C^3 : Contrastive Learning for Cross-domain Correspondence in Few-shot Image Generation

Hyuk-Gi Lee **Gi-Cheon Kang** **Chang-Hoon Jeong** **Han-Wool Sul** **Byoung-Tak Zhang**
Seoul National University, Artificial Intelligence institute (AIIS)
{hklee,gckang,chjeong,hwsul,btzhang}@bi.snu.ac.kr

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

❏ One of the biggest hurdles in few-shot image generation is that the number of given images from the target domain – typically less than ten images – is too small to approximate the true distribution of the target domain.

❏ Existing methods for few-shot image generation have shown limited domain adaptation capabilities in that they have implicitly attempted to transfer the knowledge from the source to the target domain without any direct mapping from the source to target images.

- we propose a simple yet effective approach C³, Contrastive Learning for Cross-domain Correspondence, that learns the cross-domain correspondence in an explicit way.
- we validate the effectiveness of our proposed method on photorealistic and non-photorealistic domains by comparing C³ with the state-of-the-art approaches.

Method

❏ Overview

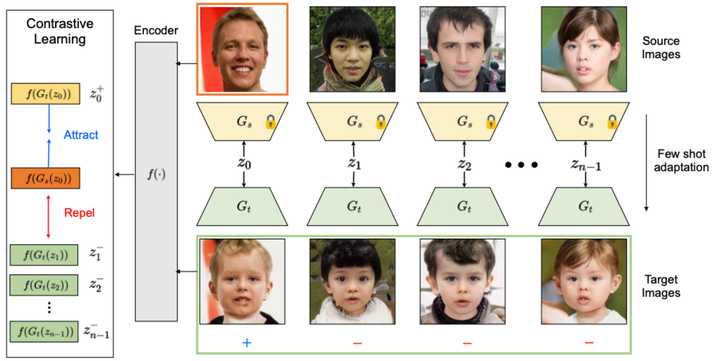

Figure: An overview of our approach, C . The source image is an anchor (orange box). Positive pair consists of an anchor and target image (+ on green box) from the same latent vector mapped to anchor. Negative pairs consist of anchor and others (- on green box) from different latent vectors within mini-batch. C³ makes semantic similarity of corresponding pair (positive pair) high during adaptation, otherwise vice versa.

❏ Contrastive loss for cross-domain correspondence

$$\mathcal{L}_{con}\left(G_s(z_i), G_t(z_i)\right) = -\log \frac{\exp\left(\cos\left(f\left(G_s(z_i)\right), f\left(G_t(z_i)\right)\right)/\tau\right)}{\sum_{j=1}^{M} \exp\left(\cos\left(f\left(G_s(z_i)\right), f\left(G_t(z_j)\right)\right)/\tau\right)}$$

Contrastive loss takes output of the encoder for source and target image from same latent vector as corresponding pair (positive pair) otherwise as negative pairs within mini-batch size M.

❏ Objective function

$$G_{s\to t} = \arg\min_{G} \max_{D_{patch}} \mathcal{L}_{adv}(G, D_{patch}) + \lambda_{con}\mathcal{L}_{con}(G, G_s)$$

L_{adv} is an adversarial loss for training GANs, and L_{con} is a contrastive loss for keeping source-target semantic similarity while adapting to target domain.

Experiments - Qualitative

❏ Comparison with other methods

Figure: Adaptation results for different main methods to target domains given on 10-shot target data:
FFHQ → FFHQ-Babies (top), FFHQ → FFHQ-Sunglasses (middle), FFHQ → Face-Sketches (bottom)

❏ Other adaptation results

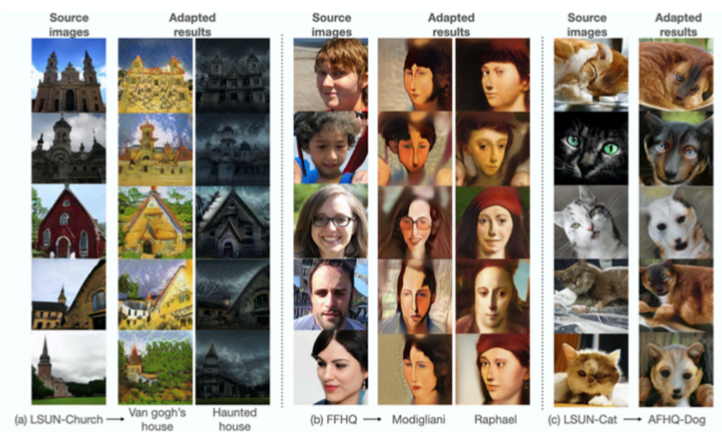

Figure: Other domain adaptation results by our method. target data is also 10-shot. In all above domains, generated images resemble structure of source images such as face pose, building structure depending on source domain.

Experiments - Quantitative

❏ FID scores

	Babies	Sunglasses	Sketches
TGAN [12]	104.79	55.61	53.41
TGAN+ADA [37]	102.58	53.64	66.99
BSA [13]	140.34	76.12	69.32
FreezeD [14]	110.92	51.29	46.54
MineGAN [18]	98.23	68.91	64.34
EWC [15]	87.41	59.73	71.25
CDC [16]	74.39	42.13	45.67
Ours	$\mathbf{67.55 \pm 2.23}$	$\mathbf{36.69 \pm 2.63}$	$\mathbf{41.50 \pm 1.64}$

FID scores (↓) for target domains with entire target data. Standard deviations are computed across 5 random runs.

❏ LPIPS distances

	Babies	Sunglasses	Sketches
MineGAN [18]	0.52 ± 0.03	0.43 ± 0.04	0.40 ± 0.05
EWC [15]	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.58 \pm 0.01}$	0.42 ± 0.03
CDC [16]	0.57 ± 0.02	0.57 ± 0.02	$\mathbf{0.45 \pm 0.02}$
Ours	$\mathbf{0.58 \pm 0.02}$	0.56 ± 0.01	$\mathbf{0.45 \pm 0.03}$

LPIPS scores(↑) for adapted results. Standard deviations is computed across the target samples (In this case 10)

Ablation studies

❏ Effect of λ_{con} value

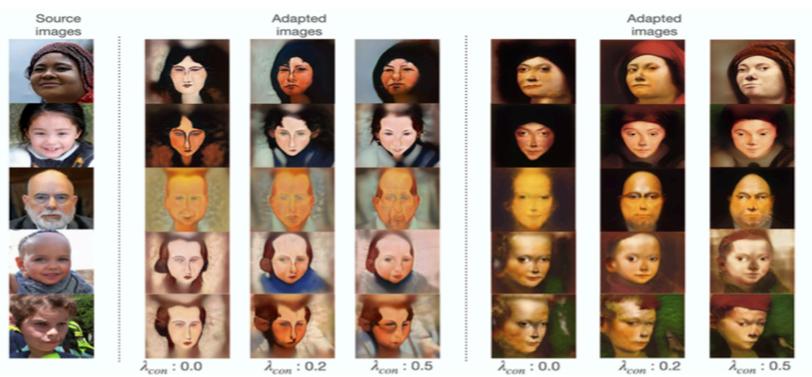

Figure: Effect of λ_{con} on adaptation results. The larger λ_{con}, visual features of source images remain strong on adapted images. Conversely, the smaller λ_{con}, the weaker the adaptation results reflect the correspondence with the source images.

❏ N-shot data setting

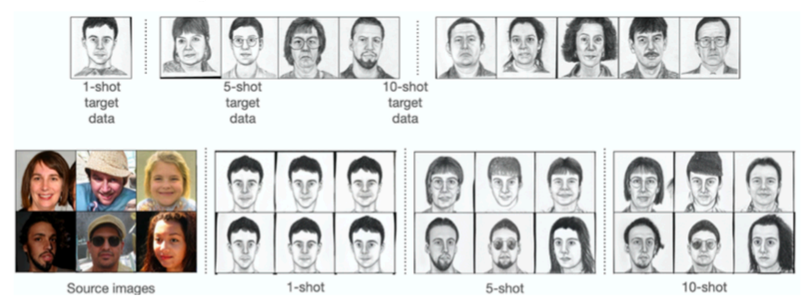

Figure: Adaptation results on different target data size. The larger target data size, our method can generate more diverse and detailed images. Even if given on 1-shot target data, generated images reflect weak correspondence like pose and expressions with source images.