CONAN: Contrastive Fusion Networks for Multi-view Clustering

Guanzhou Ke
Faculty of Intelligent Manufacturing
Wuyi University
Jiangmen, China
guanzhouk@gmail.com

Zhiyong Hong*
Faculty of Intelligent Manufacturing
Wuyi University
Jiangmen, China
hongmr@163.com

Zhiqiang Zeng
Faculty of Intelligent Manufacturing
Wuyi University
Jiangmen, China
zhiqiang.zeng@outlook.com

Zeyi Liu
Faculty of Intelligent Manufacturing
Wuyi University
Jiangmen, China
326341809@qq.com

Yangjie Sun
Faculty of Intelligent Manufacturing
Wuyi University
Jiangmen, China
1301437846@qq.com

Yannan Xie
Faculty of Electrical Engineering
Polytechnic of EPE[1]
Foshan, China
286628842@qq.com

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671851

***Abstract*—With the development of big data, deep learning has made remarkable progress on multi-view clustering. Multi-view fusion is a crucial technique for the model obtaining a common representation. However, existing literature adopts shallow fusion strategies, such as weighted-sum fusion and concatenating fusion, which fail to capture complex information from multiple views. In this paper, we propose a novel fusion technique, entitled contrastive fusion, which can extract consistent representations from multiple views and maintain the characteristic of view-specific representations. Specifically, we study multi-view alignment from an information bottleneck perspective and introduce an intermediate variable to align each view-specific representation. Furthermore, we leverage a single-view clustering method as a predictive task to ensure the contrastive fusion is working. We integrate all components into an unified framework called CONtrAstive fusion Network (CONAN). Experiment results on five multi-view datasets demonstrate that CONAN outperforms state-of-the-art methods. Our source code will be available soon at https://github.com/guanzhou-ke/conan.**

***Index Terms*—Clustering, multi-view fusion, contrastive learning**

I. Introduction

In real-world scenarios, data are frequently represented as multiple views or modalities. For example, various descriptors may characterize an image, such as SIFT [1] or histograms of oriented gradients (HoG) [2]. Multi-view learning is a promising way to obtain good representations from multiples views. Due to a large amount of unlabeled data in the real world, unsupervised multi-view learning, e.g., multi-view clustering, has attracted the attention of the machine learning community.

The significant difference between multi-view and single-view clustering is that the former needs to extract the valuable representation shared from multiple views. It means that fusion is a crucial technique for multi-view clustering. Previous works focus on some simple fusion strategies [3], such as weighted-sum fusion [4], [5] and concatenating fusion [6]–[8]. These methods are difficult to capture high-level semantics from multiple views because they ignore the relationship of common representations and view-specific representations during the fusion process. Another fusion technique, neural network fusion, focuses on capturing the latent structure shared from multiple views [9], [10]. It can capture a more complex structure than the former but could harm the view-specific representations during the fusion process, called over-fusion. To against over-fusion, they employ Generative Adversarial Nets (GAN) to the constraint condition [11] to maintain view-specific representations [4], [6], [12]. However, the trick brings an underlying risk: *inconsistent optimization objective*. We observe that clustering focuses on extracting class-level information, but GAN aims to reconstruct original space from pixel-level information. It leads to the network being forced to learn some useless information for predicting labels during the optimization process.

In this paper, we identify the major challenge in multi-view clustering fusion: *can we extract harmonious common representations from each view under maintaining the view-specific representations?* The simple idea is that we could introduce a new representations space to align each view-specific representation. For maintaining the independent representation space of each view, we block the interaction of view-specific representations. We study this idea from an information bottleneck perspective [13]. Inspired by previous works [14], [15], we can introduce an additional intermediate variable (representation space) to align each view-specific representation. We found that the contrastive learning method can be a suitable alignment method. Hence, we leverage the contrastive learning method [16] to improve multi-view fusion called contrastive fusion, discussed in section III.

However, we found that the contrastive fusion component cannot work well because it lacks the objective of extracting task-relevant information. In other words, it cannot identify

[*] Corresponding author.

[1] Guangdong Polytechnic of Environmental Protection Engineering, Foshan, China

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

what a good direction for alignment view-specific representations is. According to the suggestion of [15], we observe that a suitable predictive task, e.g., the single-view clustering method, can help the contrastive fusion component to extract harmonious common representations, discussed in section III-C. To this end, we integrate these components into a unified framework, namely CONtrAstive fusion Network (CONAN). The main contributions of this paper are summarized as follows:

1) We propose a novel fusion technique, which extracts common representation from view-specific representations and preserves the independence of each view space.
2) We integrate all components into an unified framework, as shown in Fig.1. The framework we proposed is more light-weighted than previous works [4], [6].

II. Related Work

In this section, we give a brief summary of the existing work on contrastive learning and multi-view clustering.

A. Contrastive Learning

Contrastive learning [16] is one of the unsupervised methods that have developed rapidly in recent years, especially in visual representations [17]–[19]. Contrastive learning aims to group the positive sample pairs and repulse the negative sample pairs, where the positive pair denotes the different augmentation views of an image. The negative pair is the augmentation view of other images. There is some literature to study the relation of between contrastive learning and multi-view learning from information bottleneck theory [14], [15], [20]. In [15], the authors proposed contrastive multi-view coding (CMC), which obtains robust representations by the multiple views augmentation of an image (more than two views). They found that the robustness of representations can be improved by increasing the number of views. On the other hand, the literature [14], [20] found that it can help extract consistent information by applying contrastive methods on multiple views. Based on the previous works, we introduce an intermediate variable and apply a contrastive method between this variable and each view-specific representations, called contrastive fusion. Intuitionally, the view-specific representations get aligned in the new representation space. The advantage of this is that it does not damage the view-specific representations and can achieve multi-view fusion.

B. Multi-view Clustering

The multi-view clustering taxonomy involves two categories: traditional methods and deep learning methods. Traditional methods rely on statistical theory, e.g., multi-kernel learning methods [21], [22], where the characteristics of different views can be non-linearly mapped into a common representation. Nevertheless, these methods require selecting the appropriate kernel for each view carefully. Subspace-based methods [23]–[25] project the multi-view data into a low-dimensional shared subspace to obtain the standard representation. There are also methods exploiting non-negative matrix factorization [26] and graph-based models [27]. These statistical models have a common drawback, as they cannot extract complex structures within the data. Therefore, there have been many deep methods combined with traditional methods, such as [9], [28], [29]. These extract higher semantic features through neural networks and then apply on these features statistical models to extract a common representation.

In recent years, the complete deep learning methods have been proposed, such as [4], [6]–[8], [12]. These methods have a more vital ability to express a data structure, obtaining higher performance on many benchmark datasets. The literature [4], [6], [12] adopts GAN as multi-view fusion constraint to prevent over-fusion. It may be beneficial in generative tasks, but degenerate solutions will appear in multi-view clustering tasks. Using GAN constraints seems to learn discriminative features, but it still requires the network to learn pixel-level features. This phenomenon we call inconsistent optimization objective, we mentioned before. Compared with previous methods, our proposed contrastive fusion does not require the network to learn pixel-level features, and it belongs to the same kind as clustering, i.e., discriminative model.

III. Method

Our goal is to group a set of n data points consisting of V views $\mathcal{D} = \{\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^V\}$ into c clusters, where $\mathbf{X}^v \in \mathbb{R}^{n \times d_v}$ denotes the samples of the dimension d_v from the v-th views.

A. Networks Architecture

The proposed approach, CONAN, consist of V view-specific encoder networks $e_v(\cdot)$, a fusion network $f(\cdot)$, a small neural network projection head $p(\cdot)$, and a clustering head $g(\cdot)$, as illustrated in Fig.1.

- View-specific encoder networks $e_v(\cdot)$ that extracts view-specific representations(vectors) h from each view data point, i.e., $\mathbf{h^v} = e_v(\mathbf{X_v})$. CONAN allows various forms of the network architecture as encoder, e.g., Fully-Connected Networks (FCNs) or Convolution Neural Networks (CNNs), according to the data type. In practice, we adopt CNNs to process image data and FCNs for vector-liked data.
- A fusion network $f(\cdot)$: it fusions concatenating vectors $\bar{\mathbf{h}} = cat(\mathbf{h^1}, ..., \mathbf{h^v})$ to obtain the common representation $\mathbf{z} = f(\bar{\mathbf{h}})$. We adopt two fully-connected layers with ReLU as fusion network. The major advantage of our fusion network is that can fit better any complex function [30] than the weighted-sum method.
- A projection head $p(\cdot)$: it maps representations to space where contrastive loss is applied. We adopt the same setting in previous work [17] to build our projection head. We discuss its details in the next section.
- A clustering head $g(\cdot)$: it assigns each data point to a compact space according to $\mathbf{z}$, in other words, we can obtain cluster labels from it, i.e., $Y = g(\mathbf{z})$. This component could adopt any single-view clustering methods,

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

such as Deep Divergence Clustering (DDC) [31] or Deep Embedding Clustering (DEC) [32].

In summary, our proposed framework, CONAN, is more light and flexible than the previous AE-Based [9], [29], [33], or GAN-Based [4], [6], [12] framework. Furthermore, CONAN adapts any online clustering methods thanks to the common representation $\mathbf{z}$ we obtained by contrastive fusion network.

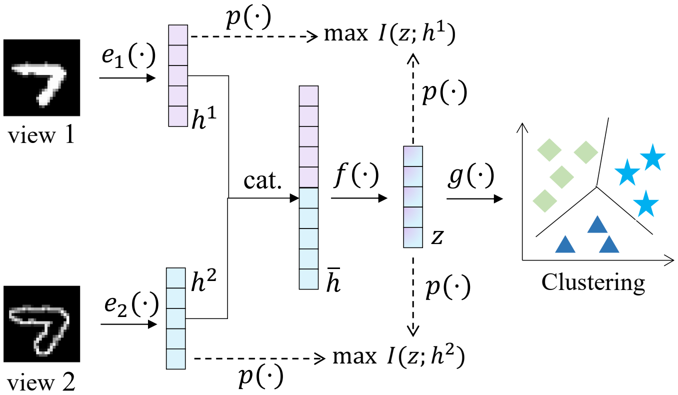

Fig. 1. Illustration of CONAN. It consist of V view-specific encoder networks $e_v(\cdot)$ to obtain view-specific representations $\mathbf{h^v}$, a fusion network $f(\cdot)$ to obtain the common representation $\mathbf{z}$ via fusion the concatenating vectors $\bar{\mathbf{h}}$, a small neural network projection head $p(\cdot)$ to map representations to space where contrastive loss is applied, and a clustering head $g(\cdot)$ to obtain soft assignment $Y = g(\mathbf{z})$. $I(\cdot;\cdot)$ represent mutual information.

B. Contrastive Fusion

Our goal is to extract the task-relevant information (consistency) shared between view-specific representations h and discard the task-irrelevant information. For convenience, we discuss two views situation, but it can be easily extended to multiple view cases. According to the conclusion of previous works [14], [15], [20], we can formula our goal from the information bottleneck perspective as the following definition:

$$I(\mathbf{X}^1;\mathbf{h}^2) = I(\mathbf{X}^2;\mathbf{h}^2) - I(\mathbf{X}^2;\mathbf{h}^2|\mathbf{X}^1) \tag{1}$$

$$I(\mathbf{X}^2;\mathbf{h}^1) = \underbrace{I(\mathbf{X}^1;\mathbf{h}^1)}_{\text{task-relevant}} - \underbrace{I(\mathbf{X}^1;\mathbf{h}^1|\mathbf{X}^2)}_{\text{task-irrelevant}} \tag{2}$$

It is benefit for extracting the task-relevant information when both maximizes (1) and (2), but we cannot maximize them directly in unsupervised setting. According to the conclusion of the literature [14], however, we can minimize the second term in (1) and (2) to maximize the lower bound on the mutual information $I(\mathbf{X}^1;\mathbf{h}^2)$ and $I(\mathbf{X}^2;\mathbf{h}^1)$. We discuss our strategy to minimize the conditional mutual information in next section.

According to above theory, we introduce an intermediate variable called $\mathbf{z}$ which can both maximize $I(\mathbf{z};\mathbf{h}^1)$ and $I(\mathbf{z};\mathbf{h}^2)$. Hence, we can rewrite the original definition as following:

$$I(\mathbf{X}^1;\mathbf{z}) = I(\mathbf{X}^2;\mathbf{z}) - I(\mathbf{X}^2;\mathbf{z}|\mathbf{X}^1) \tag{3}$$

where $\mathbf{z}$ is the common representation what we need. We call the process of maximization $I(\mathbf{z};\mathbf{h}^1)$ and $I(\mathbf{z};\mathbf{h}^2)$ as contrastive fusion, and then we focus on how to measure $\mathbf{z}$ and each $\mathbf{h}$, i.e., $\max I(\mathbf{z};\mathbf{h}^v)$. Inspired by contrastive learning methods [17], [18], we introduce cosine similarity to compute the score between $\mathbf{z}$ and $\mathbf{h}^v$, i.e., $sim(\mathbf{z},\mathbf{h}^v) = \mathbf{z}^T\mathbf{h}^v/\|\mathbf{z}\|\|\mathbf{h}^v\|$. Then, we extend the loss function in [17] to multiple views setting:

$$\mathcal{L}^{simclr}_{contrast} = -\sum_{v=1}^{V} log \frac{\exp(sim(p(z_i), p(h_j^v))/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(sim(p(z_i), p(h_k^v))/\tau)} \tag{4}$$

where $p(\cdot)$ is a projection head, (i, j) is a pair positive examples, $\mathbf{1}_{[k\neq i]} \in \{0,1\}$ denotes an indicator function evaluating to 1 if $k \neq i$, and τ represents a temperature parameter which we adopt optimal setting as $\tau = 0.1$ suggested by [17] in our experiments. According to the conclusion of [15], we can maximize $I(\mathbf{z};\mathbf{h}^v)$ by minimization the contrastive loss.

On the other hand, there is another version contrastive objective which does not require negative samples, called SimSiam [18], as following:

$$\mathcal{L}^{simsiam}_{contrast} = \sum_{v=1}^{V} -\frac{1}{2}(sim(p(\mathbf{z}), q(\mathbf{h}^v)) + sim(p(\mathbf{h}^v), q(\mathbf{z})) \tag{5}$$

where $q(\cdot)$ is a prediction head that consists of two fully-connected layers, note that the output vectors of $q(\cdot)$ can be treated as constant, which means that the output vectors of $q(\cdot)$ cannot receive any gradient during back-propagation. It seems that is the key trick of [18] to achieve state-of-the-art performance, called stop gradient. We study the impact of fusion via using two different objectives in section IV. According to the suggestion of ablation study in section IV, we adopt the default contrastive objective as (4) in CONAN.

C. Loss Function

In this section, we discuss our solution to minimize $I(\mathbf{X}^2;\mathbf{z}|\mathbf{X}^1)$ in (3). In an unsupervised setting, a common solution is to reconstruct the original data space what auto-encoder does. The reconstruction try to minimize $I(\mathbf{X};\mathbf{z}|\bar{\mathbf{X}})$, where $\bar{\mathbf{X}}$ is required to approximate the original data space $\mathbf{X}$. From a high level, the reconstruction or other downstream tasks, e.g., clustering or classification, can be roughly seem as a predictive learning task [15], [20]. According to these assumptions [15], [20], we can formula a general definition to instead $I(\mathbf{X}^2;\mathbf{z}|\mathbf{X}^1)$, as following:

$$I(\mathbf{X}^2;\mathbf{z}|T) = I(\mathbf{X}^2;\mathbf{z}|\mathbf{X}^1) + \epsilon_{info} \tag{6}$$

where T denotes an alternative predictive task, and ϵ_{info} is the gap between T and the original predictive task. However, the choice of a suitable predictive task is still an open question in unsupervised learning. We empirically chose the online clustering task as the alternative predictive task to maximize (3). There are two common clustering tasks we used in

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

this work, called DDC [31] and DEC [32]. The DDC is a divergence-based clustering method that uses Cauchy-Schwarz divergence (CS-divergence) [34] to enforce cluster separability and compactness. Another one, DEC, is a KL divergence-based method that use the Student's t-distribution as a kernel to measure the similarity between hidden representation z_i and centroid μ_j. Next, we describe the loss function of DDC and DEC, respectively.

The loss function of DDC consists of three terms as shown in the following:

$$
\begin{aligned}
\mathcal{L}_{ddc} = & \quad \textstyle\sum_{i=1}^{k-1}\sum_{j>i}^{k} \frac{\alpha_i^T \mathbf{K} \alpha_j}{\sqrt{\alpha_i^T \mathbf{K} \alpha_i \alpha_j^T \mathbf{K} \alpha_j}} \\
+ & \quad triu(AA^T) \\
+ & \quad \textstyle\sum_{i=1}^{k-1}\sum_{j>i}^{k} \frac{m_i^T \mathbf{K} m_j}{\sqrt{mi^T \mathbf{K} m_i m_j^T \mathbf{K} m_j}}
\end{aligned} \tag{7}
$$

where A denotes the cluster assignment matrix, α_i is the i-th column of matrix A; $\mathbf{K}$ denotes kernel similarity matrix computed by $\mathbf{k}_{ij} = \exp(-\|z_i - z_j\|^2/(2\sigma)^2)$, and σ is Gaussian kernel bandwidth, default as 0.15; $triu(AA^T)$ denotes the strictly upper triangular elements of AA^T; $m_ij = \exp(\|a_i - e_j\|^2)$ where e_j is corner j of the standard simplex in $\mathbb{R}^k$.

The loss function of DEC as shown in the following:

$$
\mathcal{L}_{dec} = \sum_{i=1}^{N}\sum_{j=1}^{k} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{8}
$$

where q_{ij} denotes the probability of assignment data point i to cluster j, and it compute by:

$$
q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2/\beta)^{-\frac{\beta+1}{2}}}{\sum_{j'}^{K}(1 + \|z_i - \mu_j'\|^2/\beta)^{-\frac{\beta+1}{2}}}
$$

and

$$
p_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{j'} q_{ij'}^2/\sum_i q_{ij'}}
$$

According to the suggestion of DEC, we pass the whole dataset through the initialized view-specific encoders to get common representation $\mathbf{z}$ and then perform standard k-means algorithm in $\mathbf{z}$ to obtain k initial centroids $\{\mu_j\}_{j=1}^k$.

Finally, we integrate all components into an unified framework, namely CONAN. As default, we adopt DDC as the predictive task and (4) as the contrastive measurement during fusion view-specific representations, as shown in the following:

$$
\mathcal{L}_{conan} = \mathcal{L}_{ddc} + \lambda \mathcal{L}_{contrast}^{simclr} \tag{9}
$$

where λ is a trade-off parameter to limit the contrastive loss term too large. As shown in section IV, we found that $\lambda = 0.01$ is the optimal value for all experiments. Concretely, Algorithm 1 summarizes the proposed method.

Algorithm 1: CONAN pseudocode, PyTorch-style

```
# $e_v$: $v$-th view-specific encoder.
# f: fusion module.
# g: clustering module.
# p: projection mlp.
# l1: clustering loss.
# l2: contrastive loss.

for x1,...,xv in dataloader:
   # Step 1: obtain view-specific
    representations
   h1,...,hv = e1(x1),...,ev(xv)
   hs = cat(h1,...,hv) # n-by-(v*d)
   # Step 2: fusion view-specific
    representations
   z = f(hs)
   # Step 3: soft assign z to k cluster
   logits = g(z)
   # Step 4: compute the joint loss.
   L = l1(z, logits) + $\lambda$ * C(z, hs)
   # Step 5: back-propagate loss and
   # update networks' parameters by Adam
   L.backward()
   update(e, f, g, p)

def C(z, hs) # compute contrastive loss
   hs = split(hs) # to list.
   zp = p(z) # projection.
   hps = [p(h) for h in hs]
   # compute zq = q(z) and hqs = [q(h)
   # for h in hs] if used simsiam loss.
   subloss = 0.
   # compute zp and each hp loss .
   for hp in hps:
      subloss += l2(zp, hp)
   return subloss
```

IV. Experiments

A. Dataset

We evaluate CONAN and other competitors using five well-known multi-view datasets containing raw image and vector data, and report the dataset description in Table I. There are:

- Edge-MNIST (E-MNIST) [35], which is a large-scale and widely-used benchmark dataset consisting of 70,000 handwritten digit images with 28 × 28 pixels. The views contain the original digits and the edge-detected version, respectively. In unsupervised setting, we only need the training set to evaluate models.
- Edge-FMNIST (E-FMNIST) [36], which is a more complex dataset than standard MNIST consisting of 28 × 28 grayscale images of clothing items. We synthesize the second view by running the same edge detector used to create Edge-MNIST.
- COIL-20 [37], and COIL-100 [38], which depicts from different angles containing grayscale images of 20 items and RGB images of 100 items, respectively. We create a three-view dataset by randomly grouping the images for an item into groups of three.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

- PASCAL VOC 2007 (VOC) [39]. We use the multi-view version provided by [40], which consists of GIST features and word frequency counts from manually tagged natural images.

TABLE I
DATASET DESCRIPTION

Dataset	#samples	#view	#class	#shape*
Edge-MNIST	60,000	2	10	1 × 28 × 28
Edge-FMNIST	60,000	2	10	1 × 28 × 28
COIL-20	480	3	20	1 × 128 × 128
COIL-100	2,400	3	100	3 × 128 × 128
VOC	5649	2	20	512, 399

*Single shape denotes that the input dimensionality is same for all views.

TABLE II
CLUSTERING RESULTS ON FIVE BENCHMARK DATASETS.

Dataset	E-MNIST		E-FMNIST		COIL-20		COIL-100		VOC	
Metrics	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
SC-Best$^{\alpha}$	-	-	-	-	-	-	-	-	.402	.411
DCCA$^{\beta}$	.476	.443	.426	.419	-	-	-	-	.397	.425
DCCAE$^{\beta}$	.508	.475	.458	.436	-	-	-	-	.416	.443
DMSC	.653	.614	.524	.518	.651	.720	.483	.697	.541	.538
DAMC	.646	.594	.495	.487	.672	.729	.526	.701	.560	.552
Ae^{2}-Nets	.537	.463	.410	.392	.717	**.812**	.541	.753	.571	.571
EAMC	.668	.628	.541	**.622**	.690	.753	.518	.680	.607	.615
CONAN	**.905**	**.861**	**.592**	.552	**.725**	.803	**.556**	**.774**	**.621**	**.621**

$^{\alpha}$We cannot apply spectral clustering on the dataset except VOC because the large-scale or high-dimensionality dataset occupies many computation resources. $^{\beta}$ DCCA and DCCAE can only apply on two views dataset.

B. Baseline Models

We evaluate our framework to an extensive set of baseline models, which represent state-of-the-art for multi-view clustering: 1) Deep Canonical Correlation Analysis (DCCA) [28]; 2) Deep Canonically Correlated Auto-encoders (DCCAE) [29]; 3) Deep Multi-modal Subspace Clustering (DMSC) [9]; 4) Deep Adversarial Multi-view Clustering (DAMC) [6]; 5) End-to-end Adversarial attention network for Multi–modal Clustering (EAMC) [4]; 6) Autoencoder in autoencoder networks (Ae^{2}-Nets) [33]. Furthermore, in order to demonstrate CONAN can improve the clustering performance compared with the single-view setting. We use the standard spectral clustering algorithm [41] to conduct on each view and the concatenated view, and then we report the best result in spectral clustering (SC-best).

C. Implementation Details

We implement CONAN and the competitor non-linear methods on the PyTorch [42] platform, running on CentOS Linux 7 utilizing two NVIDIA Quadro P5000 Graphics Processing Units (GPUs) with 16 GB memory size. We train our model for 100 epochs, using the Adam optimization technique [43] with default parameters, and using Cosine Annealing learning rate scheduler [44] with an initial learning rate of 0.001. We leverage the grid search to find an optimal trade-off parameter λ range from 10^{-3} to 1. We observe that sets too small value to λ would curb the fusion process, while too large is harmful to the predictive task. Hence, we set λ to 0.01 in all experiments. For Edge-MNIST and Edge-FMNIST, we adopt CNN, proposed in [9] as the view-specific encoder. We set the batch size of the training dataset to 128 and the dimensionality of hidden representation, i.e., $\mathbf{h^v}$ and $\mathbf{z}$, to 288. For COIL-20 and COIL-100, we adopt AlexNet [45] as the view-specific encoder. We set the batch size to 24 and the hidden representation to 1024. Note that the suggestion of input shape in [45] is $3 \times 144 \times 144$, but we observe that it also work well on the COIL-(20/100) dataset. For VOC, we adopt three fully-connected layers with ReLU as the view-specific encoder, i.e., *x-512-512-256*, where x denotes the input dimensionality. The batch size is set to 100. Due to the instability of unsupervised, we train CONAN 20 times and report the results from the run resulting in the lowest value of (9).

D. Evaluation Metrics

The clustering performance is measured using two standard evaluation matrices, i.e., clustering Accuracy (ACC) and Normalized Mutual Information (NMI), where a higher value indicates better performance. For further details on these evaluation metrics, the reader is referred to [46]. It should be noted that the validation process of the clustering methods involves only the cases where ground truth labels are available.

E. Compared with Baseline Models

We evaluate our model and baseline models on five well-known datasets and report the ACC and NMI results in Table II. The results illustrate that our model can have a significant improvement compared to baseline models. It is worth noting that the proposed method manages a performance improvement of $23.7\%(0.905 - 0.668)$ and $23.2\%(0.861 - 0.628)$ against the second-best method in terms of ACC and NMI metrics, respectively. We believe that the contrastive fusion we proposed outperforms the shallow fusion methods, such as concatenating (DAMC) and weight-sum (EAMC). On the other hand, our method is a friendly-resource algorithm compared with the self-expression layer in DMSC.

F. The Impact of Different Contrastive Loss

Contrastive measurement is a crucial component of our method. According to the conclusion of the state-of-the-art for contrastive learning, [18], there are two different type measurements, i.e., negative samples needed ($\mathcal{L}_{contrast}^{simclr}$) and positive samples only ($\mathcal{L}_{contrast}^{simsiam}$), that may lead to different

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

TABLE III
THE IMPACT OF DIFFERENT CONTRASTIVE LOSS ON E-FMNIST AND COIL-20/100 DATASET.

	$\mathcal{L}_{contrast}^{simclr}$	$\mathcal{L}_{contrast}^{simsiam}$	ACC	NMI
E-FMNIST	–	–	0.508	0.492
	✓	–	0.592	**0.552**
	–	✓	**0.603**	0.551
COIL-20	–	–	0.539	0.651
	✓	–	**0.725**	**0.803**
	–	✓	0.629	0.697
COIL-100	–	–	0.389	0.679
	✓	–	**0.556**	**0.774**
	–	✓	0.407	0.698

performance during the fusion process. Hence, we study the representative method of two kind measurements on E-FMNIST and COIL-20/100 dataset, as shown in Table III. The results illustrate that contrastive fusion can significantly improve the performance of clustering. Furthermore, we can see that there are no overwhelming results. Using $\mathcal{L}_{contrast}^{simclr}$ or $\mathcal{L}_{contrast}^{simsiam}$ in (9) achieve almost the same performance on the E-FMNIST dataset, but the former is almost entirely beyond the latter on the COIL-20/100 dataset.

Our conclusion is that because contrastive learning relies on the quality (data augmentation) and quantity of the dataset [47], both can achieve almost the same performance on the E-FMNIST dataset with sufficient quantity and quality. Conversely, if the data quality and quantity are insufficient, the performance depends on establishing negative samples, which is particularly obvious on COIL-100.

G. The Impact of Different Predictive Tasks

TABLE IV
THE IMPACT OF DIFFERENT PREDICTIVE TASKS ON E-FMNIST.

model	$\mathcal{L}_{ddc}$	$\mathcal{L}_{dec}$	ACC	NMI
Random	–	–	0.105	0.001
CONAN	–	–	0.117	0.002
	✓	–	**0.592**	**0.552**
	–	✓	0.494	0.486

We study the different predictive tasks on E-FMNIST in order to understand the importance of a good predictive task to our model, the results as reported in Table IV. We found that it would be close to a random model if we removed the clustering head. It is in line with our conclusion in section III-C. In addition, there is a significant gap between $\mathcal{L}_{ddc}$ and $\mathcal{L}_{dec}$. The results illustrate that a suitable predictive task is crucial to minimize (6). Nevertheless, the choice of predictive task is still an open question.

H. Visualization of The Contrastive Fusion

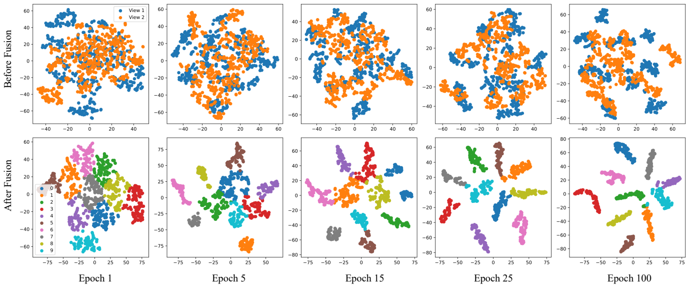

Fig. 2. Visualization of representations on the E-MNIST dataset before (first row) and after fusion (second row) using T-SNE at epoch 1st, 5th, 15th, 25th, and 100th.

We demonstrate the change of representations before and after contrastive fusion on the E-MNIST dataset, as shown in Fig.2. For convenience, we randomly choose 1K samples projected to 2-D using T-SNE [48]. We can see that the view-specific representations $\mathbf{h}^1$ and $\mathbf{h}^2$ always keep their unique characteristic. On the other hand, the common representation $\mathbf{z}$ becomes more separate and compact with increasing the training epoch. The results illustrate that contrastive fusion can improve the clustering performance and does not harm the view-specific representation.

V. Conclusion

In this work, we proposed a contrastive fusion strategy that aligns view-specific representation into a new representation space (common representations). We apply the existing single-view online clustering method to help the contrastive fusion network extracting valuable information from multiple views. Experimental results on five real-world datasets demonstrate the contrastive fusion improves the quality of unsupervised representations.

Acknowledgment

This work was supported by Natural Science Foundation of Guangdong Province, China (No. 2020A1515011468; 2016A0303010003); Guangdong University Scientific Research Project, China (No. 2019KTSCX189; 2018KTSCX235; 2018WZDXM014); Joint Research and Development Fund of Wuyi University and Hong Kong and Macau (2019WGALH21).

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[3] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, 2019. [Online]. Available: https://doi.org/10.1109/TKDE.2018.2872063

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

[4] R. Zhou and Y. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 14 607–14 616. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01463

[5] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Thirty-second AAAI conference on artificial intelligence*, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16212

[6] Z. Li, Q. Wang, Z. Tao, Q. Gao, and Z. Yang, "Deep adversarial multi-view clustering network," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 2019, pp. 2952–2958. [Online]. Available: https://doi.org/10.24963/ijcai.2019/409

[7] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2014, pp. 36–45. [Online]. Available: https://doi.org/10.3115/v1/d14-1005

[8] N. McLaughlin, J. M. del Rincón, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1325–1334. [Online]. Available: https://doi.org/10.1109/CVPR.2016.148

[9] M. Abavisani and V. M. Patel, "Deep multimodal subspace clustering networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1601–1614, 2018.

[10] X. Zhang, X. Tang, L. Zong, X. Liu, and J. Mu, "Deep multimodal clustering with cross reconstruction," in *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 12084. Springer, 2020, pp. 305–317. [Online]. Available: https://doi.org/10.1007/978-3-030-47426-3_24

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[12] H. Salman and J. Zhan, "Semi-supervised learning and feature fusion for multi-view data clustering," in *IEEE International Conference on Big Data, Big Data 2020, Atlanta, GA, USA, December 10-13, 2020*. IEEE, 2020, pp. 645–650. [Online]. Available: https://doi.org/10.1109/BigData50022.2020.9378412

[13] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[14] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=B1xwcyHFDr

[15] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, ser. Lecture Notes in Computer Science, vol. 12356. Springer, 2020, pp. 776–794. [Online]. Available: https://doi.org/10.1007/978-3-030-58621-8_45

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[18] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[20] Y. H. Tsai, Y. Wu, R. Salakhutdinov, and L. Morency, "Self-supervised learning from a multi-view perspective," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=-bdp_8Itjwp

[21] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," *Advances in neural information processing systems*, vol. 17, pp. 1537–1544, 2004.

[22] L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, W. Wang, and Y.-D. Shen, "Robust multiple kernel k-means using l21-norm," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[23] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 129–136.

[24] X. Zhao, N. Evans, and J.-L. Dugelay, "A subspace co-training framework for multi-view clustering," *Pattern Recognition Letters*, vol. 41, pp. 73–82, 2014.

[25] Y. Guo, "Convex subspace representation learning from multi-view data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013.

[26] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, 2013, pp. 252–260.

[27] Z. Xue, G. Li, S. Wang, C. Zhang, W. Zhang, and Q. Huang, "Gomes: A group-aware multi-view fusion approach towards real-world image clustering," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015, pp. 1–6.

[28] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[29] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[30] B. C. Csáji *et al.*, "Approximation with artificial neural networks," *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24, no. 48, p. 7, 2001.

[31] M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A. Salberg, and R. Jenssen, "Deep divergence-based approach to clustering," *Neural Networks*, vol. 113, pp. 91–101, 2019. [Online]. Available: https://doi.org/10.1016/j.neunet.2019.01.015

[32] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 478–487. [Online]. Available: http://proceedings.mlr.press/v48/xieb16.html

[33] C. Zhang, Y. Liu, and H. Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2577–2585.

[34] R. Jenssen, J. C. Príncipe, D. Erdogmus, and T. Eltoft, "The cauchy-schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels," *J. Frankl. Inst.*, vol. 343, no. 6, pp. 614–629, 2006. [Online]. Available: https://doi.org/10.1016/j.jfranklin.2006.03.018

[35] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 469–477.

[36] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017. [Online]. Available: http://arxiv.org/abs/1708.07747

[37] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Department of Computer Science, Columbia University, Tech. Rep. CUCS-005-96, February 1996.

[38] Nayar and H. Murase, "Columbia object image library: Coil-100," Department of Computer Science, Columbia University, Tech. Rep. CUCS-006-96, February 1996.

[39] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[40] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *British Machine Vision Conference,*

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.

BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings. British Machine Vision Association, 2010, pp. 1–12. [Online]. Available: https://doi.org/10.5244/C.24.58

[41] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[44] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2012, pp. 1106–1114.

[46] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," *Advances in neural information processing systems*, vol. 24, pp. 1413–1421, 2011.

[47] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 9929–9939. [Online]. Available: http://proceedings.mlr.press/v119/wang20k.html

[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on September 27,2022 at 02:27:33 UTC from IEEE Xplore. Restrictions apply.