JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Exploring Intra-class Variation Factors with Learnable Cluster Prompts for Semi-supervised Image Synthesis

Yunfei Zhang[1*], Xiaoyang Huo[1*], Tianyi Chen[1], Si Wu[1,2,3†], and Hau San Wong[4]

[1]School of Computer Science and Engineering, South China University of Technology

[2]Peng Cheng Laboratory

[3]PAZHOU LAB

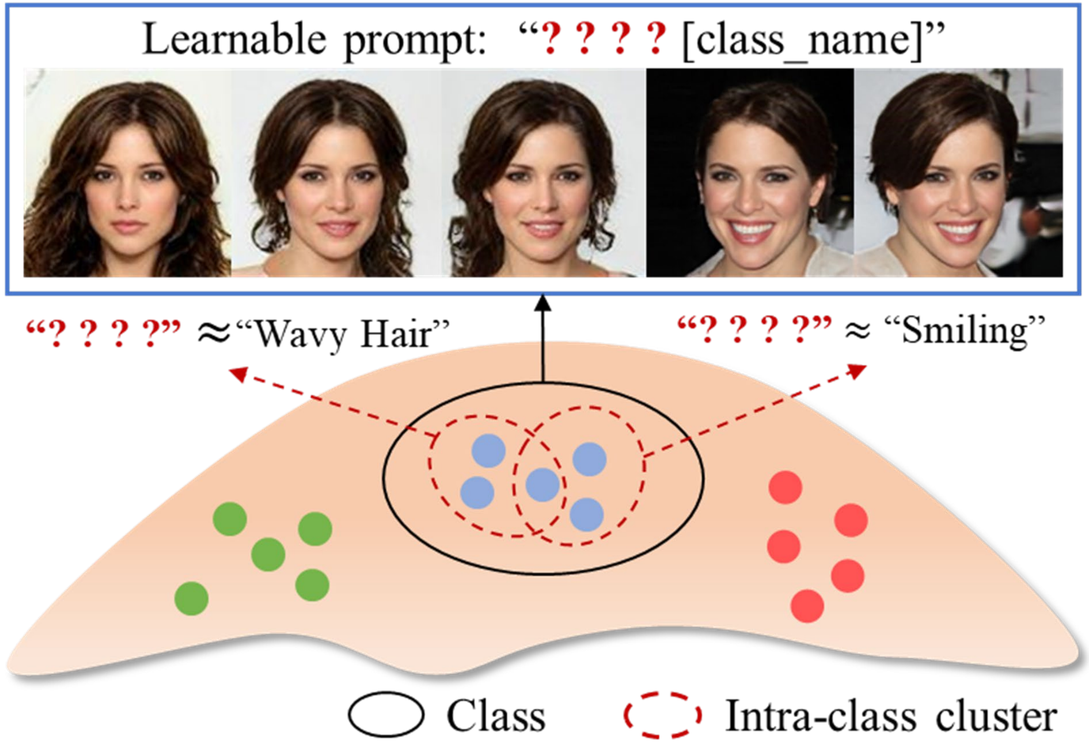
[4]Department of Computer Science, City University of Hong Kong

*{cszhangyunfei@mail.scut.edu.cn, cswusi@scut.edu.cn}

TUE-PM-312

Proposed Approach

- We associate the **intra-class cluster label embeddings** with the cluster semantics, and the expressiveness of their combination is higher than that of a single class label embedding for **capturing multiple underlying modes with diverse visual appearances**.
- We leverage the language-vision pre-training and **represent the clusters with learnable prompts**.
- To guide the generator to capture intra-class variation factors, the **cluster prompts serve as conditional information** and are jointly learnt in the adversarial training process.

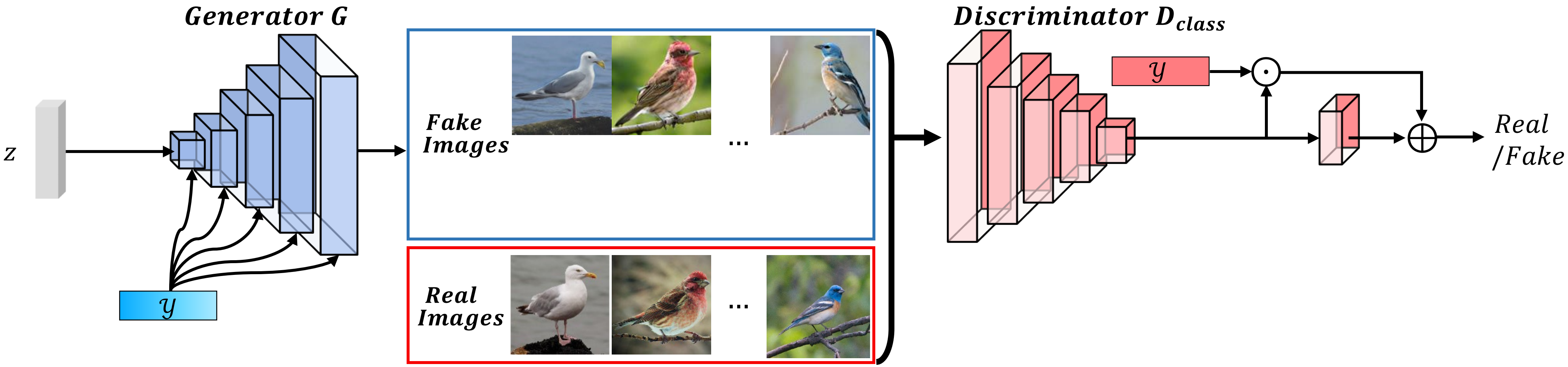


Structure of LCP-GAN

- For class-conditional image generation, we design a cluster conditional generator by injecting a combination of intra-class cluster label embeddings, and further incorporate a real-fake classification head on top of CLIP to distinguish real instances from the synthesized ones, conditioned on the combination of the learnable cluster-specific prompts.
- Our framework consists of five components: a generator G, a class-conditional discriminator D_{class}, a CLIP-based discriminator D_{prompt}, and a classifier C.

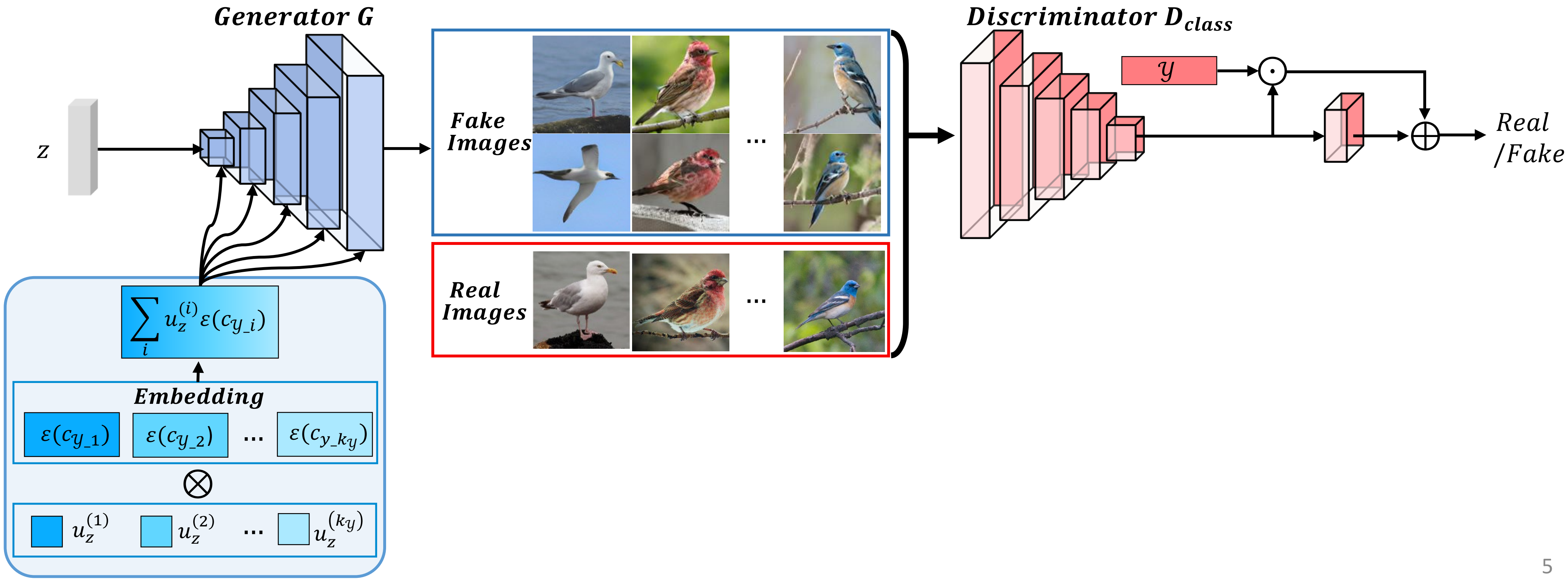
Structure of LCP-GAN

- Firstly, we implemented a general semi-supervised class-conditional generative model with a class-conditional D_{class}. G synthesizes fake images from the latent code z together with a random class label $\mathcal{Y}$. Only one single label embedding is learnt per class, which is insufficient to account for large intra-class variance.



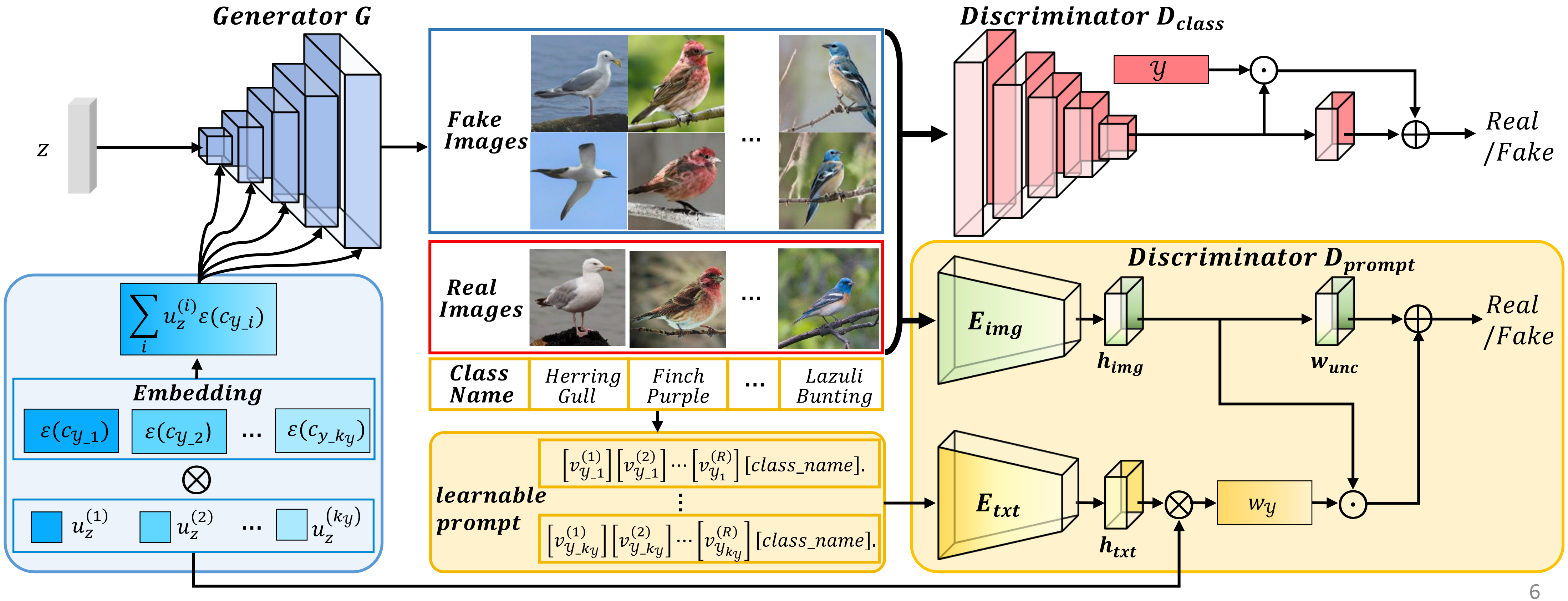
Structure of LCP-GAN

- To better match class-specific data distribution, the generator learns to synthesize high-fidelity images, conditioned on a combination of intra-class cluster label embeddings $\sum_i u_z^{(i)} \varepsilon(c_{y\_i})$, and our model can capture multiple underlying modes with diverse visual appearances.



Structure of LCP-GAN

- Considering that natural language can express a wide range of visual concepts, we build a CLIP-based discriminator D_{prompt} to learn the cluster-specific context vectors and guide generator G to capture cluster semantics.



Cluster-conditional Generation

- For labeled images, we perform intra-class data partitioning via soft k-means clustering in the feature space of a ResNet pre-trained on ImageNet.

- The images of each class $\mathcal{Y}$ are divided into $k_{\mathcal{Y}}$ clusters, and the corresponding prototypes $\boldsymbol{\rho}_{\boldsymbol{\mathcal{Y}}} = \{\rho_{\mathcal{Y}\_1}, \ldots, \rho_{\mathcal{Y}\_k_{\mathcal{Y}}}\}$ are computed by the weighted mean vectors of the embedded images as follows:

$$\rho_{\mathcal{Y}\_i} \leftarrow (1-\mu)\rho_{\mathcal{Y}\_i} + \mu u_x^{(i)} f(x)$$

- In the above equation, the degree $u_x^{(i)}$ to which x belongs to intra-class cluster i is computed as follows:

$$u_x^{(i)} = \frac{\exp(\cos(\rho_{\mathcal{Y}\_i}, f(x)/\delta))}{\sum_j \exp(\cos(\rho_{\mathcal{Y}\_j}, f(x))/\delta)}$$

Cluster-conditional Generation

- For image generation, we simulate the condition by randomly sampling a coefficient vector $u_z = [u_z^{(1)}, \dots, u_z^{(k_y)}]$ to combine the cluster label embeddings, based on which the generator synthesizes an image conditioned as follows:

$$x_z = G(z, \textstyle\sum_i u_z^{(i)} \varepsilon(c_{y\_i})).$$

- Where $\varepsilon(\cdot)$ denotes the learnable embedding layer.

- By using combination we are able to capture multiple underlying modes with diverse visual appearances.

Learning Cluster-specific Prompts

- Inspired by CoOp [1], we model the cluster-specific context words with learnable vectors, and the context words are in the form of continuous vectors that have the same dimension as the word embeddings. Specifically, we adopt the prompt form as follows:

$$t_{y\_i} = \left[v_{y\_k_y}^{(1)}\right]\left[v_{y\_k_y}^{(2)}\right]\cdots\left[v_{y_{k_y}}^{(R)}\right][class\_name].$$

- where the context vectors $\left\{v_{y_i}^{(\mathrm{r})}\right\}_{\mathrm{r}=1}^{R}$ are learnable, and the word embedding vector of the i-th class name is used in the token position $[class\_name]$.

Learning Cluster-specific Prompts

- We perform the prompt-conditional adversarial training, and the corresponding discriminator D_{prompt} is built by incorporating a real-fake classification head on top of the CLIP encoders. The conditional identification weight is defined as follows:

$$w_y = \sum_{\mathrm{i}} u_x^{(i)} h_{txt}(E_{txt}(t_{y\_i}))$$

- The prediction probability is computed as follows:

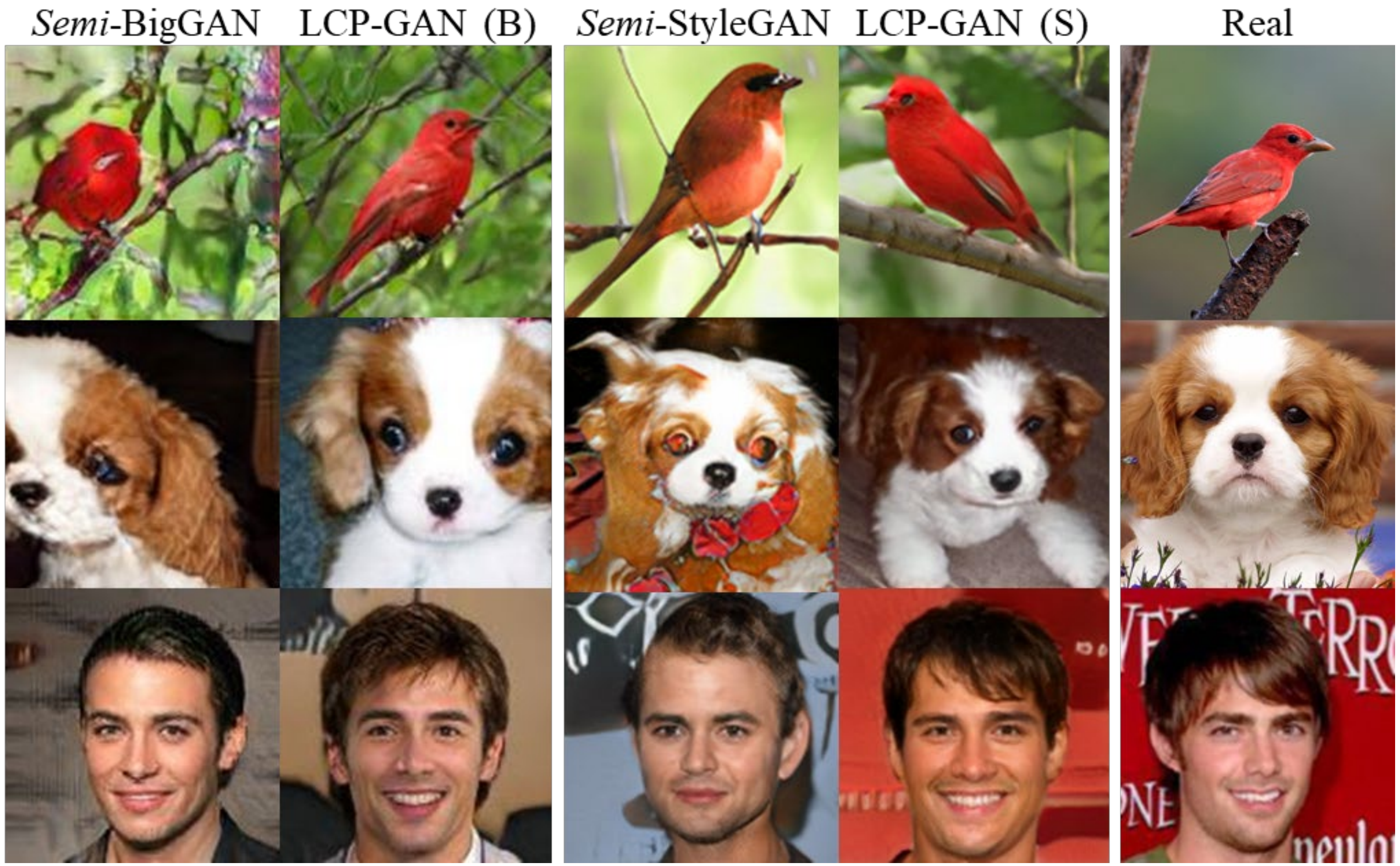
$$D_{prompt}(x, \boldsymbol{t_y}, u_x) = w_y \cdot h_{img}\left(E_{img}(x)\right) + \mathrm{w_{unc}} \cdot h_{img}(E_{img}(x))$$

- where w_{unc} denotes the unconditional identification weight.

- By incorporating the D_{prompt} , we are able to pull the representations of cluster-specific images closer to the corresponding prompt.

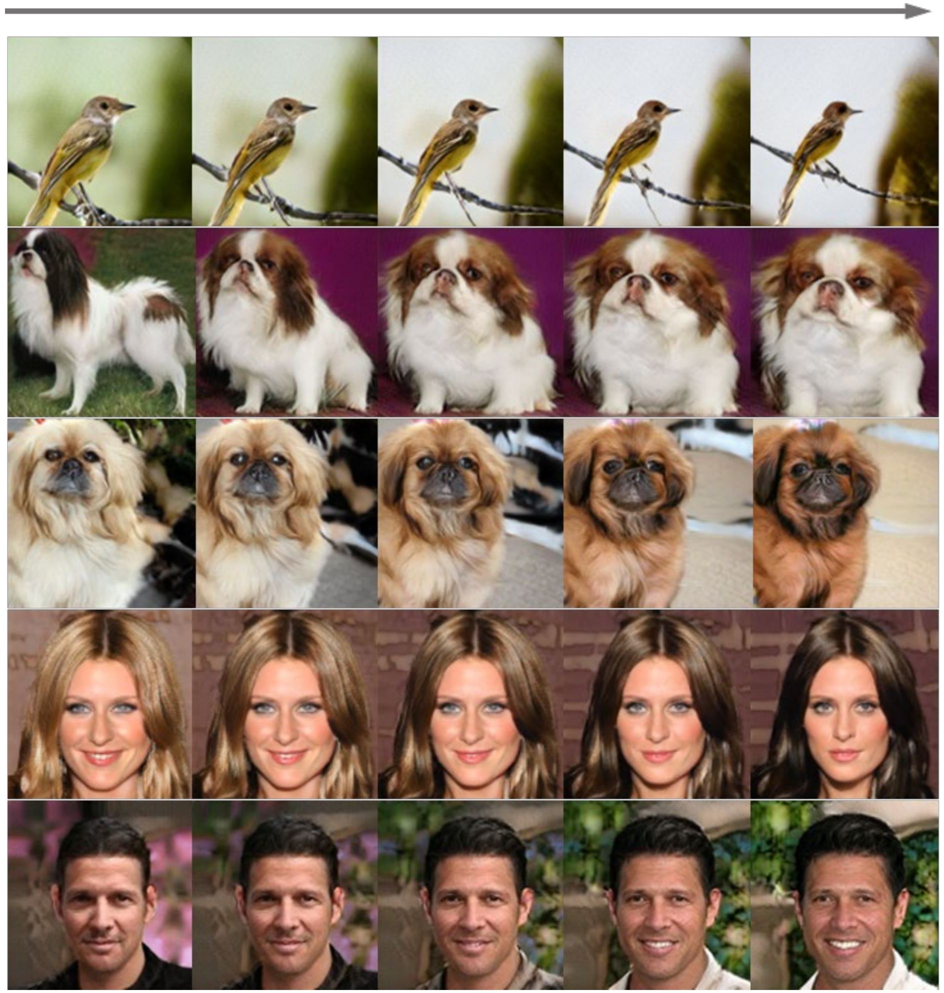
Class-Conditional Image Synthesis

- Synthesis results of LCP-GAN and the base models



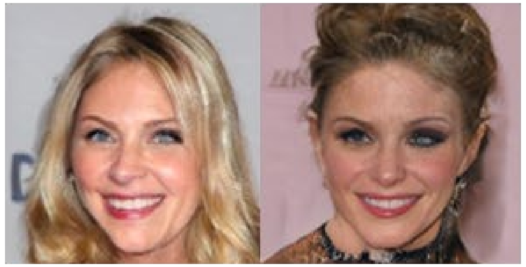
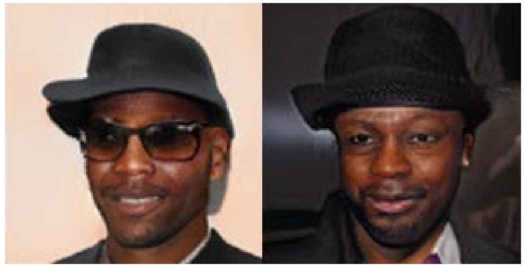
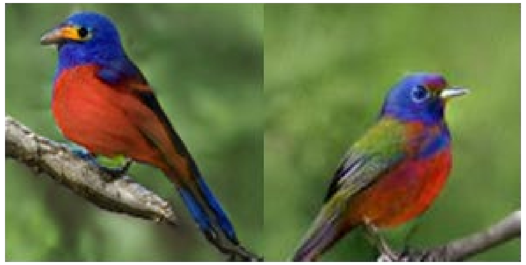
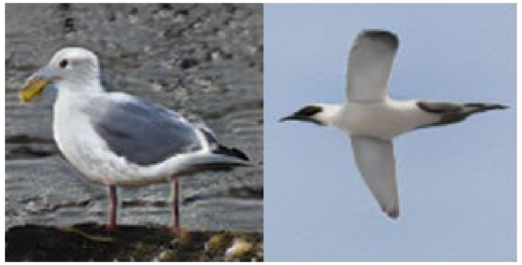
Linear interpolation

- Interpolation results of LCP-GAN by linearly combining the paired cluster label embeddings.



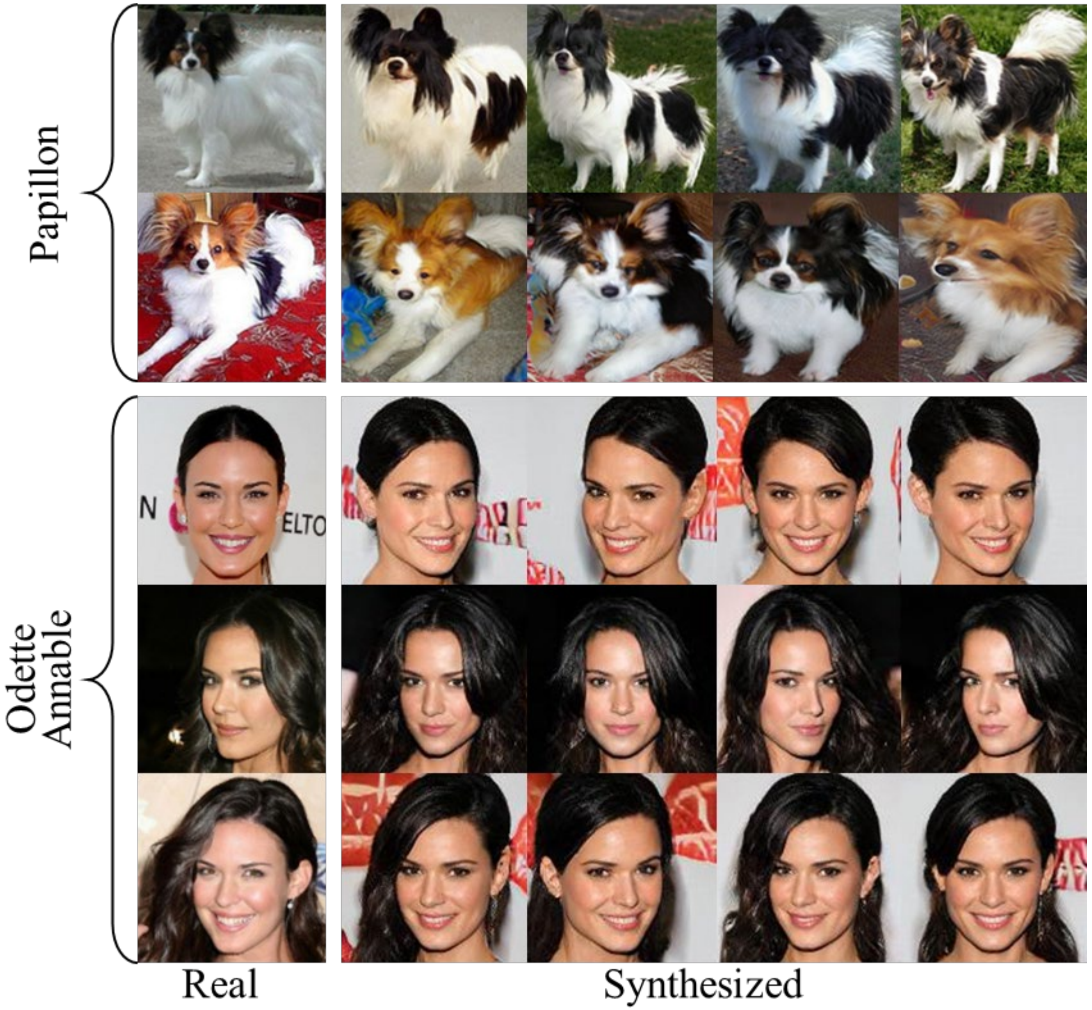
Learnt prompts

- The class-specific nearest words and synthesized images.

1. Blond Hair 2. Pale Skin 3. Smiling 4. Brown Hair 5. Heavy Makeup		1. Male 2. Wearing Hat 3. Eyeglasses 4. Big Nose 5. Mustache	
1. Blue Crown 2. Green Wings 3. Orange Chest 4. Red Breast 5. Black Eyes		1. Gray Wing 2. Yellow Bill 3. Black Tail 4. Long Leg 5. White Throat	

Semantically meaningful clusters

- The cluster-specific images synthesized by LCP-GAN.



Conclusion

- We extend a semi-supervised GAN framework to **learn from intra-class clusters**, and enable class-specific image synthesis to be **conditioned on the combination of cluster label embeddings**.
- We leverage the language vision pre-training and jointly **learn cluster-specific prompts through prompt-conditional adversarial training.**
- Our design is able to **discover a wide range of semantically meaningful intra-class variation factors** and achieve superior performance on multiple semi-supervised image synthesis tasks.

Thank you