An Introduction to Disentanglement

Paul Vicol

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Outline

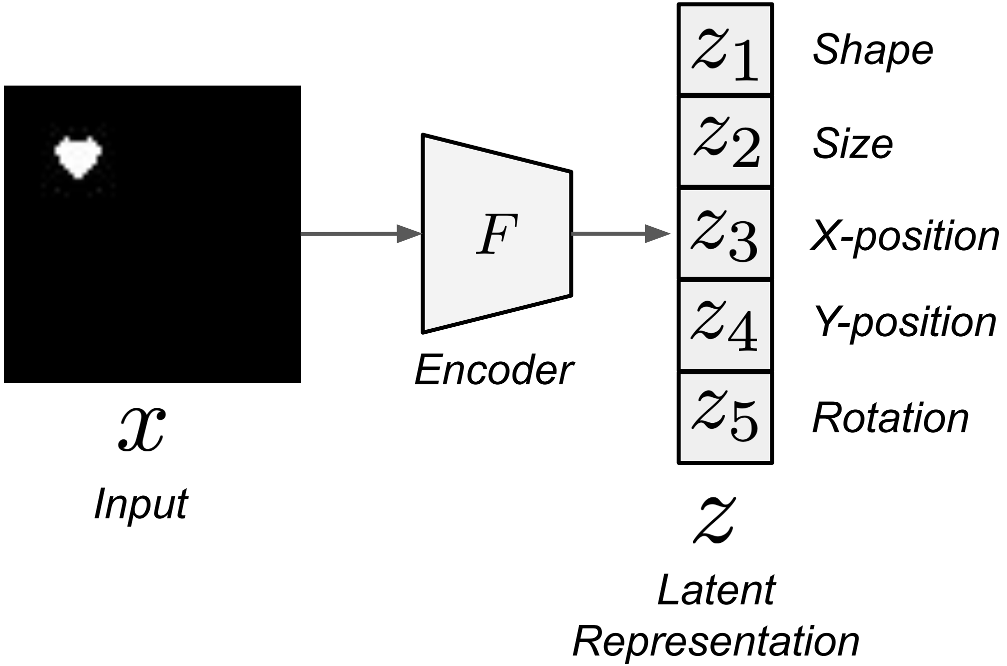
1. What are disentangled representations?

2. Why are disentangled representations *useful*?
    - Robustness on out-of-distribution data (domain adaptation & domain generalization)
    - Fairness
    - Interpretability
    - Controllable generative modelling

3. How can we *learn* disentangled representations?
    - Supervised & unsupervised
    - VAEs and friends (β-VAE, β-TCVAE, FactorVAE)
    - A few domain adaptation methods

Disentangled Representations

- A *disentangled representation* is one in which different factors of variation are represented by *different components of the representation*
  - e.g., different dimensions in the latent space



What are disentangled representations good for?

Robustness to Distribution Shifts

- We want to learn classifiers that *generalize to new domains*

Source Domains (Paintings & Sketches)

Target Domain (Real Images)

- **Approaches typically fall into two categories:**
  1) Ones that *discard domain information* from the learned representation
  2) Ones that *preserve information about both domain and class, using disentangled latent subspaces*

- **Domain adaptation** learns representations from source domains that transfer to a *specific, known target domain*
- **Domain generalization** learns representations from source domains, that can be transferred to *previously unseen domains at test time*

Fairness

- Automated systems are increasingly used to make *decisions that impact people's lives*

Healthcare

Bank Loans

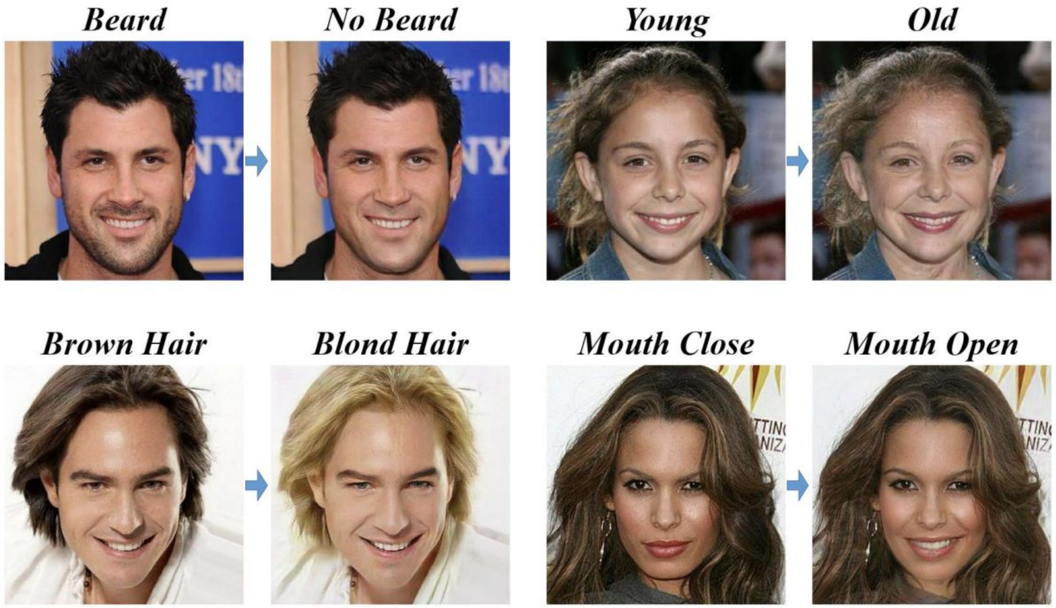
Insurance

Targeted Ads

- In order to make fair decisions, the algorithm should not depend on certain *sensitive attributes, e.g., race or gender*
- We do not want our models to perpetuate biases present in the dataset (e.g., due to historical discrimination/unfair treatment)
- We wish to *purge information about the sensitive attributes from the learned representation*

Controllable Generative Modeling

- If a representation z is disentangled, we can modify one dimension to change a single attribute, yielding a meaningful modified representation $\tilde{z}$
- This can allow for *controllable generative modeling*

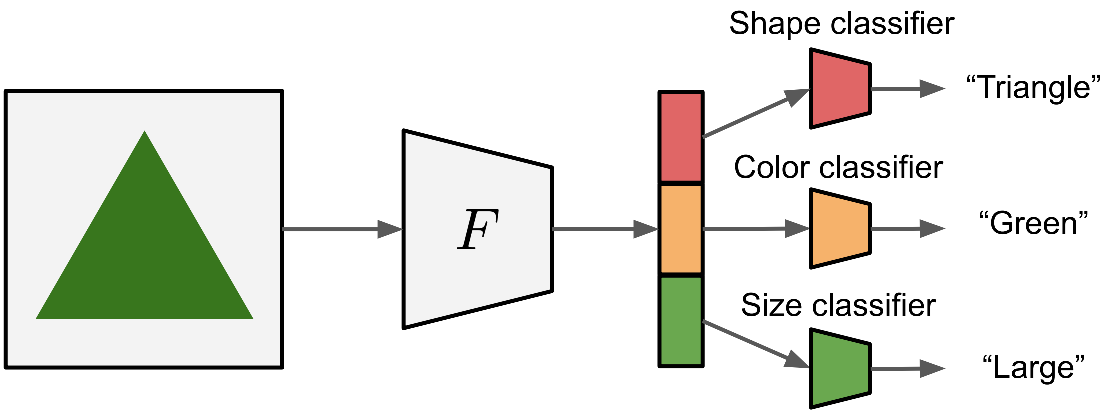


He et al.. “AttGAN.” 2017.

How can we learn disentangled representations?

Disentangling with Supervision

- Given *full supervision for the values of attributes*, you could train *classifiers on each latent subspace*
  - This would enforce that each subspace contains information about a specific attribute



- However, this *does not prevent* the encoder from simply encoding all attributes in *each* latent subspace
  - Need to explicitly enforce *independence between subspaces*

Mutual Information

- *Mutual information (MI)* measures the *statistical dependence between random variables*

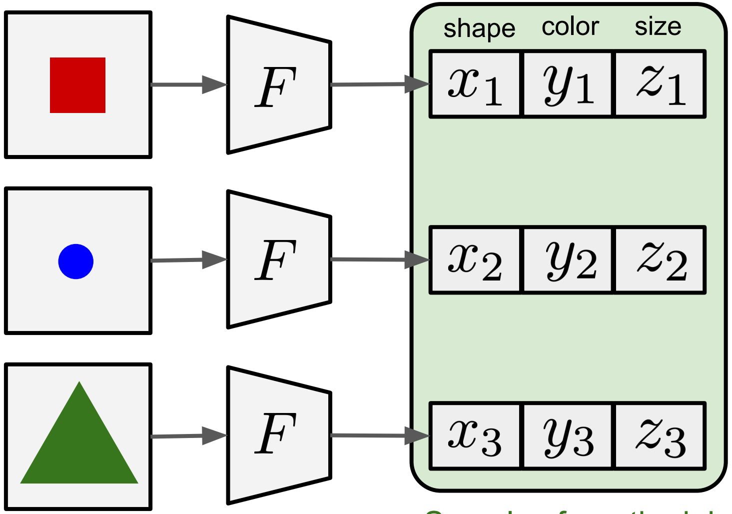
$$I(x;y) = D_{\mathrm{KL}}\underbrace{[p(x,y)||p(x)p(y)]}_{\text{The divergence between the joint distribution and the product of marginal distributions}}$$

- Recall that if x and y are *independent*, then $p(x,y) = p(x)p(y)$ and thus $I(x;y) = 0$
- *MI minimization* is at the heart of many approaches to disentanglement
- *Total correlation (TC)* is a generalization of MI between *multiple random variables*

  $C(x_1, \ldots, x_n) = D_{\mathrm{KL}}[p(x_1, \ldots, x_n)||p(x_1)\cdots p(x_n)]$

A Generic Way to Minimize Mutual Information

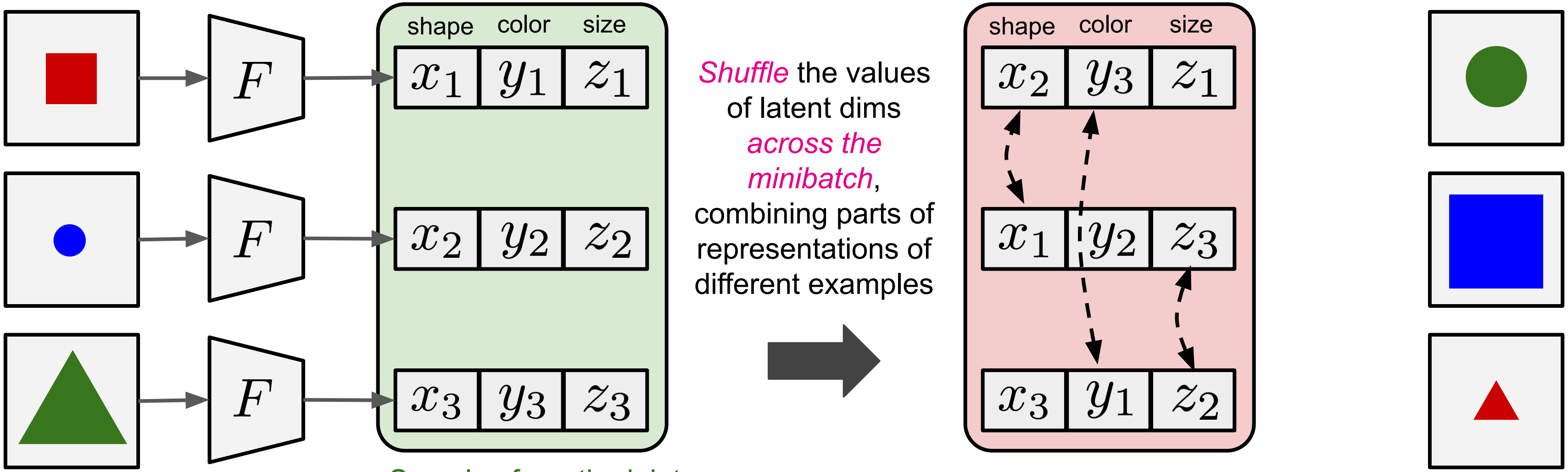
- To minimize $I(x;y)$ we want the distributions $p(x,y)$ and $p(x)p(y)$ to be close
  - Can be done using many *distribution alignment/matching techniques*



Samples from the joint distribution $p(x, y, z)$

A Generic Way to Minimize Mutual Information

- To minimize $I(x;y)$ we want the distributions $p(x,y)$ and $p(x)p(y)$ to be close
  - Can be done using many *distribution alignment/matching techniques*

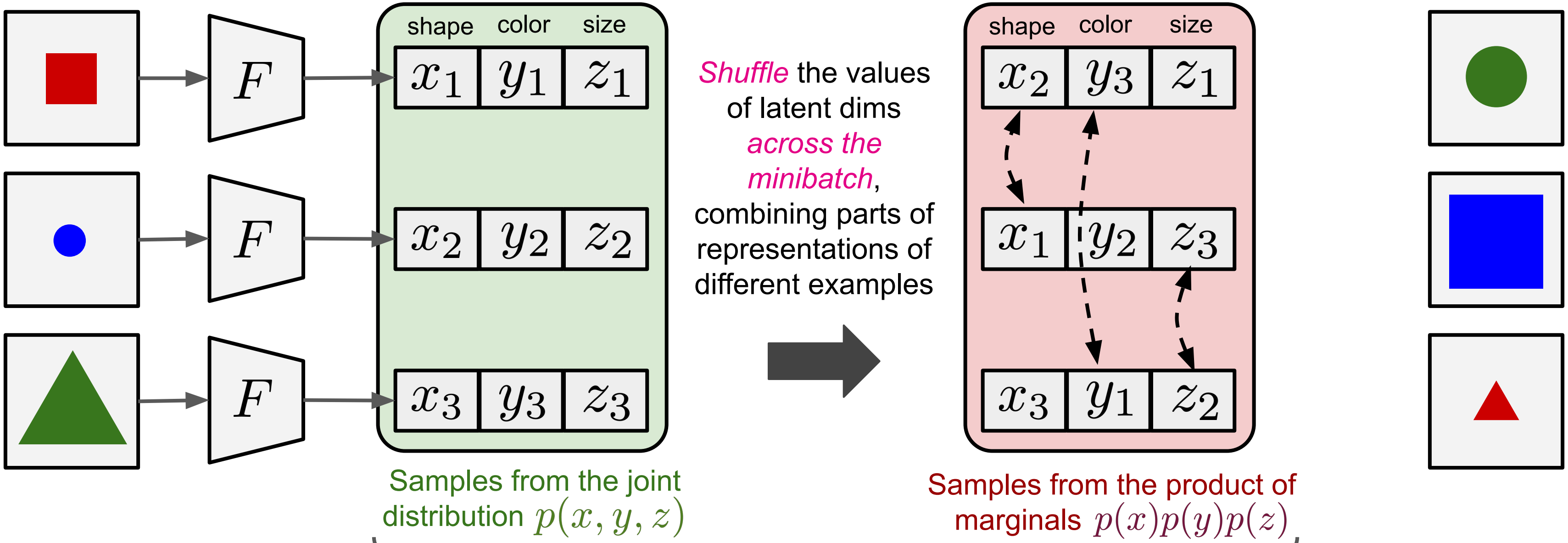


Samples from the joint distribution $p(x,y,z)$

Samples from the product of marginals $p(x)p(y)p(z)$

A Generic Way to Minimize Mutual Information

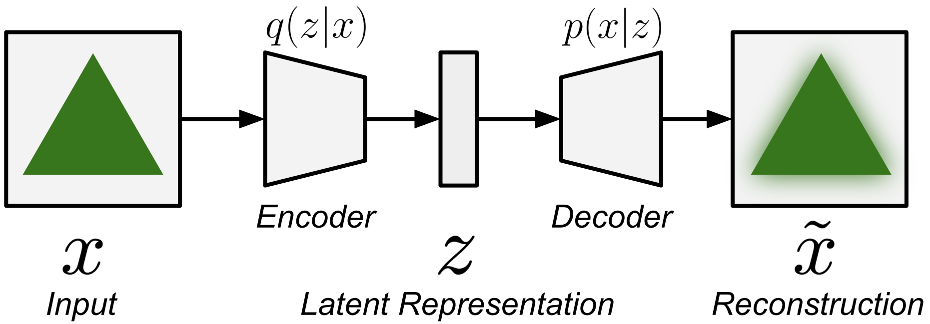
- To minimize $I(x;y)$ we want the distributions $p(x,y)$ and $p(x)p(y)$ to be close
  - Can be done using many *distribution alignment/matching techniques*



Train a *discriminator* to distinguish between samples from these distributions and train the encoder *adversarially*

Unsupervised Disentanglement: VAEs

- In *unsupervised disentanglement*, we only have samples from the data distribution *without access to the true factors of variation*
- Variational autoencoders (VAEs) are *unsupervised, latent-variable generative models*



- Trained by *maximizing the evidence lower bound (ELBO)*, which is a lower bound on the marginal likelihood $p(x) = \int p(x, z)dz$

ELBO: $$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q(z|x^{(i)})}[\log p(x^{(i)}|z)] - KL[q(z|x^{(i)})||p(z)]\right]$$

β-VAE, β-TCVAE, and FactorVAE

- β-VAE upweights the KL divergence term with β > 1:

Modified ELBO: $$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q(z|x^{(i)})}[\log p(x^{(i)}|z)] - \beta KL[q(z|x^{(i)})||p(z)]\right]$$

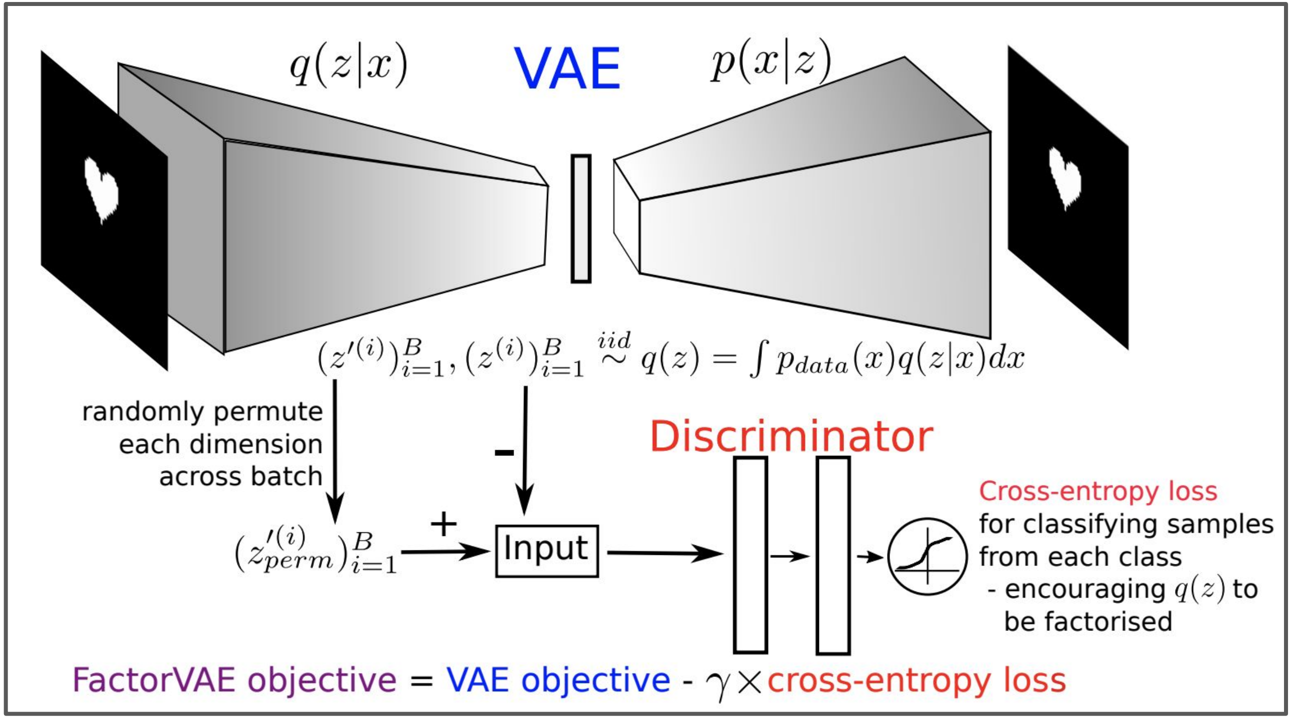
- β-VAE has a *trade-off* between reconstruction quality and disentanglement
- This is due to a *problem hidden within the KL term* of the ELBO

- The KL term can be decomposed as:

$$\mathbb{E}_{p_{data}(x)}[KL(q(z|x)||p(z))] = \underbrace{I(x;z)}_{\text{Penalizing this reduces the amount of info about x stored in z which leads to poor recons.}} + \underbrace{KL(q(z)||p(z))}_{\text{Encourages independence in the dimensions of z, by matching the prior}}$$

β-VAE, β-TCVAE, and FactorVAE

- FactorVAE combines the standard ELBO with an adversarial term *minimizing the total correlation between latent dimensions*



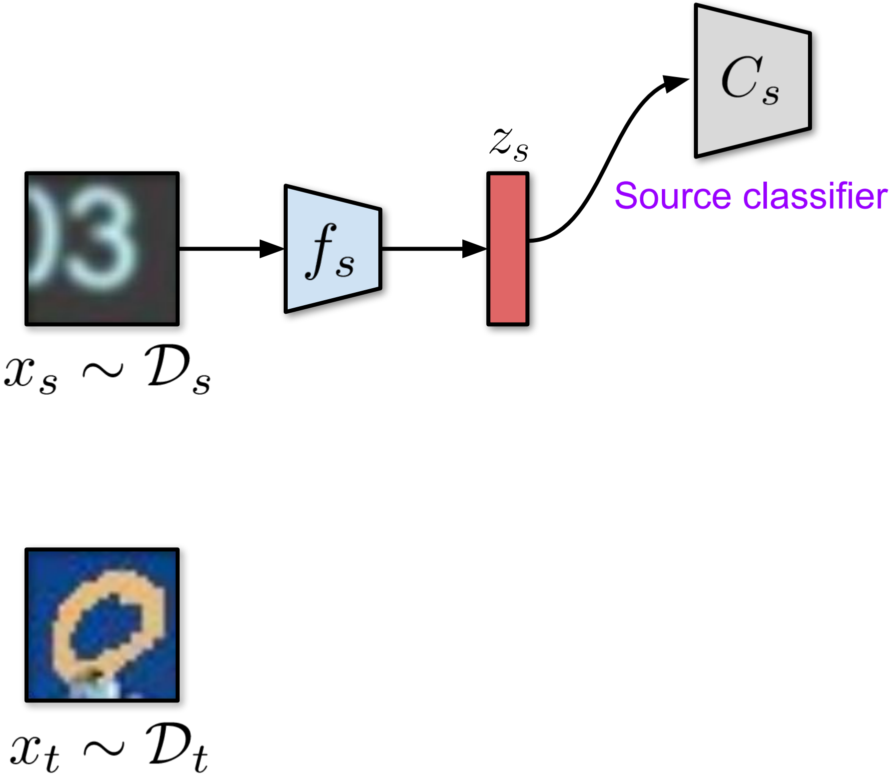
Kim & Mnih. “Disentangling by Factorizing.” ICML 2018.

Adversarial Discriminative Domain Adaptation (ADDA)

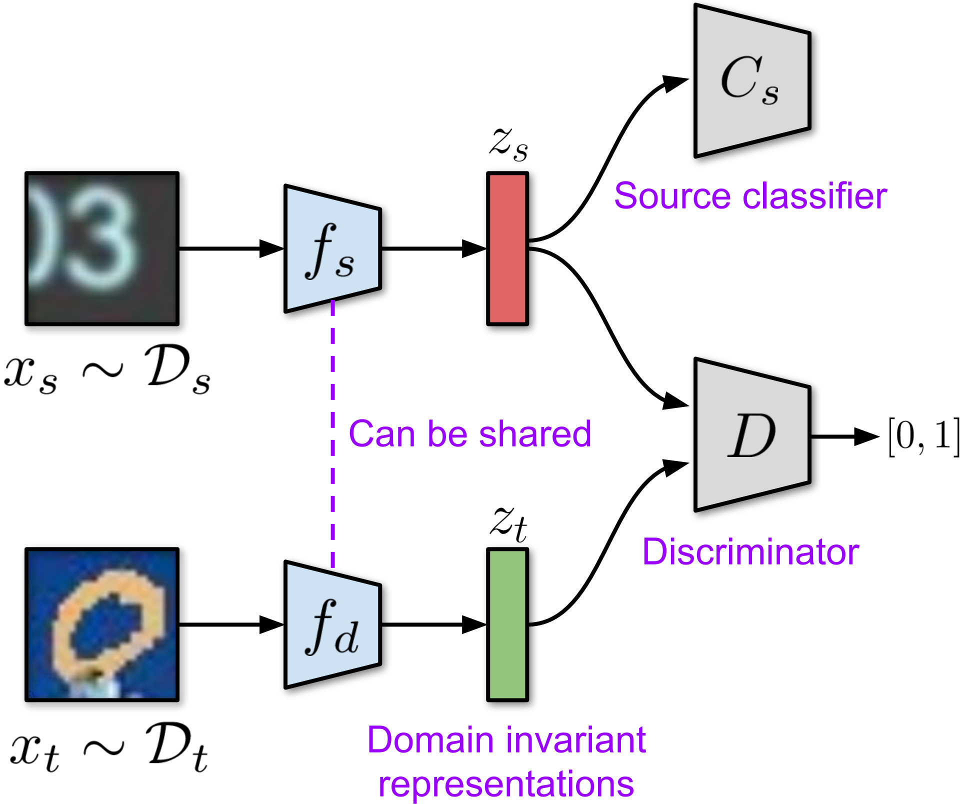
$x_s \sim \mathcal{D}_s$

$x_t \sim \mathcal{D}_t$

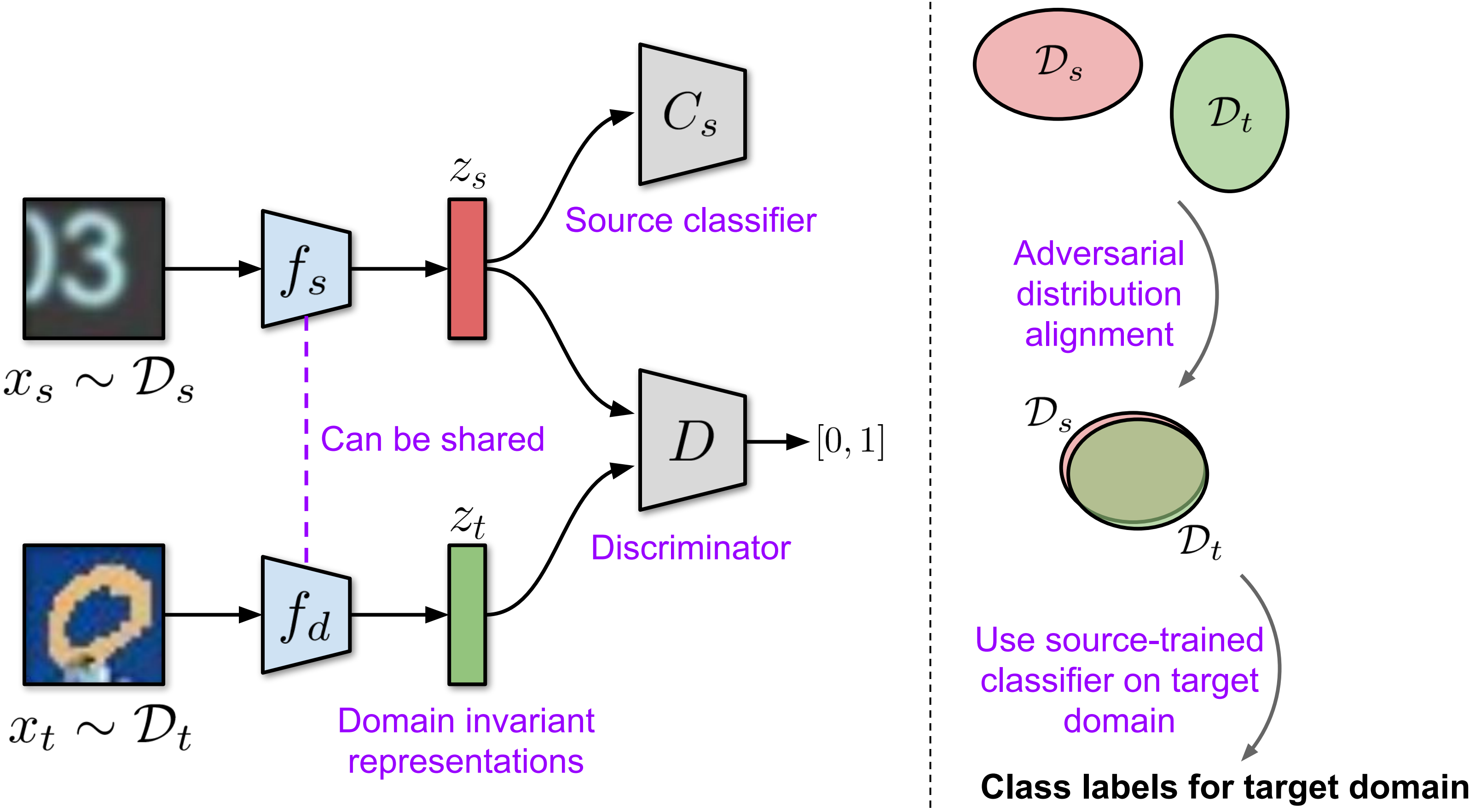
Adversarial Discriminative Domain Adaptation (ADDA)



Adversarial Discriminative Domain Adaptation (ADDA)

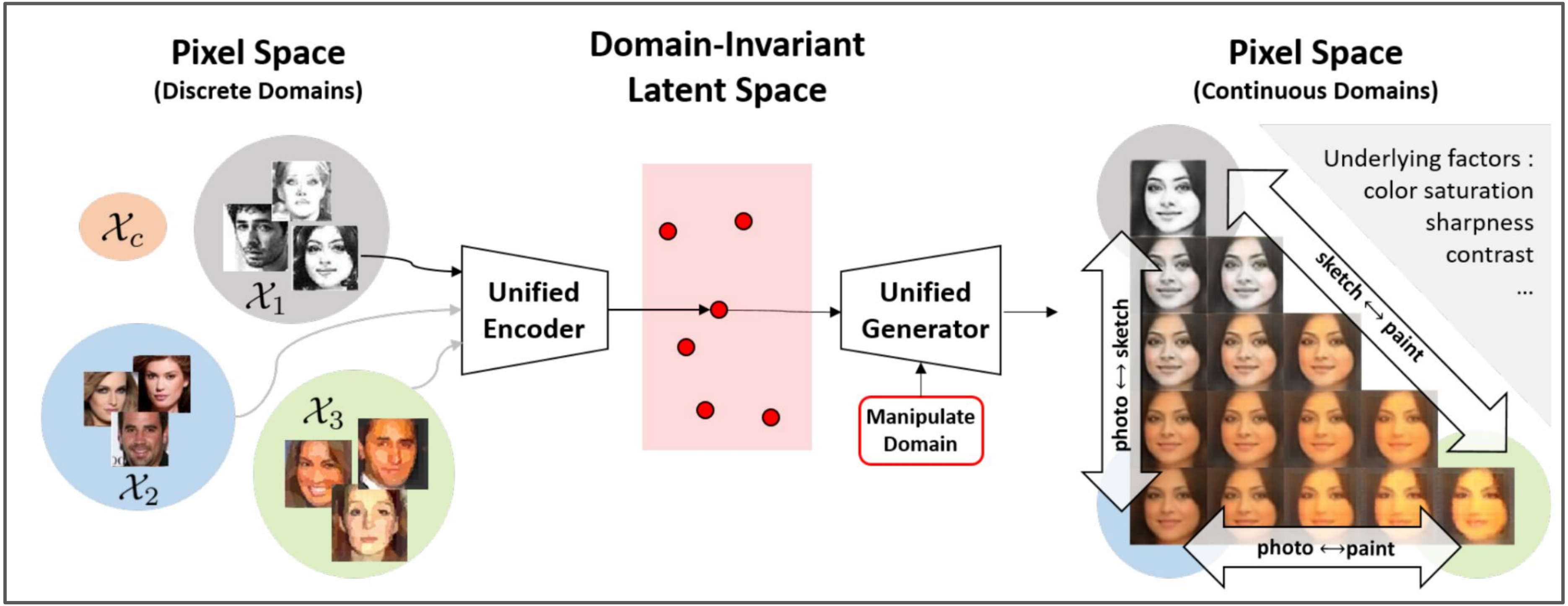


Adversarial Discriminative Domain Adaptation (ADDA)



Unified Feature Disentangler Network (UFDN)

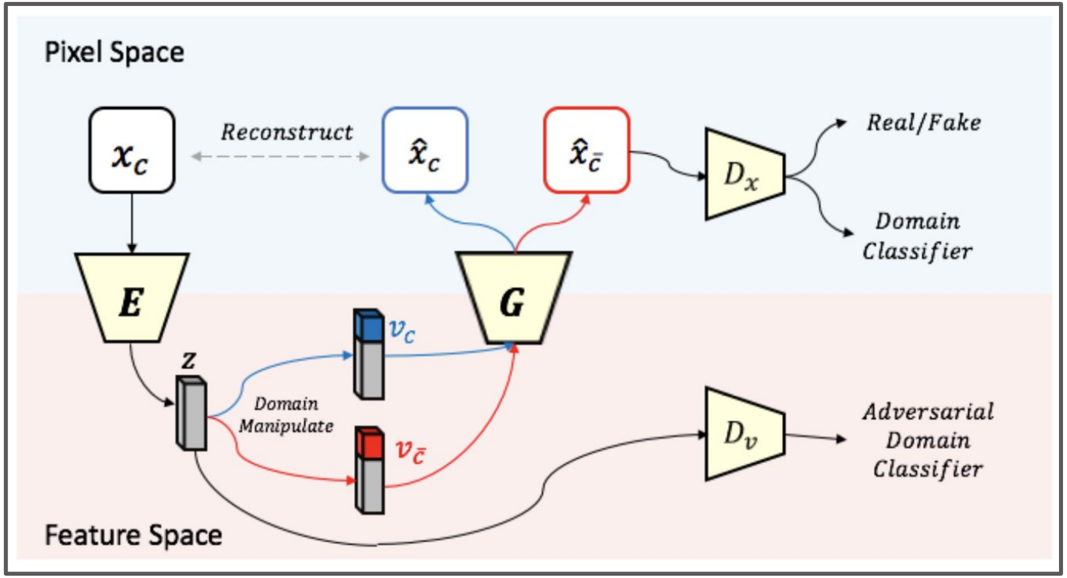
- Allows for explicit control over the domain; can interpolate between different domains



Liu et al., “A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation.” NeurIPS 2018.

Unified Feature Disentangler Network (UFDN)

- Allows for explicit control over the domain; can interpolate between different domains

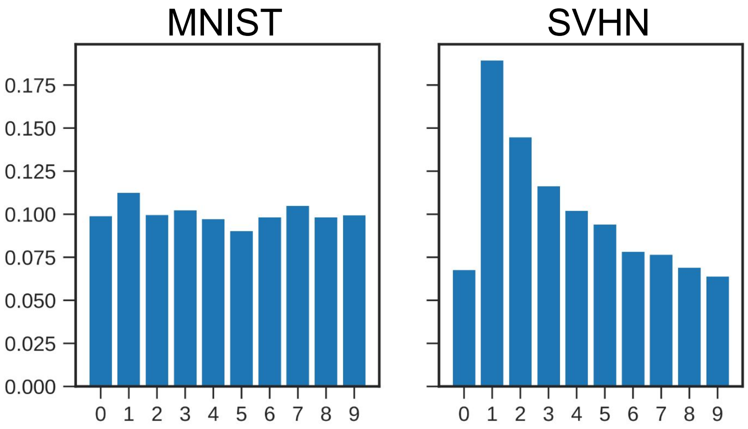


Liu et al., "A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation." NeurIPS 2018.

Correlations Between Factors

- Most work *assumes that the ground-truth factors of variation are independent*
  - That is, that there are *no correlations between attributes*
  - This holds for simple/synthetic benchmark tasks (e.g., dSprites, Shapes3D)
- But this *real data often has correlations* between attributes, breaking this assumption

Correlation Between Class & Domain



Correlation Between Foreground & Background

Q/A