NAVIGATING CONFLICTING VIEWS: HARNESSING TRUST FOR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Resolving conflicts is essential to make the decisions of multi-view classification more reliable. Much research has been conducted on learning consistent and informative representations among different views, often assuming that all views are equally important and perfectly aligned. However, real-world multi-view data may not always conform to these assumptions, as some views may express distinct information. To address this issue, we develop a computational trust-based discounting method to enhance the existing Evidential Multi-view framework in scenarios where conflicts between different views may arise. Its belief fusion process considers the reliability of predictions made by individual views via an instance-wise probability-sensitive trust discounting mechanism. We evaluate our method on six real-world datasets, using Top-1 Accuracy, Fleiss' Kappa, and a new metric called Multi-View Agreement with Ground Truth that takes into consideration the ground truth labels, to measure the reliability of the prediction. We also evaluate whether uncertainty measures can effectively indicate prediction correctness by calculating the AUROC. The experimental results show that computational trust can effectively resolve conflicts, paving the way for more reliable multi-view classification models in real-world applications.

1 INTRODUCTION

Multi-View Classification (MVC) plays a critical role in deep learning by greatly enhancing the ability to make accurate decisions through integrating multi-source information. Its effectiveness has been verified with the successful application in many domains such as autonomous driving (Yurtsever et al., 2020) and AI-assisted medical diagnostic systems (Kang et al., 2020). Most of the existing studies on MVC rely on the assumption that data from different views consistently provide reliable information about the ground truth (Liang et al., 2024; Zhang et al., 2023a; Xu et al., 2024a). Nevertheless, this assumption may not always be valid in real-world scenarios. Substantial variations in the informativeness of data from different views can produce conflicting results, thereby undermining the reliability of the model's predictions.

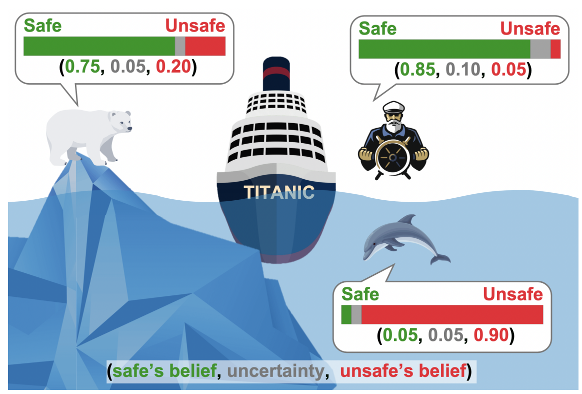

Figure 1: Example of conflicting multi-view opinions. The Titanic's route is safe in Captain's and Polar Bear's View, while unsafe in Dolphin's view.

A possible solution for resolving conflicts is to project data from different views into a shared latent space (Hardoon et al., 2004; Wang et al., 2015; Federici et al., 2020; Hjelm et al., 2019), and then draw a joint representation from the latent space for the classification task. This is achieved by integrating essential features via weighting schemes, such as attention mechanisms (Zheng et al.,

2021) and weighted fusion (Atrey et al., 2010; Zhang et al., 2019). These methods typically assign higher weights to more informative views or features, thus reducing the impact of potential conflicting information. Although these methods have achieved promising results in MVC, their focus on the joint representation can be a limitation. Solely relying on the joint representation hinders the capacity to thoroughly grasp information provided by different views. In contexts such as ocean navigation, characterized by observations sources from various views (e.g., the perspectives of the captain, dolphin and Polar Bear when observing an iceberg as shown in Figure 1), it is crucial to thoroughly analyze and comprehend each view before making the decision to cross and face or detour, as different views provide unique and complementary information.

Existing approaches to resolve conflicts build neural networks to generate view-specific predictions and then combine view-specific predictions together. As a prime example, the Evidential Multi-view framework is emerging as a promising approach, offering a reliable means for the final fusion stage. Within this framework, evidence acts as a metric of endorsement for the associated predicted label, and the evidence is collected through view-specific neural networks. Subsequently, evidence from diverse viewpoints is fused, considering their respective epistemic uncertainties. However, there may exist cases where the view-specific information is not well aligned with the ground truth, resulting in misleading predictions with high confidence (low uncertainty). For example, as shown in Figure 1, while the dolphin can clearly observe the massive structure hidden beneath the water's surface, the captain may only see the tip of the iceberg.

In this work, we take a significant step further: leveraging the Evidential Multi-view framework, we propose a new computational trust based opinion fusion method to resolve potential conflicts in MVC. Specifically, the computational trust is modelled through an evidence network that operates on a view-specific and instance-wise basis. Drawing upon the principle of trust discounting in subjective logic, it evaluates the reliability of view-specific predictions generated by existing Evidential frameworks, such as Evidential Deep Learning (EDL) (Sensoy et al., 2018). Within the proposed method, each view-specific evidence is transformed into a degree of trust using the Binomial opinion theory (Jøsang, 2018). These degrees of trust are then utilized to establish uncertainty and a trust-aware opinion, ultimately facilitating the generation of reliable predictions. In summary, the contributions of this paper include:

1. We present a novel learnable trust-discounting mechanism to extend the widely-used Evidential MVC framework, enhancing its conflict resolution capabilities. Drawing from the Binomial opinion theory within subjective logic, it operates on a view-specific and instance-wise basis, adeptly resolving conflicts among views through a probability-sensitive trust discounting rule;
2. We develop a stage-wise training strategy [1] to optimize the parameters of the proposed mechanism, which works robustly on different datasets;
3. We conduct extensive experiments on six real-world datasets, showing that our method outperforms the existing Evidential MVC methods, particularly on the datasets exhibiting large discrepancy among view-specific predictions. In addition, our method can also enhance the consistency among opinions derived from different views.

2 Related Work

Multi-View Classification leverages multiple data sources, offering varied perspectives on the same object, to enhance the classification performance. Recent advancements in MVC have focused on generating noise-robust representations through cluster-based (Huang et al., 2023; Wen et al., 2023a; Zhang et al., 2023b), self-representation-based (Hou et al., 2020), and partially view-aligned (Wen et al., 2023b; Huang et al., 2020) methods, harnessing the expressive power of deep neural networks. However, noise-robust representations may not fully resolve conflicts in opinions for a given data instance, as conflicts may arise by discrepant information from distinct views, and the discrepancy cannot be eliminated by addressing noises. Our method addresses this limitation by introducing a separate evidence network that evaluates the reliability of view-specific predictions and adjusts the final predictions according to the degree of trust.

[1]We move the detailed training algorithm to the Appendix A due to the space limitation.

Trusted Multi-View Classification has emerged as a crucial area and a pivotal domain within Multi-View Learning. This research area aims to enhance the accuracy and dependability of classification models by integrating data from multiple views, guided by their prediction confidence and epistemic uncertainty. The seminal work, Trusted Multi-View Classification (TMC) (Han et al., 2021), introduced the fusion of different views from an opinion perspective using the Dempher-Shafer Combination rule. Building upon TMC, Han et al. (2022) extended the approach by incorporating the pseudo-view, a concatenation of all other views, resulting in improved performance. Subsequent studies by Liu et al. (2022) and Xu et al. (2024a) explored alternative opinion fusion methods. Concurrent research efforts, such as those by Jung et al. (2022) and Jung et al. (2023), focus on multiview uncertainty estimation, enhancing the model's reliability. Similar to the TMC, our method is also built upon the Evidential Neural Network, and generating the fused decision by using the Dempher-Shafer Combination rule. However, we introduce a novel Trust Discounting module, which adjust the original evidence and opinions before the Dempher-Shafer Combination based on the reliability of evidence and opinion.

Conflictive Multi-View Classification argues that existing work primarily focusing on either learning joint aligned representations or better quantifying uncertainty overlook the problem of potential contradictory in the prediction space. Recognizing this gap, the pioneer work by Xu et al. (2024a) highlighted this issue and introduced the Degree of Conflict loss. This loss quantifies the disparity between different predictions in the prediction space while accounting for uncertainty, aiming to mitigate conflict-related challenges. However, this approach may inadvertently lead correct predictions to converge towards incorrect ones, potentially jeopardizing model stability. For instance, if most views are making incorrect predictions, the minority of correctly predicted views may be forced to align with the majority of incorrect ones. In contrast, our method can generate more accurate predictions with properly estimated uncertainty. As the trust discount module of our method is trained based on the correctness of the view-specific prediction and directly assess the reliability of it, instead of using other views's predictions which may provide incorrect optimization direction.

3 Trust Fusion Enhanced Evidential MVC

3.1 Preliminaries

Given training data $\mathcal{D} = \{\{\boldsymbol{x}_i^v\}_{v=1}^V, y_i\}_{i=1}^N$ where N is the number of training data, each instance $\boldsymbol{x}_i$ has V views, ground truth label y_i and an one-hot encoded label $\mathbf{y}_i$ (i.e., for a K-class classification problem, $\mathbf{y}_{i,k}$ is 1 if k is the index of ground truth label for i-th instance, otherwise it is 0). The task of MVC is to learn a function f that maps $\{\boldsymbol{x}_i^v\}_{v=1}^V$ to $\mathbf{y}_i$.

The Evidential MVC framework applies Subjective Logic (SL) to the K-class classification problem by assigning belief masses to individual class labels and computing epistemic uncertainty for the generated belief masses. The formulation links the evidence collected from instance view-specific observation to the concentration parameter of the Dirichlet Distribution. Let $f_\theta^v(\cdot)$ denote the view-specific neural network for evidence generation, where the view-specific evidence for an instance is $\boldsymbol{e}^v = f_\theta^v(\boldsymbol{x}^v)$, the association between the evidence and the Dirichlet parameters is simply $\alpha_k = e_k + 1$ (Sensoy et al., 2018; Han et al., 2021) . The belief mass on class label k, denoted as b_k, and uncertainty u are subject to the additive requirement, i.e., $u + \sum_{k=1}^K b_k = 1$. With respect to MVC, the view-specific belief mass b_k^v and uncertainty u^v can then be computed as

$$S^v = \sum_{k=1}^K \alpha_k^v, \quad b_k^v = \frac{e_k^v}{S^v} = \frac{\alpha_k^v - 1}{S^v}, \quad u^v = 1 - \sum_{k=1}^K b_k^v = \frac{K}{S^v} \tag{1}$$

To generate the final prediction, SL models the view-specific predictions as multinomial opinions, denoted as $\omega^v = [\boldsymbol{b}^v, u^v, \boldsymbol{a}^v]$, with $\boldsymbol{a}^v$ being the base rate (i.e., a prior probability distribution over classes, generally a discrete uniform distribution), and then combine them together with an appropriate belief fusion rules based on the context (Jøsang et al., 2013). The Belief Constraint Fusion (BCF) (Jøsang et al., 2013), an extension of Dempher-Shafer combination rule (Shafer, 1976), was first adopted by (Han et al., 2021) in trusted MVC. Other fusion rules, such as Aleatory Cumulative Belief Fusion (A-CBF) (Liu et al., 2022) and Averaging Belief Fusion (ABF) (Xu et al., 2024a) have also been explored. We choose to stay with BCF in our experiments due to its intuitive foundation (Jøsang et al., 2013; Jøsang, 2018) and the effectiveness demonstrated by (Han et al., 2021; 2022).

The fusion rule, $\oplus$, of BCF, among two views, i.e., $\omega = \omega^1 \oplus \omega^2$, can be formulated as follows:

$$b_k = \frac{1}{1-C}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1), \quad u = \frac{1}{1-C} u^1 u^2 \tag{2}$$

where $C = \sum_{i \neq j} b_i^1 b_j^2$ is the normalization factor, and b_k is the belief mass of label k and u is the uncertainty the fused opinion ω. Since the order of combination does not affect the final result (Jøsang, 2018), applying Eq. 2 by sequentially combining the V views in pairs, where the result of each combination is then combined with the next view, will derive the final fused opinion, which is as follows,

$$\omega = \omega^1 \oplus \omega^2 \oplus \cdots \omega^V \tag{3}$$

For the fused opinion ω, we can derive the parameters of the Dirichlet α_k by reversing the computation of Eq. 1.

Corollary 1. *An alternative representation for BCF is based on combining the evidence* [2], *from which the opinion $\omega = [\boldsymbol{b}, u, \boldsymbol{a}]$ can be derived:*

$$e_k = e_k^1 + e_k^2 + \frac{e_k^1 e_k^2}{K} \tag{4}$$

3.2 Conflict Resolving By Trust Fusion

We realize conflicts can happen when view-specific opinions express conflicting preferences, leading to ambiguity in the fused opinion, for example, two views' candidate labels has same confidence(belief), and subsequently draws potential inaccurate predictions. Based upon this, we define the conflict problem as follows:

Definition 1 (Conflicts within Multi-view Classification)**.** *In a K-class multi-class classification problem involving a multi-view dataset, a classification conflict arises when multiple views that predict different classes. This conflict leads to ambiguity in aggregating these predictions, as it becomes challenging to determine a single, coherent classification result from those inconsistent predictions.*

Although Belief Fusion has been verified effectively to fuse different opinions under SL, it still can generate unreliable fused opinions and lead to inaccurate predictions, for example, the Titanic navigation route case used in Figure 1. The data of different views' opinions have been recollected, and shown in Table 1. Besides, we also compute the fused opinion generated through BCF by substituting the data of three (i.e., Captain, Dolphin and PolarBear) functional opinions into Eq. 2 and Eq. 3 [3], and the fused opinion has also been appended to the Table 1.

Table 1: Opinions from Different views and BCF Fused opinion

View	Belief: Safe	Belief: Distrust	Uncertainty
Captain(functional)	0.85	0.05	0.10
Dolphin(functional)	0.05	0.90	0.05
PolarBear(functional)	0.75	0.20	0.05
Fused (BCF)	0.68	0.31	0.01

From Table 1, we can see that compared to the "unsafe" option, the fused opinion assigns a higher belief mass to the "safe" option (0.68 vs. 0.31). As a result, the prediction will be "safe", which is factually incorrect, as indicated in Figure 1. We attribute this error to less evidence collected, so less beleif mass to support the factual correct option "safe" in Captain's and PolarBear's view. insufficient evidence being collected, resulting in less belief mass supporting the factually correct option, "unsafe," in the opinions of both Captain and PolarBear. Additionally, the fused opinion exhibits lower uncertainty (0.01) compared to the original views' opinions (0.1, 0.05 and 0.05), however, the uncertainty is expected to be not lower than that of all views to reflect the struggle among different opinions in the presence of conflict.

[2] We provide the proof in Appendix B.2 and we implement BCF based on this equation due to its computational efficiency.

[3] We provide the detailed calculation process in Appendix

We utilize the principle of Trust Fusion (TF) by Trust Discounting (TD) (Jøsang et al., 2015) to handle the incorrect prediction caused by conflicting opinions. The basic idea of TD is to discount evidence or opinion from an individual view as a function of trust on that view. It can be used to weigh the current view-specific opinion according to the degree of trust, thus guiding the fusion process to generate more reliable prediction. Here we present a Probability-sensitive Trust Discounting rule, as show in Eq. 5, and use it in an instance-wise manner in our experiments as follows,

Definition 2 (Instance-wise Probability-Sensitive Trust Discounting)**.** *For each view of each individual instance, the trust-discounted opinion is defined as*

$$\breve{\omega}_i^v = \ddot{\omega}_i^v \otimes \acute{\omega}_i^v = \begin{cases} \breve{\boldsymbol{b}}_i^v = \ddot{p}_{t,i}^v * \acute{\boldsymbol{b}}_i^v, \\ \breve{u}_i^v = 1 - \ddot{p}_{t,i}^v * \left(\sum_{k=1}^K \acute{b}_{k,i}^v\right). \end{cases} \tag{5}$$

where i, v are the index for v-th view of i-th instance, $\otimes$ indicates the TD operator, $\breve{\omega}$ denotes the discounted opinion, and $\ddot{\omega}$, $\acute{\omega}$ denote referral opinion and functional opinion (e.g., opinions in Table 1), respectively. The scalar probability $\ddot{p}_t$ denotes the degree of trust, representing how much we are confident with the opinion given by the view-specific evidential model. Given Eq. 5, we fuse the trust-discounted opinions from V views of i-th instance with BCF by:

$$\bar{\omega}_i = \breve{\omega}_i^1 \oplus \breve{\omega}_i^2 \oplus \cdots \oplus \breve{\omega}_i^V = \left(\ddot{\omega}_i^1 \otimes \acute{\omega}_i^1\right) \oplus \left(\ddot{\omega}_i^2 \otimes \acute{\omega}_i^2\right) \oplus \cdots \oplus \left(\ddot{\omega}_i^V \otimes \acute{\omega}_i^V\right) \tag{6}$$

Note that 1) the referral opinion is different from the functional opinion shown in Table 1, which aims for assessing reliability of corresponding views' functional opinion, and 2) comparing with original Probability-Sensitive TD (Jøsang et al., 2012), our proposed instance-wise manner takes into consideration the opinions reliability of each instance, instead of global reliability of view only.

According to (Jøsang et al., 2015), the probability $\ddot{p}_t$ can be computed by $\ddot{p}_t = \ddot{b}_t + \ddot{a}_t * \ddot{u}$ [4] with $\ddot{\boldsymbol{a}}$ being the uniformly distributed base rate, i.e., $\ddot{a}_t$ = 1/2 for each individual instance on each view. Assuming we have the referral opinions for each view's functional opinion in Table 1, and defined in the Table 2.

Table 2: Referral Opinions of Different views

	Belief		Uncertainty	Trust ($\ddot{p}_t$)
View	Trust	Distrust		
Captain(referral)	0.60	0.30	0.10	0.65
Dolphin(referral)	0.90	0.00	0.10	0.95
PolarBear(referral)	0.20	0.70	0.10	0.25

By substituting trust scores $\ddot{p}_t$ with the data in Table 2 and functional beliefs $\acute{\boldsymbol{b}}$ with the data in Table 1 in Eq. 5 and Eq. 6, we effectively apply TD to original functional opinions. This process enabled us to compute the discounted opinions for each view as well as their fused opinion through BCF combination, which is shown as in Table 3.

Table 3: Discounted Opinions from Different views and BCF Fused opinion

	Belief		Uncertainty
View	Safe	Unsafe	
Captain(discounted)	0.55	0.03	0.42
Dolphin(discounted)	0.04	0.86	0.10
PolarBear(discounted)	0.19	0.05	0.76
Fused (BCF)	0.22	0.70	0.08

We can see that with the intervention of TD, the BCF fused opinion now assigns more belief mass to "unsafe," which aligns with the factual label. Additionally, the uncertainty of the fused opinion is now 0.08, which is rational given that Captain's and PolarBear's opinions have high uncertainty. Therefore, the decision aligning with Dolphin's opinion, which has significantly lower uncertainty than the others, is reasonable.

[4]We will show that $p_t = b_t + a_t * u$ is equivalent to $p_t = \alpha_2/(\alpha_1 + \alpha_2)$ with the assumption that base rate a_t is uniformly distributed in Appendix B.1.

Corollary 2. *Above Eq. 2 also corresponds to updating the Dirichlet evidence by* [5] *:*

$$\breve{e}^v_{k,i} = \frac{\ddot{p}^v_{t,i}\acute{u}^v_{t,i}}{1 - \ddot{p}^v_{t,i} + \ddot{p}^v_{t,i}\acute{u}^v_{t,i}}\,\acute{e}^v_{k,i} \tag{7}$$

The following propositions provide theoretical analysis of the proposed TD rule for achieving TF, and their detailed proof can be found in Appendix B.4.

Proposition 1. *Instance-wise Probability-Sensitive TD maximizes the belief mass of the Ground truth label after BCF, under the assumption that at least one view's prediction is correct.*

Proposition 2. *The combined opinion generated by proposed TF (TD+BCF) for conflicting views, will exhibit greater uncertainty than obtained through fusion with non-discounted functional opinions.*

3.3 Learning to Form Opinions

We depict the proposed TF (TD+BCF) along with entire Evidential MVC framework in Figure 2. The view-specific functional evidence is generated through an Evidential Neural Network (ENN), i.e., $\acute{\boldsymbol{e}}^v_i = f^v_{\acute{\theta}}(\boldsymbol{x}^v_i)$, which is same as Han et al. (2021). Similar to the functional evidence generation process, we construct another view-specific evidential network parameterized by $\ddot{\theta}$, for collecting referral evidence $\ddot{e}$, i.e., $\ddot{\boldsymbol{e}}^v_i = f^v_{\ddot{\theta}}([\boldsymbol{x}^v_i, \acute{\boldsymbol{b}}^v_i])$ [6], where both feature representation $\boldsymbol{x}^v_i$ and functional opinion $\acute{\boldsymbol{b}}^v_i$ are used as inputs.

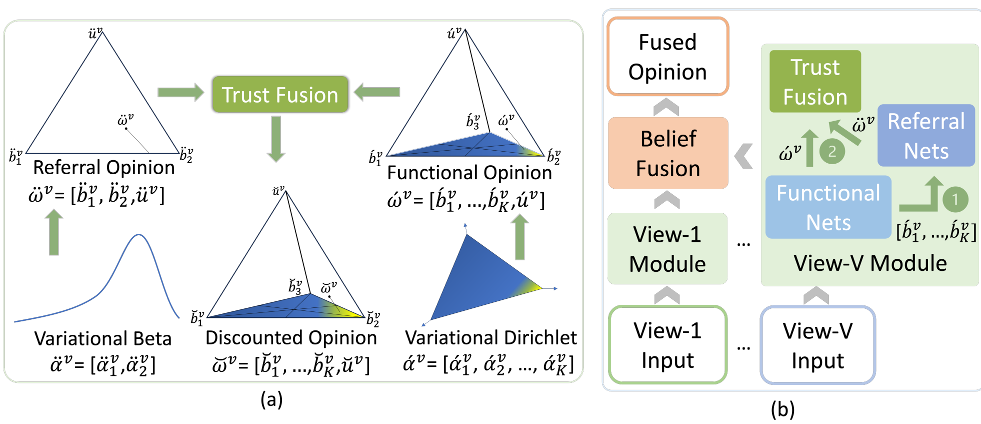

Figure 2: The TF Enhanced Evidential MVC Framework. (a) is the zoomed in view of View-specific Module, the Discounted Opinion is produced by applying Trust Fusion to discount the Functional Opinion using the Referral Opinion. (b) is the overall flow of the Evidential MVC framework.

In terms of loss function, we follow Sensoy et al. (2018); Han et al. (2021; 2022); Xu et al. (2024a) and optimize parameters of each view-specific evidential network. The loss term for i-th instance on v-th view is defined as follows,

$$L^v_i = \sum_{k=1}^{K} \mathbf{y}_{i,k}(\psi(S^v_i) - \psi(\alpha^v_{i,k})) + \lambda_o \mathrm{D}_{KL}[\mathrm{Dir}(\mathbf{p}^v_i|\tilde{\boldsymbol{\alpha}}^v_i)||\mathrm{Dir}(\mathbf{p}^v_i|\mathbf{1})] \tag{8}$$

where ψ is the digamma function, $\lambda_o = min(1.0, o/10)$ is the annealing factor, and o is the index of the current training epoch, $\tilde{\boldsymbol{\alpha}} = \mathbf{y} + (1 - \mathbf{y}) \odot \boldsymbol{\alpha}$ is the Dirichlet parameters after removing misleading evidence from predicted distribution parameters $\boldsymbol{\alpha}$, and $\mathbf{p}$ is the projected probability, i.e., $\mathbf{p} = \boldsymbol{\alpha}/S$.

Note that, 1) the loss term above is directly linked with the distribution parameters that are generated through ENN parameterized by θ, which will also be updated through back-propagation during

[5] We provide the proof in Appendix B.3.

[6] We used Bi-Linear layer instead of Dense/Linear Layer in our experiments.

Algorithm 1: Algorithm For Training (Simplified by omitting batch-wise operation)

Input: Multi-view dataset: $\mathcal{D} = \{\{\mathbf{x}_i^v\}_{v=1}^V, y_i\}_{i=1}^N$.
Initialize: Initialize the parameters $\acute{\theta}$, $\ddot{\theta}$ of Functional and Referral ENNs, respectively.
Stage-1 Warm-up Referral Network
Obtain $\{\ddot{\mathbf{e}}^v\}^V \leftarrow$ Referral ENNs outputs and $\{\ddot{\boldsymbol{\alpha}}^v\}^V$ by $\ddot{\boldsymbol{\alpha}}^v \leftarrow \ddot{\mathbf{e}}^v + 1$;
Update the parameters $\ddot{\theta}$ by Gradient Descent (GD) with loss of Eq. 10 for all $\{\ddot{\boldsymbol{\alpha}}^v\}^V$;
Stage-2 Update Functional Network
/*Substage-2a*/
Obtain $\{\acute{\mathbf{e}}^v\}^V \leftarrow$ Functional ENNs outputs and $\{\acute{\boldsymbol{\alpha}}^v\}^V$ by $\acute{\boldsymbol{\alpha}}^v \leftarrow \acute{\mathbf{e}}^v + 1$;
Update the parameters $\acute{\theta}$ by GD with loss of Eq. 8 for all $\{\acute{\boldsymbol{\alpha}}^v\}^V$;
/*Substage-2b*/
Obtain $\{\ddot{\mathbf{e}}^v\}^V \leftarrow$ Referral ENNs outputs and $\{\ddot{\boldsymbol{\alpha}}^v\}^V$ by $\ddot{\boldsymbol{\alpha}}^v \leftarrow \ddot{\mathbf{e}}^v + 1$;
Obtain $\{\acute{\mathbf{e}}^v\}^V \leftarrow$ Functional ENNs outputs and $\{\acute{\boldsymbol{\alpha}}^v\}^V$ by $\acute{\boldsymbol{\alpha}}^v \leftarrow \acute{\mathbf{e}}^v + 1$;
Obtain $\ddot{\omega}^v$ and $\acute{\omega}^v$ by Eq. 1 with $\ddot{\mathbf{e}}^v$ and $\acute{\mathbf{e}}^v$ for all views, respectively ;
Obtain BCF fused opinion $\bar{\omega}$ by Eq. 6 and corresponding $\bar{\boldsymbol{\alpha}}$ by reversing Eq. 1;
Update the parameters $\acute{\theta}$ by GD with loss of Eq. 8 for $\bar{\boldsymbol{\alpha}}$;
Stage-3 Adjust Referral Network
By repeating Stage-2b and update $\ddot{\theta}$ instead of $\acute{\theta}$ only ;
Stage-4 Adjust Functional Network
By repeating entire Stage-2;
Output: Functional and Referral networks parameters.

training stage; 2) even though we omit the notation for distinguishing the distribution parameters that govern the variational transformation of referral and functional opinions, this loss term will still be applied to the referral and functional nets respectively; 3) the above equation will be also applied to the final fused opinion since its corresponding variational Dirichlet has parameter $\bar{\boldsymbol{\alpha}}$ as well. We illustrate when and how to use the loss term in our proposed stage-wise training algorithm (Alg. 1) [7].

We also adopt a warm-up stage for the referral nets since the randomly initialized parameters of them could introduce unreliable trust scores for discounting at early training stage. The loss term used at the warm-up stage is simply the left summation term of Eq. 8 with a different target label which is defined as

$$z_i^v = \begin{cases} 1 & \text{if } \hat{y}_i^v = y_i \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $\hat{y}_i^v = \arg\max_k \acute{\boldsymbol{b}}$ which is predicted label of functional opinion, so the target label z_i^v primarily indicates the correctness of such prediction. Following Müller et al. (2019), we apply label smoothing with smoothing factor $\eta = 0.9$ to the hard label. The association between one-hot encoded hard label $\mathbf{z}_i^v$ of target z_i^v and smooth label is $\mathring{\mathbf{z}}_i^v = \mathbf{z}_i^v \odot \eta + (1-\eta)/2$. since the smoothed label could provide training signals for neurons of both target and non-target labels, we omit the KL term here. The summation term, with Beta distribution parameters $\ddot{\boldsymbol{\alpha}}_i^v$ of referral opinion, changes to follows,

$$\sum_{j=1}^{2} \mathring{\mathbf{z}}_{ij}^v(\psi(\ddot{\boldsymbol{\alpha}}_{i1}^v + \ddot{\boldsymbol{\alpha}}_{i2}^v) - \psi(\ddot{\boldsymbol{\alpha}}_{ij}^v)) \tag{10}$$

3.4 Multi-View Agreement with Ground Truth (MVAGT)

The MVAGT (Multi-View Agreement with Ground Truth) is a novel evaluation metric designed specifically for multi-view classification problems with conflicting views. It assesses the model's performance on the test set by considering the ground truth labels, thus providing a more reliable and realistic measure of the model's ability to handle view disagreements. The rationality behind MVAGT lies in its alignment with real-world scenarios, where the majority agreement among multiple views is often considered more reasonable for the final decision. In the presence of view conflicts, a model that

[7] We provide a simplified version of training algorithm here for improving the readability and we direct readers to Appendix A for the detailed training algorithm

can make predictions consistent with the majority of views is deemed more trustworthy and reliable. By evaluating models using MVAGT, we can examine the reasonableness of the fused decision and assess the model's capability to handle view conflicts effectively. Mathematically, MVAGT calculates the accuracy of the model on the test set as follows:

$$\text{MVAGT} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1} \left(\sum_{v=1}^{V} \mathbb{1}((\hat{y}_i^v = y_i) > \frac{V}{2} \right) \tag{11}$$

where M is the total number of test samples, V is the number of views, $\hat{y}_i^v$ is the predicted label of the i-th sample from the v-th view, y^i is the ground truth label of the i-th sample, and $\mathbb{1}(\cdot)$ is the indicator function that returns 1 if the condition is satisfied and 0 otherwise.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets. Following previous work (Han et al., 2021; 2022; Jung et al., 2022; Xu et al., 2024a), we conducted experiments on six benchmark datasets: Handwritten[8], Caltech101 (Fei-Fei et al., 2004), PIE [9], Scene15 (Fei-Fei & Perona, 2005), HMDB (Kuehne et al., 2011) and CUB (Wah et al., 2011) with train-test split of 80% vs. 20%. A detailed description of these datasets is provided in the Appendix, we direct readers to the Appendix C.2 for further details regarding these datasets.

Compared Methods. We aim to resolve conflicts among predictions of different views, so we consider the methods that generate view-specific predictions which could have potential conflicts, and thus included following baselines: Fusion by Majority Voting (F-Mode) and Fusion by Probability Averaging (F-Avg), which are two commonly used fusion methods in most MVC methods. We also consider existing Evidential MVC baselines, TMC (Han et al., 2021), and the conflict resolution pioneering work ECML (Xu et al., 2024a). Recent work, TMNR (Xu et al., 2024b) applied Evidential MVC for noisy label learning, and CCML (Liu et al., 2024) derived consistent evidence among shared information by dynamically decoupling the consistent and complementary evidence [10]. Our method can also be extended to leverage the pseudo view, as demonstrated by its application to ETMC (Han et al., 2022), an extended version of TMC that incorporates pseudo views. We also compare with one multi-view uncertainty estimation baseline, MGP (Jung et al., 2022), in our experiments. We term our methods as TF and ETF where E indicates the pseudo-view is incorporated. All methods were run on a single 24GB RTX3090 card for fair comparison.

Evaluation Metrics. We evaluate MVC methods based on the reliability from prediction accuracy of fused opinion and the consistency among different views predictions. Similar to Han et al. (2021; 2022); Jung et al. (2022); Xu et al. (2024a), we measure the prediction accuracy using Top-1 Classification Accuracy, which checks whether the final predicted label of fused opinion is same as ground truth. Regarding to the consistency among various views' predictions, we apply the Fleiss Kappa (Fleiss, 1971), which is a statistical measure for assessing the agreement between different raters, with scores closer to 1 indicating higher agreement among the different predictions. The intuition behind using this two metrics is a reliable prediction should not be accurate only but also from most agreements. We also evaluate the model with the newly proposed metric, MVGAT, which measures the consistency of different views predictions with the ground truth label.

4.2 EXPERIMENT RESULTS AND ANALYSIS

For each individual metric, mean and standard deviation from ten runs with ten different random seeds are reported. In all tables, the best-performing method is highlighted in bold, and the second-best method is underlined.

Predictions Accuracy via Top-1 Accuracy. Similar to Han et al. (2021; 2022); Jung et al. (2022); Xu et al. (2024a), we first evaluated the model performance on the test split by Top-1 Classification Accuracy, as shown in Table 4. Building on the strengths of pseudo view, our method (ETF)

[8] `https://archive.ics.uci.edu/ml/datasets/Multiple+Features`

[9] `http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html`

[10] We re-run the official implementation of ECML, TMNR, CCML with our data loader for fair comparison.

Table 4: Top-1 accuracy on test split.

Dataset	Handwritten	Caltech101	PIE	Scene15	HMDB	CUB	AVG
F-Mode	99.10±0.20	94.13±0.08	79.41±0.00	62.45±0.11	51.70±0.41	70.25±0.38	76.13
F-Avg	99.25±0.00	**95.59±0.06**	90.59±0.29	76.21±0.09	71.49±0.35	92.75±0.53	87.65
MGP	99.60±0.10	94.42±0.20	90.13±0.87	74.30±0.41	73.97±0.15	90.79±1.03	87.03
ECML	99.57±0.11	94.25±0.08	91.40±0.47	64.34±0.11	72.90±0.11	92.58±0.25	85.84
TMNR	99.72±0.08	94.31±0.09	89.34±0.59	74.14±0.13	73.46±0.15	92.25±0.38	87.21
CCML	99.00±0.00	94.64±0.10	93.09±0.36	73.97±0.15	72.59±0.42	<u>93.83±0.41</u>	87.91
TMC	99.63±0.13	94.30±0.13	87.43±0.90	73.99±0.19	73.30±0.18	92.50±0.37	86.60
TF(ours)	99.68±0.11	<u>95.26±0.10</u>	<u>93.31±0.40</u>	77.83±0.32	<u>74.35±0.09</u>	93.33±0.75	<u>88.96</u>
ETMC	<u>99.75±0.00</u>	94.41±0.11	91.69±0.47	<u>78.41±0.20</u>	74.01±0.19	93.67±0.41	88.74
ETF(ours)	**99.98±0.07**	95.07±0.08	**94.63±0.34**	**82.01±0.17**	**75.55±0.15**	**94.08±0.38**	**90.22**

consistently outperforms all evidential MVC methods and the multi-view uncertainty estimation method, and gains the best average performance over six datasets compared with all baselines. For example, on the PIE and Scene15 datasets, the use of referral trust boosts the accuracy of ETMC by 2.94% and 3.60%, respectively. Moreover, ETF surpasses the pioneering conflict resolving method ECML by a substantial margin of 3.23% on PIE, 9.66% on Scene15 and 2.65% on HMDB, highlighting better power of conflicts handling of our method. It is worth noting that Caltech101 inherently has lower level of conflicts, as corroborated by high accuracy and Fleiss' Kappa scores (Table 5) of all baselines. Nevertheless, ETF maintains the compatible performance with the best one, F-Avg (a minor decrease of 0.52%), and still outperforms other methods, e.g., improve the accuracy of ETMC by 0.66%.

When compared to well-established methods like TMC, MGP, and ECML without pseudo views, our method TF consistently demonstrates superior performance across all datasets. For example, our proposed trust discounting method enhance TMC's performance by 3.84% on Scene15 and 5.88% on PIE, while also achieving the highest Top-1 accuracy on other datasets. Notably, our method TF, even without incorporating pseudo views, exhibits comparable performance to ETMC with pseduo views. For instance, TF outperforms ETMC on three datasets (Caltech101, PIE, and HMDB) out of a total of six.

Table 5: Fleiss' Kappa on test splits.

Dataset	Handwritten	Caltech101	PIE	Scene15	HMDB	CUB	AVG
F-Mode	0.63±0.04	**0.97±0.00**	0.38±0.00	<u>0.42±0.00</u>	<u>0.56±0.00</u>	**0.71±0.01**	<u>0.61</u>
F-Avg	0.54±0.03	**0.97±0.00**	0.37±0.01	<u>0.42±0.00</u>	0.55±0.01	0.58±0.06	0.57
MGP	0.59±0.05	0.94±0.00	0.21±0.01	0.33±0.00	0.51±0.00	0.43±0.07	0.50
ECML	0.42±0.05	<u>0.95±0.00</u>	<u>0.40±0.01</u>	0.26±0.00	0.53±0.01	0.44±0.07	0.50
TMNR	0.59±0.02	0.94±0.01	0.29±0.02	0.30±0.00	0.53±0.00	0.37±0.06	0.50
CCML	0.64±0.04	0.91±0.01	0.39±0.01	0.36±0.01	0.53±0.01	0.63±0.04	0.58
TMC	0.54±0.07	0.94±0.01	0.23±0.02	0.30±0.01	0.52±0.01	0.37±0.19	0.48
TF(ours)	<u>0.65±0.02</u>	<u>0.95±0.00</u>	0.36±0.01	0.39±0.00	0.54±0.00	0.51±0.10	0.57
ETMC	0.66±0.01	0.84±0.00	0.28±0.04	0.37±0.00	-0.15±0.04	0.45±0.10	0.41
ETF(ours)	**0.76±0.02**	<u>0.95±0.00</u>	**0.48±0.01**	**0.48±0.01**	**0.65±0.00**	<u>0.64±0.03</u>	**0.66**

Predictions Consistency via Fleiss' Kappa and MVGAT. To further validate the effectiveness of our proposed method, we evaluate it with two additional metrics, Fleiss' Kappa Fleiss (1971) and our proposed metric, MVAGT. As depicted in Table 5, our method (ETF) achieves the highest Fleiss' Kappa score on four datasets (Handwritten, PIE, Scene15, HMDB and CUB). Even through ETF does not rank first on the remaining two datasets (the third on Caltech101 and the second on CUB), it remains the most generalizable model with the highest average Fleiss' Kappa (0.66). It's worth noting that while our methods assume the existence of conflicts, Caltech101 is a dataset with fewer conflicts, which explains the performance discrepancy in Table 4. Nevertheless, ETF still outperforms other evidential or the MGP and enhances the robustness of ETMC with an improvement of approximately 13% on Caltech101. Moreover, it's essential to highlight that ETMC exhibits extremely poor agreement on HMDB with a negative value of -0.15. However, by applying our method, ETF significantly improves performance by an absolute value of 0.8. This underscores the relative robustness of our method across different datasets.

Table 6: MVAGT on test split.

Dataset	Handwritten	Caltech101	PIE	Scene15	HMDB	CUB
F-Mode	88.87±1.73	94.13±0.08	79.41±0.00	62.54±0.11	51.70±0.41	70.25±0.38
F-Avg	18.78±5.89	93.89±0.24	17.06±1.22	27.70±0.36	51.18±0.51	59.50±5.25
MGP	81.37±5.73	91.55±0.29	63.20±2.31	52.10±0.41	50.43±0.42	42.50±9.26
ECML	74.08±0.61	91.05±0.27	78.46±1.19	41.91±0.31	50.95±0.48	48.58±5.36
TMNR	86.80±1.03	90.92±0.18	65.15±3.68	51.86±0.61	50.48±0.47	36.58±6.42
CCML	86.78±1.42	88.97±1.09	81.91±1.40	55.23±0.84	51.34±0.91	63.67±2.61
TMC	81.58±6.57	90.27±0.38	51.54±3.00	51.42±0.46	50.37±0.45	43.25±14.8
TF(ours)	88.97±0.61	92.01±0.22	80.59±0.75	60.41±0.52	52.47±0.35	54.33±7.54
ETMC	98.10±0.17	92.41±0.32	75.15±4.13	73.75±0.45	8.45±1.09	91.08±1.06
ETF(ours)	**98.53±0.08**	**94.47±0.12**	**90.37±0.40**	**79.18±0.38**	**71.43±0.32**	**91.17±0.67**

While a multi-view classification (MVC) classifier with both high accuracy and high Fleiss' Kappa score generally suggests good reliability, high Fleiss' Kappa scores without reference to the ground truth label might be misleading, particularly in cases where the majority of views agree on the same incorrect class. Therefore, we propose a new evaluation metric (MVAGT), specifically tailored for conflict MVC scenarios. MVAGT assesses correctness on the test split by verifying that more than half of the views make correct decisions. Since majority agreement is often more reasonable for final decisions in real-world scenarios, evaluating methods using MVAGT ensures the reasonableness of the fused decision. As depicted in Table. 6, ETF demonstrates superior performance compared to other methods. Moreover, ETF exhibits good generalizability across different datasets, where ETMC experiences significant decreases (e.g., HMDB) or other methods alternately occupy the second-best position.

Discussion on consistency improvement. It is worth noting that applying TD solely on existing functional opinions will not improve the consistency among different views, however, our methods show that the consistency of opinions from different views is significantly improved, as measured by Fleiss Kappa and MVGAT. We attribute this improvement to the incorporation of TD in the training stage. The functional opinion will be discounted accordingly by the referral opinion, and it thus receive larger magnitude of gradients from the loss term, e.g., at the Substage 2b in Alg. 2, due to interactions between different opinions, e.g., Eq.2. Therefore, the functional opinion will be enforced to align with the ground truth which leads to the improved consistency among different views' opinions.

5 Conclusion

In this paper, we introduced a theoretically-grounded approach for resolving conflicts in Multi-View Classification. This approach is built on top of the principle of the Trust Discounting in Subjective Logic, where the computational trust, aka referral trust, is represented as a Binomial opinion with a Beta probability density function.The functional trust is then discounted by the amount computed as a function of the degree of trust. We demonstrated through extensive experiments that the proposed trust discounting method not only benefits classification accuracy but also increases consistency among different views, providing a new reliable approach to handling conflicts in MVC.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 647–655. PMLR, 2014.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 524–531, 2005.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition workshop*, pp. 178–178, 2004.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378, 1971.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Dongdong Hou, Yang Cong, Gan Sun, Jiahua Dong, Jun Li, and Kai Li. Fast multi-view outlier detection via deep encoder. *IEEE Transactions on Big Data*, 8(4):1047–1058, 2020.

Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33:2892–2902, 2020.

Zongmo Huang, Yazhou Ren, Xiaorong Pu, Shudong Huang, Zenglin Xu, and Lifang He. Self-supervised graph attention networks for deep weighted multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7936–7943, 2023.

Audun Jøsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Audun Jøsang, Tanja Ažderska, and Stephen Marsh. Trust transitivity and conditional belief reasoning. In *Trust Management VI: 6th IFIP WG 11.11 International Conference, IFIPTM 2012, Surat, India, May 21-25, 2012. Proceedings 6*, pp. 68–83, 2012.

Audun Jøsang, Paulo CG Costa, and Erik Blasch. Determining model correctness for situations of belief fusion. In *Proceedings of the 16th International Conference on Information Fusion*, pp. 1886–1893, 2013.

Audun Jøsang, Magdalena Ivanovska, and Tim Muller. Trust revision for conflicting sources. In *2015 18th International Conference on Information Fusion (Fusion)*, pp. 550–557, 2015.

Myong Chol Jung, He Zhao, Joanna Dipnall, Belinda Gabbe, and Lan Du. Uncertainty estimation for multi-view data: the power of seeing the whole picture. *Advances in Neural Information Processing Systems*, 35:6517–6530, 2022.

Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. *Advances in Neural Information Processing Systems*, 36, 2023.

Hengyuan Kang, Liming Xia, Fuhua Yan, Zhibin Wan, Feng Shi, Huan Yuan, Huiting Jiang, Dijia Wu, He Sui, Changqing Zhang, et al. Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. *IEEE transactions on medical imaging*, 39(8): 2606–2614, 2020.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pp. 2556–2563, 2011.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR, 2014.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 2024.

Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7585–7593, 2022.

Ying Liu, Lihong Liu, Cai Xu, Xiangyu Song, Ziyu Guan, and Wei Zhao. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7269–7277, 2024.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

C Wah, S Branson, P Welinder, P Perona, and S Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pp. 1083–1092. PMLR, 2015.

Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15712–15721, 2023a.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:16129–16137, 2024a.

Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. *arXiv preprint arXiv:2404.11944*, 2024b.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2577–2585, 2019.

Chaoyang Zhang, Zhengzheng Lou, Qinglei Zhou, and Shizhe Hu. Multi-view clustering via triplex information maximization. *IEEE Transactions on Image Processing*, 2023a.

Pei Zhang, Siwei Wang, Liang Li, Changwang Zhang, Xinwang Liu, En Zhu, Zhe Liu, Lu Zhou, and Lei Luo. Let the data choose: Flexible and diverse anchor graph fusion for scalable multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11262–11269, 2023b.

Lecheng Zheng, Yu Cheng, Hongxia Yang, Nan Cao, and Jingrui He. Deep co-attention network for multi-view subspace learning. In *Proceedings of the Web Conference 2021*, pp. 1528–1539, 2021.

A PROPOSED ALGORITHM FOR TRAINING AND TESTING

Algorithm 2: Algorithm For Training

Input: Multi-view dataset: $\mathcal{D} = \{\{\mathbf{x}_i^v\}_{v=1}^V, y_i\}_{i=1}^N$.
Initialize: Initialize the parameters $\acute{\theta}$ of the Functional networks; initialize the parameters $\ddot{\theta}$ of the Referral networks.
/*Stage-1 Warm-up Referral Network*/
for *minibatch* **do**
  for $v = 1 : V$ **do**
    $\ddot{\mathbf{e}}^v \leftarrow$ Referral Evidential network batch output;
    Obtain $\ddot{\boldsymbol{\alpha}}^v \leftarrow \ddot{\mathbf{e}}^v + 1$;
  end
  Obtain overall loss by summing losses calculated by Eq. 10 of all $\{\ddot{\boldsymbol{\alpha}}^v\}_{v=1}^V$;
  Update the parameters $\ddot{\theta}$ by gradient descent with the loss from above;
end
/*Stage-2 Update Functional Network*/
for *minibatch* **do**
  /*Substage-2a*/
  for $v = 1 : V$ **do**
    $\acute{\mathbf{e}}^v \leftarrow$ Functional Evidential network batch output;
    Obtain $\acute{\boldsymbol{\alpha}}^v \leftarrow \acute{\mathbf{e}}^v + 1$;
  end
  Obtain overall loss by summing losses calculated by Eq. 8 of all $\{\acute{\boldsymbol{\alpha}}^v\}_{v=1}^V$;
  Update the parameters $\acute{\theta}$ by gradient descent with the loss from above;
  /*Substage-2b*/
  for $v = 1 : V$ **do**
    $\ddot{\mathbf{e}}^v \leftarrow$ Referral Evidential network batch output;
    $\acute{\mathbf{e}}^v \leftarrow$ Functional Evidential network batch output;
    Obtain $\ddot{\omega}^v$ and $\acute{\omega}^v$ by Eq. 1 with $\ddot{\mathbf{e}}^v$ and $\acute{\mathbf{e}}^v$, respectively ;
  end
  Obtain joint opinion $\bar{\omega}$ by Eq. 6 and $\bar{\boldsymbol{\alpha}}$ of this opinion by reversing Eq. 1;
  Obtain loss by Eq. 8 with $\bar{\boldsymbol{\alpha}}$ and update the parameters $\acute{\theta}$ with gradient descent;
end
/*Stage-3 Adjust Referral Network*/
By repeating Stage-2b only and update $\ddot{\theta}$ instead of $\acute{\theta}$;
/*Stage-4 Adjust Functional Network*/
By repeating entire Stage-2;
Output: Functional and Referral networks parameters.

Algorithm 3: Algorithm For Testing

/*Testing Phase*/
Requires: the parameters $\acute{\theta}$ of the Functional networks; the parameters $\ddot{\theta}$ of the Referral networks.
for *minibatch* **do**
  for $v = 1 : V$ **do**
    $\ddot{\mathbf{e}}^v \leftarrow$ Referral Evidential network batch output;
    $\acute{\mathbf{e}}^v \leftarrow$ Functional Evidential network batch output;
    Obtain $\ddot{\omega}^v$ and $\acute{\omega}^v$ by Eq. 1 with $\ddot{\mathbf{e}}^v$ and $\acute{\mathbf{e}}^v$, respectively ;
  end
  Obtain joint opinion $\bar{\omega}$ by Eq. 6 and $\bar{\boldsymbol{\alpha}}$ of this opinion by reversing Eq. 1;
  Obtain predicted labels of minibatch using $\arg\max$ over belief masses.
end
Output: Predicted Labels and Opinions including fused opinion, functional opinions, referral opinions, discounted opinions for each instance of each view.

B PROOFS AND DERIVATIONS

B.1 CALCULATION OF PREDICTIVE PROBABILITY

According to Subjective Logic (SL) Jøsang (2018), the predictive probability p_k for class k, can be calculated by

$$p_k = b_k + a_k * u \tag{12}$$

where b_k is the belief mass for k-th label, u is the predictive uncertainty or epistemic uncertainty Sensoy et al. (2018). We usually assume the prior a_k conforms to a uniform discrete distribution, i.e., $a_k = 1/K$, so the above equation is identical to

$$p_k = \frac{\alpha_k}{S} \tag{13}$$

where α_k is the Dirichlet concentration parameter for k-th label, and S is the Dirichlet strength, i.e., $S = \sum_k \alpha_k$.

Proof.

$$\begin{aligned} p_k &= b_k + a_k * u \\ &= b_k + \frac{1}{K} * \frac{K}{S} \\ &= \frac{e_k}{S} + \frac{1}{S} \\ &= \frac{\alpha_k}{S} \end{aligned}$$

□

Since Beta Distribution is 2-dimensional Dirichlet Distribution, above equations for calculating probabilities of multinomial opinions could also be applied to binomial opinions.

B.2 ALTERNATIVE REPRESENTATION OF BELIEF CONSTRAINT FUSION(BCF)

Proof. We the proof for Eq. 4 as follows,

$$\begin{aligned} e_k &= S * b_k \\ &= S\,\frac{1}{1-C}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1) \\ &= S\,\frac{1-\sum_k b_k}{u^1 u^2}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1) \\ &= (S - S * \sum_k b_k)\,\frac{1}{u^1 u^2}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1) \\ &= (S - \sum_k e_k)\,\frac{1}{u^1 u^2}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1) \\ &= K\,\frac{1}{u^1 u^2}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1) \\ &= K\,\frac{1}{u^1 u^2}(\frac{e_k^1 e_k^2}{S^1 S^2} + \frac{e_k^1 u^2}{S^1} + \frac{e_k^2 u^1}{S^2}) \\ &= K(\frac{e_k^1 e_k^2}{K * K} + \frac{e_k^1 u^2}{K u^2} + \frac{e_k^2 u^1}{K u^1}) \\ &= \frac{e_k^1 e_k^2}{K} + e_k^1 + e_k^2 \end{aligned}$$

□

B.3 Dirichlet Evidence Updating by Trust Discounting (TD)

As mentioned earlier, the TD in Definition 2 also corresponds to updating Dirichlet evidence using following equation,

$$\breve{e}_k = \frac{\ddot{p}_t \acute{u}}{1 - \ddot{p}_t + \ddot{p}_t \acute{u}} \acute{e}_k \tag{14}$$

where $\ddot{p}_t$ is the probability representing trust degree and $\acute{u}$ is the uncertainty for functional opinion. $\acute{e}_k$ is Dirichlet evidence of functional opinion, and $\breve{e}_k$ is Dirichlet evidence after discounting.

Proof.

$$\begin{aligned}
\breve{\boldsymbol{e}}_k &= \breve{\boldsymbol{b}}_k * \breve{S} \\
&= \frac{\ddot{p}_t \acute{b}_k K}{\breve{u}} \\
&= \frac{\ddot{p}_t \acute{b}_k K}{1 - \ddot{p}_t + \ddot{p}_t \acute{u}} \\
&= \frac{\ddot{p}_t}{1 - \ddot{p}_t + \ddot{p}_t \acute{u}} \frac{\acute{e}_k}{\acute{S}} K \\
&= \frac{\ddot{p}_t}{1 - \ddot{p}_t + \ddot{p}_t \acute{u}} \frac{K}{\acute{S}} \acute{e}_k \\
&= \frac{\ddot{p}_t \acute{u}}{1 - \ddot{p}_t + \ddot{p}_t \acute{u}} \acute{e}_k
\end{aligned}$$

□

B.4 Detailed Proof of Propositions

Proof. Proof details of Proposition 1. Recall that scalar probability $\ddot{p}_t$ represents the degree of trust as mentioned before. The belief mass for k-th label of final fused opinion is as follows,

$$\begin{aligned}
\bar{b}_k &= \frac{1}{1 - \breve{C}} (\breve{b}_k^1 \breve{b}_k^2 + \breve{b}_k^1 \breve{u}^2 + \breve{b}_k^2 \breve{u}^1) \\
&= \frac{1}{1 - \breve{C}} ((\acute{b}_k^1 \ddot{p}_t^1)(\acute{b}_k^2 \ddot{p}_t^2) + \acute{b}_k^1 \ddot{p}_t^1 \breve{u}^2 + \acute{b}_k^2 \ddot{p}_t^2 \breve{u}^1)
\end{aligned}$$

We use g to denote the index of ground-truth label, and we have

$$\bar{b}_g = \frac{1}{1 - \breve{C}} ((\acute{b}_g^1 \ddot{p}_t^1)(\acute{b}_g^2 \ddot{p}_t^2) + \acute{b}_g^1 \ddot{p}_t^1 \breve{u}^2 + \acute{b}_g^2 \ddot{p}_t^2 \breve{u}^1)$$

The discounted opinion's uncertainty $\breve{u}$ is

$$\begin{aligned}
\breve{u} &= 1 - \ddot{p}_t (\sum_k \acute{b}_k) \\
&= 1 - \ddot{p}_t (1 - \acute{u}) \\
&= 1 - \ddot{p}_t + \ddot{p}_t * \acute{u}
\end{aligned}$$

In the warm-up training stage, the Eq. 10 is used to make sure $\ddot{p}_t \to 1$ (with hard targets for simplicity here) for those views' predictions are same as the ground truth label, and $\breve{u} \to 0$ for those views' predictions are incorrect. Therefore, $\breve{u} \to \acute{u}$ when $\acute{b}_g = max(\acute{\mathbf{b}})$, and $\breve{u} \to 1$ when $\acute{b}_g \neq max(\acute{\mathbf{b}})$.

Therefore, with the assumption that at least one-view's prediction is same the ground truth (i.e., correct label, let's say view 1's prediction is correct), we have

$$\begin{aligned}
\bar{b}_g &= \frac{1}{1 - \breve{C}} ((\acute{b}_g^1 \ddot{p}_t^1)(\acute{b}_g^2 \ddot{p}_t^2) + \acute{b}_g^1 \ddot{p}_t^1 \breve{u}^2 + \acute{b}_g^2 \ddot{p}_t^2 \breve{u}^1) \\
&\geq \frac{1}{1 - \breve{C}} ((\acute{b}_k^1 \ddot{p}_t^1)(\acute{b}_k^2 \ddot{p}_t^2) + \acute{b}_k^1 \ddot{p}_t^1 \breve{u}^2 + \acute{b}_k^2 \ddot{p}_t^2 \breve{u}^1 (\text{equality holds iif. } k = g)) \\
&= \frac{1}{1 - \breve{C}} (\breve{b}_k^1 \breve{b}_k^2 + \breve{b}_k^1 \breve{u}^2 + \breve{b}_k^2 \breve{u}^1) = \bar{b}_k
\end{aligned}$$

Besides the warm-up stage, in other training stages, such as training stage 3 in Alg.2, the $\ddot{p}_t$ will also be updated to maximize $\bar{b}_g$ based on the Eq. 8, i.e., $\bar{b}_g \geq \bar{b}_k$(equality holds iif. $k = g$. Therfore, the referral opinion is learnt to maximize the belief mass of ground truth label of the final fused opinion as well.

$\square$

Proof. Proof details of Proposition 2. Let $\bar{u}$ and $\bar{u}'$ denote the uncertainty of BCF combined opinion with or without Trust Discounting, respectively.

$$
\begin{aligned}
\bar{u} &= \frac{1}{\sum_{k=1}^{K}(\frac{\breve{b}_k^1 \breve{b}_k^2}{\breve{u}^1 \breve{u}^2} + \frac{\breve{b}_k^1}{\breve{u}^1} + \frac{\breve{b}_k^2}{\breve{u}^2}) + 1} \\
&= \frac{1}{\sum_{k=1}^{K}(\frac{\acute{b}_k^1 \ddot{p}_t^1 \acute{b}_k^2 \ddot{p}_t^2}{(\acute{u}^1 \ddot{p}_t^1 + 1 - \ddot{p}_t^1)(\acute{u}^2 \ddot{p}_t^1 + 1 - \ddot{p}_t^2)} + \frac{\acute{b}_k^1 \ddot{p}_t^1}{\acute{u}^1 \ddot{p}_t^1 + 1 - \ddot{p}_t^1} + \frac{\acute{b}_k^2 \ddot{p}_t^2}{\acute{u}^2 \ddot{p}_t^1 + 1 - \ddot{p}_t^2}) + 1} \\
&= \frac{1}{\sum_{k=1}^{K}(\frac{\acute{b}_k^1 \acute{b}_k^2}{(\frac{\acute{u}^1}{\ddot{p}_t^2} + \frac{1}{\ddot{p}_t^1 \ddot{p}_t^2} - \frac{1}{\ddot{p}_t^2})(\frac{\acute{u}^2}{\ddot{p}_t^1} + \frac{1}{\ddot{p}_t^1 \ddot{p}_t^2} - \frac{1}{\ddot{p}_t^1})} + \frac{\acute{b}_k^1}{\acute{u}^1 + \frac{1}{\ddot{p}_t^1} - 1} + \frac{\acute{b}_k^2}{\acute{u}^2 + \frac{1}{\ddot{p}_t^2} - 1}) + 1} \\
&= \frac{1}{\sum_{k=1}^{K}(\frac{\acute{b}_k^1 \acute{b}_k^2}{(\frac{\acute{u}^1}{\ddot{p}^2} + \frac{1 - \ddot{p}^1}{\ddot{p}^1 \ddot{p}^2})(\frac{\acute{u}^2}{\ddot{p}^1} + \frac{1 - \ddot{p}^2}{\ddot{p}^1 \ddot{p}^2})} + \frac{\acute{b}_k^1}{\acute{u}^1 + \frac{1}{\ddot{p}^1} - 1} + \frac{\acute{b}_k^2}{\acute{u}^2 + \frac{1}{\ddot{p}^2} - 1}) + 1} \\
&\geq \frac{1}{\sum_{k=1}^{K}(\frac{\acute{b}_k^1 \acute{b}_k^2}{\acute{u}^1 \acute{u}^2} + \frac{\acute{b}_k^1}{\acute{u}^1} + \frac{\acute{b}_k^2}{\acute{u}^2}) + 1} = \bar{u}'
\end{aligned}
$$

$\square$

B.5 Loss Functions and Hyperparameters for Optimization

Recall that the probability density function (pdf) of the Dirichlet distribution, $\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha})$, is given by:

$$
\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1}
$$

where:

- $\mathbf{p} = (p_1, p_2, \dots, p_K)$ is a probability vector, such that $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$ for all k.
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)$ is a vector of concentration parameters, with $\alpha_k > 0$.
- $B(\boldsymbol{\alpha})$ is the multivariate Beta function, defined as $B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$.
- $\Gamma(\cdot)$ is the Gamma function.

Recall that our loss function for Dirichlet Parameters $\boldsymbol{\alpha}$ is

$$
L_i^v = \sum_{k=1}^{K} \mathbf{y}_{i,k}(\psi(S_i^v) - \psi(\alpha_{i,k}^v)) + \lambda_o \mathrm{D}_{KL}[\text{Dir}(\mathbf{p}_i^v | \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}^{\boldsymbol{v}}) || \text{Dir}(\mathbf{p}_i^v | \mathbf{1})]
$$

Specifically, the left summation term is derived from the Bayes risk for Cross-Entropy loss with a Dirichlet distribution, which is also dentoed as L_{ace} in previous work (Han et al., 2021). We omit the index of view v and instance i for simplicity, so L_{ace} is defined as follows,

$$
\begin{aligned}
L_{ace} &= \int \left[\sum_{k=1}^{K} -\mathbf{y}_k log(p_k) \right] \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} (p_k)^{\alpha_k - 1} d\mathbf{p} \\
&= \sum_{k=1}^{K} \mathbf{y}_k (\psi(S) - \psi(\alpha_k)) \qquad (15)
\end{aligned}
$$

Where ψ is the digamma function.

Recall that our referral network will generate the evidence for binomial opinion, and the evidence will be converted into parameters of Beta Distribution, i.e., $Beta(\alpha_0, \alpha_1)$ Subsequently, by replacing the Dirichlet Distribution with Beta Distribution, and the label y_k in above equation with another label, we can have the ace loss for Beta Distribution, as Eq. 10.

And the right term, KL divergence loss is

$$\begin{aligned} &\mathrm{D}_{KL}\left[\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) \parallel \mathrm{Dir}(\mathbf{p} \mid \mathbf{1})\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\alpha_k)}\right) + \sum_{k=1}^{K}(\alpha_k - 1)\left[\psi(\alpha_k) - \psi\left(\sum_{j=1}^{K}\alpha_j\right)\right] \end{aligned} \tag{16}$$

C Additional Details of The Experiment

C.1 Hyper-parameters of Proposed Methods

The hyper-parameters for training TF and ETF has been shown in in Table 7. Concretely, ”lr” is the learning rate for functional networks, ”rlr” indicates the learning rate for referral networks. For the ”lr”, we follow ETMC (Han et al., 2022), and used same strategy to select learning rate for the functional nets. When tuning the learning rate for referral networks, we follow a basic principle of starting with a value less than or equal to the base learning rate, and then gradually decreasing the learning rate of referral network by a factor of three. For fair comparison, we used same learning rate for functional networks for evidence-based methods, except MGP (Jung et al., 2022), for which we followed their paper.

Table 7: TF and ETF hyper-parameters

Hyper-parameter	Handwritten	Caltech101	PIE	Scene15	HMDB	CUB
lr	3e-3	1e-4	3e-3	1e-2	3e-4	1e-3
rlr	3e-4	3e-5	1e-3	3e-3	1e-4	3e-4
weight-decay	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
warm-up epochs	1	1	1	1	1	1

The Adam optimizer (Kingma & Ba, 2015) is used for updating model parameters with beta coefficients = (0.9, 0.999) and epsilon = 1e-8.

C.2 Summary of Dataset

Table 8: Summary of Datasets

Dataset	Size	K	Dimensions	#Train	#Test
HandWritten	2000	10	240/76/216/47/64/6	1600	400
Caltech101	8677	101	4096/4096	6941	1736
PIE	680	68	484/256/279	544	136
Scene15	4485	15	20/59/40	3588	897
HMDB	6718	51	1000/1000	5374	1344
CUB	600	10	1024/300	480	120

We provide the summary of the dataset in Table 8, we direct readers to Han et al. (2021) for further details regarding these datasets. The datasets used in our experiments are 1) **Handwritten** dataset has 2000 samples of 10 classes. Each class is one of the digit 0 to 9 with samples evenly distributed (i.e., 200 samples per class). We use six descriptors to represent different views, and they are Pixel averages in 2 × 3 windows (Pix) feature with 240 dimensions, Fourier coefficients of the character shapes (FOU) with 76 dimensions, Profile correlations (FAC) features with 216 dimensions, Zernike moments (ZER) with 47 dimensions, Karhunen-Love coefficients (KAR) with 64 dimensions, and Morphological (MOR) features with 6 dimensions; 2) **Caltech101** dataset has 101 classes and 8677 images in total; We used the extracted features from DECAF Donahue et al. (2014) and VGG19 Simonyan & Zisserman (2014). Both views have 4096 dimensions. 3) **PIE** dataset includes intensity (484 dimensions), Local binary patterns (LBP) (256 dimensions) and Gabor feature (279 dimensions) of 680 facial images, with 68 subjects; 4) **Scene15** dataset has 4485 images from 15 indoor and outdoor scene categories. There are 3 different views information, and they are GIST, Pyramid Histogram of Oriented Gradients (PHOG) and Local binary patterns (LBP) feature. These views are in 20, 59 and 40 dimensions respectively; 5) **HMDB** has 6718 samples of 51 categories of actions, which is consisted of Histogram of oriented gradients (HOG) feature and Motion Boundary Histograms (MBH) features as a 2-view dataset. Both views have 1000 dimensions; 6) **CUB** dataset has 200 different categories of birds and 11788 images in total. Same as Han et al. (2021), we used first 10 categories in our experiment and GoogleNet Szegedy et al. (2015) and doc2vec Le & Mikolov (2014) to extract the image features and text features to simulate a 2-view dataset. Image view and text view has 1024 and 300 dimensions respectively.

D Supplementary Insights and Additional Analysis

D.1 AUROC for Uncertainty.

The uncertainty score, as illustrated in Proposition 2, will be more accurate withou introducing biases, so it is essential to validate the increased uncertainty. Following the approach of prior work (Filos et al., 2019), we assess uncertainty to ensure a thorough evaluation. Specifically, we employed AUROC to measure the model's discriminate power in distinguishing incorrect predictions using uncertainty scores. As shown in Table 9, TF and ETF consistently demonstrate the best performance on five out of the six datasets, showcasing their robust generalizability. Despite a performance decrease on the CUB dataset, our method (ETF) still maintains the second-best result, outperforming other approaches, whether incorporating pseudo views or not. One possible reason for the decreased performance on CUB could be the unstable optimization caused by the limited number of training instances (e.g., 480), whereas other datasets, such as Scene15, contain significantly more instances (e.g., 3588).

Table 9: AUROC of uncertainty scores for identifying incorrect predictions.

Dataset	Handwritten	Caltech101	PIE	Scene15	HMDB	CUB
MGP	99.29±0.30	87.62±0.90	88.43±0.67	63.92±1.96	82.87±0.60	58.20±11.4
ECML	79.05±5.62	86.31±0.50	87.51±0.49	60.50±0.25	81.63±0.15	57.30±8.50
TMC	99.23±0.22	87.33±0.47	90.16±0.99	62.60±0.54	82.63±0.48	64.80±10.5
TF(ours)	<u>99.32±0.35</u>	**88.99±0.54**	**95.90±0.08**	64.56±2.02	83.59±0.23	53.52±14.3
ETMC	99.30±0.19	88.35±0.63	<u>93.02±1.40</u>	<u>66.49±0.44</u>	<u>85.42±0.34</u>	**72.56±8.11**
ETF(ours)	**99.90±0.30**	<u>88.70±0.54</u>	92.47±1.19	**70.44±1.10**	**86.23±0.49**	<u>64.41±3.54</u>

D.2 Ablation Study of Warm-up Epochs

In the proposed stage-wise training algorithm, we adopt a warm-up stage (i.e., training stage 1) for better initialization of referral networks. As random initialized parameters may not able to assess the reliability of corresponding functional opinions correctly. The key hyper-parameter of the warm-up stage, is the warm-up epochs. We ablate different values of this hyper-parameter and evaluate the effect of it on the performance of our method. Specially, we used an empirical value, i.e., one single epochs, for all reported results in the experiment section. And here we provide more analysis with finely grain values, starting from 0 and increasing steadily, for example, to 2, 5, and 10, that is first random initializing the parameters of the referral networks and then not warm-up training or training with 2, 5, 10, and followed by each, finish the rest training stages. Please note that if this value is set to be 0, which means we disable the warm-up stage, and reported results with warm-up epoch 1 are also included, as shown in Figure 3.

From Figure 3, we can find that incorporating warm-up stage (warm-up epochs ≥ 1) can generally results in better accuracy. For some datasets (e.g. HMDB), increasing the number of warm-up epochs further improves accuracy compared to the results previously reported. This observation suggests that adjusting this value based on the specific dataset can lead to enhanced performance.

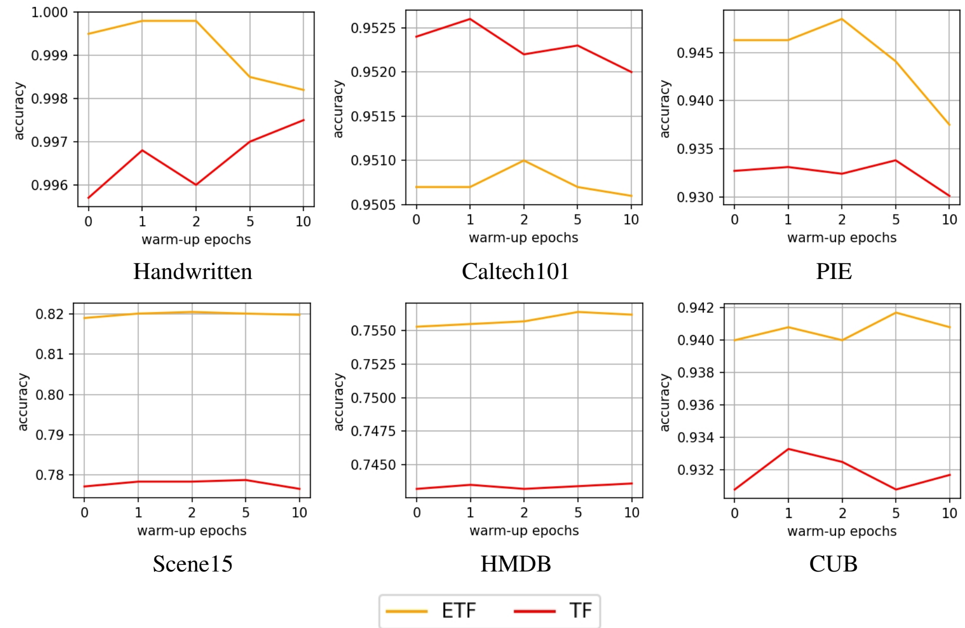

Figure 3: The effect of different warm-up epochs on testing accuracy.

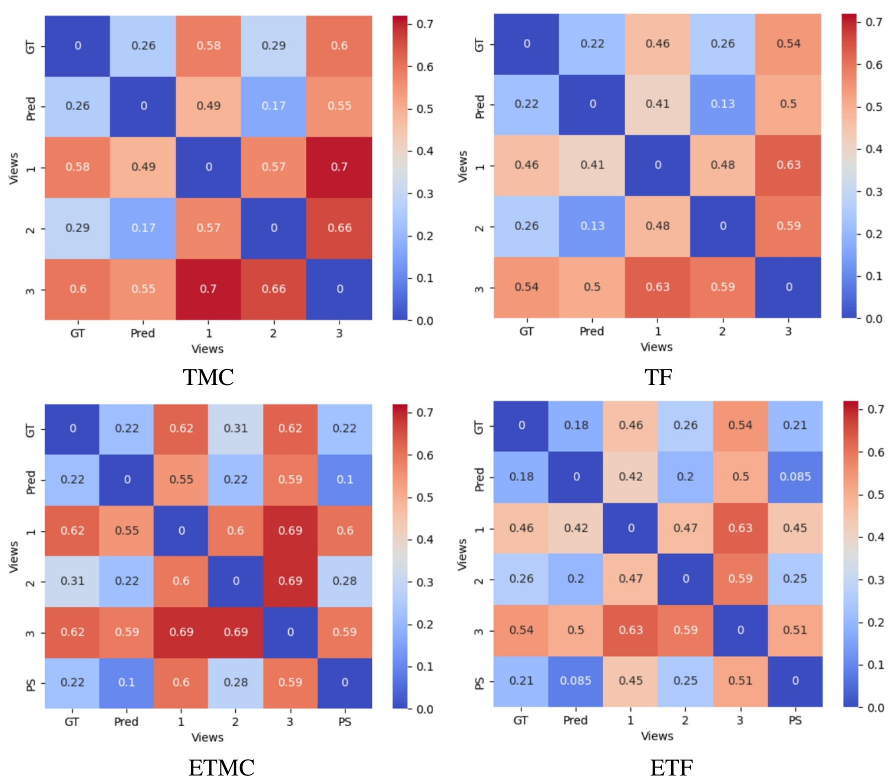

Figure 4: Conflict Ratio on Scene15, Four Methods TMC, TF, ETMC, ETF are compared. GT, Pred, 1, 2, 3 and PS are ground-truth, prediction, GIST, PHOG, LBP and pseudo view respectively.

D.3 Ablation Study of With or Without the TD module

We conduct the ablation study to validate the effectiveness of TD module. In the case without the TD module, the corresponding training stages related to TD module will be disabled, for example, the warm-up stage and training stage 2-b.

Table 10: Ablation Study of With or Without the TD module

Method	Top-1 Acc(%)	FleissKappa	MVGAT(%)	Uncer_AUROC(%)
ETF(w/ TD, reported)	82.01±0.17	0.48±0.01	79.18±0.38	70.44±1.10
ETF(w/o TD)	81.06±0.16	0.46±0.01	77.42±0.49	69.95±0.83
TF(w/ TD, reported)	77.83±0.32	0.39±0.00	60.41±0.52	64.56±2.02
TF(w/o TD)	76.82±0.33	0.37±0.01	59.04±0.71	63.54±1.50

We can see that without the core module TD, the performance over four metrics drops, which indicates the effectiveness of our proposed TD module.

D.4 Ablation Study of Smoothing Factor

We varied the smoothing factor used in the warm-up stage for ablation. we set warm-up epoch equals 1, which same as the reported results in the main text. The equation we used for smoothing hard label is $\mathring{\mathbf{z}}_i^v = \mathbf{z}_i^v \odot \eta + (1-\eta)/2$, with larger smoothing factor, the smoothed label becomes meaningless, so we vary the factor from 0.6 to 1.0 by step size 0.1.

Table 11: Ablation Study of Smoothing Factor

Method	Top-1 Acc(%)	FleissKappa	MVGAT(%)	Uncer_AUROC(%)
ETF(0.9, reported)	82.01±0.17	0.48±0.01	79.18±0.38	70.44±1.10
ETF(1.0)	82.07±0.12	0.48±0.01	79.32±0.38	71.03±0.40
ETF(0.8)	82.04±0.23	0.49±0.01	79.40±0.48	70.48±1.11
ETF(0.7)	82.07±0.10	0.48±0.01	79.42±0.28	70.53±0.82
ETF(0.6)	81.96±0.16	0.47±0.01	79.31±0.40	70.36±0.74

We can see that our method is relatively robust to different smoothing factors, and even gains performance improvement with adjusted smoothing factors on Scene15 Dataset, e.g., factor equals to 1.0, the smoothing factor we used in submission (e.g., 0.9) is the empirical value suggested in the original paper, to avoid hyper-parameters over-tuning.

D.5 The Effectiveness of Leveraging Different Views

We take the Scene15 dataset as example, and ablate the number of views to validate how the trust discounting mechanism performs with varying number of views.

Table 12: Test Accuracy by using different views on Scene15

Comb	N Views used	view 1	view 2	view 3	Top-1 Accuracy
1	1	✓	x	x	57.16±0.22
2	1	x	✓	x	75.15±0.01
3	1	x	x	✓	62.97±0.45
4	2	✓	✓	x	78.70±0.00
5	2	✓	x	✓	68.21±0.01
6	2	x	✓	✓	80.21±0.00
7	3	✓	✓	✓	82.01±0.17

Based on the table above, we observe that the effectiveness of each individual view on classification varies significantly, as reflected in the test accuracy of individual views. However, our method consistently improves accuracy by effectively incorporating different views. For instance, View 2, View 3, and View 1 rank 1st, 2nd, and 3rd, respectively, in terms of single-view accuracy. Combination

4 (View 1 & 2) outperforms Combination 5 (View 1 & 3) across comparisons, and in such case, view 1 are the common view, but view 2 is better than 3, so combination 4 is expected to outperform combination 5, and this holds true when comparing Combination 4 with Combination 6, or comparing Combination 5 with Combination 6. The highest accuracy is achieved when all views are utilized together, which also proves the effectiveness of our method.

D.6 Instance Similarity of Vector Datasets

We also calculated the pair-wise cosine similarities and provided both the results and an analysis accordingly. Specifically, we considered to calculate the instance similarity using pair-wise cosine similarity. Please note the AVG view means calculating instance similarity on each view first, then averaging over all views.

Table 13: View-Specific Pairwise Feature Similarity For Six Datasets

	View	Mean	Median	Min	Max
Handwritten	1	0.6268	0.6329	0.1249	1.0000
	2	0.8043	0.8095	0.4456	1.0000
	3	0.8586	0.8592	0.6304	1.0000
	4	0.7917	0.8038	0.2970	1.0000
	5	0.9167	0.9168	0.8137	1.0000
	6	0.7036	0.7964	0.0097	1.0000
	AVG	0.7836	0.7889	0.5350	1.0000
Caltech101	1	0.9684	0.9725	0.6968	1.0000
	2	0.9748	0.9792	0.5175	1.0000
	AVG	0.9716	0.9756	0.6263	1.0000
PIE	1	0.7518	0.7696	0.2842	0.9954
	2	0.7173	0.7203	0.4939	0.8530
	3	0.8613	0.8682	0.5598	0.9895
	AVG	0.7768	0.7829	0.5471	0.9395
Scene15	1	0.9038	0.9234	0.0538	1.0000
	2	0.8689	0.8904	0.1185	1.0000
	3	0.8133	0.8385	0.0072	1.0000
	AVG	0.8620	0.8789	0.1170	1.0000
HMDB	1	0.9372	0.9375	0.9002	1.0000
	2	0.9418	0.9418	0.8898	1.0000
	AVG	0.9395	0.9397	0.8970	1.0000
CUB	1	0.4112	0.3952	0.1346	0.9577
	2	0.9033	0.9128	0.5949	0.9972
	AVG	0.6572	0.6494	0.4153	0.9674

Based on the Table above, we can see that for some datasets, like Handwritten and CUB, different views show different statistics indicating the similarity varies significantly in different views. However, for other datasets, like HMDB and Caltech101, the instance similarity among different views are pretty similar.

As we calculated the pairwise similarity using the feature vectors of instances, this similarity also reflects the semantic similarity. Consequently, similar statistics among different views suggest that their classification performance is likely to be comparable.

1) For similar views: If one view achieves high accuracy, the other is likely to perform similarly, resulting in both high accuracy and consistency. For example, this is observed in the Caltech101 dataset (refer to Top-1 Accuracy and Fleiss Kappa). If one view performs with low accuracy, the other tends to perform similarly, leading to fused predictions that are consistently low in accuracy across views. An example of this can be seen in the HMDB dataset.

2) For dissimilar views: If one view achieves high accuracy while the other produces low-accuracy predictions, this leads to higher conflicts. But the accuracy of the fused prediction depends on the specific fusion mechanism employed by the method. Examples of this scenario can be observed in the Handwritten and CUB datasets.

D.7 End2End Training on Food101 Dataset

In order to further validate the effectiveness of our model on a large dataset, we use an additional dataset, Food101, which has both an image and text view. This is one dataset has the same number of class labels, 101, as Caltech101, and has more training (i.e., 61127), validation (i.e, 6845) and testing (i.e., 22716) instances.. We tried our best, but can only find this dataset having comparable statistics, e.g., number of class labels and instances.

Table 14: Test Performance on Food101

Method	Top-1 Acc
TMC	92.35±0.34
ETMC	92.49±0.13
ECML	92.53±0.15
CCML	92.70±0.06
TF	92.79±0.15
ETF	93.09±0.02

We trained all methods using pre-trained Resnet50 and base-uncased Bert as image and text encoder, and we adopt AdamW Optimizer for updating parameters. All other settings e.g., maximum number of epochs, are identical, and we run each method three times for reporting mean and standard deviation. We do not include TMNR here as it requires pre-extracted frozen feature vectors for computing similarity matrix working for noisy label learning, and we are not able to have frozen feature vectors in this End2End training case as the parameters of encoder will also be updated. Our method ETF consistently outperforms all other methods with regards classification accuracy as shown in the table.

D.8 Reduce Conflicts by Trust Fusion

We calculate the Conflict Ratio (CR) by normalizing the number of times that the v-th view prediction is different from w-th view, i.e., CR$(\hat{\boldsymbol{y}}^v, \hat{\boldsymbol{y}}^w) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}(\hat{y}_i^v \neq \hat{y}_i^w)$, where M is total number of test instances, $\hat{y}_i^w$ is the predicted label of i-th instance on w-th view, and $\mathbb{1}$ is the indicator function that returns 1 if the condition is satisfied and 0 otherwise. By applying Trust Discounting, both TMC's and ETMC's conflicts between different views are significant reduced. As an example, the CR on Scene15 is visualized by heatmap, shown in Figure D.1. The colors in the heatmap generated by our method are noticeably more blue (or less red) than those of the baselines, indicating that the conflict ratio has been reduced by our method.

D.9 Explanation for the Decrease of AUROC for Uncertainty

We argue the decreased performance of AUROC on whether uncertainty can indicate the correctness of predicted label in caused by insufficient training instances. As shown in Table 8, there are less than 550 training instances on PIE and CUB datasets, where our methods, ETF and TF, have decreased performance, compared to ETMC and TMC, in which the only difference is the TD module.

Besides, we also investigate a particular testing instance of CUB dataset for the decreased performance on AUROC of uncertainty. As the error case displayed in Figure 5, ETF corrects the error prediction made by ETMC. However, even though the combined prediction is correct after applying trust discounting, the predictive uncertainty is still relatively high. If ETF corrects previously incorrect predictions but assigns them relatively high uncertainty scores (e.g., 0.4), it may lead to a decrease in the AUROC for predictive uncertainty. This is because AUROC evaluates the model's ability to discriminate between correct and incorrect predictions based on uncertainty scores. Correcting predictions while maintaining high uncertainty scores can make it more challenging for the model to distinguish between correct and incorrect predictions, resulting in a lower AUROC score, even though the accuracy improves.

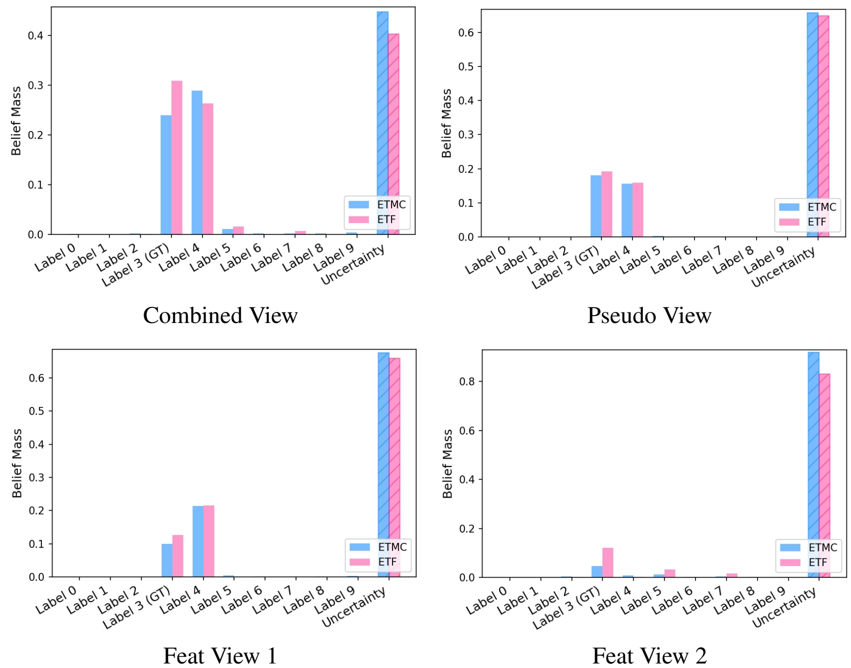

Figure 5: Bar chart for each label's belief mass and predictive uncertainty of one testing instance of CUB dataset. GT indicates the ground truth label of the selected instance.

D.10 Simulating Conflicting Predictions with Noisy Instances

We plot the model performance for Evidential MVC methods with various level of noises introduced to inputs in Figure 6 and Figure 7, for methods incorporate pseudo views and not incorporate pseudo views respectively. Our methods consistently outperforms other methods like TMC and ECML.

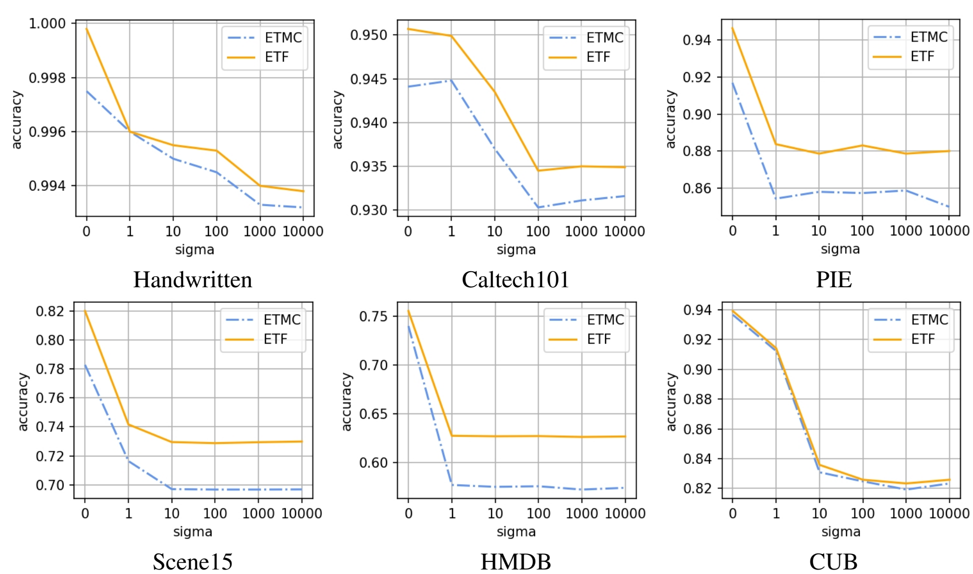

Figure 6: Performance of pseudo-view incorporated Evidential MVC methods on multi-view data with different levels of noise.

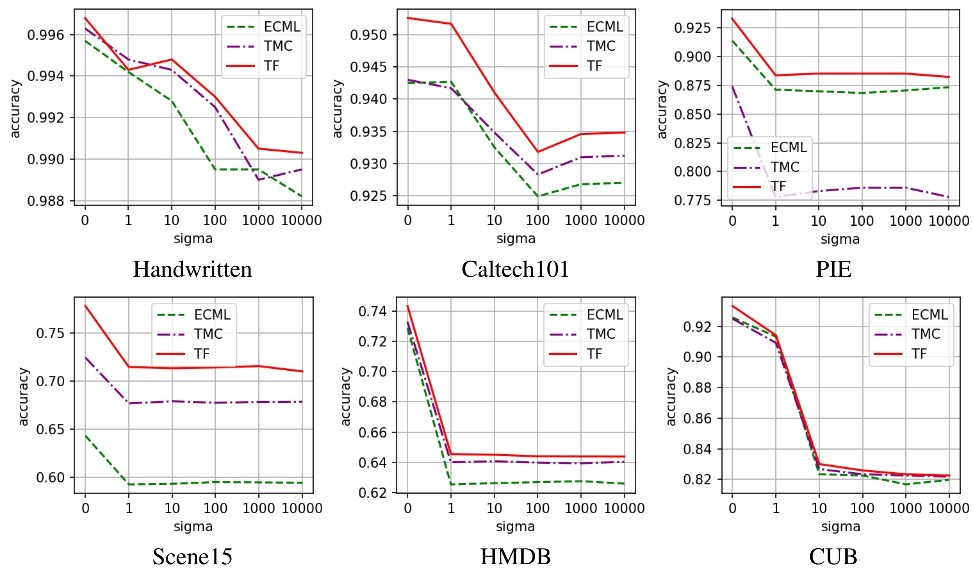

Figure 7: Performance of non pseudo-view incorporated Evidential MVC methods on multi-view data with different levels of noise.

D.11 Limitations

One possible limitation of our work is that the warm-up loss is not optimal solution, even though we explored the impact of different warm-up epochs and showed the effectiveness with using warm-up loss. Another possible limitation would be stage-wise training algorithm is time consuming, we leave it to future work for improving its efficiency.

E Technical Requirement and Execution

E.1 Execution Time

The proposed instance-wise approach does indeed introduce additional time complexity compared to the baselines, particularly compared to methods like TMC and ETMC that do not incorporate the TF Module but with same Belief Fusion method. However, our method does not rely on the dependencies between instances for computation. This allows us to perform batch-wise calculations during both training and testing, a practice widely adopted in most deep learning algorithms, which can enhance efficiency.

Table 15: Handwritten

Method	Train(Seconds)	Test(Seconds)
F-Avg	22.88±0.30	0.040±0.09
F-Mode	26.26±0.36	0.041±0.09
MGP	452.31±1.43	0.428±0.10
EMCL	52.63±1.15	0.041±0.09
TMC	55.46±0.78	0.042±0.09
TF	183.51±1.81	0.043±0.09
ETMC	62.45±0.95	0.042±0.09
ETF	202.15±2.24	0.044±0.09

Table 16: Caltech101

Method	Train(Seconds)	Test(Seconds)
F-Avg	78.62±0.95	0.063±0.09
F-Mode	94.01±0.87	0.063±0.09
MGP	2439.60±7.35	3.428±0.13
ECML	152.99±5.96	0.064±0.10
TMC	114.77±1.89	0.066±0.10
TF	463.41±10.65	0.067±0.09
ETMC	153.64±1.690	0.066±0.09
ETF	543.99±24.88	0.067±0.010

From another perspective, we can view the TF stage as an additional layer appended to the existing framework (e.g., TMC). Let h be the input vector with dimension d_h used for the classification task. For a K-class classification problem, we obtain a $K+1$-dimensional functional opinion (1 dimension for uncertainty). The weight matrix W of the proposed BiLinear layer will have dimensions d_h x d_{K+1} x d_2, and the bias vector will have dimension d_2. The time complexity for matrix multiplication is O(d_h x d_{K+1} x d_2) and the time complexity for bias addition is O(d_2). Thus, the overall time

Table 17: PIE

Method	Train(Seconds)	Test(Seconds)
F-Avg	4.94±0.26	0.033±0.09
F-Mode	6.06±0.27	0.034±0.09
MGP	123.63±2.38	0.374±0.11
ECML	12.92±1.50	0.035±0.09
TMC	11.39±0.31	0.035±0.09
TF	41.63±0.68	0.037±0.09
ETMC	10.36±0.37	0.036±0.09
ETF	50.39±0.71	0.037±0.09

Table 18: Scene15

Method	Train(Seconds)	Test(Seconds)
F-Avg	27.33±0.37	0.039±0.09
F-Mode	33.77±0.65	0.040±0.09
MGP	576.76±1.27	0.420±0.15
ECML	63.24±0.72	0.040±0.09
TMC	73.26±0.53	0.042±0.10
TF	229.05±2.86	0.042±0.09
ETMC	86.81±3.11	0.042±0.09
ETF	271.99±2.26	0.043±0.09

Table 19: HMDB

Method	Train(Seconds)	Test(Seconds)
F-Avg	38.26±0.65	0.045±0.09
F-Mode	48.86±0.64	0.048±0.09
MGP	654.42±1.35	0.971±0.13
ECML	82.32±1.17	0.047±0.09
TMC	74.62±0.65	0.047±0.09
TF	278.99±3.47	0.047±0.09
ETMC	99.54±0.93	0.046±0.09
ETF	365.94±8.12	0.047±0.09

Table 20: CUB

Method	Train(Seconds)	Test(Seconds)
F-Avg	3.57±0.29	0.033±0.09
F-Mode	4.48±0.29	0.033±0.09
MGP	136.74±0.76	0.239±0.10
ECML	8.17±0.28	0.036±0.09
TMC	7.66±0.30	0.034±0.09
TF	29.21±0.41	0.035±0.09
ETMC	13.98±0.38	0.035±0.09
ETF	37.57±0.56	0.036±0.09

complexity is O(d_h x d_{K+1} x d_2). Given the dataset for a classification task, the additional layer exhibits linear time complexity with respect to only the hidden size. Since this hidden size is relatively small and compact to the classification dimension, we argue that the increase in time complexity is not substantial as shown in following tables. We report the training and testing time by averaging 10 times running as shown in Tables 15 - 20.

E.2 Framework and Reproducibility

For experimental results to be reproducible, we will release our official implementation upon the paper's acceptance. Specifically, we used PyTorch (Paszke et al., 2019) version 1.13.0, built with CUDA 11.7, to implement our codes. The Python environment version is 3.8, and the operating system is Ubuntu 22.04.4. All Experiments are conducted on a single Nvidia RTX 3090 GPU with 24GB of memory.