MCMI: Multi-Cycle Image Translation with Mutual Information Constraints

Xiang Xu   **Megha Nawhal**   **Greg Mori**   **Manolis Savva**
Simon Fraser University
{xuxiangx, mnawhal, mori, msavva}@sfu.ca

Abstract

We present a mutual information-based framework for unsupervised image-to-image translation. Our MCMI approach treats single-cycle image translation models as modules that can be used recurrently in a multi-cycle translation setting where the translation process is bounded by mutual information constraints between the input and output images. The proposed mutual information constraints can improve cross-domain mappings by optimizing out translation functions that fail to satisfy the Markov property during image translations. We show that models trained with MCMI produce higher quality images and learn more semantically-relevant mappings compared to state-of-the-art image translation methods. The MCMI framework can be applied to existing unpaired image-to-image translation models with minimum modifications. Qualitative experiments and a perceptual study demonstrate the image quality improvements and generality of our approach using several backbone models and a variety of image datasets.

1   Introduction

Image-to-image (I2I) translation has gained prominence in recent years. The goal for I2I translation is to map an image from one domain to a corresponding image in another domain. This is particularly challenging in the unpaired case where no aligned training pairs are available. Most unpaired I2I translation models [33, 31, 17, 14, 11, 6, 15, 7] rely on cycle consistency loss to achieve realistic image generation. The underlying assumption is that semantic information is preserved across domains when the model is forced to reconstruct the original input from the translated output. However, for tasks with one-to-many mappings (e.g., photo colorization), the use of cycle consistency forces the model to translate back to only one possible outcome. This makes image generators susceptible to steganography, where information necessary to reconstruct the input is hidden in small perturbations of the output image [8]. Consequently, the output image is not constrained to be semantically related to the input in order to satisfy cycle consistency. This leads to less robust I2I models and inferior mapping functions which fail to preserve important semantic information across domains [8, 2].

In this work, we present an information-theoretic view of the problem and propose an I2I translation framework to address the issues with cycle consistency loss. Our insight is that when mapping from source to target domain, an ideal I2I model should only rely on semantic information from the source and not from any other extra data. We reformulate an I2I model that satisfy this constraint as a Markov chain and demonstrate that cycle consistency loss violates the Markov property as output and input images are forced to reconstruct each other. Ideally, we would like the image quality benefits from cycle consistency while also enforcing the Markov property to preserve semantic information.

To reconcile the Markov property and cycle consistency, we extend single-cycle translation to multi-cycle translation, proposing multi-cycle translation with mutual information constraints (MCMI). Specifically, we propose multi-cycle translation where the first translation cycle employs cycle consistency loss and we use the *data processing inequality theorem* to enforce non-increasing mutual

information (MI) for subsequent cycle translations. Any learned mapping function that does not satisfy this inequality will not correspond to a Markov chain and thus will fail to satisfy the Markov property. This avoids the function space where the Markov property will not be satisfied during training. Together with the cycle consistency loss in the first cycle, our MCMI framework results in higher quality image translation and semantically-relevant cross-domain mapping functions while adhering to the Markov property. To summarize, our contributions are as follows.

- We propose an information-theoretic formulation of unpaired image-to-image translation: the multi-cycle mutual information (MCMI) framework.
- We demonstrate the generality of MCMI by applying it to various single-cycle I2I models.
- We perform quantitative and qualitative experiments to show the MCMI framework produces higher quality images and more semantically-related cross-domain mappings compared to the backbone single-cycle I2I models.

2 Related Work

Image-to-image (I2I) translation maps an image from a source domain to an image in a target domain. Paired I2I translation refers to the setting where a dataset contains input-output pairs for learning the translation function. Prior work in this setting has used non-parametric models [9], parametric models such as CNNs[24], and conditional GANs[12, 27, 29]. Unpaired I2I translation refers to the problem setting where input-output pairs are not available. This setting implicitly requires learning relations between the source and target domains. Prior work tackling this problem spans a diverse array of techniques such as patch-based Markov Random Fields [23], cross-modal weight sharing networks [16, 1], and combination of VAEs and GANs [17, 26, 25, 5]. All these approaches focus on one-to-one mapping between the source and target domain.

Recently, there has been some work on one-to-many image translation [11, 14, 15, 32]. MUNIT [11] assumes content features share similar values across different domains; DRIT [14] further use a content discriminator to force the content features to be in the same latent space; DRIT++ [15] extends DRIT with mode-seeking loss [19]; and DMIT [28] provides an efficient model for both multi-modal and multi-domain translation.

Cycle consistency loss is widely used in the aforementioned unpaired I2I models. The issues with cycle consistency loss are first explored by Chu et al. [8] demonstrating that CycleGAN might lead to steganography using perturbations in the output to hide information on the input. Later work by Bashkirova et al. [2] formulate the problem as a self-adversarial attack. However, their method is inefficient as the model needs further training on adversarial images with injected noise.

3 MCMI: Multi-Cycle Translation with Mutual Information Constraints

Our goal is to perform image-to-image (I2I) translation in an unpaired setting where the input-output pairings are not available. Given an image x from domain X, we produce a set of generated images $\mathcal{Y}$ for a domain Y. For example, X could be the domain of *cat* images, while Y could be the domain of *dog* images. Given a particular instance of a cat image x, there are many corresponding dog images $y \in \mathcal{Y}$ that are potential translations (i.e. one-to-many mapping). Our approach focuses on ensuring two key properties in the output set $\mathcal{Y}$: i) it contains diverse elements that are realistic images of the domain Y; and ii) it preserves semantically-relevant content (e.g., animal pose or expression) during translation while allowing variation in visual content (e.g., breed of dog or fur color).

To learn semantically-relevant cross-domain mappings, the mapping of visual features should only share dependencies between the input and generated images within every cycle, but not with the sequence of generated data that preceded it from earlier cycles. Therefore, we formulate this problem as a Markov chain where the Markov property is satisfied across domain translations. Direct optimization of the Markov property is difficult. Such optimization would require enforcing conditional independence of future generated images from previously generated ones, given the current generated image. However, the inputs and outputs are conditionally dependent in an unconstrained I2I translation network. Instead, we propose the multi-cycle MI constraints that are easier to optimize.

Our framework is based on a multi-cycle Markov chain formulation of the I2I translation problem which we describe first in Section 3.1. Then, we describe the components comprising the MCMI

framework: i) a backbone single-cycle I2I module (Section 3.2); ii) an MI estimator (Section 3.3); iii) MI constraints (Section 3.4); and iv) an I2I loss function forcing the MI constraints (Section 3.5).

3.1 Multi-Cycle Image-to-image Translation as a Markov Chain

We first extend single-cycle translation to multi-cycle translation by treating an I2I translation model as a backbone module. Every translation cycle consists of $X^n \to Y^n \to X^{n+1}$. Here X, Y refer to two image domains, and the superscript indicates how many times a particular domain has been visited. At each cycle translation step, X^{n+1} is referred to as the reconstructed image of input X^n and Y^n is the translated image. The reverse translation going from Y to X is similar. For simplicity, we do not explicitly state the reverse process.

We propose to view an optimal multi-cycle I2I translation model as a Markov chain. This means that the reconstructed image X^{n+1} should only depend on the translated image Y^n and is independent of input X^n and all other images before it. Here, we treat image domains as random variables and visits to the same domain at different cycles as different random variables. This Markov chain assumption addresses the issues introduced by the cycle consistency loss as future image generations no longer rely on encoded information from the initial input. Under this assumption, the I2I model avoids encoding information from earlier translations and learning of trivial mapping functions.

However, the commonly-used cycle consistency loss violates the Markov property as it directly optimizes the reconstructed image to be close to the input with a direct pixel-level loss. Simply eliminating the cycle consistency loss causes decreased image quality [2]. To reconcile the Markov property and cycle consistency, we formulate our loss functions within a two-cycle translation: $X^1 \to Y^1 \to X^2 \to Y^2 \to X^3$. For the first cycle, we apply the cycle consistency loss (encouraging X^2 close to X^1). For the second cycle, we only use the MI constraints between output Y^2, X^2, X^3 and input X^1 to characterize and bound the cross-domain mappings and achieve the desired Markov property. The same I2I module is used in both cycles so there is no increase in network parameters. We empirically find that two cycles offer a good trade-off between performance, memory use, and efficiency. Please refer to the supplementary material for more details.

Directly optimizing towards a Markov chain is difficult. We instead use the *data processing inequality theorem* [3] which states that given a Markov chain, the MI must be non-increasing going from input to the output. MCMI enforces the non-increasing MI constraints in the second cycle with $I(X^1; X^2) \geq I(X^1; Y^2) \geq I(X^1; X^3)$. Any learned mapping function that does not satisfy this inequality will not correspond to a Markov chain and fails to satisfy the Markov property. Explicitly optimizing for non-increasing MI will train a model that avoids the function space where the Markov property is not satisfied. Combined with the cycle consistency loss in the first cycle, we obtain better quality image translations and semantically-relevant cross-domain mapping functions.

3.2 Backbone Module

MCMI starts with a backbone I2I module that can be any of the existing models in the literature. We examine three state-of-the-art models: CycleGAN [31], MUNIT [11] and DRIT++ [15]. Here, we describe the key differences between these modules.

CycleGAN consists of two image mapping functions $G : X \to Y$ and $F : Y \to X$. The self-reconstruction loss is used together with two image domain discriminators to train the model. The mapping functions G and F are deterministic and do not lead to one-to-many mapping. Refer to Zhu et al. [31] for more details.

MUNIT makes the assumption that image features can be disentangled into two latent spaces: a domain-invariant content space, and a domain-specific style space. During training, only the encoded content feature is shared across domain. The style feature is swapped at each step. Cycle consistency is enforced after one full cycle translation. At test time, style feature can be randomly sampled from $\mathcal{N}(0, I)$ and this allows one-to-many mappings. For more details refer to Huang et al. [11].

DRIT++ is based on DRIT [14] with feature disentanglement and cross-domain mapping similar to MUNIT. A content discriminator is introduced to ensure the content feature from different domains map to the same content latent space. A mode seeking loss [19] also encourages the generators to

explore more modes during training and leads to more diverse one-to-many image mappings based on style features. Refer to Lee et al. [15] for more details.

3.3 Mutual Information Estimator

The mutual information $I(X; Y)$ is equivalent to the Kullback-Leibler (KL) divergence between the joint distribution $\mathbb{P}_{XY}$ and the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Y$. It is defined as: $I(X; Y) = D_{KL}(\mathbb{P}_{XY} \| \mathbb{P}_X \otimes \mathbb{P}_Y)$. In practice, computing the exact MI between two high-dimensional random variables (e.g., two images) is intractable. Therefore, approximate algorithms have been used to compute a lower bound of the MI. MINE [4] uses the Donsker-Varadhan representation of KL divergence to derive a lower bound of the MI: $I_\theta(X; Y) \leq I(X; Y)$ expressed as $I_\theta(X; Y) = \sup \ (\mathbb{E}_{\mathbb{P}_{XY}}[T_\theta]) - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}[e^{T_\theta}])$. InfoNCE [20] is another approach for estimating the MI. With K samples of (x, y) from two different random variables X, Y, the lower bound is represented as:

$$I(X; Y) \geq \mathbb{E}\left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{f(x_i, y_i)}}{\frac{1}{K} \sum_{j=1}^{K} e^{f(x_j, y_i)}}\right] \triangleq I_{\text{NCE}}^{\text{lower}}, \tag{1}$$

where $f(x, y)$ is a critic function approximating the density ratio that preserves the MI between X, Y [20]. Alternatively, Poole et al. [22] derive an upper bound on the MI expressed as:

$$I(X; Y) \leq \mathbb{E}\left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{p(y_i|x_i)}{\frac{1}{K-1} \sum_{j \neq i} p(y_i|x_j)}\right], \tag{2}$$

where K represents the number of samples and $p(y|x)$ is the conditional probability. Given that an optimal critic for $I_{\text{NCE}}^{\text{lower}}$ is achieved when $f(x, y) = \log p(y|x)$ [20, 22], Equation (2) with near optimal critic can be re-written as:

$$I(X; Y) \leq \mathbb{E}\left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{e^{f(x_i, y_i)}}{\frac{1}{K-1} \sum_{j \neq i} e^{f(x_j, y_i)}}\right] \triangleq I_{\text{NCE}}^{\text{upper}}. \tag{3}$$

The key difference between $I_{\text{NCE}}^{\text{lower}}$ and $I_{\text{NCE}}^{\text{upper}}$ is whether or not $f(x_i, y_i)$ is in the denominator. This means that once we have a good estimate of the lower bound with a near optimal critic, the upper bound is also known and the actual MI is bounded between these two values.

We use Equations (1) and (3) based on InfoNCE to estimate the lower and upper bounds of the MI. In our setting, (x, y) are images sampled from different translation steps. In Section 3.4 we will use InfoNCE as a tight constraint to support our multi-cycle MI constraints. We decided against MINE as it only gives a lower bound on the MI, resulting in worse performance. The supplemental material provides an ablation experiment to study the impact of each of the bounds.

We define the critic function $f(x, y)$ as the cosine similarity between two image representations after passing through a shared convolutional network conv(). Concretely, $f(x, y) = |\cos(\text{conv}(x), \text{conv}(y))| \cdot s - m$, where s is a scale and m is an offset. We use the absolute cosine similarity because two image features encoded by opposite vectors (i.e. cosine similarity of -1) should have high MI as knowing one allow us to know the other with high confidence. The smallest MI value occurs when cosine similarity equals zero, indicating orthogonality or no correlation. The scale and offset are used to ensure the largest conditional probability $p(y|x)$ is at most 1. We set both s and m to 2.0 in our experiments by hyper-parameter tuning on the validation set.

Equation (3) is most accurate in the optimal critic case, thus during training we have a additional step that maximizes the lower bound $I_{\text{NCE}}^{\text{lower}}$ to ensure that the critic stays close to optimal. In summary, we can use these two bounds to estimate MI across translation steps, which can in turn be used in training our I2I model. Detailed training pipeline is provided in the supplementary material.

3.4 Multi-Cycle Mutual Information Constraints

We use the estimated MI (using neural estimation) to derive an optimization objective for non-increasing MI. We enforce the MI between generated images at the current translation step and the initial ground truth images to be non-increasing. Formally, MI at translation step n: $X^n \rightarrow Y^n \rightarrow$

X^{n+1}, can be written as $I(X^1; X^n) \geq I(X^1; Y^n)$ and $I(X^1; Y^n) \geq I(X^1; X^{n+1})$. In the previous section we discussed how MI cannot be accurately estimated and hence is difficult to use directly in a loss function. Thus, we use our estimated MI lower and upper bounds. To show that the new inequality constraints still achieve non-increasing MI, we note the following holds for any MI lower and upper bounds: $I_{(q;p)}^{\text{lower}} \geq I_{(m;n)}^{\text{upper}} \implies I(q;p) \geq I(m;n)$. That is, if the lower bound MI for (q, p) is larger or equal to the upper bound MI for (m, n), then the actual MI for (q, p) is guaranteed to be larger or equal to the actual MI for (m, n). This means that we can strictly optimize towards non-increasing MI constraints with the lower and upper bounds from Equations (1) and (3). In our experiments, we found that enforcing the MI constraints only on the second cycle is sufficient to achieve good image quality. We can now reformulate the MI constraints for $X^1 \rightarrow Y^1 \rightarrow X^2 \rightarrow Y^2 \rightarrow X^3$ as:

$$
\begin{aligned}
I_{\text{NCE}}^{\text{lower}}(X^1; X^2) &\geq I_{\text{NCE}}^{\text{upper}}(X^1; Y^2) \\
I_{\text{NCE}}^{\text{lower}}(X^1; Y^2) &\geq I_{\text{NCE}}^{\text{upper}}(X^1; X^3).
\end{aligned}
\tag{4}
$$

3.5 Overall MCMI Loss Function

The MI loss function for a two-cycle translation model is defined as:

$$
L_{\text{MI}} = \max\left(I_{\text{NCE}}^{\text{upper}}(X^1; Y^2) - I_{\text{NCE}}^{\text{lower}}(X^1; X^2), 0\right) + \max\left(I_{\text{NCE}}^{\text{upper}}(X^1; X^3) - I_{\text{NCE}}^{\text{lower}}(X^1; Y^2), 0\right),
$$

with the other translation direction swapping the order of images from the two domains. In addition to L_{MI}, we use an image adversarial loss L_{adv} to promote realism of the generated image in the second cycle. We also increase the lower bound of our estimated MI. This ensures the generated images are bounded and that we are training towards the optimal critic for I_{NCE}. The overall loss function for MCMI is:

$$
L_{\text{MCMI}} = L_{\text{orig}} + L_{\text{adv}} + \alpha L_{\text{MI}} - \beta I_{\text{lower}}.
\tag{5}
$$

Here, L_{orig} refers to the original loss function used in an I2I translation model backbone (CycleGAN, MUNIT and DRIT++). This includes the image adversarial loss, the content adversarial loss in DRIT++, and the cycle consistency loss applied to all first cycle translations. Our approach does not rely on a specific I2I translation model, as we do not make any modification to the networks or loss functions within the I2I translation module. α and β are hyper-parameters that control the trade-off between different loss terms.

4 Experiments

We demonstrate the effectiveness of the MCMI framework through quantitative and qualitative evaluation. We show that our approach can synthesize more realistic and semantically-related images for a diverse set of unpaired I2I translation models and datasets. Ablation experiments analyze the impact of the multi-cycle MI constraints versus cycle consistency loss.

4.1 Experimental Procedure

Evaluation metrics. We define three metrics to evaluate MCMI relative to baseline I2I models:

- **Quality.** We measure image quality and realism in two ways. Firstly, we use the Fréchet Inception Distance (FID) [10] which has been shown to correlate with image quality. We compute FID scores from samples generated at the first (FID1) and second (FID2) cycles. Secondly, we perform a user study that involves a 'quality' preference judgment.
- **Diversity.** We use LPIPS [30] to measure realism and diversity of the generated images. We follow the evaluation setup used for StarGAN v2 [7] where 10 images are randomly generated per test image and the average of the pairwise distances are computed among all outputs generated from the same input. We do not find significant differences in LPIPS for samples in the first and second cycles and report the LPIPS on first cycle samples. CycleGAN models do not support one-to-many translation so we compute LPIPS from 4 samples generated at different cycles of the translation (first to fourth cycles).
- **Markovness.** On datasets with ground truth image pairs, we compute the pixel-wise mean squared error ϵ_Y for the translated image in the target domain and ϵ_X for the reconstructed image in the source domain. Note that images are normalized to $[0, 1]$. We define the ratio $\epsilon_{\text{mark}} = \epsilon_Y / \epsilon_X$ as a 'Markovness error'. High ϵ_{mark} indicates the model reconstructs input X
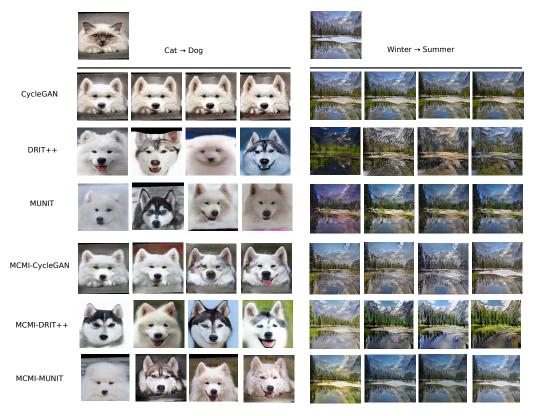
Figure 1: Images generated with MCMI (two-cycle) and corresponding single-cycle baselines. The top-left image is the input image. Samples for CycleGAN models are from successive output images after multiple translation steps. Samples for MUNIT and DRIT++ models come from different style codes at the first cycle. The MCMI-adapted models produce higher quality, diverse images. Additional examples in the supplement.

while not producing translated images that match Y well, indicating higher deviation from the Markov assumption.

Datasets. We evaluate on commonly used I2I datasets including the pets dataset (Cat→Dog) [31], Yosemite dataset (Winter→Summer scenes) [31], CelebA [18], and the Google Aerial Photo to Maps dataset [31]. These datasets span different scales and have diverse attributes and structural correspondences.

Implementation details. We implement our proposed method in PyTorch [21]. We use the official PyTorch implementations of the CycleGAN, MUNIT, DRIT++, and StarGAN models. Images are randomly cropped to 256x256 for CycleGAN and MUNIT, 216x216 for DRIT++ and 128x128 for StarGAN. The feature encoder $f(x, y)$ used to compute the MI estimates uses four convolutional layers with instance batch normalization and leaky ReLU activation. Optimizer and I2I network parameters are the same as the corresponding original implementation.

Baselines. We compare MCMI-adapted I2I translation modules against baselines using each module in isolation. We adapt each baseline to the MCMI setup according to Equation (5). CycleGAN [31], MUNIT [11] and DRIT++ [15] are the baselines using a standard single-cycle setup. MCMI-CycleGAN, MCMI-MUNIT, and MCMI-DRIT++ are the corresponding MCMI-adapted models.

4.2 Results

We discuss our extensive qualitative and quantitative evaluation here. Please refer to the supplemental material for additional results.

Table 1: Comparison of MCMI-adapted CycleGAN, MUNIT and DRIT++ models against baselines. Arrows indicate whether lower (↓) or higher (↑) value is better. Standard deviations on two trained models and 4 random seeds indicated for FID. All standard deviations for LPIPS are ≤ 0.003. We see that MCMI achieves better image quality (measured by FID), while preserving diversity (LPIPS).

Model	Cat→Dog			Winter→Summer		
	FID1↓	FID2↓	LPIPS↑	FID1↓	FID2↓	LPIPS↑
CycleGAN	78.1 ± 0.57	70.9 ± 0.61	0.09	65.0 ± 0.66	61.5 ± 0.75	0.07
MUNIT	21.4 ± 1.05	25.8 ± 1.04	0.31	55.5 ± 0.04	63.4 ± 0.05	0.22
DRIT++	21.9 ± 0.41	22.1 ± 0.35	**0.32**	39.0 ± 0.11	44.0 ± 0.09	0.24
MCMI-CycleGAN	62.9 ± 0.34	28.0 ± 0.34	0.22	63.2 ± 0.58	60.4 ± 0.45	0.18
MCMI-MUNIT	18.6 ± 0.21	21.9 ± 0.20	0.30	48.3 ± 0.22	50.2 ± 0.21	0.24
MCMI-DRIT++	$\mathbf{17.3 \pm 0.25}$	$\mathbf{17.2 \pm 0.27}$	**0.32**	$\mathbf{37.2 \pm 0.13}$	$\mathbf{41.5 \pm 0.17}$	**0.25**

Qualitative comparisons. Figure 1 shows generated images from baseline single-cycle I2I models and two-cycle MCMI-adapted counterparts. Images for CycleGAN models are sampled from different cycles (as explained in the definition of our diversity metric). Images for DRIT++ and MUNIT-based models come from the first cycle with different sampling of the style code. We observe that MCMI leads to higher image quality while preserving the structural content of the input (e.g., head pose for pets and slowly changing landscape for Yosemite images). For instance, the dog images from the MUNIT baseline have different head orientations. In contrast, MCMI-MUNIT maintains the head pose of the input. We also observe that MCMI-CycleGAN generates diverse images from different cycles despite the backbone model only supporting one-to-one mapping. For example, the dog fur color changes and the dog mouth opens up through the translation cycles.

Quantitative comparisons. Table 1 reports the quantitative evaluation of MCMI-adapted models against baseline I2I models. Here, we use all test set images for evaluation. We see that MCMI improves image quality (lower FID scores) while retaining diversity (similar LPIPS scores). The improvement is particularly pronounced for the Cat→Dog dataset where there are many high-level visual feature changes during translation. We also see that DRIT++ and MUNIT models have similar FID in the first and second cycles, whereas CycleGAN has much better FID in the second cycle where the translation is only supervised with our proposed MCMI constraints. We believe this is caused by CycleGAN I2I backbone module not supporting one-to-many translation.

User study. We generated sets of 4 images using the MCMI-DRIT++ model and corresponding DRIT++ baseline for 20 randomly sampled images cat images in the cat→dog dataset. The images in each set are the generation outputs in the target domain after each translation step (from 1 step to 4 steps). We then recruited 10 participants for our study. For each such test instance of a source image and two target generated sets (one produced by MCMI-DRIT++ and one by DRIT++), the participants were asked to select which of the two sets has: i) higher quality, ii) higher overall set diversity. The MCMI-DRIT++ image sets were chosen as having higher quality 79% on average (remaining 16% for baseline DRIT++ images, and 5% ambiguous), while they were chosen to be more diverse 46% of the time (52% for DRIT++, and 2% ambiguous). These results indicate that MCMI generates more realistic images at comparable overall diversity.
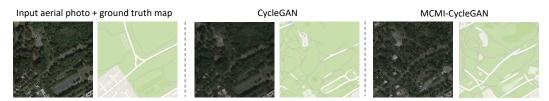

Figure 2: Aerial photo to map translations. MCMI-CycleGAN better preserves semantic information from generated maps to reconstructed photo, exhibiting more Markovian behavior.

Table 2: Average mean squared error for translated maps (ϵ_Y) and reconstructed aerial photos (ϵ_X) at first and second cycles. The Markovness error ϵ_{mark} is defined as ϵ_Y/ϵ_X for each cycle.

	First cycle			Second cycle		
	$\epsilon_Y \downarrow$	ϵ_X	$\epsilon_{\text{mark}} \downarrow$	$\epsilon_Y \downarrow$	ϵ_X	$\epsilon_{\text{mark}} \downarrow$
CycleGAN	0.0238	0.0056	4.25	0.0243	0.0076	3.20
MCMI-CycleGAN	**0.0211**	0.0141	**1.50**	**0.0201**	0.0169	**1.19**

Figure 3: Multi-domain I2I translation on the CelebA [18] dataset. Comparisons between generated images from our MCMI-adapted StarGAN and the baseline StarGAN model.

Markovness comparison. To measure the degree to which MCMI helps to enforce the Markov property, we compare the CycleGAN baseline with MCMI-CycleGAN on the Google Aerial Photo to Maps dataset where paired images are available. During training, we still train with randomly shuffled unpaired images, but at evaluation time we can compute ϵ_{mark} using the ground truth pairs. Figure 2 provides qualitative examples. CycleGAN produces map translations that differ significantly from the ground truth while the reconstructed aerial photos are similar to the input. This indicates that CycleGAN is far from Markovian. On the other hand, MCMI-CycleGAN more faithfully utilizes semantic information from the previous input. For instance, the lower left corner of the MCMI-CycleGAN translated map has three side roads that connect to the long main road rather than two. The reconstructed aerial photo also reflects this change. Quantitative evaluation in Table 2 confirms that MCMI-CycleGAN is more Markovian. On all test images, MCMI-CycleGAN achieves both the lowest pixel-wise MSEs for translated maps and Markovness errors at first and second cycle.

Multi-domain I2I translation. To further demonstrate the generality of our approach we also train a MCMI-StarGAN model on a multi-domain I2I translation task using the CelebA datasets. The CelebA [18] dataset is a face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. We follow the setup for StarGAN [6] and train on five selected attributes. Figure 3 shows the results. Overall, MCMI-StarGAN produces better quality images. For example, we see that baseline model generates a lot of hair in regions where should not be (third column), and in cases where part of the face is occluded (e.g., by a hand), StarGAN generates noisy images around the occluded region while MCMI-StarGAN does not generate black marks.

5   Conclusion

We present a mutual information-based approach for the unsupervised image-to-image (I2I) translation problem. Our MCMI framework extends several state-of-the-art I2I modules to enable multi-cycle I2I translation bounded by mutual information constraints. Qualitative and quantitative experiments show that the generated images exhibit increased quality at similar overall diversity, and a human user study further confirms the quality improvements. We show that MCMI-adapted I2I models are more Markovian in that they better utilize semantic information from input images during cross-domain mapping. Though we have focused on unsupervised I2I translation, our approach can be extended to estimate and bound mutual information changes when additional supervision is available in the form of image labels and annotated attributes. We believe our information theoretic view can be applied to problem domains other than images.

References

[1] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[2] D. Bashkirova, B. Usman, and K. Saenko. Adversarial self-defense for cycle-consistent gans. In *Advances in Neural Information Processing Systems*, pages 635–645, 2019.

[3] N. J. Beaudry and R. Renner. An intuitive proof of the data processing inequality. *arXiv e-prints*, art. arXiv:1107.0740, July 2011.

[4] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[5] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. *arXiv preprint arXiv:1912.01865*, 2019.

[8] C. Chu, A. Zhmoginov, and M. Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[9] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[11] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[14] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[15] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.

[16] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[17] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[19] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019.

[20] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[22] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

[23] R. Rosales, K. Achan, and B. J. Frey. Unsupervised image translation. In *iccv*, pages 472–478, 2003.

[24] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017.

[25] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[26] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[27] X. Wang, A. Jabri, and A. A. Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[28] X. Yu, Y. Chen, S. Liu, T. Li, and G. Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2019.

[29] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2759–2768, 2019.

[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[32] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017.

[33] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.

A   Supplementary Overview

This supplement provides the following additional information supporting the main paper.

- In Appendix B, we provide additional comparisons on the Photo→Monet and Photo→Portrait datasets. We also provide more generated images for Cat→Dog, Summer→Winter and CelebA attributes to supplement Section 4.2.

- Appendix C provides additional ablation studies to investigate the impact of the number of cycle translations, the effect of different mutual information bounds, and the impact of the margin value, to supplement Sections 3.1 and 3.3 of the main paper.

- Appendix D provides technical details for the training pipeline described in Section 3.3.

B   Additional Qualitative Results

Photo→Portrait and Photo→Monet results. We perform additional qualitative comparisons on the Photo→Portrait [31] and Photo→Monet datasets [31]. The comparisons to CycleGAN-based baselines follow the same settings as described in Section 4.1, where sampled images come from four different cycles. On portrait datasets, we follow the original CycleGAN training procedure and introduce an additional identity loss in the first cycle. See Section 5.2 and Figure 9 of the CycleGAN paper [31] for identity loss.

Generated samples for Photo→Portrait are in Figure 4. We observe that the single cycle CycleGAN baseline sometimes fails to generate portraits that have different art styles from the input photos. Results from different cycles also look similar due to the fixed one-to-one mapping network structure. However, when trained under the MCMI framework, MCMI-CycleGAN produces better translated images that look more like portrait paintings on the second cycle (second column for each input example). We observe well-preserved facial structure, while there are changes to facial features such as perceived smoothness of the skin during multi-cycle translations. Similarly to CycleGAN-based results on other datasets, the output of the first cycle supervised by self-reconstruction loss is of inferior quality as compared to the output of the second cycle.

Figure 5 shows generated samples for Photo→Monet. It is difficult to reason about image quality for Monet paintings, but we do find that both MCMI-CycleGAN and the baseline can preserve semantic structure of the input photo at the first and second cycles. One interesting observation is that in later cycles, MCMI-CycleGAN generates Monet images that look very different from the initial input. For example, the sea surface in the bottom-left sample is slowly transformed into ground and the last few photos become a grass field rather than a lake. This means that the model trained under the MCMI framework has more diverse mappings and that later cross-domain translations only rely on the semantic information from the current cycle input, not the initial input where there is a lake in the scene. This indicates more Markovian behavior for the model adapted using MCMI.

Additional Samples. We provide additional samples for the datasets introduced in Section 4.2. Figure 6 shows results for MCMI-CycleGAN and MCMI-DRIT++ on the Cat→Dog dataset. Following a similar setup as in Section 4.2, samples for MCMI-CycleGAN are from successive output images at different cycles. Section 4.2 demonstrates that first cycle MCMI-CycleGAN results are inferior. Thus, we ignore the first cycle and show results starting from the second cycle. Samples for the MCMI-DRIT++ model come from different style codes at the first cycle.

Figure 7 provides additional results for MCMI-CycleGAN and MCMI-DRIT++ on the Summer→Winter dataset. We follow the standard setup for DRIT++ [15] and use the simple concatenation model for fair comparison.

Figure 8 shows additional results for MCMI-StarGAN trained on the CelebA dataset [18]. We observe that our MCMI framework produces more realistic wrinkles on aged pictures and is better at modeling the hair.

Figure 9 provides more samples of CycleGAN-based models on the Google Aerial Photo to Maps dataset [31]. We observe that the translated map and reconstructed aerial photo from MCMI-CycleGAN have more Markovian behavior. For instance in the last row, the lower-left part of the translated map contains trees instead of lake, and the reconstructed aerial photo then reflects this
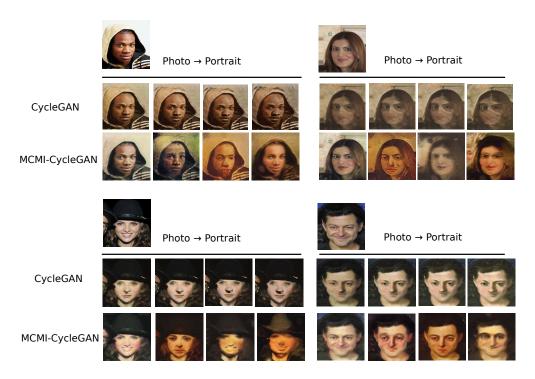
Figure 4: **Photo→Portrait results.** Comparisons between images generated using the MCMI approach and baseline modules. The top left image in each set is the input image. Then, each row shows output images corresponding to successive translation steps using the specified model.
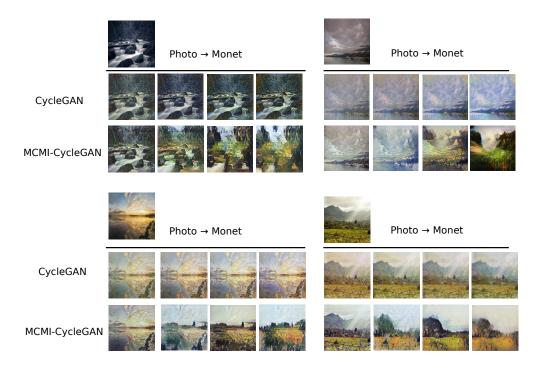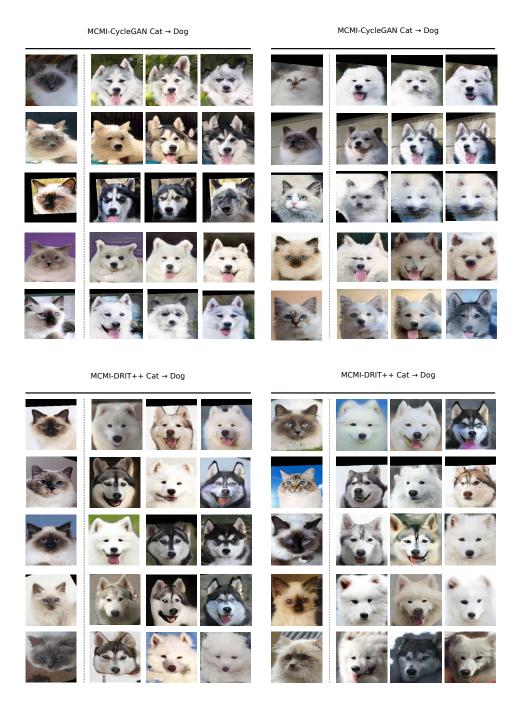

Figure 5: **Photo→Monet results.** Comparisons between images generated using the MCMI approach and baseline modules. The top left image in each set is the input image. Then, each row shows output images corresponding to successive translation steps using the specified model.

MCMI-CycleGAN Cat → Dog

MCMI-CycleGAN Cat → Dog

MCMI-DRIT++ Cat → Dog

MCMI-DRIT++ Cat → Dog

Figure 6: **Additional Cat→Dog results.** The first image in each row is the input cat image. Samples for MCMI-CycleGAN are from successive output images after multiple translation steps (ignoring the first cycle). Samples for MCMI-DRIT++ come from different style codes at the first cycle.
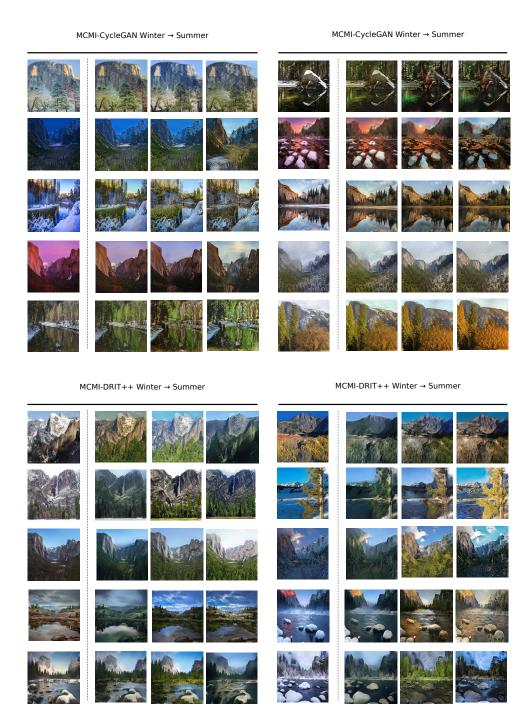
MCMI-CycleGAN Winter → Summer

MCMI-CycleGAN Winter → Summer

MCMI-DRIT++ Winter → Summer

MCMI-DRIT++ Winter → Summer

Figure 7: **Additional Summer→Winter Samples.** First image at each row is the input winter image. Samples for MCMI-CycleGAN are from successive output images after multiple translation steps (ignoring first cycle). Samples for MCMI-DRIT++ models come from different style codes at the first cycle.
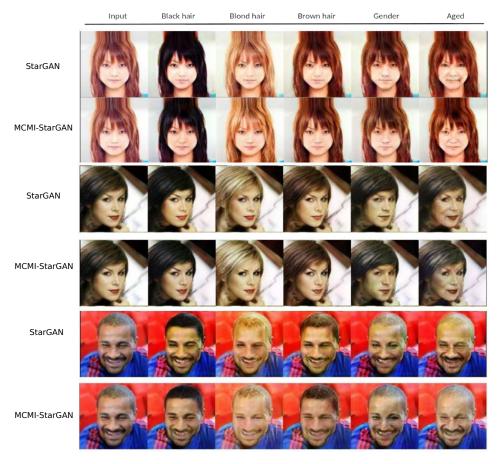
Figure 8: **Additional multi-domain I2I translation results on the CelebA [18] dataset.** Each pair of rows compares the baseline StarGAN model with an MCMI-adapted StarGAN model.
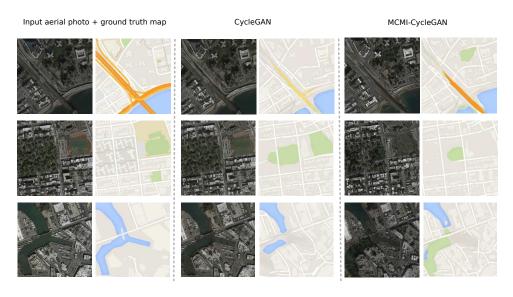

Figure 9: **Additional Aerial photo to map results**. MCMI-CycleGAN better preserves the semantic information from generated maps to reconstructed photo, exhibiting more Markovian behavior.
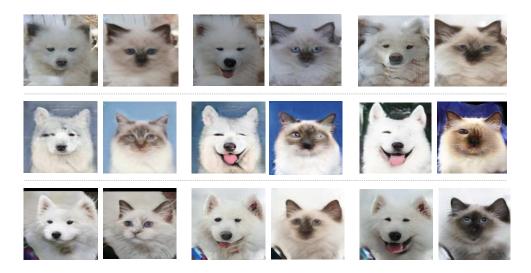
Figure 10: **Markovian behavior on Cat→Dog dataset**. Each row represents one multi-cycle translation with corresponding cat and dog images at each cycle.

difference. On the other hand, the standard CycleGAN model reconstructs the lake even though the translated map contains no lake in the bottom-left region.
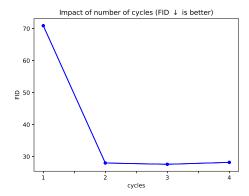
To further demonstrate Markovian behavior from our MCMI-adapted models, we also show the corresponding input cat image and generated dog image at each cycle on the Cat→Dog dataset. Figure 10 illustrates three different multi-cycle translations from MCMI-CycleGAN. We can see that when the associated cat image changes, the translated dog image is also different at each cycle. For instance, in the second row, when one of the eyes begins to disappear, subsequent results also have a disappearing eye.

C Additional Ablation Studies

Impact of the number of cycle translations. We enforce our proposed mutual information constraints on different numbers of cycle translations and study the effect on model performance. We conduct this ablation with the MCMI-CycleGAN model on the Cat→Dog dataset. Figure 11 (left) plots the FID score across different numbers of translation cycles. Note that the non-increasing mutual information constraints are enforced. We observe that the performance gain is not significant after two cycles of translations. However, GPU memory use increases for each additional translation step. This supports that our choice of two cycle MCMI in Section 4.5 offers a good trade-off between performance and memory.

Effect of the MI margin. Equation 4 from the main paper strictly enforces the non-increasing MI constraints with upper and lower bounds. However, this constraint might be too hard. It is possible to add a margin to L_{MI}. A negative margin means that we relax the constraint and a positive margin means that MI must decrease by a certain value. For this ablation study, we use different values of the margin and evaluate their effect on the model performance. Similar to before, we use CycleGAN as our backbone I2I module and train the model on the Cat→Dog dataset. Figure 11 (right) shows the FID score plotted against different margin values. We see that a small margin value between -0.4 and 0.0 results in the best performance.

Effect of different MI bounds. In Table 3, we report model performance on the Cat→Dog dataset with different versions of the MI bounds used in the training loss. We first compare to only increasing the MI lower bound without any MI constraints (CycleGAN + MI). We find that this leads to a decrease in diversity in the generated images (indicated by lower LPIPS) as well as less realistic
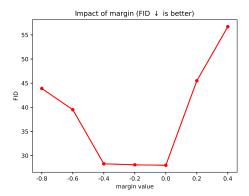
Figure 11: **Effect of number of cycle translations (left) and the MI constraint margin (right).** FID scores are reported for the MCMI-CycleGAN model trained on the Cat→Dog dataset. On the left, we observe that performance gain is not significant after two cycles. On the right, we find that a small margin between -0.4 and 0.0 offers the best performance.

Table 3: **CycleGAN-based models with different implementation of MI loss functions**. Arrows indicate whether lower ($\downarrow$) or higher ($\uparrow$) value is better. Standard deviations of test set numbers indicated for FID. All standard deviations for LPIPS values are ≤ 0.003.

	Cat→Dog		
Model	FID1$\downarrow$	FID2$\downarrow$	LPIPS$\uparrow$
CycleGAN + MI	69.6 ± 0.32	35.0 ± 0.33	0.10
MCMI-CycleGAN	62.9 ± 0.34	$\mathbf{28.0 \pm 0.34}$	**0.22**
MCMI-CycleGAN (lower bounds MI)	65.1 ± 0.28	32.3 ± 0.26	0.20
MCMI-CycleGAN (non-decreasing MI)	88.9 ± 0.44	59.6 ± 0.39	0.19
MCMI-CycleGAN (non-changing MI)	$\mathbf{58.7 \pm 0.30}$	36.0 ± 0.31	0.11

generation (shown by higher FID1 and FID2 scores). We then compare MCMI-CycleGAN with upper and lower bounds to using both lower bounds in the MI constraints (lower bounds MI), enforcing the exact opposite MI constraints (non-decreasing MI), and enforcing MI to be the same across cycles (non-changing MI). Similarly to before, all other versions result in decrease in LPIPS and worse FID scores for the translated images obtained after the first and second cycles. We only observe minor improvement in FID1 for non-changing MI. However, FID2 and LPIPS are much worse than before.

D   Technical Details

As described in Section 3.3, the mutual information evaluation module consists of four convolutional layers with instance batch normalization. The number of channels per layer is $64 \rightarrow 128 \rightarrow 128 \rightarrow 128$. We use a leaky ReLU as our activation function and max-pooling in earlier layers. The mutual information estimation network directly takes an entire image as input. During training, we separately optimize the I2I translation module and MI estimation module at two different steps. In the first step, we compute the MCMI loss L_{MI} and only optimize the I2I model with this gradient. In the second step, we compute I_{lower} and optimize the MI estimation model so that the MI lower bound is increased and the critic is closer to the optimal.

Note that the gradient from the MI estimation step does not flow back to the I2I translation model as we optimize the two modules separately. Only the MCMI loss L_{MI} optimizes the I2I model. InfoNCE [20] can not provide an accurate MI estimation with a small batch size. Thus, during training we also provide additional anchor images as input to the MI estimation network. We use a

batch size of 8 in all our experiments. For training MCMI-StarGAN on CelebA, we also found that the additional image adversarial loss leads to convergence issues, and do not use it.

We use the Adam [13] optimizer to train our models. The learning rate and decay schedule are kept the same as originally used for each baseline module. The learning rate is set to $1e^{-4}$ for DRIT++, MUNIT, and StarGAN-based models. The learning rate is $2e^{-4}$ for CycleGAN-based models. We train DRIT++ models for 1.2k epochs, MUNIT models for 50k iterations, StarGAN models for 20k iterations, and CycleGAN models for 200 epochs. The batch size for the I2I model is 1 for CycleGAN-based and MUNIT-based models and 2 for DRIT++ models. In Equation 5 from the main paper, we set α to 0.5 and β to 1.0.

For evaluation, we follow a similar approach as in previous works. For DRIT++ and MUNIT-based models, we first extract the content feature from the input image. To translate an image, we randomly sample the attribute (or style) feature and combine it with the encoded content feature before passing them to the image generator. For computing FID2, output samples from the previous translation step are used as input for the next cycle. For StarGAN and CycleGAN-based models that do not support feature disentanglement, we directly pass the output sample from the previous translation step to the image generator.