

Supplementary Material: Robust Variational Contrastive Learning for Partially View-unaligned Clustering

Changhao He
hechanghao.gm@gmail.com
Sichuan University
Chengdu, China

Peng Hu*
penghu.ml@gmail.com
Sichuan University
Chengdu, China

Hongyuan Zhu
hongyuanzhu.cn@gmail.com
Institute for Infocomm Research, A*STAR
Singapore

Xi Peng
pengx.gm@gmail.com
Sichuan University
Chengdu, China

1 Introduction

In this supplementary material, we first elaborate on the mathematical notations, dataset details, and necessary derivations for the main paper. In addition, to further validate the effectiveness of our VITAL, we conduct some additional experiments.

2 Notations

We present the primary mathematical symbols and their descriptions from the main paper in Table 1.

3 Datasets

Eight widely used multi-view datasets were used to validate the effectiveness of VITAL in both partially and fully aligned scenarios. Their details are as follows:

- CUB [17]: This dataset consists of various categories of birds and we utilized the first 10 categories. Deep visual features from GoogLeNet and text features using doc2vec [12] were employed as two views.
- Scene-15 [4]: This dataset consists of 4485 indoor and outdoor images, with a total of 15 categories. We extracted GIST and PHOG features as two views.
- Wiki [2]: This dataset consists of 2866 image-text pairs selected from Wikipedia articles, with a median text length of 200 words, covering 10 different categories.
- NUS-WIDE [1]: This dataset consists of approximately 270,000 images, divided into 81 categories based on labels. We only selected images from the top ten largest categories for experiments, resulting in 9000 image-text pairs used in the experiments.
- Deep Animal [11]: This dataset consists of 10,158 images from 50 categories. We ultimately utilized two deep image features extracted from DECAF [10] and VGG19 [15] as two views.
- Deep Caltech-101 [3]: This dataset consists of 101 categories and 1 background scene category, with each category having 40 to 800 images, totaling 9146 images. We selected 8677 images excluding the background scene category, and similar to the processing of the Deep Animal dataset, we extracted two deep image features using DECAF [10] and VGG19 [15] as two views.

Table 1: The primary set of mathematical symbols used in the main paper.

Symbol	Description
\mathcal{X}	Multi-view dataset
X^k	The sample set of the k -th view
x_i^k / z_i^k	The i -th sample/latent variable of the k -th view
N	Total number of instances
V	Total number of views
T_{inter}/T_{intra}	Ground truth of inter-/intra- view
p_{Θ}	True posterior distribution
q_{ϕ}	Recognition model (probabilistic encoder)
q_{θ}	Generative model (probabilistic decoder)
\mathcal{H}	Cross-entropy loss
I	Identity matrix
α	Sensitive parameter
l	Pair loss
w	Soft label of pair
B	Batch size
γ_k	Mixture coefficient of the k -th component
ϕ_k	Probability density of the k -th component
d	Confidence (distance between two components)
Q_d	Quantile
M_{GMM}	Positive set obtained by GMM model
$M_{confidence}$	Positive set obtained by M_{GMM} and Q_d
M	Confidence mask
C	Common representations
Σ	Specific representations

- MNIST-USPS [14]: This dataset consists of two datasets, MNIST [13] and USPS [7]. We treated them as two views and randomly selected 5000 pairs for experiments.
- NoisyMNIST [18]: This dataset was derived from MNIST [13] by randomly rotating and adding random noise, resulting in a total of 70,000 clean/noisy sample pairs. Considering the large size of this dataset and that some baselines may consume too much time, we randomly selected 30,000 pairs of samples for experiments.

*Corresponding author

4 Derivation Details

In this section, we will elaborate on the detailed derivation of the evidence lower bound (ELBO) function \mathcal{L}_{ELBO} and the Kullback-Leibler (KL) divergence loss \mathcal{L}_{KL} , corresponding to Equations (2), (3) and (4) in the main paper. Additionally, we provide detailed explanations of the optimization processes for \mathcal{L}_{intra} and \mathcal{L}_{inter} .

4.1 Derivation of \mathcal{L}_{ELBO}

Since in multi-view datasets, various views of an instance may serve as mutual priors, the objective is to minimize the Kullback-Leibler (KL) divergence between the approximate and true posterior across all views:

$$\begin{aligned}
& \sum_{m=1}^V \sum_{n=1}^V D_{KL}(q_\phi(z_i|x_i^m)||p_\Theta(z_i|x_i^n)) \\
&= \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \log \frac{q_\phi(z_i|x_i^m)}{p_\Theta(z_i|x_i^n)} dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \log \frac{q_\phi(z_i|x_i^m)p_\Theta(x_i^n)}{p_\Theta(z_i, x_i^n)} dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \left[\log \frac{q_\phi(z_i|x_i^m)}{p_\Theta(z_i, x_i^n)} + \log p_\Theta(x_i^n) \right] dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \log p_\Theta(x_i^n) dz \\
&\quad + \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \log \frac{q_\phi(z_i|x_i^m)}{p_\Theta(z_i, x_i^n)} dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \log p_\Theta(x_i^n) + \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \log \frac{q_\phi(z_i|x_i^m)}{p_\Theta(x_i^n|z_i)p_\Theta(z_i)} dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \log p_\Theta(x_i^n) \\
&\quad + \sum_{m=1}^V \sum_{n=1}^V \int_z q_\phi(z_i|x_i^m) \left[\log \frac{q_\phi(z_i|x_i^m)}{p_\Theta(z_i)} - \log p_\Theta(x_i^n|z_i) \right] dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \log p_\Theta(x_i^n) + \sum_{m=1}^V \sum_{n=1}^V D_{KL}(q_\phi(z_i|x_i^m)||p_\Theta(z_i)) \\
&\quad - \sum_{m=1}^V \sum_{n=1}^V \mathbb{E}_{q_\phi(z_i|x_i^m)} [\log p_\Theta(x_i^n|z_i)] \\
&= \sum_{m=1}^V \sum_{n=1}^V \log p_\Theta(x_i^n) + \mathcal{L}_{ELBO},
\end{aligned} \tag{1}$$

which finishes the derivation of Equation (2) to Equation (3) in the main paper. It is worth noting that Equation (1) is a multi-view version of the ELBO function, which is different from that in [9], as the latter can only handle single-view data.

4.2 Derivation of \mathcal{L}_{KL}

The KL divergence loss is used to minimize the distance between the standard Gaussian distribution $p_\Theta(z_i) \sim \mathcal{N}(z_i; \mathbf{0}, I)$ and the

approximate posterior $q_\phi(z_i|x_i^m) \sim \mathcal{N}(z_i; \mu_i^m, (\sigma_i^m)^2 I)$. Followed by [9], \mathcal{L}_{KL} can be computed as follows:

$$\begin{aligned}
\mathcal{L}_{KL} &= \sum_{m=1}^V \sum_{n=1}^V D_{KL}(q_\phi(z_i|x_i^m)||p_\Theta(z_i)) \\
&= \sum_{m=1}^V \sum_{n=1}^V \int q_\phi(z_i|x_i^m) \log \left(\frac{q_\phi(z_i|x_i^m)}{p_\Theta(z_i)} \right) dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int q_\phi(z_i|x_i^m) \log \left[\frac{\frac{1}{\sqrt{2\pi}(\sigma_i^m)^2} e^{-\frac{(z_i-\mu_i^m)^2}{2(\sigma_i^m)^2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}}} \right] dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int q_\phi(z_i|x_i^m) \log \left\{ \frac{1}{\sqrt{(\sigma_i^m)^2}} e^{\frac{1}{2} \left[z_i^2 - \frac{(z_i-\mu_i^m)^2}{(\sigma_i^m)^2} \right]} \right\} dz \\
&= \sum_{m=1}^V \sum_{n=1}^V \int q_\phi(z_i|x_i^m) \left[-\log(\sigma_i^m)^2 + z_i^2 - \frac{(z_i-\mu_i^m)^2}{(\sigma_i^m)^2} \right] dz.
\end{aligned} \tag{2}$$

According to the first moment of the Gaussian distribution, $\mathbb{E}(z_i) = \mu_i^m$, and the second moment, $\mathbb{E}(z_i^2) = (\mu_i^m)^2 + (\sigma_i^m)^2$, Equation (2) can be finally simplified as:

$$\mathcal{L}_{KL} = \sum_{m=1}^V \sum_{n=1}^V \frac{1}{2} (-\log(\sigma_i^m)^2 + (\mu_i^m)^2 + (\sigma_i^m)^2 - 1), \tag{3}$$

which finishes the derivation of Equation (4) in the main paper.

4.3 Optimization processes of \mathcal{L}_{intra} and \mathcal{L}_{inter}

According to the second term in \mathcal{L}_{ELBO} , we can obtain the expectation forms regarding intra- and inter- view as follows:

$$\begin{aligned}
& - \sum_{m=1}^V \sum_{n=1}^V \mathbb{E}_{q_\phi(z_i|x_i^m)} [\log p_\Theta(x_i^n|z_i)] \\
&= - \underbrace{\sum_{k=1}^V \mathbb{E}_{q_\phi(z_i^k|x_i^k)} [\log p_\Theta(x_i^k|z_i^k)]}_{\mathcal{L}'_{intra}} - \underbrace{\sum_{m \neq n}^V \mathbb{E}_{q_\phi(z_i^m|x_i^m)} [\log p_\Theta(x_i^n|z_i^m)]}_{\mathcal{L}'_{inter}}.
\end{aligned} \tag{4}$$

In the optimization process of the two terms mentioned above, due to the typically high variance exhibited by the naive Monte Carlo gradient estimator, we follow the approach in VAE [9] and use a Stochastic Gradient Variational Bayes (SGVB) estimator instead:

$$\mathcal{L}'_{intra} \simeq -\frac{1}{L} \sum_{l=1}^L \sum_{k=1}^V \log p_\Theta(x_i^k|z_{(i,l)}^k), \tag{5}$$

and

$$\mathcal{L}'_{inter} \simeq -\frac{1}{L} \sum_{l=1}^L \sum_{m \neq n}^V \log p_\Theta(x_i^m|z_{(i,l)}^n), \tag{6}$$

where L is the number of Monte Carlo samples in the SGVB estimator and is set to 1 in our experiment, $z_{(i,l)}^k$ and $z_{(i,l)}^n$ are the results

Algorithm 1 Robust Variational Contrastive Learning for Partially View-unaligned Clustering

```

1: Input: partially view-unaligned dataset  $\{X^k\}_{k=1}^2$ ;  $VCL\ epoch = 100$ ;  $max\ epoch = 110$ ; the recognition model  $\{q_\phi^k\}_{k=1}^2$ ; the generative model  $\{q_\theta^k\}_{k=1}^2$ .
2: for  $epoch = 1$  to  $max\ epoch$  do
3:   for sampled batch  $x$  do
4:     if  $epoch \leq VCL\ epoch$  then
5:       Compute the overall loss  $\mathcal{L} = \mathcal{L}_{inter} + \mathcal{L}_{intra} + \mathcal{L}_{KL}$  by Equation (10), Equation (8), Equation (4) in the main paper, respectively.
6:     else
7:       Compute the  $\mathcal{L}_{inter}$  by the vanilla contrastive loss.
8:       Fit the loss value set through a two-component Gaussian Mixture model.
9:       Compute the confidence quantile  $Q_d$  by Equation (15) in the main paper.
10:      Filter the confidence subset of GMM output by  $Q_d$ .
11:      Compute the  $\mathcal{L} = \mathcal{L}_{inter-DR} + \mathcal{L}_{intra} + \mathcal{L}_{KL}$  by Equation (17), Equation (8), Equation (4), respectively.
12:    end if
13:    Update  $\{q_\phi^k\}_{k=1}^2$  and  $\{q_\theta^k\}_{k=1}^2$ .
14:  end for
15: end for ▷ Training
16: for sampled batch  $x$  do
17:   Forward  $x$  through the recognition model  $\{q_\phi^k\}_{k=1}^2$  and obtain  $\{C^k\}_{k=1}^2$  and  $\{\Sigma^k\}_{k=1}^2$ .
18:   Compute the cross-view Euclidean distance matrix  $D$  of  $\{C^k\}_{k=1}^2$ .
19:   for  $x_i^1$  in  $x$  do
20:     Realign it with its category-level counterpart  $x_j^2$  through  $j = \operatorname{argmin}_{j \neq i} D_{ij}$ .
21:   end for
22: end for ▷ Inference
23: Output: Apply k-means on the fused results  $[C^k + \Sigma^k]_{k=1}^2$ .

```

obtained through the reparameterization trick [9]:

$$\begin{aligned} z_{(i,l)}^k &= \mu_i^k + \sigma_i^k \times \epsilon_l^k, & \epsilon_l^k &\sim \mathcal{N}(0, I) \\ z_{(i,l)}^n &= \mu_i^n + \sigma_i^n \times \epsilon_l^n, & \epsilon_l^n &\sim \mathcal{N}(0, I), \end{aligned} \quad (7)$$

where \times denotes the element-wise product. The purpose of transforming the process of computing expectations into a sampling process is to enable gradient descent updates on the parameters. So Equation (5) finally exhibits a typical cross-entropy form:

$$\mathcal{L}_{intra} = \sum_{k=1}^V \mathcal{H}(T_{intra}, p(x_i^k, \hat{x}_i^k)), \quad (8)$$

where $\hat{x}_i^k = q_\theta(z_{(i,l)}^k)$, \mathcal{H} denotes cross-entropy, T_{intra} represents the intra-view ground truth, and $p(\cdot, \cdot)$ is the likelihood between two points and can be computed according to Equation (9) in the main paper.

However, for Equation (6), directly computing the expectation through sampling is not feasible due to the inconsistency of cross-view variables. On the other hand, we observe that since the variance of each view is modeled as specific information, the variances of different views are independent of each other. Take another look at $q_\phi(z_i^m | x_i^m) \sim \mathcal{N}(\mu_i^m, (\sigma_i^m)^2 I)$ and $q_\phi(z_i^n | x_i^n) \sim \mathcal{N}(\mu_i^n, (\sigma_i^n)^2 I)$, as both μ_i^m and μ_i^n respectively represent the location of x_i^m and x_i^n in the semantic space, \mathcal{L}'_{inter} can be relaxed to an optimization form that only involves the means:

$$\mathcal{L}_{inter} = \sum_{m \neq n}^V \mathcal{H}(T_{inter}, p(\mu_i^m, \mu_i^n)), \quad (9)$$

where T_{inter} represents the inter-view ground truth. Based on the method proposed in [5], both terms in Equation (8) and Equation (9) can be regarded as contrastive loss, which is essentially a contrastive learning process. Significantly, for \mathcal{L}_{intra} , we adhere to the calculation format of contrastive loss because it encapsulates the semantics within the view, thus effectively serving as a reconstruction term. However, for \mathcal{L}_{inter} , due to the generation of a large number of FNPs in the category-level contrast process, we have accordingly modified it. The specific process is detailed in section 3.4 of the main paper.

5 Additional Experiments

5.1 Network Architecture

For the probability encoder, we employed a four-layer fully connected network (FCN) followed by a batch normalization layer [8] and a rectified linear unit (ReLU) activation function at the end of each layer. To prevent overfitting, we also added a Dropout layer [16] after the activation function with a probability value set to 0.2 to enhance generalization. As for the probability decoder, we employed a symmetric four-layer fully connected network (FCN), with each layer followed by a ReLU activation function. Additionally, since we preprocessed the raw data with min-max normalization before feeding all data into the neural network, we used a Sigmoid activation function in the adaptation layer of the decoder to normalize the decoder output to the range of 0 to 1, corresponding to the calculation of \mathcal{L}_{intra} . We applied this architecture to all datasets, and the overall architecture is illustrated in Table 2.

Table 2: The architecture of the probabilistic encoder/decoder. $\dim^{(k)}$ denotes the dimension of input data of the k -th view.

Encoder
Linear($\dim^{(k)}$, 1024), BatchNorm, ReLU, Dropout(0.2)
Linear(1024, 1024), BatchNorm, ReLU, Dropout(0.2)
Linear(1024, 1024), BatchNorm, ReLU, Dropout(0.2)
Linear(1024, 256)
Decoder
Linear(128, 1024), ReLU
Linear(1024, 1024), ReLU
Linear(1024, 1024), ReLU
Linear(1024, $\dim^{(k)}$), Sigmoid

Table 3: Ablation studies on \mathcal{L}_{VCL} . ✓ denotes the adoption of the component, while ✗ indicates its exclusion. The best results are indicated in bold.

Aligned	\mathcal{L}_{inter}	\mathcal{L}_{intra}	\mathcal{L}_{KL}	CUB			NUS-WIDE			Deep Animal			Deep Caltech-101		
				ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Partially	✓	✗	✗	67.43	60.95	48.36	58.49	44.93	37.64	38.70	47.08	25.57	36.30	62.61	23.67
	✓	✓	✗	75.53	70.05	59.85	58.52	45.11	37.88	41.80	48.17	27.45	41.79	66.55	26.87
	✓	✓	✓	77.83	74.40	64.25	59.77	45.58	38.60	42.99	48.01	28.42	45.09	70.17	31.62
fully	✓	✗	✗	71.50	64.31	52.18	52.28	42.40	29.83	45.94	57.15	34.32	43.04	70.65	28.8
	✓	✓	✗	79.23	72.53	62.10	59.69	48.34	40.15	47.87	58.69	36.25	47.98	74.15	31.87
	✓	✓	✓	84.70	79.34	71.72	63.04	51.42	43.90	54.02	62.58	42.28	54.06	78.46	36.84

5.2 Algorithm

The training phase of our algorithm consists of two steps: i) training using the variational contrastive learning loss \mathcal{L}_{VCL} , and ii) dynamic training using the dynamic rectification version of the variational contrastive learning loss \mathcal{L}_{VCL-DR} . In the inference phase, we first forward the test data through the probabilistic encoder to obtain the common representations $\{C^k\}_{k=1}^V$ and specific representations $\{\Sigma^k\}_{k=1}^V$. Then, we realign the embedding features through $\{C^k\}_{k=1}^V$, and finally $[C^k + \Sigma^k]_{k=1}^V$ are served as the fused results. The overall algorithm is depicted in Algorithm 1.

5.3 Category-level Alignment Rate

In addition to the three widely used metrics in the main paper, namely ACC, NMI, and ARI, we introduce another metric called Category-level Alignment Rate (CAR) [19] to quantify the performance of realignment, which can be computed as follows:

$$CAR = \frac{1}{N} \sum_{i=1}^N \delta(C(x_i^{k_1}, \hat{x}_i^{k_2})), \quad (10)$$

where δ is the Dirichlet function, $\hat{x}_i^{k_2}$ is the realigned counterpart of $x_i^{k_1}$, and $C(a, b)$ is an indicator function evaluating to 1 i.f.f. a and b belong to the same category. We present the CAR performance of VITAL and SURE (the state-of-the-art method for addressing PVP) on all datasets in Figure 1. Intuitively, for a given cross-view pair, the random probability of correctly aligning at category-level is $\frac{1}{C}$, where C is the number of categories. When there are too many categories, this makes category-level alignment challenging [19, 20]. However, from Figure 1, we observe that on the Deep Animal dataset (which contains 50 categories) and the Deep Caltech-101 dataset (which contains 101 categories), VITAL outperforms SURE by 127.30% and 100.64% respectively. This is because SURE utilizes cross-modal reconstruction to retain intrinsic semantics in the common representation, which may disrupt common information and affect alignment effectiveness. In contrast to SURE, VITAL explicitly separates common information and specific information in its modeling, which results in a better capture of category attributes compared to SURE, thus enhancing category-level alignment rate.

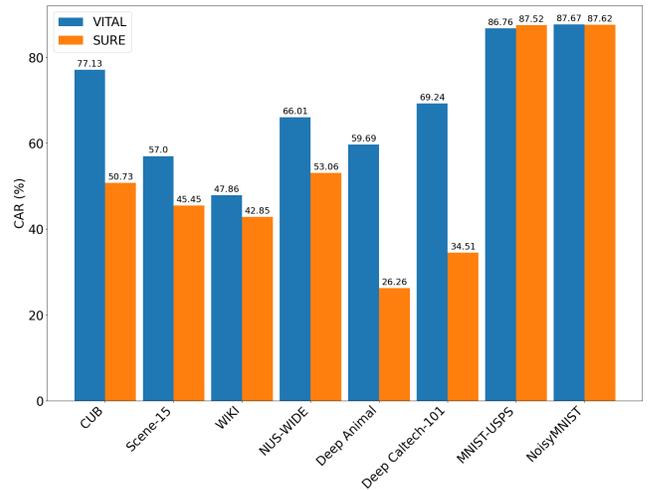


Figure 1: CAR performance of SURE and VITAL on eight widely used datasets in partially aligned (50%) scenario.

5.4 Ablation Studies

The ablation studies are conducted on \mathcal{L}_{VCL} to reveal the effectiveness of its components. To be specific, \mathcal{L}_{VCL} can be written as follows according to the main paper:

$$\mathcal{L}_{VCL} = \mathcal{L}_{inter} + \mathcal{L}_{intra} + \mathcal{L}_{KL}. \quad (11)$$

We present the ablation studies for the four datasets (CUB, NUS-WIDE, Deep Animal, Deep Caltech-101) in partially aligned (50%) and fully aligned scenarios in Table 3. According to the table, we observe that in the partially aligned scenario, removing the component \mathcal{L}_{KL} , which means imposing no constraints on the approximate posterior and true posterior, results in an average decrease of 2.01%, 2.07%, and 2.71% in ACC, NMI, and ARI, respectively. This phenomenon is more pronounced in the fully aligned scenario. On the other hand, removing the component \mathcal{L}_{intra} , which means retaining no view-specific information, leads to an average decrease of 4.18%, 3.58%, and 4.20% in ACC, NMI, and ARI, respectively. These results demonstrate the significant contribution of each component to the final results, validating the effectiveness of each component in variational contrastive learning.

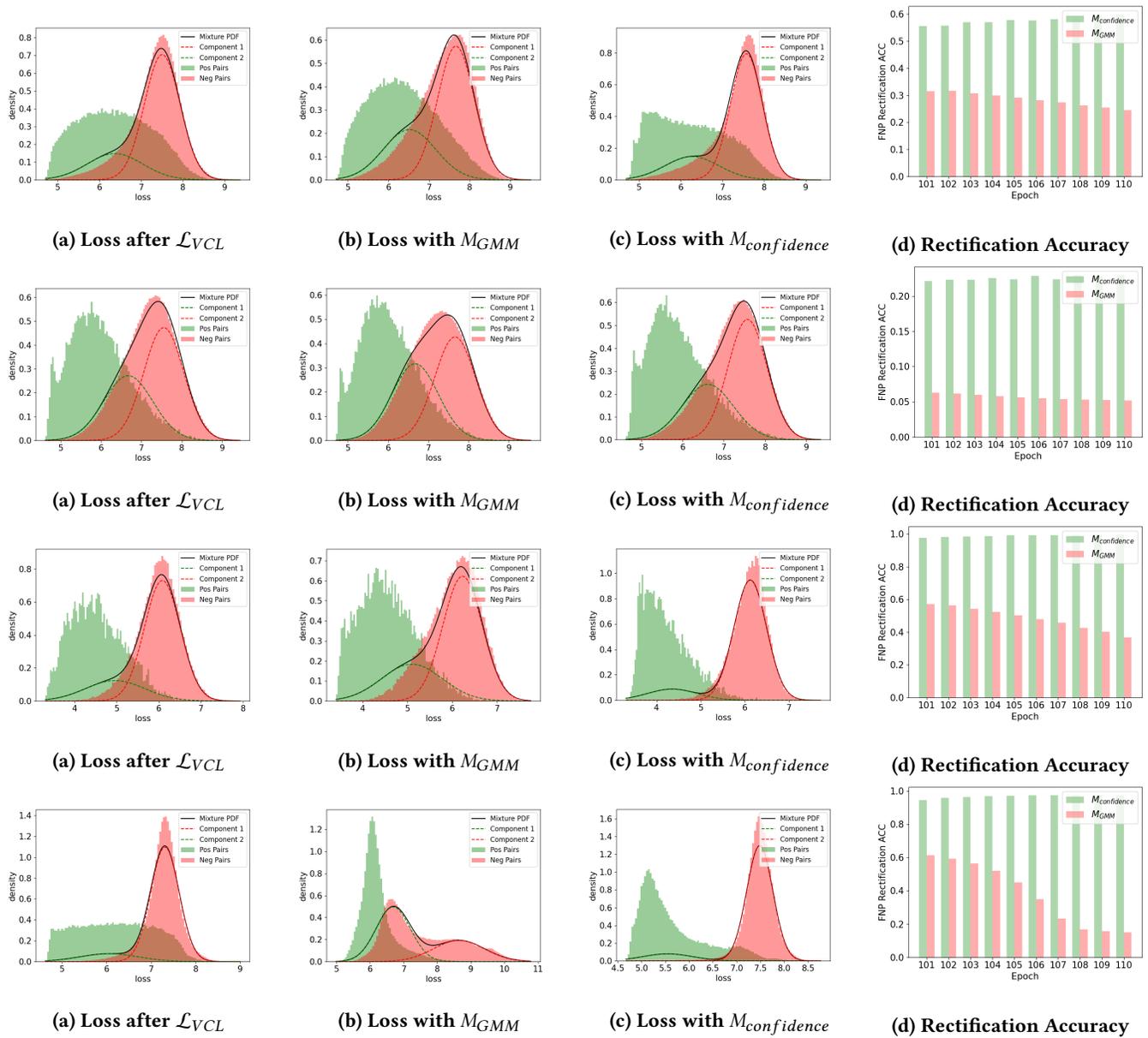


Figure 2: From top to bottom, each row represents the rectification visualization for NUS-WIDE, Deep Animal, MNIST-USPS, and NoisyMNIST, respectively. (a) Loss density plot of positive and negative pairs after training with \mathcal{L}_{VCL} . (b) Loss density plot of positive and negative pairs after rectification using M_{GMM} directly (threshold 0.5). (c) Loss density plot of positive and negative pairs after rectification using the confidence subset $M_{confidence}$ of M_{GMM} . (d) Variation of FNP rectification accuracy during the rectification process using M_{GMM} and $M_{confidence}$.

5.5 Rectification Visualization

In the partially aligned scenario (50%), we present rectification visualizations for four additional datasets, as shown in Figure 2. Similar to Figure 2 in the main paper, (a) demonstrates that considerable overlap still exists between positive and negative pairs in the loss density plot after training with \mathcal{L}_{VCL} . A direct solution is to fit the loss with a two-component Gaussian mixture model

and then separate positive and negative pairs by setting a threshold on the posterior probability for each component (usually 0.5 [6]). However, the overlap between the two components may contain a large number of confounding pairs that cannot be correctly assigned. To address this issue, we propose a confidence computation method specific to each fitting result, rather than directly using the fitting result. Detailed computational procedures are provided in

the main paper. Utilizing the confidence subset of M_{GMM} , denoted as $M_{confidence}$, can make the partition result cleaner to some extent, thus reducing the occurrence of False Negative Pairs (FNPs). This can be verified by the results in (d), which show that the FNP rectification accuracy of $M_{confidence}$ is consistently higher than that of the original partition result M_{GMM} .

References

- [1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*. 1–9.
- [2] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 521–535. <https://doi.org/10.1109/TPAMI.2013.142>
- [3] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*. IEEE, 178–178.
- [4] Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 524–531.
- [5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
- [6] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with Noisy Correspondence for Cross-modal Matching. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 29406–29419. https://proceedings.neurips.cc/paper_files/paper/2021/file/f5e62af885293cf4d511cef31e61c80-Paper.pdf
- [7] J.J. Hull. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 5 (1994), 550–554. <https://doi.org/10.1109/34.291440>
- [8] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. pmlr, 448–456.
- [9] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1312.6114>
- [10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
- [11] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 453–465. <https://doi.org/10.1109/TPAMI.2013.140>
- [12] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.
- [13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. <https://doi.org/10.1109/5.726791>
- [14] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. 2019. COMIC: Multi-view Clustering Without Parameter Selection. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, Long Beach, California, USA, 5092–5101.
- [15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
- [17] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).
- [18] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*. PMLR, 1083–1092.
- [19] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2023. Robust Multi-View Clustering With Incomplete Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2023), 1055–1069. <https://doi.org/10.1109/TPAMI.2022.3155499>
- [20] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. 2021. Partially View-Aligned Representation Learning With Noise-Robust Contrastive

Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1134–1143.