RIFT: Disentangled Unsupervised Image Translation via Restricted Information Flow

Ben Usman *
Boston University
usmn@bu.edu

Dina Bashkirova *
Boston University
dbash@bu.edu

Kate Saenko
Boston University
saenko@bu.edu

Abstract

Unsupervised image-to-image translation methods aim to map images from one domain into plausible examples from another domain while preserving the structure shared across two domains. In the many-to-many setting, an additional guidance example from the target domain is used to determine the domain-specific factors of variation of the generated image. In the absence of attribute annotations, methods have to infer which factors of variation are specific to each domain from data during training. In this paper, we show that many state-of-the-art architectures implicitly treat textures and colors as always being domain-specific, and thus fail when they are not. We propose a new method called RIFT that does not rely on such inductive architectural biases and instead infers which attributes are domain-specific vs shared directly from data. As a result, RIFT achieves consistently high cross-domain manipulation accuracy across multiple datasets spanning a wide variety of domain-specific and shared factors of variation.

1. Introduction

The goal of unsupervised image translation is to learn a mapping between two sets of images (two domains) that preserves the shared structure present in both domains without pair supervision. When one domain has *unique factors of variation* absent in the other domain, we must alter the problem definition to ensure that a unique well-defined solution exists: for an input pair consisting of a source image and a target "guide" image, the learned mapping must produce an image from the target domain, preserving all factors of the source image that are varied in both domains, and taking factors of variation specific to the target domain from the guide image. For example, in Fig. 1(a-b) the task is to preserve pose, skin tone, and background of the male source, and apply the hair color of the female guide, whereas in Fig. 1(c-d), preserve object color and shape, and use orientation and size from the guide. This problem is referred as unsupervised *many-to-many* translation [15].

Identifying and preserving shared factors *from data* is of crucial importance in many applications of image translation, such as try-on [21], since many real-world attributes (*e.g.* lighting, shape, roughness) are too difficult to annotate manually [32]. Moreover, in many interpretability and fairness applications [26], we do not know which factors are skewed in advance, and would like to infer that from data.

Unfortunately, recent work [5] on evaluation of many-to-many translation methods suggests that prior state-of-the-art methods fail to infer which attributes are domain-specific and which are domain-invariant *from data* on certain kinds of attribute combinations, and rely on heuristics that work for some dataset pairs, but fail on other. More specifically, many state-of-the-art methods [8, 15] implicitly assume that all domain-specific variations can be modeled as "style" mixed-in globally into intermediate features of image decoders via adaptive instance normalization (AdaIN) [14] originally designed for style transfer. As a result, these methods change *all* colors and textures of the source input to match the guide, even if these colors and textures are varied across both domains and therefore should be preserved. Indeed, Fig. 2a shows that even on a toy dataset pair from Fig. 1(c-d), MUNIT and StarGANv2 change the color of the source object to match color of the guide, even though it is varied in both domains and should be preserved. In Sec. 5, we show that these methods also change backgrounds and skin tones in the female-to-male setup from Fig. 1(a-b) varied in both domains. Methods based on auto-encoders and reconstruction losses [1, 6, 20] preserve shared information better, but often fail to apply correct domain-specific factors. For example, in Fig. 2a, DIDD [6] failed to extract and apply the correct orientation and size from the guide.

In this paper, we propose Restricted Information Flow for Translation **(RIFT)**, a novel approach that does not rely on a fixed inductive bias to perform disentanglement and achieves consistently high attribute manipulation accuracy across different kinds of shared or domain-specific attributes. As illustrated in Fig. 2b, our method preserves *shared factors* (background and pose) of the input male face during "brunet male-to-female translation" and encodes

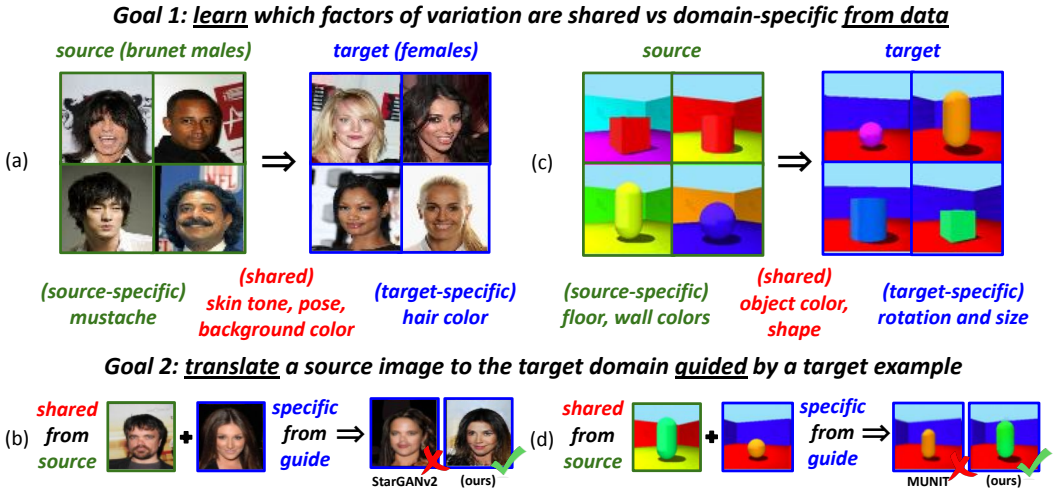
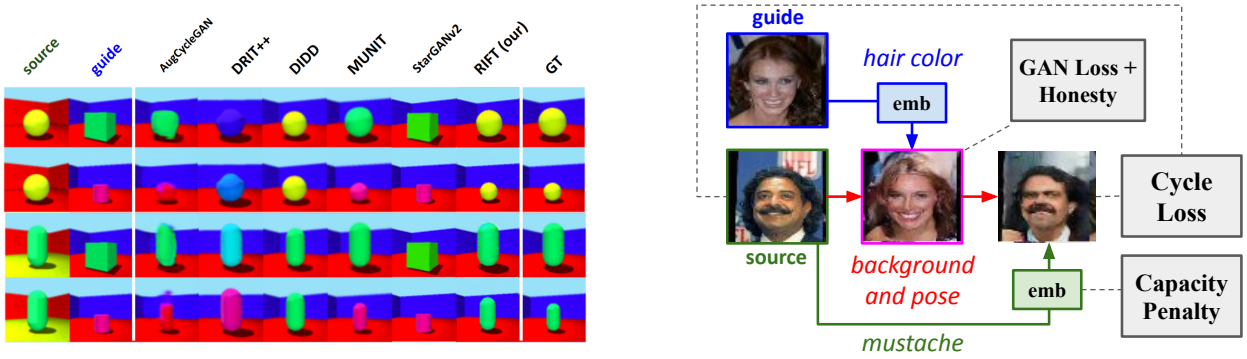


Figure 1: **Problem.** An unsupervised many-to-many image translation model must *disentangle* factors of variation shared across two domains from those specific to each domain using *unpaired* source and target images during training. At the same time, the model has to perform domain translation, preserving factors of the source image shared across two domains and applying target-specific factors from the "guide" image. We show that existing methods fail on at least one of two datasets shown above, and the proposed method excels on both.



(a) **Flaws in existing methods.**

(b) **An overview of RIFT.**

Figure 2: **(a)** All prior methods fail to either preserve shared attributes of the source (shape, object color), or apply target-specific attributes of the guide (size, orientation), while the proposed method (RIFT) succeeds at both (compared to GT), see Fig. 1(c). **(b)** To minimize the cycle-reconstruction loss, RIFT encodes source-specific factors of variation (mustache) into the source-specific embedding, because source-specific factors (mustache) can not be predicted from an source (male) image translated into the target (female) domain.

male-specific attributes (mustache) in a domain-specific embedding. Our framework defines domain-specific factors as those that *can not* be inferred from the image translated into another domain. For example, since female images never contain beards, the model would not be able to infer whether the source image had a beard without a domain-specific embedding, and therefore would fail to minimize the cycle loss. This forces the model to encode domain-specific factors into the domain-specific embedding. Unfortunately, prior work [4, 9] shows that cycle-consistent models tend to "hide" information necessary for accurate cycle-reconstruction in the form of imperceptible low-amplitude adversarial noise embedded into generated images - a so-called "self-adversarial attack". In our case, it manifests as hiding information about mustaches inside generated female images. With this in mind, we propose using the *translation honesty* loss [4] to penalize the model for hiding male-specific information (mustache) inside generated female images. On the other hand, to prevent the model from encoding information shared across two domains (*e.g.* pose, background) into domain-specific embedding, we propose an *embedding capacity* loss that penalizes the model for encoding extra information into domain-specific embedding. As a result, information about the mustache is forced out of the generated female image into the domain-specific embedding, while information about the pose and background is forced out of the domain-specific embedding into the translation result - resulting in correct disentanglement of domain-specific and shared factors. We experimentally verify that the self-adversarial attack takes place, and that the honesty loss prevents it. We also provide a bound over the effective number of bits stored in the domain-specific embedding, and show both theoretically and empirically that, as expected, disabling the capacity loss results in shared fac-

tors erroneously encoded into domain-specific embeddings.

To sum up, we propose a new method for unsupervised many-to-many image translation that does not rely on an inductive bias towards treating certain kinds of attributes as domain-specific or shared. We verify that disabling either component results in all information encoded exclusively either into the domain-specific embedding or the translated image in the form of adversarial noise. Our experiments across three splits of Shapes-3D [18], SynAction [31] and Celeb-A [19] confirm that the resulting model achieves consistently high attribute manipulation accuracy across a wide range of shared and domain-specific attributes.

2. Related work

Image-to-image translation. Unsupervised image-to-image translation methods, such as CycleGAN [34], and UNIT [22], infer semantically meaningful one-to-one cross-domain mappings from pairs of semantically related sets of images (domains) without pair supervision. The problem becomes ill-posed [5] if factors varied in one of two domains are either not present or not varied in the other.

Many-to-many translation. To account for (and enable control over) domain-specific factors, many-to-many image translation methods [1, 8, 15, 20, 23] separate domain-invariant "content" from domain-specific "style". Following Bashkirova et al. [5], we avoid terms "content" and "style" to distinguish the general many-to-many translation problem from its subtask - style transfer [11].

Adaptive instance normalization. Many state-of-art many-to-many translation methods [8, 15], use AdaIN [14], originally designed for style transfer. Some methods [24] add spatial dimension to AdaIN to distinguish colors and textures of different objects, but fundamentally still rely on re-normalization of decoder features to perform disentanglement. While effective at realistic layout-preserving texture transfer (day-to-night, summer-to-winter), this architectural choice was shown [5] to limit the range of applications of these methods to cases in which domain-specific information lies within textures and colors.

Autoencoders. In contrast, methods like Augmented CycleGAN [1], DRIT++ [20] and Domain Intersection and Domain Difference (DIDD) [6] rely on embedding losses and therefore are more general. For example, DIDD forces domain-specific embeddings of opposite domain to be zero, while DRIT++ uses adversarial training to make the source and target content embeddings indistinguishable.

Cycle losses. Most many-to-many methods [1, 15] use cycle-consistency on domain-specific embeddings to ensure that the information from the guide is not ignored, and cycle loss on reconstructed images [34] to improve semantic consistency. However, cycle-consistency on images has been shown [4, 9] to force one-to-one unsupervised translation models to "cheat" by hiding domain-specific information in generated translations in the form of imperceptible low-amplitude structured noise. Alternative consistency objectives, such as the patchwise contrastive loss [29], are designed to be invariant to differences across domains, and therefore can not be used to supervise manipulation of domain-specific factors in the many-to-many case.

Few-shot [23] and **truly unsupervised** [3] translation methods solve a related but different problem. Since these methods have either very few domain examples or no domain labels whatsoever, shared and domain-specific attributes can not be inferred (or even defined) by looking at data. To resolve this ambiguity, these methods also assume that the layout distribution is shared, and that the variability in appearance (*e.g.* colors and textures) is domain-specific.

Single-domain unsupervised disentanglement methods, such as InfoGAN [7] and β-VAE [13], tackle a different problem as well. First, many-to-many translation is not aimed at in controlled manipulation of *individual factors*, but of all domain-specific or all shared factors at once. Second, if we applied these methods to the combined source and target dataset to analyse the distribution of latent codes across each domain, the structure of this dataset would conflict with the independence assumption built into these methods, since distributions of domain specific factors are *not independent* from the distribution of domain labels.

Overall, prior methods ensure that the guide input modulates the translation result in some non-trivial way, but, to our knowledge, no prior work explicitly address adversarial embedding of domain-specific information into the translated image, or quantitatively verifies that domain-specific factors are correctly applied and shared factors are preserved during translation, and this work fills this gap.

3. Restricted Information Flow for Translation

In this section we introduce the many-to-many image translation problem, and describe how our method solves it. Our model reconstructs input images from generated translations and domain-specific embeddings (Fig. 2b), forcing domain-invariant information out of domain-specific embedding using capacity losses, and forcing domain-specific information out from the generated translation using honesty losses, ensuring correct disentanglement.

Setup. Following Huang et al. [15], we assume that we have access to two unpaired image datasets $A = \{a_i\}$ and $B = \{b_i\}$ that share some semantic structure, but differ visually (*e.g.* male and female faces with poses, backgrounds and skin color varied in both). In addition to that, each domain has domain-specific factors of variability, *e.g.* only males have variation in the amount of facial hair and only females have variation in the hair color (Fig. 1). Our goal is to find

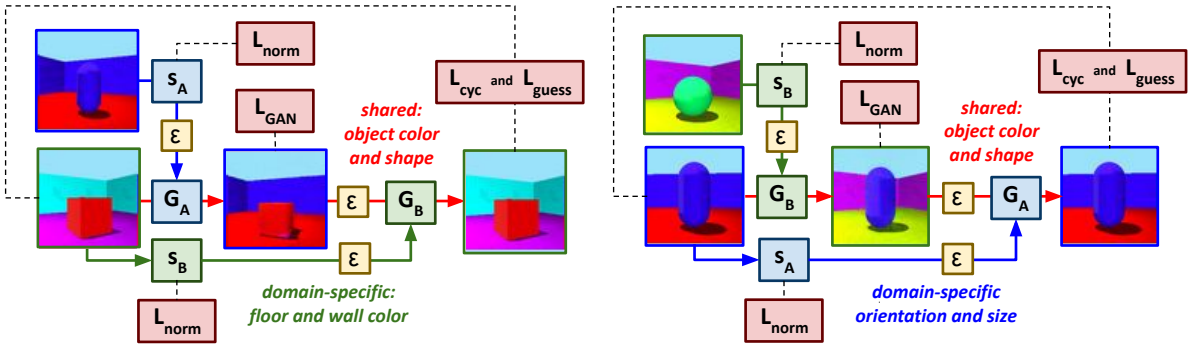


Figure 3: **Losses used to train RIFT.** For illustration, we use 3D-Shapes-A described in Sec. 4 and shown in Fig. 4 and 2a. When the model is trained, green arrows carry only B-specific factors (floor and wall color), blue arrows carry only A-specific factors (orientation and size), and red arrows carry factors shared across two domains (object color and shape).

a pair of guided cross-domain mappings $F_{\text{A2B}} : A, B \to B$ and $F_{\text{B2A}} : B, A \to A$ such that for any source inputs a_s, b_s and guide inputs a_g, b_g from respective domains, resulting guided cross-domain translations $b' = F_{\text{A2B}}(a_s, b_g)$ and $a' = F_{\text{B2A}}(b_s, a_g)$ look like plausible examples of respective output domains, share domain-invariant factors with their "source" arguments (a_s and b_s respectively) and domain-specific attributes with their "guidance" arguments (b_g and a_g respectively). For example, the correct guided female-to-male mapping F_{B2A} applied to female source image b_s and a guide male image a_g should generate a new male image a' with pose, background, and skin tone from the female input image b_s, and facial hair from the guidance input a_g, because poses, backgrounds and skin tone vary in both, while facial hair is male-specific.

Method. While it might be possible to approximate functions F_{A2B} and F_{B2A} directly, following prior work, we split each one into two learnable parts: encoders $s_A(a), s_B(b)$ that extract domain-specific information from corresponding guide images, and generators $G_{\text{A2B}}(a, s_b)$ and $G_{\text{B2A}}(b, s_a)$ that combine that domain-specific information with a corresponding source image, as illusrated in Figure 3. Final mappings are compositions of these networks:

$$F_{\text{A2B}}(a, b) = G_{\text{A2B}}(a, s_B(b)),\ F_{\text{B2A}}(b, a) = G_{\text{B2A}}(b, s_A(a))$$

Losses introduced in the remainder of this section ensure that encoders s_* extract domain-specific information from their inputs (and nothing else), and that generators G_* use the encoder outputs, (only) domain-invariant factors from their source inputs, and generate plausible images.

Noisy cycle-consistency loss. To ensure that each factor of input images is not ignored completely, we use a guided analog of the cycle consistency loss [34]. This loss ensures that any image translated into a different domain, and translated back with its original domain-specific embedding is reconstructed perfectly. Additionally, before translating images back into their original domains, we add zero-mean Gaussian noise (ε_s, ε_g) of variance σ_s and σ_g to each dimension of generated images and domain-specific embeddings - the motivation is given in two following paragraphs.

$$L^A_{\text{cyc}} = \mathbb{E}_{a,b}\,||a_{\text{cyc}} - a||_1,\ \ L^B_{\text{cyc}} = \mathbb{E}_{b,a}\,||b_{\text{cyc}} - b||_1 \quad (1)$$

$$a_{\text{cyc}} = G_{\text{B2A}}(G_{\text{A2B}}(a, s_B(b) + \varepsilon_g) + \varepsilon_s, s_A(a) + \varepsilon_g) \quad (2)$$

$$b_{\text{cyc}} = G_{\text{A2B}}(G_{\text{B2A}}(b, s_A(a) + \varepsilon_g) + \varepsilon_s, s_B(b) + \varepsilon_g) \quad (3)$$

$$a \sim A,\ b \sim B,\ \varepsilon_s \sim \mathcal{N}(0, \sigma_s),\ \varepsilon_g \sim \mathcal{N}(0, \sigma_g) \quad (4)$$

Translation honesty. Unfortunately, any form of cycle loss encourages the model to "hide" domain-specific information inside the translated image in the form of structured adversarial noise [9]. To prevent the model from "hiding" the domain-specific information, such as mustache, inside a generated female image (instead of putting it into a male-specific embedding s_a), we use two so-called "self-adversarial defences" proposed by Bashkirova et al. [4]. First, we *destroy* the structured signal by adding Gaussian noise ε_s to intermediate images before cycle reconstruction, see Eq. (2) above. Moreover, we use an additional *guess loss* to train the generator. To compute it, we train a pair of *guess discriminators* that predict which of its two inputs is a cycle-reconstruction and which is the original image. For example, if the male-to-female generator G_{A2B} is consistently adversarially embedding mustaches into all generated female images, then the cycle-reconstructed female b_{cyc} will also have traces of an embedded mustache, because it was generated using that male-to-female generator G_{A2B}, and will be otherwise identical to the input b. In this case, the guess discriminator D^{gs}_B, trained specifically to detect differences between input images and their cycle-reconstructions, will detect this hidden signal and penalize the model:

$$L^A_{\text{guess}} = [D^{\text{gs}}_A(a, a_{\text{cyc}})]^2 + [1 - D^{\text{gs}}_A(a_{\text{cyc}}, a)]^2 \quad (5)$$

$$L^B_{\text{guess}} = [D^{\text{gs}}_B(b, b_{\text{cyc}})]^2 + [1 - D^{\text{gs}}_B(b_{\text{cyc}}, b)]^2 \quad (6)$$

Domain-specific channel capacity. Unfortunately, neither of two losses described above can prevent the model

from learning to embed the entire guide image a_g into the domain-specific embeddings s_a and reconstructing it from that embedding, ignoring its first argument, *i.e.* always produce the guide input exactly. In order to prevent this from happening, we add Gaussian noise ε_g to predicted domain-specific embeddings before cycle reconstruction (see Eq. 2 above) and penalize norms of these embeddings:

$$L^A_{\text{norm}} = \mathbb{E}_a \, ||s_A(a)||_2^2, \;\; L^B_{\text{norm}} = \mathbb{E}_b \, ||s_B(b)||_2^2 \quad (7)$$

Theorem 1 in supplementary shows that this procedure constrains the *effective capacity* of domain-specific embeddings. Intuitively, the mutual information between the input guide image a_g and the predicted translation a' corresponds to the maximal amount of information that an observer could learn about translations a' by observing guides a_g . Formally, we can show that if we add Gaussian noise of amplitude σ_g and penalize the norms of embeddings $s_A(a_g)$ as described above, this mutual information is bounded by:

$$\text{MI}(a_g; a') \leq \dim(s_A(a)) \cdot \log_2 \left(1 + L^A_{\text{norm}}/\sigma_g^2\right), \quad (8)$$

$$a' = G_{\text{B2A}}(b_s, s_A(a_g) + \varepsilon_g), \; \varepsilon_g \sim \mathcal{N}(0, \sigma_g) \quad (9)$$

meaning that minimizing L^A_{norm} loss effectively limits the amount of information from the guide image a_g that G_{A2B} can access to generate a', *i.e.* the effective capacity of the domain-specific embedding. Note that disabling either the noise ($\sigma_g = 0$) or the capacity loss ($L_{\text{norm}} \to \infty$) theoretically results in effectively *infinite* capacity, so we need both. Intuitively, this bound describes the expected number of "reliably distinguishable" embeddings that we can pack into a ball of radius $\sqrt{L^A_{\text{norm}}}$ assuming that each embedding is perturbed randomly by Gaussian noise with amplitude σ_g.

Realism losses. Remaining losses are analogous to the original GAN and identity losses from CycleGAN [22] ensuring that generated images lie within respective domains:

$$L^A_{\text{GAN}} = [D_A(a)]^2 + [1 - D_A(G_{\text{B2A}}(b, s_A(a) + \varepsilon^a_s))]^2$$

$$L^B_{\text{GAN}} = [D_B(b)]^2 + \left[1 - D_B(G_{\text{A2B}}(a, s_B(b) + \varepsilon^b_s))\right]^2$$

$$L^A_{\text{idt}} = \mathbb{E}_a \, ||G_{\text{B2A}}(a, s_A(a) + \varepsilon_g) - a||_1,$$

$$L^B_{\text{idt}} = \mathbb{E}_b \, ||G_{\text{A2B}}(b, s_B(b) + \varepsilon_g) - b||_1$$

Discriminator losses We also train discriminators D_A, D_B and guess discriminators $D^{\text{gs}}_A, D^{\text{gs}}_B$ by minimizing corresponding adversarial LS-GAN [25] losses.

4. Experiments

We would like to measure how well each model can generalize across a diverse set of shared and domain-specific attributes. In this section, we discuss datasets we used and generated to achieve this goal, as well as baselines and metrics we used to compare our method to prior work.

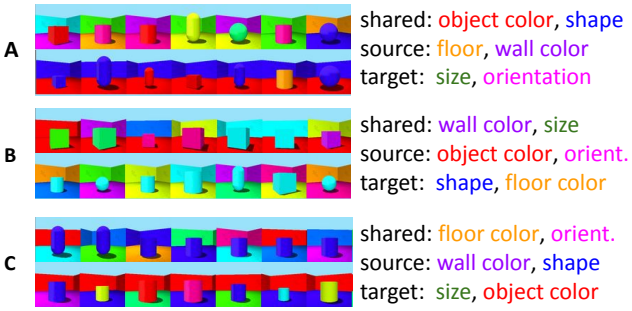


Figure 4: **Shapes-3D-ABC**: splits, shared and specific factors.

Data. Popular image translation datasets (*e.g.* summer-to-winter [22], GTA5-to-BDD, AFHQ [17]) lack attribute annotations, precluding quantitative evaluation, and focus exclusively on layout-preserving texture/palette transfer. To evaluate methods' ability to disentangle and transfer other kinds of attributes, following the protocol proposed by Bashkirova et al. [5], we re-purposed existing disentanglement datasets to evaluate the ability of our method to model different attributes as shared and domain-specific. We used 3D-Shapes [18], SynAction [31] and CelebA [19]. Unfortunately, among the three, only 3D-Shapes [18] is balanced enough and contains enough labeled attributes to make it possible to generate and evaluate all methods across several attribute splits of comparable sizes. For example, if we were to build a split of SynAction with domain-specific pose, the domain with fixed pose would contain only 90 images.

3D-Shapes-ABC. The original 3D-Shapes [18] dataset contains 40k synthetic images labeled with six attributes: floor, wall and object colors, object shape and object size, and orientation (viewpoint). There are ten possible values for each color attribute, four possible values for the shape (cyliner, capsule, box, sphere), fifteen values for orientation, and eight values for size. We used three subsets of 3D-Shapes with different attribute splits visualized in Figure 4. Three resulting domain pairs contained 4.8k/4k, 12k/3.2k, and 12k/6k images respectively.

SynAction. We used the same [5] split of SynAction [31] - with background varied in one domain (nine possible values), identity/clothing varied in the other (ten possible values), and pose varied in both (real-valued vector). The resulting dataset contains 5k images in one domain and 4.6k images in the other. We note that the attribute split of this dataset **matches** the inductive bias of AdaIN methods, since the layout (pose) is shared and textures (background, clothing) are domain-specific in both domains. We noticed that the original "fixed bg" domain [5] actually has some variation in the background, and fixed them before training both our method and all baselines (see supplementary Sec. 7.2).

CelebA-FM. We used the male-vs-female split proposed by Bashkirova et al. [5] with 25k images in each domain,

and evaluated disentanglement of six most visually prominent attributes: pose, skin and background color (shared attributes, real-valued vectors), male-specific presence of facial hair (binary), female-specific hair color (three possible values), and domain-defining gender.

Baselines. We compare the proposed method against several state-of-art AdaIN methods, namely MUNIT [15], StarGANv2 [8], MUNITX [5], and autoencoder-based methods, namely Domain Intersection and Domain Difference (DIDD) [6], Augmented CycleGAN [1] and DRIT++ [20]. We did not evaluate other AdaIN-based methods, such as EGSC-IT [24], since have all share the disentanglement strategy. We did not evaluate truly unsupervised methods [3] and other methods that *explicitly* preserve the layout and transfer the appearance [33] because they approach a *different problem*, as discussed in Sec. 2 and Sec. 7.3. For reference, we provide a random baseline (RAND) that corresponds to returning a random image from the target domain.

Metrics. In order to evaluate the performance of our method, we measured how well the domain-specific attributes were manipulated and domain-invariant attributes were preserved. Following Bashkirova et al. [5] we trained an attribute classifier $f(x)$, and for each attribute k, we measured the its *manipulation accuracy* - the probability of correctly modifying an attribute across input-guide pairs for which the value of the attribute *must change*:

$$\text{ACC}_k^{\text{A}} = p(f_k(F_{\text{A2B}}(a,b)) = y_k^* \mid f_k(a) \neq f_k(b))$$

where the "correct" attribute value equals $y_k^* = f_k(a)$ for shared attributes, and $y_k^* = f_k(b)$ otherwise. For real-valued multi-variate attributes (pose keypoints, background RGB, skin RGB, etc.) we measured the probability of generating an image with the attribute closer to the correct attribute vector y_k^* then to y_k' from the opposite domain:

$$\text{ACC}_k^{\text{A}} = p(\|f_k(F_{\text{A2B}}(a,b)) - y_k^*\| \leq \|f_k(F_{\text{A2B}}(a,b)) - y_k'\|)$$

where $y_k^* = f_k(a)$ and $y_k' = f_k(b)$ for shared attributes, and vice-versa otherwise. The manipulation accuracy in the opposite direction ACC_k^{B} was estimated analogously. For *Shapes-3D* we can report *aggregated* domain-specific and domain-invariant manipulation accuracies $\text{ACC}_k^S(s)$ and $\text{ACC}_k^C(s)$ averaged (see Sec. 7.4) across splits in which the given attribute k was shared/common (C) or domain-specific (S), and the *relative discrepancy* between them:

$$\text{RD} = 100 \cdot \frac{\sum_k |\text{ACC}_k^S - \text{ACC}_k^C|}{\sum_k (\text{ACC}_k^S + \text{ACC}_k^C)}. \qquad (10)$$

In this work, we interested in improving *not* the realism of generated images, but the disentanglement quality. Nevertheless, in supplementary Sec. 7.5 we report FID and LPIPS of compared methods, and show that our method is on par with them. More detailed description of the evaluation protocol and the architecture are given in the supp. Sec. 7.6.

Method	3DS	SA	CA	**AVG**	**RD**
StarGANv2	45	**82**	51	<u>59</u>	97
MUNIT	<u>58</u>	37	53	49	56
MUNITX	33	52	55	47	74
DRIT++	18	24	55	32	<u>20</u>
AugCycleGAN	12	37	40	29	<u>20</u>
DIDD	44	67	**64**	58	35
RIFT (ours)	**88**	<u>78</u>	<u>60</u>	**75**	**6**
RAND	12	24	49	27	9

Table 1: **Average (AVG↑) manipulation accuracy (ACC) and relative discrepancy (RD↓)** across 3D-Shapes-ABC (3DS), SynAction (SA), and CelebA-FM (CA). Notation: **best**, <u>2nd best</u>.

5. Results

In this section, we first compare our method to prior work both qualitatively and quantitatively. Then we show what happens if we remove key losses discussed Section 3. And finally, we discuss implicit assumptions made by our method, and propose several key challenges that future methods will need to address to further improve manipulation accuracy across three datasets we used in this paper.

Quantitative results. Tables 1 and 2 show that across three splits of 3D-Shapes-ABC our method achieves the highest average manipulation accuracy and the lowest relative discrepancy between accuracies of modeling same attributes as shared and specific. On SynAction, that *matches* the inductive bias of AdaIN-based methods, our method performs on-par with the AdaIN-based StarGANv2 and outperforms all non-AdaIN methods. On CelebA-FM, our method performs on par with DIDD up to a small margin and outperforms other methods. Overall, RIFT achieves best or second-best (with a small margin) performance in *each* of three dataset, whereas both runner-ups (DIDD and StarGANv2) perform poorly on at least one of three dataset (DIDD on SynAction, StarGANv2 on CelebA, both on 3D-Shapes). RIFT also achieves best average accuracy (AVG) across three datasets, and lowest relative discrepancy (RD) on 3D-Shapes.

Qualitative results. Figures 5, 7 and 2a show that, in the absolute majority of cases, the proposed method successfully preserves domain-invariant and uses domain-specific information from respective domains on 3D-Shapes and SynAction, and does so much better than all other baselines, which agrees with the quantitative evaluation above. Figure 8 shows that, on CelebA, our method preserves poses and backgrounds, and applies hair color better then other baselines. We provide a more detailed side-by-side qualitative comparison of generated images across all baselines and all datasets in the supplementary. In suppl. Fig 10 we also show how RIFT can change domain-specific factors of images while keeping them within their original domain.

Method	3D-Shapes-ABC												SynAct			CelebA-FM					
	FC		WC		OC		SZ		SH		ORI		PS	IDT	BG	HC	FH	GD	ORI	BG	SC
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	S	S	S	S	C	C	C
StarGANv2	0	99	0	99	0	78	5	56	4	99	0	**96**	**96**	**52**	**99**	**76**	15	97	87	11	22
MUNIT	5	94	0	**99**	0	97	**59**	31	96	58	**99**	61	75	28	7	45	7	90	**89**	43	44
MUNITX	1	50	2	55	8	28	12	16	95	21	**99**	7	93	26	37	64	17	75	83	50	43
DRIT++	7	12	9	19	10	10	27	14	7	15	42	51	52	6	13	23	9	96	**89**	67	44
AugCycleGAN	10	8	10	9	11	7	17	13	30	13	7	7	90	8	12	16	30	98	12	42	40
DIDD	38	81	29	22	72	18	41	20	87	43	48	34	89	12	**99**	22	**50**	91	78	**89**	56
RIFT (ours)	**100**	**100**	**100**	**100**	**100**	**100**	5	**60**	**98**	**100**	97	**96**	89	47	**99**	22	35	**99**	65	83	**57**
RAND	10	10	10	10	10	10	12	19	24	19	6	6	50	11	11	12	31	99	50	50	50

Table 2: **Manipulation accuracy** for shared/common (C) or domain-specific (S) attributes aggregated across **Shapes-3D-ABC**: floor color (FC), wall color (WC), object color (OC), size (SZ), shape (SH), room orientation (ORI); **SynAction**: pose (PS), identity/clothing (IDT), background (BG); **CelebA-FM**: hair color (HC), facial hair (FH), gender (GD), face orientation (ORI), bg (BG) and skin (SC) color.

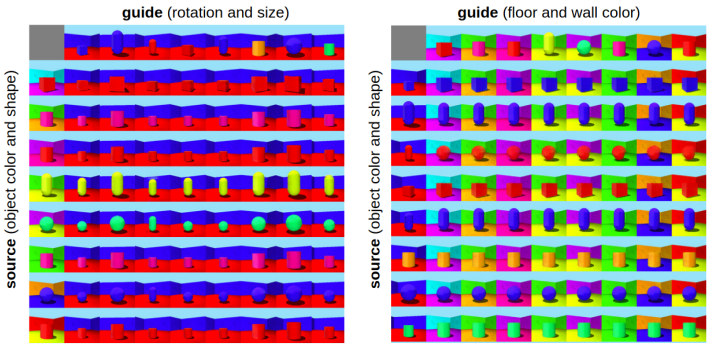
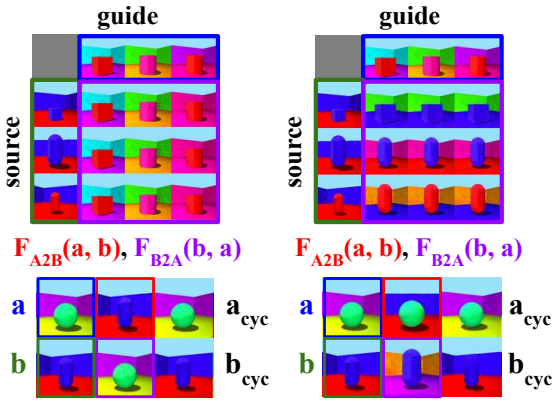


Figure 5: **Guided translations by RIFT on 3D-Shapes-A.** Our model successfully preserves shared attributes (object color and shape) of the source image and applies domain-specific attributes from the guide image (rotation and size on the left, floor and wall color on the right). Comparison to prior work can be found in Fig. 2a and in supplementary.



(a) w\o **capacity** penalty (b) w\o **honesty** loss

Figure 6: **Ablations.** Effects of disabling capacity and honesty losses on guided translations (top) and guided cycle-reconstructions (bottom) on Shapes-3D-A. Inputs images from domains **A** and **B**, **A2B** and **B2A** guided translations.

Ablations. During B2A translation on Shapes-3D-A the model trained with all losses correctly uses object color/shape from the source image and floor/wall color from the guide (Fig. 5). If we remove the penalty on the capacity of domain-specific embeddings (L_{norm}), the model ignores the source input (Fig. 6a-top): it encodes all attributes into domain-specific embeddings, and cycle-reconstructs inputs a and b perfectly from these embeddings (Fig. 6a-bottom), completely ignoring the source input: $\boldsymbol{b} = \boldsymbol{F}_{\text{A2B}}(\boldsymbol{a}, \boldsymbol{b}) = b_{\text{cyc}}$. Removing honesty losses (L_{guess}), on the other hand, results in a model that ignores the guide input altogether (Fig. 6b-top). The model "hides" domain-specific information inside generated translations instead of the domain-specific embeddings, and makes domain-specific embeddings equal zero, resulting in zero capacity loss $L_{\text{norm}} = 0$, and zero cycle reconstruction loss $L_{\text{cyc}} = 0$. For example (Fig. 6b-bottom), the size and orientation of $\boldsymbol{b}$ is hidden inside $\boldsymbol{F}_{\text{B2A}}(\boldsymbol{b}, \boldsymbol{a})$ in the form of imperceptible adversarial noise and is used to reconstruct b_{cyc} perfectly. If mapping F_{A2B} actually used size and orientation of b to generate b_{cyc}, it would have also applied that same size and orientation when generating $\boldsymbol{F}_{\text{A2B}}(\boldsymbol{a}, \boldsymbol{b})$, but it did not - so we conclude that both F_{A2B} and F_{B2A} ignore domain-specific embeddings and embed information inside generated translations - see more ablation visualizations in suppl. Fig. 12. In the supplementary Sec. 7.8 we also show that the model trained with all proposed losses does *not* hide information inside generated images: we trained a separate classification network to predict attributes of the inputs that *should have been lost* during translation from translated images. The resulting classifier was able to accurately predict hidden information from images generated *without* honesty losses, and was unable to predict them above chance from images generated by a model trained *with* honesty loss (suppl. Tab. 5). This confirms that shared attributes of the guide and domain-specific attributes of the source were indeed correctly ignored by the generator trained with proposed losses.

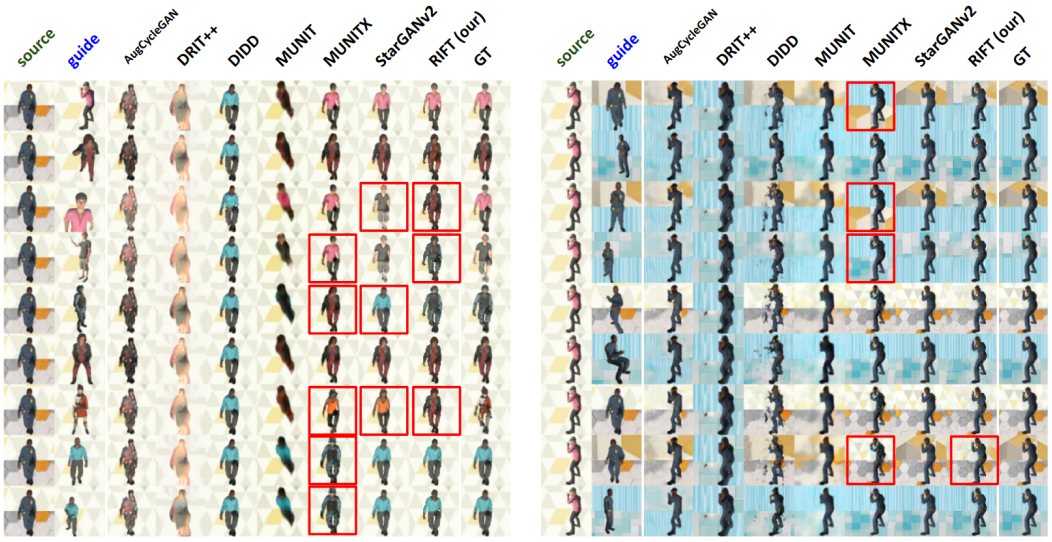


Figure 7: **Qualitative comparison to prior work on SynAction.** Our model correctly preserves shared attributes (pose) of the source image and applies domain-specific attributes of the guide domain (clothing/identity colors on the left, background texture on the right) - compare to Ground Truth (GT). Errors made by top performing methods are highlighted in red.

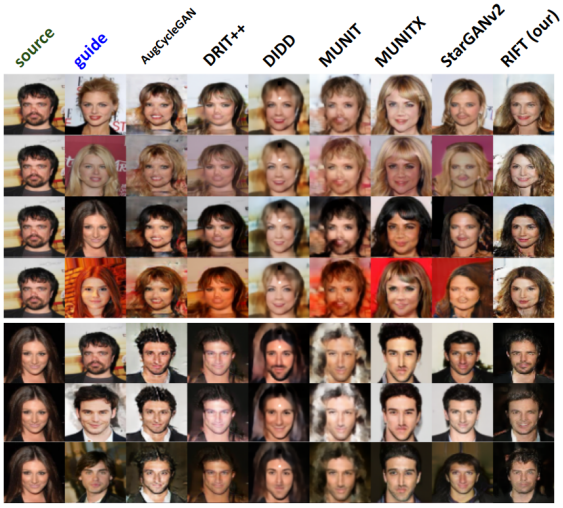


Figure 8: **Qualitative** comparison to prior work on CelebA-FM. Methods should *preserve the pose and the background* of the source, and apply *only the hair color* of the female guide during male2fem translation (top) and *only the facial hair* of the male guide during fem2male translation (bottom). Only RIFT and DIDD preserved background colors *and* applied correct target-specific hair colors and mustaches.

Challenges. We suggest two major causes of remaining errors that existing methods fail to handle at the moment, and future researchers will need to address to make further progress in this task possible. First, some attributes "affect" very different number of pixels in training images, and as a consequence contribute very differently to reconstruction losses, making the job of balancing different loss components much harder. For example, the floor color in 3D-Shapes "affects" roughly half of all image pixels, whereas size affects only one tenth of all pixels - resulting in drastically different effective weights across all losses, especially if both are either domain-specific or shared at the same time. Second, unevenly distributed shared attributes in real world in-the-wild datasets (such as CelebA) pose an even more serious challenge, rendering the many-to-many problem task *not well defined*. For example, if both male and female domains had hair color variation, but males were mostly brunet with only 3% of blondes, but females were equally likely to be blondes and brunettes - should the model preserve blonde hair when translating females to males and sacrifice the "realism" of the generated male domain, or should it treat hair-color as a domain-specific attribute despite variations present in both? This poses an open question. We also discuss the ethical aspects of unsupervised image translation in supplementary Section 7.9.

6. Conclusion

In this paper we propose RIFT - a new unsupervised many-to-many image-to-image translation method that determines which factors of variation are shared and which are domain-specific *from data*, and achieves consistently high attribute manipulation accuracy across a wide range of datasets with different kinds of domain-specific and shared attributes, and low discrepancy between these accuracies. We provide ablations confirming that the self-adversarial embedding takes place in the many-to-many setting, that the honesty loss prevents it from happening. We also show that the capacity loss restricts the effective capacity of the domain-specific embedding in agreement with the provided theoretical bound. Finally, we identify core challenges that need to be resolved to enable further development of unsupervised many-to-many image-to-image translation.

References

[1] Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented CycleGAN: Learning many-to-many mappings from unpaired data. arXiv preprint arXiv:1802.10151 (2018)

[2] Augenstein, S., McMahan, H.B., Ramage, D., Ramaswamy, S., Kairouz, P., Chen, M., Mathews, R., et al.: Generative models for effective ML on private, decentralized datasets. arXiv preprint arXiv:1911.06679 (2019)

[3] 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 (ICCV), pp. 14154–14163 (October 2021)

[4] Bashkirova, D., Usman, B., Saenko, K.: Adversarial self-defense for cycle-consistent GANs. in proceedings of the Thirty Second Conference on Advances in Neural Information Processing Systems (2019)

[5] Bashkirova, D., Usman, B., Saenko, K.: Evaluation of correctness in unsupervised many-to-many image translation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)

[6] Benaim, S., Khaitov, M., Galanti, T., Wolf, L.: Domain intersection and domain difference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3445–3453 (2019)

[7] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems (2016)

[8] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)

[9] Chu, C., Zhmoginov, A., Sandler, M.: CycleGAN, a master of steganography. arXiv preprint arXiv:1712.02950 (2017)

[10] Citron, D.K., Chesney, R.: Disinformation on steroids: The threat of deep fakes. Cyber Brief (2018)

[11] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)

[12] Grover, A., Song, J., Agarwal, A., Tran, K., Kapoor, A., Horvitz, E., Ermon, S.: Bias correction of learned generative models using likelihood-free importance weighting. arXiv preprint arXiv:1906.09531 (2019)

[13] Higgins, I., Matthey, L., Pal, A., Burgess, C.P., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)

[14] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)

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

[16] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)

[17] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4401–4410 (2019)

[18] Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning, pp. 2649–2658, PMLR (2018)

[19] Lee, C.H., Liu, Z., Wu, L., Luo, P.: MaskGAN: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[20] Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M.K., Yang, M.H.: DRIT++: Diverse image-to-image translation via disentangled representations. arXiv preprint arXiv:1905.01270 (2019)

[21] Lewis, K.M., Varadharajan, S., Kemelmacher-Shlizerman, I.: TryOnGAN: Body-aware try-on via layered interpolation. ACM Transactions on Graphics (TOG) **40**(4), 1–10 (2021)

[22] Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. arXiv preprint arXiv:1703.00848 (2017)

[23] Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz., J.: Few-shot unsueprvised image-to-image translation. In: arxiv (2019)

[24] Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. arXiv preprint arXiv:1805.11145 (2018)

[25] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2794–2802 (2017)

[26] McDuff, D., Ma, S., Song, Y., Kapoor, A.: Characterizing bias in classifiers using generative models. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc. (2019)

[27] Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection: A survey. arXiv preprint arXiv:1909.11573 (2019)

[28] Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286 (2018)

[29] Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision (2020)

[30] Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)

[31] Sun, X., Xu, H., Saenko, K.: TwoStreamVAN: Improving motion modeling in video generation. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 2744–2753 (2020)

[32] Usman, B., Dufour, N., Saenko, K., Bregler, C.: PuppetGAN: Cross-domain image manipulation by demonstration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9450–9458 (2019)

[33] Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: TransGaGa: Geometry-aware unsupervised image-to-image translation. In: CVPR (June 2019)

[34] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)