TransGaGa: Geometry-Aware Unsupervised Image-to-Image Translation

Wayne Wu[1] Kaidi Cao[2] Cheng Li[1] Chen Qian[1] Chen Change Loy[3]

[1]SenseTime Research [2]Stanford University
[3]Nanyang Technological University

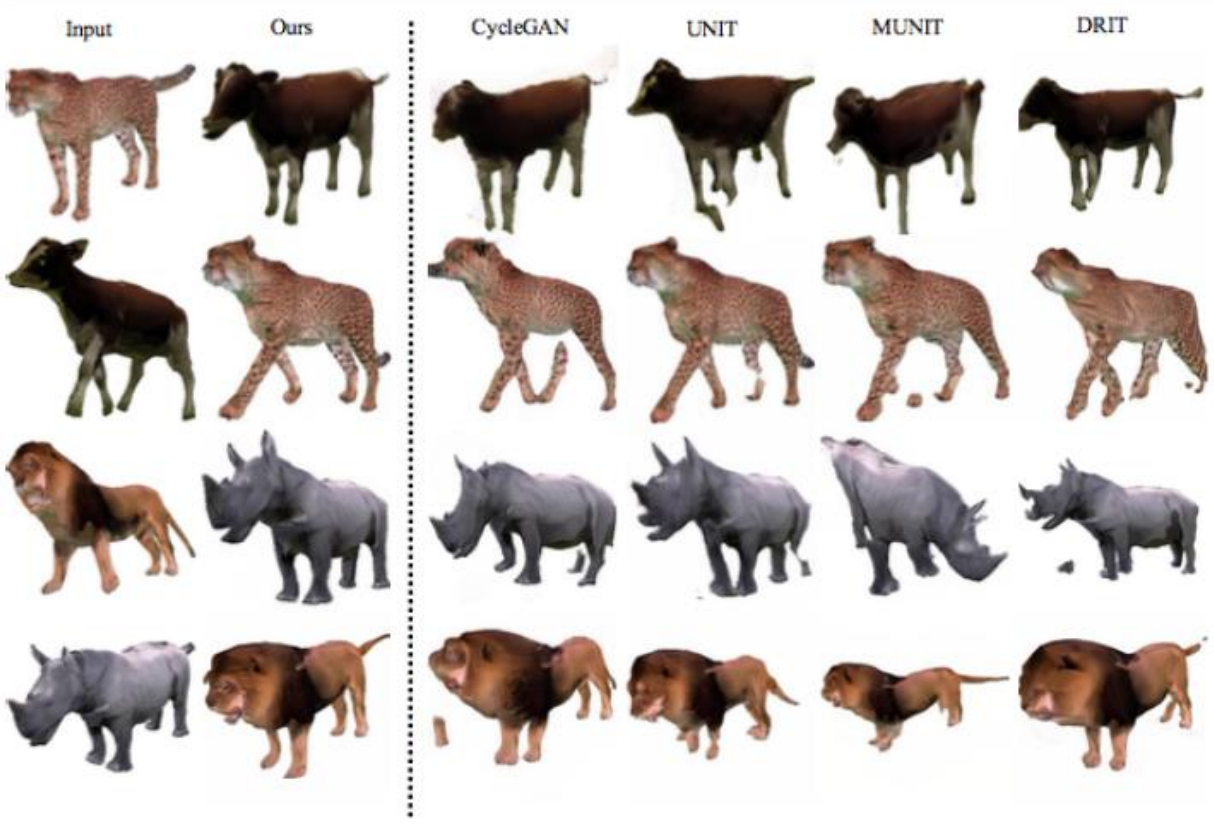
{wuwenyan, chengli, qianchen}@sensetime.com kaidicao@cs.stanford.edu
ccloy@ntu.edu.sg

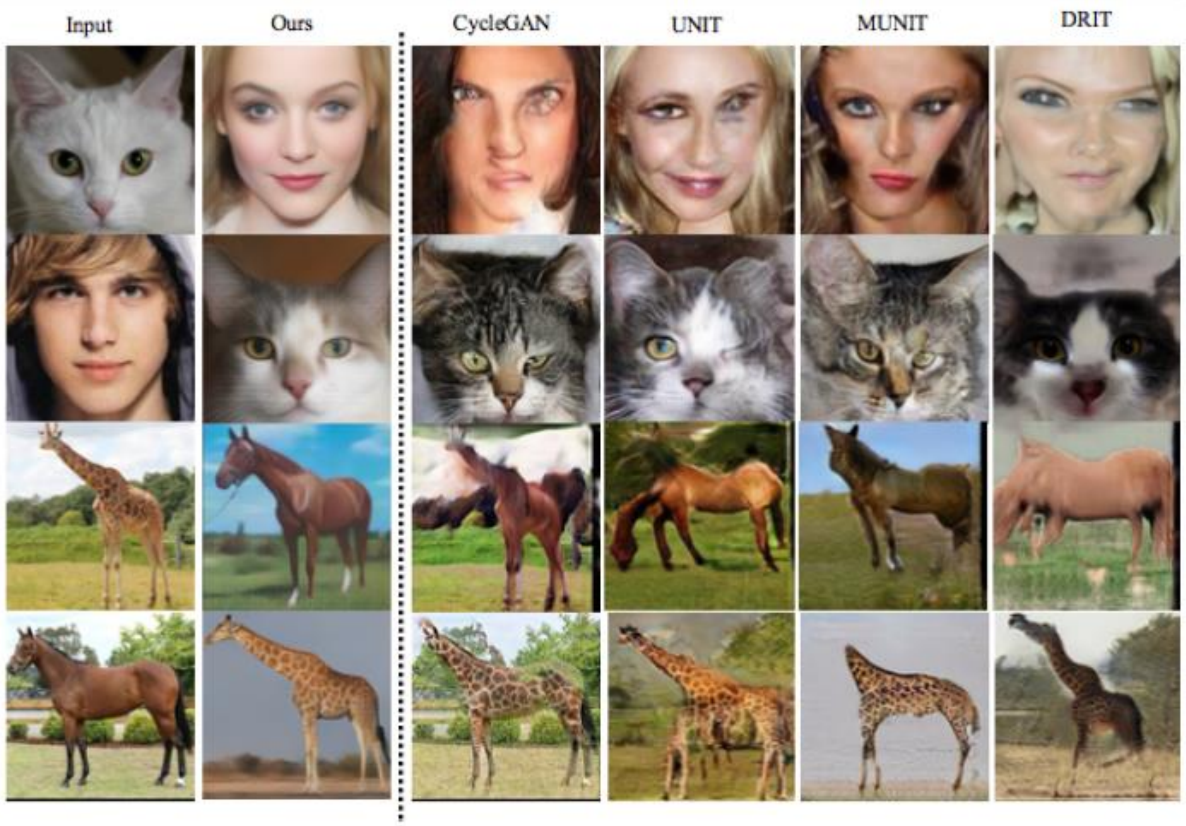
19.06.25

발표자 : 김용규

Motivation

- Learning a translation across <u>large geometry variations</u> always ends up with failure.
  So, They present a novel disentangle-and-translate framework to tackle the complex objects image-to-image translation task.



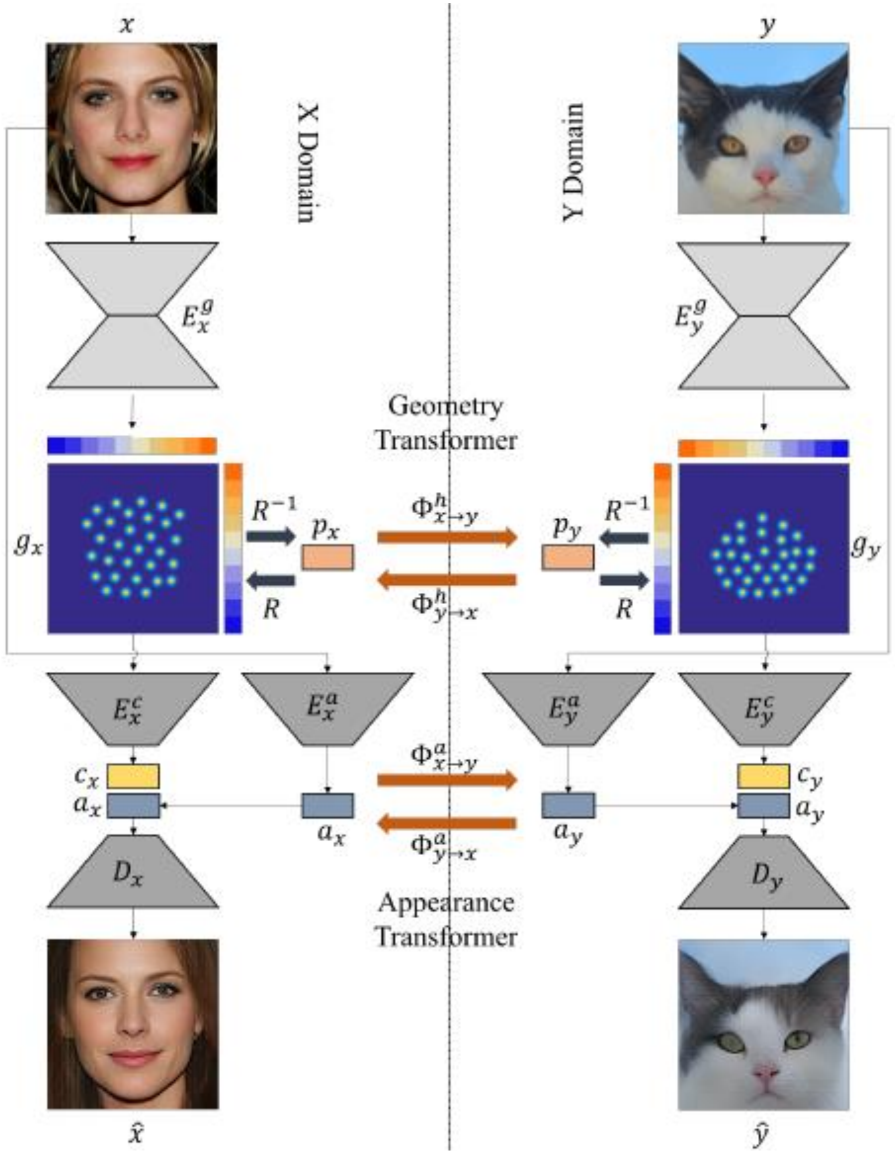


Contribution

- They propose a novel framework for unsupervised image-to-image translation. Instead of directly translating on the image space, we build the mapping between two domains on their disentangled latent appearance-geometry space.

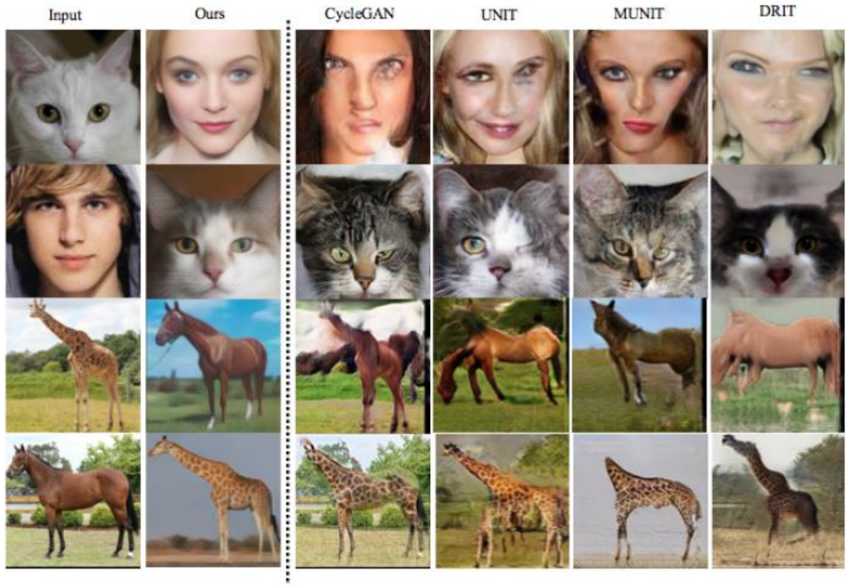
- Fine-disentangled latent space naturally endows our model with the ability of diverse and exemplar-guided generation, which is a challenging and ill-posed multimodal problem in unsupervised image-to-image translation.

Method

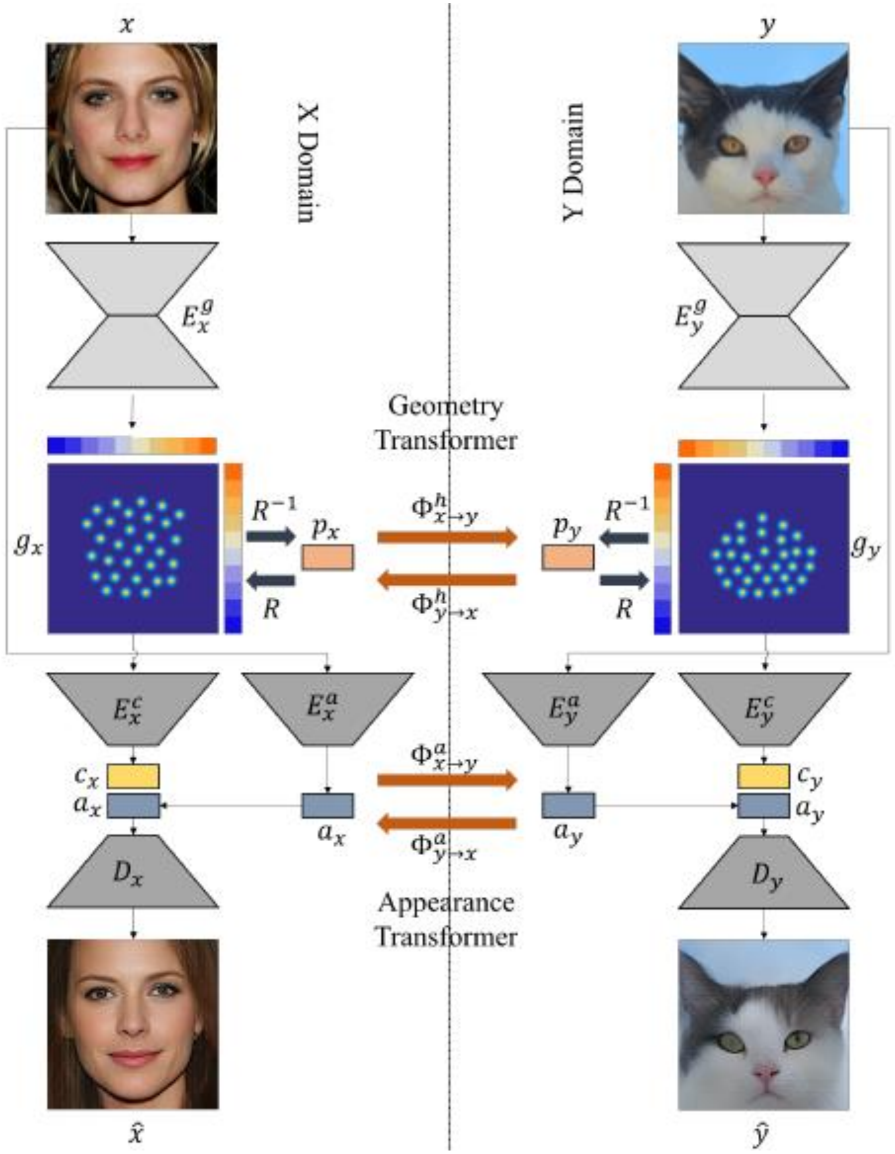


- They assume each domain can be disentangled into a <u>Cartesian product of geometry(structure) space G and appearance space A.</u>

- Geometry 역할 : Spatial distribution이 다르더라도 같은 similar semantic meaning component끼리 mapping + 위치적 정보



Method



- Conditional VAE 사용

$$\mathcal{L}_{\text{disentangle}} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{prior}}$$

$$\mathcal{L}_{\text{CVAE}}(\pi, \theta, \phi, \omega) = -KL(q_\phi(c|x, g)||p(a|x)) + \|x - D(E^c(E^g(x)), E^a(x))\|,$$

Geometric map 의 supervision 없이 학습시키기 위해서 사용

$$\mathcal{L}\text{prior} = \sum_{i \neq j} \exp(-\frac{||g^i - g^j||^2}{2\sigma^2}) + \text{Var}(g)$$

- Transformer을 통한 mapping은 visual relationship 보장 X

Cross-domain appearance consistency loss

$$\mathcal{L}^a_{\text{con}} = \|\zeta(x) - \zeta(D_y(\Phi^g_{x\to y} \cdot E^g_x(x), \Phi^a_{x\to y} \cdot E^a_x(x)))\|, \tag{4}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{prior}} + \mathcal{L}^a_{\text{con}} + \mathcal{L}^a_{\text{cyc}} + \mathcal{L}^g_{\text{cyc}} + \mathcal{L}^{pix}_{\text{cyc}} + \mathcal{L}^a_{\text{adv}} + \mathcal{L}^g_{\text{adv}} + \mathcal{L}^{pix}_{\text{adv}}$$

Experiment

Table 1: **Human perceptual study.** Pairwise A/B tests on horse→giraffe and human→cat face task.

Method	**horse → giraffe** % Testers labeled *better*	**human → cat face** % Testers labeled *better*
CycleGAN [52]	15.0%	15.4%
UNIT [27]	19.3%	18.9%
MUNIT [14]	20.4%	17.8%
DRIT [23]	16.1%	23.4%
Ours	**50.0%**	**50.0%**

(a) Score of "realism".

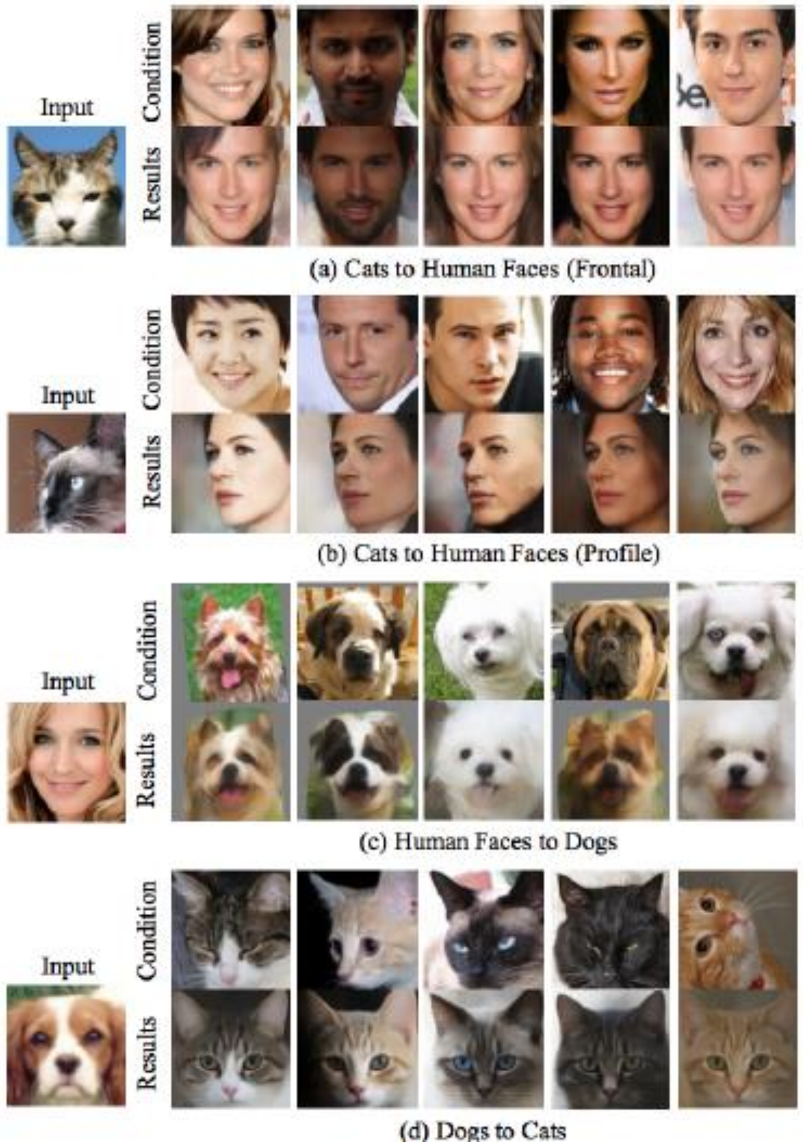
Method	**horse → giraffe** % Testers labeled *better*	**human → cat face** % Testers labeled *better*
CycleGAN [52]	11.9%	25.7%
UNIT [27]	16.5%	23.3%
MUNIT [14]	19.2%	31.7%
DRIT [23]	23.6%	34.4%
Ours	**50.0%**	**50.0%**

(b) Score of "geometry-consistency".

Table 2: **Quantitative Results.** We use FID (lower is better) and diversity (higher is better) with LPIPS distance to evaluate the quality and diversity of the generated images.

	Real Data		**CycleGAN [52]**		**UNIT [27]**		**MUNIT [14]**		**DRIT [23]**		**Ours**	
	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity
cats → human face	0.00	0.54	57.92	-	98.39	-	40.91	**0.41**	69.53	0.20	**32.25**	0.39
human face → cats	0.00	0.65	44.23	-	35.26	-	23.24	0.53	33.14	0.52	**21.88**	**0.56**
cats → dogs	0.00	0.66	143.14	-	104.32	-	100.26	0.59	67.01	0.54	**65.77**	**0.60**
dogs → cats	0.00	0.65	75.75	-	66.84	-	27.60	0.56	31.04	**0.59**	**23.23**	0.58
dogs → human face	0.00	0.54	105.09	-	103.35	-	37.84	0.40	46.70	0.32	**31.06**	**0.41**
human face → dogs	0.00	0.66	149.61	-	91.38	-	73.98	0.60	68.84	0.57	**52.20**	**0.67**
Average	0.00	0.62	95.96	-	83.26	-	50.64	0.52	52.71	0.46	**37.73**	**0.54**

Experiment



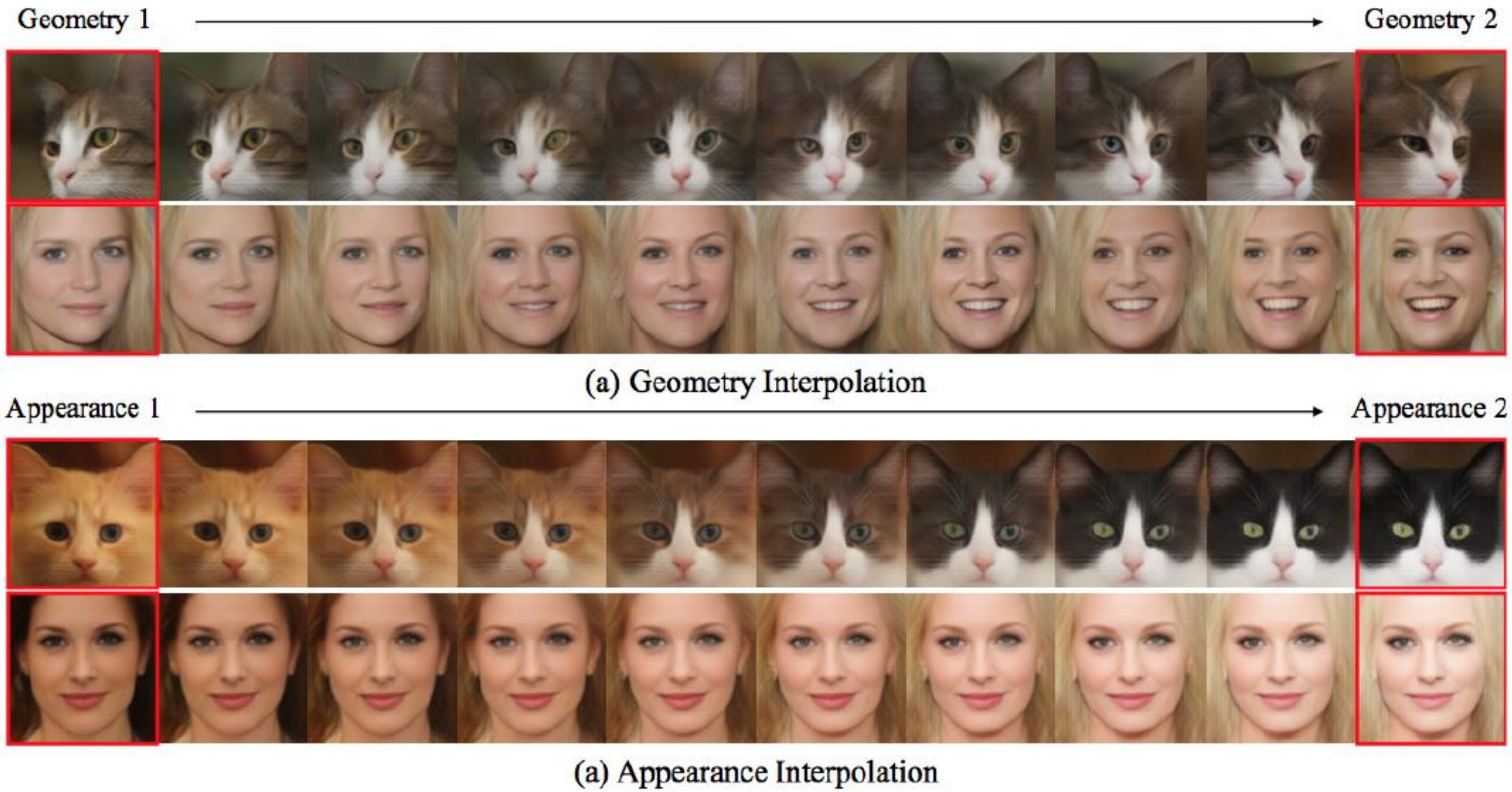


Figure 7: **Interpolation.** Linear interpolation results of geometry and appearance latent code on cat and human face datasets.

Experiment

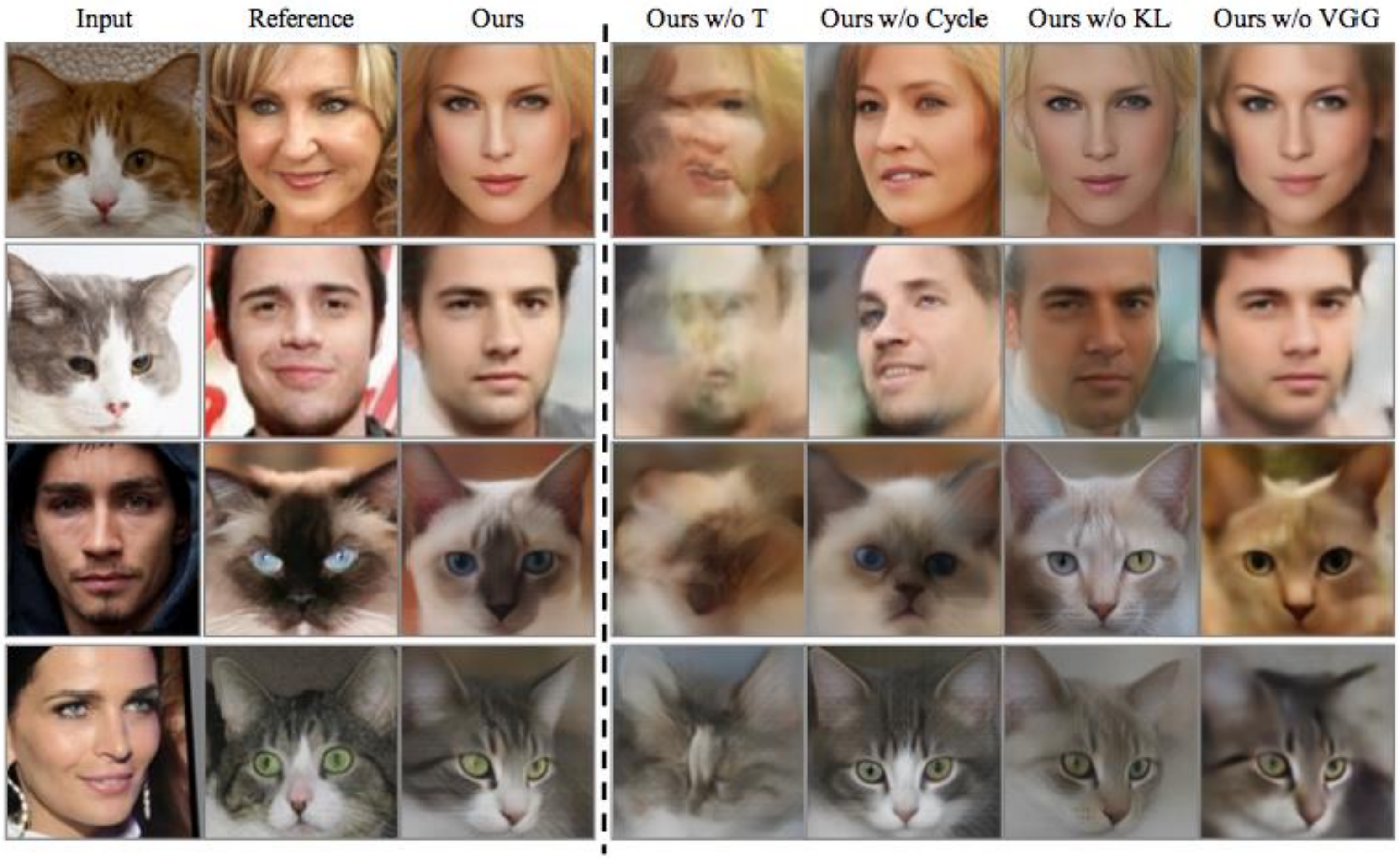


Figure 8: **Quantitative ablation study.** Visualisation results on human↔cat task.