# A Cyclically-Trained Adversarial Network for Invariant Representation Learning (Supplementary Material)

Jiawei Chen
Boston University
garychen@bu.edu

Janusz Konrad
Boston University
jkonrad@bu.edu

Prakash Ishwar
Boston University
pi@bu.edu

In this supplementary material, section 1 provides an ablation study which evaluates the impacts of various training objectives in our framework. Section 2 gives examples of synthesized images from our trained model without using an input image.

## 1. Ablation study

In this section, we perform an ablation study for various training objectives for the proposed model on three datasets (see corresponding results in Tables 1 and 2). Specifically, we examine the following cases:

1. training the model with forward cycle only (using Eq.(1) and Eq.(2) in the paper),
2. training the model with forward cycle and cycle-consistency constraint on the latent space (using Eq.(1), Eq.(2) and the first term in Eq.(3) in the paper),
3. training the model with forward cycle and cycle-consistency constraint on the image space (using Eq.(1), Eq.(2) and the second term in Eq.(3) in the paper),
4. training the complete model with both forward and backward cycles.

We first observe that the cycle-consistency constraint on the latent space alone significantly improves invariance, e.g., identity classification CCR decreases from 33.40% (forward cycle only) to 15.61% for UPNA synthetic dataset. However, it also slightly diminishes information about unspecified factors of variation, which is reflected in both latent and image spaces, e.g., head pose estimation errors computed from the latent representations and GLS values of the pose angles are $0.47^{\circ}$ and 0.63 larger than those of case 1 (in average), respectively. This is likely because the specified and certain unspecified factors of variation are not strictly independent, e.g., some subjects have a wider range of pose angles in UPNA dataset. Thus, when the model is trained to disentangle the specified factor of variation from the latent space, the correlated unspecified factors are inevitably affected.

When combining a forward cycle with cycle-consistency constraint on the image space, we observe that the learned representations and synthesized images retain more information about the original image compared to other cases, e.g., the model in this case yields the best classification CCRs and GLS scores for the viewing orientations for 3D chairs. However, it also yields the highest classification CCR for chair style, a specified factor, which indicates that although the model is provided with an input class code **c** to control the specified factor of variation in the synthesized image, the image reconstruction objective still leads the model to preserve a certain amount of the original information about the specified factor in the latent representation.

In the case of training the model with a forward cycle and complete backward cycle, we witness that the model achieves a good balance between eliminating information about the specified factor and preserving information about the unspecified factors of variation in the latent space. More importantly, the model in this case consistently outperforms the model merely trained with a forward cycle in terms of the ability to suppress classification performance for the specified factor, while ensuring that the estimation performance for the unspecified factors does not fall short. Similarly, the synthesized images from the complete model consistently have a better GLS scores than the images from the model trained with forward cycle only. Such observations justify the effectiveness and indispensability of both cycle-consistency constraints.

## 2. Image synthesis without input image

Once trained, our model can also synthesize novel images without using an input image due to the constraint imposed over the latent space. To generate a new image, we

Table 1: Comparison of the quality of invariance of representations generated by the proposed model trained with various objectives. ↑ means higher is better. ↓ means lower is better.

| Dataset | Factors of variation | Methods | | | | |
|---|---|---|---|---|---|---|
| | | Random guess/ Median | Forward cycle | Forward cycle + cycle-consistency on $\mathbf{z}$ | Forward cycle + cycle-consistency on $\mathbf{x}$ | Forward cycle + cycle-consistency on $\mathbf{x}$ & $\mathbf{z}$ |
| 3D Chairs | Chair Style ↓ | 0.07% | 3.21% | 0.30% | 3.51% | 0.79% |
| | $\theta$ ↑ | 50% | 74.37% | 72.00% | 84.40% | 78.17% |
| | $\phi$ ↑ | 3.22% | 69.54% | 70.96% | 72.60% | 71.90% |
| YaleFace | Identity ↓ | 2.63% | 12.36% | 4.47% | 17.89% | 6.97% |
| | Illumination Cond. ↑ | 1.56% | 85.40% | 82.37% | 85.73% | 85.50% |
| UPNA Synthetic | Identity ↓ | 10% | 33.40% | 15.61% | 56.48% | 18.05% |
| | Yaw ↓ | 5.10°±6.70° | 2.10°±2.08° | 2.48°±2.60° | 1.65°±1.57° | 2.12°±2.12° |
| | Pitch ↓ | 4.98°±5.02° | 2.20°±2.06° | 2.82°±3.12° | 1.73°±1.62° | 2.23°±2.10° |
| | Roll ↓ | 4.68°±6.88° | 1.29°±1.43° | 1.69°±1.99° | 0.86°±0.85° | 1.16°±1.24° |

Table 2: Comparison of *GLS* values for the proposed model trained with various objectives. ↑ means higher is better. ↓ means lower is better.

| Datasets | Factors of variation | Methods | | | |
|---|---|---|---|---|---|
| | | Forward cycle | Forward cycle + cycle-consistency on $\mathbf{z}$ | Forward cycle + cycle-consistency on $\mathbf{x}$ | Forward cycle + cycle-consistency on $\mathbf{x}$ & $\mathbf{z}$ |
| 3D Chairs | Chair Style ↑ | 0.77 | 0.76 | 0.92 | 0.87 |
| | $\theta$ ↑ | 0.66 | 0.64 | 0.78 | 0.66 |
| | $\phi$ ↑ | 0.52 | 0.50 | 0.59 | 0.57 |
| YaleFace | Identity ↑ | 0.97 | 0.98 | 0.98 | 0.98 |
| | Illumination Cond. ↑ | 0.68 | 0.59 | 0.72 | 0.70 |
| UPNA Synthetic | Identity ↑ | 0.99 | 0.99 | 1.00 | 1.00 |
| | Yaw ↓ | 2.62 | 3.13 | 2.40 | 2.55 |
| | Pitch ↓ | 2.95 | 3.58 | 2.30 | 2.46 |
| | Roll ↓ | 1.39 | 2.14 | 1.28 | 1.37 |

first sample a latent vector from a prior distribution (in our experiments: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$). Then we concatenate it with a class code and feed them into a trained decoder to synthesize a new image. As shown in Fig. 1, the synthesized images are realistic. Although our model cannot control the unspecified factors without an input image, the synthesized images from randomly sampled latent vectors are still useful for some applications, e.g., dataset augmentation.

Figure 1: Image synthesis without input image; $\mathbf{z}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.