Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation for Image Synthesis and Classification

*Tzu-Chien Fu[1], *Yen-Cheng Liu[2], Wei-Chen Chiu[1,3], and Y.-C. Frank Wang[1,2]

[1]Research Center for IT Innovation, Academia Sinica

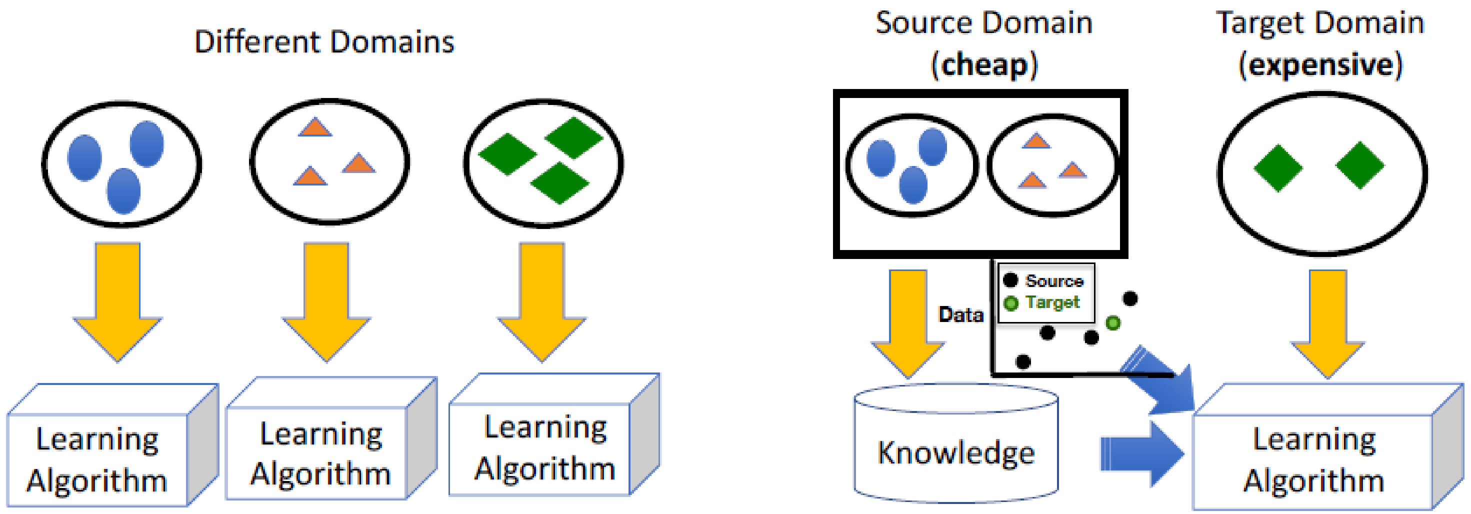
[2]Dept. EE, National Taiwan University

[3]Dept. CS, National Chiao Tung University

(* indicates equal contributions)

(Traditional) Machine Learning vs. Transfer Learning

- Transfer Learning
  - Collecting/annotating data is typically expensive.
  - Improved learning & understanding in the target domain by leveraging knowledge from the source domain

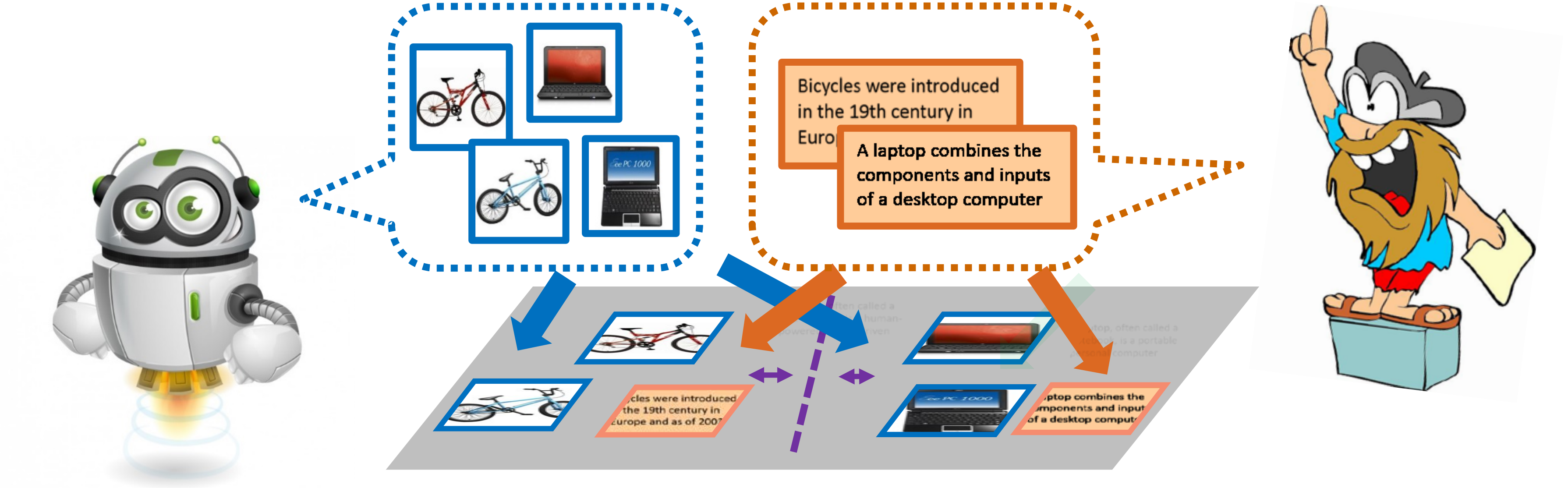


Research Focuses

- Transfer Learning for
  - Homogeneous/heterogeneous domain adaptation
  - Multi-label classification / zero-shot learning
  - Robust face recognition (e.g., cross-resolution, cross-modality, etc.)

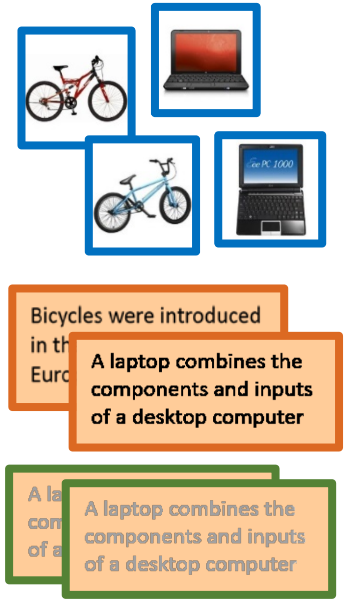
Heterogeneous Domain Adaptation

- Deep Transfer Learning for Cross-Domain Data Classification
  - Learning from source & target-domain data described by distinct types of features



Heterogeneous Domain Adaptation (cont'd)

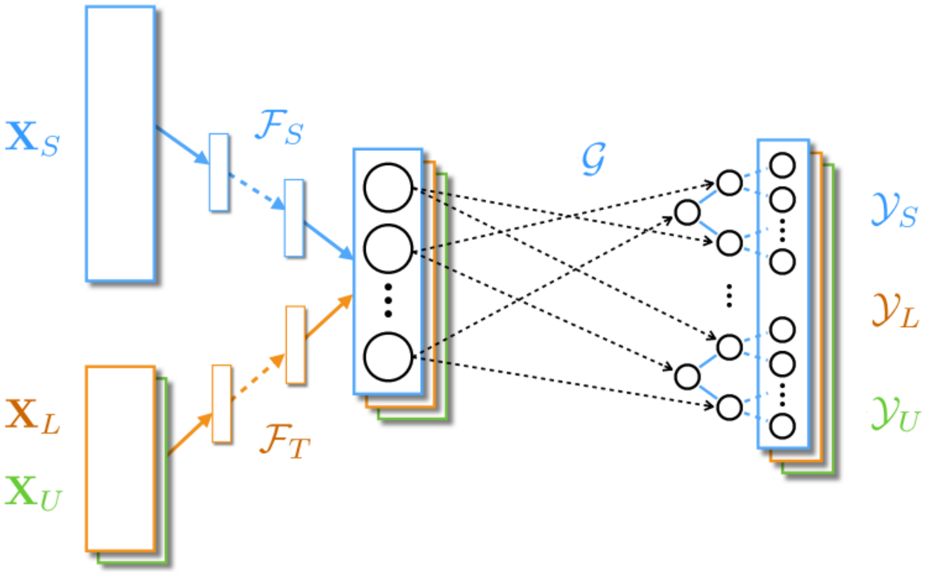
- Transfer Neural Trees (TNT)
  - Joint learning of cross-domain mapping F_S/F_T & cl. layer G (deep neural decision forest)
  - Propose stochastic pruning for G to avoid overfitting source-domain labeled data
  - Unique embedding loss for learning target-domain data in a *semi-supervised* setting



Source-domain labeled data

Target-domain labeled data

Target-domain unlabeled data



Y.-C. F. Wang et al., "Transfer Neural Trees for Heterogeneous Domain Adaptation," ECCV, 2016.

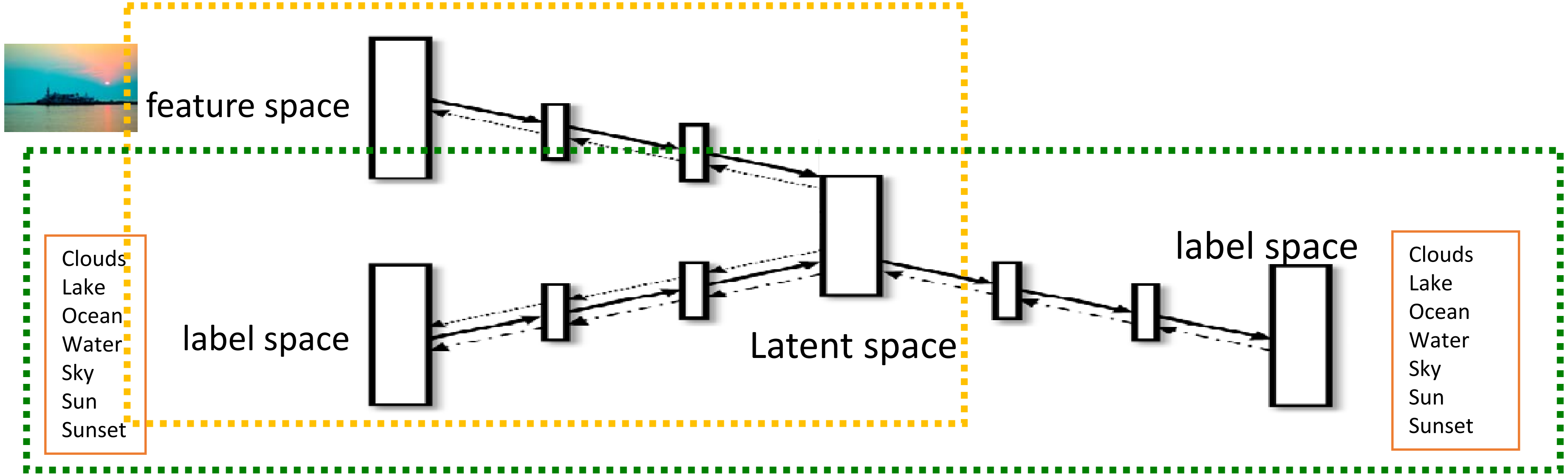
Multi-Label Classification

- Predicting multiple labels w/o using annotated ground truth info (e.g., bounding box)
- Learning across image and label-domain data + exploit label co-occurrences

Labels:
Person
Table
Sofa
Chair
TV
Lights
Carpet
…

Multi-Label Classification (cont'd)

- ***Canonical Correlated AutoEncoder (C2AE)*** [AAAI'17]
  - Unique integration of autoencoder & deep canonical correlation analysis (DCCA)
  - Autoencoder in C2AE: label embedding + label recovery + label co-occurrence
  - DCCA in C2AE: joint feature & label embedding



Y.-C. F. Wang et al., Learning Deep Latent Spaces for Multi-Label Classification, AAAI 2017

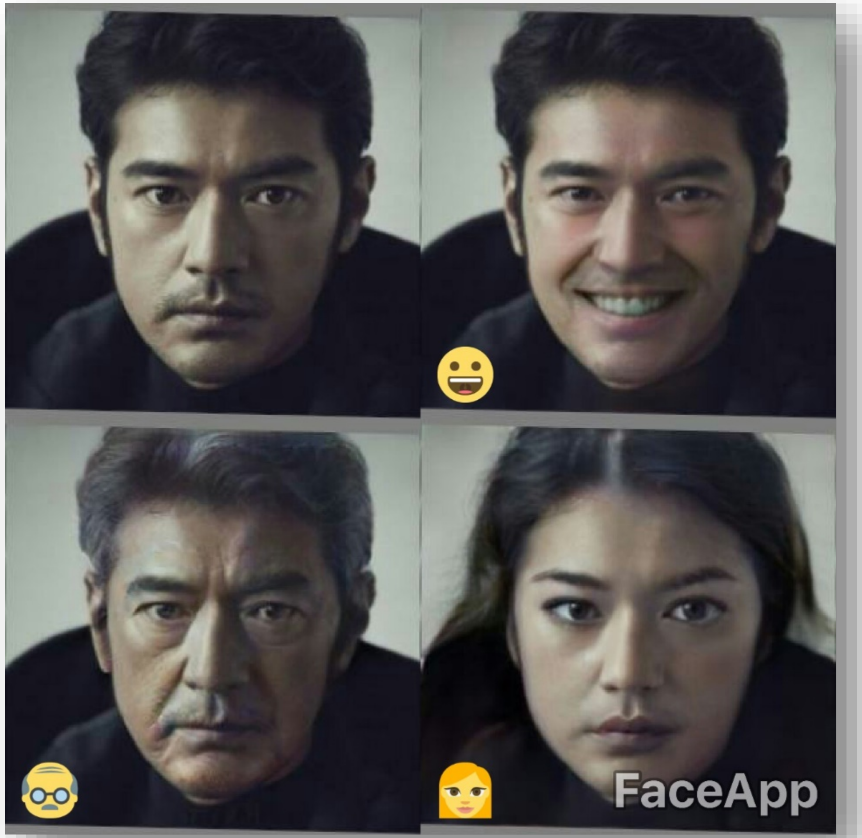
Research Focuses

- Transfer Learning for
  - Domain adaptation
    - Cross-domain image synthesis/translation/classification
  - Multi-label classification / zero-shot learning
  - Robust face recognition (e.g., cross resolution, etc.)

FaceApp

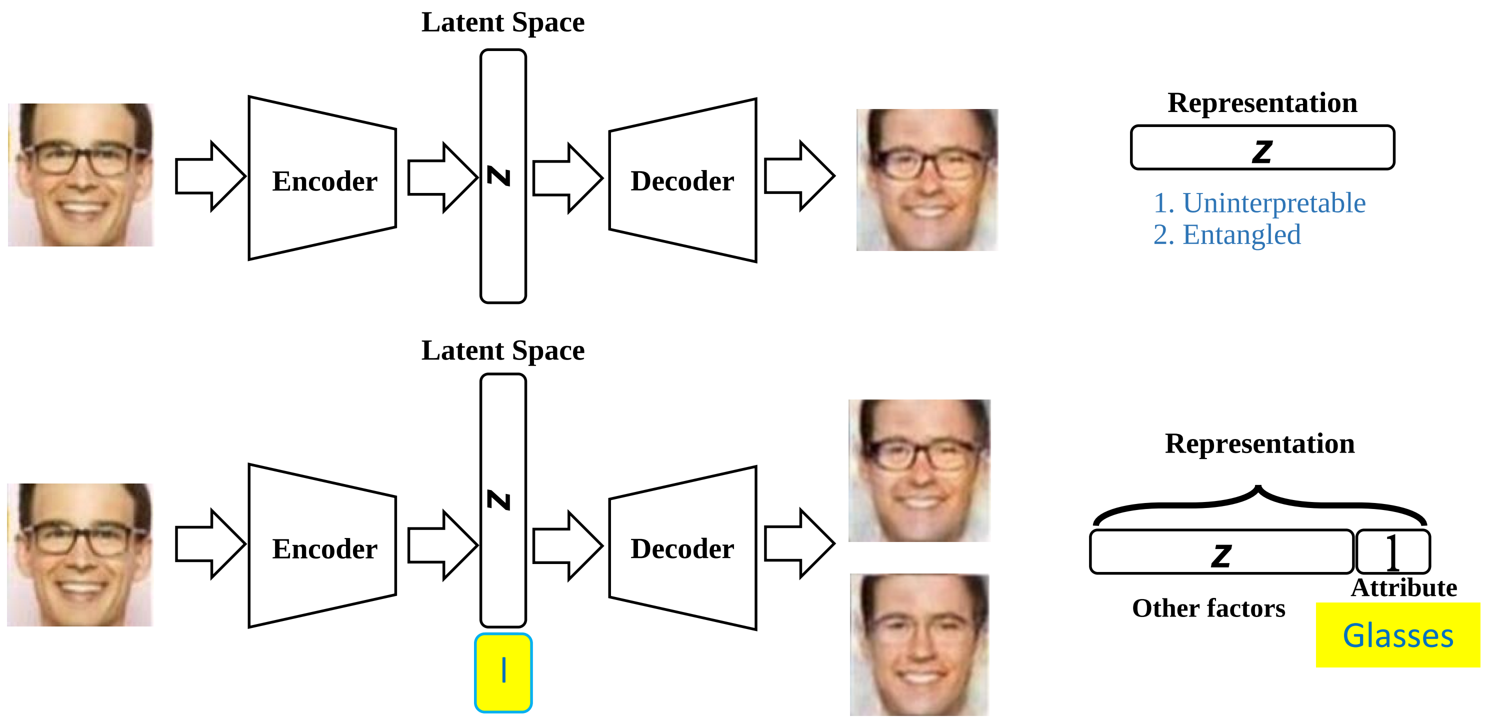
- Beyond putting a smile on your face
- Over 10M downloads

Input



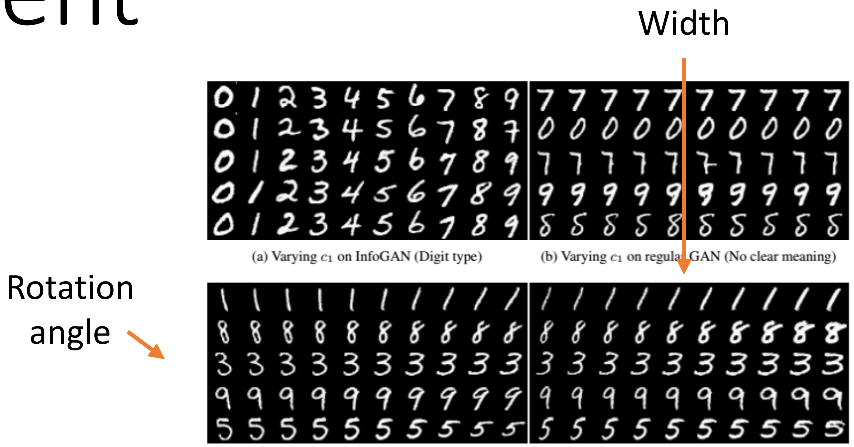
Introduction

- Feature Disentanglement:
  - Learn a latent space which factorizes the representation ***z*** into different parts (i.e., attributes) for describing the corresponding info (e.g., identity, pose, or expression of facial images).



Settings for Feature Disentanglement

- Unsupervised Learning
  - Disentangling images *without* observing attribute info
  - *No guarantee* in disentangling particular semantics
- Supervised Learning
  - With supervision of image labels, disentangle the associated factor from feature representation
  - Can manipulate the output image with label/attribute of interest accordingly.
- Ours: *Cross-Domain Feature Disentanglement*
  - Source-domain training data: existing annotated instances
  - Target-domain data: no ground truth info, to be adapted/manipulated
  - Can be viewed as either semi-supervised learning, or unsupervised domain adaptation

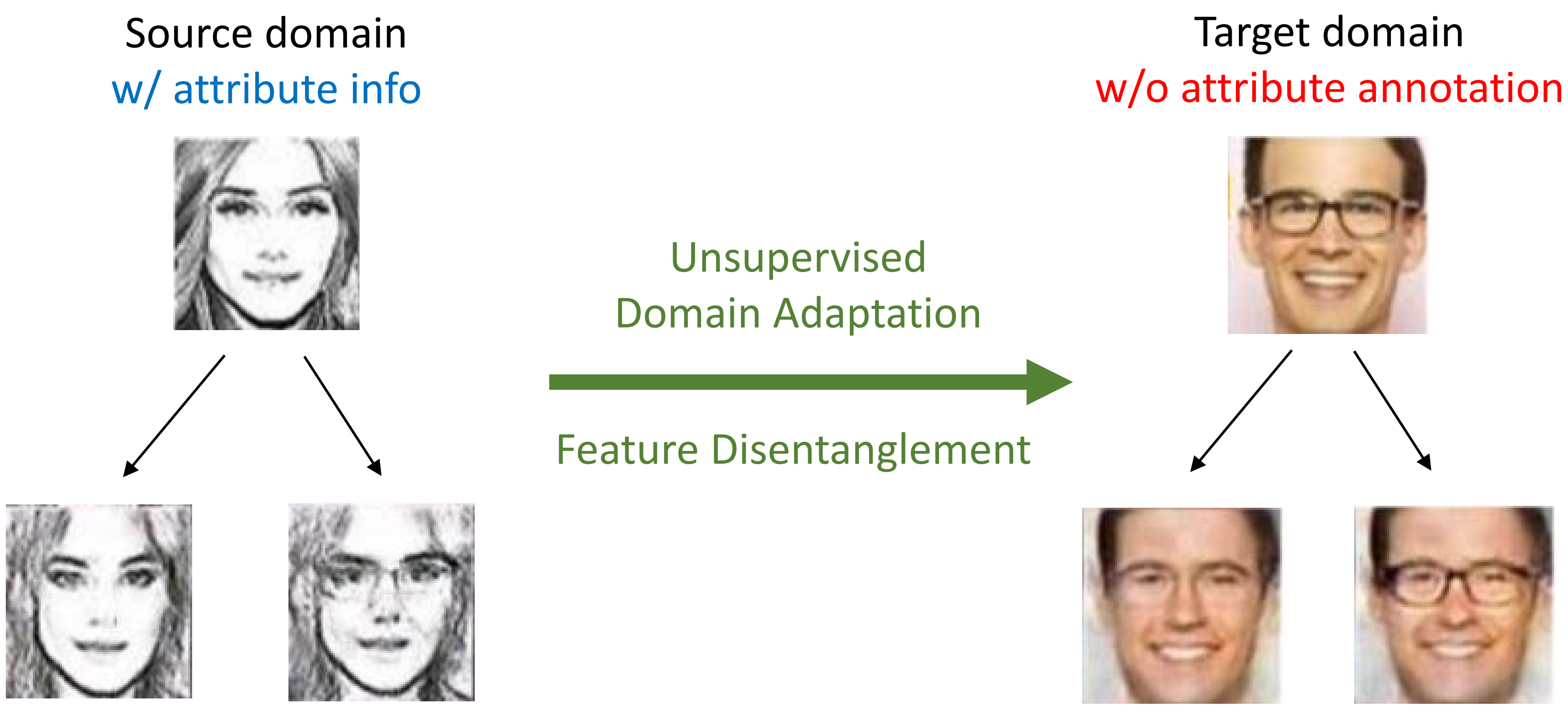


(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation) (d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Our Goal

- A unified framework for cross-domain feature disentanglement, with only attribute supervision from the source domain.



Related Works

- Feature Disentanglement
  - Unsupervised: InfoGAN [1]
  - Supervised: AC-GAN [2]
- Unsupervised Cross-Domain Image Synthesis/Translation
  - Image synthesis: CoGAN [3]
  - Image translation: UNIT [4]

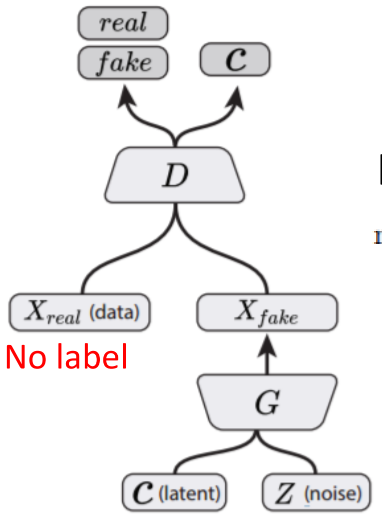
[1] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. Advances in Neural Information Processing Systems (NIPS), 2016.

[2] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. arXiv, 2016.

[3] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. Advances in Neural Information Processing Systems (NIPS), 2016

[4] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. arXiv, 2017.

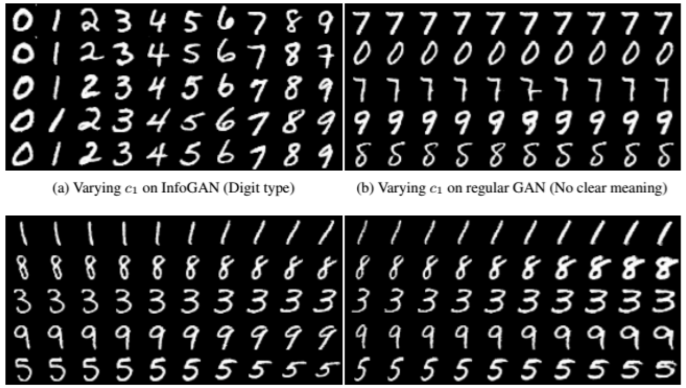
InfoGAN & AC-GAN (Unsup/Sup. Feature disentanglement)



No label

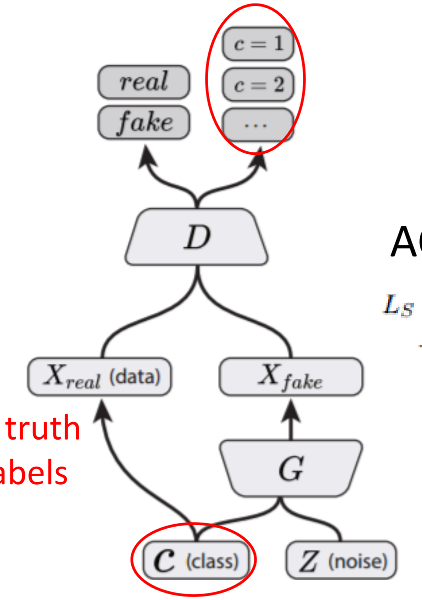
InfoGAN (unsupervised)

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$



(a) Varying c_1 on InfoGAN (Digit type)   (b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from −2 to 2 on InfoGAN (Rotation)   (d) Varying c_3 from −2 to 2 on InfoGAN (Width)



w/ ground truth attribute labels

AC-GAN (supervised)

$$L_S = E[\log P(S = real \mid X_{real})] + E[\log P(S = fake \mid X_{fake})]$$
$$L_C = \underline{E[\log P(C = c \mid X_{real})]} + E[\log P(C = c \mid X_{fake})]$$

monarch butterfly   goldfinch   daisy   redshank   grey whale

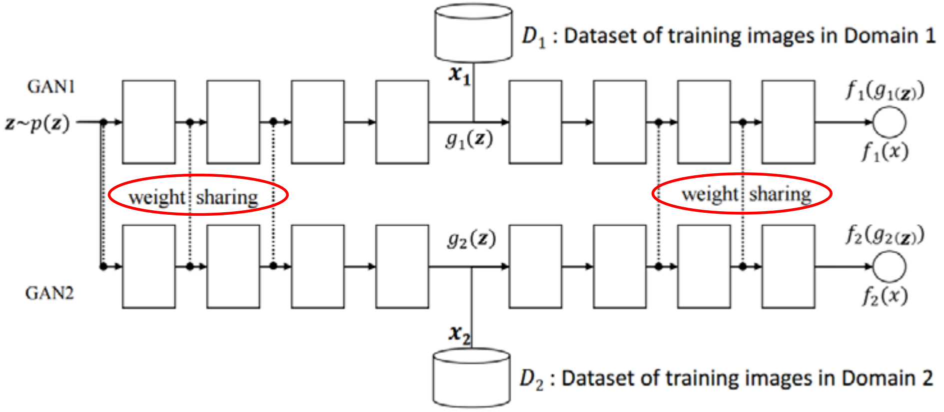
[1] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), 2016.

[2] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. arXiv preprint arXiv:1610.09585, 2016.

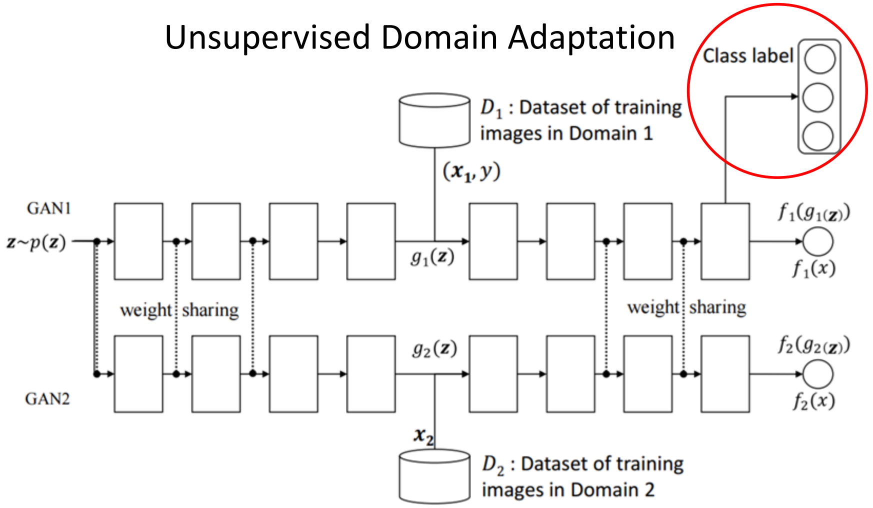
CoGAN (Unsupervised Cross-Domain Image Synthesis)

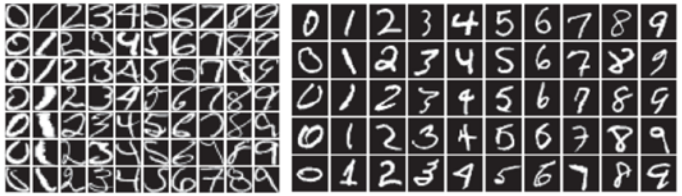
- Synthesize pairs of corresponding images
- Enforce weight-sharing constraints in high-level layers

Unsupervised Cross-Domain Image Synthesis



Unsupervised Domain Adaptation





(a) USPS (b) MNIST

Task \ Method	[18]	[19]	[20]	[21]	CoGAN
MNIST→USPS	0.408	0.467	0.478	0.607	**0.912** ±0.008
USPS→MNIST	0.274	0.355	0.631	0.673	**0.891** ±0.008
Average	0.341	0.411	0.554	0.640	**0.902**

[3] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In Advances in Neural Information Processing Systems (NIPS), 2016

UNIT (Unsupervised Cross-domain Image Translation)

UNIT learns translation functions of mapping an image in one domain to another without any corresponding images across domains.

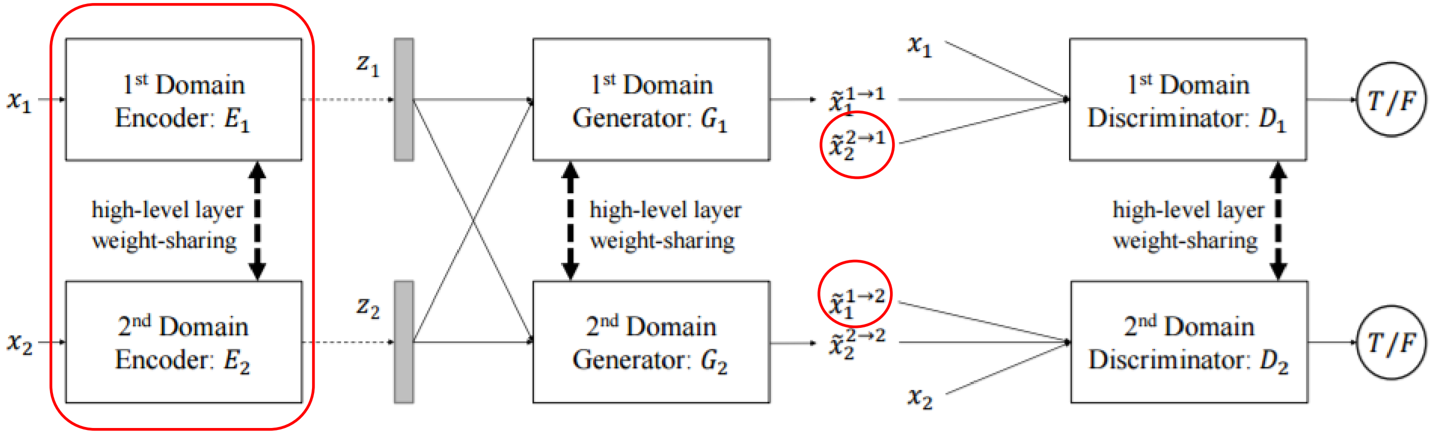


Table 2: UDA results on adapting from the SVHN domain to the MNIST domain. The results of the other algorithms were duplicated from (Taigman et al., 2017)

Method	Accuracy
SA (Fernando et al., 2013)	59.32%
DANN (Ganin et al., 2016)	73.85%
DTN (Taigman et al., 2017)	84.88%
UNIT (proposed)	**90.53%**

[4] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. arXiv, 2017.

Proposed Method - Cross-Domain Disentanglement (CDD)

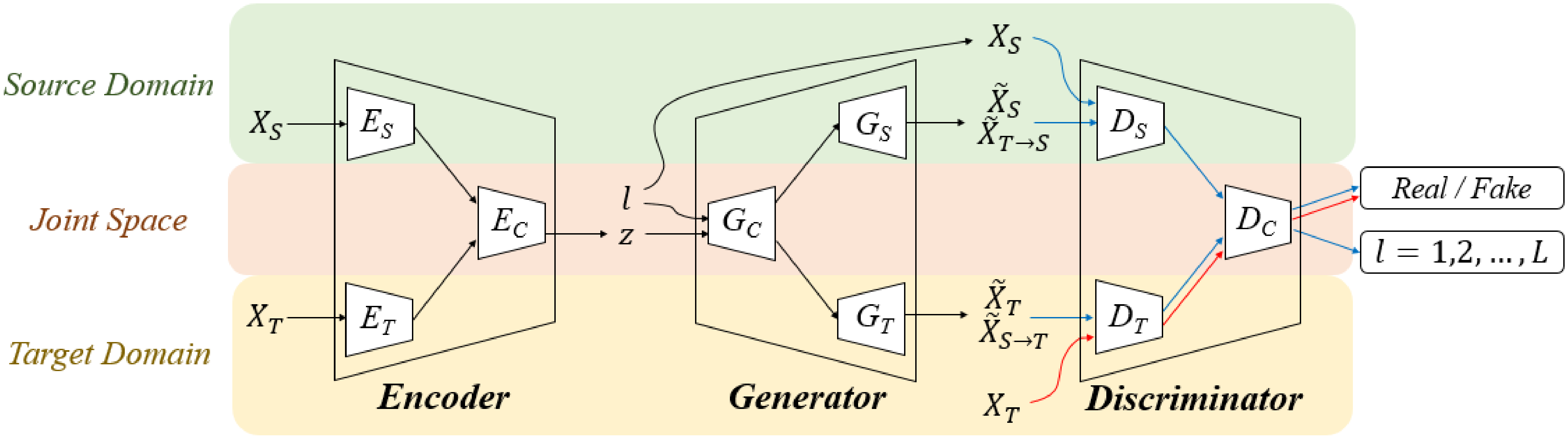
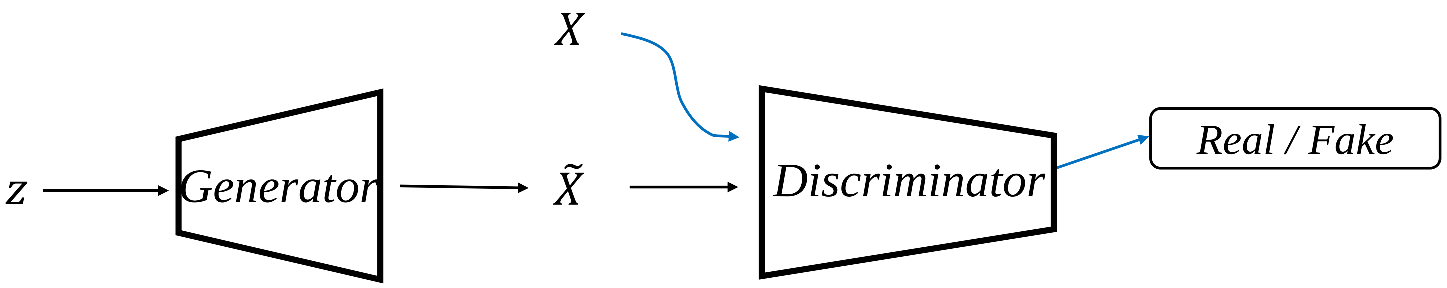


Figure: Overview of our proposed method

Proposed Method - Cross-Domain Disentanglement (CDD)

Generative Adversarial Network (GAN)



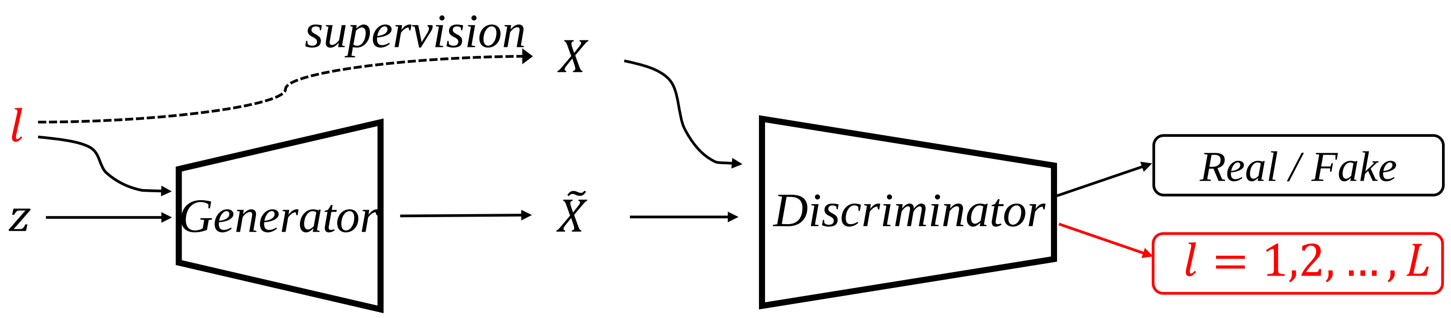
- ✓ *Synthesize* realistic images

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(\tilde{X}))] + \mathbb{E}[\log(D(X))]$$

Synthesized images $\tilde{X}$ **Real images** X

Proposed Method

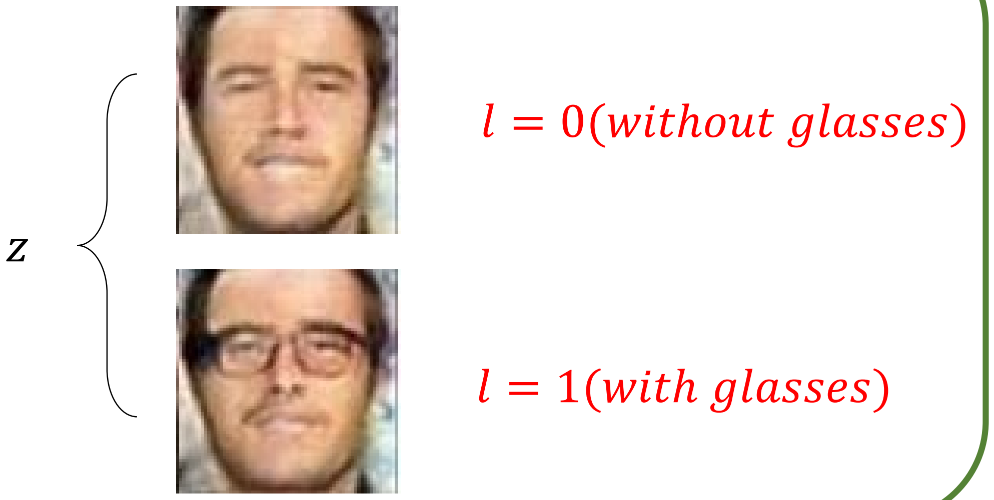
AuxiliaryClassifier-GAN (AC-GAN)



- ✓ *Synthesize* images conditioned on disentangled factor l
- ✓ Disentangle the specific factor l from the representation z

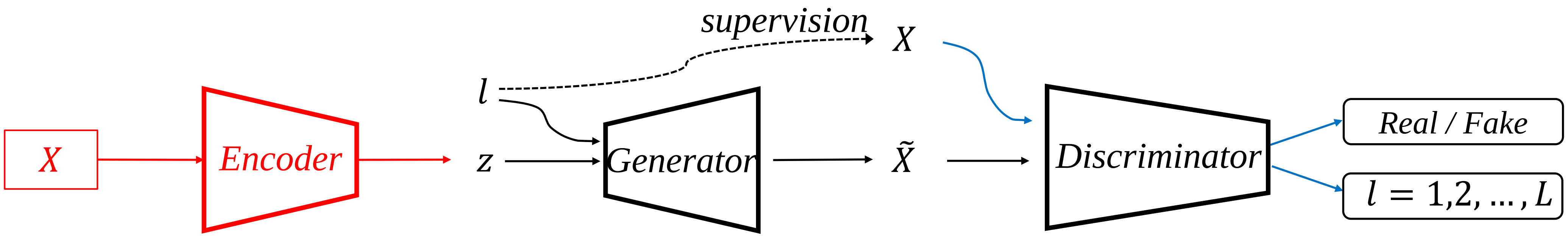
$$\mathcal{L}_{dis} = \mathbb{E}[\log(L = l|X\ )] + \mathbb{E}[\log(L = l|\tilde{X}\ )]$$

Synthesized images $\tilde{X}$



Proposed Method

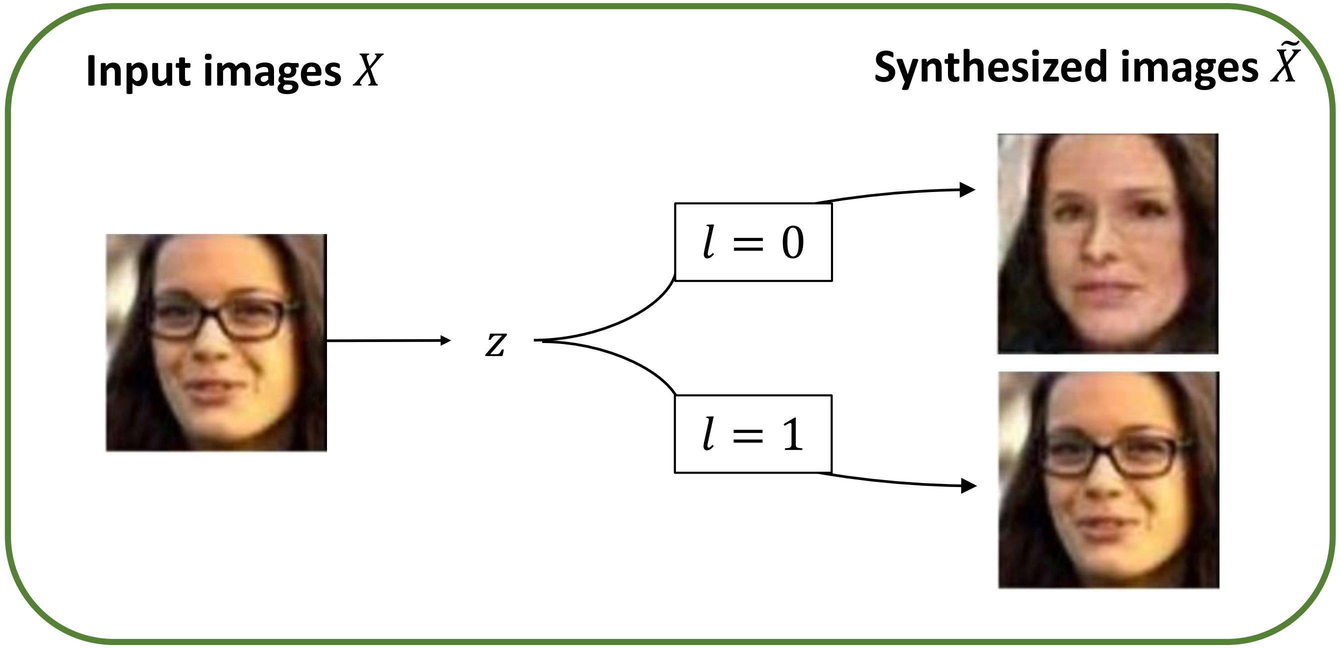
VAE + AC-GAN



- ✓ *Translate* the images conditioned on disentangled factor l

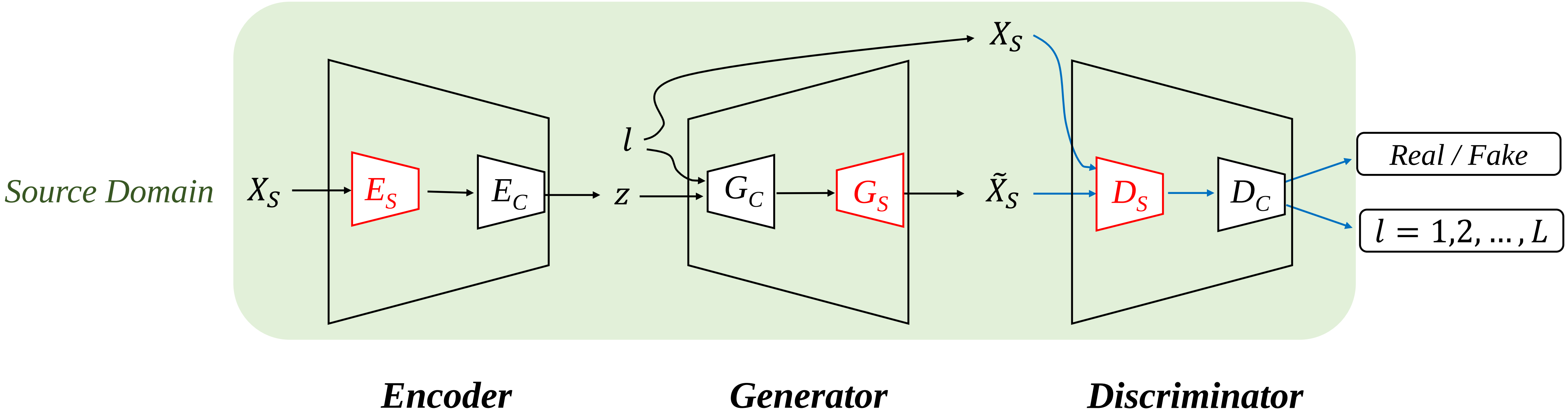
$$\mathcal{L}_{VAE} = \mathcal{L}_{perc} + KL(q_S(z|X)||p(z))$$

$$\mathcal{L}_{perc} = ||\Phi(X) - \Phi(\tilde{X})||_F^2$$



Proposed Method

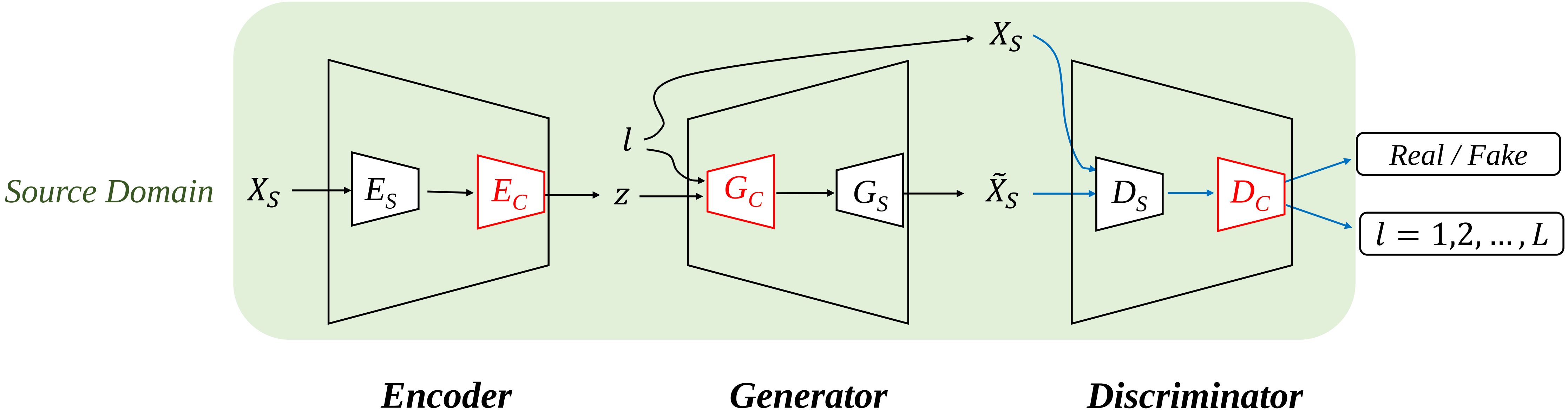
VAE + AC-GAN



✓ Divide the network into low-level and high-level layers.

Proposed Method

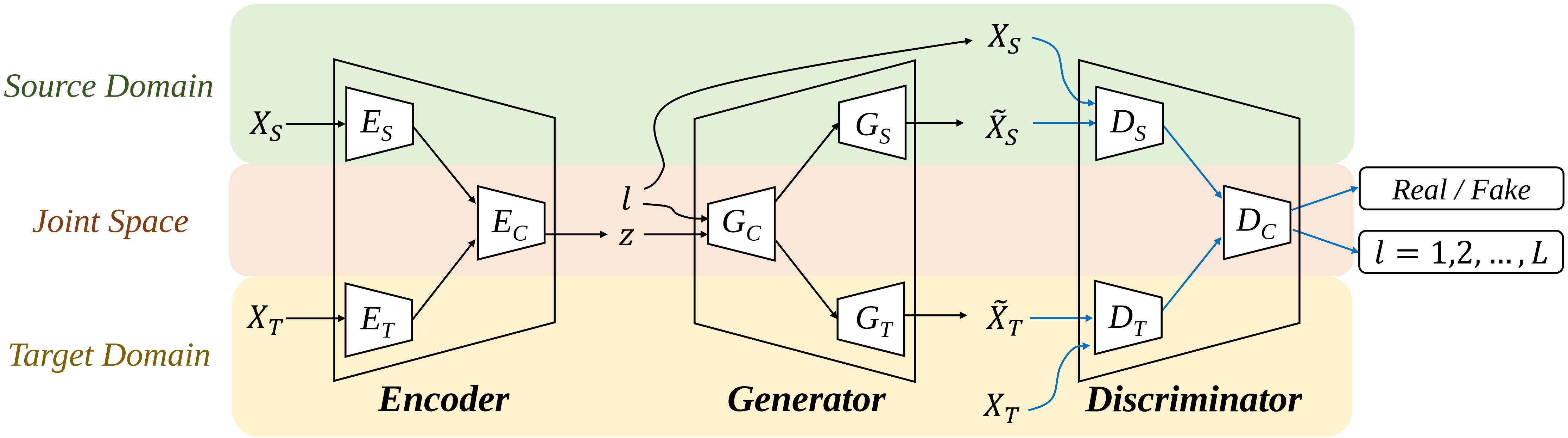
VAE + AC-GAN



✓ Divide the network into low-level and high-level layers.

Proposed Method

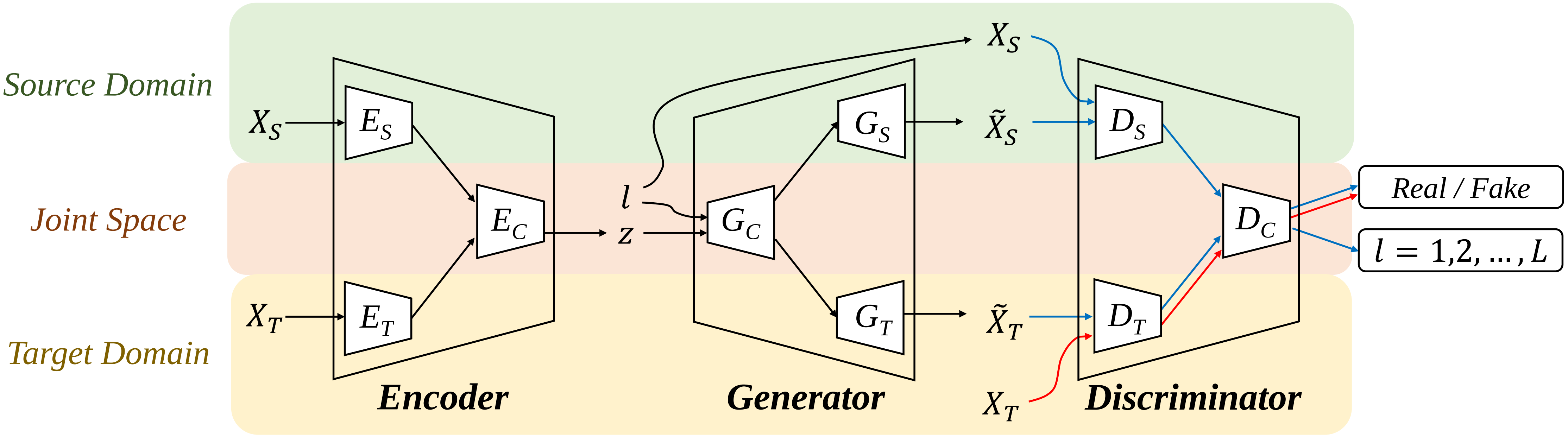
VAE + AC-GAN for cross-domain images



✓ Share the high-level layers of ***Encoder***, ***Generator***, and ***Discriminator***

Proposed Method

VAE + AC-GAN for cross-domain images

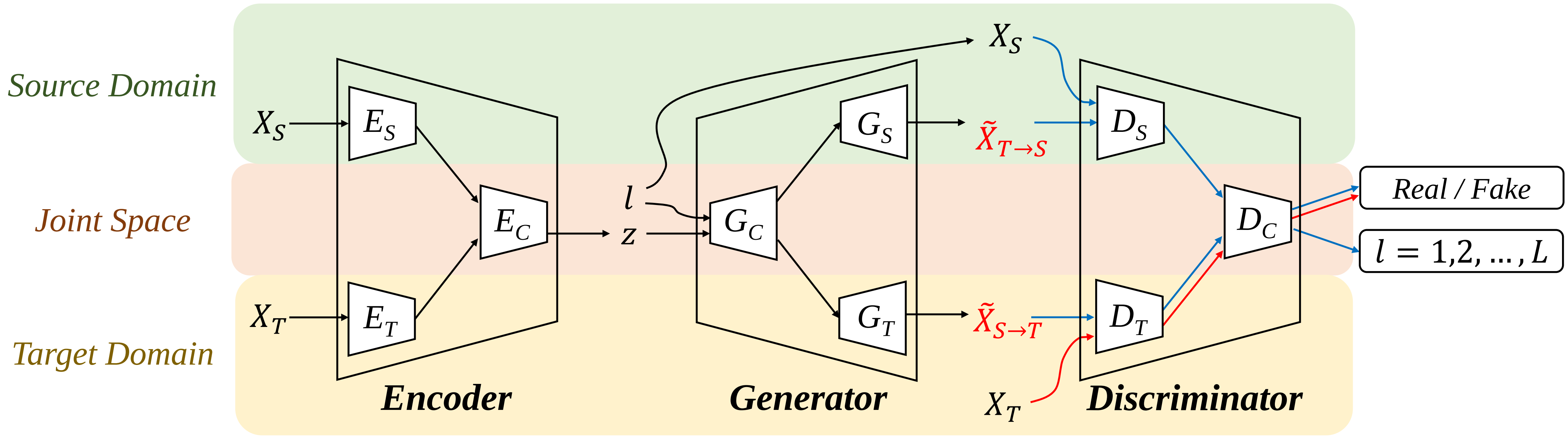


- ✓ No attribute supervision in the target domain.
- ✓ We only urge the synthesized data in target domain $\tilde{X}_T$ to be disentangled.

$$\mathcal{L}_{dis}^{T} = \mathbb{E}[\log(L = l|\tilde{X}_T)]$$

Proposed Method

VAE + AC-GAN for cross-domain images

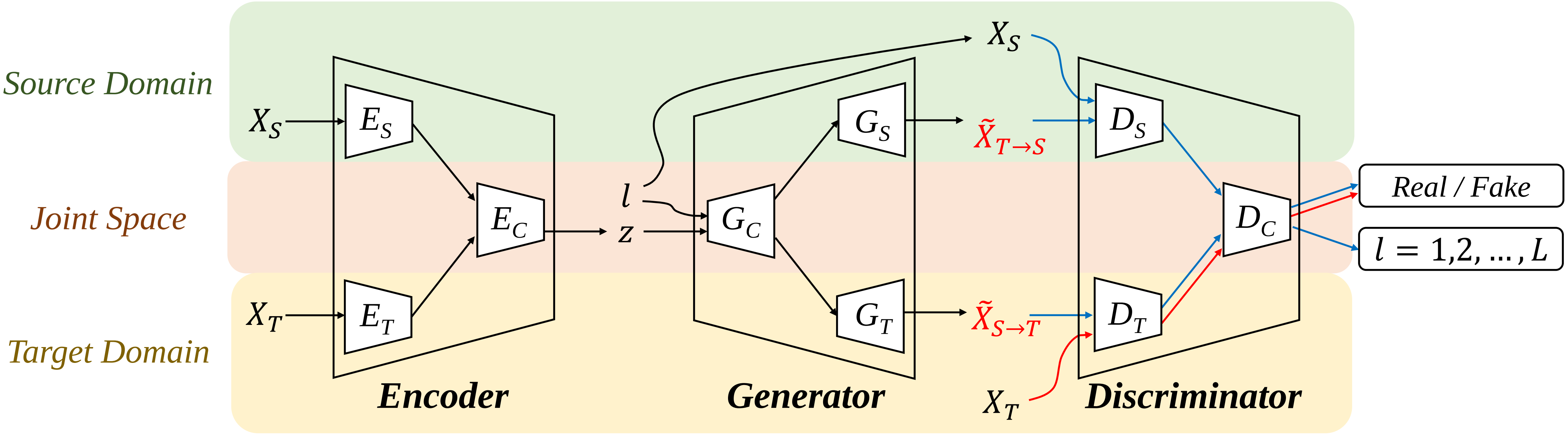


✓ Translate the images $\tilde{X}_{T\to S}$ and $\tilde{X}_{S\to T}$ across different domains.

$$\mathcal{L}_{adv}^{cd} = \mathcal{L}_{adv}^{S\to T} + \mathcal{L}_{adv}^{T\to S},$$
$$\mathcal{L}_{adv}^{S\to T} = \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_{S\to T})))] + \mathbb{E}[\log(D_C(D_T(X_T)))],$$
$$\mathcal{L}_{adv}^{T\to S} = \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_{T\to S})))] + \mathbb{E}[\log(D_C(D_T(X_S)))].$$

Proposed Method

VAE + AC-GAN for cross-domain images

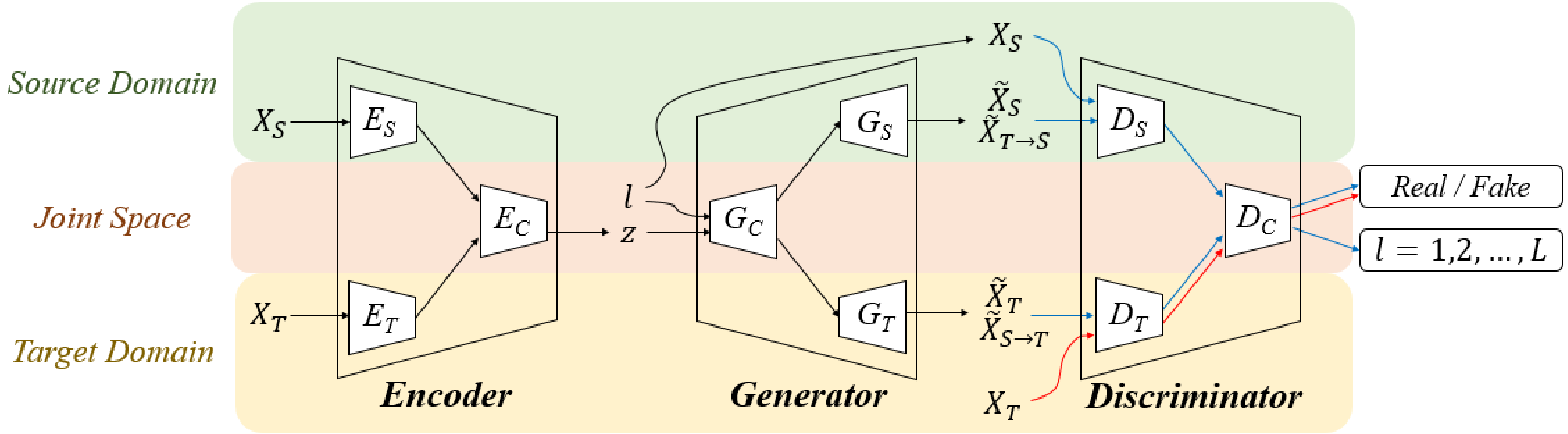


✓ Tie the disentangled factor $\boldsymbol{l}$ across domains with

$$\mathcal{L}_{dis}^{cd} = \mathbb{E}[\log(L = l|\tilde{X}_{S\to T})] + \mathbb{E}[\log(L = l|\tilde{X}_{T\to S})].$$

Proposed Method

VAE + AC-GAN for cross-domain images

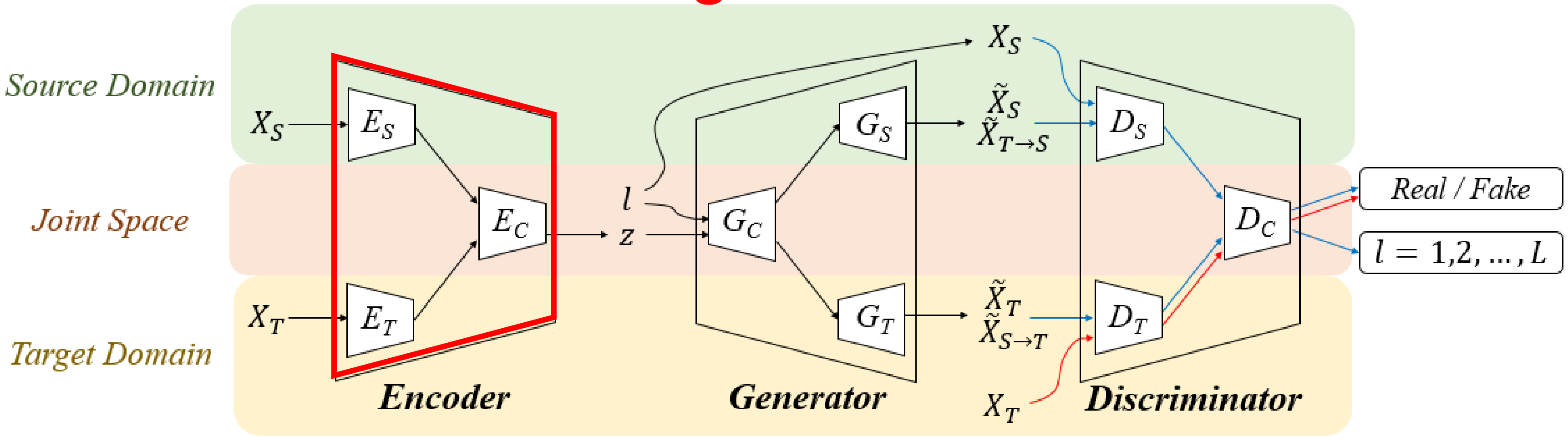


✓ Overall objective function can be defined as:

$$\mathcal{L}_E = \mathcal{L}_{VAE}$$
$$\mathcal{L}_G = \mathcal{L}^S_{perc} + \mathcal{L}^T_{perc} + \mathcal{L}_{dis} + \mathcal{L}_{adv}$$
$$\mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv}$$

Proposed Method

VAE + AC-GAN for cross-domain images



✓ Overall objective function can be defined as:

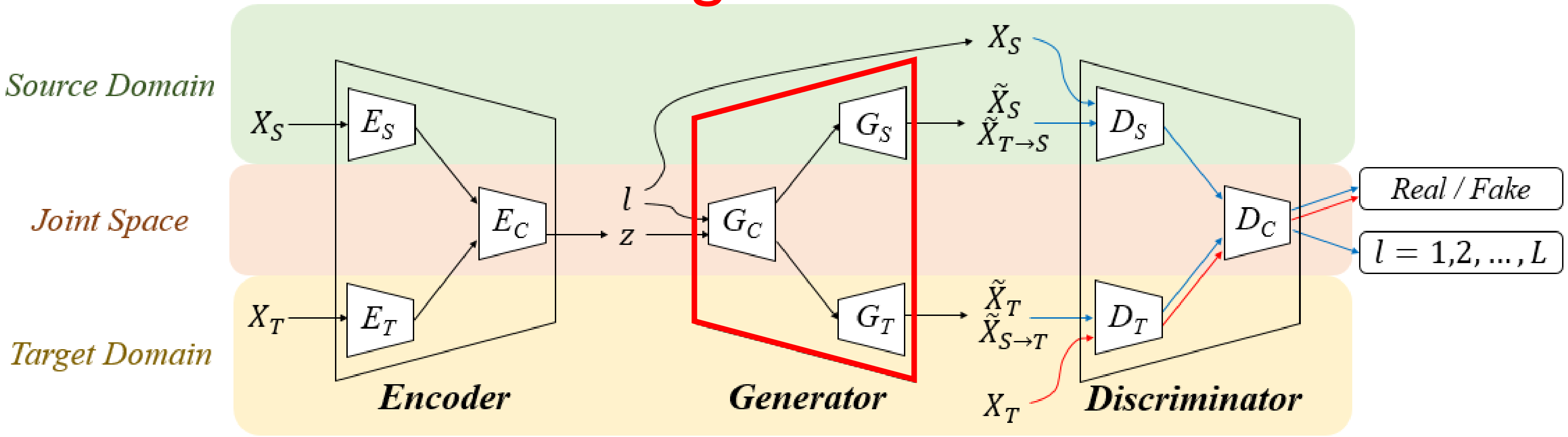
$$\boxed{\mathcal{L}_E = \mathcal{L}_{VAE}}$$

$$\mathcal{L}_G = \mathcal{L}^S_{perc} + \mathcal{L}^T_{perc} + \mathcal{L}_{dis} + \mathcal{L}_{adv}$$

$$\mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv}$$

Proposed Method

VAE + AC-GAN for cross-domain images



✓ Overall objective function can be defined as:

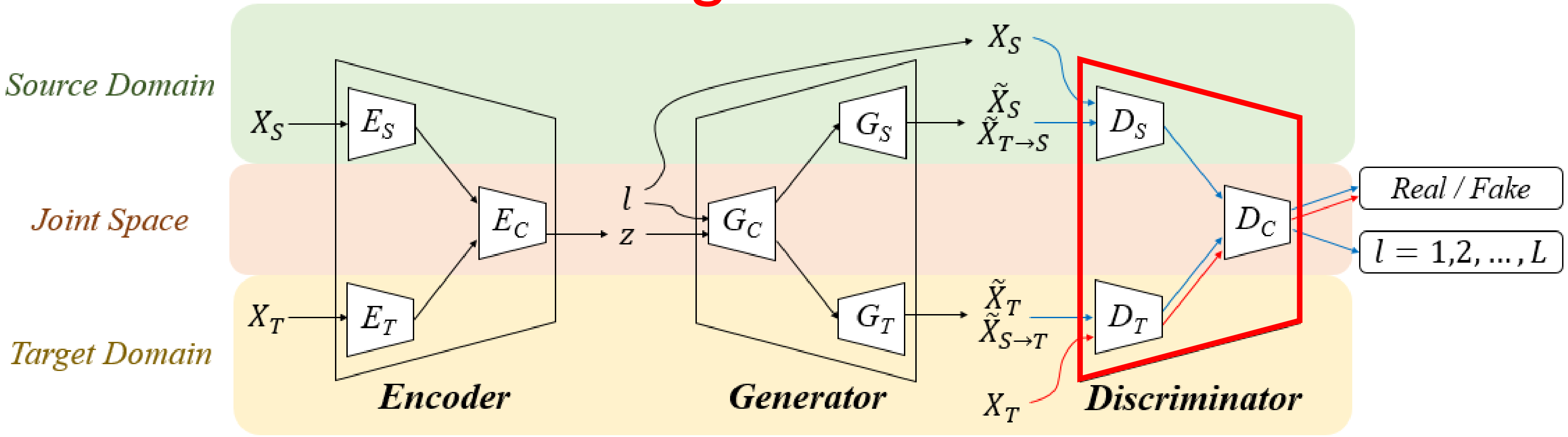
$$\mathcal{L}_E = \mathcal{L}_{VAE}$$

$$\boxed{\mathcal{L}_G = \mathcal{L}^S_{perc} + \mathcal{L}^T_{perc} + \mathcal{L}_{dis} + \mathcal{L}_{adv}}$$

$$\mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv}$$

Proposed Method

VAE + AC-GAN for cross-domain images



✓ Overall objective function can be defined as:

$$\mathcal{L}_E = \mathcal{L}_{VAE}$$

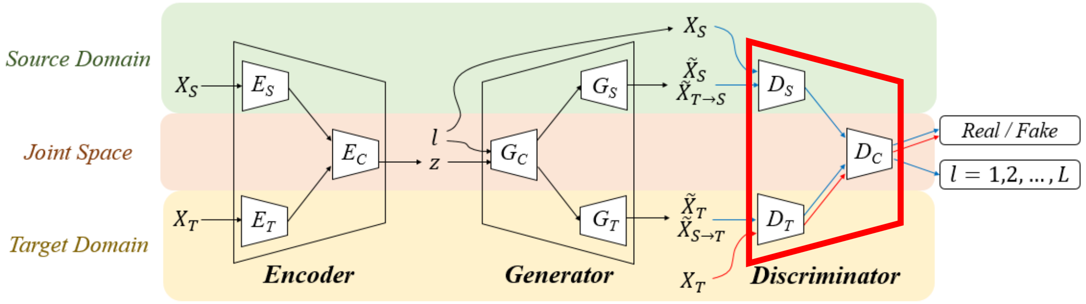
$$\mathcal{L}_G = \mathcal{L}^S_{perc} + \mathcal{L}^T_{perc} + \mathcal{L}_{dis} + \mathcal{L}_{adv}$$

$$\boxed{\mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv}}$$

Experiments

- Qualitative Evaluation:
  - Conditional image synthesis and translation
- Quantitative Evaluation:
  - Cross-domain attribute classification
- Dataset
  - CelebFaces Attributes dataset (CelebA)
  - A large-scale face dataset with 200K+ celebrity images with 40 facial annotated attributes

Results



S : faces w/o eyeglasses; T : faces w/ eyeglasses; l : attribute of smiling

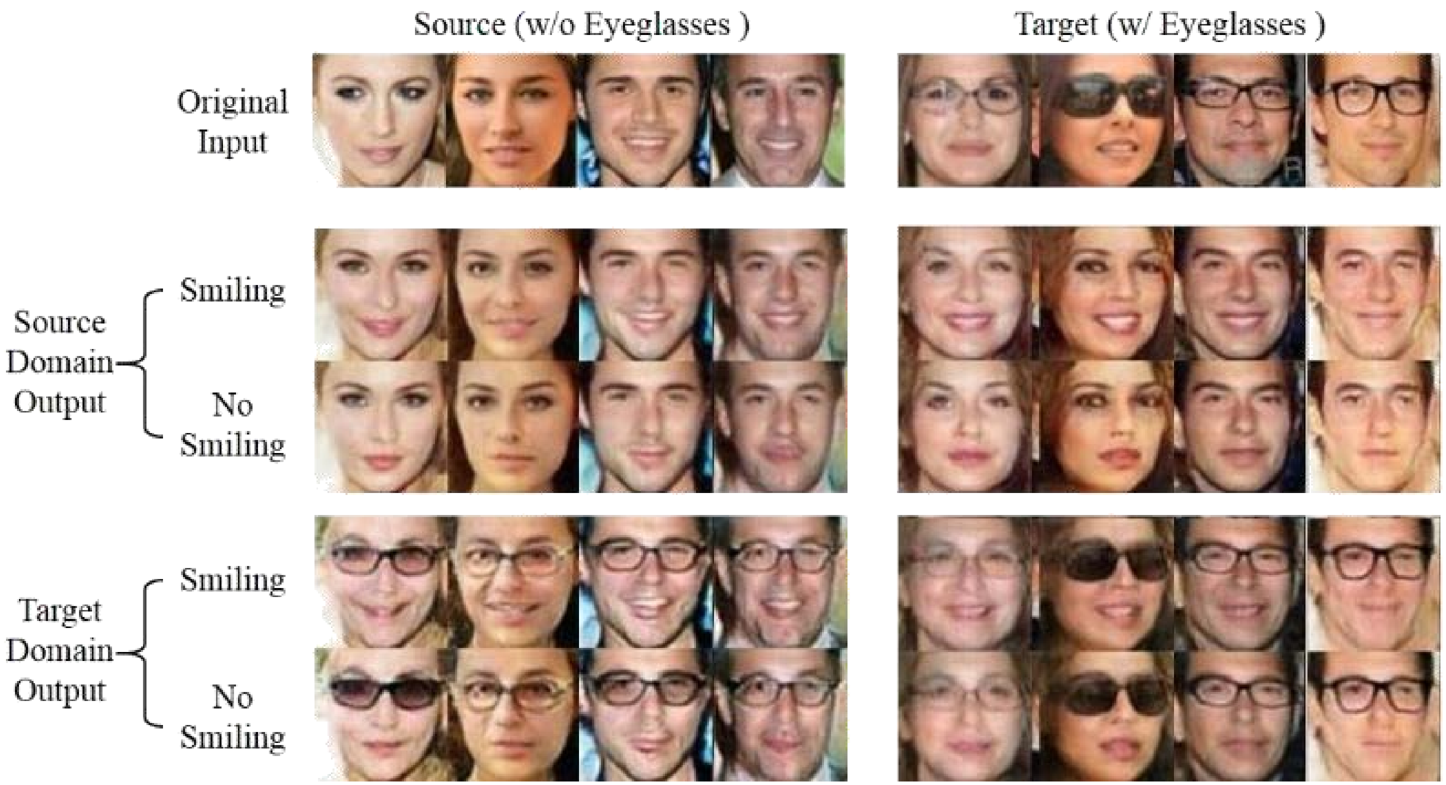


Table 1: Cross-domain classification results of face images with respect to the attribute of smiling. Source and target-domain test data are faces w/o eyeglasses and w/ eyeglasses, respectively.

Method		CoGAN	UNIT	Ours*	Ours
Accuracy (%)	Source	87.03	88.66	89.48	89.73
	Target	71.92	71.82	**83.69**	**84.43**

Results

S : real photo of faces; T : simulated sketch of faces; l : attribute of smiling

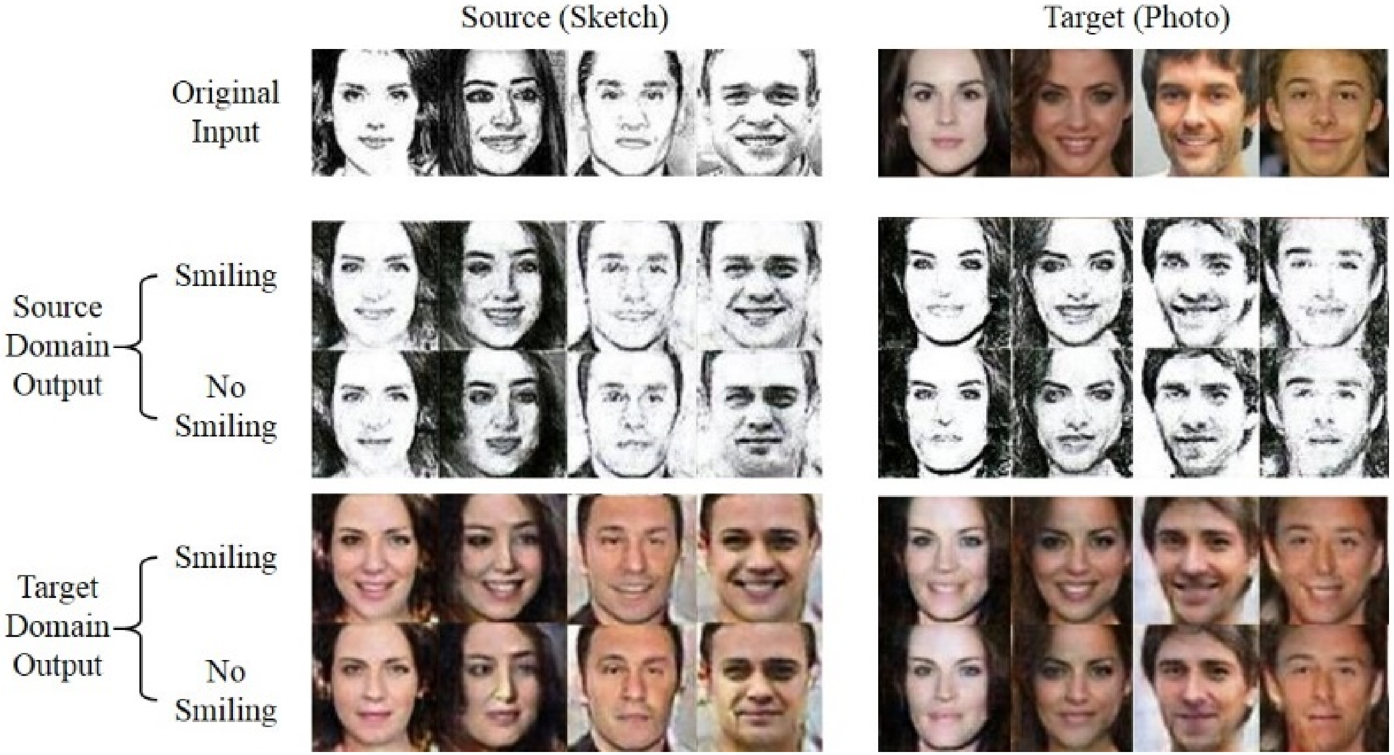


Table 2: Cross-domain classification results of face images with respect to the attribute of smiling. Source and target-domain test data are sketch and photo faces, respectively.

Method		CoGAN	UNIT	Ours*	Ours
Accuracy (%)	Source	89.50	90.10	90.19	90.01
	Target	78.90	81.04	**87.61**	**88.28**

Results

S : real photo of faces; T : simulated sketch of faces; l : attribute of eyeglasses

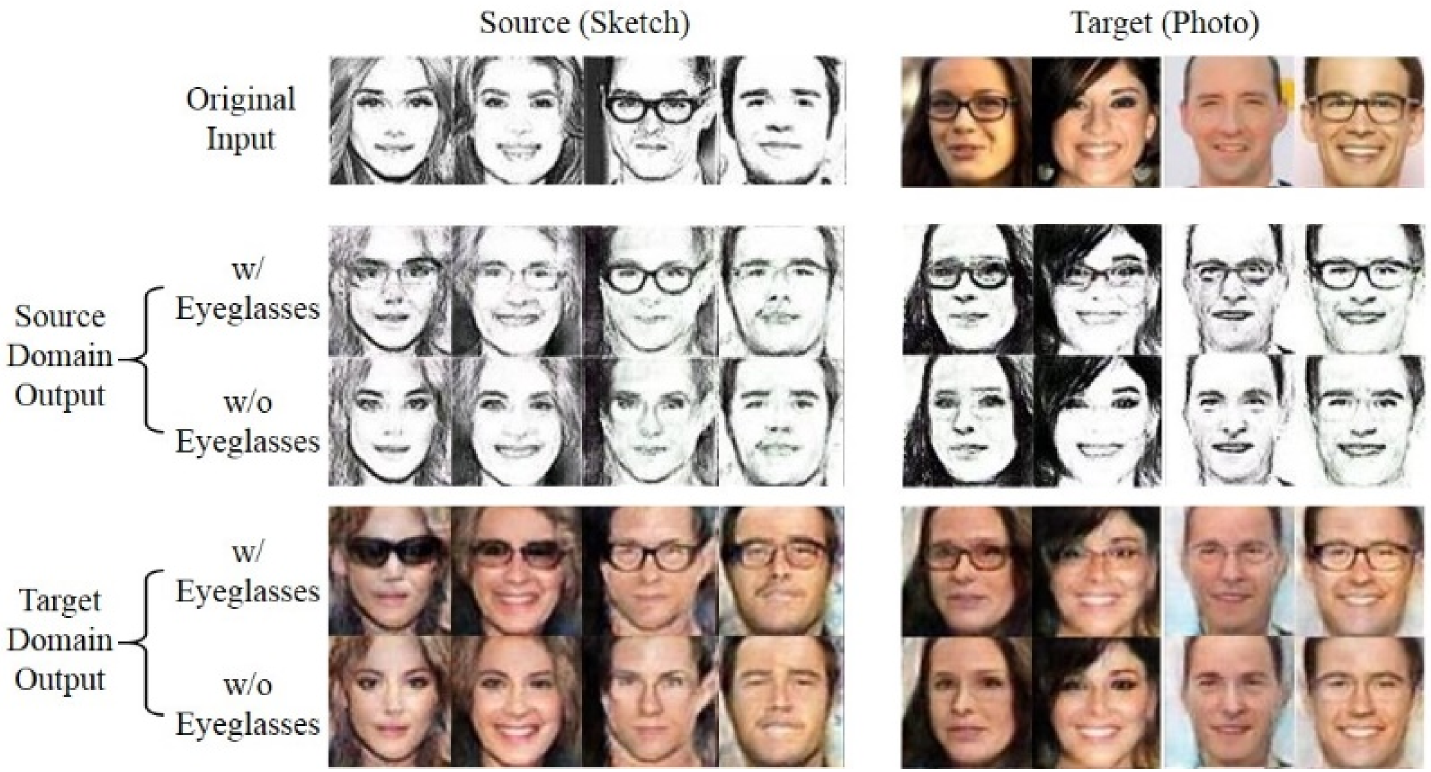


Table 3: Cross-domain classification results of face images with respect to the attribute of eyeglasses. Source and target-domain test data are sketch and photo faces, respectively.

Method		CoGAN	UNIT	Ours*	Ours
Accuracy (%)	Source	96.63	97.65	97.06	97.19
	Target	81.01	79.89	**94.49**	**94.84**

Summary

- Transfer Learning for
  - Homogeneous/heterogeneous domain adaptation
  - Multi-label classification / zero-shot learning
  - Robust face recognition (e.g., cross-resolution, cross-modality, etc.)
- Feature Disentanglement for
  - Cross-domain image synthesis/translation/classification
  - Only label supervision from a single (source) domain is needed