Analysis of cellular phenotypes with unbiased image-based generative models

Ruben D. Fonnegra
Institución Universitaria Pascual Bravo
Medellín, Colombia

Mohammad Vali Sanian
University of Helsinki
Helsinki, Finland

Zitong S. Chen
Broad Institute
Cambridge, MA

Lassi Paavolainen
University of Helsinki
Helsinki, Finland

Juan C. Caicedo
Univ. of Wisconsin-Madison
Madison, WI

Abstract

Observing changes in cellular phenotypes under experimental interventions is a powerful approach for studying biology and has many applications, including treatment design. Unfortunately, not all interventions can be tested experimentally, which limits our ability to study complex phenomena such as combinatorial treatments or continuous time or dose responses. In this work, we explore unbiased, image-based generative models to analyze phenotypic changes in cell morphology and tissue organization. The proposed approach is based on generative adversarial networks (GAN) conditioned on feature representations obtained with self-supervised learning. Our goal is to ensure that image-based phenotypes are accurately encoded in a latent space that can be later manipulated and used for generating images of novel phenotypic variations. We present an evaluation of our approach for phenotype analysis in a drug screen and a cancer tissue dataset.

1 Introduction

The study of cellular biology with microscopy images is widespread for observing phenotypes and investigating their responses to perturbations [1, 2]. With the advent of automated imaging systems and high-throughput platforms, it is now possible to create very large imaging datasets that scan a large space of biological variation [3]. Computational methodologies and machine learning are increasingly used to quantify and profile biological events in such large image collections [4–6]. However, despite increased capacity for data production, the full space of biological variation is too large to explore experimentally. For instance, millions of compounds need to be tested in thousands of diseases with specialized imaging to evaluate their individual ability for affecting cellular phenotypes.

Here, we ask the question: can generative AI augment the experimental analysis of image-based biological research? Generative models have shown remarkable success in computer vision and natural language processing, resulting in methods that can successfully generate natural images given text prompts [7, 8]. However, cellular images are rarely annotated with rich language descriptions, limiting the use of such approaches for biological analysis. In principle, unconditional generative models can be used for capturing the visual variation in an image collection and to explore new realistic images [9–11]. Unfortunately, generating realistic images alone is not informative for biological studies, and instead, the ability to control phenotypic variation is necessary for advancing discovery projects. Recently, weakly supervised generative models have been investigated to study cellular variability [12] and fluorescent channel prediction [13], by conditioning the generative process on known experimental interventions.

NeurIPS 2023 Generative AI and Biology Workshop.

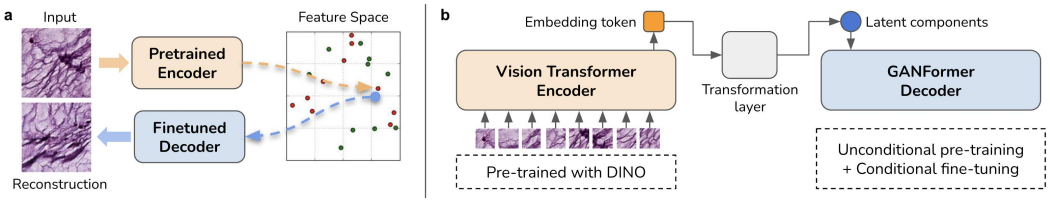

Figure 1: Illustration of the Decoupled Self-supervised Autoencoder. a) Input images are mapped to a feature space using a pre-trained encoder. Features are decoded back to reconstructed images with a finetuned decoder. b) Overview of the network architectures, trained independently and connected through a transformation layer.

In this paper, we approach the problem of unbiased image-based phenotype generation with a conditional model. Our model is entirely image-based requiring no textual or categorical annotations to fully leverage the wealth of information contained in biological images alone. We focus on unbiased, image-based models for three reasons: 1) Image-based cellular profiling has shown remarkable ability to reveal phenotypic variation in many perturbation studies [14–18], even with classical features [19, 20]. This highlights the importance of prioritizing images as a powerful and independent probe of biological activity. 2) Although experimental annotations are usually available for weakly supervised learning [21], these annotations may contain biases [16, 22] or may miss unique variation, such as cellular heterogeneity. 3) Advances in self-supervised learning in computer vision have made tremendous progress and manual annotations, captions or similar descriptions are no longer needed to achieve state-of-the-art performance [23, 24]. This is a unique opportunity to realize the full potential of images as a source of phenotypic information in biological experiments.

Our main hypothesis is that images of phenotypes not observed experimentally can be generated by manipulating observed phenotypes in a sufficiently rich feature space. To this end, we propose an approach where representation learning is decoupled from the generative modeling of image data. We first learn a feature representation in a self-supervised manner to obtain rich image features of a given phenotype. We show that the resulting feature space preserves accurate phenotypic information without using any type of semantic supervision. Next, we train a generative model to decode image features from this rich feature space, ensuring that images are reconstructed according to the true visual structures. Our approach resembles auto-encoder systems where encoder and decoder networks are trained together to learn image features and image reconstruction. Our method differs from auto-encoders in two key ways: first, we train the encoder and decoder separately, with loss functions specialized for each task following recent advances in representation learning [25, 26] and image generation [27, 28]. This brings together the best of both worlds in a new type of auto-encoder that we call **D**ecoupled **S**elf-supervised **A**uto**E**ncoder (DSAE). Our experiments show that this makes our approach more reliable for generative analysis of cellular phenotypes.

2 Approach

To efficiently analyze image-based phenotypes in an unbiased way, we aim to capture high-quality image features that could be used in many downstream tasks. Having a unified representation for image analysis facilitates consistency and interpretation across tasks, whereas training independent models specialized for each task may result in conflicting or misaligned outcomes. In addition, unbiased analysis of cellular phenotypes means that we do not constraint our models with prior knowledge about the experiments, but instead we let the image data explain the phenotypic variation on its own. Therefore, we consider the generative modeling problem as an additional downstream task that can be conditioned on unbiased image features. In that way, image-based phenotypes can be interpreted by exploring their natural transitions between perturbation states [12], or can be manipulated to estimate cellular responses in unobserved states.

To this end, we build an encoder-decoder network that learns representations and generates image samples (1). We independently train the encoder without category labels or manual annotations using self-supervised learning, which produces semantic discriminative embeddings [29]. The decoder is trained separately using unconditional generative modeling to capture the visual distribution of the image collection. Once both networks produce satisfactory results, we connect them with the goal of reconstructing images given their feature representations.

Decoupled Self-supervised Auto-Encoder (DSAE). The encoder architecture is a Vision Transformer (ViT) network [30]. To train the encoder, we follow the self-**DI**stillation with **NO** labels (DINO) [29] approach to self-supervision, which trains a student and a teacher network with the same architecture but different parameters. With a fixed teacher, the student network is trained with gradient descent to match the teacher outputs after both are presented different augmented views of the same image. The parameters of the teacher network are calculated as an exponential moving average of the student network. In our experiments, we train a ViT small 16×16 transformer network as our encoder backbone, which has been shown to exhibit excellent performance in multiple classification benchmarks [31, 32, 24]. Our decoder follows the GANformer architecture [33], which is composed of bipartite transformer layers where attention is computed between two sets of elements. The bipartite transformer computes attention between a set of input elements and a set of latent vectors, in such a way that the computational cost is reduced by controlling the number of latents. The bipartite transformer allows to compute two types of attention, simplex and duplex attention, which propagates information in one or both directions of the interacting elements. We adopt the duplex attention mechanism in our experiments. Many of these elements make the GANformer network an excellent decoder under the generative adversarial training regime, including memory efficiency for high-definition synthesis, and faster convergence.

Assembling the Encoder-Decoder Both the encoder and the decoder are first trained separately on the same image dataset, which enables independent control over their optimization choices and hyperparameters. Furthermore, our decoupled strategy can build an efficient autoencoder by reusing the weights of encoder and decoder networks that are pre-trained. Consider the encoder E and the decoder G, both pre-trained on an image collection $\mathbf{X}$ separately. We replace the input of G from a random-noise latent vector to the output of E. Specifically, let $z \in \mathbb{R}^n$ be the feature embedding (CLS token) of an image x obtained with the ViT encoder $z = E(x)$. Then, the decoder receives transformed feature embeddings $G(\tau(z))$, where τ is a linear transformation $\tau : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that maps z to the corresponding input space of G. Then, we implement an additional loss term over the feature embeddings of real and fake images that resembles a perceptual loss [34] under a cycle-consistency constraint [35] to finetune G. Images generated by the decoder $\hat{x} = G(z)$ are processed by the fixed encoder network $\hat{z} = E(\hat{x})$, which are expected to match the features of the original image x using $L1$ norm as follows:

$$\begin{aligned} \mathcal{L}_{Fcc} &= \mathbb{E}_x \left[||z - \hat{z}||_1 \right] \\ &= \mathbb{E}_x \left[||E(x) - E(G(E(x)))||_1 \right] \end{aligned} \tag{1}$$

3 Experiments

Datasets We used two bioimage datasets in our evaluation as follows: A) *Tissue dataset*: known as NCT-CRC-HE-100K dataset [36], it includes 100,000 RGB image patches from colorectal cancer samples in human tissues. We use the color-normalized image patches at 224x224 pixels, which are organized in nine classes. For validation, the CRC-VAL-HE-7K subset that includes 7,180 image patches is used. B) *Cellular dataset*: the BBBC021 [37, 38] dataset was used to study the effect of chemical compounds over various doses on human breast cancer cells. We used the subset with 103 compound-concentration pairs for the classification of 12 mechanism-of-action categories.

Baselines For evaluation, we consider a diverse set of autoencoder models: 1) *Variational Autoencoder* (VAE) [39], widely used model for images and other data types. 2) *Bidirectional GAN* (BiGAN) [40] learns the reverse mapping from pixel space to the latent space as feature extractor. 3) *Style Adversarial Latent Autoencoder* (StyleALAE) [41] improves representation learning and image generation simultaneously. 4) *Dual Contradistinctive VAE* (DC-VAE) [42] combines contrastive and discriminative loss terms to improve features and image generation.

Architectures We implemented and evaluated our approach using three network architectures to understand their impact on performance: 1) *CNN*: The convolutional architecture of StyleGANv2 [43], with a skip-type generator and style blocks for the decoder. 2) *ResNet*: This configuration only replaces the encoder from the CNN above and uses a regular ResNet50 architecture instead. 3) *ViT*: The encoder is a ViT-small network with input patches of 16×16 pixels. The transformer architecture of the decoder follows StyleGANv2 with additional bipartite attention modules in all layers. Image transformations can have a strong impact on learned representations [44, 45], so we

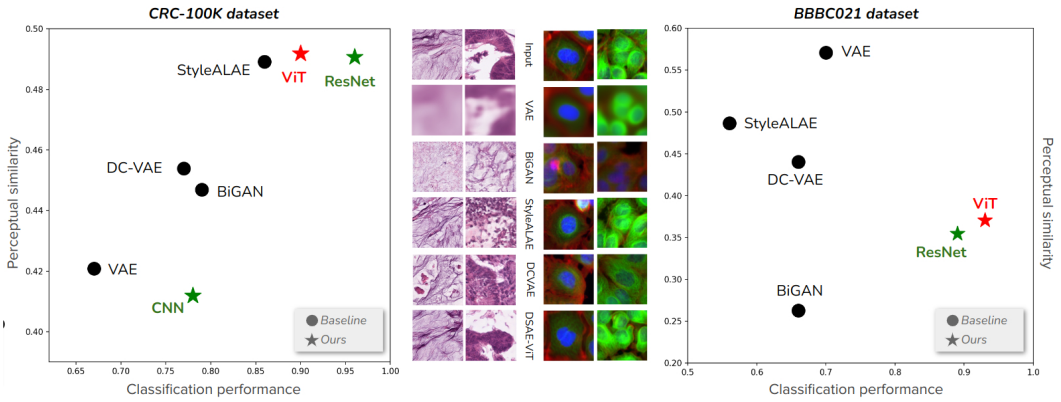

Figure 2: Performace comparison among models in two datasets: tissues (left) and cells (right). Horizontal axis: quality of representations (F1-score). Vertical axis: quality of image reconstructions (perceptual distance). Middle panel: example reconstructions for both datasets.

adjusted them for each dataset after empirical evaluations. In the tissue images case, we followed traditional RGB image augmentations to train the encoder, including random flip, color jittering, random grayscale, gaussian blurring, solarization and normalization. In the cellular images case, we implemented channel-indpendent randomized brightness and contrast changes, as well as random flips and rotations.

4 Representation and Image Synthesis Evaluation

Unbiased feature representations should capture the semantic structure of images and all their important discriminative details to support phenotyping. Thus, we conducted a quantitative evaluation of the representations in both datasets to verify whether the encoder has sufficient ability to capture biologically relevant features of cellular and tissue morphology. Given that our decoder is conditioned on image features, we want to ensure that the conditional variable is indeed meaningful for biological analysis. We follow standard practice for evaluating representations and implement linear (Tissues) and 1-NN (cells) classifiers to test performance. All classifiers were trained with the feature representations extracted from the encoder models. Results for classification performance are reported with the F1-score in the horizontal axis of Figure 2, showing that our encoder extracts richer discriminative features and outperforms all other models. The main reason for this success is that we train the encoder separately, while previous approaches train it jointly with the generator.

We next evaluated image reconstruction given its features. Ideally, we would like to reverse the process and recreate the exact same image. Realistically, we expect the reconstructions to exhibit the most significant characteristics of the input image as faithfully as possible. Thus, we employ the perceptual distance (PD) based on the VGG network [46] to evaluate reconstruction similarity. The results are reported in the vertical axis of Figure 2. In contrast to the classification tasks, image reconstruction performance results are mixed and our approach (DSAE) obtains the best reconstruction performance in the tissue dataset, but seems to underperform in the cellular images dataset. After qualitative inspection of the reconstructed images we observe that samples generated by DSAE are realistic and very consistent with the general semantic features of the original image in both datasets. One of the reasons for this discrepancy in quantitative results is that the VGG-based perceptual distance is tuned to identify differences in natural RGB images, and the images of cells are out of that distribution in many ways, making it only an approximation of generative performance.

5 Latent image manipulation

Here, we explore image-based phenotype manipulations for interpreting biological experiments. We use a subset of the BBBC021 cellular dataset, which includes images of cancer cells treated with 113 compounds across eight different concentrations. To understand how compounds change cellular structure, we follow a common interpretation procedure of treatment effect based on comparing

control cells vs. treated cells. First, we created a DSAE model, projected the images into the feature space and normalized them to minimize the effect of technical variations [14]. Figure 3 visualizes single cells treated with various chemicals using the UMAP algorithm [47], with untreated cells in red and treated cells in gray. We also color cells treated with two chemicals over eight concentrations in blue (AZ-J) and green (AZ-A). To understand variations between untreated and treated cells in the underlying latent space we estimated the trajectory from the red cloud to the blue or green clouds using linear interpolations between their central points. Any point in this trajectory can be decoded using the the generative model of the DSAE, resulting in a smooth, continuous visualization of the transition of states, even when some of these were not collected experimentally in the lab.

Note that the feature space encodes untreated cells in the center and treated cells in the periphery, which is a meaningful biological pattern automatically discovered by the self-supervised algorithm. Also, the farther away the points are from the center red cloud, the higher the dose of the compounds, as displayed by the increasingly darker shades of blue and green of the corresponding treated points going from center to periphery. Importantly, the conditional generative model is able to generate images of intermediate phenotypic states after operating this latent space with simple manipulations. While other autoencoders are able to produce smooth transitions in the latent space, our approach is based on unbiased features that can be reliably used for other downstream analyses.

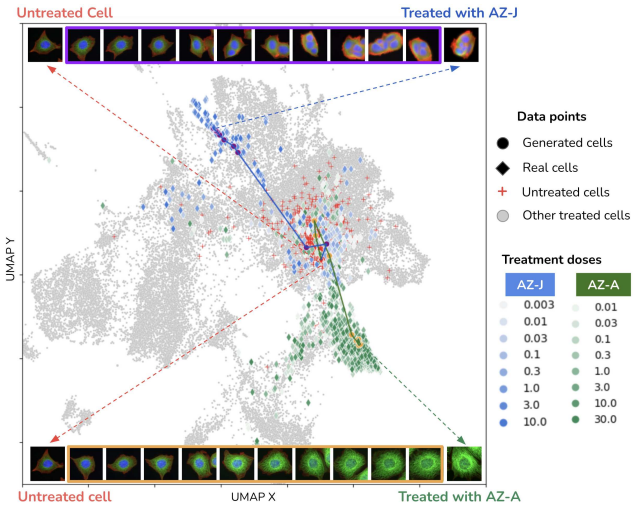

Figure 3: Effect of treatments on cells in the BBBC021 dataset. Images in the purple and yellow boxes are generated examples interpolated between the given phenotypes.

This evaluation is an average case interpolation that shows how useful the latent space and generative modeling approach can be in practice. However, image-based experiments contain more information, including a large amount of variation that is unexplained by experimental parameters, such as cell heterogeneity. Other analyses can be designed by introducing additional variables in the manipulation of the latent space, including weakly supervised feature transformations, factor analyses, and linear discriminant analysis, among others. These types of transformations can open the way to ask more sophisticated queries about the experiment, including how a single cell would respond under a different perturbation. Our method has the potential to support such developments, and we leave these possibilities open for future work.

6 Conclusions

In this work, we explore the potential of generative models for unbiased image-based phenotypic analysis and interpretation. Our approach is entirely image based and allows us to augment existing experimental image collections with synthetically generated images. An important aspect of our approach is that the generative model is conditioned on visual features obtained through self-supervision. This conditional mechanism does not require manual annotations and controls the generation of unobserved phenotypes to exist with respect to other observed phenotypes thanks to a feature space that captures relevant biological variation. We explored this property with simple latent space interpolations, and we envision the use of powerful probabilistic reasoning in the latent space as part of our future work. Our model is similar in spirit to image auto-encoders, however, we separate the representation learning from the generative modeling following a decoupled training strategy that leverages advances of both research fields to their full potential. The decoupled nature of the proposed DSAE paves the way to connect large pre-trained image (foundation) models with other successful generative models (e.g., diffusion models) in a modular way. This, together with improved latent space analysis, will enable increasingly accurate predictions of unknown cellular states.

Acknowledgments and Disclosure of Funding

This study was supported, in part, by the National Science Foundation NSF-DBI award 2134695 and Research Council of Finland awards 340273 and 346604.

References

[1] Michael D Slack, Elisabeth D Martinez, Lani F Wu, and Steven J Altschuler. Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences*, 105(49):19306–19311, 2008.

[2] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2): 145–159, 2021.

[3] Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Biorxiv*, pages 2022–01, 2022.

[4] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, page e11517, 2023.

[5] Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, mohammad lotfollahi, and Fabian Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26711–26722. Curran Associates, Inc., 2022.

[6] Jan Oscar Cross-Zamirski, Guy Williams, Elizabeth Mouchet, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Self-supervised learning of phenotypic representations from cell images with weak labels. *arXiv preprint arXiv:2209.07819*, 2022.

[7] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[8] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.

[9] Peter Goldsborough, Nick Pawlowski, Juan C Caicedo, Shantanu Singh, and Anne E Carpenter. Cytogan: generative modeling of cell images. *BioRxiv*, page 227645, 2017.

[10] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2021.

[11] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[12] Alexis Lamiable, Tiphaine Champetier, Francesco Leonardi, Ethan Cohen, Peter Sommer, David Hardy, Nicolas Argy, Achille Massougbodji, Elaine Del Nery, Gilles Cottrell, et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 14(1):6386, 2023.

[13] Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. *arXiv preprint arXiv:2303.08863*, 2023.

[14] D Michael Ando, Cory Y McLean, and Marc Berndl. Improving phenotypic measurements in high-content imaging screens. *BioRxiv*, page 161422, 2017.

[15] Katie Heiser, Peter F McLean, Chadwick T Davis, Ben Fogelson, Hannah B Gordon, Pamela Jacobson, Brett Hurst, Ben Miller, Ronald W Alfa, Berton A Earnshaw, et al. Identification of potential treatments for covid-19 through artificial intelligence-enabled phenomic analysis of human cells infected with sars-cov-2. *BioRxiv*, pages 2020–04, 2020.

[16] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *Biorxiv*, pages 2022–08, 2022.

[17] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders are scalable learners of cellular morphology. *arXiv preprint arXiv:2309.16064*, 2023.

[18] Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E Mancuso, Zachary F Phillips, Albert Kim, et al. A pooled cell painting crispr screening platform enables de novo inference of gene function by self-supervised deep learning. *bioRxiv*, pages 2023–08, 2023.

[19] Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo, Beth A Cimini, Kyle Karhohs, David J Logan, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell systems*, 13(11):911–923, 2022.

[20] Meraj Ramezani, Julia Bauman, Avtar Singh, Erin Weisbart, John Yong, Maria Lozada, Gregory P Way, Sanam L Kavari, Celeste Diaz, Marzieh Haghighi, et al. A genome-wide atlas of human cell morphology. *bioRxiv*, 2023.

[21] Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of single-cell feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9309–9318, 2018.

[22] Wolfgang M Pernice, Michael Doron, Alex Quach, Aditya Pratapa, Sultan Kenjeyev, Nicholas De Veaux, Michio Hirano, and Juan C Caicedo. Out of distribution generalization via interventional style transfer in single-cell microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2023.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[24] Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, pages 2023–06, 2023.

[25] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[29] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[32] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[33] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International conference on machine learning*, pages 4487–4499. PMLR, 2021.

[34] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[36] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

[37] Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cellsphenotypic profiling across cancer cell types. *Molecular cancer therapeutics*, 9 (6):1913–1926, 2010.

[38] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012.

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[40] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[41] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020.

[42] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 823–832, 2021.

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[44] Lorenzo Brigato and Stavroula Mougiakakou. No data augmentation? alternative regularizations for effective training on small datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 139–148, 2023.

[45] Ihab Bendidi, Adrien Bardes, Ethan Cohen, Alexis Lamiable, Guillaume Bollot, and Auguste Genovesio. No free lunch in self supervised representation learning. *arXiv preprint arXiv:2304.11718*, 2023.

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[47] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.