Generative Adversarial Image Synthesis with Decision Tree Latent Controller

Takuhiro Kaneko, Kaoru Hiramatsu, Kunio Kashino
NTT Communication Science Laboratories, NTT Corporation

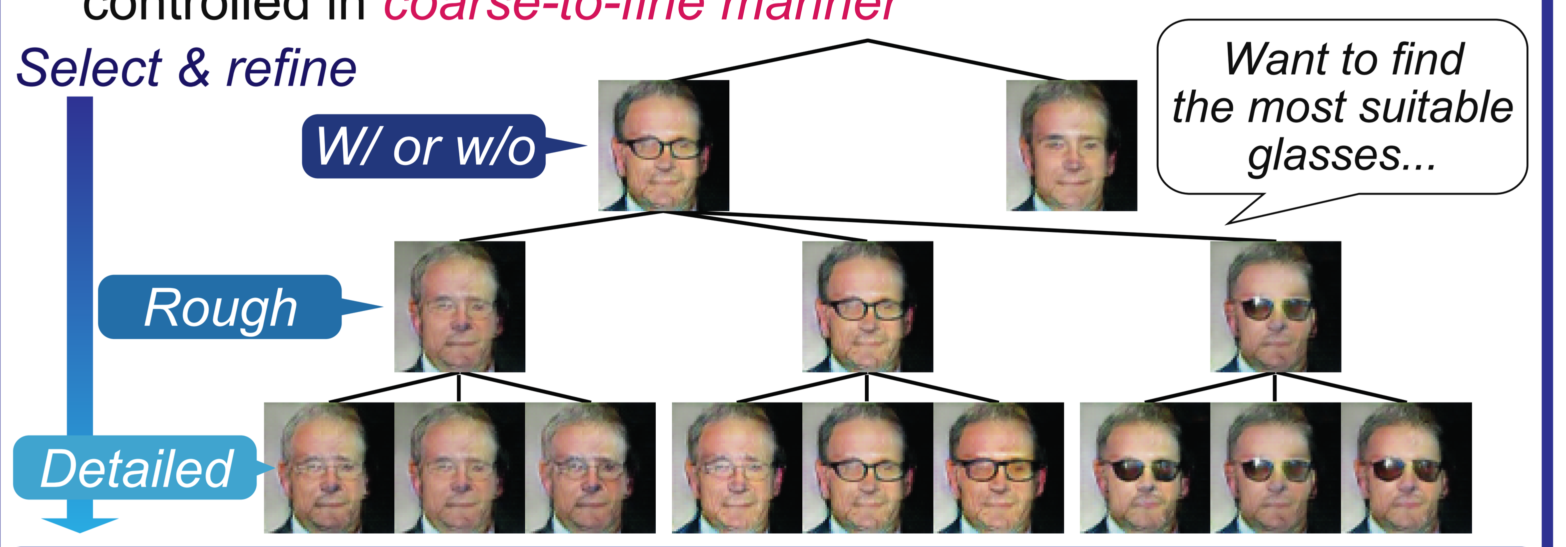
CVPR 2018 SALT LAKE CITY • JUNE 18-22

DTLC-GAN Demos

1 Introduction

Motivation

- Create generative model that enables image generation to be controlled in *coarse-to-fine manner*



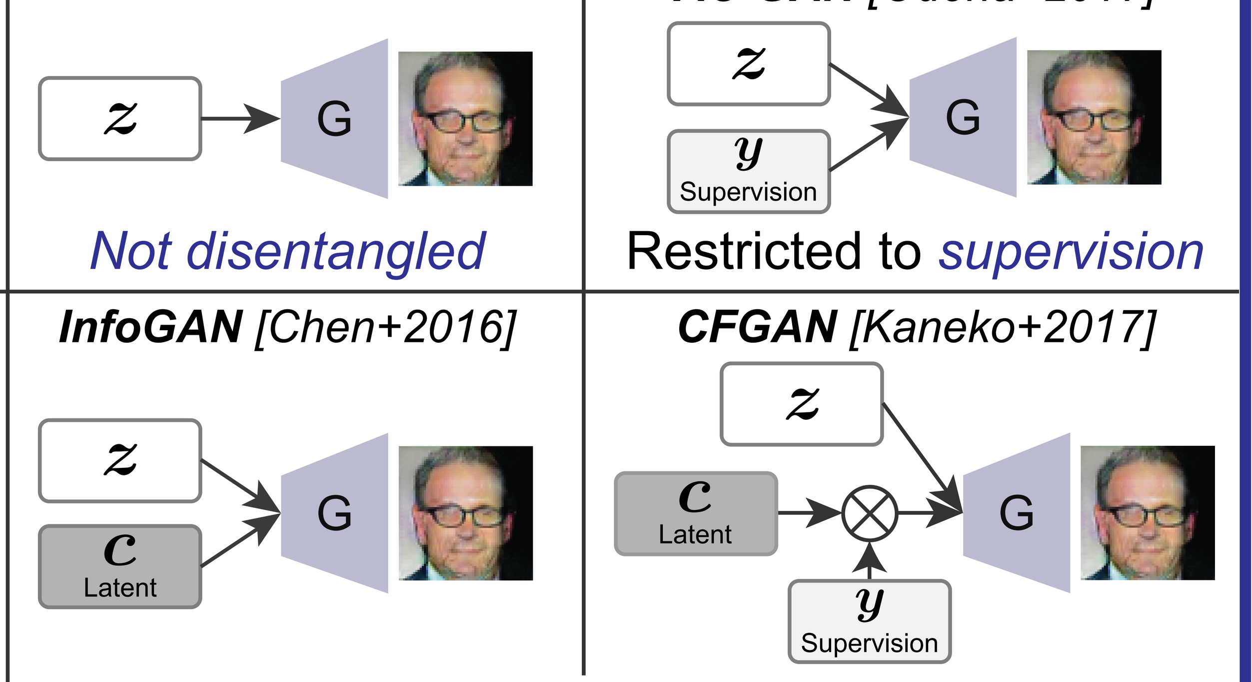
Contributions

- Derive this novel functionality in *deep generative model*
- Propose new extension of GAN called ***DTLC-GAN***
- Discover *hierarchically interpretable representations* with either *unsupervised or weakly supervised* settings

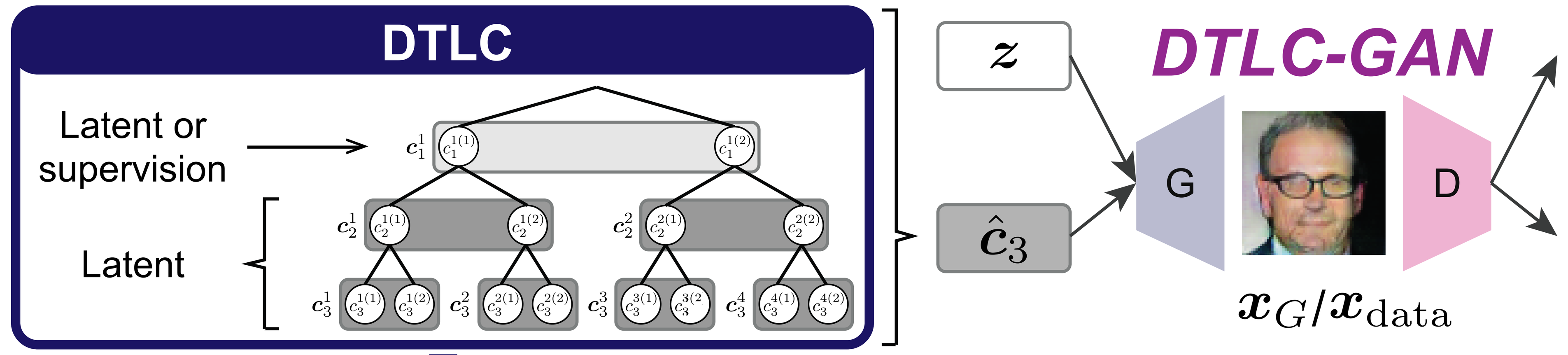
2 Related Work

Relationship to previous GANs

# of disentangled latent layers	Unsupervised	(Weakly) Supervised
0	***GAN*** *[Goodfellow+2014]* Not disentangled	***CGAN*** *[Mirza+2014]* ***AC-GAN*** *[Odena+2017]* Restricted to *supervision*
1	***InfoGAN*** *[Chen+2016]*	***CFGAN*** *[Kaneko+2017]*
	Limited to discovering *one-layer latent* representations	
2, 3, ...	***DTLC-GAN*** Discover *multi-layer latent* representations (Hierarchically interpretable)	



3 Proposed: DTLC-GAN (Decision Tree Latent Controller GAN)



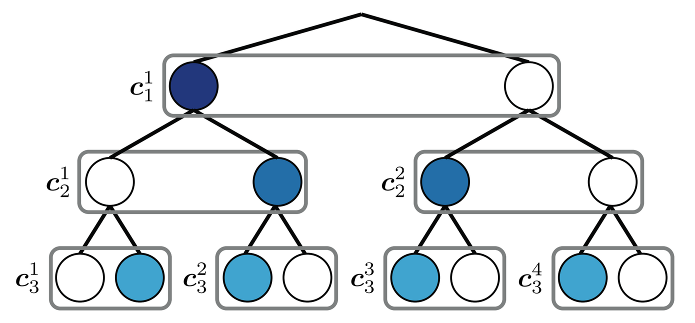
HCMI

$Q_1(c_1|x)$
$Q_2(\hat{c}_2|x, c_1)$
$Q_3(\hat{c}_3|x, \hat{c}_2)$

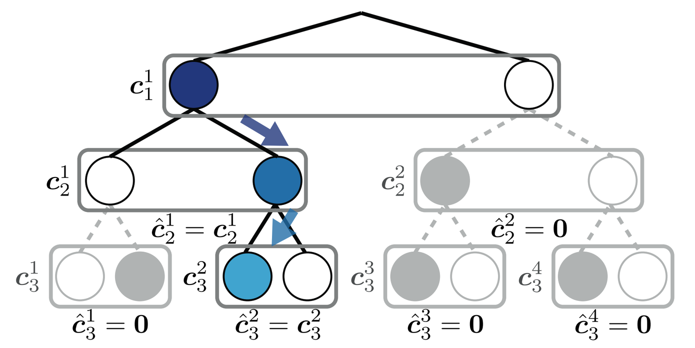
Decision Tree Latent Controller

- In training phase, latent codes are randomly sampled with hierarchical inclusion constraints

Step 1*. In each layer, each code is randomly sampled from categorical distribution independently*



Step 2*. Select child node codes conditioned on parent node codes*



Step 3*. Generate image:* $x_G = G(z, \hat{c}_3)$

Hierarchical Conditional Mutual information Regularization

- Discover hierarchically disentangled representations on basis of *information gain*

Regularization for c_1

- ***Unsupervised*** setting (c_1 is latent)

$$\mathcal{L}_{\text{MI}}(G, Q_1) = \log Q_1(c_1|x_G) \le I(c_1; x_G)$$

Mutual information (InfoGAN)

- ***Weakly supervised*** setting (c_1 is supervised)

$$\mathcal{L}_{\text{AC}}(G, Q_1) = \log Q_1(c_1|x_G) + \log Q_1(c_1|x_{\text{data}})$$

Auxiliary classifier (AC-GAN)

Regularization for $\hat{c}_2, \cdots, \hat{c}_L$

$$\mathcal{L}_{\text{HCMI}}(G, Q_l) = \log Q_l(\hat{c}_l|x_G, \hat{c}_{l-1}) \le I(\hat{c}_l; x_G|\hat{c}_{l-1})$$

Hierarchical conditional mutual information

Full objective

$$\mathcal{L}_{\text{Full}}(D, G, Q_1, \cdots, Q_L)$$
$$= \mathcal{L}_{\text{GAN}}(D, G) - \lambda_1 \mathcal{L}_{\text{MI/AC}}(G, Q_1) - \sum_{l=2}^{L} \lambda_l \mathcal{L}_{\text{HCMI}}(G, Q_l)$$

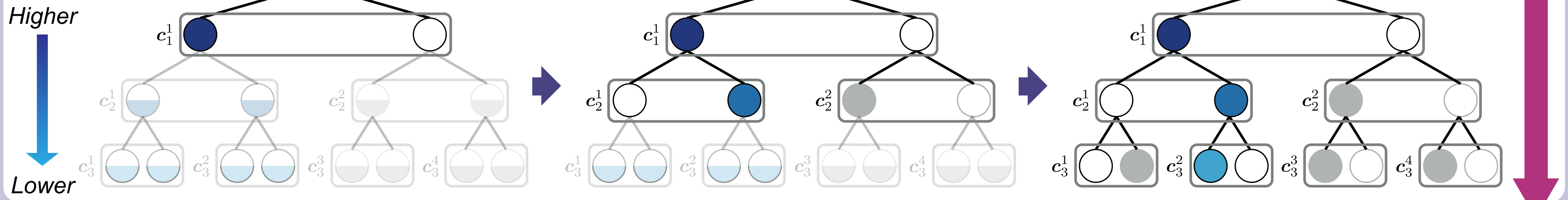
- Minimized for $G, Q_1, \cdots, Q_L$ and maximized for D

Challenge for learning

"How to avoid confusion between ***inter-layer*** *and* ***intra-layer*** *disentanglement"*

Curriculum Learning

Curriculum for DTLC: *In learning higher layer codes, fix and set average value to lower layer codes*



Curriculum for HCMI: *Add regularization from highest layer to lowest layer in step-by-step manner*

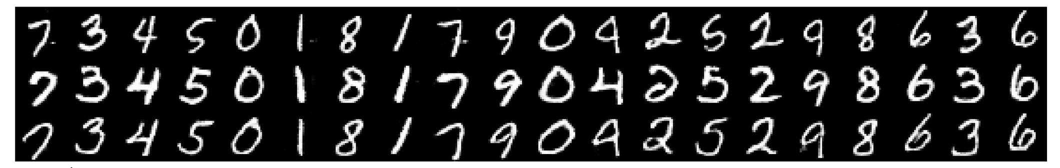
$\mathcal{L}_{\text{GAN}}(D, G) - \lambda_1 \mathcal{L}_{\text{MI}}(G, Q_1)$ ➜ Add $-\lambda_2 \mathcal{L}_{\text{HCMI}}(G, Q_2)$ ➜ Add $-\lambda_3 \mathcal{L}_{\text{HCMI}}(G, Q_3)$

4 Experiments

1. Representation comparison

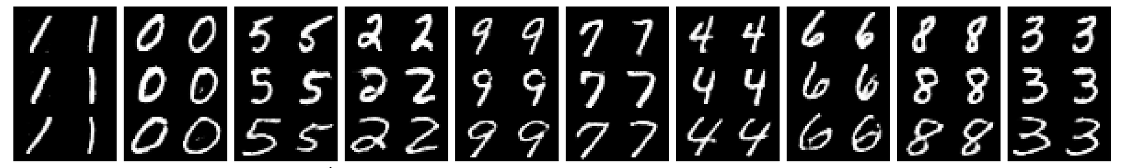
- **Dataset:** MNIST (Unsupervised)
- **Categories:** 20 (flat) vs. 10 x 2 (hierarchical)

(a) **InfoGAN:** 20 *flat* categories



c_1^1: *Fail to disentangle digit types and font styles*

(b) **DTLC-GAN (proposed):** 10 x 2 *hierarchical* categories

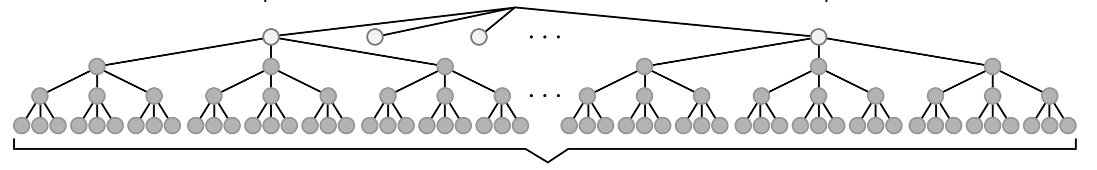


c_1^1 *: Digit types*
$c_2^1, \cdots, c_2^{10}$*: Font styles*
Hierarchically Interpretable

2. Ablation study on curriculum learning

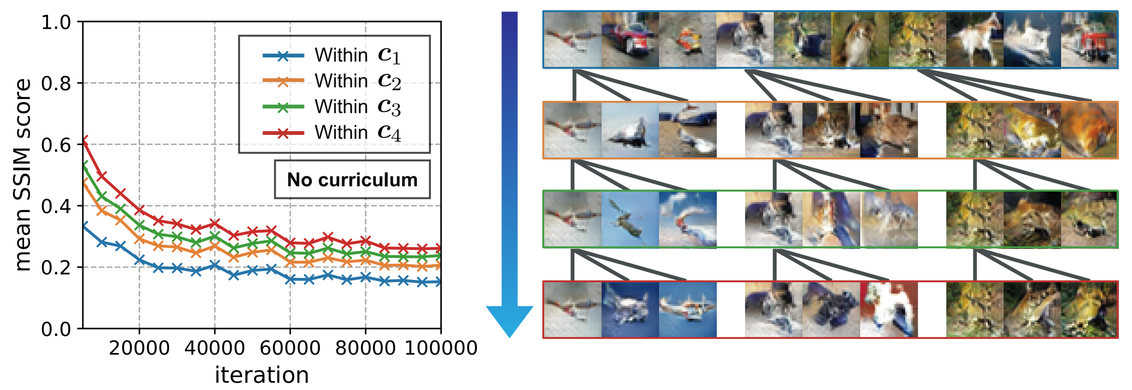
- **Dataset:** CIFAR-10 (Weakly supervised)
- **Categories:** 10 x 3 x 3 x 3 = **270**
- **Evaluation metric:** For each layer, measure inter-category similarity on basis of SSIM

Supervision: airplane, automobile, ..., truck (**10** classes)



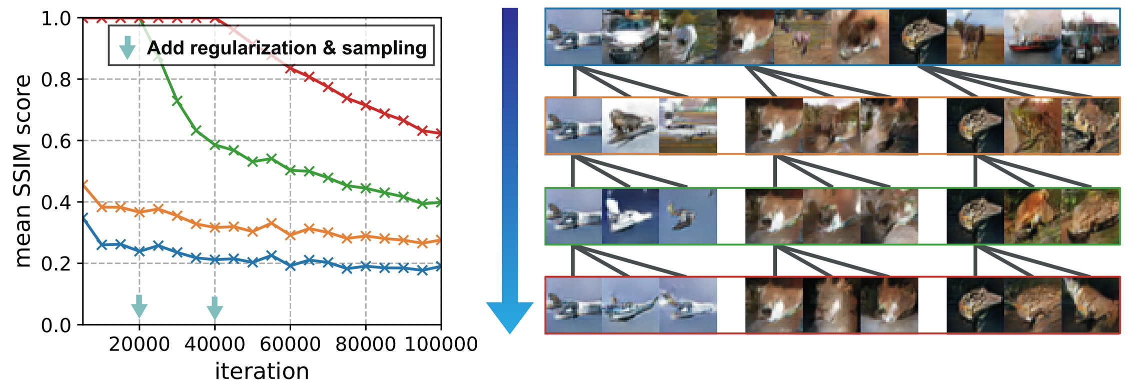
Latent: 10 x 3 x 3 x 3 = **270** categories

(a) **Without curriculum**



Confusion between inter-layer and intra-layer disentanglement

(b) **With curriculum (proposed)**



Similarity becomes larger in lower-layer codes

3. Effect on image quality (w/ WGAN-GP)

- **Dataset:** CIFAR-10 (Unsupervised/supervised)
- **Categories:** 10 x 3^L (L = 0, ..., 4) (= **810** in L = 4)
- **Evaluation metric:** Inception score *[Salimans+2016]*

Model	Unsupervised	Supervised
WGAN-GP	7.86 ± .07†	-
AC/Info-WGAN-GP	7.97 ± .09	8.42 ± .10†
DTLC[2]-WGAN-GP	8.03 ± .12	8.44 ± .10
DTLC[3]-WGAN-GP	8.15 ± .08	8.56 ± .07
DTLC[4]-WGAN-GP	**8.22 ± .11**	**8.80 ± .08**

†Baseline: WGAN-GP ResNet *[Gulrajani+2017]*

Scores improve as # of layers becomes larger

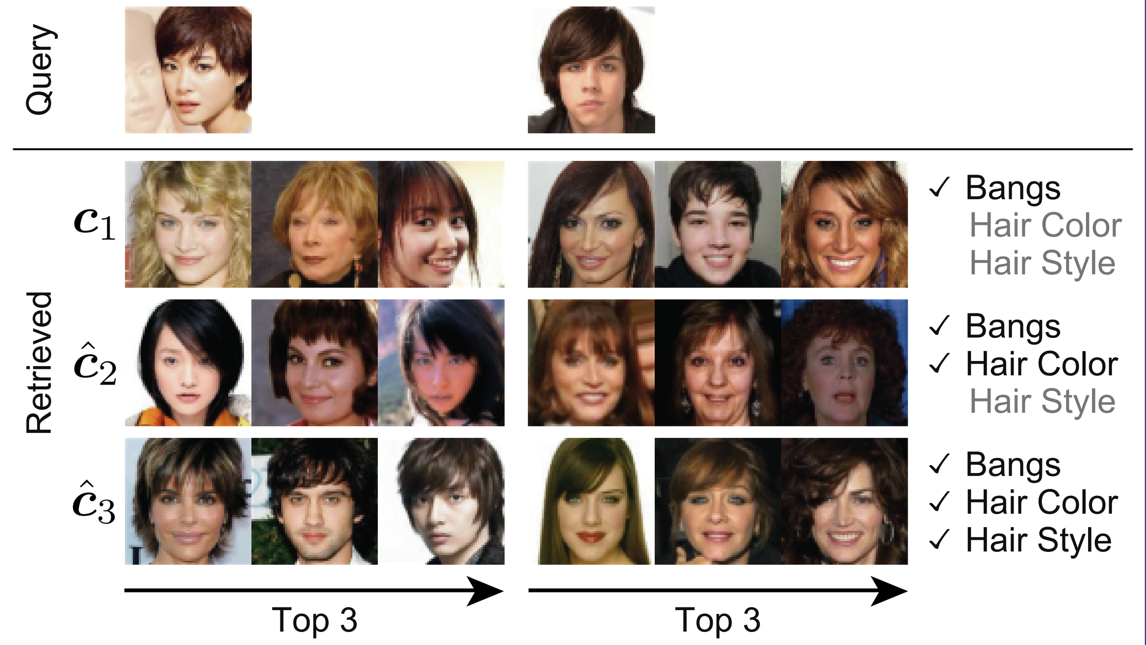
4. Extension to continuous codes

- **Dataset:** 3D Faces (Unsupervised)
- **Categories:** 5 (discrete) x 1 (continuous)

Second-layer codes learn continuous representations conditioned on first-layer codes

5. Application to image retrieval

- **Dataset:** CelebA (Weakly supervised)
- **Categories:** 1 (w/o attribute) + 1 (w/ attribute) x 3 x 3



Details match more in lower-layer codes