Safe Multi-View Deep Classification

Wei Liu[1], **Yufei Chen**[1 *], **Xiaodong Yue**[2,3,5], **Changqing Zhang**[4], **Shaorong Xie**[2]

[1] College of Electronics and Information Engineering, Tongji University, Shanghai, China
[2] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[3] Artificial Intelligence Institute of Shanghai University, Shanghai, China
[4] College of Intelligence and Computing, Tianjin University, Tianjin, China
[5] VLN Lab, NAVI MedTech Co., Ltd. Shanghai, China
ldachuan@outlook.com, yufeichen@tongji.edu.cn, yswantfly@shu.edu.cn, zhangchangqing@tju.edu.cn, srxie@shu.edu.cn

Abstract

Multi-view deep classification expects to obtain better classification performance than using a single view. However, due to the uncertainty and inconsistency of data sources, adding data views does not necessarily lead to the performance improvements in multi-view classification. How to avoid worsening classification performance when adding views is crucial for multi-view deep learning but rarely studied. To tackle this limitation, in this paper, we reformulate the multi-view classification problem from the perspective of safe learning and thereby propose a Safe Multi-view Deep Classification (SMDC) method, which can guarantee that the classification performance does not deteriorate when fusing multiple views. In the SMDC method, we dynamically integrate multiple views and estimate the inherent uncertainties among multiple views with different root causes based on evidence theory. Through minimizing the uncertainties, SMDC promotes the evidences from data views for correct classification, and in the meantime excludes the incorrect evidences to produce the safe multi-view classification results. Furthermore, we theoretically prove that in the safe multi-view classification, adding data views will certainly not increase the empirical risk of classification. The experiments on various kinds of multi-view datasets validate that the proposed SMDC method can achieve precise and safe classification results.

Introduction

In real-world scenarios, such as image analysis, computing vision, data mining and multimedia, the same object can be represented by multiple different modalities or multiple types of features, known as multi-view data (Xu, Tao, and Xu 2013), which promotes multi-view learning to design advanced methods of combining multiple views to achieve the performance improvement. Recently, joining the success of deep learning, multi-view deep learning, which aims to learn a shared representation of multiple information from different types of views with deep neural networks (DNNs) (Bachman, Hjelm, and Buchwalter 2019; Sun, Dong, and Liu 2020), has become an important research direction.

While multi-view deep learning shows excellent power in practice, theoretical safeness guarantee (*performance without degradation*) of existing multi-view deep learning is limited. Existing multi-view learning algorithms generally treat an equal value for different views or assign/learn a fixed weight for each view. However, in real-world applications, not all information from each view can contribute a good representation due to the unknown and complex correlation among different views. Many works show that sometimes the use of multiple views may degenerate the performance (Bickel and Scheffer 2004; Yang et al. 2012). This problem is more critical in multi-view deep learning due to the overconfident incorrect predictions of DNNs (Lakshminarayanan, Pritzel, and Blundell 2017), which makes the fusion result of multiple views from uncertain and inconsistent data sources unsafe. Taking the Figure 1 as example, which shows the empirical error on real-world multi-view dataset Food-101 (Wang et al. 2015b) with deep learning. We can find the image empirical error increases with the coming of the abnormal text view that contains unknown information described by error code. Such phenomena undoubtedly go against the expectation of multi-view deep learning and limit its effectiveness in a large of practical tasks, particularly safety-critical applications (e.g., medical diagnosis or autonomous driving). Thus, it is vital to have a safe multi-view learning algorithm, whose performance is never significantly worse when fusing multiple views.

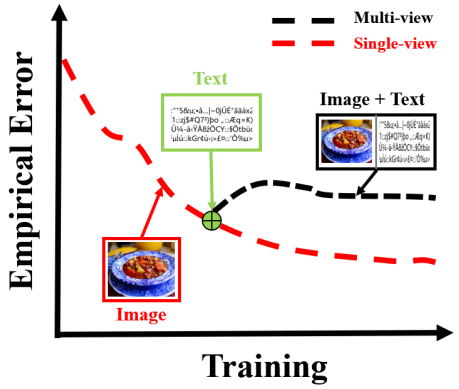

Figure 1: Empirical error for multi-view and single-view classification results with deep learning.

Though there are already many studies on multi-view learning, little work has been done explicitly about its safe-

*Corresponding author

ness. Hou et al. (Hou, Zeng, and Hu 2018) designed a stable feature selection method to guarantee performance does not become worse with more views. Tang et al. (Tang and Liu 2022) proposed a safe deep clustering method to avoid degenerating the multi-view clustering performance. However, little work focuses on the safeness of multi-view deep classification, which motivates us to devise a safe multi-view deep classification (SMDC) method from the perspective of safe learning.

Safe learning that typically appears in black-box problems recently has received a lot of attention. In black-box problems, no explicit mathematical model of the safety constraint is available and the value of the safety constraint function can only be known after a solution has been evaluated. Safe learning deals with these problems that avoid, as much as possible, the evaluation of non-safe input points, which are solutions, policies, or strategies that cause an irrecoverable loss (i.e., life threat) (Amodei et al. 2016; Kim, Allmendinger, and López-Ibáñez 2020).

Considering the non-degradation multi-view performance as the safety constraint, we can reformulate the multi-view classification problem from the perspective of safe learning. More concretely, our goal is to utilize multiple views based on the DNNs to obtain a safe result which guarantees that the performance does not deteriorate when fusing multiple data views. Specifically, our model combines different views at an evidence level instead of feature or output level as done previously, which uses the evidence theory to model the distribution of the class probabilities, and then formulates the inherent uncertainties among multiple views with different root causes as learning results. Through minimizing the uncertainties, we can increase the evidence from correct class while avoid the evidence from incorrect class and thereby increase the safeness of multi-view deep learning results. In summary, the specific contributions of this paper are:

(1) We formulate the multi-view deep classification from the perspective of safe learning aiming to provide a safe decision in an effective way, which introduces a new paradigm in multi-view deep classification.

(2) We devise a safe multi-view deep classification model that integrates each view at the evidence level, which can precisely estimates multiple kinds of uncertainties. With minimization of the uncertainties, our method can reduce the conflict among multiple views, increase the evidences support for class probabilities and thereby improve classification performance and safeness.

(3) We theoretically prove that our model can decrease the empirical risk of learning results with the increase of views, and thereby prevent performance degradation problem in multi-view deep learning. Moreover, it is theoretically guaranteed that SMDC's generalization approaches the optimal in the order $O\left(\sqrt{d\ln(n)/n}\right)$.

Related Work

Multi-View Learning: In recent years, numbers of multi-view learning methods have been proposed to extract the correlation among multiple views. Canonical Correlation Analysis (CCA) (Harold 1936) and its variants (Bach and Jordan 2002; Wang 2007; Hardoon and Shawe-Taylor 2011) are classical approaches for unsupervised cross-view representation learning. CCA finds a common latent space by maximizing the correlation among different views. Kernel CCA (Bach and Jordan 2002) uses kernels to learn the common representation, which makes the CCA more robust. Sparse CCA (Hardoon and Shawe-Taylor 2011) reduces the effect of noisy data by learning a sparse representation. Different from CCA, some methods (Liu et al. 2015; Zhao, Ding, and Fu 2017; Liu et al. 2017; Zhang et al. 2018) learn hierarchical representation through matrix factorization, and some works (Xie et al. 2018, 2020) introduce a self-representation method to better incorporate multi-view information.

Moreover, a number of multi-view deep learning works (Andrew et al. 2013; Wang et al. 2015a; Tian, Krishnan, and Isola 2020; Tao et al. 2019; Bachman, Hjelm, and Buchwalter 2019; Zhang et al. 2020; Sun, Dong, and Liu 2020; Gan et al. 2021; Wang et al. 2021; Wen et al. 2020; Liu et al. 2022) combine deep learning with multi-view learning. Deep CCA (DCCA) (Andrew et al. 2013) focuses on capturing nonlinear relationships. Deep Canonically Correlated AutoEncoder (DCCAE) (Wang et al. 2015a) trains autoencoders to obtain common representations. Cross Partial Multi-view Network (Zhang et al. 2020) focuses on learning a complete representation under complex view-missing cases. Recently, some uncertainty-based multi-view classification methods (Geng et al. 2021; Han et al. 2022) have been proposed. Dynamic Uncertainty-Aware Network (Geng et al. 2021) employs Reversal networks to learn a unified representation. Enhanced Trusted Multi-view Classification Network (Han et al. 2022) focuses on the uncertainty estimation problem and produces a reliable classification result. However, while these methods achieve a great performance on multi-view deep classification, they rarely consider the safeness of the learning results, which cannot guarantee the performance without degradation after fusion of different views. In contrast, our method revisits the multi-view deep learning from the perspective of safe learning, with the minimization of the inherent uncertainties, our method can reduce the conflict among multiple views, increase the evidences support for class probabilities and thereby produce safe multi-view learning results.

Safe Learning: Safe learning recently has received a lot of attention in various machine learning domains (Kim, Allmendinger, and López-Ibáñez 2020). Some works (El Chamie, Yu, and Açıkmeşe 2016; Achiam et al. 2017; Turchetta et al. 2020) contribute to the agent does not violate safe constraints in safe reinforcement learning (RL). Moreover, there are also many works (Turchetta, Berkenkamp, and Krause 2016; Lederer, Umlauft, and Hirche 2019; Amani, Alizadeh, and Thrampoulidis 2020; Sun et al. 2021) try to quantify the model error bounds based on the Gaussian processes (GP), which allow safe control based on these models. Some works (Schreiter et al. 2015; Zimmer, Meister, and Nguyen-Tuong 2018; Li, Rakitsch, and Zimmer 2022) combine the active learning with Gaussian processes to devise a safe active learning method for unknown environ-

ments and some works (Guo et al. 2020; Li, Guo, and Zhou 2019) focus on the safe weakly-supervised learning where a large amount of data supervision is not accessible, which never seriously hurts performance.

Evidence Theory: Evidence theory, also referred to as Dempster–Shafer Theory (DST) (Shafer 1976; Denœux, Younes, and Abdallah 2010), is a generalization of the Bayesian theory to subjective probabilities (Dempster 1968) for reasoning with partial, unreliable, incomplete, deceptive or conflicting evidence. In contrast to the Bayesian neural networks which estimate uncertainty through multiple stochastic samplings from weight parameters, DST directly models uncertainty and combines evidences from different sources with various fusion operators to produce a new representation (Jøsang 2018). Typically, the uncertainty measured by DST means vacuity (i.e., lack of evidence), which has been used as an effective method to detect out-of-distribution samples in deep learning (Şensoy, Kaplan, and Kandemir 2018). Recently, other dimensions of uncertainty have been proposed, such as dissonance (due to conflicting evidence) and consonance (due to evidence about composite subsets of state values) (Josang, Cho, and Chen 2018). In this work, considering the property of the multi-view deep classification, the vacuity and dissonance are used to model the uncertainties from multiple different views.

Safe Multi-view Deep Classification

In this section, we introduce the proposed safe multi-view deep classification model in this paper. We first give a formal definition of the safety in multi-view deep classification.

Definition 1 (Safety in Multi-view Deep Classification) *In the context of multi-view deep classification, we hope we can obtain a multi-view optima $\widehat{\boldsymbol{\alpha}}$ that minimizes the total objective loss function. To guarantee the safety in multi-view classification, the loss should be bounded by the safety constraint that avoids worsening the empirical risk $\widehat{R}(\widehat{\boldsymbol{\alpha}})$ when fusing multiple data views.*

As mentioned above, we can simply formalize our safe multi-view deep classification (SMDC) model. Given a set of n samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with labels $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ for m views, we can easily have the objective function as

$$\min \ell\left(\boldsymbol{P}\left(\widehat{\boldsymbol{\alpha}}\right), \boldsymbol{y}\right), \tag{1}$$

bounded by the *safety constraint* $\widehat{R}(\widehat{\boldsymbol{\alpha}}) - \widehat{R}(\boldsymbol{\alpha'}) \leq 0$, where $\ell(\cdot,\cdot)$ refers to a loss function, e.g., the square loss, the hinge loss, etc. $\boldsymbol{P}(\widehat{\boldsymbol{\alpha}})$ represents the probability of classes obtained by DNNs after the fusion of multiple views, $\widehat{R}(\widehat{\boldsymbol{\alpha}})$ and $\widehat{R}(\boldsymbol{\alpha'})$ indicates the empirical risk after/before the fusion of new coming data views, respectively.

Now we should consider how to construct our objective function to satisfy the *safety constraint*. To achieve this goal, we devise our framework from the view of evidence theory. Within this framework, we first propose a multi-view aggregation strategy to integrate each view at an evidence level to formalize the common multi-view representation based on the evidence theory. Then we explicitly estimate the inherent uncertainties among multiple views with different root causes as learning results. Lastly, with minimization of uncertainties, we train our model to promote the evidences from views for correct classification, and in the meantime exclude the incorrect evidences to produce the safe multi-view classification results.

Multi-View Representation in Evidence Theory

Evidence theory associates the parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution $Dir(\boldsymbol{P}|\boldsymbol{\alpha})$ with the belief distribution, where Dirichlet distribution can be considered as the conjugate prior of the categorical distribution in multi-class classification, and $\boldsymbol{P}$ is a simplex representing class assignment probabilities (Jøsang 2018). In evidence theory, traditional neural network can be naturally transformed into a evidence-based neural network with minor changes that only replace the softmax layer with an activation layer (i.e., ReLU) to guarantee a non-negative output, called as evidence (Şensoy, Kaplan, and Kandemir 2018).

Here we first give the single-view representation in the context of multi-view deep classification based on evidence theory. Given the i^{th} sample with v^{th} view, represented by $\boldsymbol{x}_i^v$, let $\boldsymbol{e}_i^v = \{e_{i1}^v, \ldots, e_{iK}^v\}$ represent the evidence vector captured by the v^{th} neural network for the K-classification problem. Then, the corresponding Dirichlet distribution has parameter $\boldsymbol{\alpha}_i^v = \boldsymbol{e}_i^v + 1 = \{\alpha_{i1}^v, \ldots, \alpha_{iK}^v\}$. Once the parameters of this distribution are calculated, its mean, i.e., $\boldsymbol{\alpha}_i^v / S_i^v$, where $S_i^v = \sum_{k=1}^{K} (e_{ik} + 1)$ is the Dirichlet strength, can be taken as an estimate of the class probabilities.

The single-view case with evidence theory has been introduced above, we now focus on the classification with multiple views. Given a set of m evidence $\boldsymbol{e} = \{\boldsymbol{e}^v\}_{v=1}^m$ from m views, we want to construct an efficient common representation to take full advantage of information from each view. Therefore, we devise a simple and efficient aggregation strategy for multi-view deep classification with evidence theory, which is shown as follows:

Definition 2 (Aggregation strategy with evidence theory) *The aggregation strategy for multi-view deep classification based on the evidence theory simply consists of evidence parameter addition. Given a data with m multiple views for K-classification problem, we can obtain a set of evidences $\{\boldsymbol{e}^v\}_{v=1}^m$, collected from m neural networks. For $k = 1, \ldots, K$, $\boldsymbol{e}^v = \{e_1^v, \ldots, e_k^v\}$. Then we have $e_k = \sum_{v=1}^m e_k^v$ represents the process of multi-view aggregation and $S = \sum_{k=1}^{K} (e_k + 1)$ is the aggregated Dirichlet strength.*

Following the Definition 2, we combine the evidence from m data views into a common aggregated representation $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$, where $\alpha_k = e_k + 1$. Then we have $\boldsymbol{P}(\boldsymbol{\alpha}) = (p_1, \ldots, p_k)$ to produce the final probability of each class, where $p_k = \frac{\alpha_k}{S}$.

Uncertainties Estimation in Multi-View Classification

Moreover, referring to the literature (Josang, Cho, and Chen 2018), we can explicitly estimate inherent uncertainties in SMDC. As we focus on the multi-class setting, no composite values (i.e. simultaneously assigned to multiple classes) are allowed, we just discuss two main uncertainties *vacuity* and *dissonance*, which correspond to the vacuous evidence and contradicting evidences. In particular, vacuity uncertainty $Vac(\boldsymbol{\alpha})$ is defined as

$$Vac(\boldsymbol{\alpha}) = \frac{K}{S}, \tag{2}$$

and the dissonance uncertainty $Diss(\boldsymbol{\alpha})$ is defined as

$$Diss(\boldsymbol{\alpha}) = \sum_{k=1}^{K} \left(\frac{b_k \sum_{j \neq k} b_j \mathrm{Bal}(b_j, b_k)}{\sum_{j \neq k} b_j} \right), \tag{3}$$

where $b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}$ and

$$\mathrm{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k} & \text{if} b_i b_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

is the relative balance function.

The main cause of $Vac(\boldsymbol{\alpha})$ is a lack of evidence, which refers to the uncertainty caused by insufficient information. The $Diss(\boldsymbol{\alpha})$ belongs to the uncertainty caused by the conflict evidence. Here we interpret these two uncertainties in terms of the class-level evidence measures of multi-view result. For a 3-classification problem, suppose we have three results $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)$. $\boldsymbol{\alpha}_1 = (256, 1, 1)$ represents low uncertainty which means the sample certainly belongs to the first class; $\boldsymbol{\alpha}_2 = (1, 1, 1)$ indicates the case of high vacuity ($Vac(\boldsymbol{\alpha}_2) = 1$) due to insufficient evidence, which commonly happens on the outlier; $\boldsymbol{\alpha}_3 = (256, 256, 256)$ shows low vacuity but high dissonance ($Diss(\boldsymbol{\alpha}_3) \approx 0.996$), which indicates that although the vacuity is close to zero, the result can not make a clear decision due to the strong conflict among each classes. In general, these two uncertainties always exist in multi-view classification, which causes the classification performance degradation. Therefore, it is necessary to decrease the *vacuity* and *dissonance* in the safe multi-view classification.

Learning to Form Safe Multi-View Results

In this section, we will discuss how to train our model to capture the evidence from multiple data views for producing a safe multi-view classification result. Within our method, for the i^{th} sample x_i with m views and one-hot label $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{iK}\}$, our goal is to find the optimal parameter $\widehat{\boldsymbol{\alpha}}_i$ that prevents the performance degradation while minimizing the inherent uncertainties (*vacuity* and *dissonance*) to provide safe classification results. Following the safety constraint, the initial loss function is designed as follows

$$\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \sum_{i=1}^{n} \ell\left(\boldsymbol{P}(\boldsymbol{\alpha}_i), \boldsymbol{y}_i\right), \tag{5}$$

where $\widehat{\boldsymbol{\alpha}}_i$ is denoted as the optimal model trained with the aggregated Dirichlet parameters $\boldsymbol{\alpha}_i$ and $\ell\left(\boldsymbol{P}(\boldsymbol{\alpha}_i), \boldsymbol{y}_i\right)$ is prediction loss term which will be discussed in the next section. Specifically, SMDC requires that the model constrained by the uncertainties, i.e.,

$$\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \sum_{i=1}^{n} Vac(\boldsymbol{\alpha}_i) + Diss(\boldsymbol{\alpha}_i). \tag{6}$$

This means our model tries to minimize the *vacuity* and *dissonance* to address the insufficient evidence and conflict evidence problems in multi-view deep learning for promoting the performance. Obviously, with the increase of data views, the vacuity of multi-view result will decrease automatically with the increase of the aggregated Dirichlet strength S. Therefore, we just need to focus on the minimization of dissonance uncertainty from the conflict in multiple data views. Then we have,

$$\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \sum_{i=1}^{n} Diss(\boldsymbol{\alpha}_i). \tag{7}$$

Moreover, we also need to make our model to capture the evidence from the correct class and avoid generating evidence for the incorrect classes. Thus we achieve this by incorporating a Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) term into our loss function that regularizes our predictive distribution by penalizing those incorrect evidence to shrink to 0. The regularizing term $\ell_{KL}(\boldsymbol{\alpha}_i)$ is

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_i &= \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{KL}(\boldsymbol{\alpha}_i) \\ &= \underset{\tilde{\boldsymbol{\alpha}}_i}{\operatorname{argmin}} \sum_{i=1}^{n} KL\left[Dir(\boldsymbol{P}_i|\tilde{\boldsymbol{\alpha}}_i) || Dir(\boldsymbol{P}_i|\langle 1, \ldots, 1\rangle)\right], \end{aligned} \tag{8}$$

where

$$\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\alpha}_i \tag{9}$$

and $Dir(\boldsymbol{P}_i|\langle 1, \ldots, 1\rangle)$ means the uniform Dirichlet distribution.

Finally, taking the Eq. (5), Eq. (7) and Eq. (8) into consideration, the objective of our framework can be formulated as the following optimization problem,

$$\begin{aligned} &\min \sum_{i=1}^{n} \ell\left(\boldsymbol{P}(\widehat{\boldsymbol{\alpha}}_i), \boldsymbol{y}_i\right) \\ &s.t. \\ &\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \sum_{i=1}^{n} \ell\left(\boldsymbol{P}(\boldsymbol{\alpha}_i), \boldsymbol{y}_i\right) + \lambda_1 Diss(\boldsymbol{\alpha}_i) + \lambda_2 \ell_{KL}(\boldsymbol{\alpha}_i), \end{aligned} \tag{10}$$

where λ_1 and λ_2 are balance factors. In general, we set $\lambda_1, \lambda_2 = \min(1.0, t/10) \in [0, 1]$ as the annealing coefficients to prevent the network from paying too much attention to the dissonance uncertainty and KL divergence in the initial stage of training, t indicates the index of the current training epoch.

Algorithm 1: Algorithm for Safe Multi-View Deep Classification (SMDC)

/*Training*/
Input: Multi-view dataset: $D = \{\{X_n^v\}_{v=1}^m, y_n\}_{n=1}^N$.
Initialize: Initialize the parameters of the neural network.
while not converged **do**
  for $v = 1 : m$ **do**
    $\boldsymbol{e}^v \leftarrow$ non-negative output in neural network;
  end for
  $\boldsymbol{e} \leftarrow$ aggregation in terms of Definition 2;
  $\boldsymbol{\alpha} \leftarrow \boldsymbol{e} + 1$;
  Obtain the overall loss by updating $\boldsymbol{\alpha}$ with Eq 10;
  Update the neural networks with gradient descent according to Eq 10;
end while
Output: parameters of neural networks.
/*Test*/
Obtain the class probability and corresponding uncertainty degree.

Eq. (10) can be understood as: SMDC minimizes the uncertainties and shrinks the evidence from incorrect classes to 0 to promote the model for seeking the optimal evidence from each view to support the sample can be classified into right class and thereby make the learned $\widehat{\boldsymbol{\alpha}}_i$ to achieve a better safe performance. The optimization process for the proposed model is given in Algorithm 1.

Theoretical Studies

In this section, we will theoretically prove our model can achieve a safe result, whose performance is never worse when fusing multiple data views.

Given a training example $\boldsymbol{x}_i$ with one-hot label $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{iK}\}$. Let $\text{Cat}\left(\widehat{\boldsymbol{y}}_i = k \,|\boldsymbol{P}_i\right)$ be the likelihood, where $\boldsymbol{P}_i \sim Dir\left(\boldsymbol{P}_i \,|\boldsymbol{\alpha}_i\right)$, $\boldsymbol{P_i} = (P_{i1}, ..., P_{ik})^T$. The expected sum of squares loss of the aggregated multi-view representation $\boldsymbol{\alpha}_i$ is defined as

$$\begin{aligned}\ell\left(\boldsymbol{P}\left(\boldsymbol{\alpha}_i\right), \boldsymbol{y}_i\right) &= \mathbb{E}_{\boldsymbol{P}_i \sim Dir\left(\boldsymbol{P}_i|\boldsymbol{\alpha}_i\right)} \left\|\boldsymbol{y}_i - \boldsymbol{P}_i\right\|_2^2 \\ &= \sum_{j=1}^{K} \left(y_{ij}^2 - 2y_{ij}\mathbb{E}\left[P_{ij}\right] + \mathbb{E}\left[P_{ij}^2\right]\right). \end{aligned} \tag{11}$$

Then, in order to show the safeness of SMDC, we analyze the empirical risk of SMDC compared with the empirical risk learned before fusion of the new coming view and obtain the following theorem.

Theorem 1 (Safeness) *Let $\widehat{\boldsymbol{\alpha}}_i$ be the optimal multi-view decision learned from SMDC after the v^{th} view coming, i.e., $\widehat{\boldsymbol{\alpha}}_i = \min \sum_{i=1}^{n} \ell_{err}\left(\boldsymbol{P}\left(\widehat{\boldsymbol{\alpha}}_i\right), \boldsymbol{y}_i\right)$, where $\ell_{err}\left(\boldsymbol{P}\left(\widehat{\boldsymbol{\alpha}}_i\right), \boldsymbol{y}_i\right)$ indicates the prediction error. And let $\boldsymbol{\alpha}'$ be the optima learned in SMDC before the v^{th} view coming. Then our SMDC model satisfies the safety constraint $\widehat{R}\left(\widehat{\boldsymbol{\alpha}}\right) \leq \widehat{R}\left(\boldsymbol{\alpha}'\right)$ to guarantee the safeness of multi-view classification, where the empirical risk of $\boldsymbol{\alpha}$ can be defined as*

$$\widehat{R}\left(\boldsymbol{\alpha}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\ell_{err}\left(\boldsymbol{P}\left(\boldsymbol{\alpha}_i\right), \boldsymbol{y}_i\right)\right]. \tag{12}$$

Theorem 1 theoretically guarantees the safeness of the empirical risk in our method, we further analyze the generalization risk of SMDC to better understand the effect of our model parameter to $\boldsymbol{\alpha}$ and drive the following theorem.

Theorem 2 (Generalization) *Let X be a set of n samples with label Y, $\boldsymbol{\alpha} \in \mathbb{B}^d$ be the parameter of loss function in a finite d-dimensional unit ball. Define generalization risk as:*

$$R\left(\boldsymbol{\alpha}\right) = \mathbb{E}_{(X,Y)}\left[\ell_{err}\left(P\left(\boldsymbol{\alpha}\right), Y\right)\right]. \tag{13}$$

Let $\boldsymbol{\alpha}^* = \text{argmax}_{\boldsymbol{\alpha}\in\mathbb{B}^d} R\left(\boldsymbol{\alpha}\right)$ be the optimal parameter in the unit ball, $\widehat{\boldsymbol{\alpha}}$ be the optimal parameter of empirical risk among a candidate set A. With probability at least $1-\delta$ we have,

$$R\left(\boldsymbol{\alpha}^*\right) \leq R\left(\widehat{\boldsymbol{\alpha}}\right) + \frac{\left(12 + \sqrt{4d\ln\left(n\right) + 8\ln\left(2/\delta\right)}\right)}{\sqrt{n}}. \tag{14}$$

Theorem 2 shows the generalization of SMDC can approach the optimal result in the order $O\left(\sqrt{d\ln\left(n\right)/n}\right)$, where d indicates the number of parameters in our model and n denotes the number of samples. In summary, Theorem 1 and Theorem 2 indicate the latent representation learned from multiple views in SMDC is closer to the true representation than the representation learned before fusion of data views, which can produce more safe results.

Experiments

In this section, we extensively evaluate the proposed method on real-world multi-view datasets and compare it with existing multi-view classification methods. Furthermore, we also provide the analysis of safeness estimation on noisy data. Experimental results show that our algorithm achieves the state-of-the-art performance on various multi-view datasets.

Experimental Setup

Datasets and Comparative Methods We conduct experiments on six real-world multi-view datasets as follows: **Handwritten** (Van Breukelen et al. 1998), **Scene15** (Fei-Fei and Perona 2005), **Animal** (Lampert, Nickisch, and Harmeling 2013), **Caltech101** (Fei-Fei, Fergus, and Perona 2004), **CUB** (Wah et al. 2011) and **HMDB** (Kuehne et al. 2011). Also compare our method with existing state-of-the-art multi-view classification methods: **DCCA** (Andrew et al. 2013), **DCCAE** (Wang et al. 2015a), **CPM-Nets** (Zhang et al. 2020), **MVTCAE** (Hwang et al. 2021), **DUA-Nets** (Geng et al. 2021) and **ETMC** (Han et al. 2022).

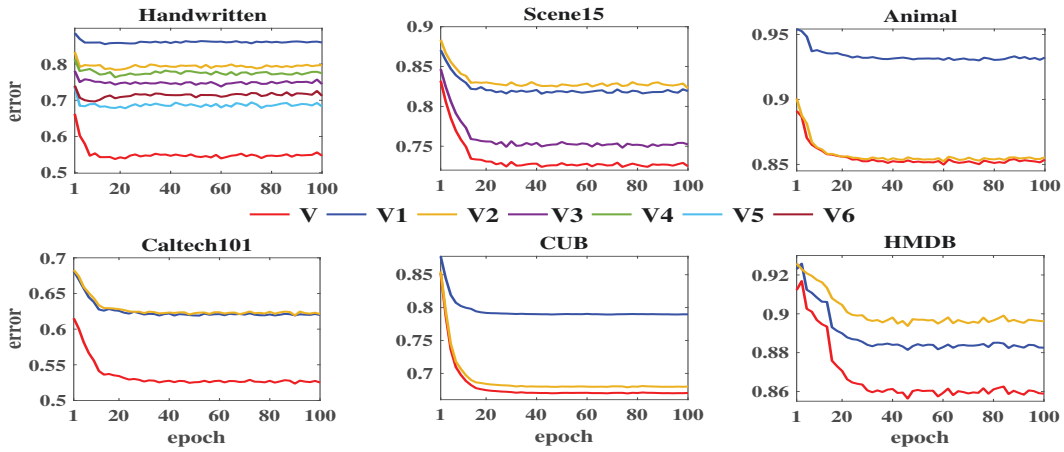

Figure 2: Comparison with the prediction error of single-view.

Implementations For our algorithm, we conduct the fully connected networks with Batch Normalization for all datasets. The Adam optimizer (Kingma and Ba 2014) is used to train the network, where l_2-norm regularization is set to $1e^{-5}$. We then use 5-fold cross-validation to select the learning rate from $\{1e^{-4}, 3e^{-4}, 1e^{-3}, 3e^{-3}\}$. For all datasets, 20% samples are used as test sets. Furthermore, we run 5 times for each method to report the average values in Figures or the mean values and standard deviations in Tables. The model is implemented by PyTorch on one NVIDIA A100 with GPU of 40GB memory.

Experimental Results

Ablation Studies In this subsection, we first conduct a detailed ablation study to clearly demonstrate the effectiveness of our major technical components, which consist of evaluation of multi-view fusion strategy, evaluation of dissonance uncertainty loss and evaluation of KL-divergence loss. Except for the result of first row in Table 1, that indicates the best accuracy among each single-view, we evaluate these three components on Caltech101 dataset with all the views. Since the uncertainties loss and KL-divergence loss are not applied to the single-view classification, thus there are five combinations between these major components. As shown in Table 1, our SMDC outperforms all other combinations in terms of the average accuracy over 5 runs, which verifies the effectiveness of our major technical components.

Performance Evaluation

In this subsection, we conduct two tests to evaluate the performance of our method. The first is to verify the effectiveness of our method and the second is to overall evaluate the superiority of our method by comparing it with state-of-the-art multi-view classification methods.

Effectiveness evaluation We first compare the average prediction error for multi-view results with each single-view on all datasets. Figure 2 shows the average classification error on testing set for multi-view (red line, termed as V) compared with the average error from each single view (termed as V1-V6) over 5 runs. We can find our multi-view classification errors are always smaller than single-view errors with the increase of epoch. Moreover, taking the test accuracy results on Handwritten dataset that contains six views (termed as V1-V6) as examples, Table 2 shows the multi-view classification accuracy increases with adding multiple views. Both of the experimental results shown in Figure 2 and Table 2 validate SMDC is effective and safe.

Main Components			Metric
Fusion	Uncertainties	KL-divergence	ACC (%)
	-	-	86.20±0.80
✓			96.00±0.12
✓	✓		96.89±0.20
✓		✓	97.33±0.12
✓	✓	✓	97.78±0.01

Table 1: Ablation study on Caltech101, ''✓'' means SMDC with the corresponding component, ''-'' means ''not applied''.

Views	Accuracy (%)
V1+V2	85.25±0.04
V1+V2+V3	96.75±0.01
V1+V2+V3+V4	97.50±0.02
V1+V2+V3+V4+V5	98.75±0.02
V1+V2+V3+V4+V5+V6	99.00±0.01

Table 2: Accuracy on Handwritten with increase of views.

Comparison with the methods Then we overall evaluate the performance of our model. The detailed results are shown in Table 4. We can clearly observe that SMDC consistently achieves better performance than other methods. Taking the results on HMDB as examples, our method improves the accuracy by about 16% compared to the second-best model. All of these results verify the improved performance of our SMDC method.

Noise ratio	HMDB	CUB	Scene15	Caltech101	Animal	Handwritten
$\lambda = 0.0$	90.76±0.01	96.65±0.01	72.76±0.06	97.78±0.01	94.10±0.01	99.00±0.01
$\lambda = 0.1$	88.56±0.01	83.33±0.00	68.32±0.02	96.65±0.10	87.23±0.02	98.33±0.01
$\lambda = 0.2$	88.46±0.01	79.33±0.00	67.52±0.10	96.02±0.12	86.98±0.01	98.00±0.01
$\lambda = 0.3$	88.44±0.00	78.47±0.05	68.00±0.15	95.11±0.02	86.78±0.00	98.00±0.01
$\lambda = 0.4$	88.33±0.01	78.46±0.03	67.01±0.02	95.01±0.10	86.00±0.01	97.50±0.02
$\lambda = 0.5$	88.40±0.00	78.31±0.01	65.33±1.12	94.73±0.12	83.67±0.01	97.25±0.02
Baseline	85.28±0.01	74.17±0.01	56.96±1.23	90.05±0.02	77.51±0.25	96.00±0.01

Table 3: Evaluation of safeness with different noise ratios (λ) based on classification accuracy ($\%$).

Data	DCCA	DCCAE	CPM-NetS	MVTCAE	DUA-Nets	ETMC	SMDC
CUB	82.03±2.40	85.50±1.37	89.44±0.06	92.00±0.04	81.42±1.15	91.23±1.21	**96.65±0.01**
HMDB	45.71±1.51	49.12±1.00	66.84±1.21	74.84±1.24	63.05±0.53	74.98±1.02	**90.84±0.11**
Scene15	54.77±1.13	55.12±0.23	67.09±0.05	66.43±0.06	68.43±0.02	68.30±0.01	**72.80±0.13**
Caltech101	84.00±0.15	90.03±0.11	90.05±1.42	91.76±0.01	93.83±0.34	93.41±0.22	**97.78±0.01**
Handwritten	94.55±2.01	97.01±0.23	94.45±1.11	97.00±0.23	98.10±0.32	98.51±0.13	**99.00±0.01**
Animal	83.33±1.25	85.80±0.51	86.59±0.05	86.32±0.16	89.05±1.22	89.71±0.34	**94.10±0.01**

Table 4: Comparison with state-of-the-art multi-view learning methods based on classification accuracy ($\%$).

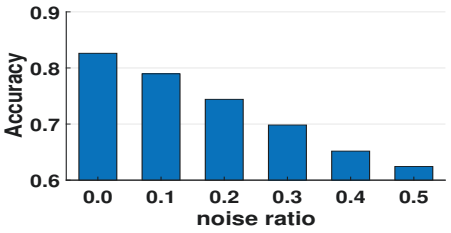

Figure 3: Classification accuracy of noise view on HMDB with different noise ratios (λ).

Safeness Estimation

In this part, we conduct qualitative experiments to overall evaluate the safeness of our model.

Similar to the work (Geng et al. 2021), we first add noise to the λ training data in one view, where λ is ranged from 0 to 0.5. Specifically, we generate λ noise vectors (denoted as $\boldsymbol{\epsilon}$) that are sampled from Gaussian distribution $\mathcal{N}(0, \mathrm{I})$. Then we add these noise vectors $\boldsymbol{\epsilon}$ to pollute λ training data in one view, i.e., $\widetilde{\boldsymbol{x}}^{(1)} = \boldsymbol{x}^{(1)} + \boldsymbol{\epsilon}$. Figure 3 shows the average testing accuracy on noise view over 5 runs. From the Figure 3, we observe that with the increasing of the noise ratio, the polluted data are increased, then the classification accuracy of the data in noise view has significant decrease, which means the noise can hurt the classification performance.

Then we test the classification accuracy of SMDC with different noise ratios λ on various datasets to provide overall safeness guarantee. Table 3 summarizes the experimental results using different noise ratios (λ). In Table 3, the 'Baseline' means the multi-view results before adding new coming noisy view and the others mean multi-view results with the coming of polluted view. We can find that SMDC always achieves better performance than the results without new coming view even if the view is polluted by noise, which verifies our method can guarantee that the classification performance does not deteriorate when fusing multiple data views. All of these experimental results validate the effectiveness and safeness of our model.

Conclusion

In this paper, we tackle an important problem of multi-view deep classification, that is, performance degradation in the presence of uncertain and inconsistency data sources. We propose an efficient safe multi-view deep classification method SMDC. The effectiveness of our proposal is demonstrated both theoretically and empirically. In theory, adding data views in our model is never worse than before in terms of the empirical risk, and the generalization analysis guarantees the generalization achieves the optimal in a faster order $O\left(\sqrt{d\ln(n)/n}\right)$. Empirical studies show that, our method can still achieve better performance even if the view is polluted by noise, which is in line with the theoretical results.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Serial Nos. 62173252, 61991410, 61976134), and Natural Science Foundation of Shanghai (Serial No. 21ZR1423900).

References

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained policy optimization. In *International conference on machine learning*, 22–31. PMLR.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2020. Regret bound for safe Gaussian process bandit optimization. In *Learning for Dynamics and Control*, 158–159. PMLR.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Bach, F. R.; and Jordan, M. I. 2002. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*.

Bickel, S.; and Scheffer, T. 2004. Multi-view clustering. In *ICDM*, volume 4, 19–26. Citeseer.

Dempster, A. P. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Denœux, T.; Younes, Z.; and Abdallah, F. 2010. Representing uncertainty on set-valued variables using belief functions. *Artificial Intelligence*, 174(7-8): 479–499.

El Chamie, M.; Yu, Y.; and Açıkmeşe, B. 2016. Convex synthesis of randomized policies for controlled Markov chains with density safety upper bound constraints. In *2016 American Control Conference (ACC)*, 6290–6295. IEEE.

Fei-Fei, L.; Fergus, R.; and Perona, P. 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*, 178–178. IEEE.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 524–531. IEEE.

Gan, Y.; Han, R.; Yin, L.; Feng, W.; and Wang, S. 2021. Self-supervised Multi-view Multi-Human Association and Tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, 282–290.

Geng, Y.; Han, Z.; Zhang, C.; and Hu, Q. 2021. Uncertainty-Aware Multi-View Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7545–7553.

Guo, L.-Z.; Zhang, Z.-Y.; Jiang, Y.; Li, Y.-F.; and Zhou, Z.-H. 2020. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, 3897–3906. PMLR.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted Multi-View Classification with Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hardoon, D. R.; and Shawe-Taylor, J. 2011. Sparse canonical correlation analysis. *Machine Learning*, 83(3): 331–353.

Harold, H. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4): 321–377.

Hou, C.; Zeng, L.-L.; and Hu, D. 2018. Safe classification with augmented features. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2176–2192.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-View Representation Learning via Total Correlation Objective. *Advances in Neural Information Processing Systems*, 34.

Jøsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer.

Josang, A.; Cho, J.-H.; and Chen, F. 2018. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, 1998–2005. IEEE.

Kim, Y.; Allmendinger, R.; and López-Ibáñez, M. 2020. Safe learning and optimization techniques: Towards a survey of the state of the art. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, 123–139. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3): 453–465.

Lederer, A.; Umlauft, J.; and Hirche, S. 2019. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32.

Li, C.-Y.; Rakitsch, B.; and Zimmer, C. 2022. Safe Active Learning for Multi-Output Gaussian Processes. *arXiv preprint arXiv:2203.14849*.

Li, Y.-F.; Guo, L.-Z.; and Zhou, Z.-H. 2019. Towards safe weakly supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(1): 334–346.

Liu, H.; Liu, L.; Le, T. D.; Lee, I.; Sun, S.; and Li, J. 2017. Nonparametric sparse matrix decomposition for cross-view dimensionality reduction. *IEEE Transactions on Multimedia*, 19(8): 1848–1859.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Liu, W.; Yue, X.; Chen, Y.; and Denoeux, T. 2022. Trusted Multi-View Deep Learning with Opinion Aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7585–7593.

Schreiter, J.; Nguyen-Tuong, D.; Eberts, M.; Bischoff, B.; Markert, H.; and Toussaint, M. 2015. Safe exploration for active learning with Gaussian processes. In *Joint European conference on machine learning and knowledge discovery in databases*, 133–149. Springer.

Şensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton university press.

Sun, D.; Khojasteh, M. J.; Shekhar, S.; and Fan, C. 2021. Uncertain-aware Safe Exploratory Planning using Gaussian Process and Neural Control Contraction Metric. In *Learning for Dynamics and Control*, 728–741. PMLR.

Sun, S.; Dong, W.; and Liu, Q. 2020. Multi-view representation learning with deep gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tang, H.; and Liu, Y. 2022. Deep Safe Multi-View Clustering: Reducing the Risk of Clustering Performance Degradation Caused by View Increase. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 202–211.

Tao, Z.; Liu, H.; Li, J.; Wang, Z.; and Fu, Y. 2019. Adversarial graph embedding for ensemble clustering. In *International Joint Conferences on Artificial Intelligence Organization*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *European conference on computer vision*, 776–794. Springer.

Turchetta, M.; Berkenkamp, F.; and Krause, A. 2016. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29.

Turchetta, M.; Kolobov, A.; Shah, S.; Krause, A.; and Agarwal, A. 2020. Safe reinforcement learning via curriculum induction. *Advances in Neural Information Processing Systems*, 33: 12151–12162.

Van Breukelen, M.; Duin, R. P.; Tax, D. M.; and Den Hartog, J. 1998. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4): 381–386.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.

Wang, C. 2007. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3): 905–910.

Wang, J.; Zheng, Y.; Song, J.; and Hou, S. 2021. Cross-View Representation Learning for Multi-View Logo Classification with Information Bottleneck. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4680–4688.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015a. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015b. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Wen, J.; Zhang, Z.; Zhang, Z.; Wu, Z.; Fei, L.; Xu, Y.; and Zhang, B. 2020. Dimc-net: Deep incomplete multi-view clustering network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3753–3761.

Xie, Y.; Liu, J.; Qu, Y.; Tao, D.; Zhang, W.; Dai, L.; and Ma, L. 2020. Robust kernelized multiview self-representation for subspace clustering. *IEEE transactions on neural networks and learning systems*.

Xie, Y.; Tao, D.; Zhang, W.; Liu, Y.; Zhang, L.; and Qu, Y. 2018. On unifying multi-view self-representations for clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, 126(11): 1157–1179.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Yang, Y.; Song, J.; Huang, Z.; Ma, Z.; Sebe, N.; and Hauptmann, A. G. 2012. Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE Transactions on Multimedia*, 15(3): 572–581.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep Partial Multi-View Learning. *IEEE transactions on pattern analysis and machine intelligence*.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018. Binary multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1774–1782.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zimmer, C.; Meister, M.; and Nguyen-Tuong, D. 2018. Safe active learning for time-series modeling with gaussian processes. *Advances in neural information processing systems*, 31.