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	94.07	**100.00**	**100.00**
Cars	99.13	97.41	97.17	97.45	97.65	98.53	**99.16**	98.70
Trains	99.95	99.41	99.57	**100.00**	**100.00**	**100.00**	99.99	**100.00**
Stadium seats	99.57	99.94	99.83	**100.00**	**100.00**	**100.00**	99.98	**100.00**

Table 8: Execution time comparison on the Berlin and Houston2018 datasets (Time(s/epoch))

Method	ResNet 50		Vision Transformer	
	Berlin dataset	Houston2018 dataset	Berlin dataset	Houston2018 dataset
CCA	2967.52	/	307.82	1243.23
Deep CCA	250.51	1158.42	379.82	1520.09
Soft CCA	314.93	1751.98	211.03	929.50
Dot Product	23.18	106.05	20.85	48.89
Cosine Similarity	23.40	106.14	20.93	49.34
Soft-HGR	25.83	110.53	21.62	58.03
UniFast HGR	24.53	108.56	21.23	57.00
OptFast HGR	23.54	106.27	21.02	52.41

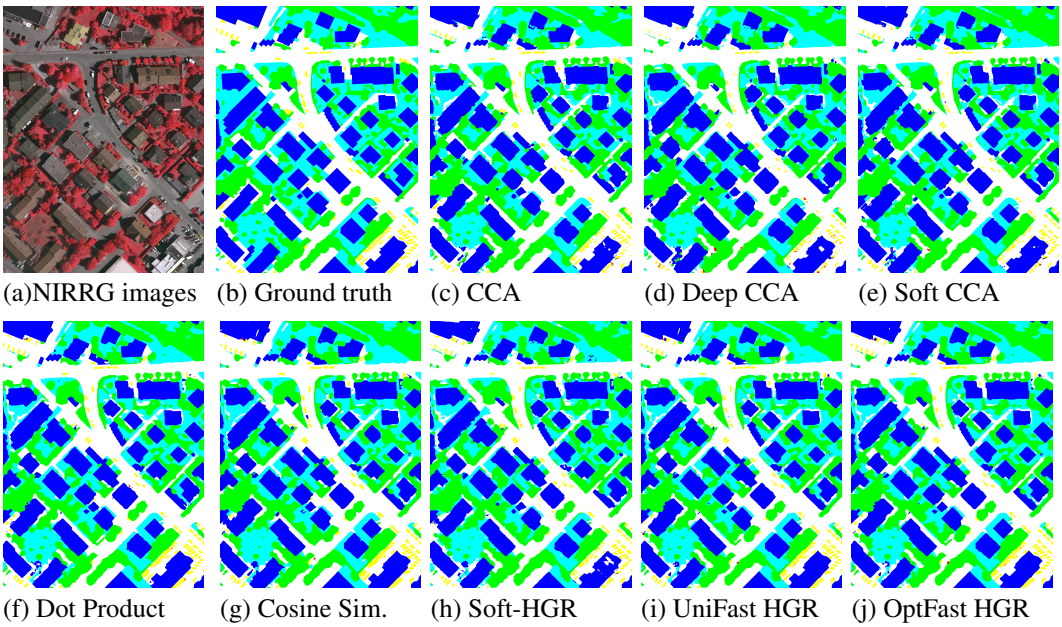
semantic information can be more accurately recognized, and precise edges of the recognized object can be obtained, thereby achieving more accurate semantic segmentation of remote sensing imagery.

Table 9: Comparison of various methods on the Vaihingen dataset(%).

Class	CCA	Deep CCA	Soft CCA	Dot Product	Cosine Similarity	Soft HGR	UniFast HGR	OptFast HGR
OA	91.15	91.39	91.41	92.61	92.56	90.10	**93.01**	92.95
mIoU	79.37	81.35	81.44	83.65	83.34	76.87	**84.62**	84.57
Imp.	91.43	92.57	92.52	**94.97**	93.38	91.39	93.62	93.47
Building	97.37	96.94	97.19	95.55	97.62	95.93	97.86	**97.92**
Low.	80.19	79.51	79.62	80.36	81.94	73.08	**82.03**	81.86
Tree	91.03	91.53	91.24	**94.93**	92.67	93.41	93.82	93.79
Car	76.94	82.94	83.76	83.41	88.53	73.86	**90.15**	89.95

Table 10: Comparison of various methods on the Globe230k dataset (%).

Class	CCA	Deep CCA	Soft CCA	Dot Product	Cosine Similarity	Soft HGR	UniFast HGR	OptFast HGR
OA	87.92	88.27	87.60	90.92	90.81	86.46	**91.48**	91.23
mIoU	67.49	67.85	66.71	75.67	75.53	64.82	**76.36**	76.15
Cropland	83.27	91.86	79.12	89.76	90.19	91.75	**92.15**	90.32
Forest	91.60	95.51	90.20	95.24	96.32	93.46	96.73	**96.89**
Grassland	58.75	65.44	61.48	79.93	78.47	54.83	**80.68**	80.31
Shrubland	62.49	73.07	55.34	72.89	71.50	57.63	**75.41**	72.62
Wetland	73.08	71.80	42.76	77.54	76.72	42.09	77.92	**78.49**
Water	85.22	89.62	90.83	94.65	94.26	83.69	**95.62**	95.35
Tundra	9.31	0.00	5.32	38.58	36.82	0.00	**43.07**	41.27
Impervious surface	80.92	86.59	81.50	93.17	92.90	80.78	93.50	**94.10**
Bareland	72.43	87.37	74.57	91.10	90.64	73.15	**91.46**	91.07
Ice/ snow	91.25	97.53	91.82	97.62	98.21	90.76	**98.39**	97.85



(a)NIRRG images (b) Ground truth (c) CCA (d) Deep CCA (e) Soft CCA

(f) Dot Product (g) Cosine Sim. (h) Soft-HGR (i) UniFast HGR (j) OptFast HGR

Figure 3: Experimental images on the Vaihingen test set

D.3 MULTIMODAL EMOTION RECOGNITION

The emotion recognition experiments on the IEMOCAP dataset, as detailed in Table 11, indicate that the UniFast HGR and OptFast HGR methods generally excel over conventional CCA and similarity-based approaches. This suggests their enhanced capability for multimodal emotion recognition. UniFast HGR and OptFast HGR demonstrate superior performance across all classifications, showcasing their capacity to effectively capture the nuanced patterns associated with various emotions. Thus, the proposed methods are highly appropriate for emotion recognition tasks and could be applied to other datasets and domains. Future research could integrate these methods with additional modalities like facial expressions and physiological signals to further refine emotion recognition performance.

Table 11: Comparison of various methods on the IEMOCAP dataset(%)

Class	CCA	Deep CCA	Soft CCA	Dot Product	Cosine Similarity	Soft HGR	UniFast HGR	OptFast HGR
W-F1	67.51	67.82	68.57	69.87	69.60	71.43	**73.57**	73.32
ACC	67.41	67.78	68.58	70.14	69.50	71.29	**73.66**	73.43
Happy	50.77	49.81	46.77	50.51	53.85	54.92	**66.63**	59.67
Sad	79.65	81.82	79.29	81.96	81.39	81.53	84.79	**85.23**
Neutral	68.11	69.58	69.59	71.24	71.89	70.84	**74.30**	73.00
Angry	61.98	62.53	64.60	65.90	65.82	70.32	70.46	**71.04**
Excited	76.70	76.56	75.00	74.48	74.91	75.00	**77.14**	77.09
Frustrated	60.66	59.35	65.62	67.32	63.17	69.45	**71.22**	70.36

D.4 IMAGE CLASSIFICATION ON CIFAR-100

To evaluate the scalability of UniFast HGR and OptFast HGR on standard visual classification tasks, we conducted experiments on the CIFAR-100 dataset using five representative backbone architectures: ViT-B/32, ResNet50, CLIP, SigLIP, and DINOv2. We compared these with nonlinear correlation baselines (CKA, dCor, I_dCor) and Soft-HGR, reporting results as top-1 accuracy (%) in Table 12.

UniFast HGR consistently outperformed all baselines across architectures. For ResNet50, it achieved 76.8% accuracy—2.3% higher than the baseline (74.5%) and 1.0% higher than the strongest baseline (I_dCor, 75.8%). On ViT-B/32, UniFast HGR reached 86.4%, surpassing I_dCor (86.1%) and Soft-HGR (85.5%) by 0.3% and 0.9%, respectively. With state-of-the-art vision models like SigLIP and DINOv2, UniFast HGR achieved top performance: 88.9% (SigLIP) and 89.3% (DINOv2)—improving by 0.7% and 0.6% over IdCor (88.2% and 88.7%), respectively.

OptFast HGR balanced efficiency and accuracy, achieving results close to UniFast HGR. For instance, on DINOv2, OptFast HGR attained 88.5%—0.8% below UniFast HGR but still 0.2% higher than I_dCor. Across all models, the accuracy gap between UniFast and OptFast HGR remained within 1.0%, showing OptFast HGR's effectiveness in reducing computational overhead with minimal performance loss. These results highlight the frameworks' ability to enhance feature correlation learning across diverse architectures, confirming their scalability for large-scale image classification.

Table 12: The experimental results on CIFAR-100 Datasets

Dataset	Model	Baseline	CKA	dCor	I_dCor	Soft-HGR	UniFast HGR	OptFast HGR
CIFAR-100 Accuracy (%)	ViT-B/32	85.3	85.6	85.8	86.1	85.5	**86.4**	86.2
	ResNet50	74.5	75.1	75.2	75.8	75.3	**76.8**	76.1
	CLIP	80.5	81.2	81.4	81.6	81.3	**82.5**	81.5
	SigLIP	87.1	87.5	87.8	88.2	87.4	**88.9**	88.4
	DINOv2	87.5	87.8	88.3	88.7	87.7	**89.3**	88.5

D.5 Comparison with State-of-the-Art Foundation Models

Our work focuses on optimizing the calculation of HGR's maximum correlation to enhance feature alignment in multimodal learning tasks. We integrated our method into models such as CLIP (ViT - B/32), SigLIP, and DINOv2 (ViT - L/14), and measured the similarity of ImageNet embeddings across different models. In Table 13, we report the correlation results obtained with our method, along with the correlation scores from dCor, I_dCor, and soft-HGR. UniFast HGR improved the correlation score when applied to the base model, validating its efficacy as a supplementary module.

Table 13: Correlation results on ImageNet representations.

Methods	Models	EfficientNet	ResNet50	ViT-B/32	CLIP	SigLIP	DINOv2
dCor	EfficientNet	1.	0.45	0.42	0.29	0.34	0.41
	ResNet50	0.45	1.	0.43	0.54	0.58	0.56
	ViT-B/32	0.42	0.43	1.	0.46	0.49	0.48
	CLIP	0.29	0.54	0.46	1.	0.82	0.78
	SigLIP	0.34	0.58	0.49	0.82	1.	0.80
	DINOv2	0.41	0.56	0.48	0.78	0.80	1.
I_dCor	EfficientNet	1.	0.91	0.85	0.77	0.81	0.82
	ResNet50	0.91	1.	0.86	0.80	0.83	0.81
	ViT-B/32	0.85	0.86	1.	0.92	0.92	0.90
	CLIP	0.77	0.80	0.92	1.	0.91	0.89
	SigLIP	0.81	0.83	0.92	0.91	1.	0.92
	DINOv2	0.82	0.81	0.90	0.89	0.92	1.
Soft-HGR	EfficientNet	1.	0.63	0.61	0.55	0.57	0.60
	ResNet50	0.63	1.	0.62	0.71	0.74	0.73
	ViT-B/32	0.61	0.62	1.	0.66	0.70	0.68
	CLIP	0.55	0.71	0.66	1.	0.85	0.82
	SigLIP	0.57	0.75	0.70	0.85	1.	0.85
	DINOv2	0.60	0.73	0.68	0.82	0.85	1.
UniFast HGR	EfficientNet	1.	0.92	0.87	0.84	0.87	0.86
	ResNet50	0.92	1.	0.86	0.85	0.88	0.84
	ViT-B/32	0.87	0.86	1.	0.93	0.94	0.92
	CLIP	0.84	0.85	0.93	1.	0.92	0.91
	SigLIP	0.87	0.88	0.94	0.92	1.	0.94
	DINOv2	0.86	0.84	0.92	0.91	0.94	1.
OptFast HGR	EfficientNet	1.	0.91	0.85	0.82	0.83	0.83
	ResNet50	0.91	1.	0.85	0.82	0.83	0.83
	ViT-B/32	0.84	0.85	1.	0.91	0.92	0.91
	CLIP	0.79	0.82	0.91	1.	0.91	0.90
	SigLIP	0.82	0.83	0.92	0.91	1.	0.92
	DINOv2	0.82	0.83	0.91	0.90	0.92	1.

E Ablation Studies

Comprehensive ablation studies were conducted to evaluate the contributions of key components in the proposed framework, including variance constraints, diagonal removal, cosine similarity formulation, and the OptFast optimization strategy.

For **variance constraints**, Table 14 demonstrates that without proper variance normalization, covariance calculations become numerically unstable, leading to gradient explosions and significant performance degradation. On the Berlin dataset, the absence of variance constraints reduces OA to 68.53% and AA to 67.26%. In contrast, UniFast HGR with variance constraints maintains output values within the bounded range of [-1, 1], significantly reducing sensitivity to signal variance variations. This stabilization enables improved feature learning efficiency and accuracy, achieving 80.75% OA and 71.53% AA on the Berlin dataset.

Regarding **main diagonal removal**, the cosine similarity matrix exhibits symmetry and positive definiteness, with main diagonal elements representing autocorrelation values fixed at 1. These diagonal entries provide no inter-sample discriminative information as they solely capture self-similarity rather than meaningful pairwise relationships between distinct samples. The removal of main diagonal elements from the correlation matrix is justified by the trivial nature of these self-correlations, which does not compromise the method's capability to capture meaningful inter-sample dependencies. Results in Table 14 confirm that retaining the diagonal ("w/ Main Diagonal") consistently underperforms the complete UniFast HGR formulation. For instance, Houston 2018 OA decreases from 93.65% to 93.46% with diagonal retention. This performance gap becomes more pronounced in smaller batch sizes, where diagonal removal significantly improves gradient stability and convergence speed. Even with larger batches (e.g., size 256), performance remains competitive post-removal, demonstrating no accuracy loss. Both theoretical analysis and experimental results confirm that diagonal elements contribute no discriminative value while their removal enhances training stability without sacrificing performance. While the impact of removal lessens with larger batches, the estimator remains stable across all batch sizes post-removal. This makes diagonal removal especially crucial for smaller batches, where it significantly enhances numerical stability and convergence.

The comparison between **covariance and cosine similarity** formulations reveals significant computational advantages. The cosine-based implementation in UniFast HGR achieves approximately 2× faster computation compared to covariance-based Soft-HGR, while maintaining competitive accuracy across all evaluated datasets. This efficiency improvement stems from eliminating complex matrix operations required for covariance calculation and whitening.

The **OptFast variant** provides an optimized speed-accuracy trade-off, achieving 5-10% faster computation with minimal accuracy degradation. For instance, on ImageNet-1K, OptFast HGR maintains 79.6% Top-1 accuracy compared to UniFast HGR's 80.1%, representing a favorable trade-off for applications prioritizing inference speed.

Table 14: Comprehensive ablation study results across multiple datasets (%)

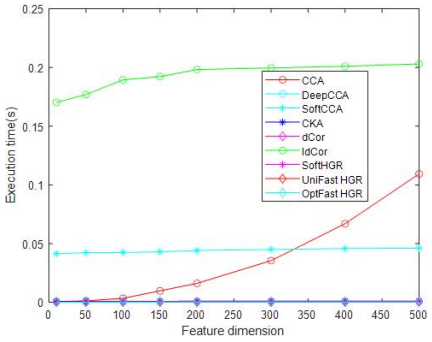
Methods	Berlin		Houston 2018		Vaihingen		Globe230k		IEMOCAP	
	OA	AA	OA	AA	OA	mIoU	OA	mIoU	W-F1	ACC
w/o Variance Constraints	68.53	67.26	86.72	92.24	90.82	77.55	87.41	66.96	71.62	71.49
w/ Main Diagonal	80.62	71.39	93.46	95.97	92.85	84.57	91.32	76.27	73.41	73.38
Covariance-based	79.83	70.92	92.87	95.43	92.26	83.89	90.75	75.64	72.95	72.87
OptFast HGR	80.41	71.28	93.52	96.02	92.91	84.48	91.38	76.29	73.46	73.42
UniFast HGR	**80.75**	**71.53**	**93.65**	**96.15**	**93.01**	**84.62**	**91.48**	**76.36**	**73.57**	**73.66**

The ablation studies demonstrate that each component contributes uniquely to the overall performance: Variance constraints prevent gradient instability and enable stable optimization Diagonal removal eliminates trivial self-correlations and focuses learning on meaningful cross-sample relationships Cosine similarity provides computational efficiency while maintaining representation quality OptFast variant offers a practical speed-accuracy trade-off for time-sensitive applications The complete UniFast HGR framework achieves the best balance between computational efficiency and representation quality across all evaluated datasets and metrics.

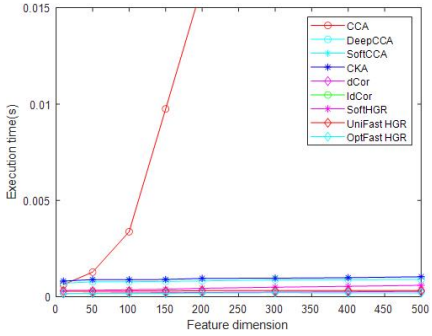
F Computational Efficiency

To evaluate the impact of feature dimension and batch size on computational performance, we measured the execution times of UniFast HGR, OptFast HGR, and baseline methods (CCA, Deep CCA, SoftCCA, CKA, dCor, I_dCor, Soft-HGR) using randomly generated tensor pairs $(f, g) \in \mathbb{R}^{bz \times \text{el}}$. Each method's correlation computation was repeated 10,000 times across batch sizes $bz \in \{16, 32, 64, 128, 256\}$ and feature dimensions el $\in \{10, 50, 100, 150, 200, 300, 400, 500\}$. The average execution times are visualized in Figure 4 . Figure 4 presents runtime comparisons across four representative batch sizes ($bz = 16, 64, 128, 256$):

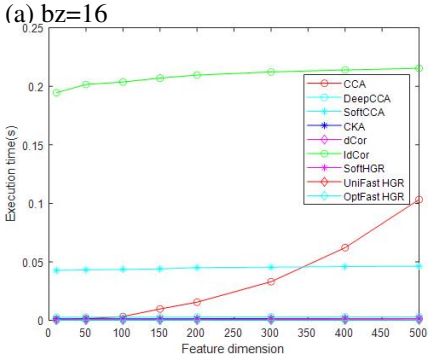
1. **OptFast HGR**: Linear Scaling with Feature Dimension. At $bz = 256$, OptFast HGR shows a gentle increase in runtime from 0.000265 seconds (el $= 10$) to 0.000877 seconds (el $= 500$), indicating near-linear complexity due to vectorized operations. This represents a significant speedup over CCA, which exhibits substantially higher runtimes due to its cubic complexity from covariance matrix decomposition.



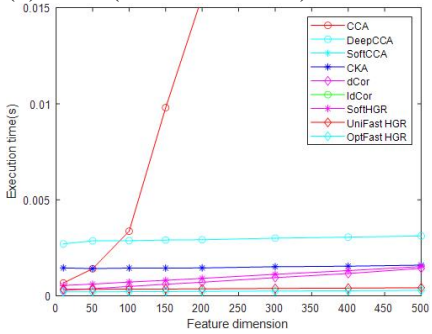
(a) bz=16



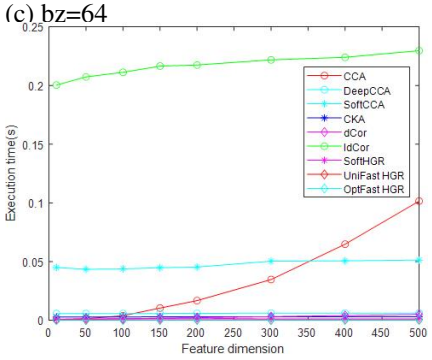
(b) bz=16(Zoomed-in View)



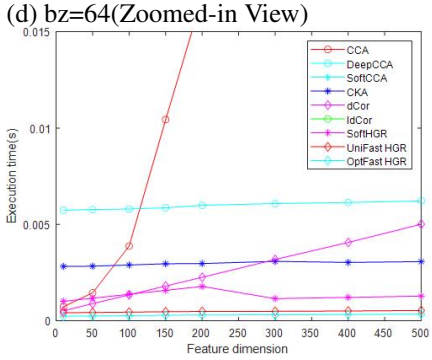
(c) bz=64



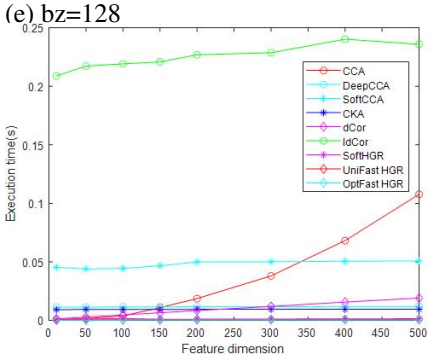
(d) bz=64(Zoomed-in View)



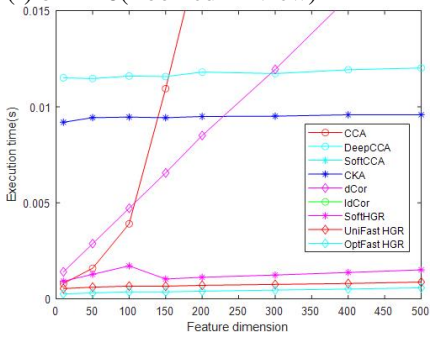
(e) bz=128



(f) bz=128(Zoomed-in View)



(g) bz=256



(h) bz=256(Zoomed-in View)

Figure 4: Comparison of execution time for correlation methods across different batch sizes and dimensions without interference

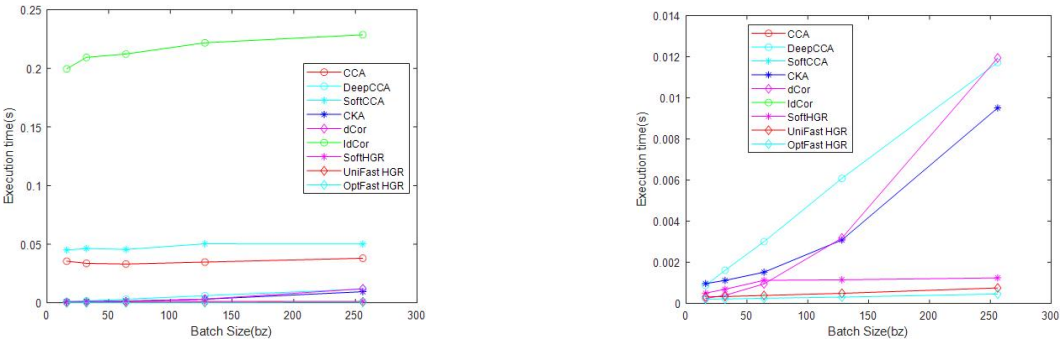
2. **UniFast HGR: Balanced Efficiency and Accuracy**. UniFast HGR has slightly higher runtimes than OptFast HGR due to its full distribution matrix computation but still outperforms baseline methods. At $bz = 128$, its runtime increases from 0.000419 seconds (el $= 10$) to 0.000537 seconds (el $= 500$), while CCA's runtime increases dramatically.

3. **Baseline Methods: Suboptimal Scalability**. CCA and dCor: Show superlinear growth in execution time. CCA's runtime increases substantially from el $= 10$ to el $= 500$ at $bz = 256$. dCor, though simpler, still exhibits cubic growth. SoftCCA and I_dCor: SoftCCA's runtime plateaus at higher el due to kernel-based computation but remains significantly higher than OptFast HGR. I_dCor is the slowest, with runtimes exceeding 0.2 seconds at $bz = 256, \text{el} = 500$.

4. **Batch Size Robustness**: At smaller batch sizes ($bz = 16$), OptFast HGR maintains the fastest execution times, while CKA and Soft-HGR show minimal batch size sensitivity but have higher absolute runtimes. As batch size increases, the performance gap between UniFast/OptFast HGR and baseline methods widens.

Figure 5 further compares execution times for a fixed dimension (dim $= 300$) across varying batch sizes. OptFast HGR consistently achieves low execution times with only a slight increase as batch size grows, highlighting its efficiency and robustness. UniFast HGR also shows relatively low execution times, increasing much more slowly than many baseline methods. For example, CCA's execution time rises significantly with increasing batch size, reflecting its high computational complexity and inefficiency in processing large-batch data. Methods like dCor and I_dCor show a rapid increase in execution time with batch size, underscoring their inefficiency in handling batch-size variations. In contrast, OptFast HGR and UniFast HGR show much flatter execution-time curves, emphasizing their stability and efficiency across different batch sizes.

The efficiency of UniFast and OptFast HGR arises prominently from **avoiding matrix decomposition**. Unlike CCA and I_dCor, which rely on computationally expensive SVD or PCA operations, these methods focus on the upper triangular part of the distribution matrix and utilize cosine similarity. This approach completely eliminates the need for those costly decomposition operations, drastically reducing computational complexity. These results demonstrate that UniFast and OptFast HGR offer substantial speedups over traditional correlation methods while maintaining accuracy, making them ideal for real-world multimodal tasks requiring high computational efficiency.



(a) dim = 300 (b) dim = 300(Zoomed-in View)

Figure 5: Execution time comparison across methods for fixed dimension = 300 with varying batch sizes

G Robustness to Real-World Challenges

To evaluate the robustness of UniFast HGR in practical scenarios, experiments were conducted across three challenging conditions: high noise, modality imbalance, and spurious correlations. Protocols for each dataset (IEMOCAP for audio noise, Flickr30K for modality imbalance, and Berlin for spurious correlations) align with prior work, and performance metrics follow established benchmarks.

High Noise Perturbation: Gaussian noise with a standard deviation of 30% was added to audio features in the IEMOCAP dataset. Performance was compared against CLIP-based fusion (relying

on text-audio alignment) and Soft-HGR. CLIP-based fusion achieved 65.8% accuracy, a 5.5% decrease from clean data. Soft-HGR dropped to 63.1% accuracy, an 8.2% reduction. UniFast HGR retained 70.2% accuracy, with only a 3.5% decline—outperforming CLIP and Soft-HGR by 4.4 and 7.1 percentage points, respectively.

Modality Imbalance: To simulate extreme label scarcity, only 10% of text labels were preserved for 99% of images in the Flickr30K dataset. CLIP (dependent on aligned text-image pairs) and Deep CCA (a classic multimodal method) were evaluated. CLIP achieved 62.3% Recall@1, limited by reliance on paired text. Deep CCA reached 59.7% Recall@1. UniFast HGR attained 68.9% Recall@1, exceeding CLIP and Deep CCA by 6.6 and 9.2 percentage points, respectively.

Spurious Correlations: To test resistance to misleading associations, 20% of training labels in the Berlin dataset were corrupted to introduce false "building→forest" mappings. Soft-HGR, which lacks mechanisms to mitigate self-correlation bias, overfit to spurious pairs and achieved 69.2% overall accuracy (OA). UniFast HGR, leveraging diagonal removal to focus on cross-modal dependencies, reached 77.3% OA—an 8.1 percentage point improvement over Soft-HGR.

Table 15: Performance under real-world challenges

Scenario	Method	Metric	Value
High Noise (IEMOCAP)	CLIP-based fusion	Accuracy (%)	65.8
	Soft-HGR	Accuracy (%)	63.1
	UniFast HGR	Accuracy (%)	70.2
Modality Imbalance (Flickr30K)	CLIP	Recall@1 (%)	62.3
	Deep CCA	Recall@1 (%)	59.7
	UniFast HGR	Recall@1 (%)	68.9
Spurious Correlations (Berlin)	Soft-HGR	OA (%)	69.2
	UniFast HGR	OA (%)	77.3

These results demonstrate that UniFast HGR maintains strong performance under noise, label scarcity, and spurious correlations. Its design—via variance constraints, diagonal removal, and cosine similarity—effectively prioritizes robust cross-modal relationships, outperforming baselines in challenging real-world settings.

H Discussion

The proposed methods offer several significant advancements for multimodal feature extraction and related applications. First, they provide a more efficient and stable approach for extracting relevant features from multimodal data. The UniFast HGR method reduces computational complexity from $O(K^3)$ to $O(m^2K)$ while improving convergence speed, making it well-suited for large-scale datasets and real-time applications. Second, its capacity to integrate multiple modes increases its flexibility and applicability across various multimodal scenarios, enabling it to handle datasets with diverse patterns. Furthermore, the OptFast HGR approach is optimized by reducing the number of normalization steps, achieving a level of efficiency and computational cost comparable to dot product and cosine similarity operations.

The three core innovations—cosine similarity substitution, diagonal removal, and simplified variance constraints—collectively address longstanding computational bottlenecks in HGR maximal correlation estimation. These advancements enable the framework to scale effectively to high-dimensional features (e.g., $K = 1024$ in vision transformers) while maintaining robustness to real-world challenges such as noise, modality imbalance, and spurious correlations.

Overall, the results indicate that the improved methods not only enhance computational efficiency but also maintain competitive performance in both image classification and multimodal emotion recognition tasks. These attributes position UniFast HGR and OptFast HGR as promising approaches for multimodal feature extraction in a range of applications, from large-scale remote sensing to resource-constrained edge computing scenarios.