Class-Guided Image-to-Image Diffusion: Cell Painting from Brightfield Images with Class Labels

Jan Oscar Cross-Zamirski[1,2] Praveen Anand[2] Guy Williams[2] Elizabeth Mouchet[2] Yinhai Wang[2]
Carola-Bibiane Schönlieb[1]
[1] DAMTP, University of Cambridge, [2] Discovery Sciences, R&D, AstraZeneca
https://github.com/crosszamirski/guided-I2I

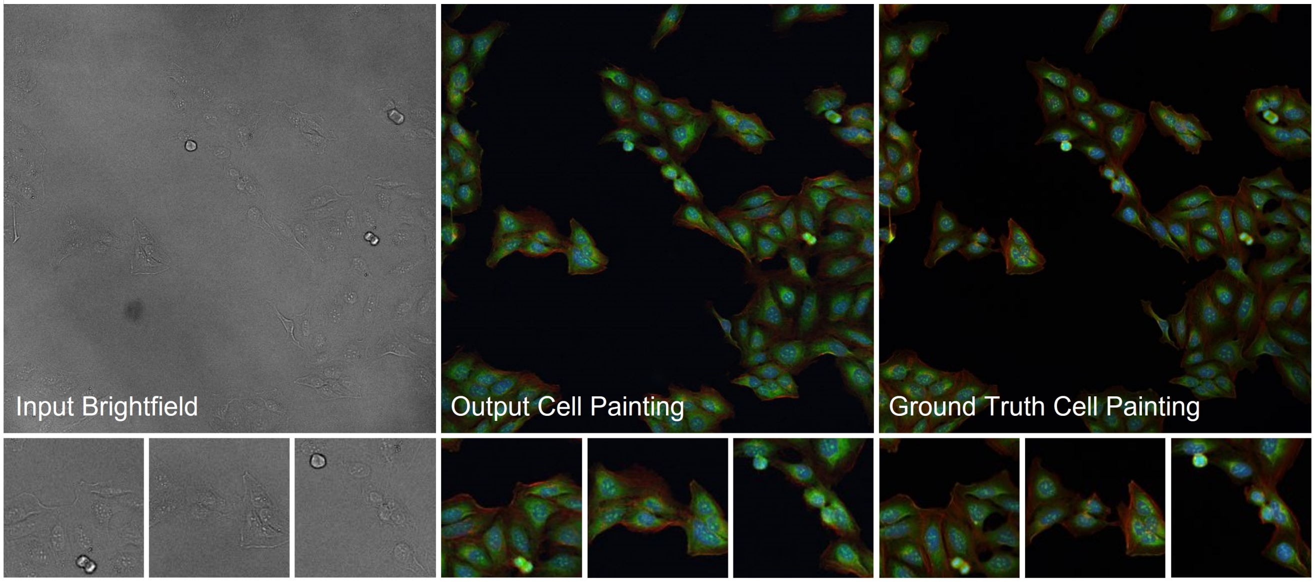

Figure 1: A colour composite example of three channels from a test plate: red (AGP), green (ER) and blue (DNA).

Abstract

Image-to-image reconstruction problems with free or inexpensive metadata in the form of class labels appear often in biological and medical image domains. Existing text-guided or style-transfer image-to-image approaches do not translate to datasets where additional information is provided as discrete classes. We introduce and implement a model which combines image-to-image and class-guided denoising diffusion probabilistic models. We train our model on a real-world dataset of microscopy images used for drug discovery, with and without incorporating metadata labels. By exploring the properties of image-to-image diffusion with relevant labels, we show that class-guided image-to-image diffusion can improve the meaningful content of the reconstructed images and outperform the unguided model in useful downstream tasks.

1. Introduction

Conditional denoising diffusion probabilistic models (DDPMs) [24, 19] are trained to learn a probability distribution capable of generating realistic samples from an input condition. These constructions typically fall into one of two categories: models conditional on an input image (image-to-image) [34] **or** models conditional on a class label [19, 36]. While many other diffusion models exist which incorporate natural language text encoders such as CLIP [32] (text-to-image) [33, 35], there has been much less attention on advancing models with both paired image **and** class label information. This can be attributed to a lack of generalist datasets which have both class labels and paired images, as this information can be expensive, sparse or narrow in application [47, 51].

Despite this, image-to-image problems with discrete metadata appear often in biological and medical image reconstruction. Examples of these inverse problems include PET reconstruction from MRI [43], predicting fluorescent labels from transmitted light microscopy [12], sparse-view CT reconstruction and artifact removal [41]. Through the nature of image acquisition there is often additional inexpensive *side* [43] or *weak label* [7] information which can be incorporated to guide the training of the inverse process towards the main task. For biological and medical datasets, class labels have been used in deep learning architectures to learn more faithful and generalisable representations [45, 31], and as extra information in image-to-image tasks [43].

These problems are dataset specific, and well-established databases of natural images [18, 53] and their associated labels are rarely analogous to the challenges presented by biological and medical datasets. These real-world datasets can have unique types of metadata labels, and application-specific ways to evaluate performance in downstream tasks. Using the predicted images in such tasks to quantify performance may be more important and informative than benchmarking with metrics such as Fréchet Inception Distance (FID) [23] and structural similarity index (SSIM) [25]. Accurately capturing a distribution of images is complementary, if not subsidiary, to being able to differentiate between images and their features [10].

We investigate the utility of image-to-image diffusion with class labels using a subset of Target2 data generated as part of the the JUMP-CP effort[1] to predict Cell Painting [5] images from paired brightfield images [16]. We find that the quality of extracted morphological features from the predicted images, and their performance on downstream mechanism of action prediction and clustering tasks can be boosted with relevant labels. This type of approach may lead to increased clinical success of image-to-image methods in drug discovery [10] and related medical reconstruction tasks [49], as a way to guide the image generation with biologically informative class information. In this study we make the following contributions:

- We introduce and implement a general framework for class-guided image-to-image diffusion, our model building upon the *Palette* image-to-image framework [34] and guided diffusion [19] .

- We apply our model to the prediction of 5-channel Cell Painting fluorescent microscopy from 3-channel brightfield images, and show that incorporating label information can improve performance. We evaluate the images with extracted biological features and a transfer learning approach to simulate image-based profiling in a drug discovery pipeline.

2. Related work

Generative adversarial networks (GANs) [21] have been the prevailing method for image-to-image translation tasks since the introduction of *pix2pix* [27] in 2016. GAN based methods have been widely adopted in medical imaging for a variety of tasks [50] including PET denoising, PET-CT translation and correction of magnetic resonance motion artefacts [3]. GANs are used in cell microscopy for cross-modality prediction [4] and super resolution [52].

Other models used for reconstruction tasks include variational auto-encoders (VAEs) [30] and normalizing flows [29]. VAEs have been used to learn or approximate the joint distribution of multiple modalities [47], sometimes with a product or mixture of experts approach to combine the distributions [38]. Product of experts have also be used for multimodal conditional image synthesis with GANs [26]. Flow-based models for modality transfer (such as MRI to PET) have outperformed conditional GANs and VAEs while leveraging *side* information [43].

Diffusion models are growing in popularity in medical imaging and have been used predominantly for MRI and CT modalities in reconstruction problems [28]. Diffusion-based generative models can achieve state of the art image quality without suffering from problems such as mode collapse, training instability, or not allowing for likelihood estimation. By comparison, GANs can suffer from training instability and mode collapse [15] in addition to feature hallucinations which are particularly undesirable in medical applications [14]. VAEs do not produce high quality image samples and flow-based models have restrictions such as the requirement for invertibility of the network.

Weakly-supervised diffusion models have been used in medical imaging, notably in anomaly detection [37, 46]. However, these models are not strictly guided image-to-image models and instead use the difference between the ground truth and reconstructed image for anomaly detection. This method would not generalize beyond anomaly detection. *InstructPix2Pix* [6] combines a text-guided conditional diffusion model with an image-to-image framework using text based prompts. However, text encoders for style-transfer are not appropriate in datasets where metadata labels are discrete classes.

Hence there is scope for developing a diffusion model for image reconstruction with discrete metadata. Examples of image-to-image problems with (often under-utilised) data include: prediction of fluorescent image channels from transmitted light images for drug discovery [12] where freely available weak labels include treatment and compound [7], as well as batch information. Experimental batch effects can be significant, and batch information has been integrated into a number of machine learning models in image-based profiling [31, 45, 2]. PET reconstruction from much cheaper MRI scans for Alzheimer's prediction also

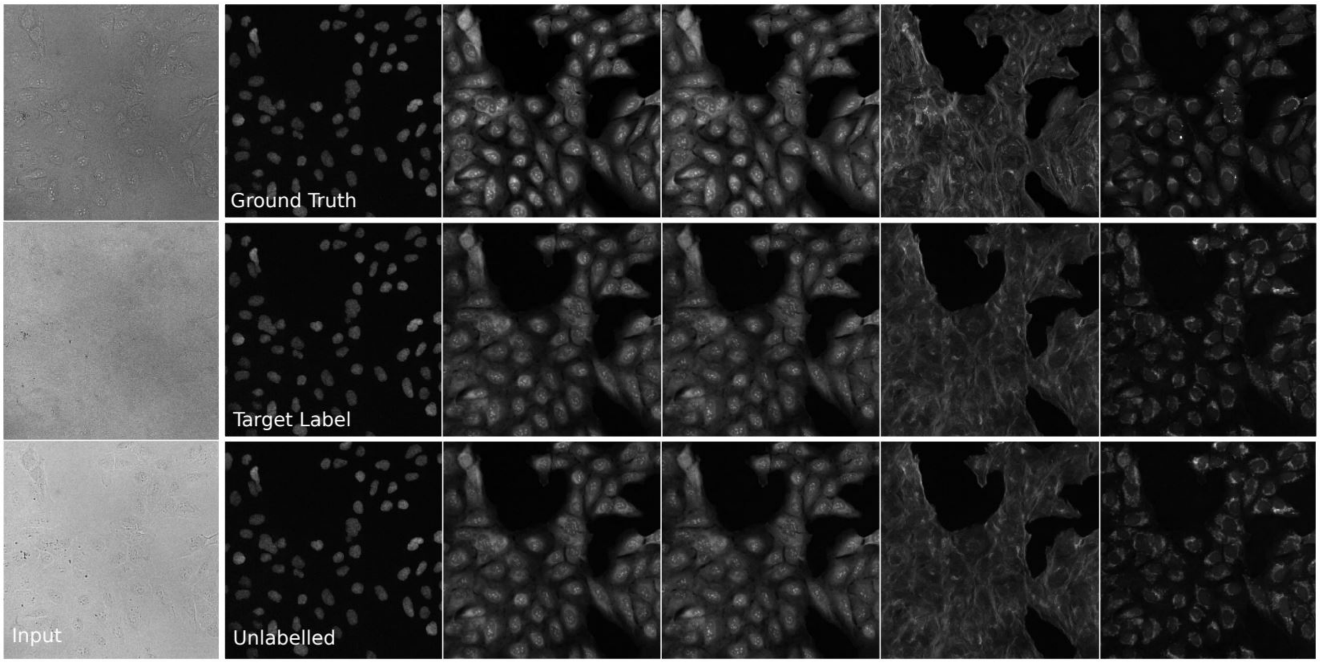

Figure 2: Given input Brightfield (3 channels) our model is able to generate 5 Cell Painting channels. Incorporating meaningful labels can improve biological feature quality and performance on downstream tasks without significantly reducing image quality or adding background noise. Columns left to right: Brightfield (input), DNA, RNA, ER, Mito, AGP.

has inexpensive and relevant metadata such as patient age, sex, disease status and genotype which has been incorporated into improving image reconstruction quality [43].

To the best of the authors' knowledge, our work is the first to use a diffusion model for fluorescent microscopy prediction. We build upon existing studies using deep learning to predict fluorescent labels [12] from transmitted light images such as brightfield, a cheaper and less invasive modality for imaging cells which can still capture meaningful information [22]. Specifically, we predict Cell Painting [5] image channels which capture rich cell morphology information which can be used in a variety of tasks in image-based profiling including bioactivity, cytotoxicity and mechanism of action prediction [10].

2.1. Image-to-image conditional diffusion

We base our model on the *Palette* framework for image-to-image diffusion from Saharia *et al.* [34]. Their model outperforms GANs on four tasks: colorization, inpainting, uncropping and JPEG restoration. *Palette* is a denoising diffusion probabilistic model [24] of the form $p(\boldsymbol{y} \mid \boldsymbol{x})$ which is trained to predict the output image $\boldsymbol{y}$ conditional on the input image $\boldsymbol{x}$. The noisy image $\widetilde{\boldsymbol{y}}$ is given by:

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\boldsymbol{y} + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{1}$$

for Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and noise level indicator γ. A neural network f_θ is trained to denoise $\widetilde{\boldsymbol{y}}$ for a given $\boldsymbol{x}$ with the loss function:

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\mathbb{E}_{\gamma}\left\| f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\boldsymbol{y} + \sqrt{1-\gamma}\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon} \right\|_p^p \tag{2}$$

where p is the chosen norm (L_1 or L_2). Eq. (2) is the image-conditional version of L_{simple} from Ho *et al.* [24].

The reverse diffusion process is computed step-by-step as:

$$\boldsymbol{y}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \tag{3}$$

for $t = T, \ldots, 1$ steps. The noise level indicator γ_t is a function of t, and α_t is the noise variance scale parameter (also timestep-dependent).

2.2. Conditional image synthesis

For conditional image synthesis with class labels, Dhariwal and Nichol [19] introduced two modifications to unconditional DDPM from Ho *et al.* [24]: adaptive group normalization (AdaGN) and classifier guidance (CG). AdaGN is a modification to the architecture which incorporates the class information into normalization layers in training, while classifier guidance exploits the gradients of a pre-trained classifier to guide the inference process (note: in this section we change $y \to k$ and $\boldsymbol{x} \to \boldsymbol{y}$ from the original paper to be consistent with the notation used in this paper)

2.2.1 Adaptive group normalization

AdaGN is a layer used to incorporate the timestep and class embedding into the residual blocks following a group normalization operation [48]. It is defined as:

$$\text{AdaGN}(h, k_{\text{proj}}) = k_s \text{GroupNorm}(h) + k_b \tag{4}$$

where $k_{\text{proj}} = [k_s, k_b]$ is a linear projection of the timestep and class embedding, and h is the activations of the residual block after the first convolution. This layer can be incorporated in the absence of class labels with just the timestep embedding: $\text{AdaGN} = k_s \text{GroupNorm}(h)$.

2.2.2 Classifier guidance

Classifier guidance enables the use of class information in inference of the trained diffusion model. Sohl-Dickstein *et al.* [39] and Song *et al.* [42] showed this can be achieved using pre-trained classifier gradients to condition the sampling of the diffusion model. First, the classifier $p_\phi(k \mid \boldsymbol{y}_t)$ is pre-trained to predict the class k from noisy images $\boldsymbol{y}_t$.

The aim is to sample each transition from the distribution:

$$p_{\theta,\phi}(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, k) = Z p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}) p_\phi(k \mid \boldsymbol{y}_t) \tag{5}$$

where $p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1})$ is the unconditional reverse noising process and Z is a normalizing constant. Although it is intractable to sample from the distribution in Eq. (5), it can be approximated as a perturbed Gaussian distribution [39]:

$$\log(p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}) p_\phi(k \mid \boldsymbol{y}_t)) \approx \log p(\boldsymbol{z}) + C, \tag{6}$$

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma} g, \boldsymbol{\Sigma}), \quad g = \nabla_{\boldsymbol{y}_t} \log p_\phi(k \mid \boldsymbol{y}_t)|_{\boldsymbol{y}_t = \mu} \tag{7}$$

where $g = \nabla_{\boldsymbol{y}_t} \log p_\phi(k \mid \boldsymbol{y}_t)$ are the gradients of the classifier and C is a constant which can be ignored. In inference, this shifts the mean of the sampled Gaussian to guide the denoising process towards the given class label k. The relative weighting of the classifier guidance term can be scaled with a constant s.

3. Class-guided image-to-image diffusion

In *Palette*, Saharia *et al.* [34] removed both classifier guidance and the class embedding of the AdaGN layer introduced by Dhariwal and Nichol [19]. In this study we reintroduce the class label k while retaining the conditional dependence on input image $\boldsymbol{x}$. We redefine the input conditions for image and associated class label as $(\boldsymbol{x}_k, k)$.

Algorithm 1: Training the denoising model f_θ

```
repeat
    $(\boldsymbol{x}_k, \boldsymbol{y}_0, k) \sim p(\boldsymbol{x}_k, \boldsymbol{y}, k)$
    $\gamma \sim p(\gamma)$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
    Take a gradient descent step on
        $\nabla_\theta \left\| f_\theta(\boldsymbol{x}_k, \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, k, \gamma) - \boldsymbol{\epsilon} \right\|_p^p$
until converged
```

Algorithm 2: Classifier guided diffusion sampling, given a diffusion model $(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t))$, classifier $p_\phi(k \mid \boldsymbol{y}_t)$, and gradient scale s.

```
Input: class label $k$, input image $\boldsymbol{x}_k$, gradient scale $s$
$\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
for $t = T, \dots, 1$ do
    $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu} + s\Sigma\nabla_{\boldsymbol{y}_t} \log p_\phi(k \mid \boldsymbol{y}_t), \boldsymbol{\Sigma})$ if $t > 1$,
    else $\boldsymbol{z} = \boldsymbol{0}$
    $\boldsymbol{y}_{t-1} =$
    $\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}_k, \boldsymbol{y}_t, k, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{z}$
end for
return $\boldsymbol{y}_0$
```

We summarise the training scheme for class guided image-to-image diffusion in Algorithm 1, and the sampling scheme in Algorithm 2. Each iteration of the reverse process of class-guided image-to-image diffusion can be computed as:

$$\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}_k, \boldsymbol{y}_t, k, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{z} \tag{8}$$

for $t = T, \dots, 1$. Here:

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu} + s\boldsymbol{\Sigma}\nabla_{\boldsymbol{y}_t} \log p_\phi(k \mid \boldsymbol{y}_t), \boldsymbol{\Sigma}) \quad \text{if} \quad t > 1, \quad \text{else} \quad \boldsymbol{z} = \boldsymbol{0} \tag{9}$$

The f_θ dependence on k is achieved with the AdaGN layer. It is optional to use k in sampling, as the AdaGN layer does not need to see the label. We test this and find that although it is possible to exclude k in sampling, it is necessary to include for improved performance. It is also possible to sample without classifier guidance by setting $s = 0$.

The training objective follows the form of Eq. (2) with L_2 norm. Our network is a U-Net architecture [24] which is based on the modified 256×256 class-conditional U-Net model used in *Palette* [19, 40]. The network is adapted to take images of size 512×512 with 3 input channels and 5 output channels in order to fit the requirements of the brightfield and Cell Painting channels.

4. Experiments

4.1. Dataset

We used a subset of one of the publicly-available JUMP Cell Painting dataset cpg0000 [11], available from the Cell Painting Gallery on the Registry of Open Data on AWS (https://registry.opendata.aws/cellpainting-gallery/). 10 plates (experimental replicates) were chosen to ensure a variety of biological phenotypes were present. They contain pairs of compounds associated by the genes they target, in addition to 46 controls compounds with a variety of mechanisms. In total, every plate contains around 2000 images - each with 5 Cell Painting channels and 3 brightfield channels. These plates were screened regularly throughout data production to enable downstream assessment of connectivity of perturbations between batches of compounds screen. Every plate contains treated cells representing 290 perturbations, each with paired perturbation with a matching target (145 targets total). Imaging details are provided in the supplementary material.

4.2. Pre-processing

To ensure that systematic variations in pixel intensity were not present in input images, we used a standardised CellProfiler [9] pipeline to perform illumination correction on all images. A smoothing function of filter size 249 pixels was used to generate an illumination correction function per imaging channel for each plate. The pixel intensities of all images were then divided by their respective correction function. This methodology is consistent with best practice established during the JUMP Cell Painting consortium [13]. After illumination correction, all images were re-sized to 512×512 pixels using bicubic interpolation. The images were all normalised to have a standard deviation of 1 and a mean of 0 with a maximum pixel intensity cutoff of 15 enforced to exclude extreme outliers.

4.3. Model training

We trained each model using 9 training plates and evaluated on a single, unseen test plate. For each model this was done twice, learning weights for 2 different, randomly selected test plates (the same 2 plates for each model). This is equivalent to k-fold cross validation - although we trained 2 versions of each model rather than 10, as producing 10 full plates per model was not possible due to the computationally intensive nature of sampling DDPMs.

Using the full plates, we trained models with no labels (*Palette*), perturbation (pert) as a weak label, target as a label. The labels were included through the AdaGN layer. We compared using the labels in training and inference against using labels in training but not in inference through the AdaGN layer to test if the labels were required in sampling. Target as a label was included as a proof of concept of the method, but it is expensive information not freely available in a practical setting (compared to the perturbation which is free information). Classifier guidance was not used to generate entire plates due to the extreme computational demands (over 500 GPU hours per model, per plate). This training regime is equivalent to Algorithm 2 with $s = 0$.

The active subset was around one third of the full plate, and this allowed us to sample using classifier guidance with $s = 1$. Additionally, training with known active compounds would provide more meaningful class labels for the model. The classifier $p_\phi(k \mid \boldsymbol{y}_t)$ was the downsampling branch of the U-Net with an additional output layer (as introduced by Ho *et al.* [24]). Training images were noised with the timestep dependent noise distribution, and the model was trained until the loss converged. The full training and sampling schemes are presented in Algorithms 1 and 2.

In training, the images were subject to random horizontal and vertical flips and 90 degree rotations each with probability $p = 0.5$. Models were trained until the loss appeared to stop decreasing, which was typically around $250,000$ iterations. Even though the quality of cellular structures appeared to improve beyond this, we found overfitting to be a problem for larger number of epochs as phantom structures appeared on the empty background. All models were trained with a batch size of 2 and the Adam optimizer with a learning rate of $8e^{-5}$. The linear noise schedule of $(10e^{-6}, 0.001)$ (as in *Palette*) with $T = 2000$ was used in training and inference. We provide the code and parameters to replicate these models in our GitHub repository. All the models were trained on the AstraZeneca Scientific Computing Platform (SCP) with 32GB GPUs. Total training time was around 24 hours on a single GPU, and sampling from the trained model was around 4 minutes per 5 channel image (increasing to 15 minutes with classifier guidance).

4.4. Post-processing

The model outputted channels were re-normalised as in the pre-processing. CellProfiler [9] was used to segment nuclei, cells and cytoplasm, then extract morphological features from each of the channels. Single cell measurements of fluorescence intensity, texture, granularity, density, location and various other features were calculated as feature vectors. Features were aggregated for each perturbation using the median value per image.

The Pycytominer package (`https://github.com/cytomining/pycytominer`) was used to normalise the cell-painting features generated for the synthetic images. The features derived for synthetic images generated by each model were normalized using all the samples. All the features generated from the ground truth data were also used for the prediction feature selection operation to allow for a fair comparison. These included dropping na columns, variance thresholding, correlation thresholding and dropping blocklisted features. Approximately 650-700 cell-painting features were selected for each plate. Features were aggregated to the perturbation level, giving 290 features per plate.

In order to segregate the active perturbations from inactives, PCA was performed using 1262 Cell Painting features that remained after CellProfiler feature selection (from the ground truth images). The top 100 dimensions of the PCA were then used to evaluate the cosine-distances between all-pairs of data points (well). An average cosine-distance score against the negative DMSO controls across all replicates was used as a score to segregate out the actives from inactives using 1D C-kmeans clustering algorithm with $k = 3$ for 3 clusters. There were a total of 118 perturbations representing 59 targets selected for the active subset, with the remaining perturbations (inactives) showing no phenotypic divergence from negative controls. We provide visualisations of the dataset and the active subset in the supplementary material.

4.5. Transfer learning with DINO

We used the self-supervised learning algorithm DINO [8] pre-trained with ImageNet [18] weights to profile the images with transfer learning, following the methodology of previous studies [17]. The backbone of the network is a vision transformer (ViT-S/8) with a 3-layer multi-layer perceptron head, from which the embeddings are extracted. The median feature embedding was taken from four 224×224 crops around the centre of each image (equivalent to a 448×448 pixel centre crop split into four non-overlapping crops). Embeddings of size 384 were extracted for each channel then concatenated to a size of 1920 for 5-channel Cell Painting (1152 for brightfield) followed by L_2 normalization. These feature representations, like the CellProfiler features, were then used for target prediction.

5. Results

5.1. Evaluation

We evaluated our models with image-level and feature-level metrics, which are presented in Table 1 (the entire plate) and Table 2 (the active subset). We compared Pearson correlation coefficient (PCC), Fréchet Inception Distance (FID) [23], structural similarity index measure (SSIM) [25] and mean-squared and mean-absolute error (MSE/MAE). The values in the tables were calculated by comparing the predicted images with the ground truth images for each model. The values presented are the mean values of all the images. Examples of the images are presented in Figures 1, 2 and 3. We also compare the FID scores and feature values between the two ground truth plates as the limit of a perfect reconstruction (each plate is meant to be an experimental replication of the same cells and treatments). The feature-level metrics were chosen to be representative of downstream applications which would be performed with real Cell Painting images in a drug discovery pipeline.

5.1.1 NN matching / NN top 5

We searched the feature spaces of each plate - both CellProfiler (CP) and transfer learning (TL) spaces - for the nearest neighbours by cosine distance. The values reported in the tables are the total number of matching targets which are nearest neighbours in the feature space of the model or ground truth plate feature space (for both plates). We repeated this analysis but for each point searching for the 5 nearest neighbours, and reporting a match if one of the 5 perturbations shared a target with the chosen point.

5.1.2 Matching target distance (MTdist)

Since there hundreds of targets and perturbations in this dataset, even searching the top 5 nearest neighbours is not sufficient to evaluate the relationships between targets. We propose the mean matching target distance (MTdist) as an informative metric. For each pair of perturbations sharing a matching target, the cosine distance is calculated between the points in feature space. The mean distance for all 290 perturbations (118 in the active subset) is presented for each model. Since the models feature spaces are normalised this should be a fair comparison between the models.

5.1.3 CellProfiler feature correlation (CPcor)

Following the methodology of previous Cell Painting prediction studies [16], we correlated each model's CellProfiler features to the ground truth CellProfiler features, and report the mean value. We also correlate the features between the ground truth replicates as a baseline (0.569 for the whole plate and 0.615 for the active subset). We would not expect the model generating features from an unseen batch to exceed this value. We include a breakdown of the features by group and channel in the supplementary material, alongside two-dimensional t-SNE plots of the features.

6. Discussion and conclusion

The purpose of this study was to explore how metadata in the form of discrete classes can be used to guide image-to-image translation tasks. DDPMs and other generative models have been successful in achieving state of the art FID scores, however learning details which differentiate between images based on biology and structure is less studied when compared to generating realistic images which could have been sampled from a training distribution.

All the models achieved very low FID scores. For entire plates (Table 1), incorporating labels through AdaGN

Label		**PCC ↑**	**FID ↓**	**SSIM ↑**	**MSE / MAE ↓**	**NN matches ↑**	**NN Top 5 ↑**	**MTdist ↓**	**CPcor ↑**
Training	**Sampling**					**CP / TL**	**CP / TL**	**CP / TL**	
None	None	**0.793**	3.54	**0.350**	**0.400 / 0.338**	6 / **8**	**23 / 25**	**0.886 / 0.0729**	**0.430**
Pert	None	0.760	**3.26**	0.267	0.465 / 0.380	**7** / 5	20 / **25**	0.910 / 0.0852	0.384
Pert	Pert	0.752	3.49	0.260	0.481 / 0.392	**7** / 3	22 / 22	0.919 / 0.0936	0.381
Target*	None	0.741	3.69	0.239	0.489 / 0.402	4 / 4	15 / 17	0.888 / 0.0834	0.283
Target*	Target*	0.745	3.86	0.228	0.495 / 0.408	**17 / 15**	**32 / 51**	**0.791** / 0.0836	0.327
GT Cell Painting		–	1.55†	–	–	12 / 13	31 / 28	0.868 / 0.0924	0.569†
GT Brightfield		–	–	–	–	– / 12	– / 28	– / (0.0551)	–

Table 1: Mean image and feature metrics for class-guided image-to-image models for two full plates, each generated with a different model. Note the brightfield feature space (3 channels) is a different size to the Cell Painting feature space (5 channels). *Target is not a freely available label and is included as a proof of concept. †We provide FID and CPcor values calculated between the two ground truth (GT) test plates, which are prepared and treated as identical replicates.

Label		**PCC ↑**	**SSIM ↑**	**MSE / MAE ↓**	**NN matches ↑**	**NN Top 5 ↑**	**MTdist ↓**	**CPcor ↑**
AdaGN	**CG**				**CP / TL**	**CP / TL**	**CP / TL**	
None	None	**0.773**	0.294	**0.423** / 0.320	4 / 2	12 / 16	0.971 / 1.533	0.386
Pert	None	0.762	**0.379**	0.444 / **0.310**	6 / 4	18 / **18**	0.939 / **1.357**	**0.507**
Pert	Pert	0.752	0.338	0.463 / 0.330	**7 / 7**	**24 / 18**	**0.929** / 1.541	0.504
Target*	None	0.730	0.235	0.506 / 0.375	9 / 14	27 / **27**	0.883 / 1.405	0.404
Target*	Target*	0.696	0.202	0.573 / 0.408	**11** / 6	**31** / 21	**0.879** / 1.579	0.355
GT Cell Painting		–	–	–	9 / 13	21 / 26	0.919 / 0.233	0.615†
GT Brightfield		–	–	–	– / 16	– / 26	– / (1.148)	–

Table 2: The analysis of Table 1 is repeated for the active subset only. There are too few images in the active subset to calculate FID. *Target is not a freely available label and is included as a proof of concept. †We provide the CPcor value calculated between actives in the two ground truth (GT) test plates, which are prepared and treated as identical replicates.

generally reduced the performance of the pixel-level metrics, although using the perturbation as a label in training resulted in the lowest FID score. Some images from the labelled models had some background noise which was not present in the unlabelled model, and this is reflected in the image-level metrics. We present an example of this effect in the supplementary material. While it is possible that class labels can improve certain aspects of the images, they may also reduce the image quality by fitting to unwanted background noise if there are uninformative class labels or no signal to be found in the training set. This was particularly notable using target as the label, which moved matching targets closer in feature space, but reduced the faithfulness of the generated image. This effect is likely amplified in high-content microscopy images where over 50% of the pixels are irrelevant background with no cellular structures.

The results for the model trained with full plates of images (290 perturbations) suggest that the image-to-image model is capable of capturing strong phenotypic signals (true positives) but struggled with noisy, lower signal images (the inactives). We may have led the model astray with uninformative labels in training. To test this theory, we repeated the analysis with the active subset (Table 2), which represents the images of cells with meaningful and quantifiable phenotypic differences from the control group (untreated cells). This resulted in a significant improvement over the unlabelled model (*Palette*) in SSIM, target matching and CellProfiler feature correlation. Furthermore, the pixel correlations and errors were not significantly reduced, so unlike when using the whole plate, there was less of a cost to the improved performance. Incorporating classifier guidance improved target matching but also at a small cost to image and feature quality. Our results show that class-guided image-to-image diffusion improves upon the naive model under well-chosen conditions, and highlight how crucial the quality of labels and training data is.

The values in Tables 1 were produced by models trained and tested on the whole plate, while Table 2 presents results from images trained with the smaller active subset. The smaller training set of the active subset reduced the qual-

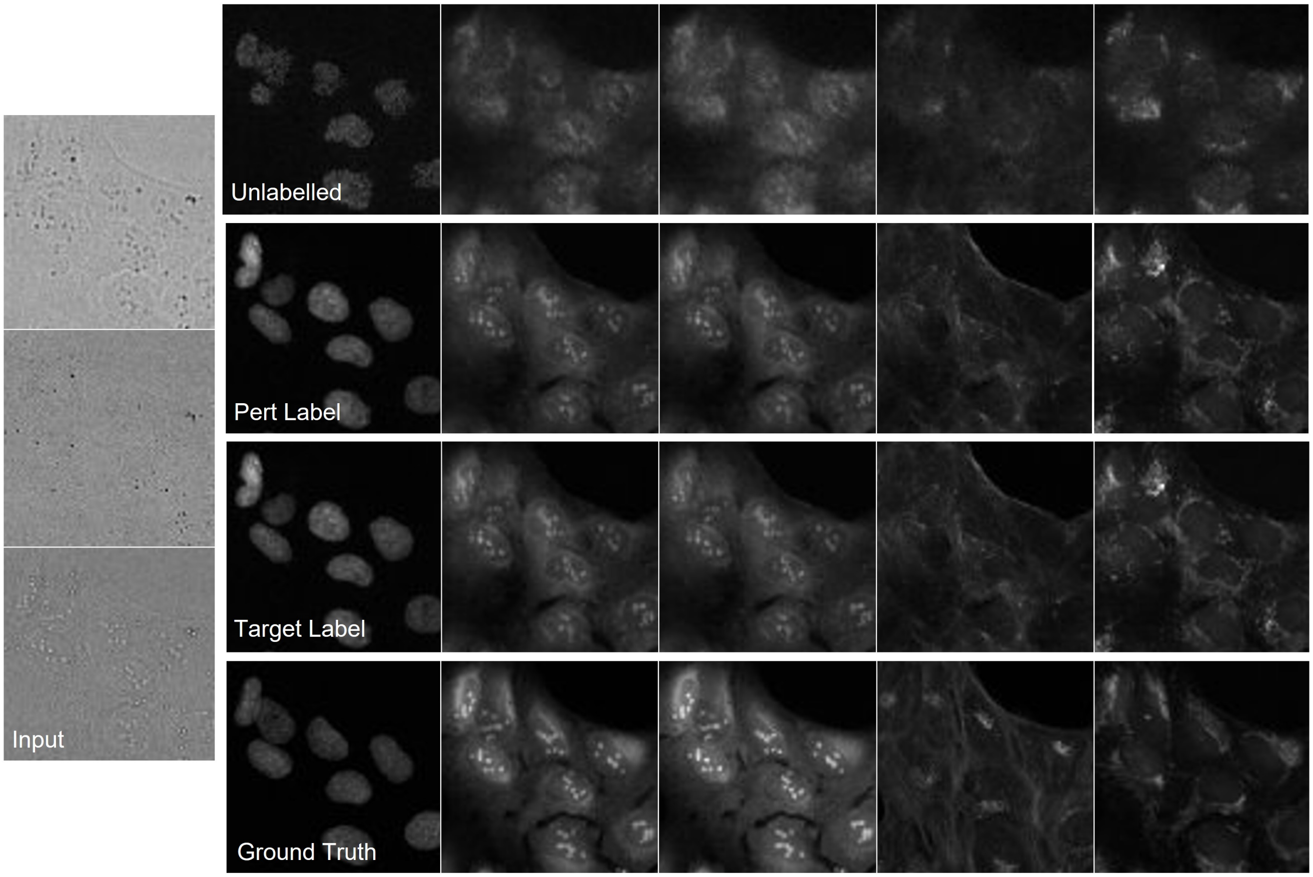

Figure 3: Images generated by models trained with the active subset. The labelled images were sampled with both AdaGN and CG. Cropped to 100×100 pixels. Columns left to right: Brightfield (input), DNA, RNA, ER, Mito, AGP.

ity of the unlabelled model. However, incorporating labels produced the highest correlations of features, and the highest SSIM even in the low training data regime. These effects were observed in both training splits. Very recently, Cell Painting datasets of immense scale with millions of images across thousands of compounds and over 50 batches have become public, and hold great promise for machine learning in drug discovery [44, 20]. The batch effect is a large part of this, and we explore the batch effect properties of our models in the supplementary material.

This study provides a valuable comparison of methods employing brightfield image channels as an input for image-based profiling. Recent studies have explored this under-utilised modality which may contain as much predictive power as fluorescent stained images [22]. Our results further reveal the potential of brightfield both as an input for cross modality prediction and as a competitive profiling modality in itself. This success may also be attributed to powerful, pretrained attention based architectures [8], which can overcome the traditional drawbacks of brightfield and are able to find meaningful structures from noisy images (we present self-attention maps of transfer learning with brightfield images in the supplementary material). Brightfield and transmitted light has traditionally been seen as less informative than fluorescent staining, but the limit of brightfield may be higher than previously thought. Furthermore, we have presented a way to use brightfield to generate full plates of model-generated Cell Painting, from which existing software can extract hand-crafted features for a greater level of interpretability.

In conclusion, we present a novel way to use discrete metadata to guide image-to-image translation. We predict unseen batches of Cell Painting from brightfield, and surpass the performance of previous methods in multiple metrics [16]. We perform image-based profiling predictions with the model predicted plates and achieve stronger results when using the freely available perturbation label with the active subset. This includes phenotypic feature correlations, SSIM and target matching, a common task in drug discovery. We propose our method could have impact in other biomedical fields to guide learning meaningful features and structures with multimodal data.

References

[1] JUMP-Target, JUMP-Cell Painting Consortium, The Broad Institute, 2022. `https://github.com/jump-cellpainting/JUMP-Target`.

[2] DM. Ando, CY. McLean, and M. Berndl. Improving phenotypic measurements in high-content imaging screens. *BioRxiv preprint*, page 161422, 2017. `https://www.biorxiv.org/content/10.1101/161422v1.abstract`.

[3] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[4] C. Belthangady and LA. Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat Methods*, 16:1215–1225, 2019.

[5] MA. Bray, S. Singh, H. Han, CT. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, SM. Gustafsdottir, CC. Gibson, and AE. Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

[6] T. Brooks, A. Holynski, and AA. Efros. Instructpix2pix: Learning to follo w image editing instructions. *arXiv preprint*, 2022. `https://arxiv.org/abs/2211.09800`.

[7] JC. Caicedo, C. McQuin, A. Goodman, S. Singh, and AE. Carpenter. Weakly supervised learning of single-cell feature embeddings. *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 9309–9318, 2018.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A Joulin. Emerging properties in self-supervised vision transformers. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[9] AE. Carpenter, TR. Jones, MR. Lamprecht, C. Clarke, IH. Kang, O. Friman, DA. Guertin, JH. Chang, RA. Lindquist, J. Moffat, Golland P., and DM. Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7:R100, 2006.

[10] SN. Chandrasekaran, H. Ceulemans, JD. Boyd, and AE. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat Rev Drug Discov*, 20:145–159, 2021.

[11] Chandrasekaran, SN. and Cimini, BA. and Goodale, A. and Miller, L. and Kost-Alimova, M. and Jamali, N. and Doench, J. and Fritchman, B. and Skepner, A. and Melanson, M. and Arevalo, J. and Caicedo, JC. and Kuhn, D. and Hernandez, D. and Berstler, J. and Shafqat-Abbasi, H. and Root, D. and Swalley, S. and Singh, S. and Carpenter, AE. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations *bioRxiv*, 2022.01.05.475090 2022

[12] EM. Christiansen, SJ. Yang, DM. Ando, A. Javaherian, G. Skibinski, S. Lipnick, E. Mount, A. O'neil, K. Shah, AK. Lee, and P. Goyal. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803, 2018.

[13] BA. Cimini, SN. Chandrasekaran, M. Kost-Alimova, L. Miller, A. Goodale, B. Fritchman, P. Byrne, S. Garg, N. Jamali, DJ. Logan, HB. Concannon, C-H. Lardeau, E. Mouchet, S. Singh, SH. Abbasi, P. Aspesi Jr, JD. Boyd, T. Gilbert, D. Gnutt, S. Hariharan, D. Hernandez, G. Hormel, K. Juhani, M. Melanson, L. Mervin, T. Monteverde, JE. Pilling, A. Skepner, SE. Swalley, A. Vrcic, E. Weisbart, G. Williams, A. Yu, B. Zapiec, and AE. Carpenter. Optimizing the cell painting assay for image-based profiling. *BioRxiv preprint*, 2022. `https://doi.org/10.1101/2022.07.13.499171`.

[14] JP. Cohen, M. Luck, and S. Honari. Distribution matching losses can hallucinate features in medical image translation. *In International conference on medical image computing and computer-assisted intervention*, pages 529–536, 2018. Springer, Cham.

[15] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and AA. Bharath. Generative adversarial networks: An overview. *in IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[16] JO. Cross-Zamirski, G. Williams, E. Mouchet, C-B. Schönlieb, R. Turkki, and Y. Wang. Label-free prediction of cell painting from brightfield images. *Sci Rep*, 12(10001), 2022.

[17] JO. Cross-Zamirski, G. Williams, E. Mouchet, C-B. Schönlieb, R. Turkki, and Y. Wang. Self-supervised learning of phenotypic representations from cell images with weak labels. *arXiv preprint arXiv:2209.07819*, 2022.

[18] J. Deng, W. Dong, R. Socher, L-J. Li, Li K., and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[19] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint*, 2021. `https://arxiv.org/abs/2105.05233`.

[20] MM. Fay, O. Kraus, M. Victors, L. Arumugam, K. Vuggumudi, J. Urbanik, K. Hansen, S. Celik, N. Cernek, G. Jagannathan, and J. Christensen. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023.

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

[22] A. Gupta, PJ. Harrison, H. Wieslander, J. Rietdijk, JC. Puigvert, P. Georgiev, C. Wählby, O. Spjuth, and Sintorn I-M. Is brightfield all you need for mechanism of action prediction? *bioRxiv preprint*, 2022. `https://doi.org/10.1101/2022.10.12.511869`.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems, 30*, 2017.

[24] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint*, 2020. `https://arxiv.org/abs/2006.11239`.

[25] A. Horé and D. Ziou. Image quality metrics: Psnr vs. ssim. *20th International Conference on Pattern Recognition*, page 2366–2369, 2010.

[26] X. Huang, A. Mallya, TC. Wang, and MY. Liu. Multimodal conditional image synthesis with product-of-experts gans. *In European Conference on Computer Vision*, pages 91–109, 2022. Springer, Cham.

[27] P. Isola, JY. Zhu, T. Zhou, and AA. Efros. Image-to-image translation with conditional adversarial networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[28] A. Kazerouni, EK. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint*, 2022. `https://arxiv.org/abs/2211.07804`.

[29] DP. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems, 31*, 2018.

[30] DP. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint*, 2013. `https://arxiv.org/abs/1312.6114`.

[31] A. Lin and AX. Lu. Incorporating knowledge of plates in batch normalization improves generalization of deep learning for microscopy images. *bioRxiv preprint*, 2022. `https://doi.org/10.1101/2022.10.14.512286`.

[32] A. Radford, JW. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. Learning transferable visual models from natural language supervision. *In International conference on machine learning*, pages 8748–8763, 2021.

[33] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. `https://arxiv.org/abs/2204.06125`.

[34] C. Saharia, W. Chan, H. Chang, CA. Lee, J. Ho, T Salimans, DJ. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint*, 2021. `https://arxiv.org/abs/2111.05826`.

[35] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, SKS. Ghasemipour, BK. Ayan, SS. Mahdavi, RG. Lopes, T. Salimans, J. Ho, DJ. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint*, 2022. `https://doi.org/10.48550/arxiv.2205.11487`.

[36] C. Saharia, J. Ho, W. Chan, T. Salimans, DJ. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *arXiv preprint*, 2021. `https://arxiv.org/abs/2104.07636`.

[37] P. Sanchez, A. Kascenas, X. Liu, AQ. O'Neil, and SA. Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. *arXiv preprint*, 2022. `https://arxiv.org/abs/2207.12268`.

[38] Y. Shi, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems, 32*, 2019.

[39] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *In International Conference on Machine Learning. PMLR*, pages 2256–2265, 2015.

[40] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint*, 2021.

[41] Y. Song, L. Shen, L. Xing, and S. Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint*, 2021. `https://arxiv.org/abs/2111.08005`.

[42] Y. Song, J. Sohl-Dickstein, DP. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*, 2020. `https://arxiv.org/abs/2011.13456`.

[43] H. Sun, R. Mehta, HH. Zhou, Z. Huang, SC. Johnson, V. Prabhakaran, and V. Singh. Dual-glow: Conditional flow-based generative model for modality transfer. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10611–10620, 2019.

[44] M. Sypetkowski, M. Rezanejad, S. Saberian, O. Kraus, J. Urbanik, J. Taylor, B. Mabey, M. Victors, J. Yosinski, AR. Sereshkeh, and I. Haque. Rxrx1: A dataset for evaluating experimental batch correction methods. *arXiv preprint arXiv:2301.05768*, 2023.

[45] S. Wang, M. Lu, N. Moshkov, JC. Caicedo, and BA. Plummer. Anchoring to exemplars for training mixture-of-expert cell embeddings. *arXiv preprint*, 2021. `https://doi.org/10.48550/arxiv.2112.03208`.

[46] J. Wolleb, F. Bieder, R. Sandkühler, and PC. Cattin. Diffusion models for medical anomaly detection. *arXiv preprint*, 2022. `https://arxiv.org/abs/2203.04306`.

[47] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[48] Y. Wu and K. He. Group normalization. *arXiv preprint*, 2018. `https://doi.org/10.48550/arxiv.1803.08494`.

[49] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint*, 2022. `https://arxiv.org/abs/2209.00796`.

[50] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[51] F. Zhan, Y. Yu, R. Wu, J. Zhang, and S. Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint*, 2022. `https://arxiv.org/abs/2112.13592`.

[52] H. Zhang, C. Fang, X. Xie, Y. Yang, W. Mei, D. Jin, and P. Fei. High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. *Biomedical optics express*, 3(10):1044–1063, 2019.

[53] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.