Multimodal Unsupervised Image-to-Image Translation - Supplementary Material

Xun Huang[1], Ming-Yu Liu[2], Serge Belongie[1], Jan Kautz[2]

Cornell University[1] NVIDIA[2]

1 Proofs

Proposition 1. *Suppose there exists* E_1^*, E_2^*, G_1^*, G_2^* *such that: 1)* $E_1^* = (G_1^*)^{-1}$ *and* $E_2^* = (G_2^*)^{-1}$, *and 2)* $p(x_{1\to 2}) = p(x_2)$ *and* $p(x_{2\to 1}) = p(x_1)$. *Then* E_1^*, E_2^*, G_1^*, G_2^* *minimizes* $\mathcal{L}(E_1, E_2, G_1, G_2) = \max_{D_1, D_2} \mathcal{L}(E_1, E_2, G_1, G_2, D_1, D_2)$ *(Eq. (5)).*

Proof.

$$\mathcal{L}(E_1, E_2, G_1, G_2) = \max_{D_1, D_2} \mathcal{L}(E_1, E_2, G_1, G_2, D_1, D_2) = \max_{D_1} \mathcal{L}_{\text{GAN}}^{x_1} + \max_{D_2} \mathcal{L}_{\text{GAN}}^{x_2} + \lambda_x(\mathcal{L}_{\text{recon}}^{x_1} + \mathcal{L}_{\text{recon}}^{x_2}) + \lambda_c(\mathcal{L}_{\text{recon}}^{c_1} + \mathcal{L}_{\text{recon}}^{c_2}) + \lambda_s(\mathcal{L}_{\text{recon}}^{s_1} + \mathcal{L}_{\text{recon}}^{s_2})$$

As shown in Goodfellow *et al.* [1], $\max_{D_2} \mathcal{L}_{\text{GAN}}^{x_2} = 2 \cdot \text{JSD}(p(x_2)|p(x_{1\to 2})) - \log 4$ which has a global minimum when $p(x_2) = p(x_{1\to 2})$. Also, the bidirectional reconstruction loss terms are minimized when E_i and G_i are inverses. Thus the total loss is minimized under the two stated conditions.

In the following, we will assume the networks have sufficient capacity and the optimality is reachable as in prior works [1, 2]. That is $E_1 \to E_1^*$, $E_2 \to E_2^*$, $G_1 \to G_1^*$, and $G_2 \to G_2^*$.

Proposition 2. *When optimality is reached, we have:*

$$p(c_1) = p(c_2),\ p(s_1) = q(s_1),\ p(s_2) = q(s_2)$$

Proof. Let z_1 denote the latent code, which is the concatenation of c_1 and s_1. We denote the encoded latent distribution by $p_E(z_1)$, which is defined by $z_1 = E_1(x_1)$ and x_1 sampled from the data distribution $p(x_1)$. We denote the latent distribution at generation time by $p(z_1)$, which is obtained by $s_1 \sim q(s_1)$ and $c_1 \sim p(c_2)$. The generated image distribution $p_G(x_1) = p(x_{2\to 1})$ is defined by $x_1 = G_1(z_1)$ and z_1 sampled from $p(z_1)$. According to the change of variable formula for probability density functions:

$$p_G(x_1) = |\frac{\partial G_1^{-1}(x_1)}{\partial x_1}| p(G_1^{-1}(x_1))$$

$$p_E(z_1) = |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| p(E_1^{-1}(z_1))$$

According to Proposition 1, we have $p_G(x_1) = p(x_1)$ and $E_1 = G_1^{-1}$ when optimality is reached. Thus:

$$\begin{aligned}
p_E(z_1) &= |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| p(E_1^{-1}(z_1)) \\
&= |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| p_G(E_1^{-1}(z_1)) \\
&= |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| |\frac{\partial G_1^{-1}(E_1^{-1}(z_1))}{\partial E_1^{-1}(z_1)}| p(G_1^{-1}(E_1^{-1}(z_1))) \\
&= |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| |\frac{\partial G_1^{-1}(G_1(z_1))}{\partial E_1^{-1}(z_1)}| p(G_1^{-1}(G_1(z_1))) \\
&= |\frac{\partial E_1^{-1}(z_1)}{\partial z_1}| |\frac{\partial z_1}{\partial E_1^{-1}(z_1)}| p(G_1^{-1}(G_1(z_1))) \\
&= p(z_1)
\end{aligned}$$

Similarly we have $p_E(z_2) = p(z_2)$, which together prove the original proposition. From another perspective, we note that $\mathcal{L}_{\text{recon}}^{c_2}, \mathcal{L}_{\text{recon}}^{s_1}, \mathcal{L}_{\text{GAN}}^{x_1}$ coincide with the objective of a WAE [3] or AAE [4] in the latent space, which pushes the encoded latent distribution towards the latent distribution at generation time.

Proposition 3. *When optimality is reached, we have* $p(x_1, x_{1\to 2}) = p(x_{2\to 1}, x_2)$.

Proof. For the ease of notation we denote the joint distribution $p(x_1, x_{1\to 2})$ by $p_{1\to 2}(x_1, x_2)$ and $p(x_{2\to 1}, x_2)$ by $p_{2\to 1}(x_1, x_2)$. Both densities are zero when $E_1^c(x_1) \neq E_2^c(x_2)$. When $E_1^c(x_1) = E_2^c(x_2)$, we have:

$$\begin{aligned}
p_{1\to 2}(x_1, x_2) &= p_G(x_2|E_1^c(x_1))p(x_1) \\
&= |\frac{\partial E_2^s(x_2)}{x_2}| q(E_2^s(x_2))p(x_1) \\
&= p(x_2|E_1^c(x_1))p_G(x_1) \\
&= p_{2\to 1}(x_1, x_2)
\end{aligned}$$

Proposition 4. *Denote* $h_1 = (x_1, s_2) \in \mathcal{H}_1$ *and* $h_2 = (x_2, s_1) \in \mathcal{H}_2$. h_1, h_2 *are points in the joint spaces of image and style. Our model defines a deterministic mapping* $F_{1\to 2}$ *from* $\mathcal{H}_1$ *to* $\mathcal{H}_2$ *(and vice versa) by* $F_{1\to 2}(h_1) = F_{1\to 2}(x_1, s_2) \triangleq (G_2(E_1^c(x_1), s_2), E_1^s(x_1))$. *When optimality is achieved, we have* $F_{1\to 2} = F_{2\to 1}^{-1}$.

Proof.

$$\begin{aligned}
F_{2\to 1}(F_{1\to 2}(x_1, s_2)) &\triangleq F_{2\to 1}(G_2(E_1^c(x_1), s_2), E_1^s(x_1)) && (1)\\
&\triangleq (G_1(E_2^c(G_2(E_1^c(x_1), s_2)), E_1^s(x_1)), E_2^s(G_2(E_1^c(x_1), s_2))) && (2)\\
&= (G_1(E_2^c(G_2(E_1^c(x_1), s_2)), E_1^s(x_1)), s_2) && (3)\\
&= (G_1(E_1^c(x_1), E_1^s(x_1)), s_2) && (4)\\
&= (x_1, s_2) && (5)
\end{aligned}$$

And we can prove $F_{1\to2}(F_{2\to1}(x_2, s_1)) = (x_2, s_1)$ in a similar manner. To be more specific, (3) is implied by the style reconstruction loss $\mathcal{L}^s_{\text{recon}}$, (4) is implied by the content reconstruction loss $\mathcal{L}^c_{\text{recon}}$, and (5) is implied by the image reconstruction loss $\mathcal{L}^x_{\text{recon}}$. As a result, style-augmented cycle consistency is implicitly implied by the proposed bidirectional reconstruction loss.

Proposition 5 (Cycle consistency implies deterministic translations). *Let $p(x_1)$ and $p(x_2)$ denote the data distributions. $p_G(x_1|x_2)$ and $p_G(x_2|x_1)$ are two conditionals defined by generators in the CycleGAN work [5]. Given*

1. *matched marginals:*

$$p(x_1) = \int p_G(x_1|x_2)p(x_2)\,dx_2,\ p(x_1) = \int p_G(x_2|x_1)p(x_1)\,dx_1,$$

2. *cycle consistency:*

$$\mathbb{E}_{x_1^*\sim p(x_1), x_2\sim p_G(x_2|x_1^*)}[p_G(x_1|x_2)] = \delta(x_1 - x_1^*),$$
$$\mathbb{E}_{x_2^*\sim p(x_2), x_1\sim p_G(x_1|x_2^*)}[p_G(x_2|x_1)] = \delta(x_2 - x_2^*)$$

then $p_G(x_1|x_2)$ and $p_G(x_2|x_1)$ collapse to deterministic delta functions.

Proof. Let x_1^* be a sample from $p(x_1)$. x_2', x_2'' are two samples from $p_G(x_2|x_1^*)$. Due to cycle consistency in $\mathcal{X}_1 \to \mathcal{X}_2 \to \mathcal{X}_1$, we have $p_G(x_1|x_2') = p_G(x_1|x_2'') = \delta(x_1 - x_1^*)$. Also, $x_2' \in \mathcal{X}_2$ and $x_2'' \in \mathcal{X}_2$ because of matched marginals. Due to cycle consistency in $\mathcal{X}_2 \to \mathcal{X}_1 \to \mathcal{X}_2$, we have $p_G(x_2|x_1^*) = \delta(x_2 - x_2') = \delta(x_2 - x_2'')$. Thus $p_G(x_2|x_1)$ collapses to a delta function, similar for $p_G(x_1|x_2)$. This proposition shows that cycle consistency [5] is too strong a constraint for multimodal image translation.

2 Training Details

2.1 Hyperparameters

We use the Adam optimizer [6] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate of 0.0001. The learning rate is decreased by half every $100,000$ iterations. In all experiments, we use a batch size of 1 and set the loss weights to $\lambda_x = 10$, $\lambda_c = 1$, $\lambda_s = 1$. We use the domain-invariant perceptual loss with weight 1 in the street scene and Yosemite datasets. We choose the dimension of the style code to be 8 across all datasets. Random mirroring is applied during training.

2.2 Network Architectures

Let `c7s1-k` denote a 7×7 convolutional block with k filters and stride 1. `dk` denotes a 4×4 convolutional block with k filters and stride 2. `Rk` denotes a residual block that contains two 3×3 convolutional blocks. `uk` denotes a $2\times$ nearest-neighbor upsampling layer followed by a 5×5 convolutional block with

k filters and stride 1. `GAP` denotes a global average pooling layer. `fck` denotes a fully connected layer with k filters. We apply Instance Normalization [7] to the content encoder and Adaptive Instance Normalization [8] to the decoder. We use ReLU activations in the generator and Leaky ReLU with slope 0.2 in the discriminator. We use multi-scale discriminators with 3 scales.

- Generator architecture
  - Content encoder: `c7s1-64, d128, d256, R256, R256, R256, R256`
  - Style encoder: `c7s1-64, d128, d256, d256, d256, GAP, fc8`
  - Decoder: `R256, R256, R256, R256, u128, u64, c7s1-3`
- Discriminator architecture: `d64, d128, d256, d512`

3 Additional Results

3.1 User studies

We additionally performed a user study to evaluate diversity. We presented users two pool of translations, one from our method and another from a compared method. Each pool contains 9 random translations. We then ask the users which pool is more diverse. As shown in Table 1, users find our method to generate more diverse translations than other methods.

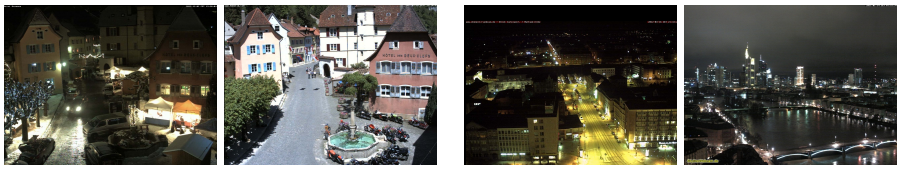
We run a user study to compare the image quality generated by different methods to ground truth images. At each time, users are shown the input image, the ground truth image and an output image from an algorithm. They are then asked to select the image that looks more realistic, given unlimited time. Table 2 shows the comparison results. We find that images generated from our method are more realistic than those from unsupervised baselines (UNIT, CycleGAN), and on par with results from the supervised BicycleGAN.

Table 1. Results of the user study on diversity evaluation. The number is the percentage each method is preferred over MUNIT.

	edges → shoes	edges → handbags
UNIT	1.1%	2.0%
CycleGAN	2.6%	4.9%
CycleGAN* with noise	2.4%	2.6%
MUNIT w/o $\mathcal{L}^x_{\text{recon}}$	19.0%	31.8%
MUNIT w/o $\mathcal{L}^c_{\text{recon}}$	36.5%	28.3%
MUNIT w/o $\mathcal{L}^s_{\text{recon}}$	8.4%	11.6%
MUNIT	50.0%	50%
BicycleGAN	36.3%	36.8%

Table 2. Results of the user study on quality evaluation. The number is the percentage each method is preferred over real images.

	edges → shoes	edges → handbags
UNIT	16.4%	9.2%
CycleGAN	13.6%	13.4%
MUNIT	20.2%	**21.0**%
BicycleGAN	**24.6**%	16.6%



(a) Image pairs from the same scene. (b) Image pairs from the same domain.

Fig. 1. Example image pairs.

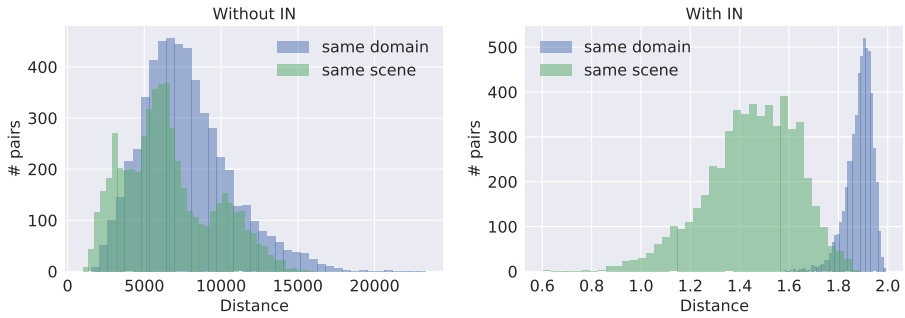


Fig. 2. Histograms of the VGG feature distance. Left: distance computed without using IN. Right: distance computed after IN. Blue: distance between image pairs from the same domain (but different scenes). Green: distance between image pairs from the same scene (but different domains).

3.2 Domain-invariant Perceptual Loss

We conduct an experiment to verify if applying IN before computing the feature distance can indeed make the distance more domain-invariant. We experiment on the day ↔ dataset used by Isola *et al.* [9] and originally proposed by Laffont *et al.* [10]. We randomly sample two sets of image pairs: 1) images from the same domain (both day or both night) but different scenes, 2) images from the same scene but different domains. Fig. 1 shows examples from the two sets of image pairs. We then compute the VGG feature (`relu4_3`) distance between each image pair, with IN either applied or not before computing the distance. In Fig. 2, we show histograms of the distance computed either with or without IN, and from

Fig. 3. Comparison with style transfer methods.

image pairs either of the same domain or the same scene. Without applying IN before computing the distance, the distribution of feature distance is similar for both sets of image pairs. With IN enabled, however, image pairs from the same scene have clearly smaller distance, even they come from different domains. The results suggest that applying IN before computing the distance makes the feature distance much more domain-invariant.

3.3 Qualitative Comparisons on Style Transfer

In Fig. 3, we compare our method with classical style transfer algorithms including Gatys *et al.* [11], Chen *et al.* [12], AdaIN [8], and WCT [13]. Our method produces style transfer results that are significantly more faithful than existing works. Also, our results appear to be much more realistic, since our method learns the distribution of target domain images using GANs. While CycleGAN [5] and other image-to-image translation methods can also learn the image distribution, they are not able to perform example guided style transfer since they do not learn a disentangled representation of content and style.

3.4 Failure Cases

Our method often fails when the content of the input image significantly deviates from the content distribution of the target domain. For example, when a jaguar opens its mouth wide, it is difficult to transfer it into house cats since very few house cats are in this pose (Fig. 4 (a)). While our algorithm can handle

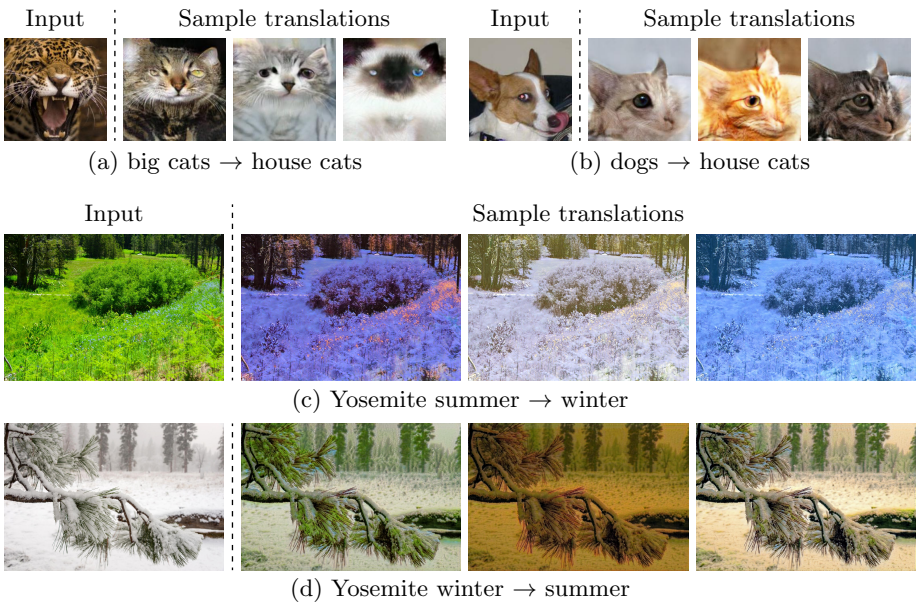


Fig. 4. Failure cases.

shape transformations in some cases, it could still fail when the required shape transformations are too large (Fig. 4 (b)). Also, it is challenging to remove flowers during Yosemite summer → winter (Fig. 4 (c)), and to remove large areas of snow during Yosemite winter → summer (Fig. 4 (d)). In addition, we find that our method does not preserve the image background during animal image translation, since it considers the background as part of the style. Using object masks as in Liang *et al.* [14] might resolve this problem.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)
2. Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., Carin, L.: Alice: Towards understanding adversarial learning for joint distribution matching. In: NIPS. (2017)
3. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: ICLR. (2018)
4. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
5. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. (2017)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2015)
7. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: CVPR. (2017)

8. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. (2017)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. (2017)
10. Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. TOG (2014)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. (2016)
12. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. arXiv preprint arXiv:1612.04337 (2016)
13. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: NIPS. (2017) 385–395
14. Liang, X., Zhang, H., Xing, E.P.: Generative semantic manipulation with contrasting gan. arXiv preprint arXiv:1708.00315 (2017)