Adversarially Robust Deep Multi-View Clustering: A Novel Attack and Defense Framework

Haonan Huang [1 2] **Guoxu Zhou** [1 3] **Yanghang Zheng** [1] **Yuning Qiu** [2] **Andong Wang** [2] **Qibin Zhao** [2 1]

Abstract

Deep Multi-view Clustering (DMVC) stands out as a widely adopted technique aiming at enhanced clustering performance by leveraging diverse data sources. However, the critical issue of vulnerability to adversarial attacks is unexplored due to the lack of well-defined attack objectives. To fill this crucial gap, this paper is the first work to investigate the possibility of adversarial attacks on DMVC models. Specifically, we introduce an adversarial attack with Generative Adversarial Networks (GANs) with the aim to maximally change the complementarity and consistency of multiple views, thus leading to wrong clustering. Building upon this adversarial context, in the realm of defense, we propose a novel Adversarially Robust Deep Multi-View Clustering by leveraging adversarial training. Based on the analysis from an information-theoretic perspective, we design an Attack Mitigator that provides a foundation to guarantee the adversarial robustness of our DMVC models. Experiments conducted on multi-view datasets confirmed that our attack framework effectively reduces the clustering performance of the target model. Furthermore, our proposed adversarially robust method is also demonstrated to be an effective defense against such attacks. This work is a pioneer in exploring adversarial threats and advancing both theoretical understanding and practical strategies for robust multi-view clustering. Code is available at https://github.com/libertyhhn/AR-DMVC.

[1]School of Automation, Guangdong University of Technology, Guangzhou, CHINA [2]RIKEN AIP, Tokyo, JAPAN [3]Key Laboratory of Intelligent Detection and the Internet of Things in Manufacturing, Ministry of Education, Guangzhou, CHINA. Correspondence to: Guoxu Zhou <gx.zhou@gdut.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

The increasing accumulation of real-world data from various sources, coupled with diverse feature extractors, highlights the critical role of multi-view learning (Fang et al., 2023). Integrating varied features, ranging from heterogeneous data to visual descriptors, for the same object has become pivotal. In unsupervised scenarios, multi-view clustering (MVC) has become an important tool aiming at the exploration of complementary and consistent information between different views to partition data samples (Huang et al., 2023c). Traditional MVC techniques, including matrix factorization (Huang et al., 2023a;b), spectral methods (Lu et al., 2022), and subspace clustering (Cao et al., 2015), focus on minimizing a predefined clustering objective function using specific distance metrics. However, these methods often underperform with high-dimensional data and demand extensive computational resources (Yan et al., 2021).

To overcome these limitations, early deep clustering models utilized deep neural networks for dimensionality reduction, facilitating more efficient clustering (Wang et al., 2015; Huang et al., 2019). This evolution has led to significant advancements in deep multi-view clustering (DMVC), where state-of-the-art models now consistently outperform traditional methods on various benchmarks (Hassani & Khasahmadi, 2020; Lin et al., 2023; Cui et al., 2023). Despite the demonstrated effectiveness of DMVC models across diverse domains, its vulnerability to adversarial attacks has not been well understood and explored yet. This issue is particularly prominent in safety-critical applications, where real-world data often faces threats from adversarial entities determined to deceive or disrupt machine learning models (Madry et al., 2018; Croce & Hein, 2020). Thus, our work is motivated by the first scientific question *Q1: How to effectively attack DMVC models?*

When facing attacks, it is equally important to study the adversarial robustness of the models. However, there is little research on the adversarial robustness of DMVC models. Although a series of adversarial DMVC algorithms (Li et al., 2019; Zhou & Shen, 2020; Wang et al., 2023) has been studied, they are not related to adversarial attack and defense of DMVC. Its primary focus has been on addressing clustering challenges with clean multi-view data and does

not thoroughly investigate the model's adversarial robustness. Therefore, our study is also motivated by the second scientific question: *Q2: How to develop a robust DMVC model to defend the attack?*

We re-emphasize that, despite extensive research exploring multi-view clustering methods from various perspectives, there has been a notable absence of prior investigations into the systematic handling of adversarial attacks and defense mechanisms designed specifically for DMVC models. This research gap has left existing DMVC models exposed to potential attacks, rendering them susceptible to clustering failures and undermining their overall reliability. Our objective is to address this gap by introducing adversarial attacks within the multi-view data space based on a well-defined adversarial threat strategy, while also developing a robust DMVC model under adversarial training paradigm for defensive purposes. In essence, our study aims to underscore the critical need for the development of adversarially robust DMVC models, emphasizing their practical utility across various applications. In summary, our contributions to addressing the above questions include:

1. We first formulate the adversary's goal for DMVC models and subsequently develop the adversarial attack framework based on GANs for DMVC models. Our method addresses the potential of multi-view model attacks for the first time, leveraging the distinct characteristics of the DMVC model to design targeted complementary and consistent attack strategies.
2. To defend against adversarial attack, we analyze the attack scenario and integrate DMVC with adversarial training to improve its robustness, which is a novel Adversarially Robust Deep Multi-View Clustering method (AR-DMVC).
3. Formalizing the problem of adversarially robust multi-view learning through an Information-Theoretic Perspective, we propose a new objective (AR-DMVC-AM) with an explicit regularizer aiming to mitigate attacks by minimizing the mutual information between the adversarial examples and clustering assignments.
4. We perform a thorough empirical evaluation and comparisons with the state-of-the-art deep multi-view clustering models on diverse benchmark datasets, including RegDB, NoisyFashion, NoisyMNIST, and PatchedMNIST. Our experimental analysis demonstrates the effectiveness of our adversarial attack and defense methods across various models and datasets.

Notations. $\{\mathbf{x}_1^v, \cdots, \mathbf{x}_n^v\}_{v=1}^V$ denotes multi-view data of V views sampled from the input data distribution $\mathcal{X}$. δ denotes the perturbations. $\mathbf{z}$ and $\mathbf{a}$ denote the learned representation and clustering assignment, respectively. $\tilde{\mathbf{x}}$, $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{a}}$ denote the adversarial examples, representation, and clustering assignment, respectively. $\|\cdot\|$ denotes the ℓ_2-norm of a vector.

2. Related Works

2.1. Deep Multi-view Clustering

Deep learning architectures have become widely used in the field of MVC, leading to the emergence of the Deep MVC subfield. The classic framework for DMVC is implemented using architectures based on adversarial networks (Xu et al., 2019; Li et al., 2019; Zhou & Shen, 2020), and autoencoders (Xu et al., 2021; Lin et al., 2023; Huang et al., 2023c). More recently, SOTA DMVC models have shifted from using prior approaches to employing self-supervised and contrastive learning methods for clustering (Trosten et al., 2021; Pan & Kang, 2021; Lin et al., 2022; Liu et al., 2023; Xu et al., 2023a). Techniques created in this specific area have demonstrated cutting-edge clustering accuracy, far surpassing conventional methods that do not utilize deep learning (Trosten et al., 2023). While there are existing methods focusing on robust multi-view clustering algorithms (Yang et al., 2020; Zeng et al., 2023), their primary emphasis is on robustness to incomplete multi-view data, and the model's inherent fragility to the well-designed perturbations is not considered. Meanwhile, although there have been related works dedicated to analyzing the trustworthiness of multi-view models (Han et al., 2021; Tang & Liu, 2022; Zhang et al., 2023), they only considered the internal missing or damaged data and did not take into account the model being attacked by adversarial samples.

2.2. Adversarial Clustering

The emergence of adversarial attacks on clustering algorithms has sparked significant interest, particularly considering their widespread utilization in computer security systems. The inaugural exploration of clustering algorithms in the context of adversarial attacks was undertaken in (Skillicorn, 2009; Dutrisac & Skillicorn, 2008; Crussell & Kegelmeyer, 2015). In these seminal works, the authors delved into the impact of adversarial samples on misclustering, highlighting the phenomenon of these samples gravitating towards the boundary of clustering centers, thereby generating new fringe clusters. Building on this foundation, (Biggio et al., 2013; 2014) approached the adversarial clustering problem from a theoretical standpoint, presenting two distinct attack strategies: poisoning and obfuscation. The former aims to compromise system availability, while the latter seeks to undermine system integrity. Additionally, (Chhabra et al., 2020) extended the understanding of adversarial samples to encompass metric-based clustering methods, such as K-Means clustering and Ward's clustering algorithms, demonstrating their existence and potential

impact.

Different from the existing clustering algorithm, deep clustering methodologies predominantly focus on minimizing reconstruction loss, aiming to enhance the discriminative nature of the target embedding space, which plays a pivotal role in determining clustering quality. Despite these efforts, embedded features are remarkably susceptible to small perturbations, leading to divergent clustering outcomes. (Yang et al., 2020) introduced an adversarial attack strategy for manipulating clustering results, complemented by an adversarial training algorithm. (Park et al., 2021) leveraged pseudo-labels generated by existing clustering algorithms to retrain deep clustering models, thus the robustness and clustering performance can be further enhanced. (Chhabra et al., 2022) devised a black-box attack utilizing GANs to generate adversarial samples based on the query output of deep clustering models. More recently, a robust fair clustering framework has been crafted to safeguard against poisoned samples that could skew the fairness of the clustering process (Chhabra et al., 2023).

However, DMVC methods have not yet been subjected to such adversarial attacks. We believe this is due to the increased difficulty of attacks for multiple views simultaneously compared to attacks for single-view data alone (e.g. (Chhabra et al., 2022)). Unlike (Chhabra et al., 2022)'s single-view focus, our work extends to multi-view clustering, not just by summation over views but by leveraging inter-view complementarity and consistency. Our tailored loss functions specifically address the multi-view challenge, enabling more effective adversarial perturbations. Meanwhile, when an attack occurs, how to develop a novel adversarial defense algorithm against malicious attacks is an important issue—a dimension entirely absent in (Chhabra et al., 2022). Due to the unsupervised nature of DMVC, directly applying adversarial training techniques is challenging (Dong et al., 2020). Our work operates within the same domain but distinguishes itself from existing works in several ways: (1) we tackle the formidable challenge of adversarial attacks in DMVC models. The intrinsic properties of multi-view data, such as complementarity and consistency across different views, usually confer heightened robustness to clustering outcomes; (2) we introduce the concept of DMVC methods with adversarial training, providing a novel approach to enhance the security and robustness of multi-view models.

3. Attack: Adversaries to DMVC Models

DMVC, a methodology that harnesses multiple data perspectives to enhance sample grouping, has garnered increasing attention in the contemporary data-driven landscape. In this context, ensuring the robustness of DMVC models against adversarial attacks is of paramount importance, as it is essential to preserve trust in clustering outcomes and amplify the practical utility of MVC models across a diverse range of real-world scenarios. To embark on this investigation, we first establish a clear understanding of the adversary's objectives within the multi-view setting:

Definition 3.1. (Adversary's Goal) The attack aims to introduce *minimal perturbations* to images used as input for the MVC model while staying within a defined noise threshold. This intentional perturbation is designed to cause *misclustering* of these samples by the model, leading to a notable decrease in performance, as quantified by various evaluation metrics.

In the realm of DMVC models, particularly the widely adopted contrastive-based ones (Xu et al., 2022; Trosten et al., 2023), their fundamental objective is to acquire concise multi-view representations that enhance clustering effectiveness. Consequently, our role as attackers involves undermining the model's clustering performance by strategically targeting the learned multi-view representations. Specifically, our aim is to ensure that the representation $\mathcal{C}\left(\mathbf{x}^v+\delta\right)$ obtained after the attack is maximally dissimilar to the representation $\mathbf{z}^v$ obtained before the attack. The optimization for the multi-view attack is thus outlined as follows:

$$\sum_{v=1}^{V}\sum_{i=1}^{N}\max_{\delta}\left\|\mathbf{z}_i^v-\mathcal{C}\left(\mathbf{x}_i^v+\delta\right)\right\|^2 \quad \text{s.t. } \|\delta\|^2\leq\epsilon, \tag{1}$$

where the $\mathcal{C}$ represents the deep multi-view clustering model, i denotes the index of individual data points within the dataset, and ϵ is imposed to prevent the adversarial sample from having excessive noise and to maintain its realism for human observers. Considering the characteristics of multi-view learning, attack models should aim to maximize disruption to multi-view learning models by targeting both complementarity and consistency. Based on this premise, we propose the following definition:

Definition 3.2. (Attacking Multi-view Complementarity and Consistency) The complementarity in multi-view learning refers to the uniqueness of view-specific representations $\mathbf{z}^v$ for each view $\mathbf{x}^v$, while consistency pertains to shared properties in the multi-view consensus representation. A successful attack method disrupts both *complementarity* and *consistency* when it induces noticeable differences between the pre-attack and post-attack states of learned view-specific representations and the consensus representation through the target model.

Based on Definition 3.2, our attack model incorporates two specific objective functions $\mathcal{L}_{\text{a-com}}$ and $\mathcal{L}_{\text{a-con}}$ to disrupt multi-view complementarity and consistency, respectively. These functions strategically introduce perturbations into the learned representations, compromising both aspects of

the multi-view clustering model. Specifically, we introduced an attack objective function $\mathcal{L}_{\text{a-com}}$ that targets the unique complementary information within each view. Motivated by GANs, this objective function is formulated as follows:

$$\mathcal{L}_{\text{a-com}} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v} \left\| \mathcal{C}^v\left(\mathbf{x}^v\right) - \mathcal{C}^v\left(\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\right)\right\|^2, \quad (2)$$

where $\mathcal{C}^v\left(\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\right)$ represents the learned view-specific representation post the attack and $\mathcal{G}$ denotes the generator model to generate the adversarial perturbation δ for a given input image $\mathbf{x}^v$. Secondly, considering the inherent consistency properties of multi-view data, we devised an additional attack objective function $\mathcal{L}_{\text{a-con}}$ aimed at undermining the consistency information across multiple views, as follows:

$$\mathcal{L}_{\text{a-con}} := \mathbb{E}_{\{\mathbf{x}^v\}_v} \|\mathcal{C}(\{\mathbf{x}^v\}_v) - \mathcal{C}\left(\{\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\}_v\right)\|^2. \quad (3)$$

In addition, the restriction on the adversarial noise norm can be succinctly reformulated as follows:

$$\mathcal{L}_{\text{constraint}} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v} \left[\min\left\{\epsilon - \left\|\mathcal{G}\left(\mathbf{x}^v\right)\right\|^2, 0\right\}\right]. \quad (4)$$

Moreover, we utilize the vanilla minimax GAN loss (Goodfellow et al., 2014) as follows:

$$\mathcal{L} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v}[\log(\mathcal{D}(\mathbf{x}^v)) + \log(1 - \mathcal{D}\left(\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\right))]. \quad (5)$$

Finally, combining Eq. 2 to 5, we formulate a saddle-point problem to train the Generator $\mathcal{G}$ and Discriminator $\mathcal{D}$:

$$\max_{\mathcal{D}} \min_{\mathcal{G}} \; \mathcal{L} - \mu_1 \mathcal{L}_{\text{a-com}} - \mu_2 \mathcal{L}_{\text{a-con}} - \mu_3 \mathcal{L}_{\text{constraint}}, \quad (6)$$

where μ_1, μ_2, μ_3 are hyper-parameters to control trade-off. In Eq. 6, the Generator $\mathcal{G}$ creates adversarial perturbations on the input examples to mimic real data and disrupt DMVC, while the Discriminator $\mathcal{D}$ attempts to distinguish real from generated adversarial examples. This leads to a competition where $\mathcal{G}$ aims to generate more realistic adversarial examples, and $\mathcal{D}$ tries to improve its ability to detect adversaries. This ongoing contest drives the evolution of both $\mathcal{G}$ and $\mathcal{D}$, with the ultimate goal of making $\mathcal{G}$'s adversarial examples indistinguishable from real data to $\mathcal{D}$. Subsequently, we use these adversarial examples as inputs for the pre-trained DMVC model to obtain representations post-attack. In summary, Algorithm 1 describes how to attack DMVC models.

Remark 3.3. If we solely attack the complementarity of multiple views (i.e., optimizing only Eq. 2), we may fail to disrupt the final learned consensus representation, potentially yielding identical results before and after the attack. Similarly, if we exclusively target the consistency of multiple views (i.e., optimizing only Eq. 3), we cannot ensure that each view has been adequately attacked, potentially affecting only a subset of views. Therefore, our model is rational, as it ensures that each view is attacked while preserving a consensus representation of changes.

Algorithm 1 Algorithm for Attacking DMVC

Input: The target model $\mathcal{C}$, unlabeled training set A, total training epochs E, batch size B, adversarial budget $\epsilon > 0$, hyperparameters μ_1, μ_2 and μ_3.
Output: The target model's clustering results, the trained GAN attack models.
Initialize parameters of GANs.
for $e = 0$ **to** $E - 1$ **do**
  for batch $m = 1, \ldots, [|U|/B]$ **do**
    Sample a minibatch B_m from U.
    Query the pre-attack and post-attack representations through $\mathcal{C}^v\left(\mathbf{x}^v\right)$, $\mathcal{C}\left(\{\mathbf{x}^v\}_v\right)$, $\mathcal{C}^v\left(\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\right)$ and $\mathcal{C}\left(\{\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\}_v\right)$.
    Update $\mathcal{G}$ and $\mathcal{D}$ by optimizing Eq. 6.
  end for
end for

4. Defense: Adversarially Robust DMVC

4.1. Basic DMVC Model

In recent years, there has been a burgeoning interest in multi-view learning employing contrastive learning (CL) (Trosten et al., 2023; Xu et al., 2023a). This approach is valuable as it can extract significant information from multi-view data and generate a succinct representation conducive to clustering. In this section, we present a straightforward DMVC model designed with CL. The conventional CL loss (Chen et al., 2020) formulation for a positive pair $(\mathbf{z}_i^u, \mathbf{z}_i^v)$ is expressed as follows:

$$\mathcal{L}_{\text{CL}} := \sum_{v=1}^{V} \sum_{u=1}^{V} -\log \frac{\exp\left(s_{ii}^{(uv)}\right)}{\sum_{s' \in \text{Neg}\left(\mathbf{z}_i^u, \mathbf{z}_i^v\right)} \exp\left(s'\right)}, \quad (7)$$

where $s_{ii}^{(uv)} = \frac{1}{\tau} \frac{(\mathbf{z}_i^u)^\top \mathbf{z}_i^v}{\|\mathbf{z}_i^u\| \|\mathbf{z}_i^v\|}$ represents the cosine similarity between the embeddings $\mathbf{z}_i^u$ and $\mathbf{z}_i^v$, and τ is a hyper-parameter set to 0.1 in all experiments. In addition, the set $\text{Neg}\left(\mathbf{z}_i^u, \mathbf{z}_i^v\right)$ is the set of similarities of negative pairs for the positive pair, which consists of $s_{ij}^{(uv)}, s_{ij}^{(uu)}$, and $s_{ij}^{(vv)}$, for all $j \neq i$. Then, the consensus representation of the multi-view embeddings $\{\mathbf{z}_i^v\}_{v=1}^V$ is modeled as their weighted sum,*i.e.*, $\mathbf{z}_i^* = \sum_{v=1}^V \omega^v \mathbf{z}_i^v$, where $\{\omega^v\}_{v=1}^V$ are trainable parameters.

In recent developments, the Deep Divergence-Based Clustering (DDC) module has been incorporated into various SOTA Deep Clustering models (Trosten et al., 2023). Thus, the fusion of views is achieved through a meticulously weighted

summation, followed by the application of the DDC clustering module to cluster the amalgamated representations. We provide a detailed composition of DDC in the Appendix A.1. Subsequently, we formulate a basic model termed Contrastive Learning-based Multi-view Clustering (CL-MVC), outlined as follows:

$$\mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u, \mathbf{x}_i^v; \theta\right) := \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{DDC}}, \tag{8}$$

where θ denotes the parameter of the network. Although many CL-based MVC models have been designed to improve clustering accuracy, as mentioned in our related works analysis, our focus in this article is to explore attacks and defenses against multi-view models. Hence, we opt not to incorporate intricate regularization model frameworks as the fundamental structure.

4.2. Adversarial Training in Multi-view Setting

To ensure the effectiveness of adversarial training, we took into account the consistency of attack structures across different views. Unlike traditional single-view or supervised adversarial training methods, attacks from distinct views may deceive the DMVC model into different targets. We note that attacks lacking alignment across views may weaken their effectiveness. For instance, attacks in the first view might try to mislead the model into categorizing them as the first category, while attacks in the second view may seek to have the model classify them as the second category. In our paper, we focus on the *worst-case* scenario, *i.e.*, disparate adversarial multi-view data may result in one consistent embedding. To achieve this goal, we introduce the contrastive loss for the adversarial multi-view data:

$$\begin{aligned}&\mathcal{L}_{\text{CL-MVC}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right)\\ &\text{where } \left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v\right) = \underset{\substack{\tilde{\mathbf{x}}_i^u \in \mathcal{B}_\epsilon[\mathbf{x}_i^u]\\ \tilde{\mathbf{x}}_i^v \in \mathcal{B}_\epsilon[\mathbf{x}_i^v]}}{\arg\max}\, \mathcal{L}_{\text{CL}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right),\end{aligned} \tag{9}$$

where $\tilde{\mathbf{x}}_i^u$ and $\tilde{\mathbf{x}}_i^v$ are adversarial data for the uth and vth views, respectively, $\mathcal{B}_\epsilon[\mathbf{x}] = \{\mathbf{x}' \in \mathcal{X} \mid d_\infty(\mathbf{x}, \mathbf{x}') \leq \epsilon\}$ are the closed ball of radius $\epsilon > 0$ centered at $\mathbf{x} \in \mathcal{X}$, and $(\mathcal{X}, d_\infty)$ denotes the input space $\mathcal{X}$ with the infinity distance metric $d_\infty(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_\infty$. In our approach, we adopt the widely used projected gradient descent (PGD) (Zhang et al., 2019) within the ϵ-balls centered at $\mathbf{x}_i^u$ and $\mathbf{x}_i^v$ to generate the adversarial data.

Remark 4.1. Note that maintaining a consistent attack embedding in adversarial training is crucial for effective defense against attacks. Our empirical findings highlight that a weaker consistency attacks regularization in adversarial training typically leads to a more vulnerable model. These results further indicate that the DMVC model is challenging to defend the adversarial perturbation.

Combing with Eq. 8, the objective function for adversarial training is given by

$$\begin{aligned}\mathcal{L}_{\text{AR-DMVC}} &= \mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u, \mathbf{x}_i^v; \theta\right)\\ &\quad + \lambda \mathcal{L}_{\text{CL-MVC}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right)\\ \text{where } \left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v\right) &= \underset{\substack{\tilde{\mathbf{x}}_i^u \in \mathcal{B}_\epsilon[\mathbf{x}_i^u]\\ \tilde{\mathbf{x}}_i^v \in \mathcal{B}_\epsilon[\mathbf{x}_i^v]}}{\arg\max}\, \mathcal{L}_{\text{CL}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right),\end{aligned} \tag{10}$$

where λ is a hyperparameter. By minimizing the above loss function, an adversarially robust DMVC model can be obtained since it simultaneously learns the succinct representation for clustering, as well as the consistent attack information from adversarial data.

The objective function $\mathcal{L}_{\text{AR-DMVC}}$ shares some similarities with CL techniques in adversarial pre-trained models (Jiang et al., 2020; Luo et al., 2023; Xu et al., 2023b). However, there are fundamental distinctions between our approach and theirs. Firstly, while adversarial contrastive learning focuses on pre-training models, our method is an end-to-end clustering framework, that introduces different learning objectives. Secondly, our method tackles the challenges of multi-view clustering problems where input data are entirely disparate, making it more complex compared to single-view input.

4.3. Attack Mitigator for Adversarial Training

Minimizing the objective function $\mathcal{L}_{\text{AR-DMVC}}$ yields the final clustering assignments predicted by the clean data and adversarial data as $\mathbf{a}$ and $\tilde{\mathbf{a}}$, respectively. However, a critical question persists: *does the clustering assignment generated by adversarial data retain the adversarial information in* $\tilde{\mathbf{x}}$*?* In this part, we answer this question from the information-theoretic perspective (Federici et al., 2020), and show that the adversarial attack can be eliminated through a simple and effective regularizer. We first introduce the conditional mutual information to measure the information between the adversarial input and the corresponding predictive clustering assignment, *i.e.*,

$$I\left(\tilde{\mathbf{x}}; \tilde{\mathbf{a}} \mid \mathbf{x}\right). \tag{11}$$

This conditional mutual information serves to quantify the information shared between adversarial data and clustering assignment when observing the clean data. Essentially, it measures the preservation of adversarial information in the clustering assignment. Therefore, reducing this conditional mutual information ensures the mitigation of the adversarial impact of perturbations in $\tilde{\mathbf{x}}$. However, direct minimization of the conditional mutual information is rather complicated. In the subsequent theorem, we illustrate that it can be upper-bounded in a more simplified formulation.

Theorem 4.2. *Given clean data* $\mathbf{x}$ *and adversarial data* $\tilde{\mathbf{x}}$ *and their corresponding cluster assignments* $\mathbf{a}$ *and* $\tilde{\mathbf{a}}$*, and let the KL divergence between* $p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right)$ *and* $p\left(\mathbf{a} \mid \mathbf{x}\right)$*:*

$$\mathcal{L}_{\text{AM}} := \mathcal{D}_{\text{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right). \tag{12}$$

Then, the conditional mutual information in Eq. 11 is upper-bounded:

$$I\left(\tilde{\mathbf{x}};\tilde{\mathbf{a}}\mid\mathbf{x}\right)\leq\mathcal{L}_{\mathrm{AM}}. \tag{13}$$

Therefore, as shown in Theorem 4.2, we can minimize the mutual information in Eq. 11 by minimizing its upper bound. To conserve space, please refer to the Appendix A.2 for detailed proofs.

4.4. Overall Adversarial Training Framework

By incorporating adversarial contrastive-based multi-view clustering loss (Eq. 10) with Attack Mitigator (Eq. 12), the objective function of the proposed method is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{AR-DMVC-AM}} &= \mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u,\mathbf{x}_i^v;\theta\right)\\ &\quad+\lambda\mathcal{L}_{\text{CL-MVC}}\left(\tilde{\mathbf{x}}_i^u,\tilde{\mathbf{x}}_i^v;\theta\right)\\ &\quad+\gamma\mathcal{D}_{\mathrm{KL}}\left(p\left(\tilde{\mathbf{a}}\mid\tilde{\mathbf{x}}\right)\|p\left(\mathbf{a}\mid\mathbf{x}\right)\right)\\ \text{where }\left(\tilde{\mathbf{x}}_i^u,\tilde{\mathbf{x}}_i^v\right) &= \underset{\substack{\tilde{\mathbf{x}}_i^u\in\mathcal{B}_\epsilon[\mathbf{x}_i^u]\\ \tilde{\mathbf{x}}_i^v\in\mathcal{B}_\epsilon[\mathbf{x}_i^v]}}{\arg\max}\,\mathcal{L}_{\mathrm{CL}}\left(\tilde{\mathbf{x}}_i^u,\tilde{\mathbf{x}}_i^v;\theta\right).\end{aligned} \tag{14}$$

Here γ parameterizes the AM regularization (see Figure 1 for framework illustration). In addition, the final cluster label $Y_i = \arg\max \mathbf{a}_i$. The learning algorithm of AR-DMVC-AM is presented in Algorithm 2.

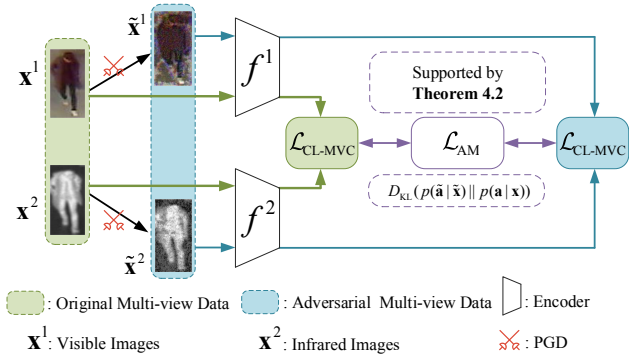


Figure 1. The framework diagram of AR-DMVC-AM utilizing a two-view dataset (RegDB) as an example.

5. Experimental Results

For evaluation, we utilize the following four benchmark multi-view datasets: RegDB (Nguyen et al., 2017), Noisy-Fashion, NoisyMNIST, and PatchedMNIST (Trosten et al., 2023). We provide detailed information about the dataset in the Appendix A.3. For all datasets, we randomly split 50% of the data for training and the remaining 50% for testing. We employ the training set to train our adversarial defense models and subsequently evaluate their performance on the testing set, comparing them with other open-source models. For the comparison methods, we take into consideration several SOTA open source deep multi-view clustering models, including EAMC (Zhou & Shen, 2020), SiMVC/CoMVC (Trosten et al., 2021), Multi-VAE (Xu et al., 2021), AECoDDC/InfoDDC (Trosten et al., 2023) and SEM (Xu et al., 2023a). We offer detailed network structures for CL-MVC and GAN in Appendix A.4.

Algorithm 2 Algorithm of Training AR-DMVC-AM

Input: Unlabeled training set U, total training epochs E, learning rate, batch size B, adversarial budget $\epsilon > 0$, hyperparameters λ and γ.
Output: The pre-trained model.
Initialize parameters of the model.
for $e = 0$ **to** $E-1$ **do**
  for batch $m = 1, \ldots, [|U|/B]$ **do**
    Sample a minibatch B_m from U.
    Compute $\tilde{\mathbf{x}}$ via Eq. 9.
    Sample $\tilde{\mathbf{a}} \sim p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right)$ and $\mathbf{a} \sim p\left(\mathbf{a} \mid \mathbf{x}\right)$.
    Update θ by minimizing the problem in Eq. 14.
  end for
end for

5.1. Attacking Results

Table 1 displays the pre-attack (original images) and post-attack (adversarial images) performance for each of the previously described models and datasets on three clustering metrics (ACC, NMI, ARI). In the table, we emphasized the results of each attacked method in *italics* and highlighted the best-performing method on the same dataset after the attack in ***bold***. For all the results, note that the GAN network produces consistent results without any variation since it creates the same constant noise for a given input. The table reveals that the model's results have experienced varying degrees of decrease after the attack, indicating that our suggested attack architecture has effectively targeted the deep multi-view clustering approaches. Notably, our method outperforms all other models in terms of post-attack data, underscoring the effectiveness of our robust model in mitigating attacks.

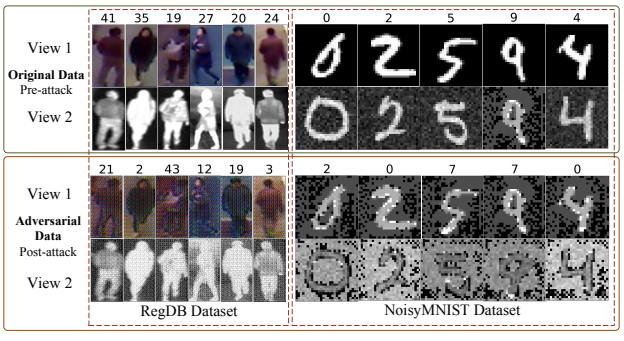


Figure 2. Adversarial samples generated by our attack on RegDB and NoisyMNIST dataset (correspond to EAMC).

Table 1. Pre-attack (PRE) and post-attack (POST) performance for deep multi-view clustering models on four datasets.

MODEL		REGDB			NOISYFASHION			NOISYMNIST			PATCHEDMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
EAMC (CVPR'20)	PRE	0.64	0.86	0.62	0.57	0.70	0.52	0.74	0.88	0.75	0.62	0.17	0.20
	POST	*0.33*	*0.57*	*0.23*	*0.30*	*0.20*	*0.11*	*0.25*	*0.11*	*0.06*	*0.53*	*0.13*	*0.15*
SIMVC (CVPR'21)	PRE	0.56	0.86	0.54	0.54	0.53	0.37	0.91	0.94	0.90	0.79	0.44	0.49
	POST	*0.30*	*0.61*	*0.24*	*0.30*	*0.25*	*0.13*	*0.29*	*0.20*	*0.12*	*0.49*	*0.13*	*0.12*
COMVC (CVPR'21)	PRE	0.45	0.73	0.38	0.69	0.71	0.59	0.99	0.99	0.99	0.81	0.48	0.52
	POST	*0.25*	*0.47*	*0.14*	*0.40*	*0.35*	*0.25*	*0.31*	*0.20*	*0.13*	*0.61*	*0.20*	*0.21*
MULTI-VAE (ICCV'21)	PRE	0.47	0.76	0.40	0.64	0.66	0.54	0.84	0.89	0.82	0.76	0.41	0.45
	POST	*0.43*	*0.71*	*0.33*	*0.47*	*0.43*	*0.30*	*0.46*	*0.39*	*0.28*	*0.51*	*0.19*	*0.16*
AECODDC (CVPR'23)	PRE	0.43	0.72	0.36	0.78	0.78	0.70	0.99	0.99	0.99	0.65	0.21	0.29
	POST	*0.23*	*0.46*	*0.13*	*0.39*	*0.39*	*0.23*	*0.24*	*0.11*	*0.06*	*0.46*	*0.11*	*0.10*
INFODDC (CVPR'23)	PRE	0.26	0.58	0.20	0.46	0.42	0.26	0.78	0.86	0.75	0.61	0.28	0.67
	POST	*0.22*	*0.50*	*0.11*	*0.25*	*0.19*	*0.10*	*0.33*	*0.22*	*0.14*	*0.53*	*0.13*	*0.15*
SEM (NEURIPS'23)	PRE	0.40	0.67	0.30	0.85	0.85	0.79	0.62	0.61	0.42	0.48	0.26	0.22
	POST	*0.33*	*0.63*	*0.21*	*0.31*	*0.30*	*0.14*	*0.21*	*0.11*	*0.07*	*0.45*	*0.16*	*0.14*
AR-DMVC (OURS)	PRE	0.55	0.84	0.48	0.68	0.69	0.56	0.99	0.99	0.99	0.83	0.52	0.58
	POST	*0.42*	*0.66*	*0.31*	*0.54*	*0.48*	*0.33*	*0.90*	*0.79*	*0.80*	*0.65*	*0.34*	*0.36*
AR-DMVC-AM (OURS)	PRE	0.54	0.85	0.50	0.69	0.73	0.59	0.99	0.99	0.99	0.81	0.46	0.52
	POST	***0.52***	***0.79***	***0.42***	***0.67***	***0.67***	***0.55***	***0.93***	***0.85***	***0.85***	***0.74***	***0.35***	***0.40***

In Figure 2, we visualize the original clean images and adversarial samples obtained by our attack methods on EAMC. For the RegDB dataset, we observe that the adversarial images retain similarity to the clean images while being significantly different from the targeted class. In the case of the NoisyMNIST dataset, the image from view 2, generated by adding Gaussian noise to the original image, appears slightly blurry after incorporating adversarial perturbations. However, we posit that the human eye can still discern the original category. In addition, we present a substantial quantity of adversarial images produced by our attack model in Appendix A.7.

We also demonstrate confusion matrices for the AECoDDC and our method (AR-DMVC-AM) on dataset NoisyMNIST in Figure 3. The horizontal axis denotes the category of clustering results, the vertical axis indicates the correct category, and the diagonal values represent the number of correctly clustered categories. By comparing Figures 3(a) and 3(b), it is evident that AECoDDC incorrectly clusters all numbers except for 1 after the attack. This discrepancy may be attributed to the simplicity of the number 1, making it less susceptible to attacks. From Figures 3(c) and 3(d), despite a slight decrease after the attack, our method maintained commendable performance, demonstrating its robustness to adversarial perturbations. We illustrate the confusion matrices for the other methods in the Appendix A.5 due to limited space.

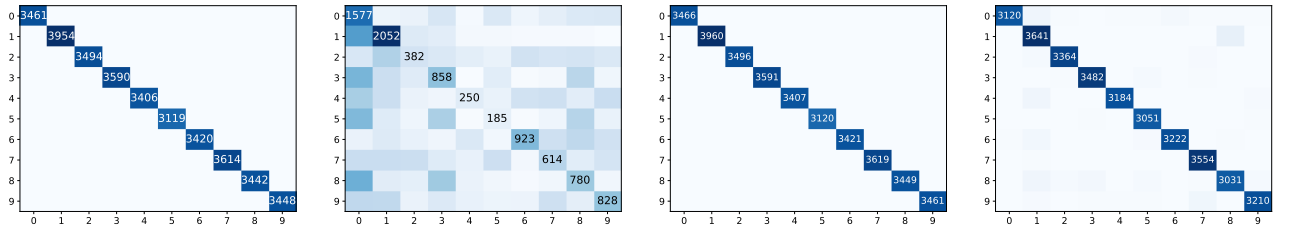
5.2. Evaluation of Robustness Transferability

As illustrated in Table 2, we can observe that AR-DMVC-AM significantly enhances AR-DMVC's robustness against adversarial attacks and demonstrates improved generalization to other datasets. This underscores the efficacy of AM regularization in augmenting robustness transferability against incremental data.

5.3. Hyperparameters Analysis

Concerning the adversarial attack hyperparameters in Eq. 6, we adhere to the configuration outlined in (Chhabra et al., 2022), and the values of μ_1, μ_2, and μ_3 are set to 5, 5, and 1, respectively. Regarding ϵ, the assigned values are 0.2 for RegDB, 0.15 for NoisyFashion, 0.3 for NoisyMNIST, and 0.3 for PatchedMNIST. In this subsection, we first explore the impact of varying the noise penalty parameter ϵ on the extent to which the attack degrades the performance of the DMVC models, as depicted in Table 3. It is evident that as the ϵ threshold increases, the efficacy of the attack escalates while the performance of the models diminishes. Meanwhile, our proposed adversarial defense method consistently preserves better clustering results, further substantiating its efficacy in ensuring adversarial robustness.

In Eq. 14, the trade-off coefficient λ is introduced to regulate varying levels of the strength of adversarial training, while γ is incorporated to govern the contribution of predictive consistency in our framework. As depicted in Figure 4(a), we vary both coefficients within the range from 0.001 to 1000



(a) Pre-attack (AECoDDC) (b) Post-attack (AECoDDC) (c) Pre-attack (AR-DMVC-AM) (d) Post-attack (AR-DMVC-AM)

Figure 3. Confusion matrices illustrating the effect of the attack for the AECoDDC and AR-DMVC-AM models on the NoisyMNIST.

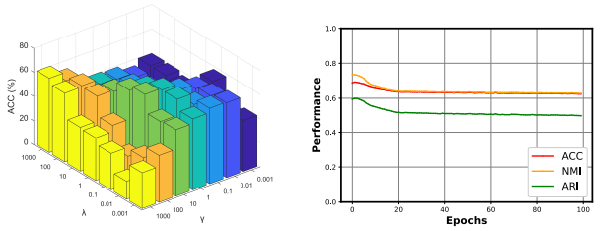
Table 2. Adversarial robustness transferability of AR-DMVC and AR-DMVC-AM on NoisyFashion and NoisyMNIST under four distinct conditions. $D1 \rightarrow D2$ denotes training and testing are executed on the different datasets, *i.e.*, $D1$ and $D2$. NoisyFashion(0-4) means that we use the 0-4 classes in NoisyFashion for training or testing, and so on.

$D1 \rightarrow D2$	MODEL	ACC	NMI
NOISYFASHION(0-4)	AR-DMVC	0.40	0.18
$\rightarrow$ NOISYFASHION(5-9)	AR-DMVC-AM	0.52	0.34
NOISYFASHION(5-9)	AR-DMVC	0.35	0.10
$\rightarrow$ NOISYFASHION(0-4)	AR-DMVC-AM	0.40	0.15
NOISYMNIST(0-4)	AR-DMVC	0.40	0.17
$\rightarrow$ NOISYMNIST(5-9)	AR-DMVC-AM	0.41	0.19
NOISYMNIST(5-9)	AR-DMVC	0.40	0.15
$\rightarrow$ NOISYMNIST(0-4)	AR-DMVC-AM	0.40	0.16
NOISYFASHION(0-4)	AR-DMVC	0.36	0.14
$\rightarrow$ NOISYMNIST(0-4)	AR-DMVC-AM	0.40	0.20
NOISYFASHION(5-9)	AR-DMVC	0.35	0.10
$\rightarrow$NOISYMNIST (5-9)	AR-DMVC-AM	0.37	0.15
NOISYMNIST(0-4)	AR-DMVC	0.42	0.24
$\rightarrow$ NOISYFASHION(5-9)	AR-DMVC-AM	0.44	0.29
NOISYMNIST(5-9)	AR-DMVC	0.32	0.10
$\rightarrow$ NOISYFASHION(0-4)	AR-DMVC-AM	0.36	0.10

and report the clustering results tested under post-attack conditions. We can conclude that a stronger intensity of adversarial training, indicated by relatively large parameters (optimal at $\lambda = 100$), results in better defense against attacks. The parameter γ exhibits relative stability, and thus, we consistently set it to 1 in all experiments. From Figure 4(b), it can be observed that AR-DMVC-AM reaches a stable state at epoch 30. Consequently, we set the epoch to 30 in our experiments under attack. Owing to space constraints, we provide the hyperparameter results for the remaining datasets in Appendix A.6.

Table 3. ACC versus perturbation parameters ϵ varies within the range of 0.1 to 0.3.

MODEL	NOISYMNIST			PATCHEDMNIST		
	0.1	0.2	0.3	0.1	0.2	0.3
EAMC	0.72	0.44	0.23	0.45	0.51	0.49
SIMVC	0.91	0.41	0.29	0.72	0.72	0.49
COMVC	0.94	0.43	0.27	0.79	0.77	0.61
MULTI-VAE	0.85	0.55	0.47	0.56	0.53	0.51
AECODDC	0.93	0.39	0.24	0.50	0.50	0.44
INFODDC	0.77	0.31	0.33	0.80	0.69	0.65
SEM	0.44	0.24	0.20	0.52	0.54	0.58
AR-DMVC	**0.99**	**0.99**	0.54	**0.82**	0.62	0.49
AR-DMVC-AM	**0.99**	**0.99**	**0.94**	0.79	**0.78**	**0.70**



(a) NoisyFashion (b) NoisyFashion

Figure 4. (a) ACC versus parameters λ and γ of AR-DMVC-AM. (b) The clustering results with respect to the epoch of AR-DMVC-AM under attack.

6. Conslusion

In this paper, we present the first adversarial attack against deep multi-view clustering models, which simultaneously targets the complementarity and consistency of multiple views through a GAN-based architecture. More importantly, driven by the concern for adversarial fragility, we employ adversarial training to bolster the adversarial robustness of the DMVC model and propose Attack Mitigator regularization to enhance the AR-DMVC. Empirically, comprehensive experiments show that previous DMVC approaches fail to detect and mitigate our adversarial attacks, whereas our proposed methods are more robust against adversarial attacks.

Limitations Our work also has a few limitations: 1) Our attack method requires indexing the representations learned

from each view. While it can be applied to most current DMVC models, it may not maximize its effectiveness for methods that only learn unified representations; 2) Our AR-DMVC-AM requires multiple backward propagations on all training data to generate adversarial variants, followed by training the model with these adversarial data. This process is computationally expensive, particularly when dealing with large-scale training sets. Developing a more efficient and effective defense method is identified as part of our future work.

Impact Statement

Our work addresses the critical need for robust defenses against adversarial attacks in Deep Multi-View Clustering (DMVC) models. By introducing an adversarially robust framework, we aim to enhance model security and reliability, providing a foundation for further research in this area.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62073087, 62071132, 62103110, and 62203124; and in part by the China Scholarship Council (CSC) under Grant 202208440315. We also thank Guang Lin and the anonymous reviewers for their helpful comments.

References

Biggio, B., Pillai, I., Rota Bulò, S., Ariu, D., Pelillo, M., and Roli, F. Is data clustering in adversarial settings secure? In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, pp. 87–98, 2013.

Biggio, B., Bulò, S. R., Pillai, I., Mura, M., Mequanint, E. Z., Pelillo, M., and Roli, F. Poisoning complete-linkage hierarchical clustering. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pp. 42–52. Springer, 2014.

Cao, X., Zhang, C., Fu, H., Liu, S., and Zhang, H. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–594, 2015.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Chhabra, A., Roy, A., and Mohapatra, P. Suspicion-free adversarial attacks on clustering algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3625–3632, 2020.

Chhabra, A., Sekhari, A., and Mohapatra, P. On the robustness of deep clustering models: Adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 35:20566–20579, 2022.

Chhabra, A., Li, P., Mohapatra, P., and Liu, H. Robust fair clustering: A novel fairness attack and defense framework. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Crussell, J. and Kegelmeyer, P. Attacking dbscan for fun and profit. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 235–243. SIAM, 2015.

Cui, C., Ren, Y., Pu, J., Li, J., Pu, X., Wu, T., Shi, Y., and He, L. A novel approach for effective multi-view clustering with information-theoretic perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331, 2020.

Dutrisac, J. and Skillicorn, D. B. Hiding clusters in adversarial settings. In *IEEE International Conference on Intelligence and Security Informatics*, pp. 185–187. IEEE, 2008.

Fang, U., Li, M., Li, J., Gao, L., Jia, T., and Zhang, Y. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Hassani, K. and Khasahmadi, A. H. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pp. 4116–4126. PMLR, 2020.

Huang, H., Zhou, G., Liang, N., Zhao, Q., and Xie, S. Diverse deep matrix factorization with hypergraph regularization for multiview data representation. *IEEE/CAA Journal of Automatica Sinica*, 10(11):2154–2167, 2023a.

Huang, H., Zhou, G., Zhao, Q., He, L., and Xie, S. Comprehensive multiview representation learning via deep autoencoder-like nonnegative matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Huang, W., Yang, S., and Cai, H. Generalized information-theoretic multi-view clustering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.

Huang, Z., Zhou, J. T., Peng, X., Zhang, C., Zhu, H., and Lv, J. Multi-view spectral clustering network. In *International Joint Conference on Artificial Intelligence*, volume 2, pp. 4, 2019.

Jiang, Z., Chen, T., Chen, T., and Wang, Z. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.

Li, Z., Wang, Q., Tao, Z., Gao, Q., Yang, Z., et al. Deep adversarial multi-view clustering network. In *International Joint Conference on Artificial Intelligence*, pp. 2952–2958, 2019.

Lin, F., Bai, B., Guo, Y., Chen, H., Ren, Y., and Xu, Z. Mhcn: A hyperbolic neural network model for multi-view hierarchical clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16525–16535, 2023.

Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., and Peng, X. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

Liu, C., Wen, J., Liu, Y., Huang, C., Wu, Z., Luo, X., and Xu, Y. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lu, Z., Nie, F., Wang, R., and Li, X. A differentiable perspective for multi-view spectral clustering with flexible extension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7087–7098, 2022.

Luo, R., Wang, Y., and Wang, Y. Rethinking the effect of data augmentation in adversarial contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nguyen, D. T., Hong, H. G., Kim, K. W., and Park, K. R. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

Pan, E. and Kang, Z. Multi-view contrastive graph clustering. *Advances in Neural Information Processing Systems*, 34:2148–2159, 2021.

Park, S., Han, S., Kim, S., Kim, D., Park, S., Hong, S., and Cha, M. Improving unsupervised image clustering with robust learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12278–12287, 2021.

Skillicorn, D. B. Adversarial knowledge discovery. *IEEE Intelligent Systems*, 24(6):54, 2009.

Tang, H. and Liu, Y. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pp. 21090–21110. PMLR, 2022.

Trosten, D. J., Lokse, S., Jenssen, R., and Kampffmeyer, M. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1255–1265, 2021.

Trosten, D. J., Løkse, S., Jenssen, R., and Kampffmeyer, M. C. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23976–23985, 2023.

Wang, Q., Tao, Z., Xia, W., Gao, Q., Cao, X., and Jiao, L. Adversarial multiview clustering networks with adaptive fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7635–7647, 2023.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. On deep multi-view representation learning. In *International Conference on Machine Learning*, pp. 1083–1092. PMLR, 2015.

Xu, C., Guan, Z., Zhao, W., Wu, H., Niu, Y., and Ling, B. Adversarial incomplete multi-view clustering. In *International Joint Conference on Artificial Intelligence*, volume 7, pp. 3933–3939, 2019.

Xu, J., Ren, Y., Tang, H., Pu, X., Zhu, X., Zeng, M., and He, L. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9234–9243, 2021.

Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., and He, L. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022.

Xu, J., Chen, S., Ren, Y., Shi, X., Shen, H. T., Niu, G., and Zhu, X. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Xu, X., Zhang, J., Liu, F., Sugiyama, M., and Kankanhalli, M. Efficient adversarial contrastive learning via robustness-aware coreset selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=fpzA8uRA95.

Yan, X., Hu, S., Mao, Y., Ye, Y., and Yu, H. Deep multi-view learning methods: A review. *Neurocomputing*, 448: 106–129, 2021.

Yang, X., Deng, C., Wei, K., Yan, J., and Liu, W. Adversarial learning for robust deep clustering. *Advances in Neural Information Processing Systems*, 33:9098–9108, 2020.

Zeng, P., Yang, M., Lu, Y., Zhang, C., Hu, P., and Peng, X. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning*. PMLR, 2023.

Zhou, R. and Shen, Y.-D. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14619–14628, 2020.

A. Appendix for Adversarially Robust Deep Multi-View Clustering: A Novel Attack and Defense Framework

A.1. The detail of DDC

As introduced in (Trosten et al., 2021), DDC has three terms. The first term is used to ensure the clusters are separable, as follows:

$$\mathcal{L}_1 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\binom{k}{2}^{-1} \sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{a}_{ai} \kappa_{ab} \mathbf{a}_{bj}}{\sqrt{\sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{a}_{ai} \kappa_{ab} \mathbf{a}_{bi} \sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{a}_{aj} \kappa_{ab} \mathbf{a}_{bj}}} \tag{15}$$

where k denotes the number of clusters, $\kappa_{ij} = \exp\left(-\|\mathbf{z}_i^* - \mathbf{z}_j^*\|^2 / \left(2\sigma^2\right)\right)$, and σ is a hyperparameter. In our paper, σ is set to 15% of the median pairwise distance between hidden representations within a mini-batch, following the approach outlined in (Trosten et al., 2021).

The second term promotes orthogonality among the cluster assignment vectors for different objects:

$$\mathcal{L}_2 = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{a}_i{}^{\top} \mathbf{a}_j. \tag{16}$$

Finally, the third term constrains the cluster assignment vectors to be close to the standard simplex in $\mathbb{R}^k$:

$$\mathcal{L}_3 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\binom{k}{2}^{-1} \sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{m}_{ai} \kappa_{ab} \mathbf{m}_{bj}}{\sqrt{\sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{m}_{ai} \kappa_{ab} \mathbf{m}_{bi} \sum_{a=1}^{n} \sum_{b=1}^{n} \mathbf{m}_{aj} \kappa_{ab} \mathbf{m}_{bj}}} \tag{17}$$

where $\mathbf{m}_{ij} = \exp\left(-\left\|\mathbf{a}_i - \mathbf{e}_j\right\|^2\right)$, and $\mathbf{e}_j$ is corner j of the standard simplex in $\mathbb{R}^k$.

The final DDC clustering loss that we minimize is the sum of the above three terms:

$$\mathcal{L}_{\text{DDC}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{18}$$

A.2. Proof to Theorem 4.2

Proof. Given the clean data $\mathbf{x}$ and adversarial data $\tilde{\mathbf{x}}$, consider their corresponding cluster assignments $\mathbf{a}$ and $\tilde{\mathbf{a}}$. The definitions of mutual information and KL divergence yield

$$\begin{aligned} I\left(\tilde{\mathbf{x}}; \tilde{\mathbf{a}} \mid \mathbf{x}\right) &= \mathbb{E}_{\tilde{\boldsymbol{x}}, \boldsymbol{x} \sim p(\tilde{\mathbf{x}}, \mathbf{x})} \mathbb{E}_{\boldsymbol{a} \sim p(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}})} \left[\log \frac{p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right)}{p\left(\tilde{\boldsymbol{x}} \mid \mathbf{x}\right)}\right] \\ &= \mathbb{E}_{\tilde{\boldsymbol{x}}, \boldsymbol{x} \sim p(\tilde{\mathbf{x}}, \mathbf{x})} \mathbb{E}_{\boldsymbol{a} \sim p(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}})} \left[\log \frac{p\left(\tilde{\mathbf{a}} \mid \tilde{\boldsymbol{x}}\right) p\left(\mathbf{a} \mid \boldsymbol{x}\right)}{p\left(\mathbf{a} \mid \boldsymbol{x}\right) p\left(\tilde{\mathbf{a}} \mid \boldsymbol{x}\right)}\right] \\ &= \mathcal{D}_{\text{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right) - \mathcal{D}_{\text{KL}}\left(p\left(\mathbf{a} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right) \\ &\leq \mathcal{D}_{\text{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right). \end{aligned} \tag{19}$$

This completes the proof. □

A.3. The detail of dataset

Here, we provide details on the four datasets utilized in this paper:

1. RegDB (Nguyen et al., 2017) is collected from a pair of aligned cameras, comprising one in the visible spectrum and another in the infrared spectrum. We randomly selected data from 50 individuals, treating each individual as a category, and augmented each category to 100 samples using three different enhancement methods: grayscale, Gaussian noise addition, and inversion.

2. NoisyFashion/NoisyMNIST: a noisy variant of FashionMNIST/MNIST is employed, where the initial view encompasses the original image, and the second view incorporates an image sampled from the same class as the first image, with the Gaussian noise (Trosten et al., 2023).

3. PatchedMNIST is a subset of MNIST, encompassing the initial three digits, where three views are extracted as 7×7 non-overlapping patches from the original images (Trosten et al., 2023).

A.4. Network Architecture

We showcase the model framework for the proposed GAN attack model and CL-MVC as illustrated in Table 4 and Table 5. The GAN attack model includes an adversarial perturbation generator and a discriminator for each view. The CL-MVC comprises view-specific encoders and the DDC module.

Table 4. The network architecture of CL-MVC.

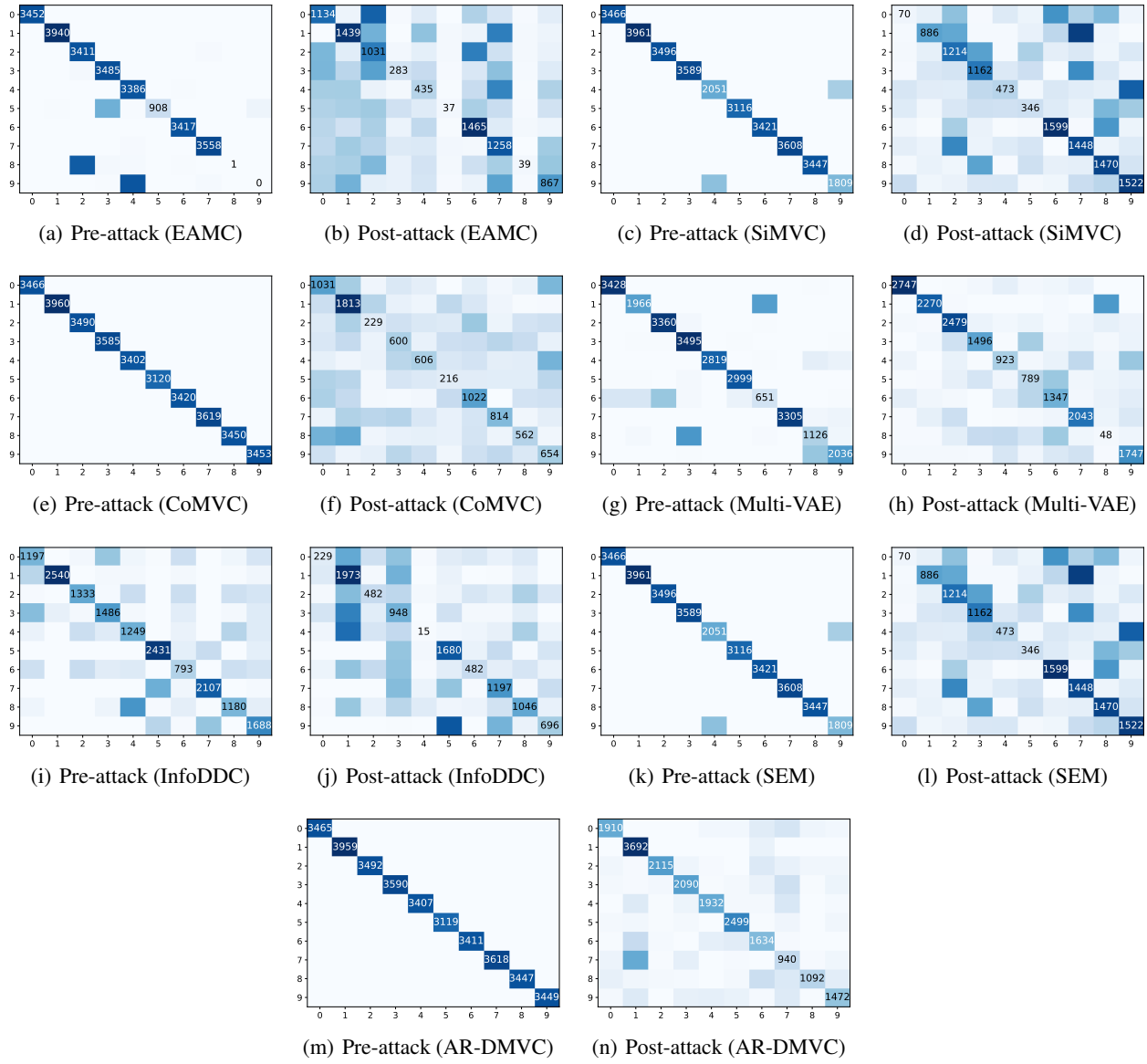
CNN Encoder	DCC Module
Conv($64 \times 3 \times 3$)	Dense(100)
ReLU	ReLU
Conv($64 \times 3 \times 3$)	BatchNorm
BatchNorm	Dense(k)
ReLU	Softmax
MaxPool(2×2)	
Conv($64 \times 3 \times 3$)	
ReLU	
Conv($64 \times 3 \times 3$)	
BatchNorm	
ReLU	
MaxPool(2×2)	

Table 5. The network architecture of the GAN attack model.

Generator	Discriminator
Conv($8 \times 3 \times 3$)	Conv($8 \times 4 \times 4$)
InstanceNorm	LeakyReLU(0.2)
ReLU	Conv($16 \times 4 \times 4$)
Conv($16 \times 3 \times 3$)	BatchNorm
InstanceNorm	LeakyReLU(0.2)
ReLU	Conv($32 \times 4 \times 4$)
Conv($32 \times 3 \times 3$)	BatchNorm
InstanceNorm	LeakyReLU(0.2)
ReLU	Dense(1)
TransposeConv($16 \times 3 \times 3$)	Sigmoid
InstanceNorm	
ReLU	
TransposeConv($8 \times 3 \times 3$)	
InstanceNorm	
ReLU	
TransposeConv(input channel $\times 3 \times 3$)	
Tanh	

A.5. Additional Confusion Matrices

We provide the remaining 7 datasets' confusion matrices results in Figure 5.



(a) Pre-attack (EAMC) (b) Post-attack (EAMC) (c) Pre-attack (SiMVC) (d) Post-attack (SiMVC)

(e) Pre-attack (CoMVC) (f) Post-attack (CoMVC) (g) Pre-attack (Multi-VAE) (h) Post-attack (Multi-VAE)

(i) Pre-attack (InfoDDC) (j) Post-attack (InfoDDC) (k) Pre-attack (SEM) (l) Post-attack (SEM)

(m) Pre-attack (AR-DMVC) (n) Post-attack (AR-DMVC)

Figure 5. Confusion matrices illustrating the effect of the attack for the EAMC, SiMVC, CoMVC, Multi-VAE, InfoDDC, SEM, and AR-DMVC models on the NoisyMNIST.

A.6. Hyperparameters Results

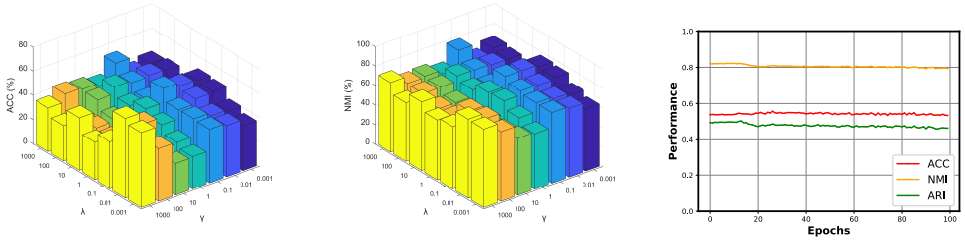
As shown in Figures 6-8, we provide the remaining hyperparameters results.

A.7. Visualising Generated Adversarial Images

We present sample adversarial images for all models across all datasets. The predominant cluster labels for each image are displayed directly above the image. Additionally, we showcase the adversarial noise generated by our generators for each individual sample.

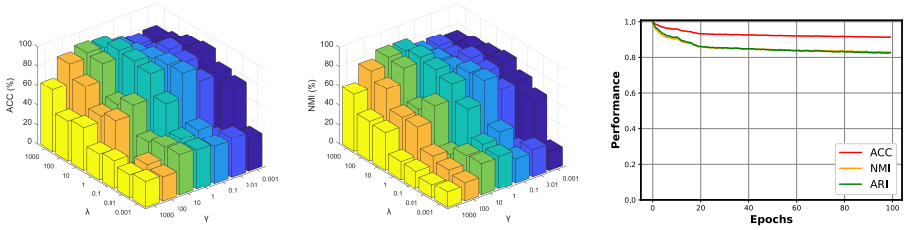
A.7.1. RegDB

Please refer to Figures 9-17 for the RegDB dataset.



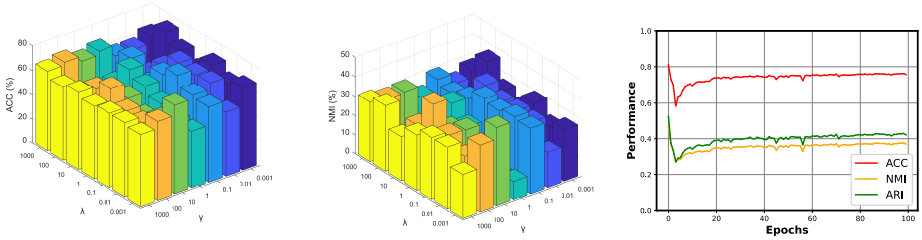
(a) RegDB (b) RegDB

Figure 6. (a) ACC and NMI versus parameters λ and γ of AR-DMVC-AM. (b) The clustering results with respect to the epoch of AR-DMVC-AM under attack.



(a) NoisyMNIST (b) NoisyMNIST

Figure 7. (a) ACC and NMI versus parameters λ and γ of AR-DMVC-AM. (b) The clustering results with respect to the epoch of AR-DMVC-AM under attack.



(a) PatchedMNIST (b) PatchedMNIST

Figure 8. (a) ACC and NMI versus parameters λ and γ of AR-DMVC-AM. (b) The clustering results with respect to the epoch of AR-DMVC-AM under attack.

A.7.2. NoisyFashion

Please refer to Figures 18- 26 for the NoisyFashion dataset.

A.7.3. NoisyMNIST

Please refer to Figures 27- 35 for the NoisyMNIST dataset.

A.7.4. PatchedMNIST

Please refer to Figures 36- 44 for the PatchedMNIST dataset.

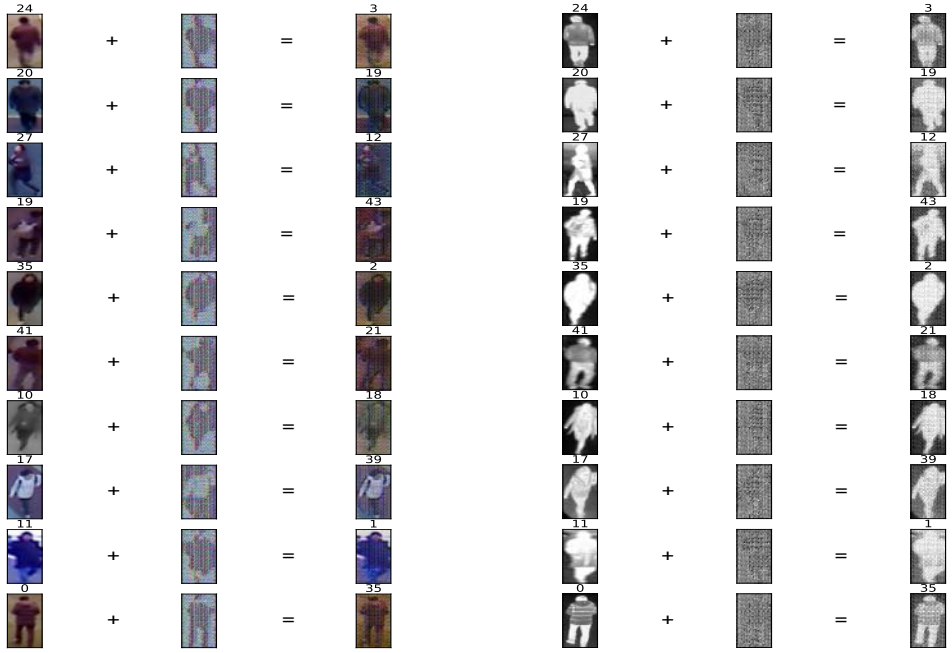


Figure 9. EAMC (RegDB)

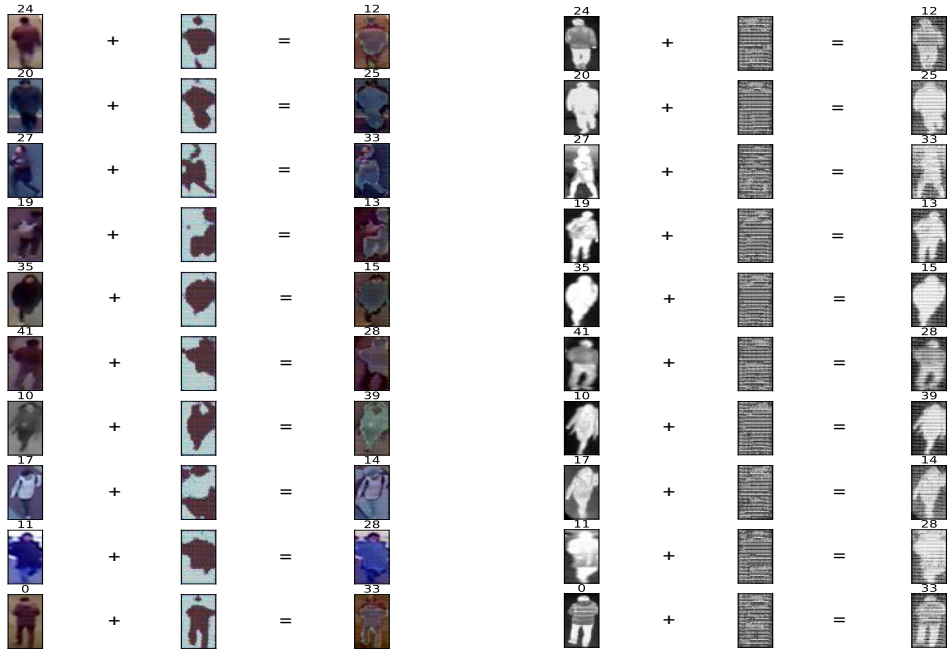


Figure 10. SiMVC (RegDB)

Figure 11. CoMVC (RegDB)

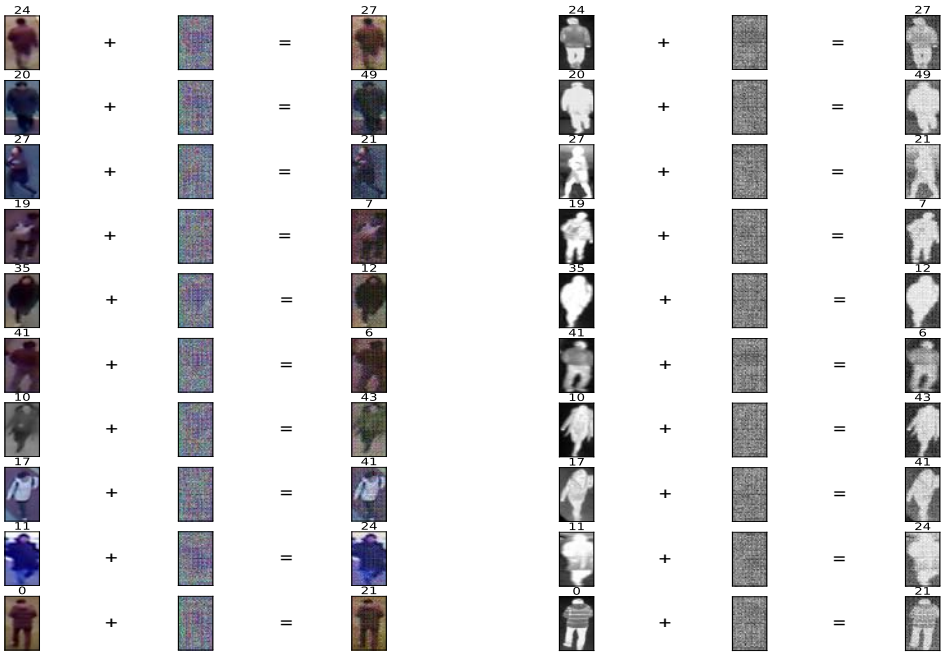


Figure 12. Multi-VAE (RegDB)

Figure 13. AECoDDC (RegDB)

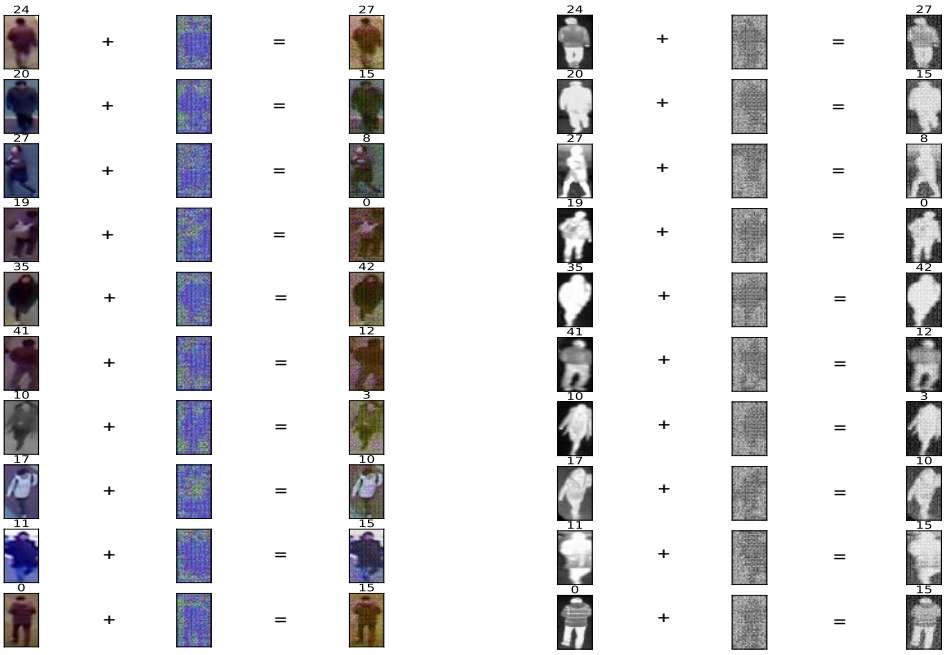


Figure 14. InfoDDC (RegDB)

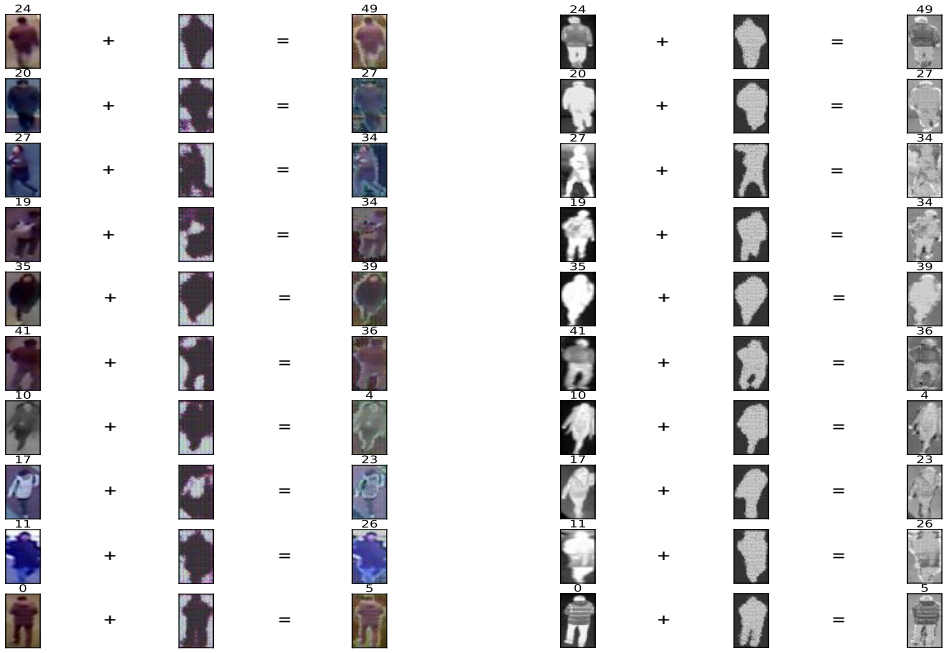


Figure 15. SEM (RegDB)

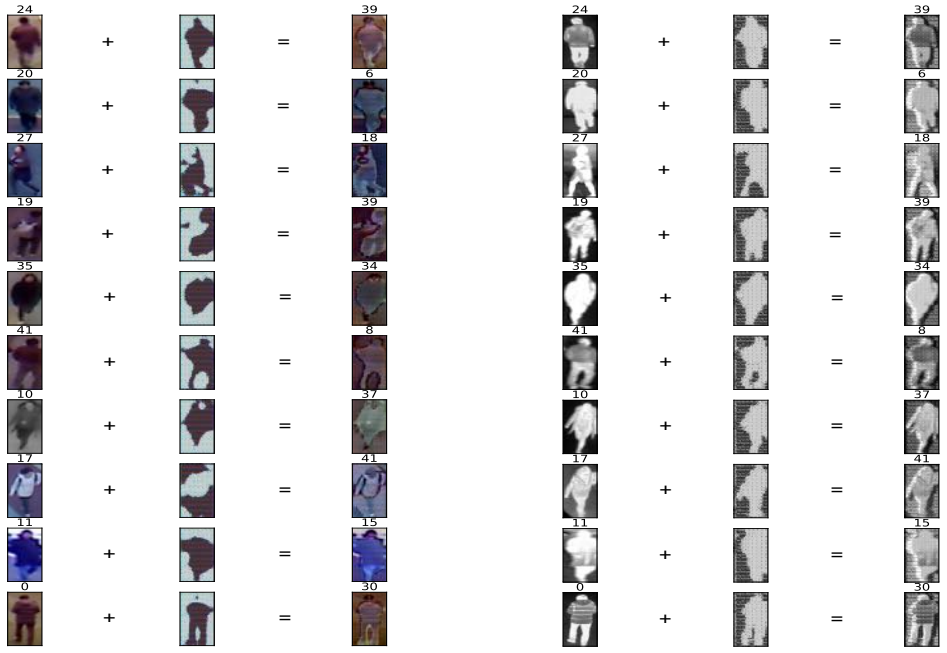


Figure 16. AR-DMVC (RegDB)

Figure 17. AR-DMVC-AM (RegDB)

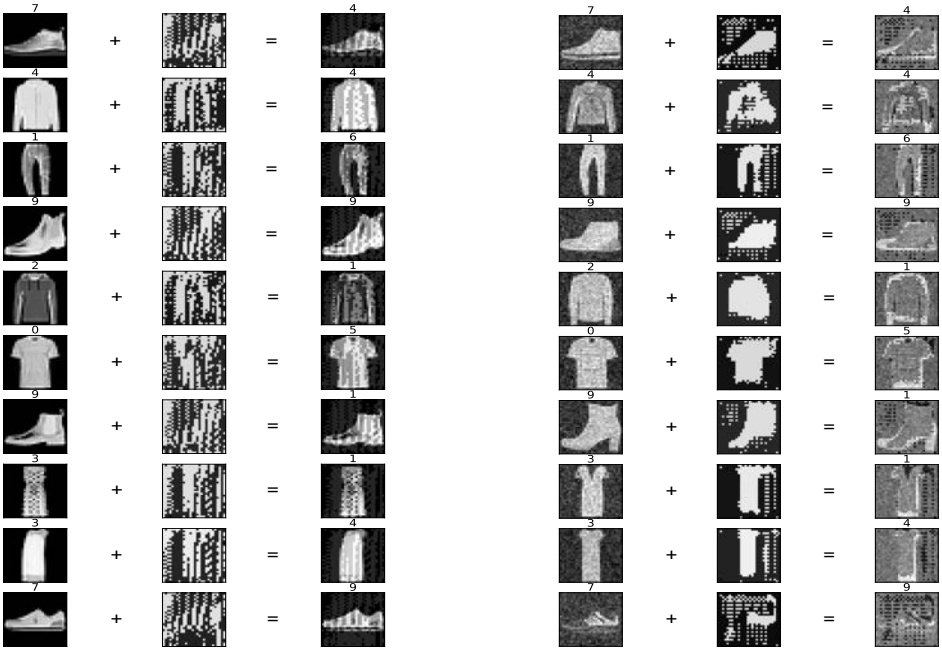


Figure 18. EAMC (NoisyFashion)

Figure 19. SiMVC (NoisyFashion)

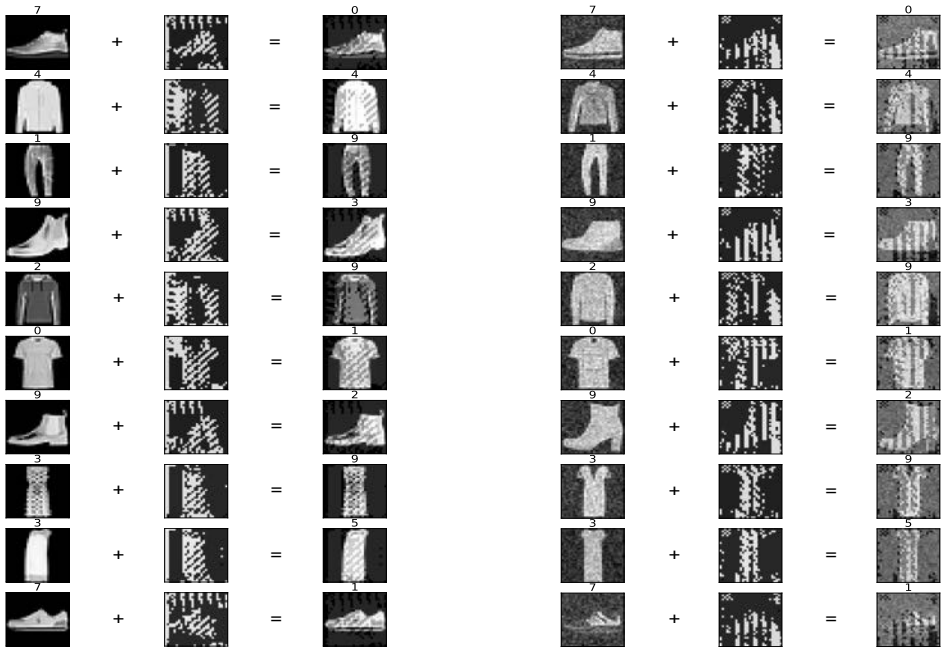


Figure 20. CoMVC (NoisyFashion)

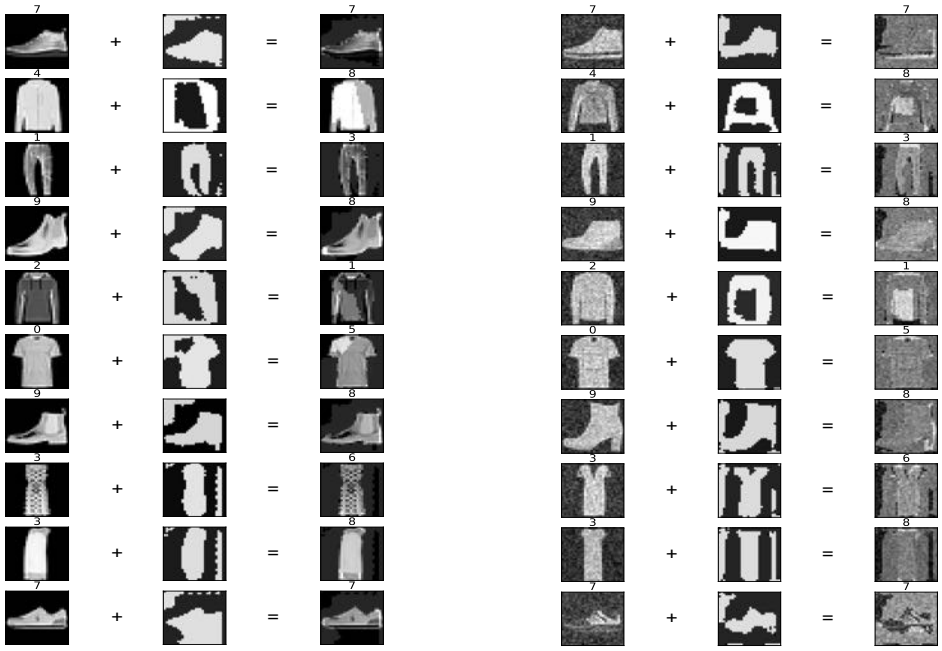


Figure 21. Multi-VAE (NoisyFashion)

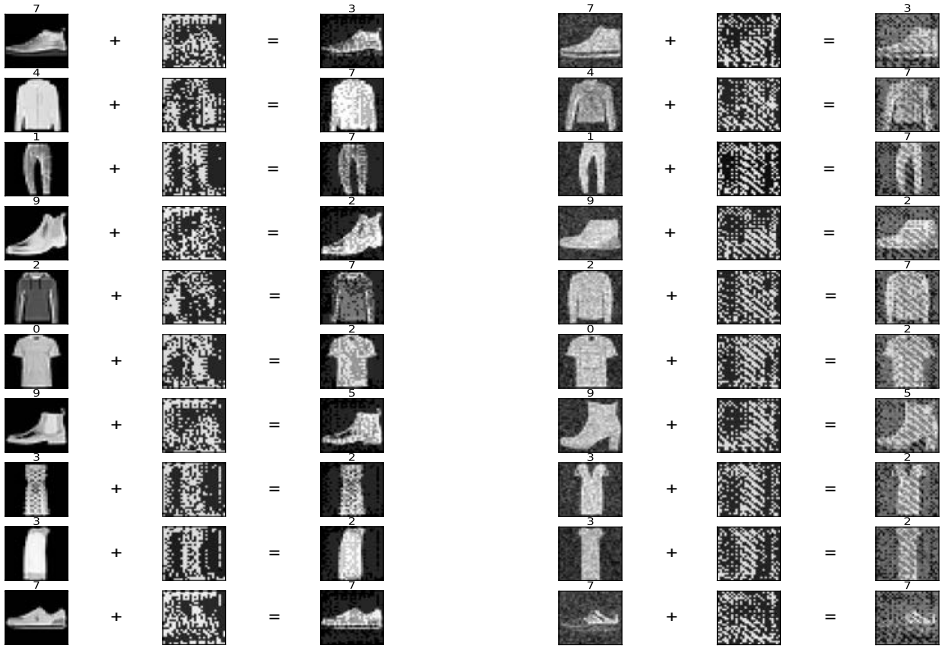


Figure 22. AECoDDC (NoisyFashion)

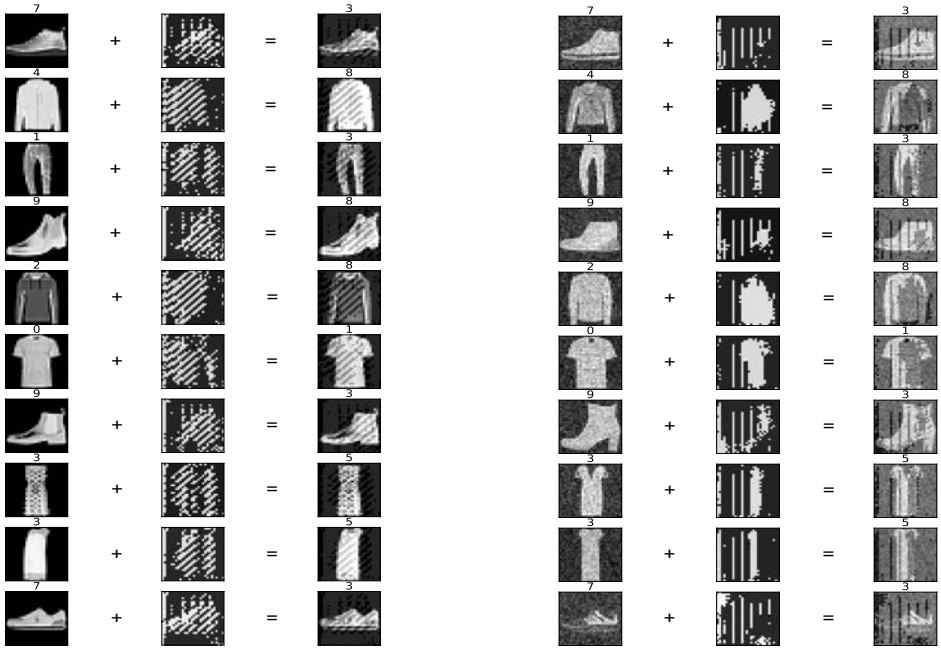


Figure 23. InfoDDC (NoisyFashion)

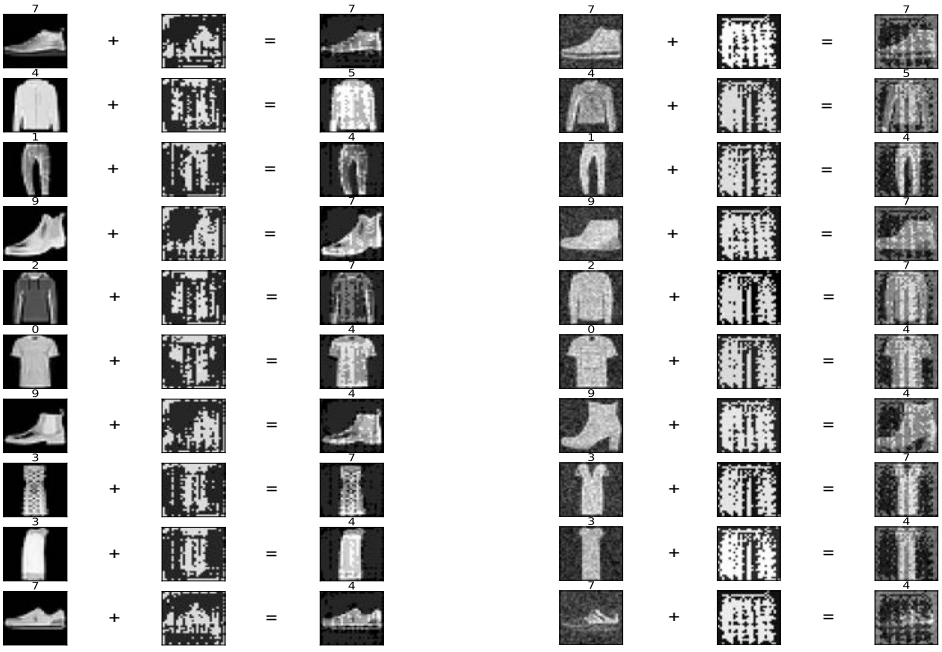


Figure 24. SEM (NoisyFashion)

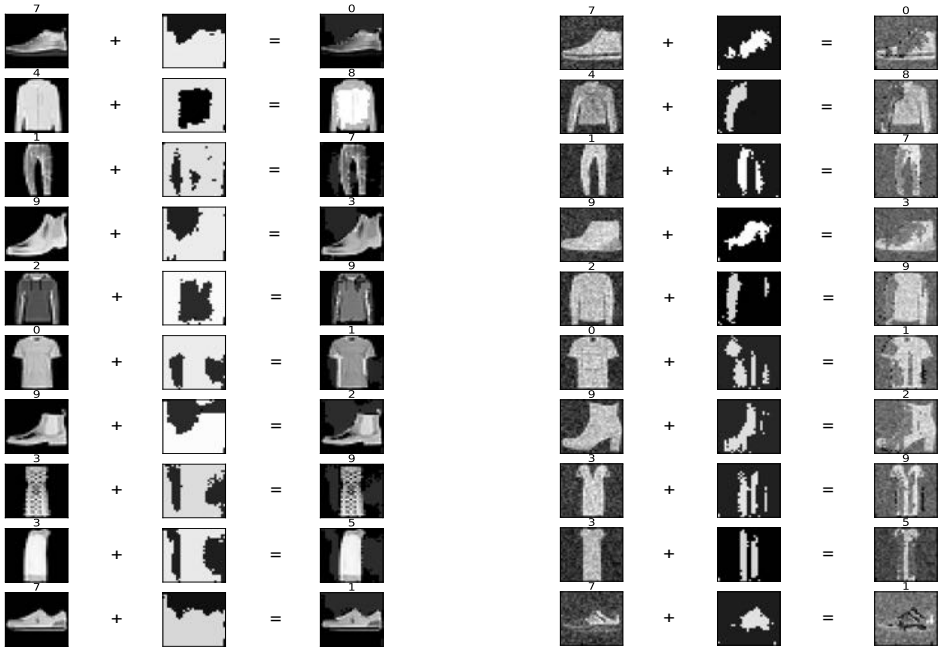


Figure 25. AR-DMVC (NoisyFashion)

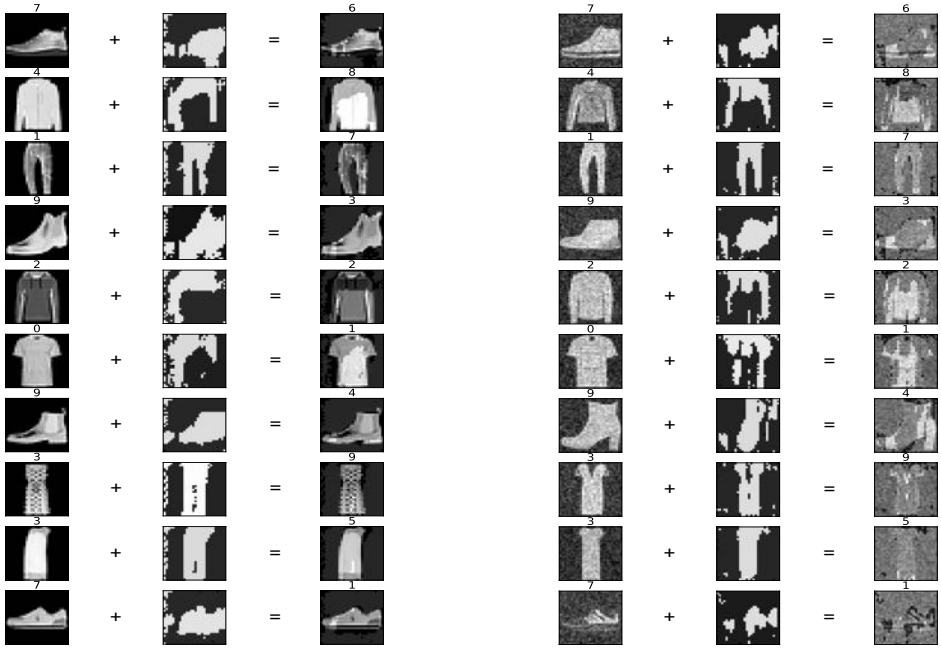


Figure 26. AR-DMVC-AM (NoisyFashion)

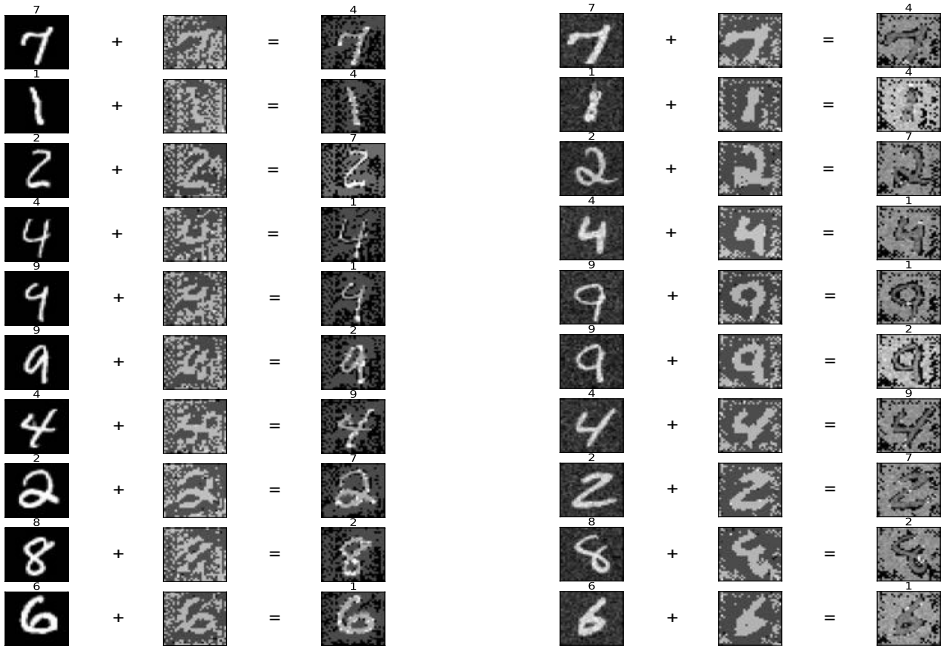


Figure 27. EAMC (NoisyMNIST)

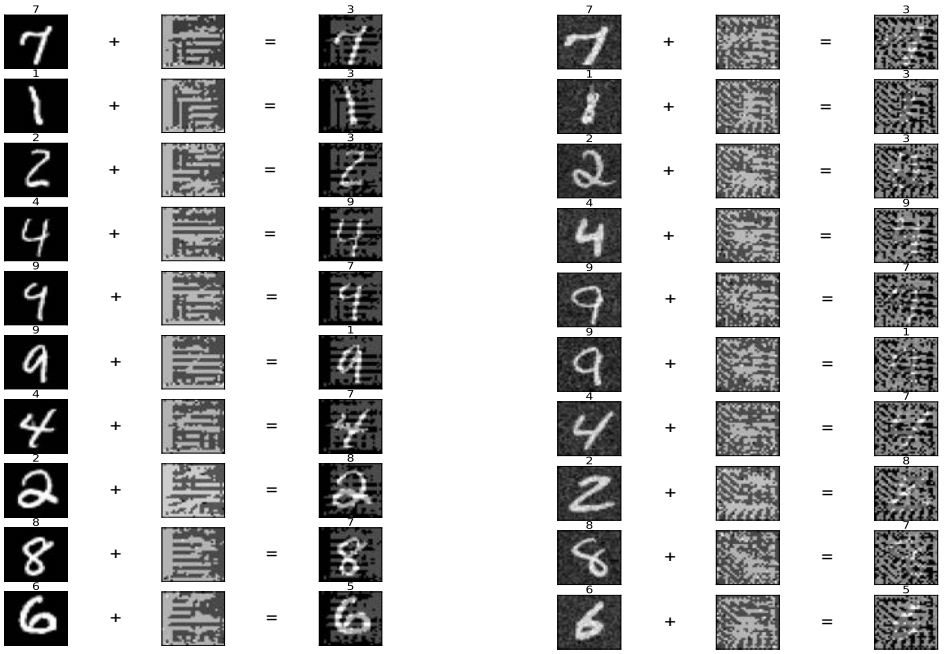


Figure 28. SiMVC (NoisyMNIST)

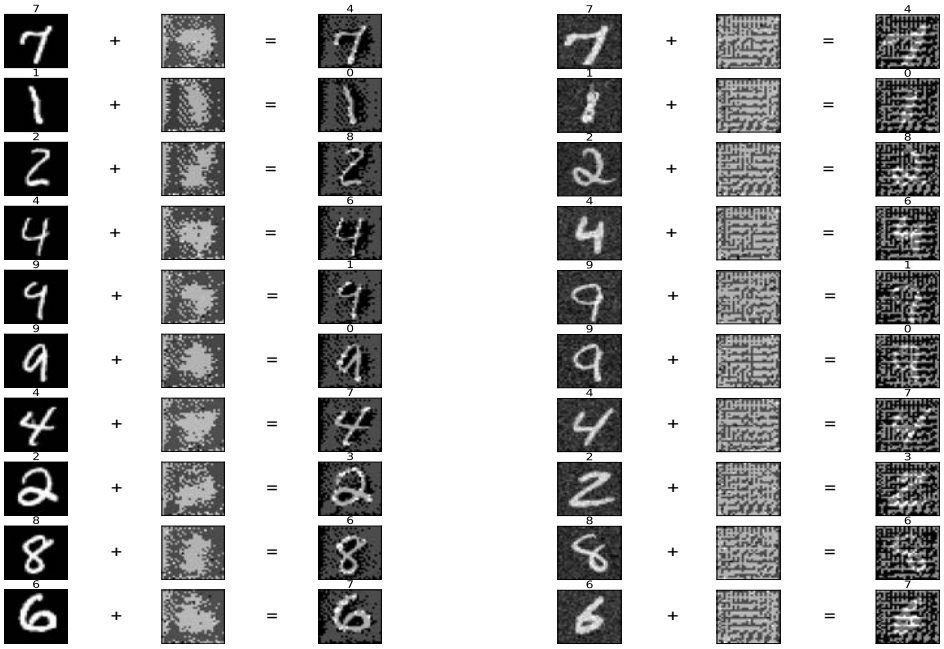


Figure 29. CoMVC (NoisyMNIST)

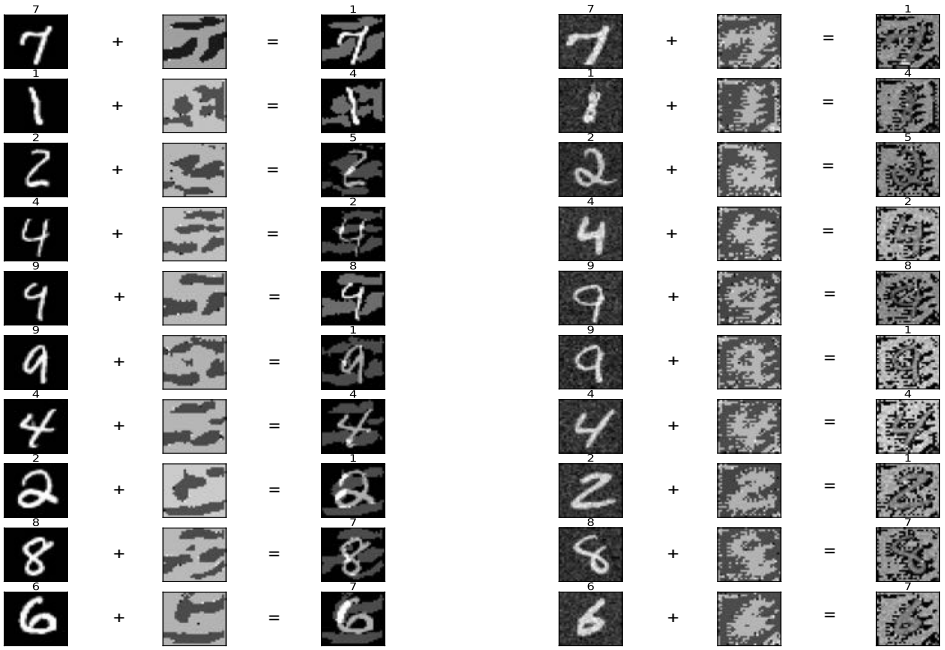


Figure 30. Multi-VAE (NoisyMNIST)

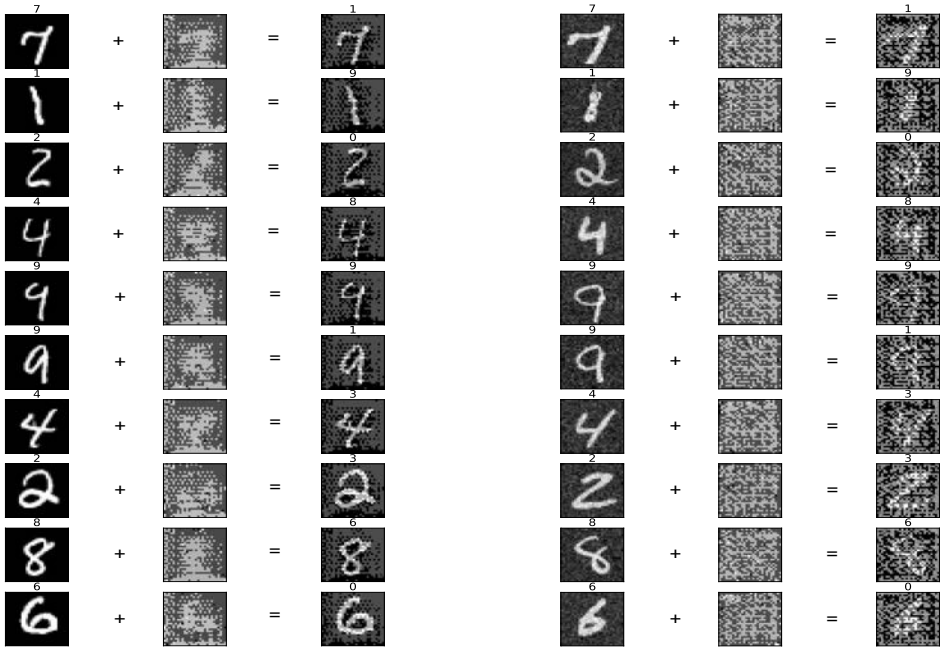


Figure 31. AECoDDC (NoisyMNIST)

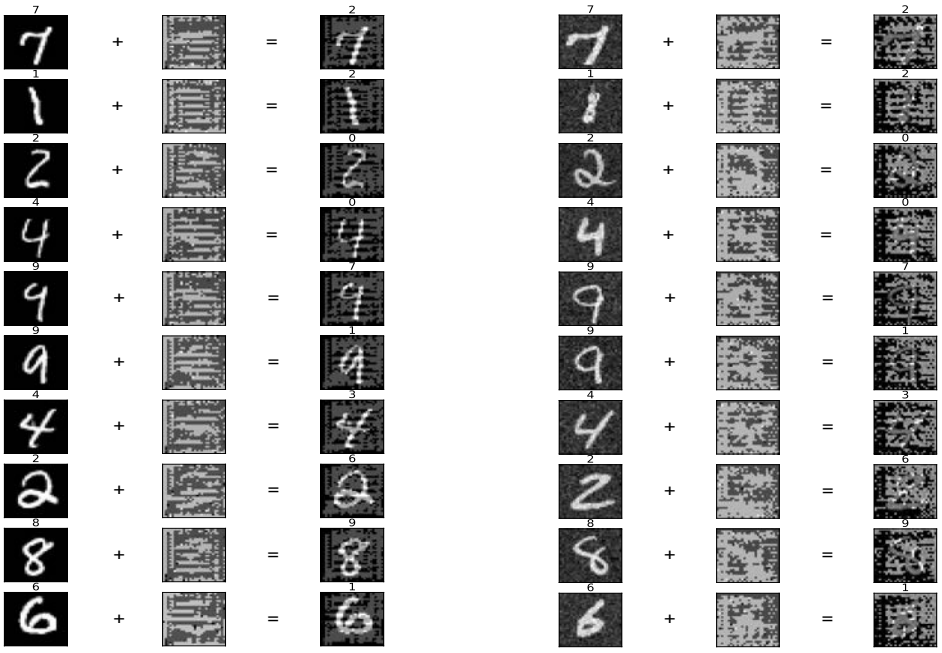


Figure 32. InfoDDC (NoisyMNIST)

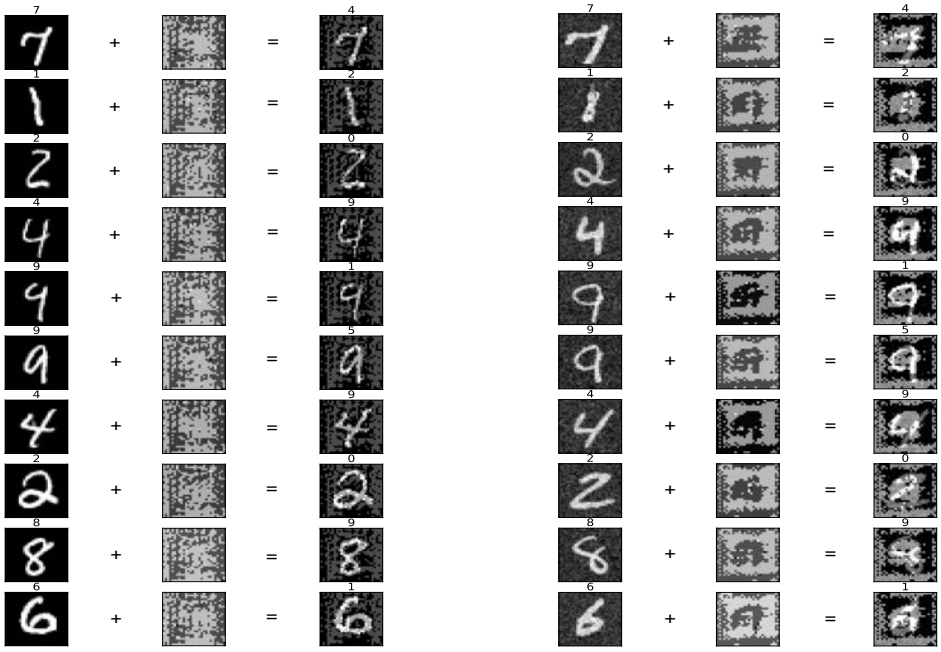


Figure 33. SEM (NoisyMNIST)

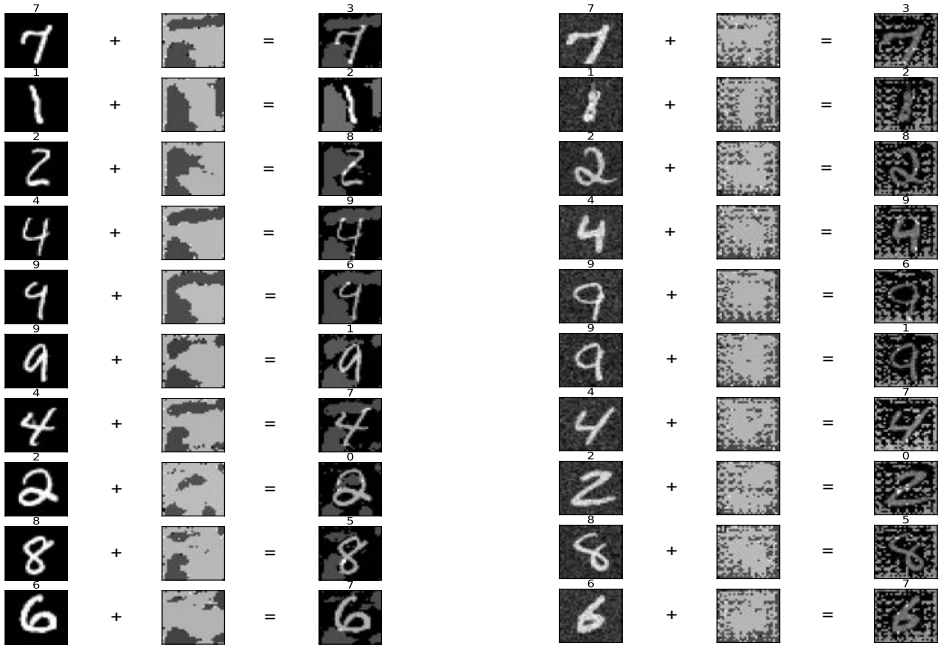


Figure 34. AR-DMVC (NoisyMNIST)

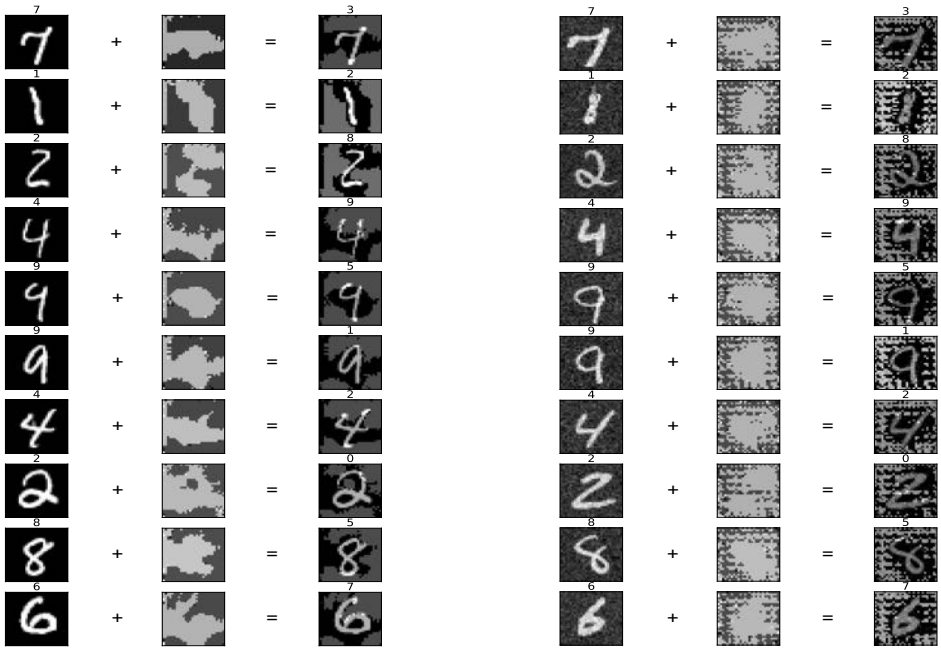


Figure 35. AR-DMVC-AM (NoisyMNIST)

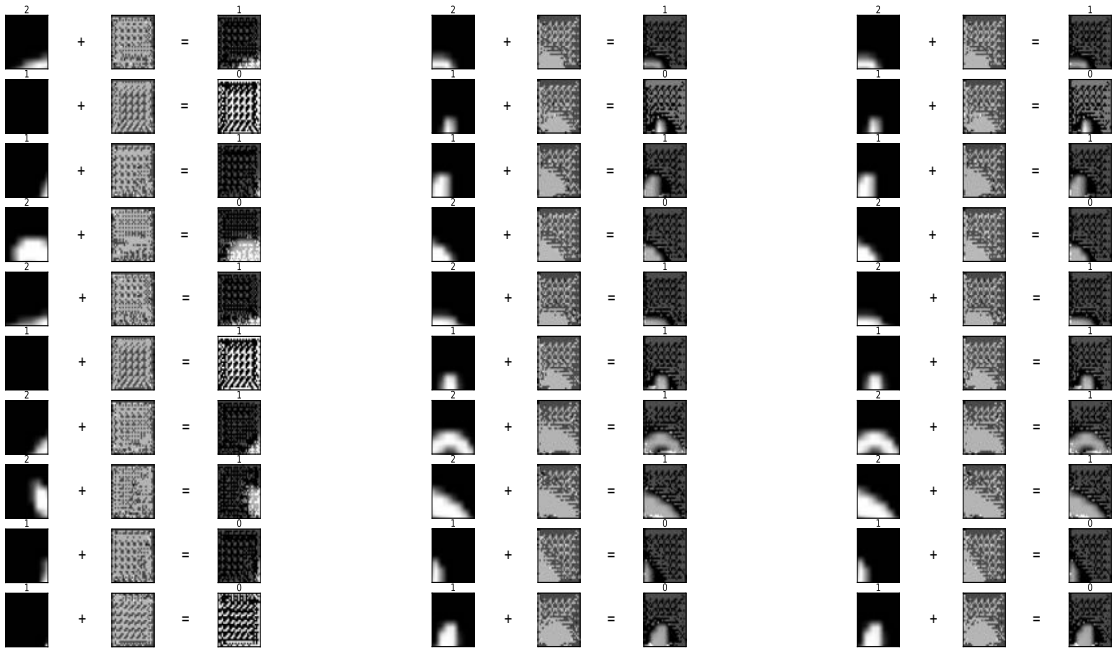


Figure 36. EAMC (PatchedMNIST)

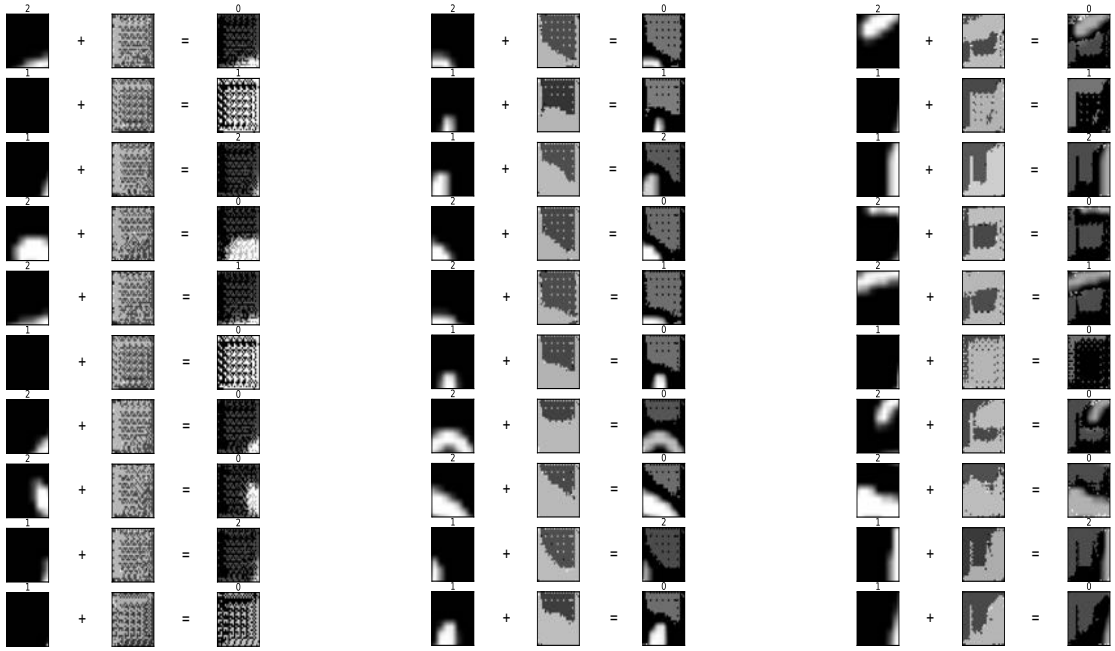


Figure 37. SiMVC (PatchedMNIST)

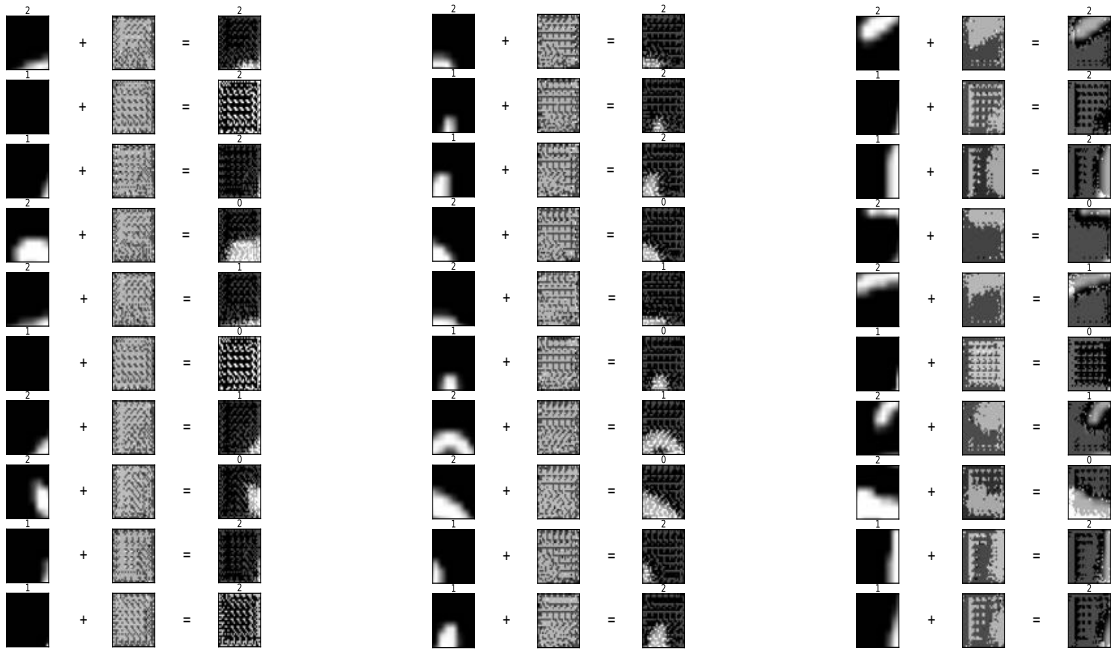


Figure 38. CoMVC (PatchedMNIST)

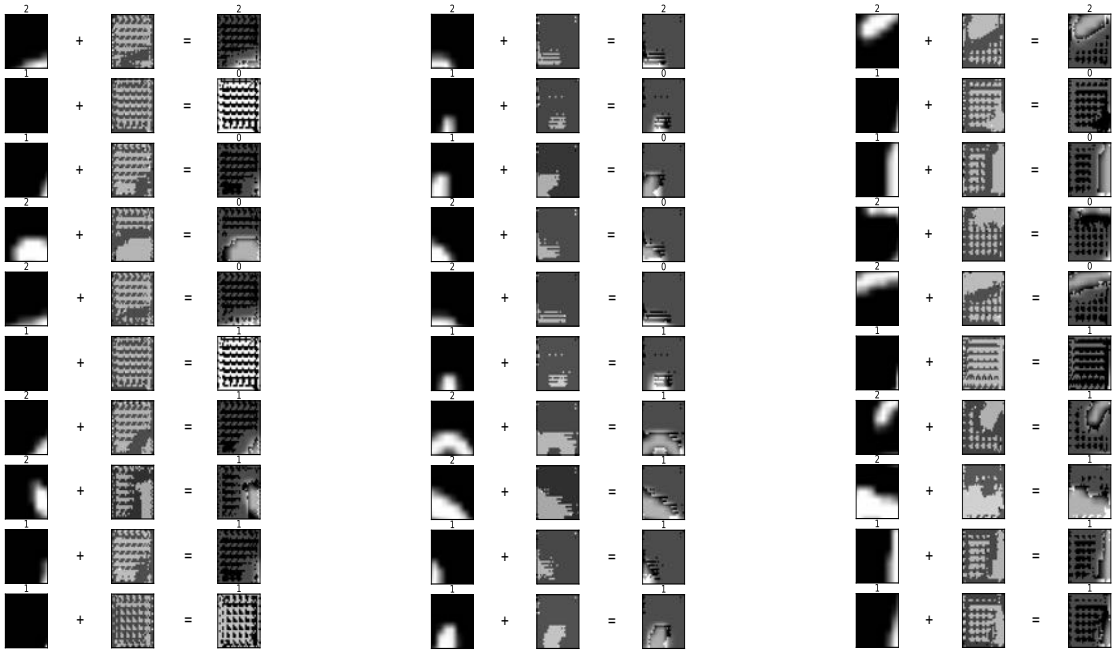


Figure 39. Multi-VAE (PatchedMNIST)

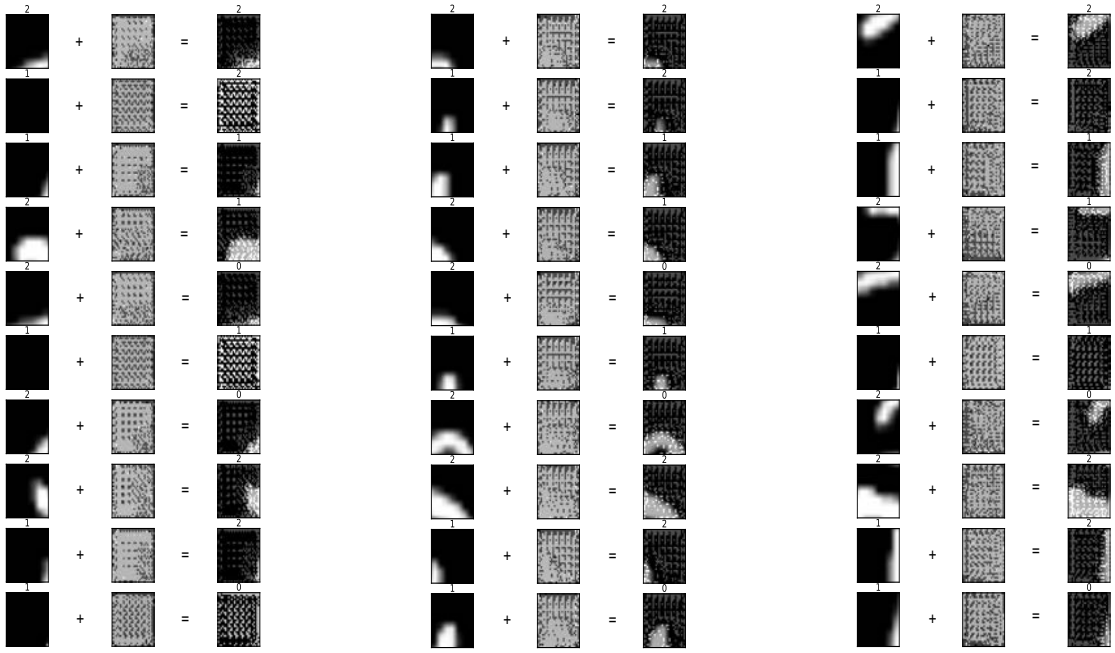


Figure 40. AECoDDC (PatchedMNIST)

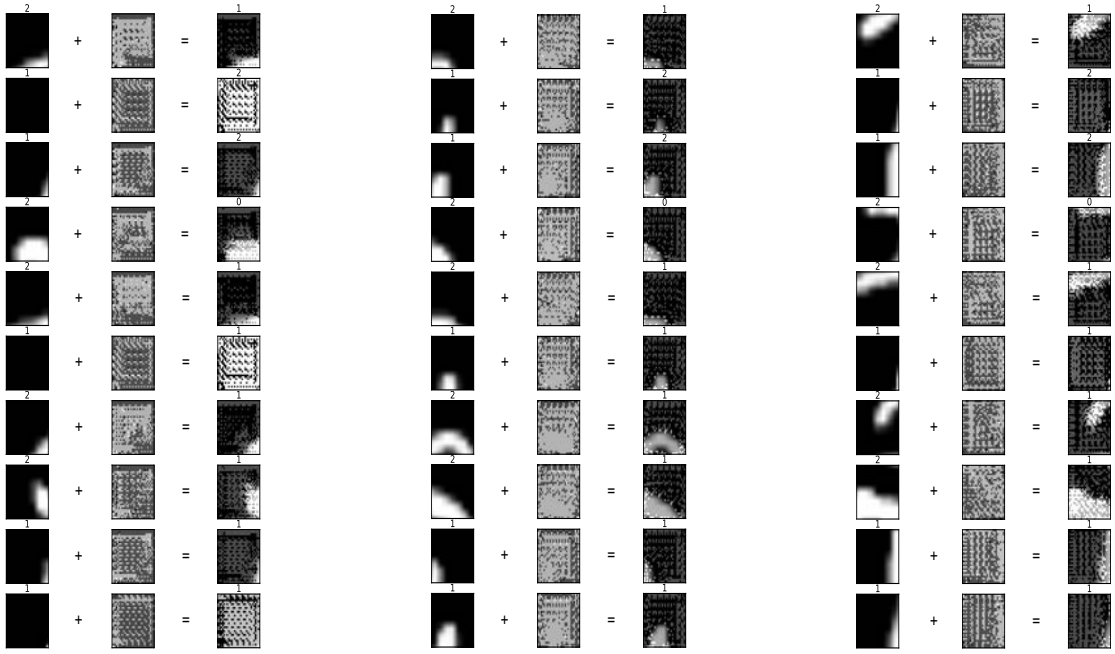


Figure 41. InfoDDC (PatchedMNIST)

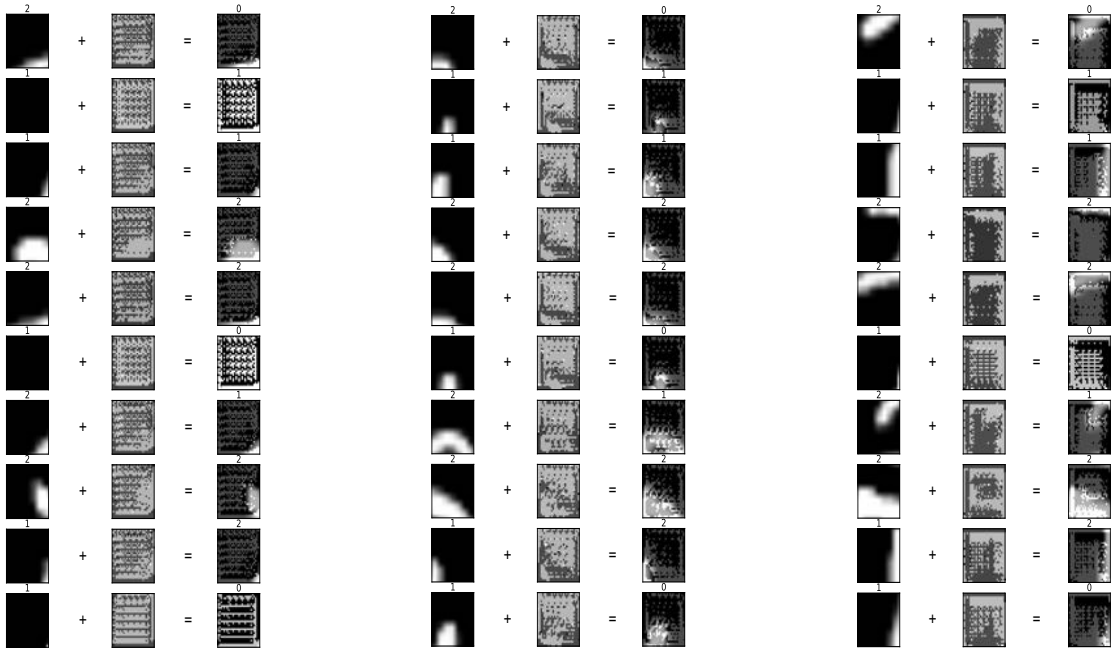


Figure 42. SEM (PatchedMNIST)

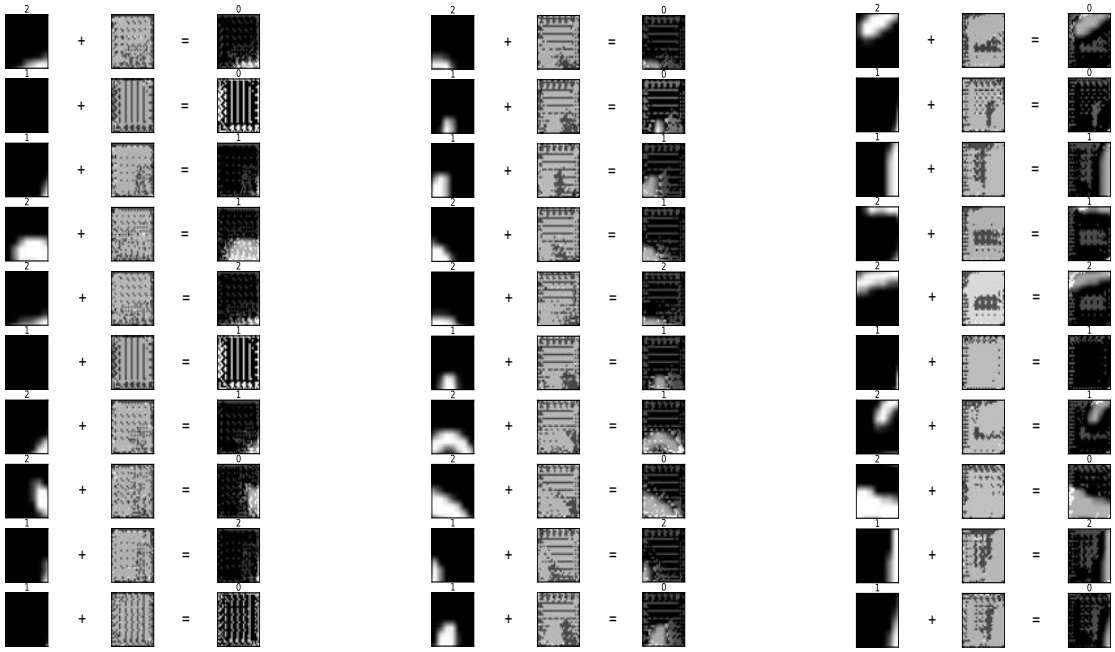


Figure 43. AR-DMVC (PatchedMNIST)

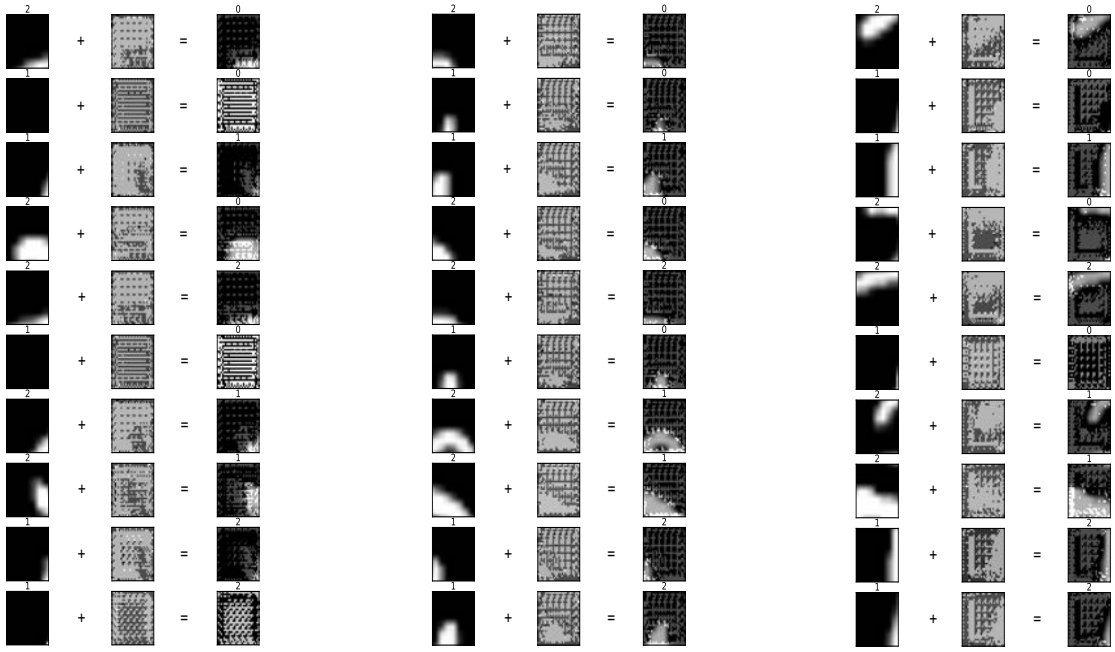


Figure 44. AR-DMVC-AM (PatchedMNIST)