Semi-Supervised Contrastive VAE for Disentanglement of Digital Pathology Images

Mahmudul Hasan*[1], Xiaoling Hu[2], Shahira Abousamra[1], Prateek Prasanna[1], Joel Saltz[1], and Chao Chen[1]

[1] Stony Brook University, Stony Brook, NY, USA
[2] Harvard Medical School, Boston, MA, USA

Abstract. Despite the strong prediction power of deep learning models, their interpretability remains an important concern. Disentanglement models increase interpretability by decomposing the latent space into interpretable subspaces. In this paper, we propose the first disentanglement method for pathology images. We focus on the task of detecting tumor-infiltrating lymphocytes (TIL). We propose different ideas including cascading disentanglement, novel architecture, and reconstruction branches. We achieve superior performance on complex pathology images, thus improving the interpretability and even generalization power of TIL detection deep learning models. Our codes are available at https://github.com/Shauqi/SS-cVAE.

Keywords: Disentanglement · Contrastive VAE · Digital Pathology.

1 Introduction

Although deep learning models have strong predictive power, their interpretability remains a significant concern. This is especially important in the medical domain to ensure trust for clinicians and patients. Disentanglement models [11,12,7] increase interpretability through an encoder-decoder framework. By applying additional regulations such as total correlation to the latent space, one can obtain disentangled latent dimensions corresponding to different interpretable factors, which can be the grounding for training interpretable classifiers [20].

However, the power of these unsupervised disentanglement methods [11,12,7] are restricted to relatively simple images, e.g., images with a single object, and will fail to provide satisfying disentanglement for complex images with multiple objects. Recently, a new category of disentanglement methods called contrastive analysis has been proposed [1,21,6]. These methods focus on a discriminative setting. Instead of trying to disentangle the whole latent space, they decompose the latent space into the group of features differentiating different populations, and the group of features that is common across populations.

In this paper, we propose the first contrastive analysis method for disentangling complex digital pathology images. In particular, we focus on the task

* Email: mahhasan@cs.stonybrook.edu.

of detecting Tumor-Infiltrating Lymphocytes (TILs) [3]. TILs are essential in cancer diagnosis and prognosis. Clinical studies have shown that the density of TILs is correlated with overall survival and the length of disease-free survival in multiple cancer types [19,4,16]. Developing interpretable models will improve transparency and thus trustworthiness of AI methods, facilitating their deployment in clinical scenarios. Furthermore, as we will demonstrate empirically, these interpretable features also have better generalization power.

Contrastive analysis for TIL prediction is not trivial. Pathology images are highly diverse in tissue types and contain different types of cells, which exhibit a large variation of density and spatial configurations. As demonstrated in Tab. 1, the direct application of the existing contrastive analysis method does not work. To address the challenge, we propose a novel cascaded contrastive analysis method named *Semi-Supervised Contrastive VAE (SS-cVAE)*, consisting of the following technical contributions:

1. We introduce a cascaded disentangling approach that first differentiates background (tissue) patches and patches with cells and then differentiates patches with low TIL counts and patches with many TILs. This novel cascaded approach proves powerful in disentangling the complex cell spatial configurations in tumor microenvironment.
2. Furthermore, to handle the complex pathology images, we propose substantial architectural improvements from our baseline model MM-cVAE [21], including residual blocks in encoders and decoders, as well as three reconstruction branches: the salient reconstruction branch, background reconstruction branch, and classification branch – collectively enhancing the disentanglement performance.
3. To further improve reconstruction performance (thus the quality of the latent space), we introduce a novel ID-GAN [14] reconstruction module.

Empirical results demonstrate superior disentangling performance of our method, compared with SOTA approaches. We also show that the disentangled latent features provide strong generalization power across different datasets.

2 Related Works

Unsupervised disentangling methods. Existing unsupervised disentanglement approaches like β-VAE [11], Factor VAE [12], β-TCVAE [7] are designed to disentangle low-level concepts like position, shape, color, orientation, style, etc. These methods are mostly applied to simple image datasets like dSprites [17], MNIST [8], or face image datasets such as CelebA [15]. They work when the image only contains a single object/face. When applied to digital pathology images, these unsupervised disentangling VAEs are limited to disentangling factors like hue and saturation and fail to capture the complex spatial relationship between cells. Additionally, their reconstruction performance often results in significant blurriness, making it challenging to interpret the disentangled factors from these models. The unique challenge presented by digital pathology images in disentanglement tasks involves multiple concepts, including spatial and morphological

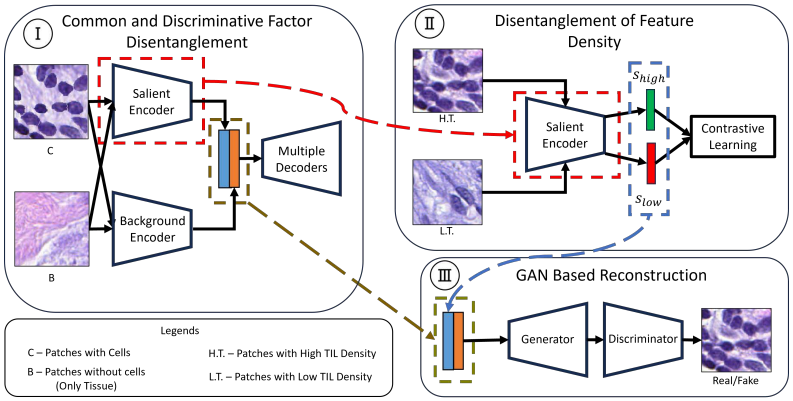

Fig. 1. Whole SS-cVAE Framework. I) Module to disentangle discriminative factors from common factors. II) Module for disentangling factors related to density. Note that, the same salient encoder from (I) is fine-tuned here. III) Module for image reconstruction from the disentangled latent generated from (I) and (II).

variations in diverse objects, as well as variations in tissue types, posing a much more complex scenario than what these methods can handle.

Contrastive analysis based disentangling methods. In contrast to unsupervised disentanglement methods, contrastive analysis approaches incorporate supervision using classifiers to decompose the latent space dimensions into two groups. One group contains latent features shared by different classes. Whereas the other group contains latent features differentiating different classes. Existing methods include Contrastive Analysis VAE (CA-VAE) methods (including cVAE [1], MM-cVAE [21]) and CA-GAN based methods (such as Double InfoGan [6]). These methods have demonstrated their ability to disentangle on datasets such as Gene Expression [10,22], CIFAR-10 [13], MNIST [8], and medical image datasets like BRATS [18]. However, they are primarily focusing on single-object images. When applied to complex pathology images containing multiple objects having complex spatial or topological relationships, the disentanglement and reconstruction performance is unsatisfactory.

3 Method

Our disentanglement framework, illustrated in Fig. 1, is comprised of three modules. In the first module (Fig. 1(I)), disentanglement occurs on synthetic patches containing cells, C, and real tissue-only patches, B. Utilizing a copy-paste mechanism in synthetic data creation, pairwise patches from both C and B are sent to two encoders: a salient encoder and a background encoder. These encoders separate patch embeddings into salient latent space (the discriminative factor, i.e. the cells) and background latent space (the common factor, i.e. the tissue).

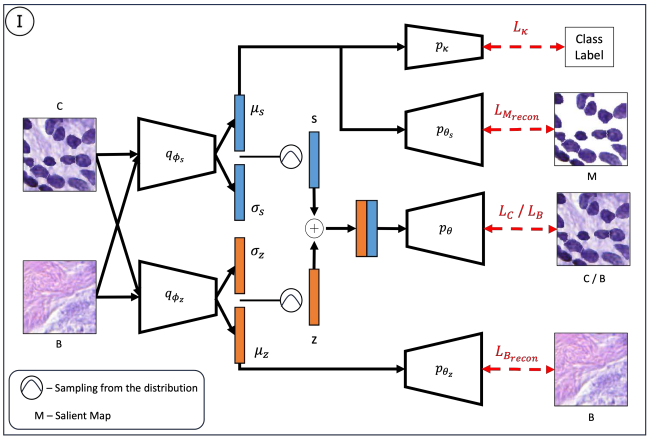

Fig. 2. The first module of SS-cVAE. It performs disentanglement of discriminative factor from common factor given any patch of C or B.

The use of pairwise synthetic images enhances disentanglement compared to unpaired real images, aiding the model in understanding the appearance with and without the discriminating factor.

In the second module (Fig. 1(II)), the salient encoder is fine-tuned to disentangle the density of a specific type of object - lymphocytes in our case. The rationale behind the two-stage disentanglement is to not only capture the existence of lymphocytes but also the variation in lymphocyte density. A two-stage disentanglement makes disentangling complex factors easier, so each module focuses on one pair of populations (with vs. without cells, low vs. high lymphocytes). Compressing the disentanglement into a single step could result in noisy latent factorization involving variations in other cell types and tissue characteristics. The two-stage disentanglement addresses this issue by eliminating tissue variation in the initial stage, leaving only other cell variations to be considered.

The third module is an image reconstruction model. Given the observed poor reconstruction performance of the VAE decoder, a GAN-based module is introduced (Fig. 1(III)). This module takes as input the latent embeddings generated by the salient and background encoders and attempts to reconstruct the original image. Further details about each module are provided in the subsequent sections.

Module 1: common and discriminative factor disentanglement. Fig. 2 illustrates the architecture of this disentanglement module for C and B. It comprises two probabilistic encoders, q_{ϕ_s} and q_{ϕ_z} approximating the salient likelihood $\mathcal{N}(\mu_s, \sigma_s)$ and background likelihood $\mathcal{N}(\mu_z, \sigma_z)$, respectively. Salient latent code, s, and background latent code, z, are sampled from these distributions. The network is trained with three reconstruction branches: image reconstructor, P_θ, salient reconstructor, P_{θ_s}, and background reconstructor, P_{θ_z}. We also have a classification branch, P_κ. P_θ reconstructs the original image from the concate-

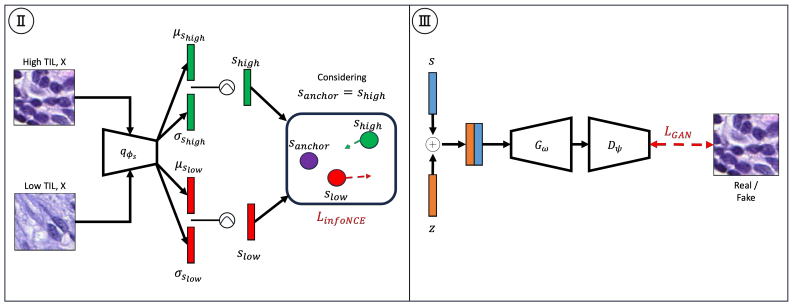

Fig. 3. 2nd and 3rd modules of SS-cVAE. II) Module to disentangle the factor corresponding to density. During InfoNCE loss calculation in (II) we consider one of the s_{high} samples as s_{anchor} III) GAN-based reconstruction using disentangled latent.

nation of sampled latent encodings s and z. P_{θ_s} reconstructs the salient map, M from μ_s. P_{θ_z} reconstructs the background image B from μ_z. The module is trained with the following objective function:

$$\mathcal{L}_c + \mathcal{L}_b + \mathcal{L}_{M_{recon}} + \mathcal{L}_{B_{recon}} + \mathcal{L}_\kappa \tag{1}$$

where $\mathcal{L}_c$ is the Evidence Lower Bound (ELBO) for C, $\mathcal{L}_b$ is the ELBO for B, $\mathcal{L}_{M_{recon}}$ is the loss for salient reconstruction, $\mathcal{L}_{B_{recon}}$ is the loss for background reconstruction, and $\mathcal{L}_\kappa$ is the binary cross entropy loss for classifying whether a patch contains cells or not. For the reconstructors, we use mean squared error (MSE) to calculate the losses. The equations for the loss functions are in Eq. 3-6 in the supplementary materials.

Module 2: disentanglement of feature density. This module focuses on creating an embedding that discerns varying densities of a cell-containing image, such as the presence level of a specific cell type in a patch. To achieve this, we fine-tune the previously trained salient encoder, resulting in a salient latent distribution $\mathcal{N}(\mu_s, \sigma_s)$ that captures feature density. During training, latent encodings s_{high} and s_{low} are sampled from patches with high and low feature densities, respectively. InfoNCE loss is applied to s_{high} and s_{low} to push the two distributions apart. Additionally, KL-Divergence loss is incorporated to fine-tune the distribution itself. Fig. 3(II) illustrates this module. The training objective function is as follows:

$$\mathcal{L}_{infoNCE} - \mathcal{L}_{KL_s} \tag{2}$$

The equations for $\mathcal{L}_{infoNCE}$ and $\mathcal{L}_{KL_s}$ are in Eq. 7-8 in supplementary material.

Module 3: GAN-based reconstruction. While VAEs serve as effective latent distribution encoders, their image generation quality can be suboptimal. To enhance image reconstruction, we employ a GAN-based model, specifically IDGAN [14]. Training IDGAN involves reconstructing images from the concatenation of disentangled salient and background embeddings, yielding superior reconstruction compared to the VAE decoder. Fig. 3(III) presents the concep-

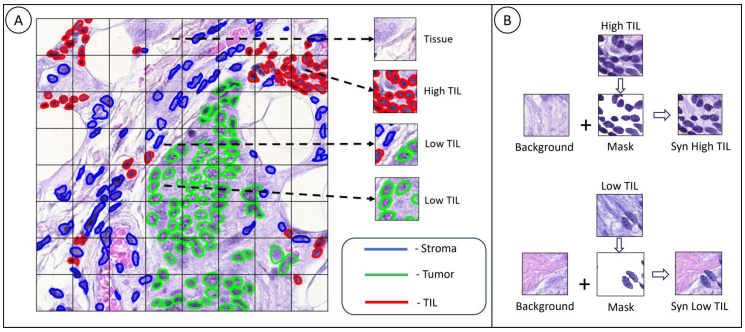

Fig. 4. Data Preprocessing. A) Patch extraction and patch labeling using ground truth cell annotation. B) Synthetic Data Creation using a copy-paste mechanism.

tual diagram of IDGAN. The loss function for IDGAN, $\mathcal{L}_{GAN}(D, G)$, is detailed in Eq. 9 in the supplementary materials.

4 Experiments and Results

Datasets and preprocessing. For the disentanglement task, synthetic data is generated using the TCGA BRCA dataset [2] and the CoNSeP dataset [9]. Both datasets consist of 1000×1000 pixel tiles in native magnification close to $40\times$. The data statistics are in Tab. 4 of the supplementary materials. Fig. 4(A) illustrates the data preprocessing, including the extraction and labeling of training patches. Tiles are resized to 1024×1024 images, and non-overlapping 128×128 patches are extracted. Both datasets contain ground truth cell annotations with their corresponding class. Using these annotations we assign patch-level labels based on the frequency of cells. For the first-stage disentanglement, patches are divided into two groups: *patches with cells* (C) and *patches without cells* (B). The cell-to-patch area ratio determines the grouping, with patches having a ratio greater than 10% assigned to C; the remaining patches are grouped into B. This partition is balanced (1:1) across training, validation, and test sets (Tab. 4, Supp.). For the second-stage disentanglement, patches are labeled as *low TIL density* ($L.T.$) or *high TIL density* ($H.T.$) based on TIL instances. Patches with at least 5 TIL instances are categorized as $H.T.$, while others are labeled as $L.T.$ (Tab. 4, Supp.). The C and B disentanglement model is trained on a synthetic dataset created using a copy-paste mechanism. A random cell mask from C patches is copied and pasted onto random B patches (Fig. 4(B)). For downstream tasks, three classes are considered: patches with only tissue, patches with only TILs, and patches with other cells (i.e., stroma, tumor). Statistics for the downstream task data are detailed in Tab. 4 (Supp.).

Implementation details. Our model employs residual blocks in encoders and transposed residual blocks in decoders. A batch size of 64 is consistently used for training, validation, and test set loading across all models, including baselines.

Table 1. Disentanglement Evaluation. The models are trained on synthetic datasets and evaluated against real datasets. B = Patches without cells, C = Patches with cells, H.T. = High TIL, L.T. = Low TIL, SS = Silhouette Score, S = Salient Latent Space, Z = Background Latent Space, CLF = Classification Score

Model	B vs C		H.T. vs L.T.	B vs C		H.T. vs L.T.
	$SS_S \uparrow$	$SS_Z \downarrow$	$SS_S \uparrow$	$CLF_S \uparrow$	$CLF_Z \downarrow$	$CLF_S \uparrow$
Dataset: BRCA						
β-VAE	0.05	0.01	-	0.64±0.07	0.49±0.08	-
Factor-VAE	0.06	0.07	-	0.61±0.06	0.60±0.08	-
β-TCVAE	0.13	0.08	-	0.70±0.06	0.62±0.10	-
Double InfoGAN	0.04	0.01	0.005	0.83±0.07	0.70±0.06	0.65±0.05
cVAE	0.10	0.20	0.10	0.89±0.03	0.92±0.06	0.76±0.02
MM-cVAE	0.14	0.13	0.06	0.86±0.04	0.92±0.03	0.75±0.04
SS-cVAE	0.32	0.10	0.27	0.92±0.04	0.92±0.02	0.87±0.05
Dataset: CoNSeP						
β-VAE	0.08	0.007	-	0.65±0.09	0.53±0.06	-
Factor-VAE	0.009	0.007	-	0.57±0.08	0.53±0.02	-
β-TCVAE	0.10	0.006	-	0.66±0.04	0.57±0.09	-
Double InfoGAN	0.006	0.009	0.005	0.64±0.07	0.55±0.06	0.57±0.09
cVAE	0.16	0.12	0.04	0.66±0.10	0.62±0.08	0.67±0.13
MM-cVAE	0.06	0.03	0.11	0.64±0.12	0.64±0.06	0.69±0.06
SS-cVAE	0.18	0.04	0.13	0.85±0.02	0.75±0.07	0.78±0.09

The Adam optimizer is employed with a learning rate of 0.001. Hyperparameters in Eq. 3 and Eq. 4 (refer to Supplementary Materials) are configured as $\lambda_1 = 10$ and $\lambda_2 = 10$. For both $\lambda_1 = 10$ and $\lambda_2 = 10$, we tested different values ranging from 10^{-3} to 10^3 with steps of multiples of 10 on the BRCA dataset and selected the one with the best validation performance. The sizes of the salient and background latent vectors are set to $s = 128$ and $z = 128$.

Evaluation metrics. For evaluation, the silhouette score (SS) is employed, as described in [6] and [21], on a real test set. Additionally, the mean of the latent is computed, augmented with a simple linear classifier, and subjected to a 5-fold cross-validation binary classification on the test set. Reconstruction performance is assessed using Frechet Inception Distance (FID). Lastly, for downstream task evaluation, multiclass classification scores are utilized.

Quantitative evaluations. Tab. 1 presents the disentanglement evaluation of our model against SOTA baselines. A higher value for the salient latent indicates superior encoding of discriminative factors, while a lower value for the background latent suggests an effective representation of a common factor without discriminative power. Our model excels in salient disentanglement, outperforming baselines, and demonstrates comparable performance in background latent disentanglement. For High TIL and Low TIL density, our model surpasses the baselines, too. Reconstruction evaluation against contrastive analysis (CA) baselines is detailed in Tab. 2. The results indicate that our model's reconstruction performance is superior to baselines, except for Double InfoGan. Importantly,

Table 2. Reconstruction Evaluation and Downstream Task Evaluation

Reconstruction		Downstream 3 Class Classification on CoNSeP dataset	
Model	FID	Model	Accuracy
MM-cVAE	293.71	ResNet-18 (without pre-train)	0.71
Double InfoGAN	214.59	ResNet-18 (BRCA pretrain)	0.76
SS-cVAE	221.15	SS-cVAE (BRCA pretrain)	0.79

our model achieves better disentanglement than Double InfoGan. Finally, utilizing our model as a pre-trained feature extractor on the BRCA dataset and finetuning on the CoNSeP dataset for multiclass classification as a downstream task, we compare our model with ResNet18 in Tab. 2. For the downstream task, we have used 3 classes: patches with only tissue, patches with only TILs, and patches with other cells (i.e., stroma, tumor). The findings suggest that our model can outperform the ResNet18 model in multiclass classification within the transfer learning framework. Apart from the abovementioned experiments we also present ablation experiments with different architectural changes and reconstructions in Tab. 3 (Supp.).

Qualitative evaluations. To visually depict what is encapsulated in the salient and background latent spaces, we conduct latent swapping between two samples. Fig. 5 exhibits salient latent swaps between patches from C and B (rows 1 and 2) and between $H.T.$ and $L.T$ patches (rows 3 and 4). The results demonstrate that our model's generations accurately capture the content of the original image, encompassing both content and cell/TIL density. In contrast, other baselines exhibit complete failure in this regard. This emphasizes the superior quality of the embeddings generated by our proposed SS-cVAE. Note that we omit results for β-VAE, Factor VAE, and β-TCVAE as they produce highly blurry and non-interpretable reconstructions. Additionally, we explore latent space interpolation through Fig. 6 in the supplementary materials. This demonstrates synthetic image generation via latent space traversal/interpolation between tissue/background images and cell images, as well as between low and high TIL images. This experimentation showcases our ability to synthesize new images while controlling discriminating factors, common factors, and density factors.

5 Conclusions

In this paper, we introduced a novel framework for disentangling concepts, including discriminative, common, and density factors, within distinct populations of digital pathology samples. Our proposed framework demonstrated notable enhancements, showcasing substantial qualitative and quantitative improvements over existing approaches in disentanglement tasks.

Acknowledgement. This research was partially supported by the National Science Foundation (NSF) grant CCF-2144901, the National Institute of General Medical Sciences (NIGMS) grant R01GM148970, the National Institute of

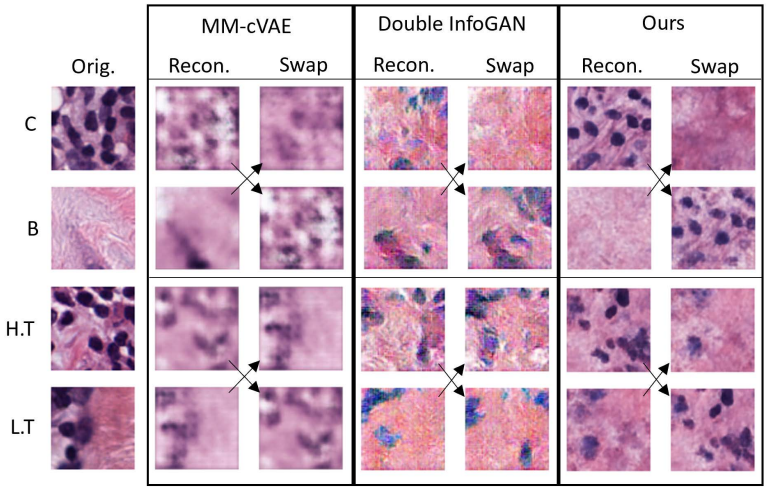

Fig. 5. Example of Latent Swap. Row 1 and 2 represent salient latent swapping between the samples of C and B. Row 3 and 4 represent the salient latent swapping between the samples of $H.T$ and $L.T$.

Health (NIH) grant 5R21CA258493-02, the National Cancer Institute (NCI) grant UH3-CA225021 and the Stony Brook Trustees Faculty Award. This work used Bridges-2 [5] at Pittsburgh Supercomputing Center through allocation [asc13007p] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. Co-Author Joel Saltz is a co-founder of Chilean Wool LLC.

References

1. Abid, A., Zou, J.: Contrastive variational autoencoder enhances salient features. arXiv preprint arXiv:1902.04601 (2019)
2. Abousamra, S., Belinsky, D., Van Arnam, J., Allard, F., Yee, E., Gupta, R., Kurc, T., Samaras, D., Saltz, J., Chen, C.: Multi-class cell detection using spatial context representation. In: ICCV (2021)
3. Abousamra, S., Gupta, R., Hou, L., Batiste, R., Zhao, T., Shankar, A., Rao, A., Chen, C., Samaras, D., Kurc, T., Saltz, J.: Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. In: Frontiers in oncology. p. 5971 (2022), https://www.frontiersin.org/articles/10.3389/fonc.2021.806603/full
4. Angell, H., Galon, J.: From the immune contexture to the immunoscore: the role of prognostic and predictive immune markers in cancer. Current opinion in immunology **25**(2), 261–267 (2013)

5. Brown, S.T., Buitrago, P., Hanna, E., Sanielevici, S., Scibek, R., Nystrom, N.A.: Bridges-2: A platform for rapidly-evolving and data intensive research. In: Practice and Experience in Advanced Research Computing (2021)
6. Carton, F., Louiset, R., Gori, P.: Double infogan for contrastive analysis. arXiv preprint arXiv:2401.17776 (2024)
7. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: NeurIPS (2018)
8. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine (2012)
9. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis (2019)
10. Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al.: A single-cell survey of the small intestinal epithelium. Nature (2017)
11. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2016)
12. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)
13. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
14. Lee, W., Kim, D., Hong, S., Lee, H.: High-fidelity synthesis with disentangled representation. In: ECCV (2020)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
16. Loi, S., Drubay, D., Adams, S., Pruneri, G., Francis, P., Lacroix-Triki, M., Joensuu, H., Dieci, M., Badve, S., Demaria, S., Gray, R., Munzone, E., Lemonnier, J., Sotiriou, C., Piccart, M., Kellokumpu-Lehtinen, P.L., Vingiani, A., Gray, K., Andre, F., Michiels, S.: Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers. Journal of Clinical Oncology **37**, JCO.18.01010 (01 2019). https://doi.org/10.1200/JCO.18.01010
17. Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging (2014)
19. Mlecnik, B., Bindea, G., Pagès, F., Galon, J.: Tumor immunosurveillance in human cancers. Cancer and Metastasis Reviews **30**(1), 5–12 (2011)
20. Soelistyo, C.J., Lowe, A.R.: Discovering interpretable models of scientific image data with deep learning. arXiv preprint arXiv:2402.03115 (2024)
21. Weinberger, E., Beebe-Wang, N., Lee, S.I.: Moment matching deep contrastive latent variable models. arXiv preprint arXiv:2202.10560 (2022)
22. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al.: Massively parallel digital transcriptional profiling of single cells. Nature communications (2017)