# Truly Unsupervised Image-to-Image Translation with Contrastive Representation Learning (Supplementary Materials)

Zhiwei Hong[1], Jianxing Feng[2], and Tao Jiang[1,3] ✉

[1] Tsinghua University, Beijing 100084, China
hzw17@mails.tsinghua.edu.cn
[2] Haohua Technology Co., Ltd, Shanghai, China
[3] University of California, Riverside, CA 92521, USA
jiang@cs.ucr.edu

## 1 Implementation details

Our model consists of three modules, the embedding network $E$, the generator $G$ and the discriminator $D$. Firstly, we train the embedding network $E$ individually for 60K iterations. Then, we train the entire model (including all three neural networks $E$, $G$ and $D$) 100K more iterations. During the training phase, the batch size is set to 32 and the input image resolution is set to $128 \times 128$. The Adam [4] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ is used for optimizing the embedding network $E$ and the RMSProp [3] optimizer with $\alpha = 0.99$ is used for optimizing the generator $G$ and discriminator $D$. All learning rates are set to 0.0001 with a weight decay $1 \times 10^{-4}$. The hinge version adversarial loss [10, 5] with $R_1$ regularization [6] using $\gamma = 10$ (Eq. 3) is adopted the same as in [1]. The hyper-parameters in the loss function of the generator $G$ are set as $\lambda^G_{style} = 0.01$ and $\lambda_{rec} = 0.1$. The hyper-parameters in the loss function of the embedding network $E$ are set accordingly in different training stages. More precisely, they are set as $\lambda_{MI} = 5.0$ and $\lambda_{co} = 1.0$ in the first 60K iterations when $E$ is trained individually and are decreased as $\lambda_{MI} = 0.5$ and $\lambda_{co} = 0.1$ in the next 100K iterations when $E$ is simultaneously trained with the generator $G$. We use grid-search to tune the hyper-parameters in our experiments and have observed that the performance of our model is pretty robust with respect to some of the hyper-parameters. We initialize the weights of the convolutional layers with the He initialization [2] and the weights of the linear layers using random numbers from $N(0, 0.01)$ with zero biases. The entire model is implemented in PyTorch [8]. It takes about 40 hours to train our model on a single NVIDIA GeForce RTX 3090 GPU.

# 2 More discussions on the hyper-parameter $\hat{K}$ and real number of domains $K$

## 2.1 Discussion on the search for an optimal/reasonable $\hat{K}$

As described in the *Sensitivity to* $\hat{K}$ part of Sec 4.3, our model is relatively robust to the hyper-parameter $\hat{K}$ when the real number of domains $K$ is not too large. The selection of an optimal $\hat{K}$ is generally hard especially in real scenarios when we are not given much meta information about the dataset. While it might be tempting to automate the choice of $\hat{K}$ into the learning process, we do not have a good solution at the moment and intend the study in the future. However, we might consider the following alternative in practice. We may choose an initial $\hat{K}$ by running an existing unsupervised image clustering method (*e.g.*, [7]) or from a small range (*e.g.*, $7 \sim 30$) based on empirical experience. Note that our current model works well only for a small or moderate number of domains anyway. We then try to improve $\hat{K}$ by checking if some of the clusters obtained by our model could be very close to each other and thus merged. This is clearly a heuristic and may not always lead to a good choice of $\hat{K}$. We plan to explore more rigorous approaches to optimize $\hat{K}$ in our future work.

## 2.2 Discussion on how the real number of domains $K$ affects the performance of different models

As stated the subsection 4.3, it is difficult for truly unsupervised methods (TUNIT and our model CUNIT) to perform well compared to fully supervised methods (such as COCO-FUNIT) when the real domain number $K$ is very large. To better understand how $K$ affects the performance of CUNIT, we have also done an experiment to show how our model performs in comparison to COCO-FUNIT on the AnimalFaces dataset with moderate domain numbers $K = 20, 30, 50, 70, 100$ (in addition to $K = 10, 149$), as shown in Table S1. It can be observed from the table that the gap between CUNIT and COCO-FUNIT increases as $K$ gets bigger and becomes quite significant when $K \geq 50$ (more than 6%).

# 3 Additional visual results

## 3.1 Truly unsupervised image-to-image translation

Some additional truly unsupervised image-to-image translation visual results of CUNIT are shown below. Here, $x_A$ denotes the input content (or source) image and $x_B$ denotes the input style (or reference) image.

**Table S1.** Quantitative evaluation (based on mFID) of different methods on the AnimalFaces dataset with different numbers of domains.

| number of domains $K$ | Truly Unsupervised | | Fully Supervised |
|---|---|---|---|
| | TUNIT [1] | CUNIT (ours) | COCO-FUNIT [9] |
| $K = 10$ | 47.9 | 45.2 | 44.8 |
| $K = 20$ | 51.7 | 49.3 | 48.3 |
| $K = 30$ | 56.9 | 54.7 | 53.1 |
| $K = 50$ | 64.3 | 62.5 | 58.9 |
| $K = 70$ | 74.1 | 72.6 | 66.6 |
| $K = 100$ | 84.2 | 83.4 | 75.1 |
| $K = 149$ | 106.3 | 106.9 | 92.4 |

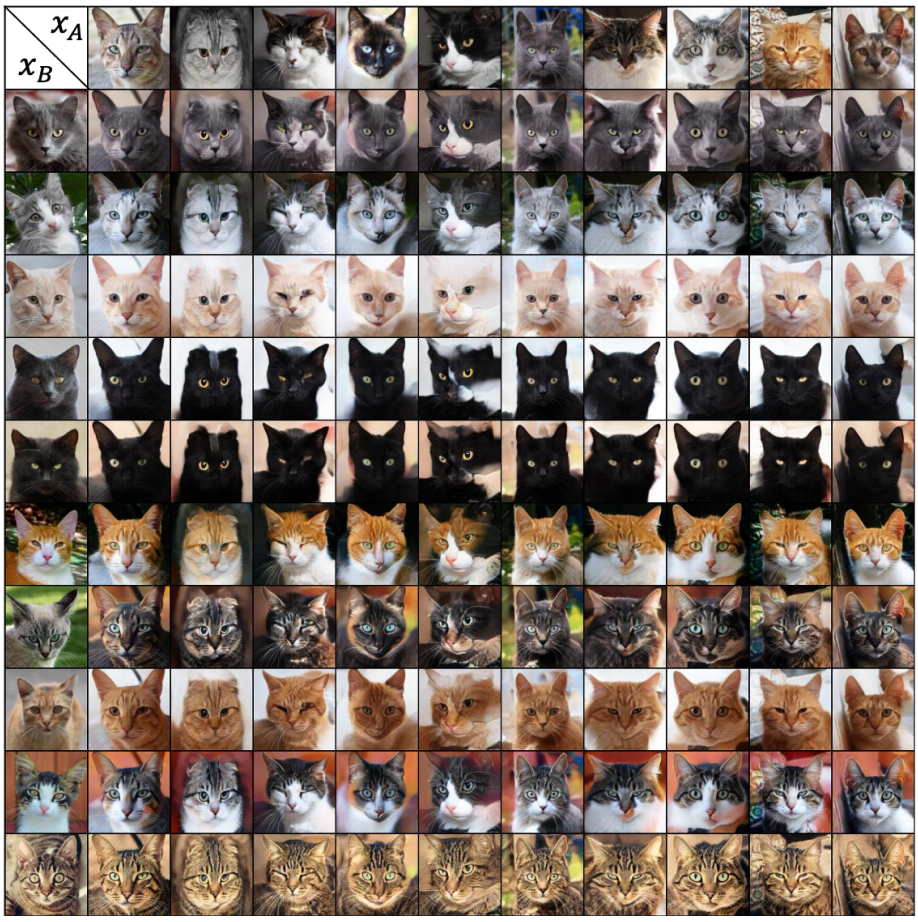


**Fig. S1.** Additional unsupervised image-to-image translation results on an AFHQ dataset (AFHQ-cats).

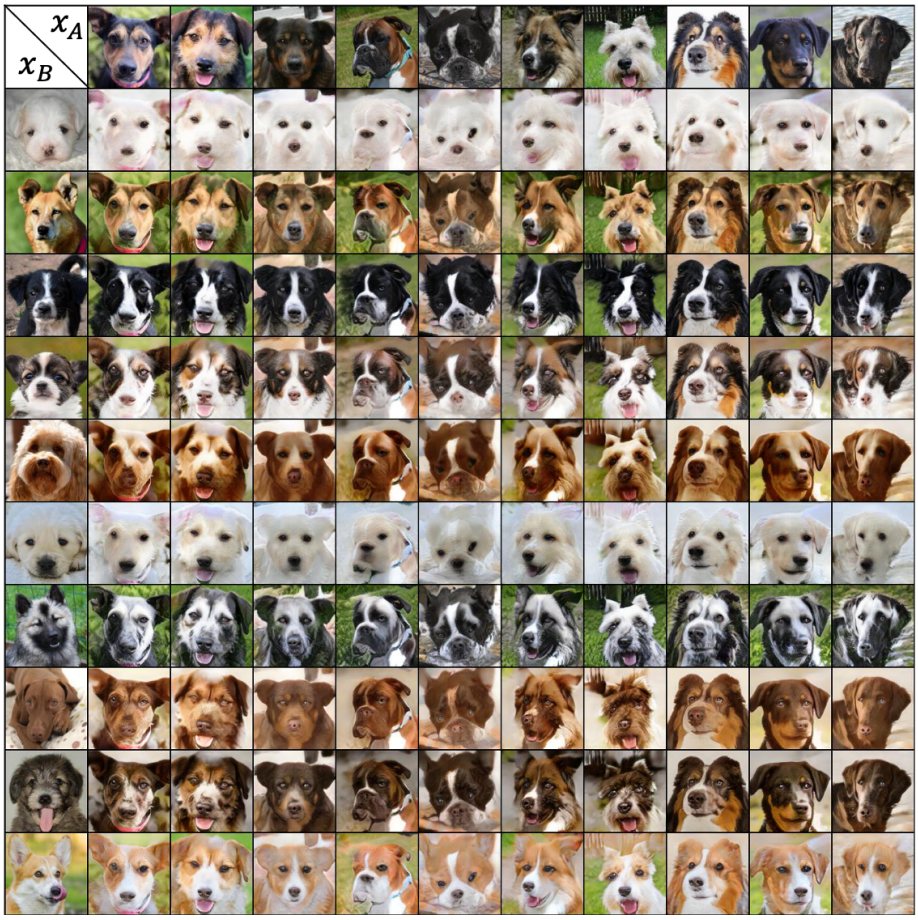


**Fig. S2.** Additional unsupervised image-to-image translation results on an AFHQ dataset (AFHQ-dogs).


**Fig. S3.** Additional unsupervised image-to-image translation results on an AFHQ dataset (AFHQ-wild animals).

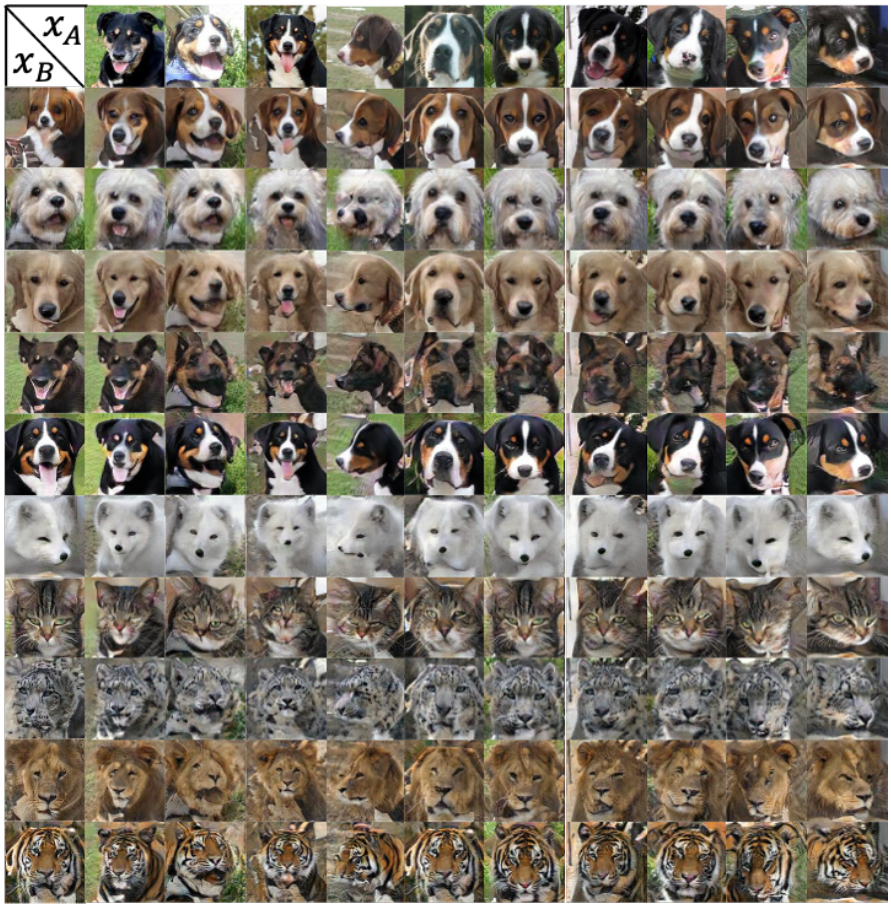


**Fig. S4.** Additional unsupervised image-to-image translation results on the AnimalFaces-10 dataset.

### 3.2 Domain-level supervised or semi-supervised image translation

Some additional visual results of various methods on cross-domain and multi-domain semi-supervised image-to-image translation are shown below.

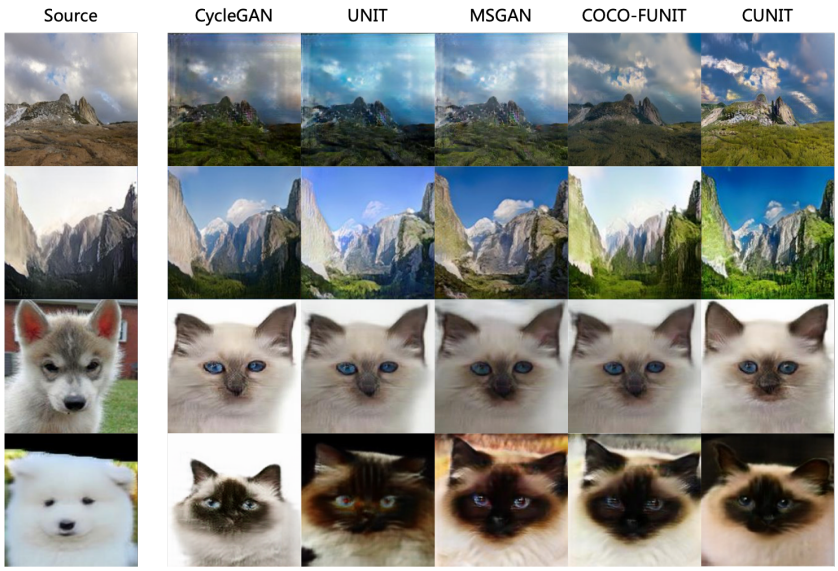


**Fig. S5.** Some cross-domain image-to-image translation visual results on the Summer2Winter and Dog2Cat datasets (the top two rows : winter → summer and the bottom two rows : dogs → cats). Clearly, the translated images generated by CUNIT have the best clarity. Moreover, it was able to acquire the target style (*i.e.*, summer) better in the top two rows and maintain the source content (*e.g.*, ear shapes) better in the bottom two rows.

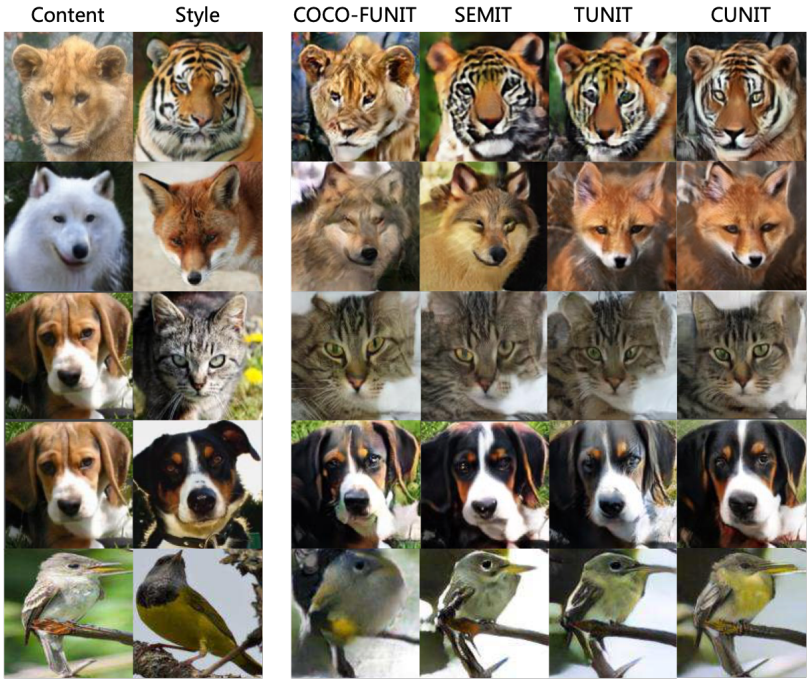


**Fig. S6.** Some visual results of CUNIT and the other compared methods under the semi-supervised setting on the AnimalFaces-10 and Birds-10 datasets (20% of the labeled samples are used). Again, CUNIT was able to generate translated images with the best clarity, and its results generally resemble the target images the most in terms of style and resemble the source images the most in terms of content.

## References


1. Baek, K., Choi, Y., Uh, Y., Yoo, J., Shim, H.: Rethinking the truly unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14154–14163 (2021)
2. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
3. Hinton, G., Srivastava, N., Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on **14**(8), 2 (2012)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017)
6. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
7. Park, S., Han, S., Kim, S., Kim, D., Park, S., Hong, S., Cha, M.: Improving unsupervised image clustering with robust learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12278–12287 (2021)
8. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
9. Saito, K., Saenko, K., Liu, M.Y.: Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 382–398. Springer (2020)
10. Tran, D., Ranganath, R., Blei, D.M.: Hierarchical implicit models and likelihood-free variational inference. arXiv preprint arXiv:1702.08896 (2017)