UDT: Unsupervised Discovery of Transformations between Fine-Grained Classes in Diffusion Models

Reality Lab, Soongsil University
https://reality.ssu.ac.kr
https://ssu-reality-lab.github.io/UDT/
Soongsil University
Youngjae Choi*, Hyunsuh Koh*, Hojae Jeong*, Byungkwan Chae*, Sungyong Park, Heewon Kim

BMVC 2025

Introduction

Background

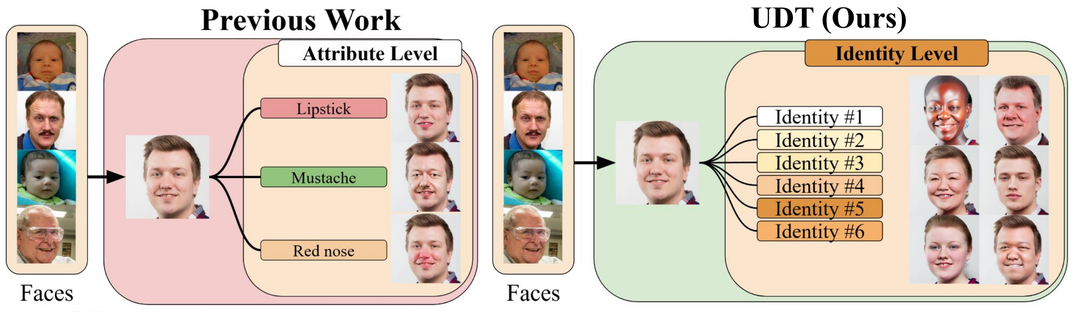
- Diffusion models are a core generative AI technology for creating high-fidelity, photorealistic images from text prompts.
- A key research goal is to explore their latent spaces to find methods for precise, controllable editing of image attributes.

Motivation / Contribution

- Unsupervised methods often fail to maintain consistency in fine-grained edits (e.g., breed-to-breed or species-to-species translations).
- UDT enables reliable child-class translations that produce consistent images of the same class across diverse input images.

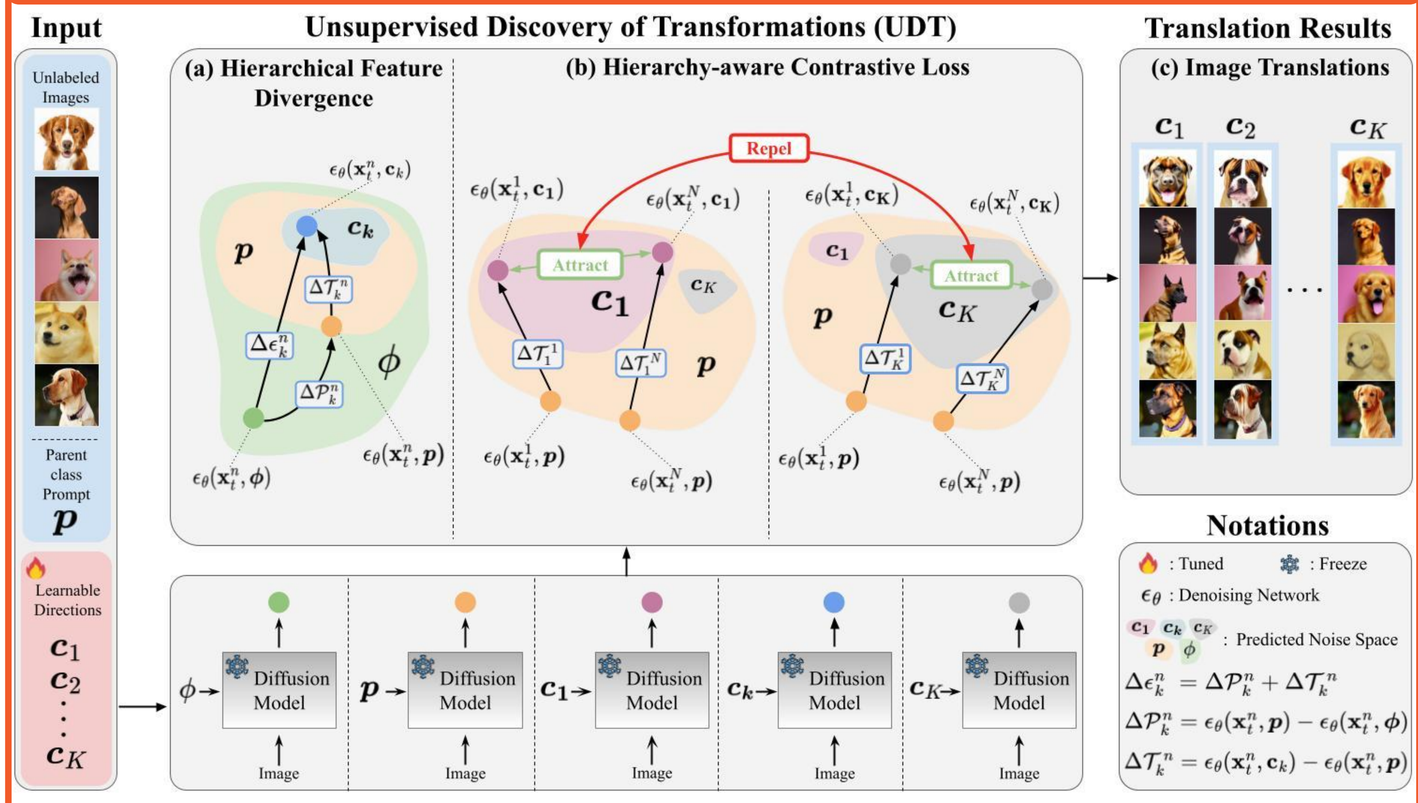
Previous Work

- Existing methods are often limited to simple attribute changes and lack the consistency and diversity required for fine-grained class transformations.



➔ UDT upgrades unsupervised exploration from attribute-level change to class-level control (e.g., identity-to-identity translation for faces)

UDT Framework



➔ UDT uses a Hierarchical Structure to find consistent and diverse ways to transform images.

(a) Hierarchy Feature Divergence

- Decomposes the predicted noise $\Delta\epsilon_k^n$ into a parent-class $\Delta\mathcal{P}_k^n$ and a child-class $\Delta\mathcal{T}_k^n$.

$$\Delta\epsilon_k^n = \underbrace{\epsilon_\theta(\boldsymbol{x}_t^n, \boldsymbol{c}_k) - \epsilon_\theta(\boldsymbol{x}_t^n, \boldsymbol{p})}_{\Delta\mathcal{T}_k^n} + \underbrace{\epsilon_\theta(\boldsymbol{x}_t^n, \boldsymbol{p}) - \epsilon_\theta(\boldsymbol{x}_t^n, \boldsymbol{\phi})}_{\Delta\mathcal{P}_k^n}.$$

(b) Hierarchy-aware Contrastive Loss

- Applies contrastive learning to child-class $\Delta\mathcal{T}_k^n$.

$$\mathcal{L}_{\mathcal{T}} = -\log \frac{\sum_{a=1}^{|\boldsymbol{X}'|}\sum_{b=1}^{|\boldsymbol{X}'|}\mathbf{1}[a\neq b]\exp\left(\mathrm{sim}(\Delta\mathcal{T}_j^a, \Delta\mathcal{T}_j^b)\right)}{\sum_{a=1}^{|\boldsymbol{X}'|}\sum_{i=1}^{|C'|}\mathbf{1}[i\neq j]\exp\left(\mathrm{sim}(\Delta\mathcal{T}_j^a, \Delta\mathcal{T}_i^a)\right)},$$

Training Objective

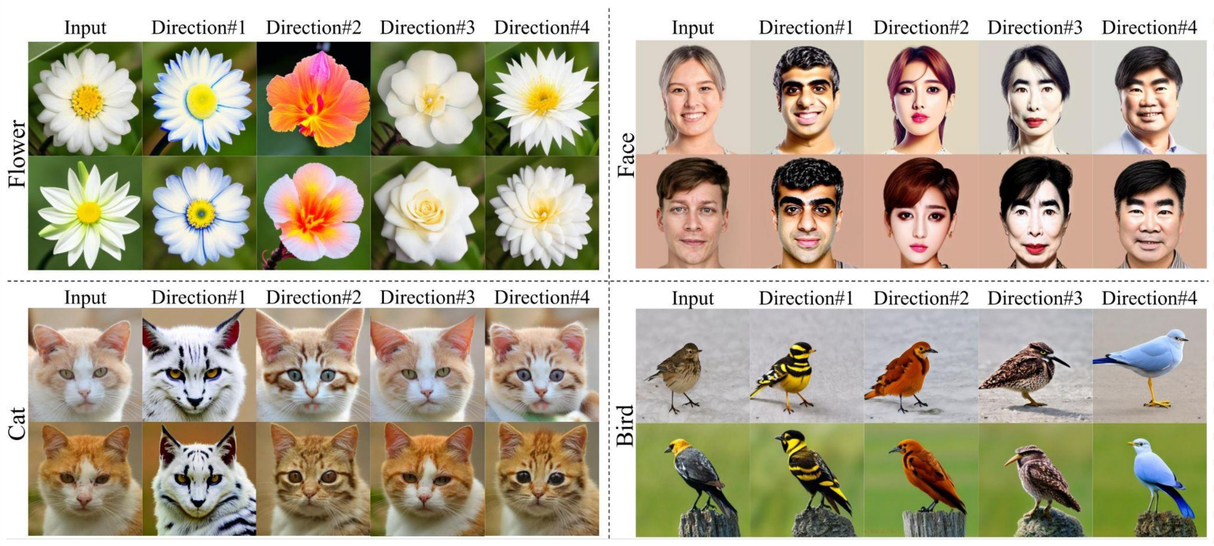
- Balances transformation consistency $\mathcal{L}_{\mathcal{T}}$ with training stability $\mathcal{L}_{\text{Reg}}$:

$$\mathcal{L} = \mathcal{L}_{\mathcal{T}} + \lambda_{\text{Reg}}\mathcal{L}_{\text{Reg}},$$

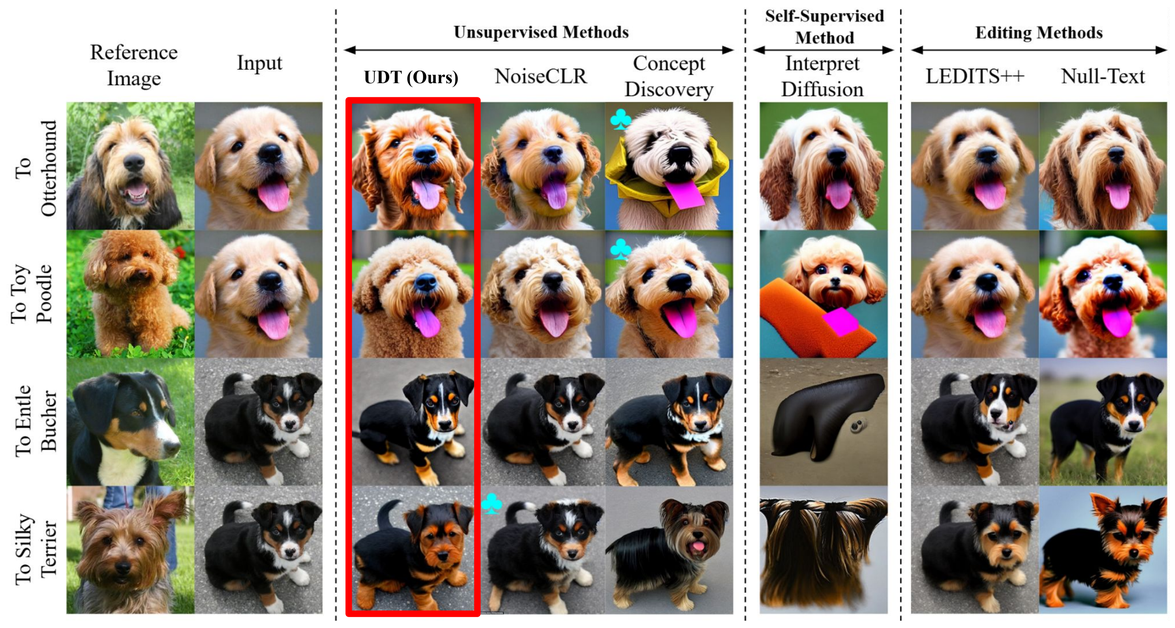
Experiments

Qualitative Comparison

- Generalization Across Diverse Domains



- UDT successfully performs target breed transformations.



Quantitative Evaluation (CLIP-based)

- Classification Accuracy on Class Translation.

	Boxer		Beagle		Golden Retriever		Siberian Husky		Miniature Poodle	
	Previous Work	Ours	Previous Work	Ours	Previous Work	Ours	Previous Work	Ours	Previous Work	Ours
Boxer	-13.90	**11.20**	-4.83	9.35	0.29	0.57	1.30	11.68	1.43	-6.43
Beagle	-16.35	4.83	-6.94	**16.52**	-12.33	3.35	-3.16	8.91	-3.44	0.63
Golden Retriever	-8.12	-6.24	-1.67	2.98	-1.59	**40.07**	5.73	12.70	2.02	3.29
Siberian Husky	-14.10	-11.63	-17.15	-7.58	-13.45	0.95	-8.56	**51.27**	-4.86	-0.60
Miniature Poodle	-20.45	-8.86	-16.36	-4.92	-17.38	17.79	-5.39	6.90	3.18	**49.69**

➔ CLIP Score↑

- Class Diversity on Class Translation.

Breed	# Predicted Breeds: NoiseCLR	Ours
Appenzeller	17	**44**
Chihuahua	11	**46**
Eng. Foxhound	20	**58**
Fox Terrier	30	**59**
Golden Ret.	7	**42**
Labrador Ret.	6	**38**
Yorkshire Ter.	18	**66**
Avg.	15.57	**50.43**

➔ Class Diversity↑

Conclusion

- **UDT** uses unsupervised learning to **discover fine-grained transformation** directions without labels.
- Through **hierarchical noise decomposition** and **contrastive learning**, it **achieves consistent and diverse class-level edits.**
- Experiments across animals, faces, and flowers confirm its **effectiveness in label-free fine-grained translation.**

Future Works

- **Fine-grained data augmentation**
  UDT enables **diverse, copyright-free character and asset creation** for media production.
- **Copyright-free character & asset generation**
  UDT for automatic, copyright-free character and asset generation in media.