Cross-View Distillation and Adaptive Masking for Incomplete Multi-View Multi-Label Classification

Supplementary Material

Table 4. Details about six multi-view multi-label datasets.

Dataset	# Sample	# Label	# Label/#Sample
Corel5k	4999	260	3.40
Iaprtc12	19627	291	5.72
Espgame	20770	268	4.69
Pascal07	9963	20	1.47
Mirflickr	25000	38	4.72
ODIR	10000	8	1.05

7. Details of Six Datasets

In this section, we provide a detailed description of the six datasets used in our experiments. Tab. 4 summarizes the statistics for each dataset, including the number of samples and labels. Notably, the Corel5k, Pascal07, Iaprtc12, Mirflickr, and Espgame datasets each contain data from six views, namely: GIST, HSV, DenseHue, DenseSIFT, RGB, and LAB, while the ODIR dataset contains five views: RGB, LAB, VGG, CLIP, and Wavelet.

8. Details of Comparison Methods

In this section, we provide detailed descriptions of the baseline methods used for comparison. The baselines cover both traditional and deep learning models. These methods are specifically designed for the dual incompleteness setting (missing views and labels).

The iMVWL [30] approach is a typical standard method. It combines two weighted matrix factorization (MF) models into one structure. The first MF model learns a common representation from all views of the incomplete multi-view data. The second MF model gets a classification mapping matrix to keep local label relationships from incomplete multi-label data. In addition, a low-rank restriction is placed on the label relationship matrix.

NAIM3L [15] builds an index matrix and uses it to learn a consistent representation across different views. It learns consistent representations and trains the classifier at the same time within one framework. However, NAIM3L does not look at possible representation degeneration and lacks a dynamic fusion at the sample level.

Inspired by the remarkable success of deep learning, DIMC [39] is the first deep learning method designed for the iM3C problem. The key advance of DIMC is the use of two missing indicator matrices, one for views and another for labels, which mitigates the negative impact of dual incompleteness issue by explicitly modeling the missing patterns. The method is suitable for both supervised and semi-supervised learning situations.

DICNet [17] builds an end-to-end multi-view feature extraction architecture based on stacked autoencoders. Its main contribution is a novel incomplete instance-level contrastive learning strategy. Guided by the multi-view consensus assumption, this strategy drives the model to maximize consistency across different views, thereby facilitating the extraction of multi-view consensus information.

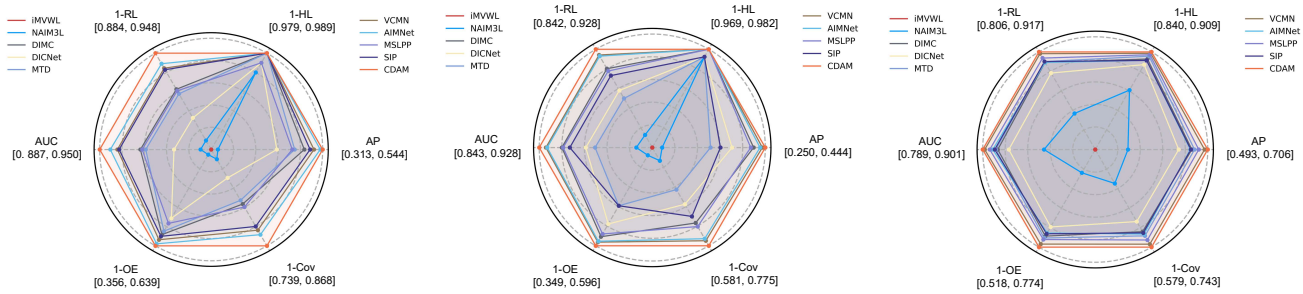
To fully leverage the complementary information in multi-view data, MTD [16] leverages a dual-channel encoder to disentangle multi-view data into shared and private representations, a process driven by a novel cross-channel contrastive loss. To further improve feature learning, it also introduces a Random Patch Masking strategy to reduce input redundancy.

Inspired by the Mixer architecture, VCMN [5] is designed for both structural simplicity and efficient information mixing. It departs from the conventional multi-stage pipeline by integrating view-specific feature learning and cross-view interaction into a single, unified framework. This allows the model to concurrently perform high-level feature abstraction within each view and information exchange across them.

An alternative strategy is to address data missingness through view imputation, where the primary goal is to reconstruct the missing views from available information before proceeding to the downstream task. AIMNet [18] introduces a graph attention-induced imputation to approximate missing features in the latent space. Critically, it reuses the guiding attention scores as confidence weights for a dynamic late fusion stage, enhancing the reliability of the final prediction.

To maintain the data's original topological structure, MSLPP [22] is designed to fuse two distinct geometric perspectives. It maps the neighborhood structures from the Euclidean space (distance-based) and the cosine space (similarity-based) into a common subspace. The model then employs a regularization term to ensure that both distance and similarity information are jointly retained in the learned representation.

Recognizing the redundancy inherent in multi-view data, SIP [19] model operates on the premise that the shared semantics among views are all that's needed for label prediction. It thus frames the learning problem within the Information Bottleneck theory, striving to distill a shared rep-



(a) Corel5k (b) Iaprtc12 (c) Mirflickr

Figure 6. Performance comparison on complete datasets (i.e., 0% missing views and 0% missing labels). The radar charts compare CDAM against state-of-the-art baselines on (a) Corel5k, (b) Iaprtc12, and (c) Mirflickr across six evaluation metrics. The results demonstrate the robustness of CDAM, showing competitive performance even in the fully-observed setting.

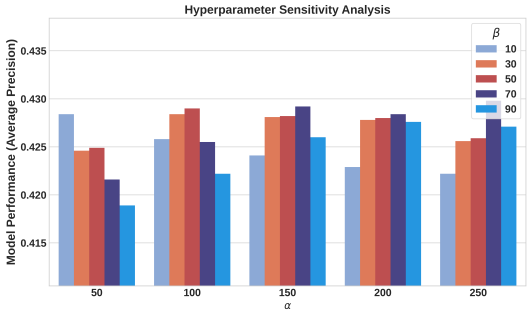
resentation that is both predictive and maximally compact. This involves a dual objective: maximizing the captured shared information while purging any view-specific, non-shared content. Uniquely, SIP also learns to model multi-label prototypes in a data-driven fashion directly in this distilled latent space.

9. Extra Experimental Results on complete datasets

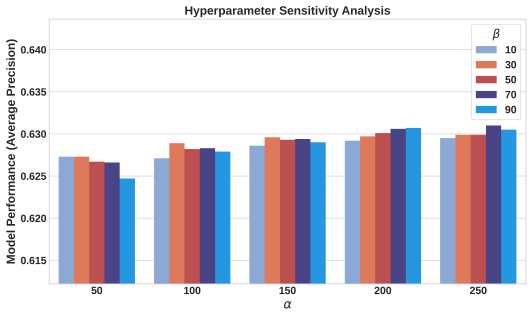
In this section, we further evaluate the performance of our CDAM in the complete data setting. As shown in Fig. 6, we compared CDAM with several baseline models in the Corel5k, Iaprtc12, and Mirflickr datasets, presenting the results in a radar chart. The results clearly indicate that CDAM achieves highly competitive performance even in this ideal scenario with no missing data. This provides strong evidence that our model is a versatile framework and that its mechanisms designed for handling incomplete data do not compromise its effectiveness on complete data.

10. Additional hyperparameter analysis

In this section, we conduct a further grid search for hyperparameters on additional datasets, presenting the results in Fig. 7. We observe that although the optimal hyperparameter configurations vary slightly across different datasets, the pattern of their impact on model performance is highly consistent. Specifically, we find that model performance is significantly hampered when α is too small, and the optimal choice of β is strongly dependent on the value of α.



(a) Corel5k



(b) Mirflickr

Figure 7. Hyperparameter sensitivity analysis for α (reconstruction loss weight) and β (distillation loss weight). The plots illustrate the impact of varying α and β on the final model performance (AP) for (a) Corel5k and (b) Mirflickr.