DA-AE: Disparity-Alleviation Auto-Encoder Towards Categorization of Heritage Images for Aggrandized 3D Reconstruction.

Dikshit Hegde, Tejas Anvekar, Ramesh Ashok Tabib, Uma Mudengudi
Center of Excellence in Visual Intelligence (CEVI), KLE Technological University.
Vidya Nagar, Hubballi, Karnakata, India.
dikshithegde@gmail.com, anvekartejas@gmail.com, ramesh_t@kletech.ac.in, uma@kletech.ac.in

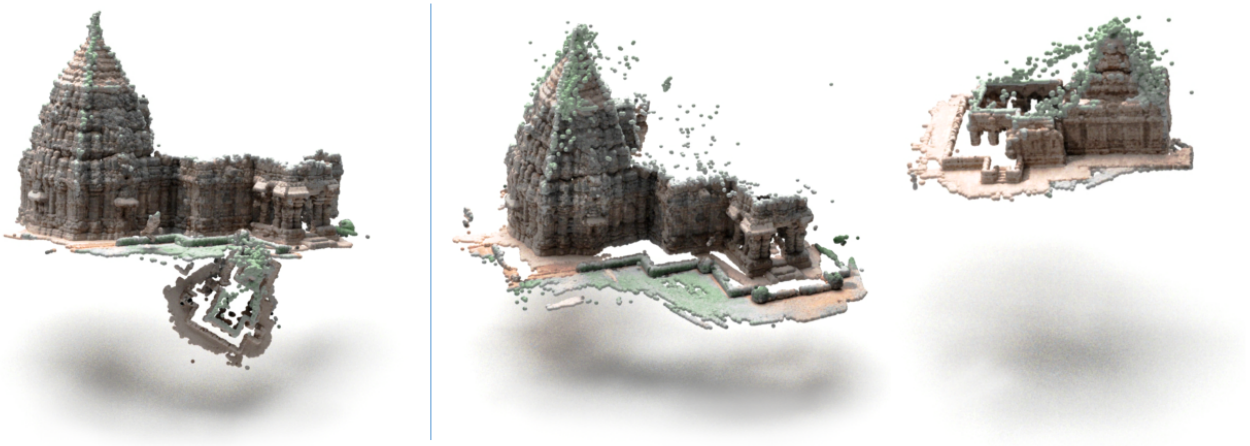

Figure 1. Comparison of 3D models produced by the MVG-MVS pipeline before and after categorization. The categorizations are derived using the proposed framework **DA-AE**: Disparity-Alleviation Auto-encoder that allows for improved data representation in latent space for better clustering (categories). The left image depicts a 3D model of heritage images before categorization with approx **8** million points. The right images are 3D models of heritage images that have been categorized with approx **4.7** million points individually using DA-AE thus yielding better 3D reconstruction.

Abstract

In this paper, we propose DA-AE: Disparity Alleviation AutoEncoder for categorization of heritage images towards 3D reconstruction. Recent survey on preservation of heritage shows demand for the digitization and conservations of heritage sites owing to their susceptibility to natural disasters and human acts. Digital conservation can be facilitated via crowdsourcing of data useful for construction of 3D models. Data from multiple sites sourced may result in elimination of relevant images due to the limitations of the pipeline. Curation and categorization of the crowdsourced data enables better 3D reconstruction. 3D reconstruction pipelines demand correlation between the data and also tries to eliminate the irrelevant information. The reconstruction pipeline is sensitive to selection of initial pair for reconstruction. By categorising individual sites, crowdsourced data can be used to create better 3D reconstructed models. Categorization of crowdsourced data demands learning robust representations of data. Towards this, we propose DA-AE for improved representation and categorization of data in latent space, along with a disparity alleviation loss. We demonstrate categorization as an event, with clustering as a downstream task. We compare our results of clustering with state-of-the-art methods on benchmark datasets (MNIST, FashionMNIST, and USPS). We demonstrate the effects of our categorization using custom dataset IDH10 and compare the results with state-of-the-art methods. We show a systematic and qualitative influence of the proposed method on 3D reconstruction of data.

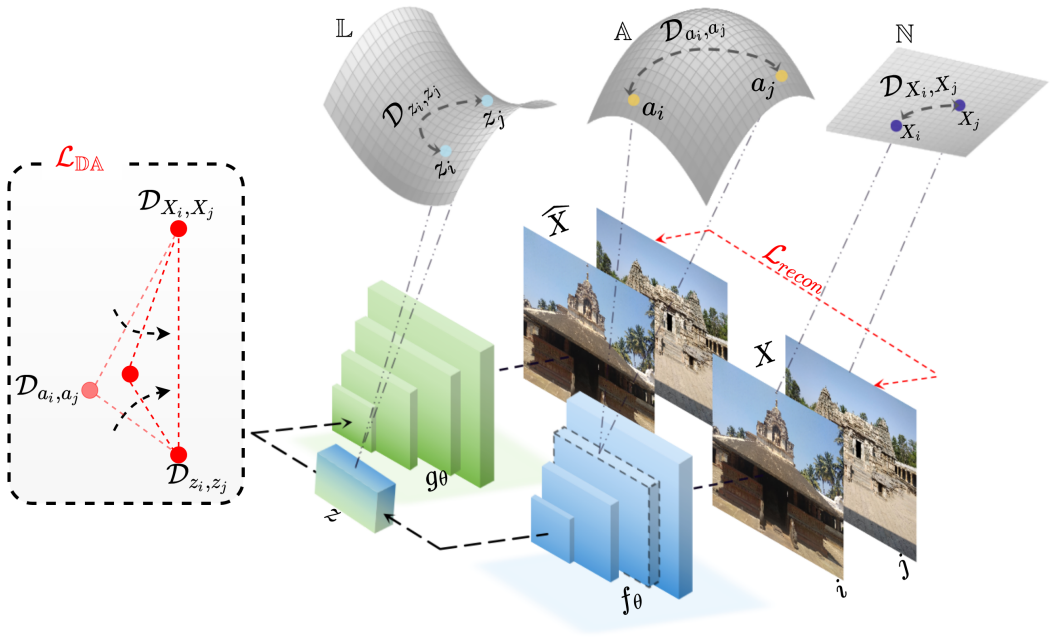

Figure 2. Disparity-Alleviation Auto-Encoder (DA-AE): $\boldsymbol{f_\theta}$ represents DA Encoder, $\boldsymbol{g_\theta}$ represents Decoder, $\boldsymbol{z}$ is a latent in Latent Space $\mathbb{L}$, $\boldsymbol{X}$ is input in Natural Space $\mathbb{N}$, and intermediate layer is tapped in Neural Representation space $\mathbb{A}$.

1. Introduction

In this paper, we propose to learn representations of cultural heritage data towards categorization and 3D reconstruction towards presentation in digital space. Cultural heritage refers to the physical artefacts and intangible features of a group or culture that are passed down from previous generations, conserved in the present, and left to future generations. Cultural heritage deteriorates over time or is destroyed by natural calamities or human acts. Cultural heritage preservation is both a goal and a need for the whole globe. Conservation entails more than simply maintaining a facade or immobilising a monument in time. We prefer restoration over replacement to maintain the ambiance, character, and a living piece of our past. With modern technology like photographs, video, audio, text, and photogrammetry, conservation of cultural heritage is a simple and effective task.

Photogrammetry (3D reconstruction) is the art, science, and technology of obtaining reliable information about physical objects and the environment through processes of recording, measuring, and interpreting photographic images. 3D reconstruction is a three-dimensional coordinate measuring approach that use pictures as the primary metrology medium. Creating 3D models of cultural heritage sites necessitates a massive quantity of data. Crowdsourced data collection is the process of constructing photographs with the assistance of the community. The crowd-sourced data incorporates information from numerous sources, which affects the 3D reconstruction. The crowdsourced data comprises blur, occlusions, watermarks and so on, necessitating curation for better 3D reconstruction [27]. The number of categories in crowdsourced data is uncertain and demands unsupervised categorization for improved 3D reconstruction [1] [28] [16].

Categorization of high-dimensional data is a time-consuming and imprecise operation. Finding the data manifold where visualisation of the data and its attributes are disentangled in such a manner that learning the mapping function to a lower dimension becomes easy is what representation learning is all about. The representation learning approach allows the system to turn raw data into an inferable state, which can then be utilised for downstream tasks like categorization, classification, and clustering. Data clustering has been studied for many years. Several methods for distinguishing data based on intrinsic similarities have been developed. Conventional clustering, subspace clustering, and deep clustering are the three primary types of clustering algorithms. As a downstream task, clustering is susceptible to data representations obtained from handcrafted features or data-dimensionality reduction methods (PCA, kernel-PCA, Auto-encoder, Nearest-Neighbor based matrix factorization).

Current works in deep clustering (LSNMF [18], EnSC [35], AECL [26], DeepClustering [2], DCN [31], DEC [30], IDEC [7], SR-KMeans [10], VaDE [12], JULE [33], DEPICT [3], DynAE [20]) represents data in latent space us-

ing an autoencoder and clustering loss over the representation learning that help deep learning architectures to categorize the data. Due to unsupervised optimization of reconstruction loss, an autoencoder's latent representations are directly unreliable. Applying a clustering loss to latent space of autoencoder may not generalize the given data.

Towards this, our contributions are:

- DA Encoder for extraction of disentangled latent space supporting theory of indiscernibles via three point approximation.
- A novel optimization for unsupervised feature distillation of latent space using Difference in Difference technique
- A novel loss function $\mathbb{DA}$ Loss ($\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$) to register the variations in the latent space with Jacobian and nonlinear independent component analysis.
- We demonstrate the results of our methodology on MNIST, FashionMNIST, USPS benchmark datasets with a customized dataset IDH10, and compare the results with state-of-the-art methods using quantitative metrics.
- We show the systematic and qualitative impact of our proposed method on 3D reconstruction.

In Section 2, we discuss the proposed methodology DA autoencoder. We discuss the results and its effect on 3D reconstruction of heritage sites in Section 3, and we conclude in Section 4.

2. DA-AE: Disparity-Alleviation Auto-Encoder

We model, DA-AE: Disparity-Alleviation Auto-Encoder towards categorization of heritage images in order to facilitate 3D reconstruction. Towards categorization, we propose a process Disparity-Alleviation, to model DA encoder with $\mathbb{DA}$ loss ($\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$) as shown in Figure 2. DA encoder and $\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$ are inspired from triangular inequality [4].

Preposition 1. Let us assume a Vanilla encoder $\boldsymbol{f_\theta}$ and tap an intermediate layer, by calling it as DA Encoder. The DA Encoder facilitates to form three point approximations of the input $\boldsymbol{X}$. The three approximations being the image space (input) or natural space $\mathbb{N}$, the neural representation space $\mathbb{A}$, and the latent space $\mathbb{L}$ as shown in Figure 2.

Preposition 2. Assume two arbitrary inputs $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ to DA encoder yielding ($\boldsymbol{a_i}$, $\boldsymbol{a_j}$) in neural representation space $\mathbb{N}$ and ($\boldsymbol{z_i}$, $\boldsymbol{z_j}$) in latent space $\mathbb{L}$ respectively. $\mathbb{DA}$ Loss ($\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$) models the per point difference $\boldsymbol{\mathcal{D}}$ between the arbitrary representation $(\boldsymbol{i}, \boldsymbol{j})$ as a point in abstract space. We propose to minimize the disparity-alleviation (triangular inequality) among $\boldsymbol{\mathcal{D}}_{\boldsymbol{X_i},\boldsymbol{X_j}}$, $\boldsymbol{\mathcal{D}}_{\boldsymbol{a_i},\boldsymbol{a_j}}$ and $\boldsymbol{\mathcal{D}}_{\boldsymbol{z_i},\boldsymbol{z_j}}$ as shown in Figure 2.

Intuitively, the proposed DA autoencoder facilitates to register the variations among spaces (image space $\mathbb{N}$, neural representation space $\mathbb{A}$, and latent space $\mathbb{L}$). The registration aids the encoder to learn an abstract feature space that has discriminative flavour to it unlike AE [26], VAE [15], stacked AE [37] and contractive AE [23]. The latent representations of DA-AE are generalized for both abstraction and generation. In this work, we demonstrate the event of categorization with clustering as a downstream task using the abstraction of DA-AE.

2.1. DA Encoder

DA Encoder faclitates to extract representations that satisfies the $\mathbb{DA}$ Loss ($\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$) as explained in Preposition 2 of Section 2. Towards representation of three point approximations of input, $\boldsymbol{a}$ and $\boldsymbol{z}$ are tapped from the DA Encoder $\boldsymbol{f_\theta}$ as shown in Figure 3. Intuitively we perform three point approximation so that the proposed theory of indiscernibles [5] is sufficed. Indiscernibles induces a contractive-discriminative optimization approach to a autoencoder. The Computation of $\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$ towards optimization of indiscernibility, demands for a maximum variational representation of latent $\boldsymbol{z}$ and neural representation $\boldsymbol{a}$. Towards this, we propose to compute Jacobian of $\boldsymbol{z}$ and $\boldsymbol{a}$ with respect to weight space $\boldsymbol{f_\theta}$. The Jacobian introduces nonlinear independent component analysis (nonlinear ICA) of the respective representations [6].

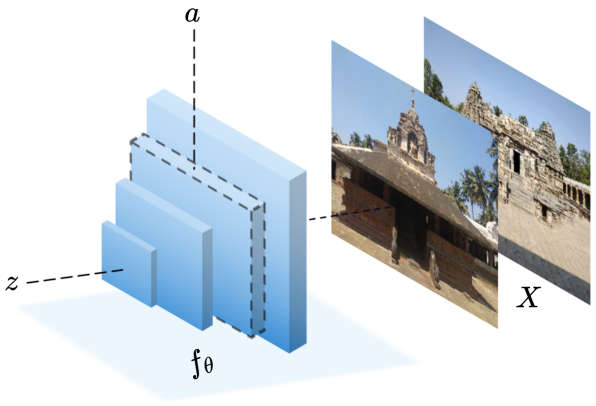

Figure 3. DA Encoder

The over all effect of $\boldsymbol{\mathcal{L}}_{\mathbb{DA}}$ along with nonlinear ICA is to mimic the technique Difference in Difference which is stastistical technique for evaluation of policies, trade patterns in time series data.

Unlike Vanilla autoencoder where the representations

have a direct flow that yield randomness in the latent space, our proposed methodology rectifies, registers and disentangles the randomness from the latent. The disentanglement is achieved from $\mathcal{L}_{\mathbb{DA}}$ yielding a pure approximation of the latent. The pure latents facilitaes better deep clustering performance towards aggrendiced 3D reconstruction of crowdsourced data.

Towards cluster assignment and cluster hardening, we propose to use IDEC [7] as a plugin to DA-AE. IDEC with DA-AE performs better on unsupervised deep clustering with clustering loss $\mathcal{L}_{cl}$as an optimization approach.

2.1.1 Loss Functions

In this section, we propose **SD-MSE** structurally dissimilar mean squared error and consider it as a reconstruction loss ($\mathcal{L}_{recon}$) towards optimizing DA-AE weights. We propose $\mathbb{DA}$**-Loss** ($\mathcal{L}_{\mathbb{DA}}$) towards minimizing the loss between per point difference in abstract space optimizing the DA Encoder weights.

- **Reconstruction Loss** ($\mathcal{L}_{recon}$): As shown in Figure 2, we reconstruct the image using a decoder g_{θ} and compute the loss $\mathcal{L}_{recon}$ between the reconstructed image $\hat{X}$ and the original image X. This loss function ensures that the latent space has the right representation of image X to up-sample the latent z.
  - **MSE** (mean square error / L_2 norm)

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(y - \hat{y}_i)^2 \tag{1}$$

  - **SDSIM** (structural dissimilarity) [21]

$$\mathcal{L}(y, \hat{y}) = 1 - \mathcal{L}_{ssim}(y, \hat{y}) \tag{2}$$

$$\mathcal{L}_{ssim}(y, \hat{y}) = \frac{(2\mu_y\mu_{\hat{y}}+c_1)(2\sigma_{y\hat{y}}+c_2)}{(\mu_y^2+\mu_{\hat{y}}^2+c_1)(\sigma_y^2+\sigma_{\hat{y}}^2+c_2)} \tag{3}$$

    where,

    * $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$.
    * L is dynamic range of pixel values.$(2^n - 1)$
    * k_1 = 0.01, k_2 =0.03 by default.

  - **SD-MSE** (structural dissimilarity with MSE) The proposed loss for reconstructing DA-AE's output leverages the advantages of both MSE and SDSIM loss. The term $\alpha \in (0, 1)$ is used for normalizing the effects of combined loss.

$$\mathcal{L}_{recon}(y, \hat{y}) = \alpha * \mathcal{L}_{mse}(y, \hat{y}) + (1 - \alpha) * (\mathcal{L}_{SDSIM}(y, \hat{y})) \tag{4}$$

- $\mathbb{DA}$ **Loss**:

$$\mathcal{L}_{\mathbb{DA}} = |\mathcal{D}_{X_i, X_j} - \mathcal{D}_{\mathcal{J}_{a_i}, \mathcal{J}_{a_j}}| + |\mathcal{D}_{\mathcal{J}_{a_i}, \mathcal{J}_{a_j}} - \mathcal{D}_{\mathcal{J}_{z_i}, \mathcal{J}_{z_j}}| \tag{5}$$

  Here $|\cdot|$ represents absolute value and $\mathcal{J}$ represents Jacobian with respect to DA Encoder f_{θ}. The per point difference $\mathcal{D}$ between arbitary representation (i, j) is one of the below functions representing them in abstract space.

  - **MSE** (mean square error / L_2 norm)

$$\mathcal{D}(m_i, m_j) = (m_i - m_j)^2 \tag{6}$$

  - **MAE** (mean absolute error / L_1 norm)

$$\mathcal{D}(m_i, m_j) = |m_i - m_j| \tag{7}$$

    here, $|\cdot|$ represents absolute value.

  Here m is a point in Natural Space $\mathbb{N}$, Neural representation space $\mathbb{A}$, or Latent space $\mathbb{L}$.

- **Clustering Loss** ($\mathcal{L}_{cl}$): $\mathcal{L}_{cl}$ is used to assign cluster assignments to data points directly. Therefore, no additional clustering algorithm on top of the learned latent representations is required. K-means loss [32], cluster assignment hardening loss [30], and agglomerative clustering loss [33] are few examples.
  - **Kullback-Leibler (KL) divergence [14]:** The KL divergence is defined as the negative sum of each event's probability in P multiplied by the log of the event's probability in Q over the probability of the event in P [30].

$$KL(P||Q) = \sum_{x \in \mathcal{X}} P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{8}$$

3. Results and Discussions

In this section, we demonstrate the results of our proposed methodology and compare them with state-of-the-art method on categorization using appropriate quality metrics. This section includes the dataset used for the Experimentations, Experimental setup and details of the evaluation metrics.

3.1. Dataset

To evaluate our proposed methodology with state-of-the-art KMeans [13], GMM [22], LSNMF [18], AC [11], SSC-OMP [36], EnSC [35], SC [25], AECL [26], DeepCluster [2], DCN [31], DEC [30], IDEC [7], SR-KMeans [10], VaDE [12], JULE [33], DEPICT [3], DynAE [20] we consider following datasets.

Table 1. DA-AE Architecture details, Here DA encoder architecture is represented in and decoder architecture is represented in .

Type	Input → Output	Activation	Normalization
Linear	784 → 500	GELU	Layernorm
Linear	500 →500	-	-
Linear	500 →2000	GELU	Layernorm
Linear	2000 → 10	-	-
Linear	10 → 2000	GELU	Layernorm
Linear	2000 → 500	GELU	Layernorm
Linear	500 → 500	GELU	Layernorm
Linear	500 → 784	Sigmoid	-

- **MNIST**:- dataset consists of 70000 handwritten digits which are centered and normalized into a dimension of 28-by-28 pixel size [17].
- **Fashion MNIST**:- dataset consists of 10 classes of different Fashion dress. There are 70000 images in total which are centered and normalized into a dimension of 28-by-28 pixel size [29].
- **USPS**:- is a digitally scanned dataset from envolpes by U.S. postal services. It contais 9298 gray scaled images of 16-by-16 resolution whihc are centered and normalized [9].
- **IDH10**:- dataset consists of curated 2944 images out of imbalanced samples of 7000 images of 10 Indian Heritage sites with structural similarities.

3.2. Experimental setup

In this section, we elaborate about the architectural design and optimization over loss functions discussed in Section 2.1.1. We show ablation on benchmark-datasets with its hyper-parameter settings.

3.2.1 Modeling of DA-AE

The architecture of DA-AE consists of an DA encoder and decoder as shown in Figure 2. For comparison with IDEC [7] and DEC [30] we keep the neural architecture same as them with some minor changes as explained in Table 1. DA-AE's parameters are optimised over loss function $\mathcal{L}_{recon}$ given in Equation 4 and $\mathcal{L}_{\mathbb{DA}}$ given in Equation 5 with adam optimizer. In Table 2 we show the selections of hyperparameters obtained over different benchmark dataset as shown in Algorithm 1.

3.2.2 DA-AE with IDEC plugin

While Training DA-AE for IDEC as a plug-in. DA-AE's hyperparameters are set with-respect-to model with best clustering acc using Grid-search algorithm [24]. The optimization strategies and neural architecture are explained in Section 3.2.1. Consider a dataset $\mathbb{X}$ with n samples and each sample $\boldsymbol{X_i} \in \mathbb{R}^{\boldsymbol{d}}$, where $\boldsymbol{d}$ is dimension of data. The number of clusters $\boldsymbol{k}$ is a prior knowledge and the $\boldsymbol{j^{th}}$ cluster-centroid is denoted by $\boldsymbol{\mu_{ij}} \in \mathbb{R}^{\boldsymbol{d}}$. Let the value of $\boldsymbol{s_i} \in$ [0, 1,...,$\boldsymbol{k}$] represent the cluster index assigned to sample $\boldsymbol{X_i}$. With these settings we train DA-AE with IDEC as a plguin with clustering loss Function explained in Section 2.1.1.

Table 2. Hyper-parameters for training DA-AE on benchmark dataset

Dataset	α	lr	batch size	$\mathcal{D}$
MNIST	0.75	0.0001	512	MAE
FMNIST	0	0.0001	512	MAE
USPS	1.0	0.0001	256	MSE
IDH10	0.5	0.0001	32	MSE

Algorithm 1: Pre-Training DA-AE

Input: Dataset → $\boldsymbol{X}$
Output: DA-AE's best weights; best hyperparameter h

Set grid $\boldsymbol{H}$ for Hyperparameters from Table 2
/* $\alpha \in [0, 1], batch - size = [32, 256, 512], lr = [0.0001]$ */
for *h in* $\boldsymbol{H}$ **do**
  Train DA-AE for 200 epochs with loss as shown in Equation 4.
  if $ACC_{new} >= ACC_{old}$ **then**
    RETURN h, weights f_θ, g_θ

3.3. Evaluation Metrics

We evaluate the performance of proposed methodology using appropriate quantitative metrics considering clustering as a downstream task towards categorization of images. We consider Unsupervised Clustering Accuracy(ACC) and Normalized Mutual Information(NMI) as a quantitative metrics

- **Unsupervised Clustering Accuracy(ACC)**:It employs a mapping function $\boldsymbol{h}$ to determine the optimal mapping between the algorithm's cluster assignment output $\boldsymbol{z}$ and the ground truth $\boldsymbol{y}$, which is specified as

$$ACC = max_h \frac{\sum_{i=1}^{N} 1\{y = h(z_i)\}}{N} \quad (9)$$

- **Normalized Mutual Index(NMI)**: It evaluates the mutual information I(y,z) between the ground truth labels y and the cluster assignments z, then normalises it using the average entropy of both ground labels H(y)

and cluster assignments H(z), which is defined as

$$NMI = \frac{I(y,z)}{\frac{1}{2}[H(y) + H(z)]} \tag{10}$$

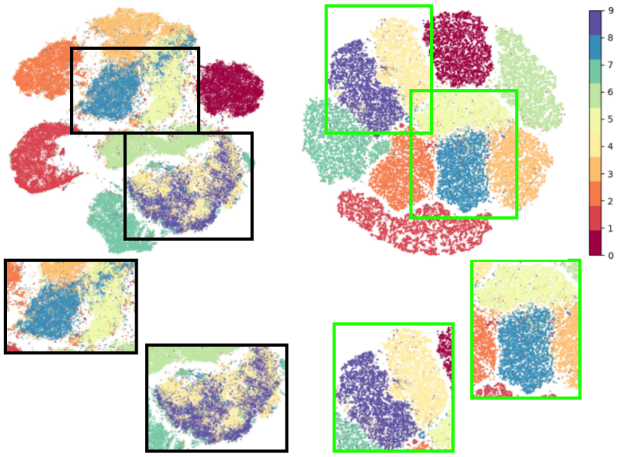

Figure 4. We visualize the latent space z of autoencoder and DA-AE on MNIST dataset. 1^{st} columns from left represents latent space of AE. 2^{nd} columns represents latent space of DA-AE. The highlighted region in GREEN illustrates the influence of the DA-AE encoder, which has the maximum interclass variance when compared to the one highlighted in **BLACK** from the autoencoder. The highligthed region depicts the impact of DA-AE encoder, which is able to distinguish classes with a large margin, as compared to AE, with minimal interclass variations.

3.4. Results

In this section, we analyse the inference of our experiments. The statistical analysis of our experiments and ablation are summarized in Table 3.

We demonstrate the effect of Jacobian and nonlinear ICA on DA-AE latent space as shown in Figure 4 and Figure 5.

We can infer that from Figure 4 (TSNE plots) that unlike autoencoder, DA-AE has superior ICA and representation of four classes (9, 8, 5, and 4) of MNIST datatset have high intra class variance. There are modest differences in the latent representation of DA-AE and autoencoder in two dimensions space, as shown in Figure 5 (PCA plots) , but there are significant differences in higher dimensions (10 dimension) space, as demonstrated through clustering accuracy in Table 3.

We demonstrate clustering accuracy of DA-AE, and compare with state-of-the-art methods on benchmark dataset as shown in Table 3. We can infer that our methodology DA-AE's clustering accuracy has outperformed on all the benchmark datasets without clustering loss. We can also infer that our proposed DA-AE with IDEC (with clustering loss) as a plugin has outperformed on FashionMNIST dataset.

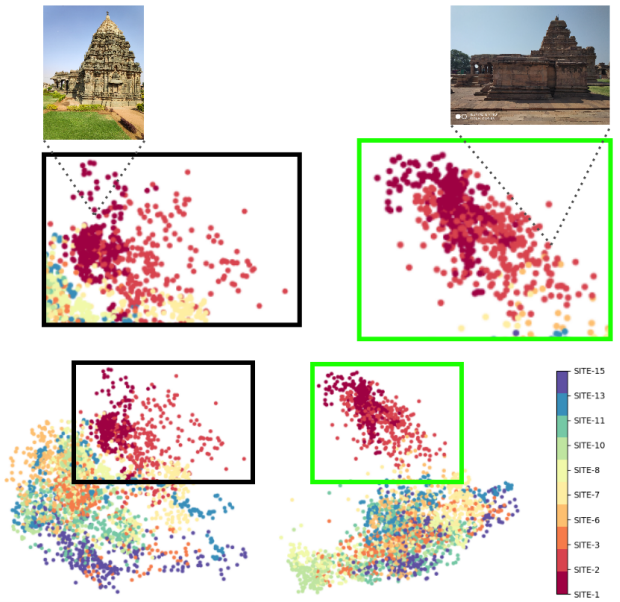

Figure 5. We visualize the latent space z of autoencoder and DA-AE on IDH10 dataset. 1^{st} columns represents latent space of AE. 2^{nd} columns represents latent space of DA-AE. 1^{st} row represents PCA analysis of the latent space. 2^{nd} row represents TSNE plots of the latent space. Highlighted region in GREEN shows the effect DA-AE encoder who's manifold exhibits maximum interclass variance compared to the one which is highlighted in **BLACK** from autoencoder.

We demonstrate the effects of categorization on 3D reconstruction using MVG-MVS pipeline [19]. We compare the 3D models generated before and after categorization of images through our proposed method DA-AE as shown in Figure 1. We infer that, the reconstruction pipeline is sensitive to selection of initial pair for reconstruction before categorization of data, as shown by 3D reconstruction in Figure 1. Our proposed framework is robust towards the sensitivity of initial pair selection as demonstrated in Figure 1.

Using the Pytorch framework, the experiments are carried out on an NVIDIA RTX 3090 GPU with 24GB RAM and an AMD RYZEN threadripper 3970x CPU.

4. Conclusions

In this paper, we have proposed DA-AE: Disparity Alleviation AutoEncoder for categorization of heritage images in order to facilitate 3D reconstruction. We have proposed disparity alleviation loss $\mathcal{L}_{\mathbb{DA}}$ that facilitates DA Encoder for better representation and categorization of data in latent space. We have demonstrated the event of categorization with clustering as a downstream task. We have demonstrated the impact of DA-AE with disparity alleviation loss and compared it with state-of-the-art methods (deep clustering) on benchmark datasets (MNIST, FashionMNIST, and

Table 3. The clustering performance of proposed methodology quantitatively using standard benchmark dataset in comparison with state of the art methods. Highest is represented in **Bold** and second highesr is represented in Underlined

	Methods	MNIST		FMNIST		USPS		IDH10	
		ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Without $\mathcal{L}_{cl}$	**SC** [25] (2002)	0.656	0.656	0.508	0.575	0.649	0.794	NA	NA
	LSNMF [18] (2007)	0.540	0.455	0.549	0.523	0.575	0.551	NA	NA
	GMM [22] (2008)	0.433	0.366	0.556	0.557	0.551	0.530	-	-
	KMeans [13] (2010)	0.532	0.500	0.474	0.512	0.668	0.627	0.378	0.390
	AC [11] (2010)	0.621	0.682	0.500	0.564	0.683	0.725	NA	NA
	SSC-OMP [36] (2016)	0.309	0.315	0.100	0.007	0.447	0.503	NA	NA
	EnSC [35] (2016)	0.111	0.014	0.629	0.636	0.610	0.684	NA	NA
	AE+KMeans	0.807	0.730	0.582	0.614	0.720	0.698	0.422	0.439
	DA-AE + KMeans (Ours)	**0.904**	**0.805**	**0.668**	**0.647**	**0.776**	**0.782**	**0.605**	**0.551**
With $\mathcal{L}_{cl}$	**DEC** [30] (2016)	0.843	0.797	0.518	0.546	0.762	0.767	NA	NA
	DCN [31] (2016)	0.830	0.810	0.501	0.558	0.688	0.683	NA	NA
	JULE [33] (2016)	0.964	0.931	0.563	0.608	0.950	0.913	NA	NA
	IDEC [7] (2017)	0.88	0.867	0.529	0.557	0.761	0.785	NA	NA
	VaDE [12] (2017)	0.945	0.876	0.578	0.630	0.566	0.512	NA	NA
	DEPICT [3] (2017)	0.965	0.917	0.392	0.392	0.899	0.906	NA	NA
	Best of DEC-DA [8] (2018)	**0.986**	**0.962**	0.580	0.650	**0.987**	0.967	-	-
	SR_KMeans [10] (2019)	0.939	0.866	0.507	0.548	0.936	**0.974**	NA	NA
	DA-AE + IDEC (Ours)	0.958	0.902	**0.678**	**0.673**	0.788	0.816	**0.657**	**0.623**
With $\mathcal{P}_l$	**DeepCluster** [2] (2018)	0.797	0.661	0.542	0.510	0.562	0.540	NA	NA
	DynAE [20] (2020)	0.987	0.964	0.591	0.642	0.981	0.948	NA	NA
	DAC [34] (2020)	0.935	0.945	0.678	0.674	-	-	NA	NA

USPS). We have achieved 10% increase on MNIST dataset, 18.3% increase on IDH10 dataset in comparison with autoencoder using DA-AE and 9.8% increase on Fashion-MNIST in comparison with DEC-DA using DA-AE with IDEC as a plugin. We have also shown, how representation using DA-AE aids categorization and helps the MVG-MVS pipeline overcome its absurdity when used in wild.

5. Acknowledgement

This work is partly carried out under the Department of Science and Technology (DST) through the ICPS programme - Indian Heritage in Digital Space for the project "Crowd-Sourcing" (DST/ ICPS/ IHDS/ 2018 (General))

References

[1] Deep features for categorization of heritage images towards 3d reconstruction. *Procedia Computer Science*, 171:483–490, 2020. Third International Conference on Computing and Network Communications (CoCoNet'19). 2

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. 07 2018. 2, 4, 7

[3] Kamran Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. pages 5747–5756, 10 2017. 2, 4, 7

[4] Charles Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, pages 147–153, 2003. 3

[5] Peter Forrest. The Identity of Indiscernibles. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020. 3

[6] Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, and Aapo Hyvarinen. Relative gradient optimization of the jacobian term in unsupervised deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16567–16578. Curran Associates, Inc., 2020. 3

[7] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. IJCAI'17, page 1753–1759. AAAI Press, 2017. 2, 4, 5, 7

[8] Xifeng Guo, En Zhu, Xinwang Liu, and Jianping Yin. Deep embedded clustering with data augmentation. In Jun Zhu

and Ichiro Takeuchi, editors, *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pages 550–565. PMLR, 14–16 Nov 2018. 7

[9] J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. 5

[10] Mohammed Jabi, Marco Pedersoli, Amar Mitiche, and Ismail Ben Ayed. Deep clustering: On the link between discriminative models and k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 12 2019. 2, 4, 7

[11] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR). 4, 7

[12] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. pages 1965–1972, 08 2017. 2, 4, 7

[13] Xin Jin and Jiawei Han. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA, 2010. 4, 7

[14] James M. Joyce. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. 4

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 3

[16] Shashidhar Veerappa Kudari, Akshaykumar Gunari, Adarsh Jamadandi, Ramesh Ashok Tabib, and Uma Mudenagudi. Augmented data as an auxiliary plug-in towards categorization of crowdsourced heritage data. *CoRR*, abs/2107.03852, 2021. 2

[17] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. 5

[18] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19:2756–2779, 2007. 2, 4, 7

[19] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. pages 3248–3255, Dec 2013. 6

[20] Nairouz Mrabah, Naimul Khan, Riadh Ksantini, and Zied Lachiri. Deep clustering with a dynamic autoencoder: From reconstruction towards centroids construction. *Neural Networks*, 130, 07 2020. 2, 4, 7

[21] Mykola Ponomarenko, Karen Egiazarian, Vladimir Lukin, and Victoriya Abramova. Structural similarity index with predictability of image blocks. In *2018 IEEE 17th International Conference on Mathematical Methods in Electromagnetic Theory (MMET)*, pages 115–118, 2018. 4

[22] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, 01 2008. 4, 7

[23] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. 01 2011. 3

[24] B. H Shekar and Guesh Dagnew. Grid search-based hyperparameter tuning and classification of microarray cancer data. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–8, 2019. 5

[25] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 05 2002. 4, 7

[26] Chunfeng Song, Feng Liu, Yongzhen Huang, Liang Wang, and Tieniu Tan. Auto-encoder based data clustering. 2, 3, 4

[27] Ramesh Ashok Tabib, Sujaykumar Kulkarni, Abhay Kagalkar, Vaishnavi Hurakadli, Abhijeet Ganapule, Rohan Raju Dhanakshirur, and Uma Mudenagudi. *Deep Learning-Based Filtering of Images for 3D Reconstruction of Heritage Sites*, pages 147–156. Springer International Publishing, Cham, 2021. 2

[28] Ramesh Ashok Tabib, T. Santoshkumar, Varad Pradhu, Ujwala Patil, and Uma Mudenagudi. *Categorization and Selection of Crowdsourced Images Towards 3D Reconstruction of Heritage Sites*, pages 133–146. Springer International Publishing, Cham, 2021. 2

[29] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 5

[30] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. 11 2015. 2, 4, 5, 7

[31] Bo Yang, Xiao Fu, N.D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. 2016. 2, 4, 7

[32] Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering, 2017. 4

[33] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. pages 5147–5156, 06 2016. 2, 4, 7

[34] Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, and Wei Liu. Adversarial learning for robust deep clustering. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9098–9108. Curran Associates, Inc., 2020. 7

[35] Chong You, Chun-Guang Li, Daniel P. Robinson, and René Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. *CoRR*, abs/1605.02633, 2016. 2, 4, 7

[36] Chong You, Daniel P. Robinson, and Rene Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit, 2016. 4, 7

[37] Changfan Zhang, Xiang Cheng, Jianhua Liu, Jing He, and Guangwei Liu. Deep sparse autoencoder for feature extraction and diagnosis of locomotive adhesion status. *Journal of Control Science and Engineering*, 2018:1–9, 07 2018. 3