Paper Note

Semi-supervised Learning for Few-shot Image-to-Image Translation

Project Link: Github

November 16, 2020

1 Background

Previous Work

- one-shot I2I translation by first training a variational autoencoder for the seen domain and then adapting those layers related to the unseen domain.
- zero-shot I2I translation, employ- ing the annotated attributes of unseen categories instead of the labeled images.
- few-shot I2I translation in a multi-class setting. These models, however, need to be trained using large amounts of hand-annotated ground-truth labels for images of the source domain

Limitations

Labeling large-scale datasets is costly and time consuming, making those methods less applicable in practice. In this paper, they overcome this limitation and explore a novel setting, few-shot I2I translation in which only limited labeled data is available from the source classes during training.

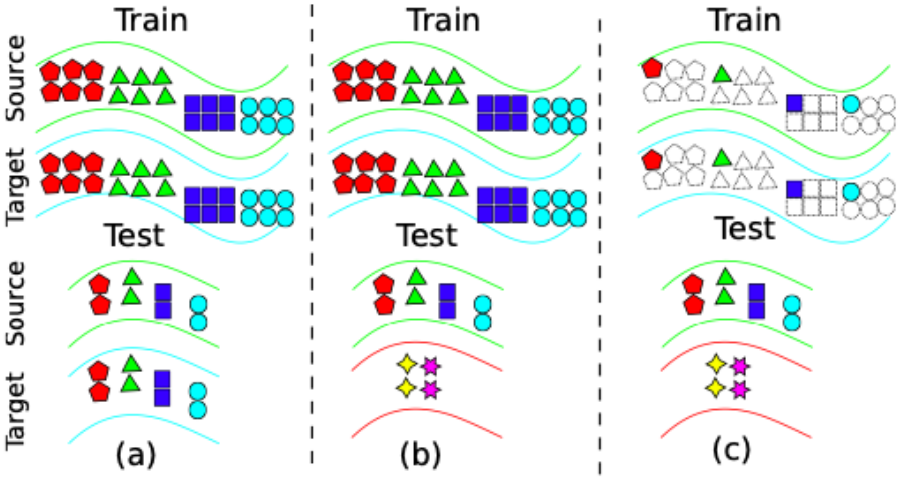

Figure 1. Comparison between unpaired I2I translation scenarios. Each colored symbol indicates a different image label, and dashed symbols represent unlabeled data. (a) *Standard* [9, 18, 46]: target classes are the same as source classes and all are seen during training. (b) *Few-shot* [28]: actual target classes are different from source classes and are unseen during training. Only a few examples of the unseen target classes are available at test time. For training, source classes act temporarily as target classes. (c) *Few-shot semi-supervised* (Ours): same as few-shot, but the source domain has only a limited amount of labeled data at train time.

Contribution: SEMIT

Model Overview

The model architecture is shown as follows:

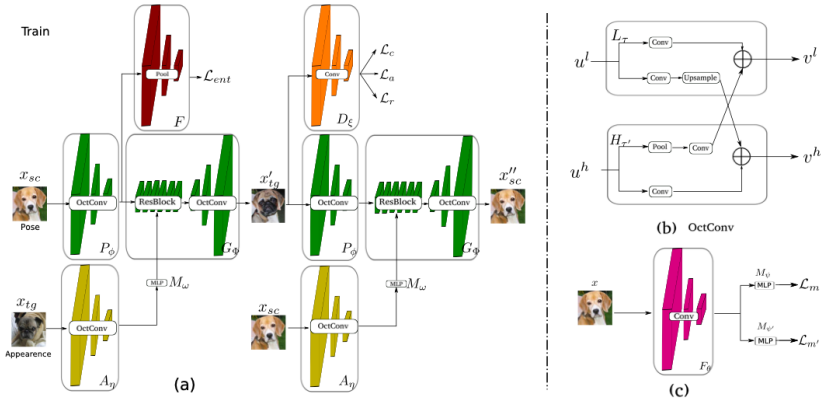

Figure 2. Model architecture for training. (a) The proposed approach is composed of two main parts: Discriminator D_ξ and the set of Pose encoder P_ϕ, Appearance encoder A_η, Generator G_Φ, Multilayer perceptron M_ω and feature regulator F. (b) The OctConv operation contains high-frequency block ($H_{\tau'}$) and low-frequency block (L_τ). (c) Noise-tolerant Pseudo-labeling architecture.

The total loss function is:

$$\min_{P_\phi, A_\eta, M_\omega, G_\Phi} \max_{D_\xi} \lambda_a \mathcal{L}_a + \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_e \mathcal{L}_{ent}$$

Here,

P_ϕ – *Pose encoder*
A_η – *Appearance encoder*
G_Φ – *Generator*
M_ω – *Multilayer perceptron*
$\mathcal{L}_a$ – *Adversarial loss*
$\mathcal{L}_c$ – $Classfication\ loss(aux-GAN,\ to\ generate\ target-specific\ images)$
$\mathcal{L}_r$ – *Reconstruction loss*
$\mathcal{L}_{ent}$ – *Entropy regulationloss*

(1)

Experiment

Datasets

- Animals
- Birds
- Flowers

- Foods

Randomly, sample 25,000 source images from the training set and translate them to each target domain (not seen during training)
1, 5, 20-shot settings for the target set.

Evaluation

IS (*Inception Score*)
FID (*Fréchet Inception Distance*)
Translation Accuracy (evaluate whether a model is able to generate images of the target class)

Representative Results

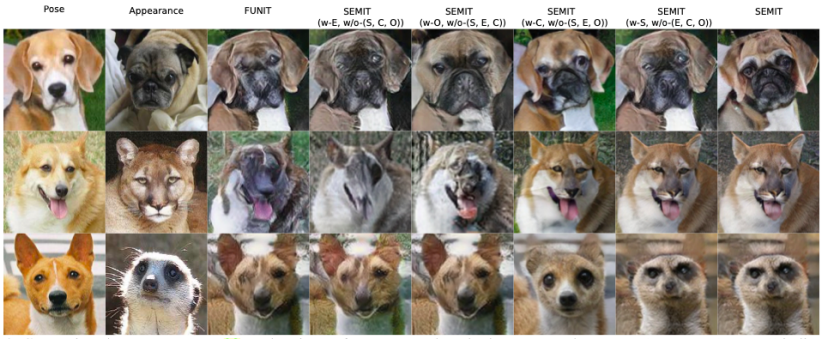

Figure 3. Comparison between FUNIT [28] and variants of our proposed method. For example, *SEMIT (w-E, w/o-(S, C, O))* indicates the model trained with only entropy regulation. More examples are in Suppl. Mat. (Sec. 1).

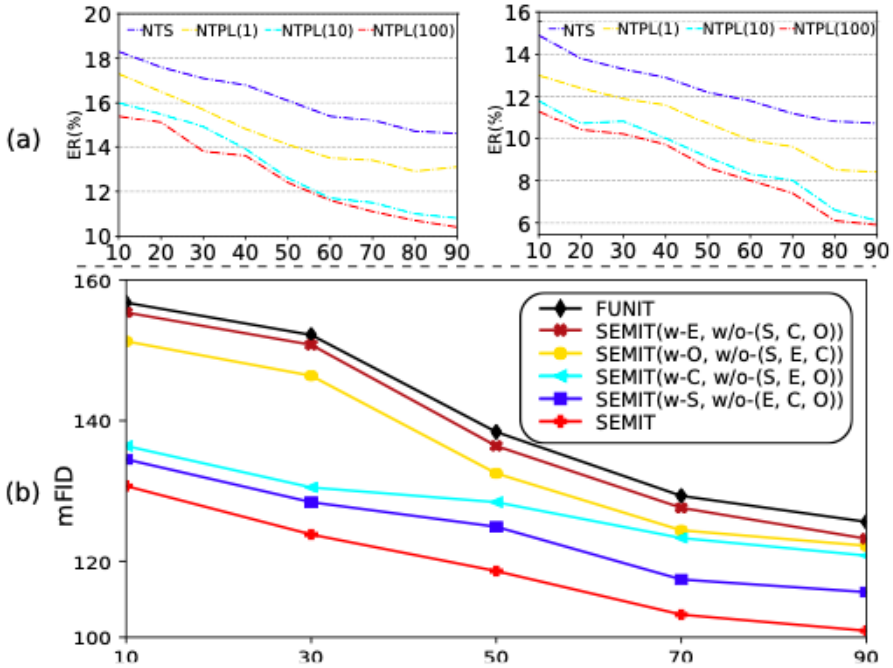

Figure 4. (a) Ablation study on classification for (left) Animals-69 and (right) Birds, measured by Error Rate (ER). (b) Ablation study of the variants of our method for one-shot on Animals-69. The x-axis shows the percentage of the labeled data used for training.