Recent Progress on Utilizing Tag Information with GANs

StarGAN & TD-GAN

Hao-Wen Dong

Dec 26, 2017

Research Center for IT Innovation, Academia Sinica

Table of contents

Introduction

Main Idea

How to utilize tag information?

Straightforward Approach

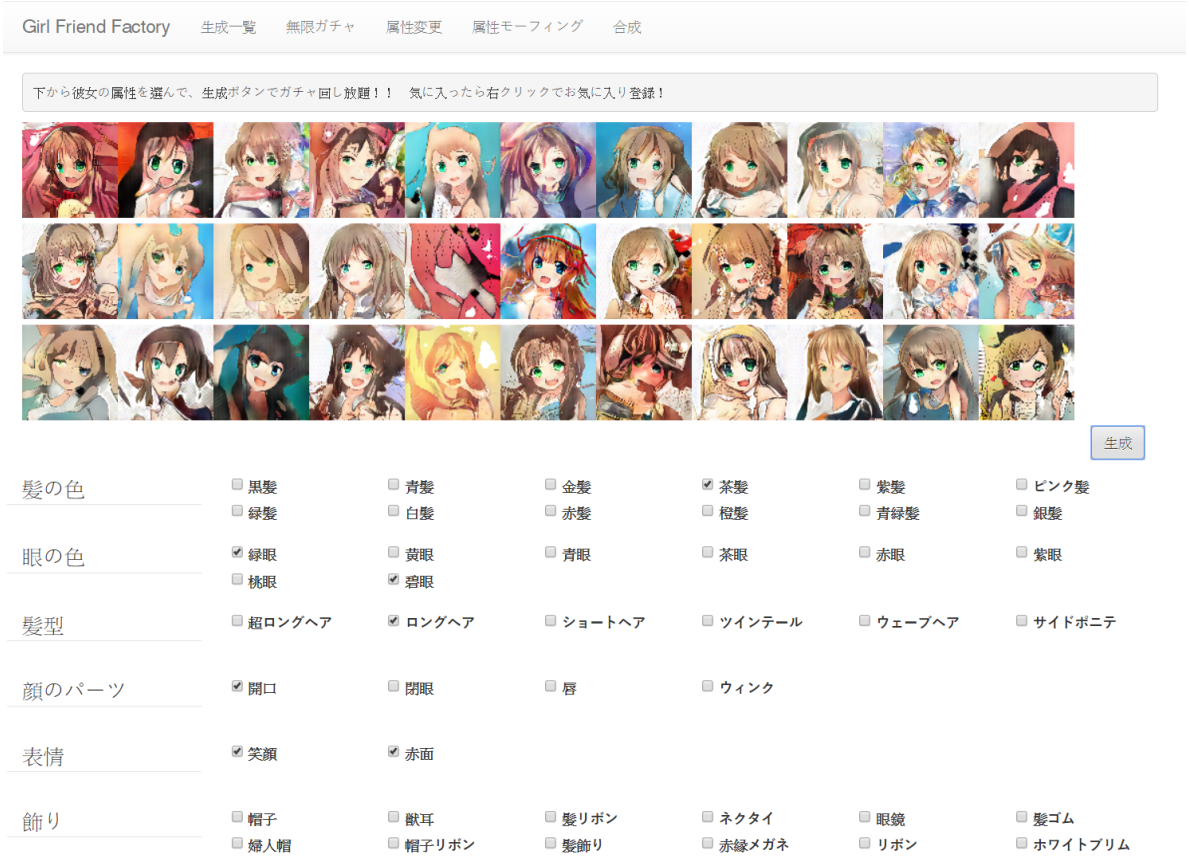

Fig. 1: *Girl Friend Factory**: generating anime characters with specific attributes by conditional GANs

*https://hiroshiba.github.io/girl_friend_factory/index.html

Inspiring Viewpoints

StarGAN [1]

Key Assumption: images of different tags can be viewed as different domains

Approach: multi-domain image-to-image translation

Paper:

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "**StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation**", *arXiv preprint arXiv:1711.09020*, 2017.

Inspiring Viewpoints

TD-GAN [2]

Key Assumption: images and tags record the same object from two different perspectives

Approach: enforce consistency between learned disentangled representations for images and tags

Paper:

Chaoyue Wang, Chaohui Wang, Chang Xu, and Dacheng Tao, "**Tag Disentangled Generative Adversarial Networks for Object Image Re-rendering**", in *Proc. 36th Int. Joint Conf. on Artificial Intelligence* (IJCAI), 2017.

StarGAN

StarGAN - Motivation

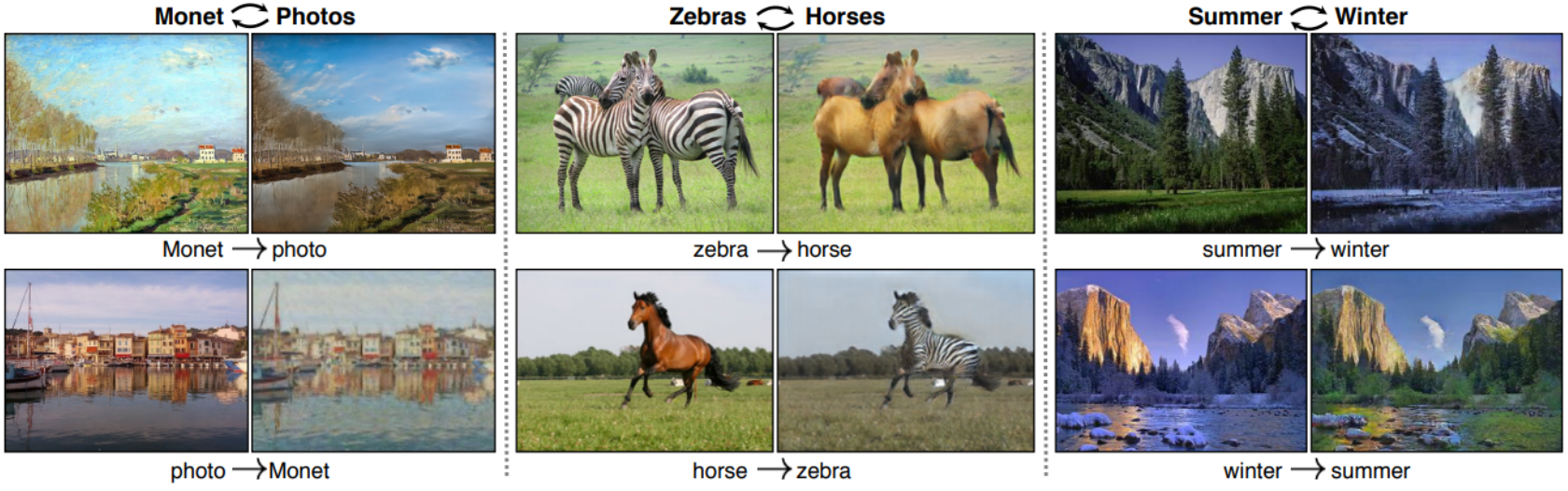

Fig. 2: CycleGAN [3]: Cycle-Consistent Adversarial Networks

StarGAN - Motivation

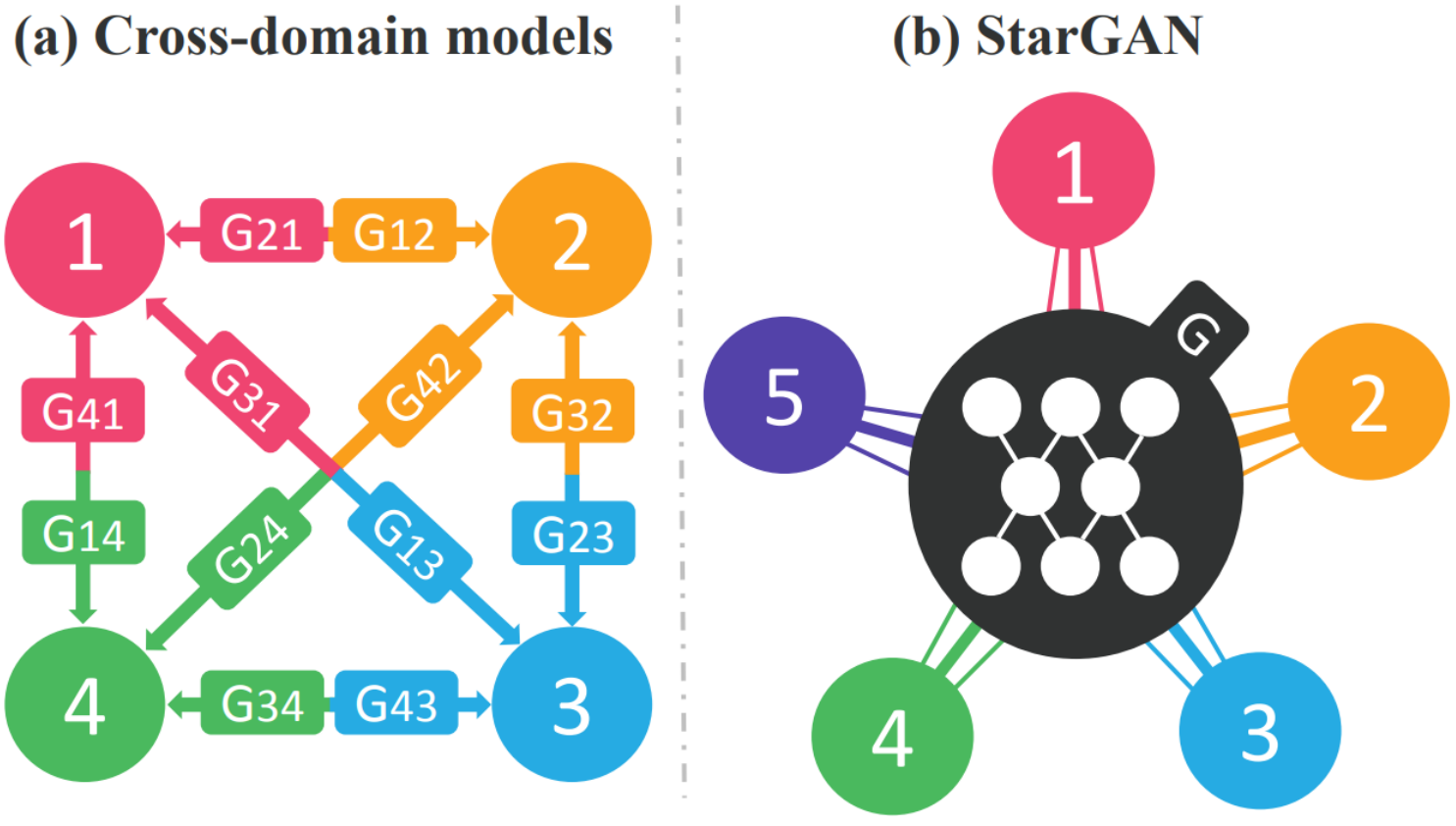

Fig. 3: Comparison of cross-domain and multi-domain models

StarGAN - Qualitative Results

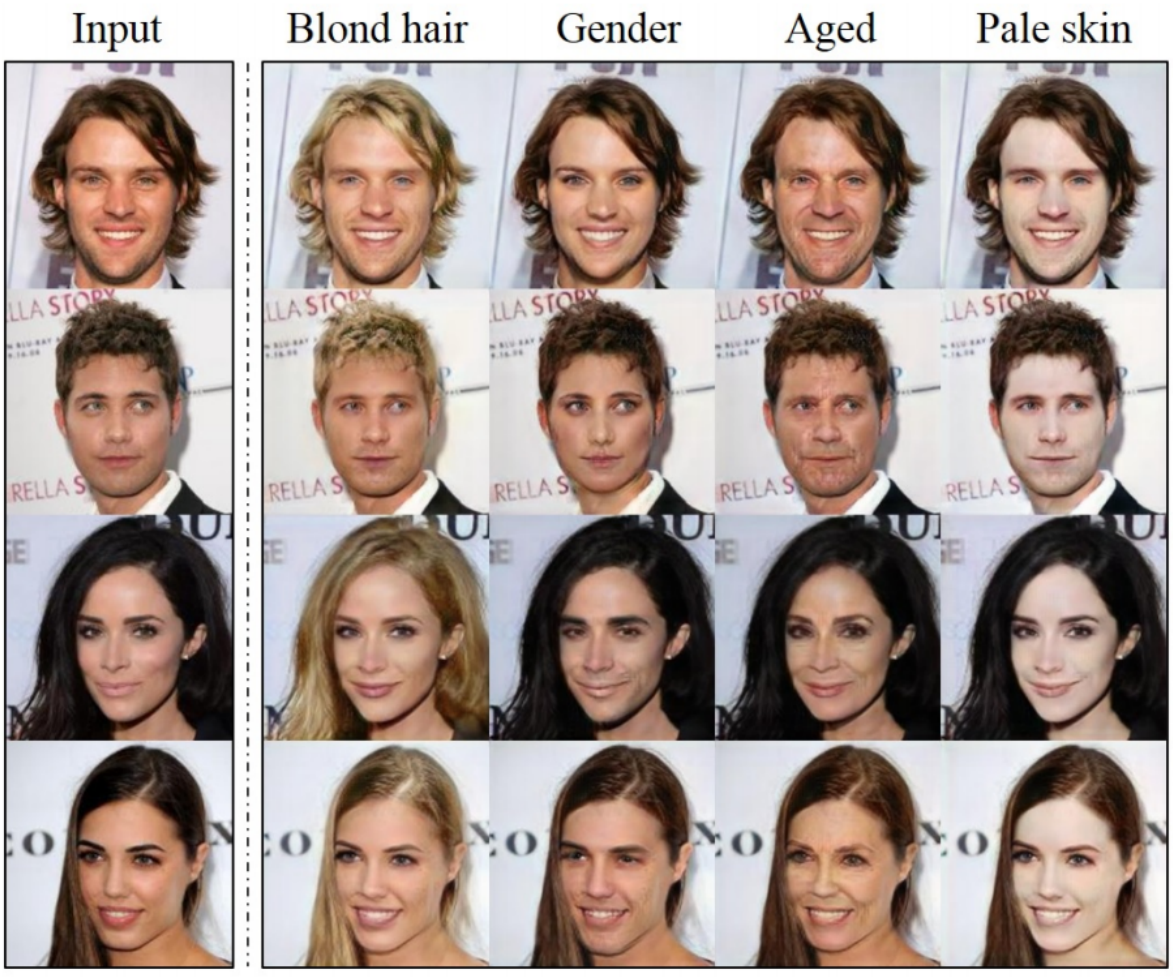

Fig. 4: Results on CelebA

StarGAN - Qualitative Results

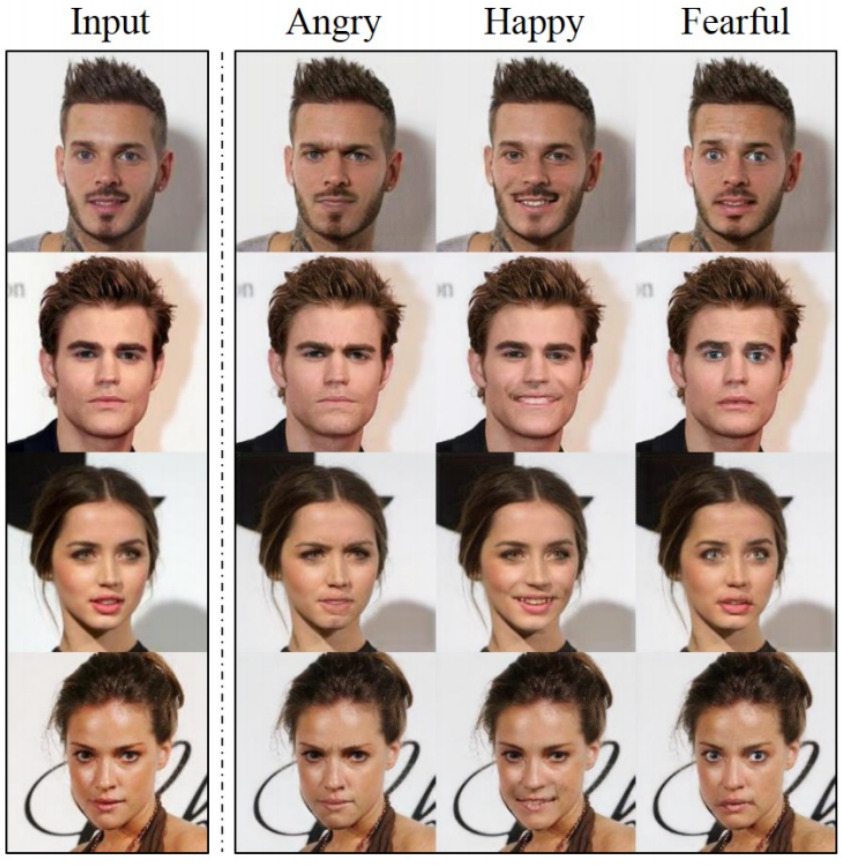

Fig. 5: Results on CelebA via **transferring knowledge** learned from RaFD

StarGAN - System Overview

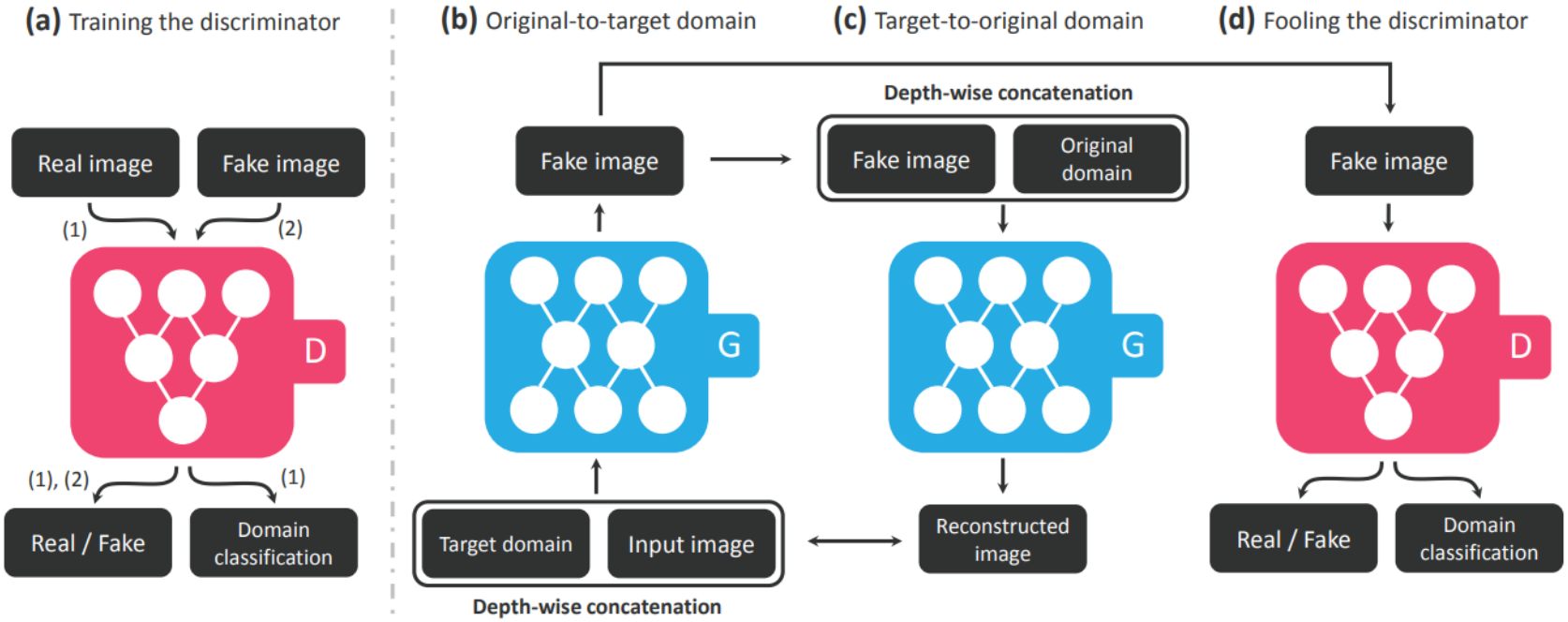

Fig. 6: System overview

StarGAN - Formulation

$$G : (x, c) \rightarrow y$$

$$D : x \rightarrow (D_{src}(x), D_{cls}(x))$$

x : input image, c : domain label, y : output image

StarGAN - Objective Functions

Adversarial Loss

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x, c)))]$$

StarGAN - Objective Functions

Adversarial Loss

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x, c)))]$$

Domain Classification Loss

$$\mathcal{L}_{cls}^{r} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)] \quad \text{(real images)}$$

$$\mathcal{L}_{cls}^{f} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x, c))] \quad \text{(fake images)}$$

StarGAN - Objective Functions

Adversarial Loss

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x, c)))]$$

Domain Classification Loss

$$\mathcal{L}_{cls}^{r} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)] \quad \text{(real images)}$$

$$\mathcal{L}_{cls}^{f} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x, c))] \quad \text{(fake images)}$$

Reconstruction Loss

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[||x - G(G(x, c), c')||_1]$$

StarGAN - Objective Functions

Full Objective Functions

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls}$$

$$\mathcal{L}_G = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec}$$

StarGAN - Qualitative Results

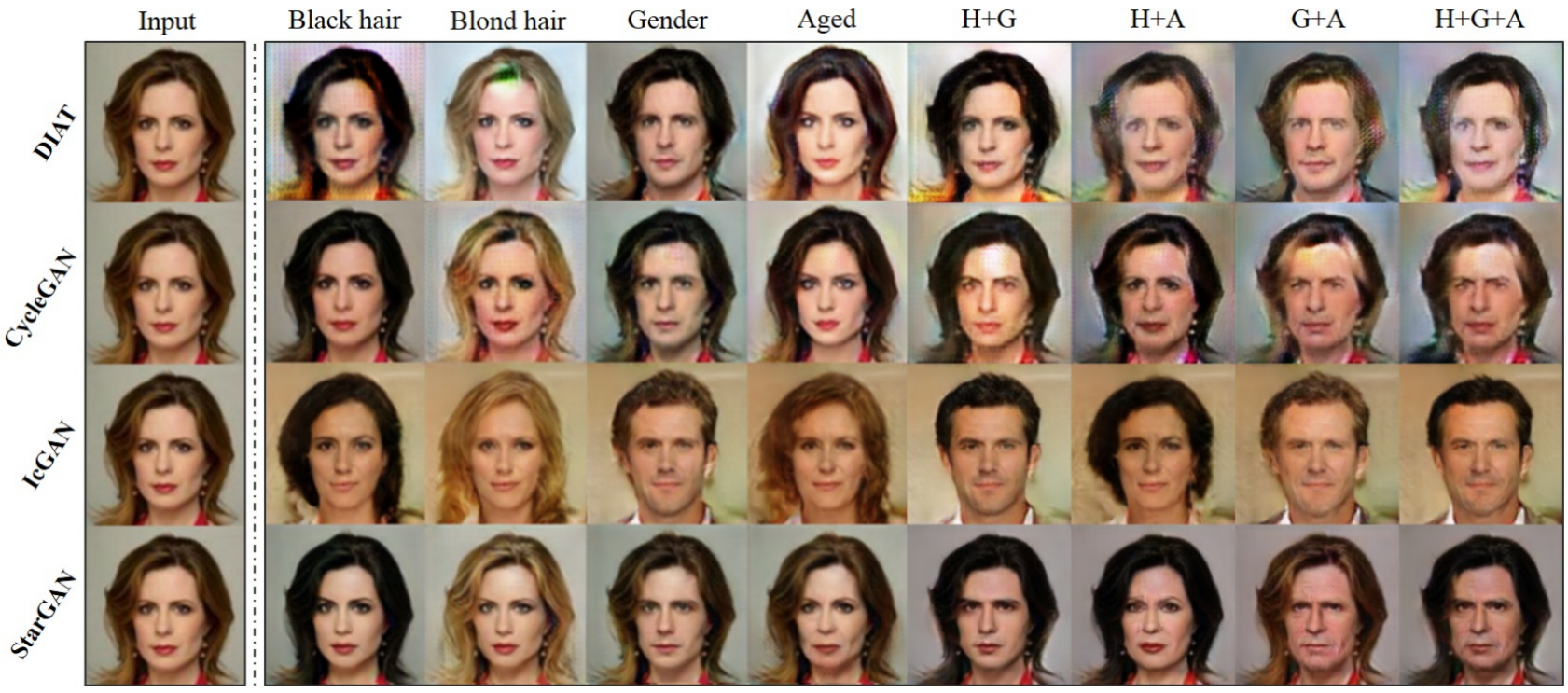

Fig. 7: Qualitative comparison among different models

StarGAN - Multiple Datasets

CelebA dataset

(binary) Black, Blond, Brown, Male, Young, etc.

RaFD dataset

(categorical) Angry, Fearful, Happy, Sad, Disgusted, etc.

StarGAN - Multiple Datasets

CelebA dataset

(binary) Black, Blond, Brown, Male, Young, etc.

RaFD dataset

(categorical) Angry, Fearful, Happy, Sad, Disgusted, etc.

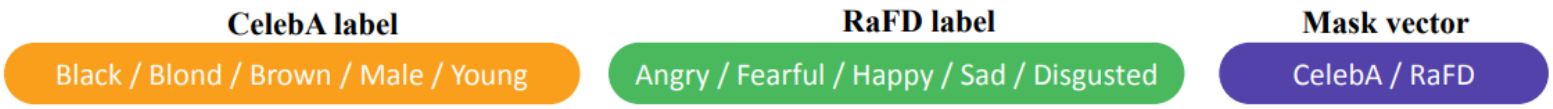

Fig. 8: Label vectors and mask vector

StarGAN - Training on Multiple Datasets

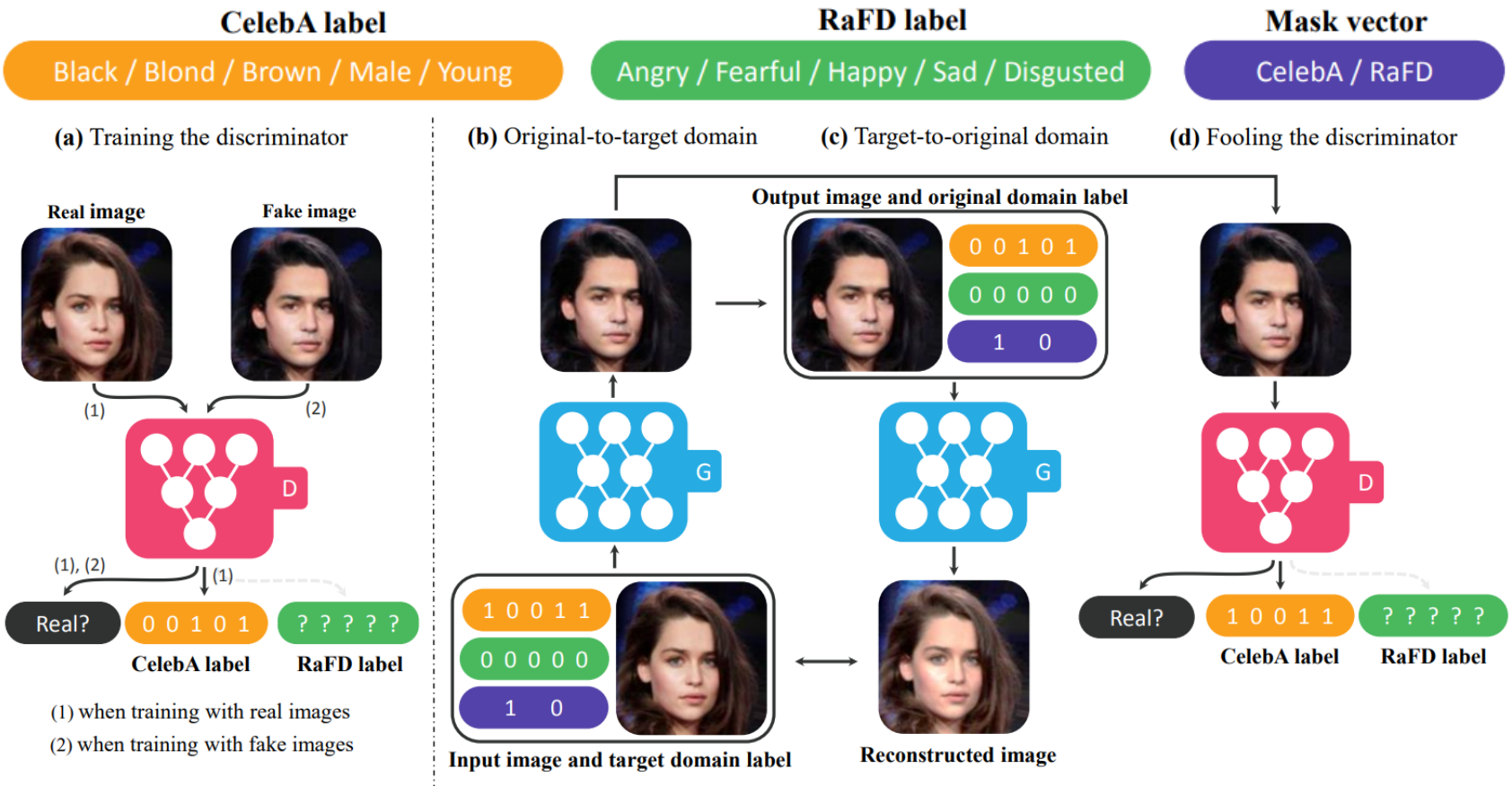

Fig. 9: System overview - *multiple datasets* (I)

StarGAN - Training on Multiple Datasets

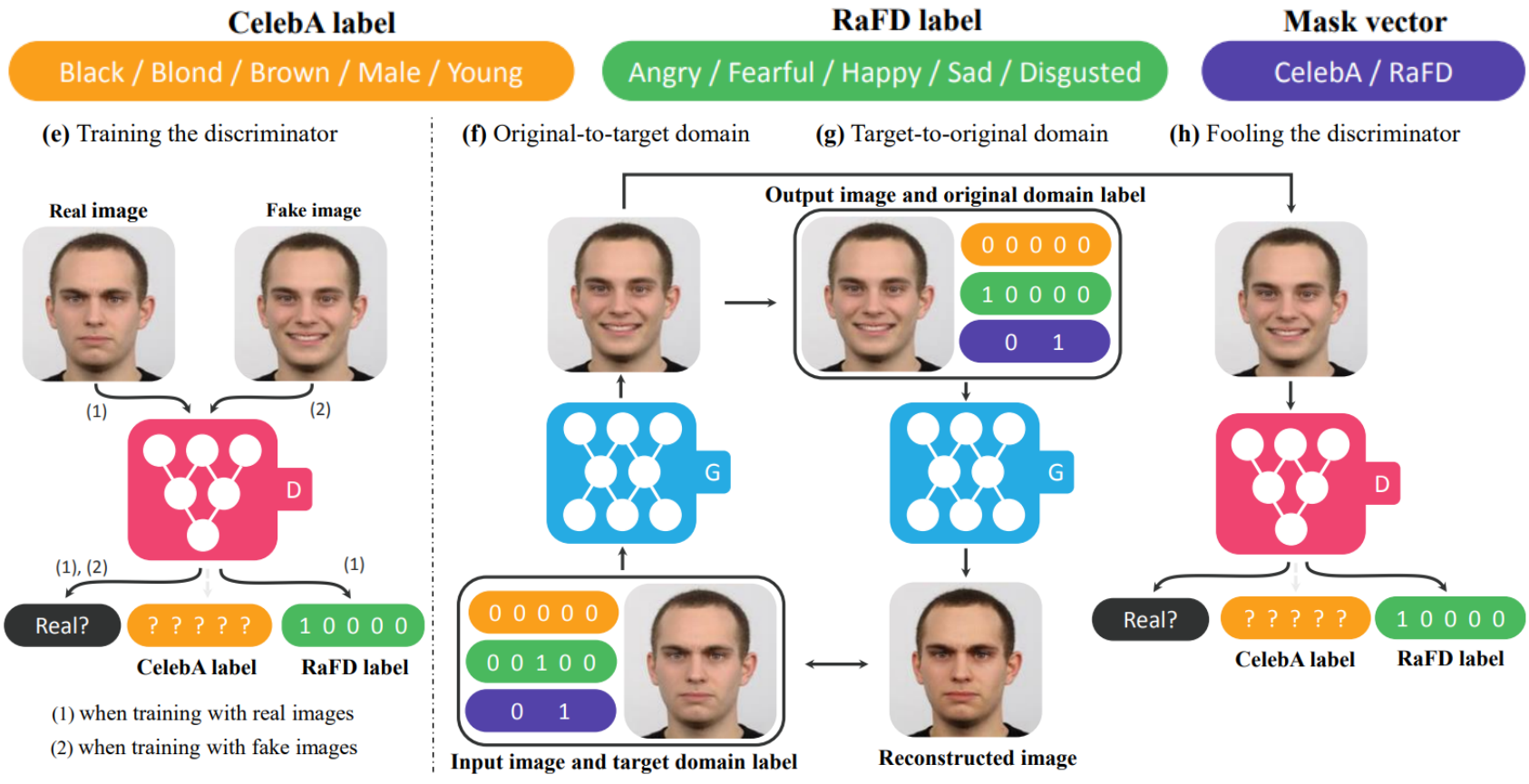

Fig. 9 (cont.): System overview - *multiple datasets* (II)

StarGAN - Qualitative Results

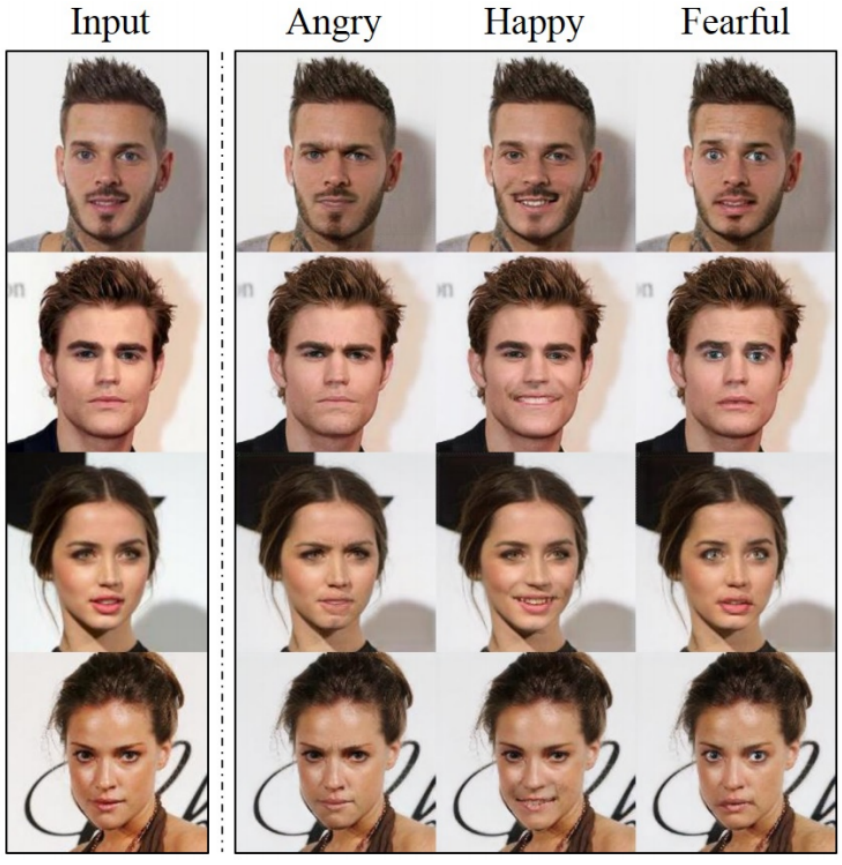

Fig. 10: Results on CelebA via **transferring knowledge** learned from RaFD

StarGAN - Mask Vector

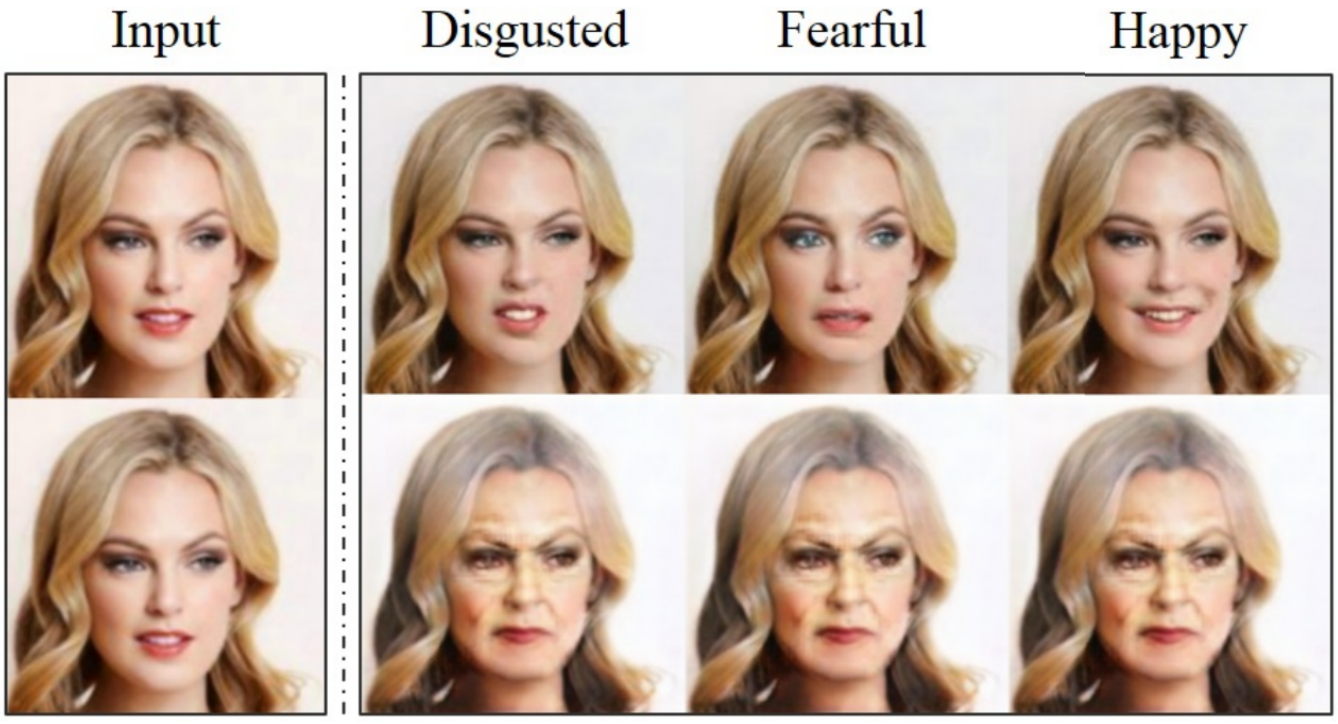

Fig. 11: Learned role of mask vector

StarGAN - Quantitative Experiments

Amazon Mechanical Turk (AMT)

vote for the best generated images based on:

- perceptual realism
- quality of transfer in attribute(s)
- preservation of a figure's original identity

StarGAN - Quantitative Results

Method	Hair color	Gender	Aged
DIAT	9.3%	31.4%	6.9%
CycleGAN	20.0%	16.6%	13.3%
IcGAN	4.5%	12.9%	9.2%
StarGAN	**66.2%**	**39.1%**	**70.6%**

Table 1: AMT perceptual evaluation (by votes)

StarGAN - Quantitative Results

Method	H+G	H+A	G+A	H+G+A
DIAT	20.4%	15.6%	18.7%	15.6%
CycleGAN	14.0%	12.0%	11.2%	11.9%
IcGAN	18.2%	10.9%	20.3%	20.3%
StarGAN	**47.4%**	**61.5%**	**49.8%**	**52.2%**

Table 1 (cont.): AMT perceptual evaluation (by votes)

TD-GAN

TD-GAN - Motivation & Goals

Key Assumption

the image and its tags **record the same object from two different perspectives**, so they should **share the same disentangled representations**

Goals

- to extract disentangled and interpretable representations for both image and its tags
- to explore the consistency between the image and its tags by integrating the tag mapping net

TD-GAN - Qualitative Results

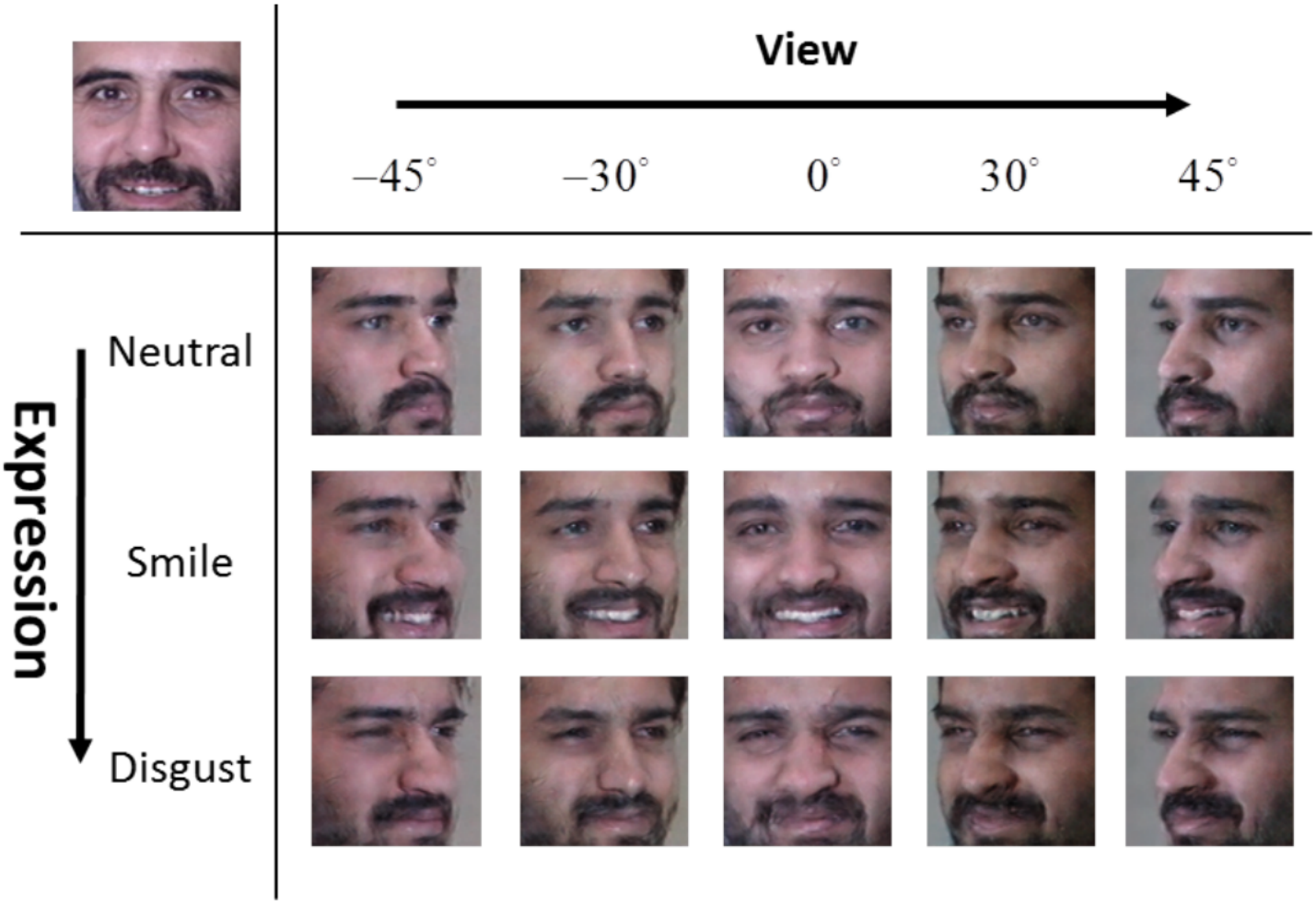

Fig. 12: Multi-factor transformation

TD-GAN - System Overview

Disentangling network R

$$\mathbf{x} \rightarrow R(\mathbf{x})$$

image $\rightarrow$ disentangled representations

TD-GAN - System Overview

Disentangling network R

$$\mathbf{x} \rightarrow R(\mathbf{x})$$

image $\rightarrow$ disentangled representations

Tag mapping net g

$$\mathbf{C} \rightarrow g(\mathbf{C})$$

tag code $\rightarrow$ disentangled representations

TD-GAN - System Overview

Generative network G

$$g(\mathbf{C}) \text{ or } R(\mathbf{x}) \rightarrow G(g(\mathbf{C})) \text{ or } G(R(\mathbf{x}))$$

disentangled representations $\rightarrow$ re-rendered image

TD-GAN - System Overview

Generative network G

$$g(\mathbf{C}) \text{ or } R(\mathbf{x}) \rightarrow G(g(\mathbf{C})) \text{ or } G(R(\mathbf{x}))$$

disentangled representations $\rightarrow$ re-rendered image

Discriminative network D

adversarial training with G and R

TD-GAN - System Overview

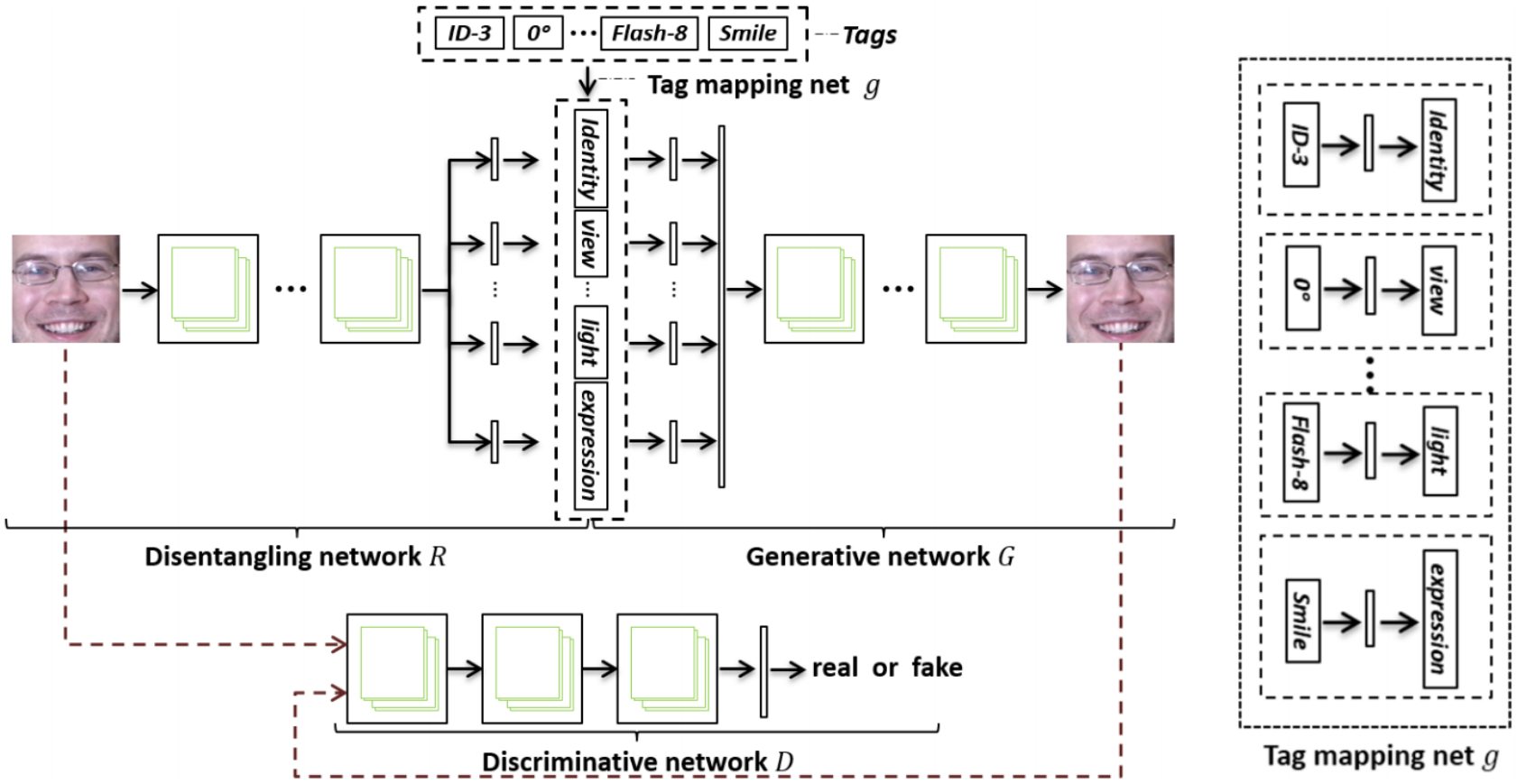

Fig. 13: System overview

TD-GAN - Formulation

Let the **tagged training dataset** be

$$\mathcal{X}^L = \{(\mathbf{x}_1, \mathbf{C}_1), \ldots, (\mathbf{x}_{|\mathcal{X}^L|}, \mathbf{C}_{|\mathcal{X}^L|})\},$$

where *tag codes*

$$\mathbf{C}_i = (\mathbf{c}_i^{\text{ide}}, \mathbf{c}_i^{\text{view}}, \mathbf{c}_i^{\text{exp}}, \ldots)$$

and $\mathbf{c}_i^{\text{ide}}, \mathbf{c}_i^{\text{view}}, \mathbf{c}_i^{\text{exp}}, \ldots$ are one-hot encoding vectors.

Let the **untagged training dataset** be $\mathcal{X}^U$.

TD-GAN - Objective Functions

Discrepency between disentangled representations

$$f_1(R, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} ||R(\mathbf{x}_i) - g(\mathbf{C}_i)||_2^2$$

TD-GAN - Objective Functions

Discrepency between disentangled representations

$$f_1(R, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} ||R(\mathbf{x}_i) - g(\mathbf{C}_i)||_2^2$$

Discrepency between real and rendered images

$$f_2(G, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} ||G(g(\mathbf{C}_i)) - \mathbf{x}_i||_2^2$$

TD-GAN - Objective Functions

Discrepency between disentangled representations

$$f_1(R, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} ||R(\mathbf{x}_i) - g(\mathbf{C}_i)||_2^2$$

Discrepency between real and rendered images

$$f_2(G, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} ||G(g(\mathbf{C}_i)) - \mathbf{x}_i||_2^2$$

TD-GAN - Objective Functions

Reconstruction Loss for untagged images

$$\tilde{f}_1(G, R) = \frac{1}{|\mathcal{X}^U|} \sum_{\mathbf{x}_i \in \mathcal{X}^U} ||G(R(\mathbf{x}_i)) - \mathbf{x}_i||_2^2$$

Adversarial Loss

$$f_3(R, G, D) = \mathbb{E}[\log D(\mathbf{x})] + \mathbb{E}[\log(1 - D(G(R(\mathbf{x}))))]$$

TD-GAN - Objective Functions

Full Objective Functions

$$\mathcal{L}_R = \min_R \ \lambda_1 f_1(R, g^*) + \lambda_3 f_3(R, G^*, D^*)$$

$$\mathcal{L}_g = \min_g \ \lambda_1 f_1(R^*, g) + \lambda_2 f_2(G^*, g)$$

$$\mathcal{L}_G = \min_G \ \lambda_2 f_2(G, g^*) + \lambda_3 f_3(R^*, G, D^*)$$

$$\mathcal{L}_D = \max_D \ \lambda_3 f_3(R^*, G^*, D)$$

(R^*, g^*, G^*, D^* are fixed as the configurations obtained from the previous iteration)

TD-GAN - Qualitative Results

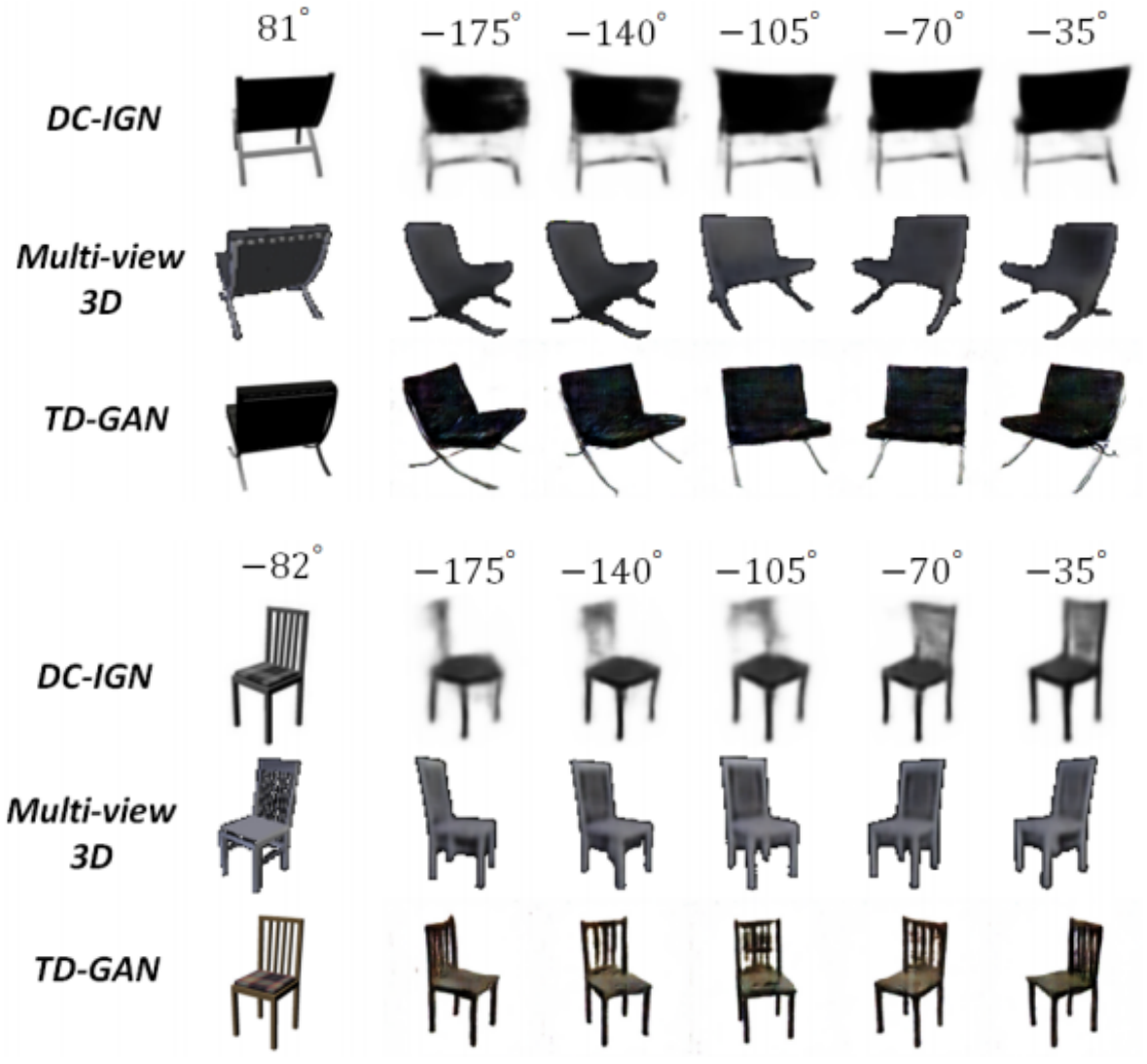

Fig. 14: Novel view synthesis results on 3D-chairs dataset

TD-GAN - Semi-supervised Extension

Test three representative training settings:

- *fully-500*: all the 500 training models and their tags
- *fully-100*: only the first 100 models and their tags
- *semi-(100,400)*: all the 500 training models but only the tags of the first 100 models

TD-GAN - Qualitative Results

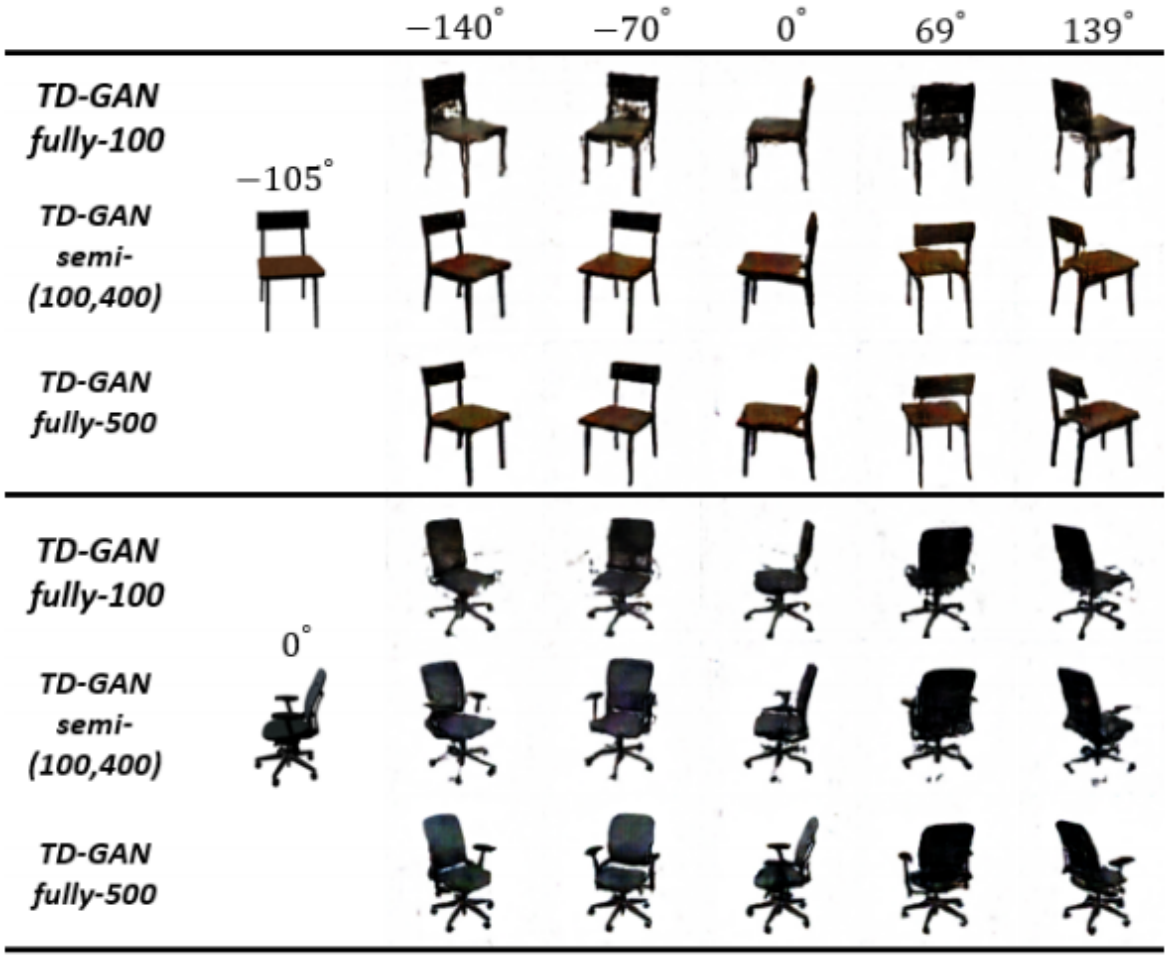

Fig. 15: Novel view synthesis results trained in three settings

TD-GAN - Quantitative Results

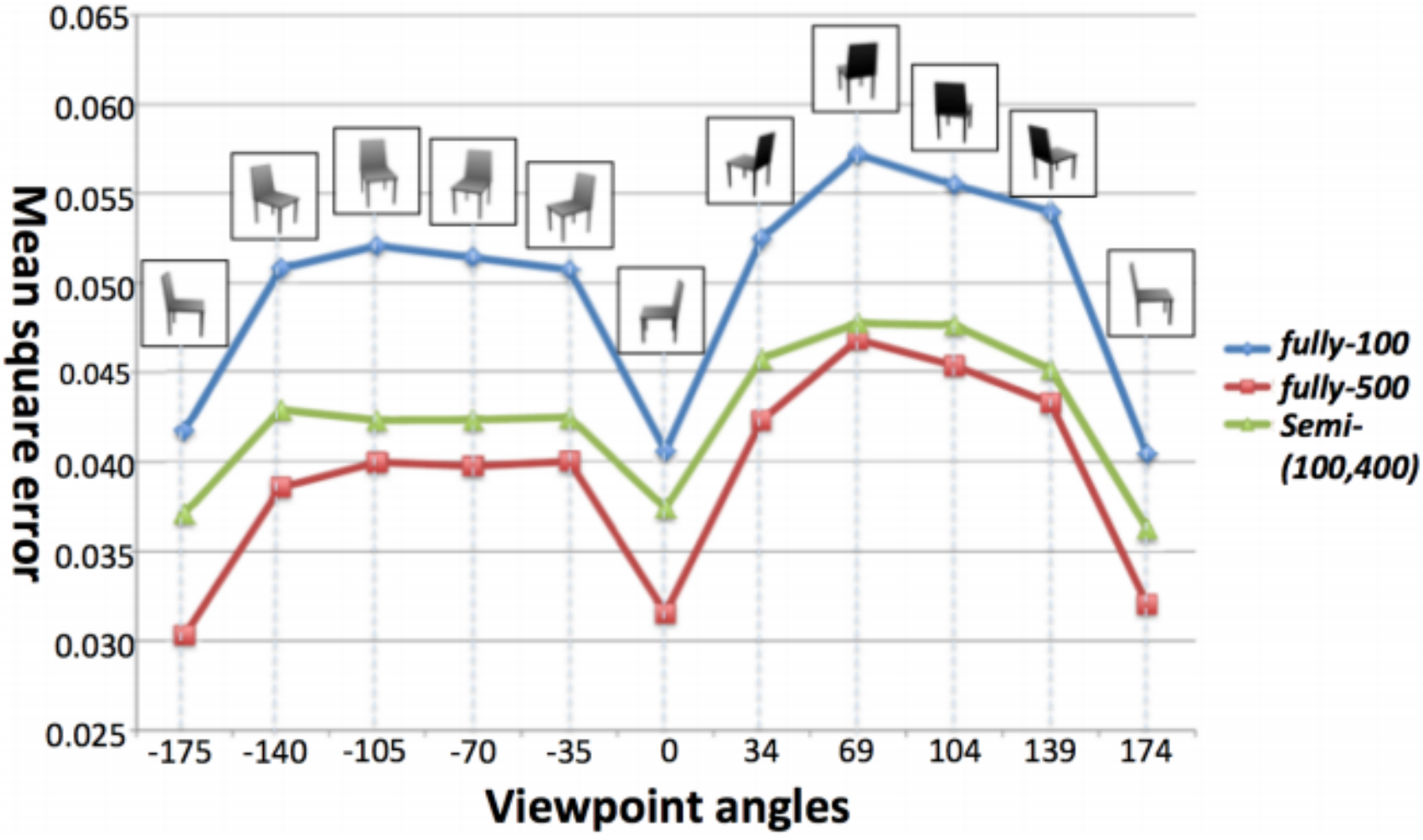

Fig. 16: MSE of novel view synthesis results trained in three settings (using testing chair images under 0° as inputs)

TD-GAN - Qualitative Results

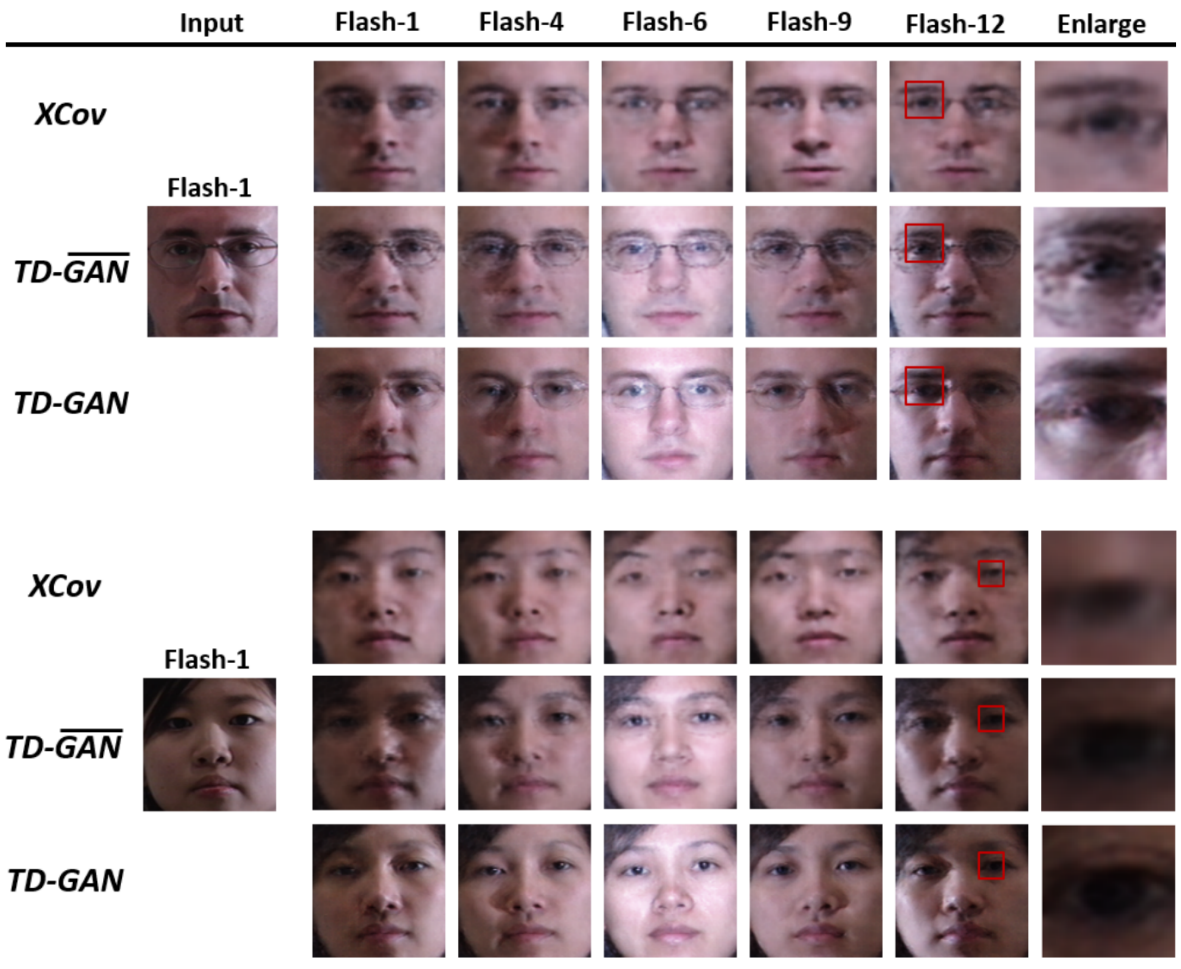

Fig. 17: Illumination transformation of human face results on Multi-PIE dataset

TD-GAN - Quantitative Results

	XCov	*TD-$\overline{GAN}$*	*TD-GAN*
Flash-1	0.5776	0.5623	0.5280
Flash-4	5.0692	0.8972	0.8818
Flash-6	4.3991	1.2509	0.1079
Flash-9	3.4639	0.6145	0.5870
Flash-12	2.4624	0.7142	0.6973
All Flash (mean)	3.8675	0.6966	0.6667

Table 2: MSE ($\times 10^{-2}$) of illumination transformation results

TD-GAN - Possible Applications

- **virtual reality systems**
  e.g. to naturally ‘paste’ persons into a virtual environment by re-rendering of faces for the continuous pose, illumination directions, and various expressions
- **architecture**
- **simulators**
- **video games**
- **movies**
- **visual effects**

Conclusions & Discussions

Conclusions & Discussions

- Assumptions matter.

Conclusions & Discussions

- Assumptions matter. (*are they reasonable?*)
  - images of different tags can be viewed as different domains
  - images and tags record the same object from two different perspectives

Conclusions & Discussions

- Assumptions matter. (*are they reasonable?*)
  - images of different tags can be viewed as different domains
  - images and tags record the same object from two different perspectives
- How to utilize tag information?

Conclusions & Discussions

- Assumptions matter. (*are they reasonable?*)
  - images of different tags can be viewed as different domains
  - images and tags record the same object from two different perspectives
- How to utilize tag information? (*plenty of ways*)
  - multi-domain image-to-image translation
  - enforce consistency between learned disentangled representations for images and tags

Conclusions & Discussions

- Assumptions matter. (*are they reasonable?*)
  - images of different tags can be viewed as different domains
  - images and tags record the same object from two different perspectives
- How to utilize tag information? (*plenty of ways*)
  - multi-domain image-to-image translation
  - enforce consistency between learned disentangled representations for images and tags
- Does disentangling make sense?

Conclusions & Discussions

- Assumptions matter. (*are they reasonable?*)
  - images of different tags can be viewed as different domains
  - images and tags record the same object from two different perspectives
- How to utilize tag information? (*plenty of ways*)
  - multi-domain image-to-image translation
  - enforce consistency between learned disentangled representations for images and tags
- Does disentangling make sense? (*some trade-offs?*)
  - interpretability vs. effectiveness (efficiency)
  - information underneath entangled factors

Questions?

ICGAN - System Overview

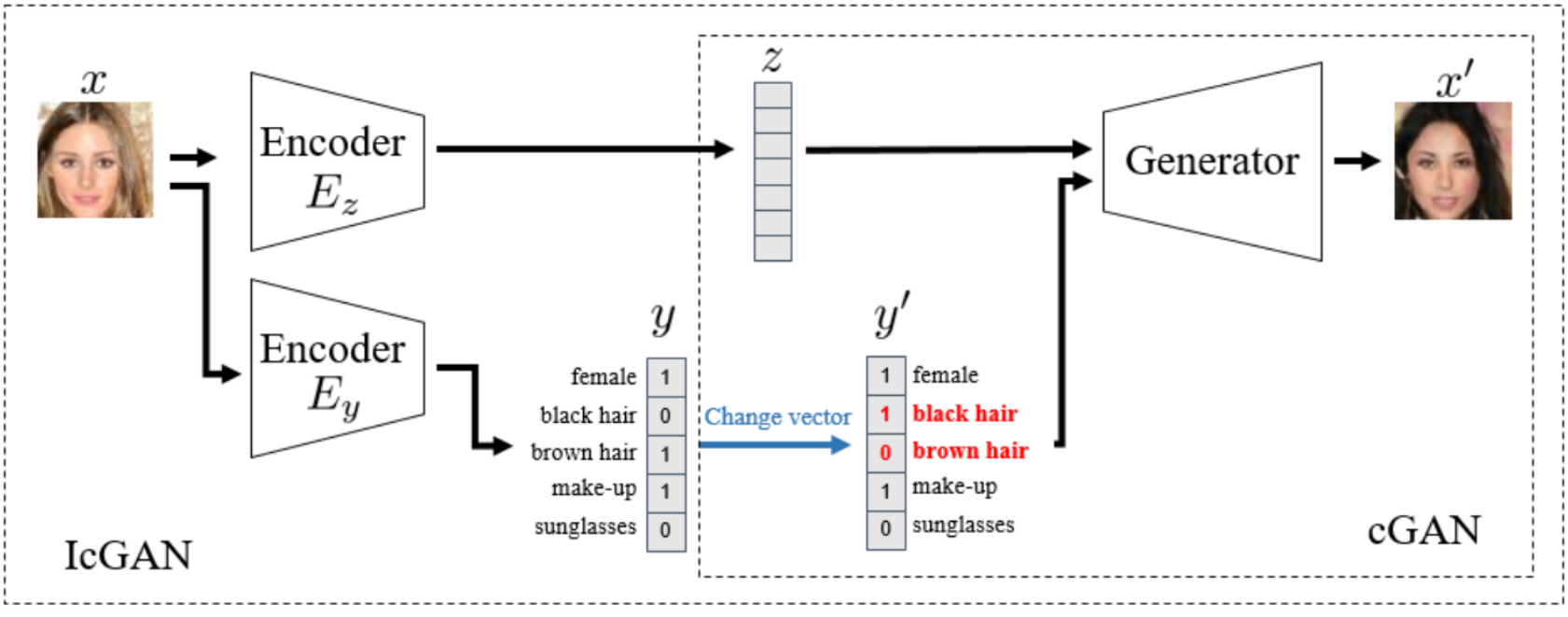

Fig. 18: ICGAN [4] - system overview

DC-IGN - System Overview

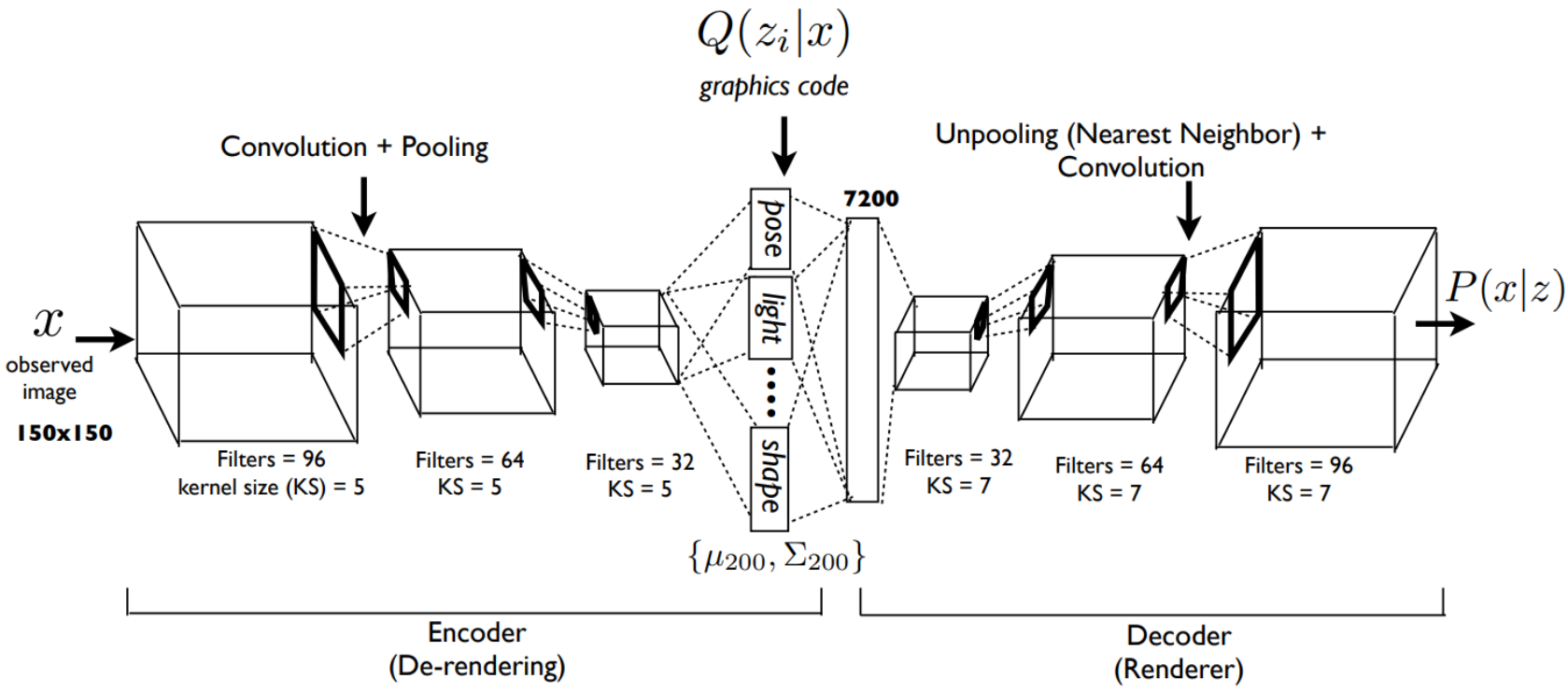

Fig. 19: DC-IGN [5] - system overview

References

[1] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-Domain image-to-image translation," *ArXiv preprint arXiv:1711.09020*, 2017.

[2] C. Wang, C. Wang, C. Xu, and D. Tao, "Tag disentangled generative adversarial networks for object image Re-rendering," in *Proc. 36th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2017.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," in *Proc. NIPS 2016 Workshop on Adversarial Training*, 2016.

[5] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.