

Evaluation of Correctness in Unsupervised Many-to-Many Image Translation

Dina Bashkirova
Boston University
dbash@bu.edu

Ben Usman
Boston University
usmn@bu.edu

Kate Saenko
Boston University and MIT-IBM Watson AI Lab
saenko@bu.edu

Abstract

Given an input image from a source domain and a “guidance” image from a target domain, unsupervised many-to-many image-to-image (UMMI2I) translation methods seek to generate a plausible example from the target domain that preserves domain-invariant information of the input source image and inherits the domain-specific information from the guidance image. For example, when translating female faces to male faces, the generated male face should have the same expression, pose and hair color as the input female image, and the same facial hairstyle and other male-specific attributes as the guidance male image. Current state-of-the-art UMMI2I methods generate visually pleasing images, but, since for most pairs of real datasets we do not know which attributes are domain-specific and which are domain-invariant, the semantic correctness of existing approaches has not been quantitatively evaluated yet. In this paper, we propose a set of benchmarks and metrics for the evaluation of semantic correctness of UMMI2I methods. We provide an extensive study how well the existing state-of-the-art UMMI2I translation methods preserve domain-invariant and manipulate domain-specific attributes, and discuss the trade-offs shared by all methods, as well as how different architectural choices affect various aspects of semantic correctness.

1. Introduction

Unsupervised image-to-image translation aims to map an input image from a source domain to a target domain so that the output looks like a valid example of the target domain while preserving the semantics of the input image, given only two sets of source and target images and no supervision. First attempts at the unsupervised image-to-image translation [40, 25, 18, 36, 12] assumed one-to-one relation between the domains, meaning that for each image in the source domain there exists exactly one corresponding image in the target domain. Unfortunately, this assumption does not hold for most applications, as in most cases there are multiple correct solutions in the target domain corre-

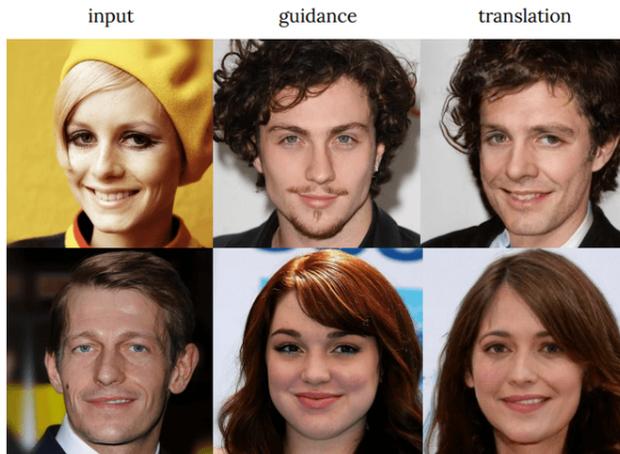


Figure 1. Given a pair of input image from source domain and a guidance image from target domain (e.g. male and female photos), unsupervised many-to-many image translation (UMMI2I) methods aim to map them to target domain so that the factors varied in both domains (such as pose, facial expression, hair and skin color, etc.) are taken from the input image, while the factors that are varied only in the target domain (such as facial hairstyle) are taken from the guidance image. Current state-of-the-art UMMI2I method StarGANv2 [4] on the CelebAHQ [22] dataset, however, falsely considers hair color a domain-specific variation and ignores the domain-specific information about facial hair from the guidance image in the top row. We introduce a set of metrics for evaluation of semantic correctness in UMMI2I translation and provide an analysis of current state-of-the-art methods.

sponding to each example from the source domain. This makes the problem ill-posed and results in unreliable translation and unpredictable behaviour on the new unseen examples from the source domain.

One natural solution to this problem is to consider many-to-many mappings [14, 28, 2, 23], in which case the translation is guided by an example from the target domain that specifies domain-specific factors of the generated target image. For example, when an image of a female face is translated to the male domain, it is not clear whether the translation result should have a beard or not. In the one-to-one setting, the model would either chose the most likely fa-

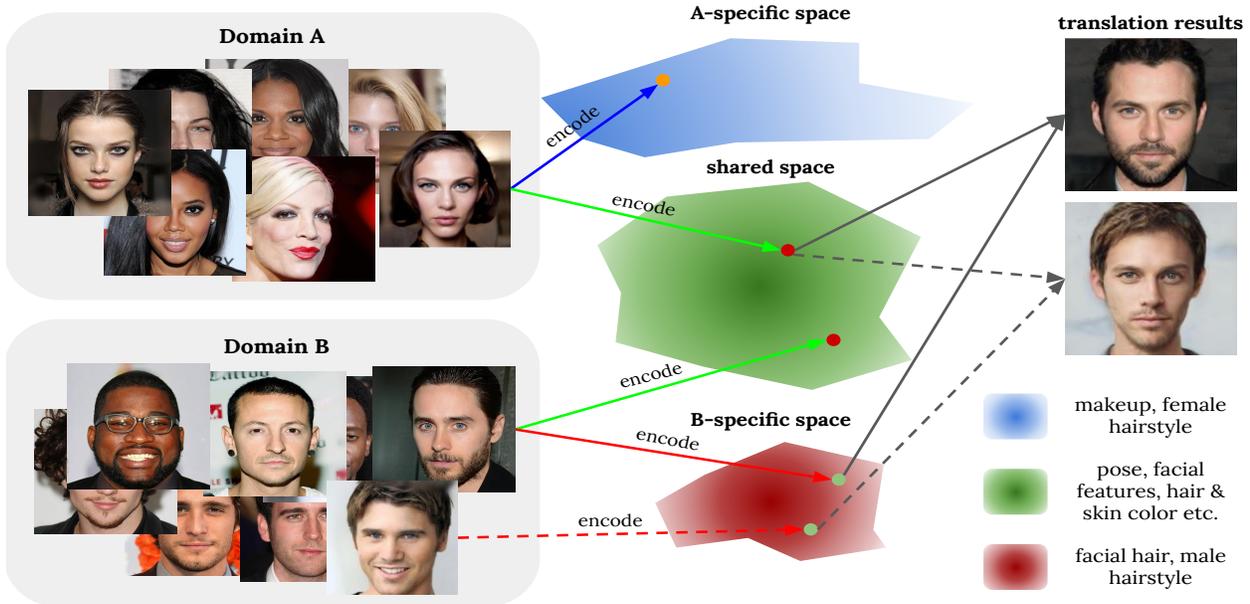


Figure 2. Overview of the common approach to disentanglement in the UMMI2I translation. Given an input image from the source domain and a guidance example from the target domain, the UMMI2I method returns an image from target domain containing content information from the input image and domain-specific attributes (e.g. facial hair) from the guidance example. In order to allow control over domain-specific attributes of the translated images while preserving the content information of the input image, a common approach is to encode images to the domain-specific and domain-invariant embedding spaces. All source examples were taken from CelebA [22] dataset; the translation examples are generated with StarGANv2 [4] (*best viewed in color*).

cial hair option or rely on some unrelated property of the input image (e.g. the value of the top-left pixel) to make this decision. In the many-to-many setting, however, the information about the desired facial hairstyle is taken from a guidance image from the target domain. The resulting translation problem for a given pair of input and guidance images has a single correct solution, and therefore lets us reason about the correctness of translation models.

Measuring which architecture choices yield semantically correct UMMI2I translations, *i.e.* provide better disentanglement of domain-specific and domain-invariant factors of variation, is crucial, yet the majority of current state-of-the-art UMMI2I translation methods are evaluated using only Fréchet inception distance (FID) [10], learned perceptual similarity (LPIPS) [39] and other metrics that measure diversity and realism, but do not take the disentanglement quality into consideration. Unfortunately, existing disentanglement metrics proposed in the representation learning literature [16, 11, 17, 7, 3, 21, 6] are not suitable for evaluation of UMMI2I translation since they study the quality of disentanglement of latent embeddings, not images, and were designed for single-domain disentanglement, therefore not taking into account which variations are shared and which are domain-specific.

In this paper, we propose a new, data-driven approach for evaluation of unsupervised cross-domain disentanglement quality in UMMI2I methods. We designed three evaluation protocols based on the synthetic 3D Shapes [17] dataset (originally designed for evaluation of single-domain disen-

tanglement), a more challenging synthetic SynAction [33] pose dataset, and a widely used CelebA [27] dataset of faces, for all of which some form of ground truth annotation is available. Our contributions can be summarized as follows:

1. To the best of our knowledge, we are the first to propose a set of metrics for evaluation of the semantic correctness of UMMI2I translation. Our metrics evaluate how well the shared attributes are preserved, how reliably the domain-specific attributes are manipulated, whether the translation result is a valid example of the target domain and whether the attribute mode-collapse had happened.
2. We create three evaluation protocols based on 3D-Shapes, SynAction and CelebA datasets, and measure disentanglement quality of the current state-of-the-art UMMI2I translation methods on them.
3. We show that 1) there is a clear trade-off between the preservation of content information and manipulation of the domain-specific variations; 2) All methods we tested manipulate very poorly the attributes associated with adding or changing certain parts of the objects, e.g. facial hair or smile; 3) AdaIN-based [13] methods showed an inductive bias towards treating spatial attributes, such as poses and position of objects in the scene, as the domain-invariant factors, and colors and textures as the domain-specific sources of variation, irrespective of which attributes were actually shared and domain-specific.

2. Related Work

2.1. Many-to-Many Image Translation

Style transfer. Perhaps one of the first steps towards producing diverse translation results using deep neural networks was the development of neural style transfer methods [8, 34, 9, 35, 15, 38, 20]. Conventionally, these methods perform texture transfer by matching the statistics of the input image output features at particular layers of the pre-trained CNN with those of the guidance image. While style transfer methods show remarkable results for generating artistically stylized images, they are not applicable for domain adaptation since they only change the texture of a given image which is often not sufficient to minimize domain shift and does not guarantee content preservation.

Latent embedding sharing and reconstruction-based methods. Another line of work used reconstruction losses of the latent embeddings for disentanglement of domain-specific and shared information. For example, Almahairi et al. [2] proposed an augmented version of CycleGAN [40] with two additional domain-specific encoders. The disentanglement is enforced by the reconstruction loss of the latent representations of a pair of source and target examples from the corresponding translation results. The disentanglement strategy introduced in DRIT[23], DRIT++[24] and DMIT[37], involves encoding of images from both domains to the shared latent space (content space) in addition to the domain-specific encoding. Additionally, those methods share part of the weights of the content encoders and have an additional content discriminator to force the disentanglement of the shared variations.

Latent embedding reconstruction and AdaIN-based manipulation. Other UMMI2I translation methods, such as MUNIT[14], FUNIT[26] for few-shot translation, EGSCIT[28] and StarGANv2 [4], used the Adaptive Instance Normalization (AdaIN) [13] to manipulate the translation results using a guidance example. The AdaIN was originally used for style transfer, thus those methods share an assumption that the domain-specific factors of variation lie within texture and color variation. In addition to the AdaIN-based manipulation, these methods use separate encoders for the domain-specific information and fully or partially shared encoders for the domain-invariant information and reconstruction losses for the latent representations.

2.2. Translation and disentanglement metrics

Translation metrics. The metrics commonly used for the evaluation of the UMMI2I methods, include Inception score [31] and Frechet Inception score [10] that measure the similarity of the distributions of the translation results and the target domain; learned perceptual similarity score [31] that measures the visual diversity of the translation results; and human evaluation where independent participants

are asked to evaluate the visual quality and overall semantic consistency of the results. While the latter is the closest option for the evaluation of semantic consistency and reliability of UMMI2I translation, 1) this approach is expensive and 2) for the majority of the datasets used for translation, it is unclear what the semantically correct translation is since there is no ground truth information of which sources of variation are shared and which ones are domain-specific; therefore it is difficult - if not impossible - to instruct the participants to identify the semantically incorrect translation.

Disentanglement Metrics. Metrics that measure the quality of disentanglement have been extensively explored in the recent literature on representation learning [16, 11, 17, 7, 3, 21, 6]. However, previous methods developed for representation learning are not suitable for evaluating the image-to-image translation disentanglement that combines representation learning with a challenging task of pixel-level domain adaptation. Perhaps the closest to our proposed idea is the work by [17, 7] and [11] that use toy data with known generative factors along with the learned factor regressor to infer the target factors from the latent embedding and comparing these predictions to the expected values. Directly inferring the generative factors from the latent representations, however, can be insufficient for evaluation of disentanglement in the UMMI2I task since the disentanglement in the latent space does not necessarily imply that the translation result is also disentangled. Additionally, the above mentioned metrics are developed for a single-domain representation learning and are not directly suitable for disentanglement of *shared* and *domain-specific* factors. In this paper, we extend the attribute regression-based approach to the cross-domain translation case and propose three datasets to utilize those metrics.

3. Problem statement

Consider a set of images $A = \{a_i\}_{i=1}^{N_A}$ from the source domain \mathcal{A} , and a set of images $B = \{b_i\}_{i=1}^{N_B}$ from the target domain \mathcal{B} . The main assumption about \mathcal{A} and \mathcal{B} is that the images in both domains are semantically similar but visually different, i.e. daytime and nighttime surveillance photos, horses and zebras etc. The task of multi-modal image-to-image translation is to learn two mappings $M_{A2B} : \mathcal{A}, \mathcal{B} \rightarrow \mathcal{B}$ and $M_{B2A} : \mathcal{B}, \mathcal{A} \rightarrow \mathcal{A}$ that preserve the content information of the input image and allows to vary the domain-specific variations using the guidance image from the target domain. According to the definition provided in MUNIT [14] and later used in other UMMI2I methods, for a given pair of domains sharing common semantic information, the images in \mathcal{A} and \mathcal{B} can be mapped into the content space shared across both domains and representing the information of the image that should be preserved in the translation result, and domain-specific spaces

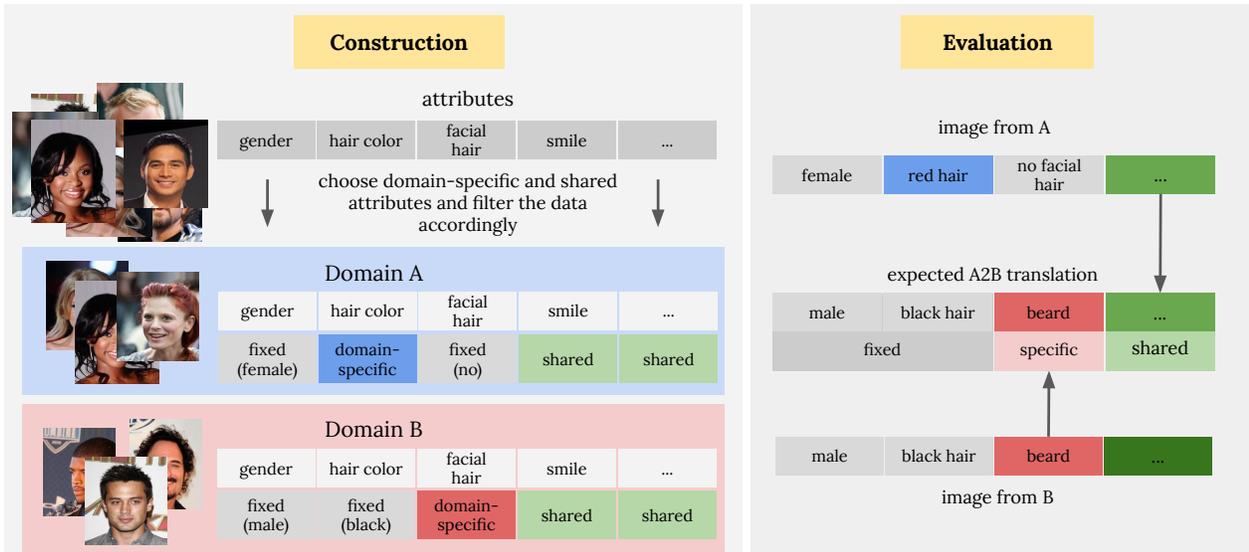


Figure 3. **Left:** Overview of the disentanglement dataset construction process. Given a dataset of images with some ground truth attribute annotation, we first pick a domain-splitting attribute (e.g. gender); then we pick two non-overlapping sets of the domain-specific attributes (e.g. facial hairstyle for male domain and hair color for female domain); we filter the original datasets so that the domain-specific attributes are varied only in one of the domains (e.g. only kept male photos with black hair). The resulting domains have known domain-specific attributes (hair color vs facial hairstyle) and shared attributes (everything else). **Right:** To evaluate the disentanglement of the translation example, we compare its attributes to those of the perfectly disentangled result. The perfectly disentangled translation result has the same content attributes values as the input image, the target domain’s attributes same as in the guidance image, and the correct values of the fixed attributes (domain-splitting attribute and the opposite domain’s specific attributes).

that encode the domain-specific factors of variation. The UMMI2I methods aim to implicitly learn representations $R_c^A : \mathcal{A} \rightarrow \mathcal{C}$ and $R_c^B : \mathcal{B} \rightarrow \mathcal{C}$ that map the input images to the shared subspace \mathcal{C} of the information shared across domains, and $R_s^A : \mathcal{A} \rightarrow \mathcal{S}_A$ and $R_s^B : \mathcal{B} \rightarrow \mathcal{S}_B$ mapping the images to the domain-specific subspaces \mathcal{S}_A and \mathcal{S}_B that reflect the domain-specific factors of variation, s.t. \mathcal{C} , \mathcal{S}_A and \mathcal{S}_B are mutually independent w.r.t. the mappings M_{A2B} and M_{B2A} . Intuitively, this means that the representation is correctly disentangled – and therefore, the translation is semantically correct – if and only if *all* information of the domain-specific factors is reflected in the domain-specific representation of the given image, *all* the information of the common factors of variation is encoded into the shared representation, and the shared and domain-specific representations are *mutually independent*, meaning that the changes in former factors should not affect the representation of the latter and vice versa.

4. Image Translation Disentanglement Metrics

In this section, we describe our proposed UMMI2I disentanglement metrics based on the definition in Section 3. We assume that all the factors of variation in the dataset can be approximated by a set of *attributes* $\mathcal{Z} = \{\zeta_i\}_{i=1}^M$ and we have access to a fully-annotated dataset D for which the values of all attributes are known, i.e. for each example $d_k \in D$ there is a corresponding vector $\mathbf{y}_k = (y_k^1, y_k^2, \dots, y_k^M)$ describing the values of attributes in \mathcal{Z} . In real image-

to-image translation applications, this assumption does not hold, as the only information available for most I2I datasets is either the correspondence to source or target domain or pairwise correspondence of samples in domains. For this reason, it is close to impossible to measure UMMI2I disentanglement empirically for an arbitrary dataset, and we propose to leverage the existing complex datasets such as CelebA, 3D-Shapes and SynAction, and modify these datasets so that the domain-specific and shared attributes are known. Figure 3 gives a schematic overview of the dataset construction and evaluation processes.

4.1. Creating a disentanglement dataset

Without loss of generality, we assume that all attributes in \mathcal{Z} are binary for simplified notation, but our approach can be naturally extended to categorical or continuous attributes. For a fully-annotated dataset $\mathcal{D} = \{d_k\}_{k=1}^N$ according to a set of attributes \mathcal{Z} , we select an attribute z_d that splits it into two domains (e.g. gender attribute for CelebA): $A = \{d_k | y_k^d = 0\}$, $B = \{d_k | y_k^d = 1\}$. After that, the domain-specific attributes \mathcal{Z}_s^A and \mathcal{Z}_s^B are selected for each domain so that $\mathcal{Z}_s^A \cap \mathcal{Z}_s^B = \emptyset$. According to the definition of the domain-specific attributes in Section 3, the domain-specific attributes of domain \mathcal{A} vary only in domain \mathcal{A} and are fixed for domain \mathcal{B} . Therefore, for domain \mathcal{A} , we pick a set of values $T_A = \{t_i\}_{i \in \mathcal{Z}_s^B}$ and filter the dataset so that the domain-specific attributes of domain \mathcal{B} are fixed for all examples in domain \mathcal{A} : $A^* = \{d_i \in$

$A|y_i^k = t_k \forall k \in \mathcal{Z}_s^B\}$. We perform the same procedure to obtain B^* . As a result, we can use the subsets A^* and B^* as a pair of domains to measure the UMMI2I disentanglement, since their domain-specific (\mathcal{Z}_s^A and \mathcal{Z}_s^B) and shared ($\mathcal{Z}_c = \mathcal{Z}/\{\mathcal{Z}_s^A \cup \mathcal{Z}_s^B \cup z_d\}$) attributes are known by design.

Considering that the UMMI2I translation models $M_{A2B} : \mathcal{A}, \mathcal{B} \rightarrow \mathcal{B}$ and $M_{B2A} : \mathcal{B}, \mathcal{A} \rightarrow \mathcal{A}$ are trained on the datasets A^* and B^* , we use them to acquire the translation results $a\hat{2}b = M_{A2B}(a, b)$, for pairs of images $a \in A^*$, $b \in B^*$. In practice, for each content example, we perform the translation with only a few guidance images due to memory and computation time limitations.

4.2. Computing Disentanglement Metrics

After the cross-domain translation results \hat{A} and \hat{B} are generated, we can use them to estimate quality of UMMI2I disentanglement. Ideally, the translation $a\hat{2}b = M_{A2B}(a, b)$ of input image $a \in A^*$ guided with image $b \in B^*$ must preserve shared attributes of image a while matching the domain B-specific attributes to those of image b . Namely, the attributes of a perfectly disentangled translation y_{a2b} should be defined as follows:

$$y_{a2b}^k = \begin{cases} y_a^k, & \text{if } k \in \mathcal{Z}_c \\ y_b^k, & \text{if } k \in \mathcal{Z}_s^B \\ T_B^k, & \text{if } k \in \mathcal{Z}_s^A \\ 1, & \text{if } k = z_d \end{cases} \quad (1)$$

Similarly, we can define the attributes of the opposite translation. To estimate the values of attributes \mathcal{Z} for the newly generated images \hat{A} and \hat{B} , we train a predictor model $R : \mathcal{D} \rightarrow \mathcal{Z}$, i.e. $y_{a\hat{2}b} = R(a\hat{2}b)$. In practice, unless stated otherwise, to take into account the error of attribute prediction by R , we compute the perfect disentanglement attributes y_{a2b}^k using attributes of input images a and b predicted by R . In all experiments, we use CNN-based predictors.

Given Equation 1, we formulate the following disentanglement metrics for the mapping M_{A2B} , the metrics for the opposite mapping M_{B2A} are defined similarly. Importantly, if both the input and the guidance images have the same value of a particular attribute $y_a^k = y_b^k$, we cannot make any judgements about the disentanglement quality of the translation result y_{a2b} for the attribute k . Therefore, we compute all our proposed metrics only on the pairs of images with different values of a needed attribute.

First, we introduce the overall translation quality metric Q_{tr} that measures the *semantic* quality of domain translation as a likelihood of the translated image to have the corresponding fixed attribute values of the target domain given that the original input image had different attribute values:

$$Q_{tr}^{A2B} = \mathbb{E}_{k \sim \mathcal{Z}_s^A \cup z_d} \mathbb{P}(y_{a2b}^k = T_B^k \mid y_a^k \neq T_B^k) \quad (2)$$

Here, we consider that $T_B^{z_d} = 1$.

In order to estimate the quality of content preservation, we introduce the metric D_c that measures the average likelihood of a shared attribute to be preserved after translation to the target domain:

$$D_c^{A2B} = \mathbb{E}_{k \sim \mathcal{Z}_c} \mathbb{P}(y_{a2b}^k = y_a^k \mid y_a^k \neq y_b^k), \quad (3)$$

The quality of domain-specific manipulation D_s^{A2B} is defined as:

$$D_s^{A2B} = \mathbb{E}_{k \sim \mathcal{Z}_s^B} \mathbb{P}(y_{a2b}^k = y_b^k \mid y_a^k \neq y_b^k), \quad (4)$$

Additionally, we can estimate the model bias for certain attributes by computing how likely the translation changes an attribute given that it is same for both input and guidance images:

$$B^{A2B} = \mathbb{E}_{k \sim \mathcal{Z}} \mathbb{P}(y_{a2b}^k \neq y_b^k \mid y_a^k = y_b^k) \quad (5)$$

For all of the metrics mentioned above, we combine the results for mappings M_{A2B} and M_{B2A} by averaging them.

Finally, the overall disentanglement quality D is computed as the average of the content preservation and domain-specific manipulation metrics $D = \frac{1}{2}(D_s + D_c)$.

4.3. How to interpret the results

Q_{tr} measures whether the translation result belongs to the target domain according to the fixed attributes, i.e., whether the translation result is a male etc., and *not* the visual quality of the generated images.

D_s and D_c measure how reliably the translation can be manipulated by the guidance example and how well the content of the input image is preserved respectively.

D measures the overall disentanglement quality.

Bias metric B , which is the likelihood of the attribute to change if it was same for both the input and guidance image, can be high only in two cases: 1) if a mode collapse appeared and the results are biased towards certain attribute values or 2) if the error of the attribute prediction is high on the translated images. It is important to detect the inaccurate attribute prediction to avoid noise in the other metrics, but it is impossible to do explicitly since the ground truth attributes for the translations are unknown. Therefore, we visually inspect whether the high-bias result is caused by the mode collapse and if not, we consider the attribute predictions for these translation results as low-confidence and report them in gray color. These results cannot be used to judge about the disentanglement quality.

5. Disentanglement Datasets

In this section, we describe the datasets we modified to quantitatively evaluate UMMI2I disentanglement quality and translation correctness: 3D Shapes, SynAction and CelebA.

Attributes	Male	Female
Hair color	fixed (black)	varied
Age	varied	fixed (young)
Smile	varied	fixed (yes)
Facial hair	varied	fixed (no)
Makeup	varied	fixed (yes)
Facial attributes*	content	content

Table 1. Short description of the domain attribute splitting used to assemble the CelebA-D dataset. *The full list of content attributes can be found in Section 9.1 of the supplementary material.

3D-Shapes We modified the 3D Shapes [17] dataset commonly used to evaluate disentanglement in representation learning. The original dataset contains the following attributes: object shape, object hue, object size, wall hue, floor hue, and orientation. We modified the 3D-Shapes dataset according to the protocol in Section 4.1 as follows: *Content attributes*: shape and hue of the foreground object. *Domain A-specific attributes*: floor hue (fixed to red in domain B), wall hue (fixed to blue in domain B). *Domain B-specific attributes*: object size (fixed to 5 out of 8 in domain A) and orientation (fixed in domain A to -30). Due to the very limited number of attributes in this dataset, we omitted the domain-splitting attribute and considered only the fixed attributes for the evaluation of domain translation quality. The resulting domains A and B contain 4000 and 4800 images respectively.

SynAction The SynAction dataset is a synthetic dataset containing videos of 10 different actors (identities) performing the same set of actions on 5 various backgrounds. We extended the dataset by introducing 5 more backgrounds by cropping and stitching the available backgrounds to make the dataset more balanced. For this dataset, the available attributes are: identity, pose and background. We created the disentanglement dataset by assigning the pose as the *shared* attribute, the background as the *domain A-specific attribute* and the identity as the *domain B-specific attribute*. To compare the pose attribute, we count the translation pose attribute as 1 if its pose is closer to that of the input image and 0 otherwise. The resulting SynAction dataset contains 6720 images in the domain A and 7560 images in the domain B.

CelebA-D To perform the evaluation on a more challenging and commonly used image translation dataset, we modified the CelebA [27] dataset containing centered photos of celebrities annotated with 40 attributes, such as hair color, gender, age etc. First, we chose the domain-splitting attribute to be "Male", i.e. the dataset was split into Male and Female subsets. We chose hair color as the varied attribute for Female domain, and presence or absence of facial hair, smile and age for Male domain; the hair color for Male domain was fixed to black. We considered the attributes associated with facial features, lighting and pose as the shared

attributes (see Table 1 for a short description and Section 9.1 of the supplementary material for a detailed list of shared attributes). Our modified CelebA-D subset contains 24661 and 29627 images in domains A and B respectively.

6. Experiments

We present our results on the three disentanglement datasets from Section 5 for Augmented CycleGAN, DRIT++, FUNIT, MUNIT, and StarGANv2. Additionally, we noticed that the original implementation of MUNIT performs poorly when a the domain-specific embedding is taken from the real guidance image and not sampled randomly. We fixed this issue by adding the domain-specific embedding reconstruction loss and GAN loss for the non-random embedding and reported these results under the name "MUNITX". For all methods, we used the official implementations and default hyperparameters. We did not split the data into train and test subsets in order to exclude the effect of possible differences in the train/test distributions on the disentanglement results. For the detailed description of the attribute prediction models, please refer to Section 9.2 of the supplementary material.

To provide a sense of scale, we added four "naive" baselines: "Rand. triplets" - completely random triplets of images as inputs and outputs, "Rand. target" - random pairs of images from source and target and a random image from the target domain as a translation result, "Content idt" - always returning the input (content) image, and "Guidance idt" - always returning the guidance image.

7. Results

The results on on Tables 2 and 3 as well as the per-attribute results (see section 9.3 of the supplementary material) indicate the following:

1. AdaIN-based methods (MUNIT, MUNITX, FUNIT and StarGANv2) tend to treat color- and texture- related information as domain-specific factors regardless of whether these attributes are actually domain-specific or shared. For example, MUNITX and StarGANv2 perform hair color change correctly with more than a 80% accuracy on average, but fail to correctly change the smile with the accuracy 34% and 22% respectively or add a correct facial hairstyle (below 28% and 16.6% respectively). In the 3D-Shapes experiments, MUNIT, MUNITX and StarGANv2 were able to correctly change the wall and floor hue, but also changed the object hue which is the content information. MUNITX performed well on the SynAction translation and was able to correctly change the identity and background, but interestingly, often mistook the positional information in the domain-specific attributes, e.g. swapped top and bottom parts of the background (see Figure 5 for visual examples).
2. The DRIT++ model preserves the content well across all

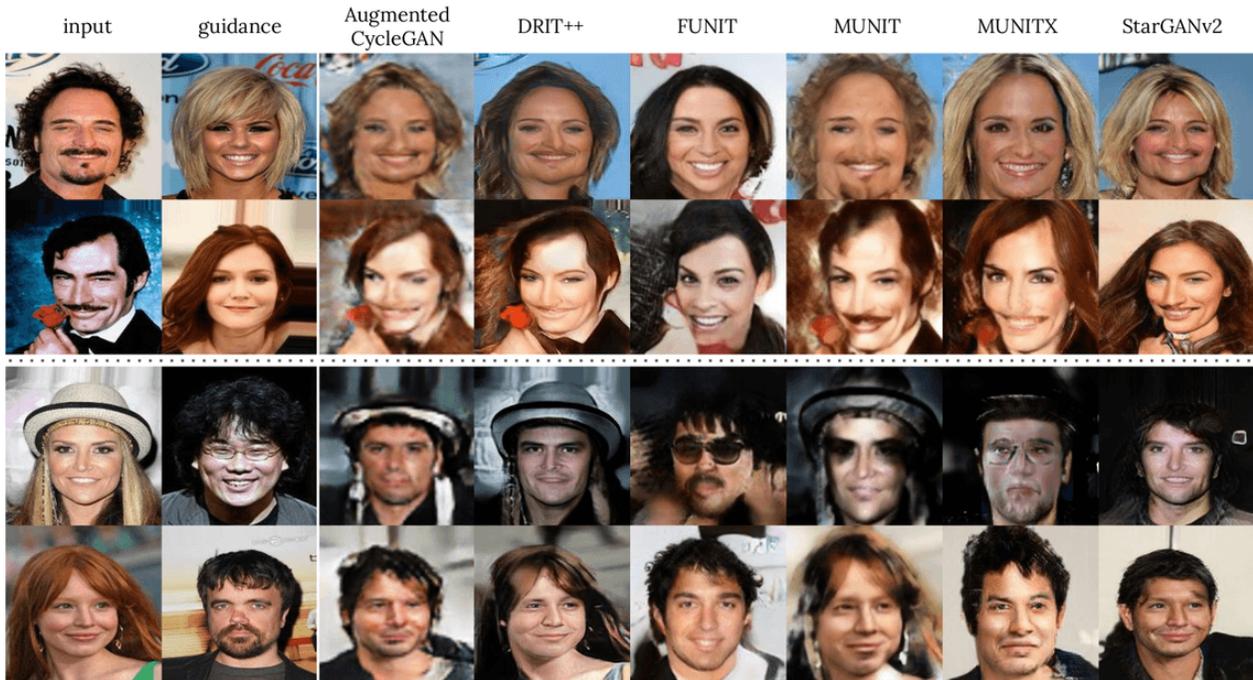


Figure 4. UMMI2I translation results on the CelebA subset. Domain-specific attributes: facial hair, smile and age (M) and hair color (F).

Model	$Q_{tr} \uparrow$	$D \uparrow$	$D_s \uparrow$	$D_c \uparrow$	$B \downarrow$
DRIT++	87.84	45.27	37.15	53.38	12.84
MUNIT	79.75	44.27	32.33	56.22	11.14
MUNITX	81.43	47.45	43.63	51.26	12.38
FUNIT	92.12	39.14	31.53	46.75	16.02
AugCycleGAN	90.92	42.10	38.94	44.25	14.86
StarGANv2	94.47	50.09	52.48	47.70	12.25
Rand. triplets	72.43	37.56	25.22	49.90	20.54
Rand. target	96.07	41.68	39.88	43.49	18.46
Content idt	0.0	50.0	0.0	100.0	0.0
Guidance idt	96.07	50.0	100.0	0.0	0.0

Table 2. UMMI2I disentanglement results of DRIT++[23], MUNIT[14], FUNIT[26], Augmented CycleGAN[2] and StarGANv2 [4] on **CelebA-D** subset. The results include domain translation quality Q_{trans} , overall disentanglement score D , domain-specific attribute disentanglement score D_s , shared attribute disentanglement score D_c and domain-specific model bias B . The bottom part contains the results on the naive baselines described in Section 6. We marked the results for which the attribute prediction accuracy was low gray to indicate the low confidence in the corresponding disentanglement scores. The results interpretation guide can be found in Section 4.3.

experiments (e.g. correctly preserves eyeglasses (82%) and hats (81%) from the content image compared to MUNITX (54% and 57% respectively). However, it shows poor performance on the domain-specific manipulation even on the color- and texture- based attributes with the average accuracy of hair color change $\approx 33\%$ and accuracy of correct wall and floor hue change on 3D-Shapes below 16%.

3. While showing remarkable success in few-show image

translation, FUNIT experiments on all three datasets resulted in moderate to severe mode-collapse, and therefore both the content preservation and domain-specific manipulation of this method are very unreliable.

4. All methods we tested preserve well the spatial information, such as head or body pose, with the translation pose being closer to that of the input image than the guidance image in more than 91% of the cases for all models that did not collapse and the attribute predictor was accurate. The best models in this aspect are MUNIT and MUNITX (98% and 76% on CelebA and SynAction respectively).

8. Conclusion

We believe that one can only improve upon the aspects that can be quantitatively evaluated, and our work provides a much needed way to measure the semantic correctness of UMMI2I translation. In this paper, we formalized the notion of semantic correctness of image translation in terms of shared and domain-specific attributes and proposed three protocols for the evaluation of UMMI2I methods. The results of five state-of-the-art UMMI2I translation methods (MUNIT, FUNIT, DRIT++, Augmented CycleGAN and StarGANv2) show that there is a clear trade-off between the content preservation and domain-specific manipulation quality. Our results indicate that current SoTA UMMI2I methods perform well on a very limited range of applications, and more effort should be made to create the UMMI2I disentanglement techniques that allow more flexible separation of domain-specific and shared factors of variation.



Figure 5. Illustration of many-to-many image translation results on 3D-Shapes (top) and SynAction (bottom) subsets.

Model	3D Shapes					SynAction				
	$Q_{tr} \uparrow$	$D \uparrow$	$D_s \uparrow$	$D_c \uparrow$	$B \downarrow$	$Q_{tr} \uparrow$	$D \uparrow$	$D_s \uparrow$	$D_c \uparrow$	$B \downarrow$
DRIT++	93.58	32.72	12.77	52.67	43.746	89.67	55.78	19.89	91.66	25.77
MUNIT	94.93	68.11	88.06	48.17	5.12	86.88	66.94	35.62	98.26	23.83
MUNITX	92.91	33.26	14.46	52.06	38.24	92.19	69.49	41.06	97.91	20.55
FUNIT	25.00	14.65	11.83	17.48	66.53	37.41	27.19	3.42	50.98	15.2
AugCycleGAN	50.24	16.53	12.42	20.63	63.65	99.92	37.65	25.28	50.01	33.33
StarGANv2	95.96	29.04	49.79	8.28	26.06	96.82	31.38	13.17	49.60	33.11
Rand. triplets	54.77	11.79	6.16	17.43	59.41	49.91	28.05	6.08	50.01	36.52
Rand. target	99.92	14.94	11.86	18.02	49.51	99.94	30.98	11.93	50.02	18.06
Content idt	0.0	50.0	0.0	100.0	0.0	0.0	50.0	0.0	100	0.0
Guidance idt	99.89	50.0	100.0	0.0	0.0	99.89	50.0	100	0.0	0.0

Table 3. UMMI2I disentanglement results on our modification of the **3D-Shapes** (left) and **SynAction** (right) datasets. The notation is same as in Table 2.

References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] Amjad Almahairi, Sai Rajeswar, Alessandro Sordani, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.
- [3] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31:2610–2620, 2018.
- [4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.
- [5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
- [6] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.
- [7] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.
- [8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [9] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017.
- [10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.
- [11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.
- [12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.
- [13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.
- [14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.
- [15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [16] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.
- [17] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.
- [18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4422–4431, 2019.
- [21] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.
- [22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.
- [24] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Dri++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.
- [25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017.
- [26] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *arxiv*, 2019.
- [27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of*

- International Conference on Computer Vision (ICCV)*, December 2015.
- [28] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018.
- [29] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. 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)*, pages 269–286, 2018.
- [30] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.
- [32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [33] Ximeng Sun, Huijuan Xu, and Kate Saenko. Twostreamvan: Improving motion modeling in video generation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2744–2753, 2020.
- [34] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016.
- [35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.
- [36] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.
- [37] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2994–3004, 2019.
- [38] Lvmin Zhang, Yi Ji, Xin Lin, and Chunping Liu. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 506–511. IEEE, 2017.
- [39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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.
- [40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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.
- [41] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

9. Supplementary material

9.1. More details on the datasets

CelebA The list of content attributes: “5_o_Clock_Shadow”, “Arched_Eyebrows”, “Bags_Under_Eyes”, “Big_Lips”, “Big_Nose”, “Blurry”, “Bushy_Eyebrows”, “Chubby”, “Double_Chin”, “Eyeglasses”, “High_Cheekbones”, “Narrow_Eyes”, “Oval_Face”, “Pale_Skin”, “Pointy_Nose”, “Straight_Hair”, “Wavy_Hair”, “Wearing_Hat”.

9.2. Attribute prediction networks

If not stated otherwise, the attribute prediction networks are implemented in Tensorflow [1].

3D-Shapes For wall hue, floor hue, object hue, size and shape classification, we used the CNNs with the following architecture: 2D convolution with $16 \ 3 \times 3$ filters followed by a ReLU activation function and max pooling layer with pooling stride 2×2 , another convolution layer with $32 \ 3 \times 3$ with ReLU activation and 2×2 max pooling; dropout layer with the drop probability 0.2, flattening layer, a dense layer with 128 units and a final dense prediction layer with the number of units equal to the number of classes in the task. The networks are trained with Adam optimizer [19] using a sparse categorical cross-entropy loss for until convergence. All classifiers achieve nearly 100% test accuracy for all tasks. For the orientation regression task, we use the following architecture: 2D convolution with $32 \ 3 \times 3$ filters followed by a ReLU activation function and max pooling layer with pooling stride 2×2 , another convolution layer with $16 \ 3 \times 3$ with ReLU activation and 2×2 max pooling; dropout layer with the drop probability 0.2, flattening layer, and the dense layer with a single unit for the final prediction. We use the mean squared error loss and Adam optimizer to train the network. The resulting accuracy on the orientation task is $> 98\%$ on test set.

SynAction To predict the identity and background, we use the following classification network architecture: three convolution layers with 16, 32 and 64 filters 3×3 filters respectively all followed by a ReLU activation function and max pooling layer with pooling stride 2×2 , dropout layer with the drop probability 0.2, flattening layer, a dense layer with 128 units and a final dense prediction layer with the number of units equal to the number of classes in the task. The networks are trained with Adam optimizer [19] using a sparse categorical cross-entropy loss for until convergence. The classifiers achieve $> 98\%$ test accuracy for both tasks. For pose estimation, we use the pretrained Personlab [29] model from Tensorflow Lite (see pose estimation visualization on SynAction in Section 9.5).

CelebA For attribute on the CelebA dataset, we used the MobileNetv2 [32] feature extractor followed by two dense layers with 1024 and 512 units respectively and ELU non-linearity [5], and the last dense layer with 40 units and the sigmoid non-linearity. The average attribute classification accuracy of this network is 92%, see the detailed information on per-attribute accuracy on Figure 11. Additionally, we measured how well the translation preserves the pose with the HopeNet [30] model pretrained on the 300W LP dataset [41] and reported the results in Table 6.

9.3. Tables

Please see Tables 4 and 5 for attribute-wise results and Table 6 for the pose estimation results on the CelebA-D subset, and Tables 7 and 8 for the attribute-wise results on the 3D-Shapes and SynAction datasets respectively.

9.4. More translation examples

For more illustrations of the UMMI2I translation results, please see Figures 6, 7, 8, 9 and 10.

9.5. Pose estimation examples

Please see the illustration of pose estimation results on Figures 12 and 13.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
DRIT	52	82	44	48	55	37	46	54	46	61	61	81	69	59	38	42	49	40
MUNIT	51	89	48	46	51	55	64	68	51	54	58	65	66	53	45	50	51	50
MUNITX	55	54	50	49	55	35	53	55	43	54	49	57	55	54	44	54	55	53
FUNIT	51	52	41	49	52	34	39	52	43	54	51	53	51	55	38	52	41	35
AugCycleGAN	53	40	49	49	47	21	38	46	32	52	46	46	43	53	47	50	47	38
StarGANv2	54	77	44	35	41	43	50	53	21	55	50	51	71	57	36	36	44	40

Table 4. CelebA content attribute results. The attribute indices correspond to the attributes as follows: 1. “5_o_Clock_Shadow”, 2. “Arched_Eyebrows”, 3. “Bags_Under_Eyes”, 4. “Big_Lips”, 5. “Big_Nose”, 6. “Blurry”, 7. “Bushy_Eyebrows”, 8. “Chubby”, 9. “Double_Chin”, 10. “Eyeglasses”, 11. “High_Cheekbones”, 12. “Narrow_Eyes”, 13. “Oval_Face”, 14. “Pale_Skin”, 15. “Pointy_Nose”, 16. “Straight_Hair”, 17. “Wavy_Hair”, 18. “Wearing_Hat”.

Method	Blond	Brown	Black	Young	Smile	Beard	Sideburns	Mustache	Goatee
DRIT	31	36	80	20	33	23	4	1	2
MUNIT	31	13	86	20	35	6	< 1	< 1	< 1
MUNITX	76	35	87	26	34	16	2	4	9
FUNIT	28	7	59	22	62	28	1	2	7
AugCycleGAN	23	16	76	32	92	51	4	6	6
StarGANv2	83	68	90	17	22	28	2	2	4

Table 5. Per-attribute domain-specific manipulation results on CelebA. Left part: male2female domain-specific attributes (hair color); right part: female2male domain-specific attributes.

Model	Y ↓	P ↓	R ↓	D_p ↓	PM ↑
DRIT++	3.49	4.73	1.57	3.26	0.76
MUNIT	3.17	3.09	1.19	2.51	0.79
MUNITX	3.27	3.09	1.19	2.52	0.79
FUNIT	5.47	6.16	1.82	4.48	0.66
AugCycleGAN	16.95	8.55	3.53	9.68	0.29
StarGANv2	4.27	4.72	1.69	3.56	0.71
Random Pairs	20.6	9.14	3.52	11.09	0.50

Table 6. Pose preservation metrics for CelebA-D translation with DRIT++[23], MUNIT[14], FUNIT[26], Augmented CycleGAN[2] and StarGANv2 [4]. The results include mean yaw, pitch and roll distance of the translated image to the content image, overall mean pose distance D_p and pose match score PM . Distances of random pairs of images are included for comparison. All pose estimation results are estimated by HopeNet [30] model.

Method	Content	A-specific	B-specific
	Pose	Background	Identity
DRIT	92	13	27
MUNIT	98	19	52
MUNITX	98	24	58
FUNIT	52	0	7
AugCycleGAN	50	36	14
StarGANv2	50	9	18

Table 8. Per-attribute results on the SynAction subset.

Method	Content		A-specific		B-specific	
	Obj. hue	Obj. shape	Floor hue	Wall hue	Size	Orientation
DRIT	10	95	13	17	14	7
MUNIT	< 1	96	100	100	88	65
MUNITX	9	95	11	10	29	8
FUNIT	10	25	0	11	29	7
AugCycleGAN	11	30	10	10	22	8
StarGANv2	5	12	89	89	14	7

Table 7. Per-attribute results on 3D-Shapes subset.



Figure 6. Illustration of many-to-many image translation results on Celeba-D subset.

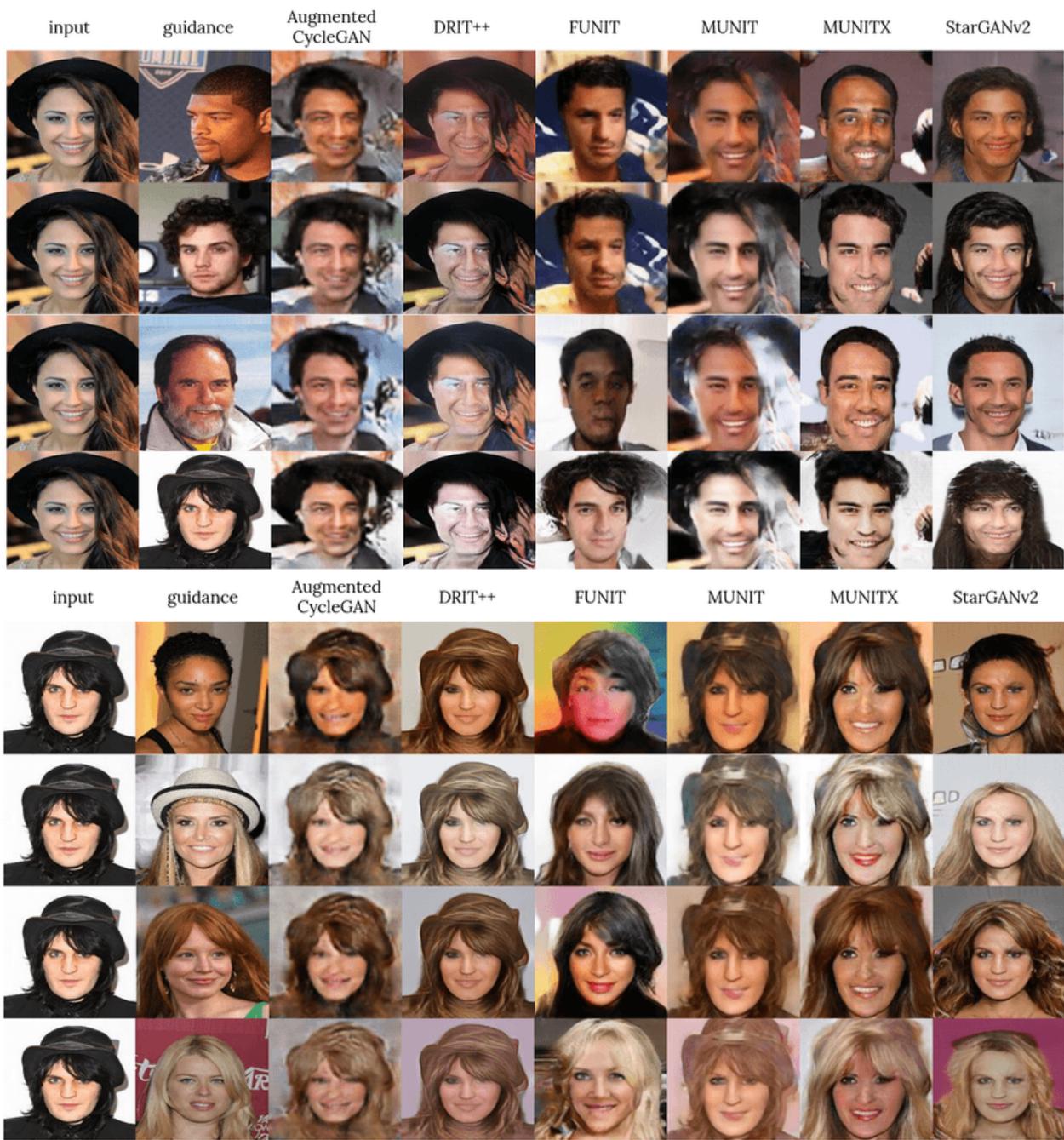


Figure 7. Illustration of many-to-many image translation results on Celeba-D subset.



Figure 8. Illustration of many-to-many image translation results on Celeba-D subset.

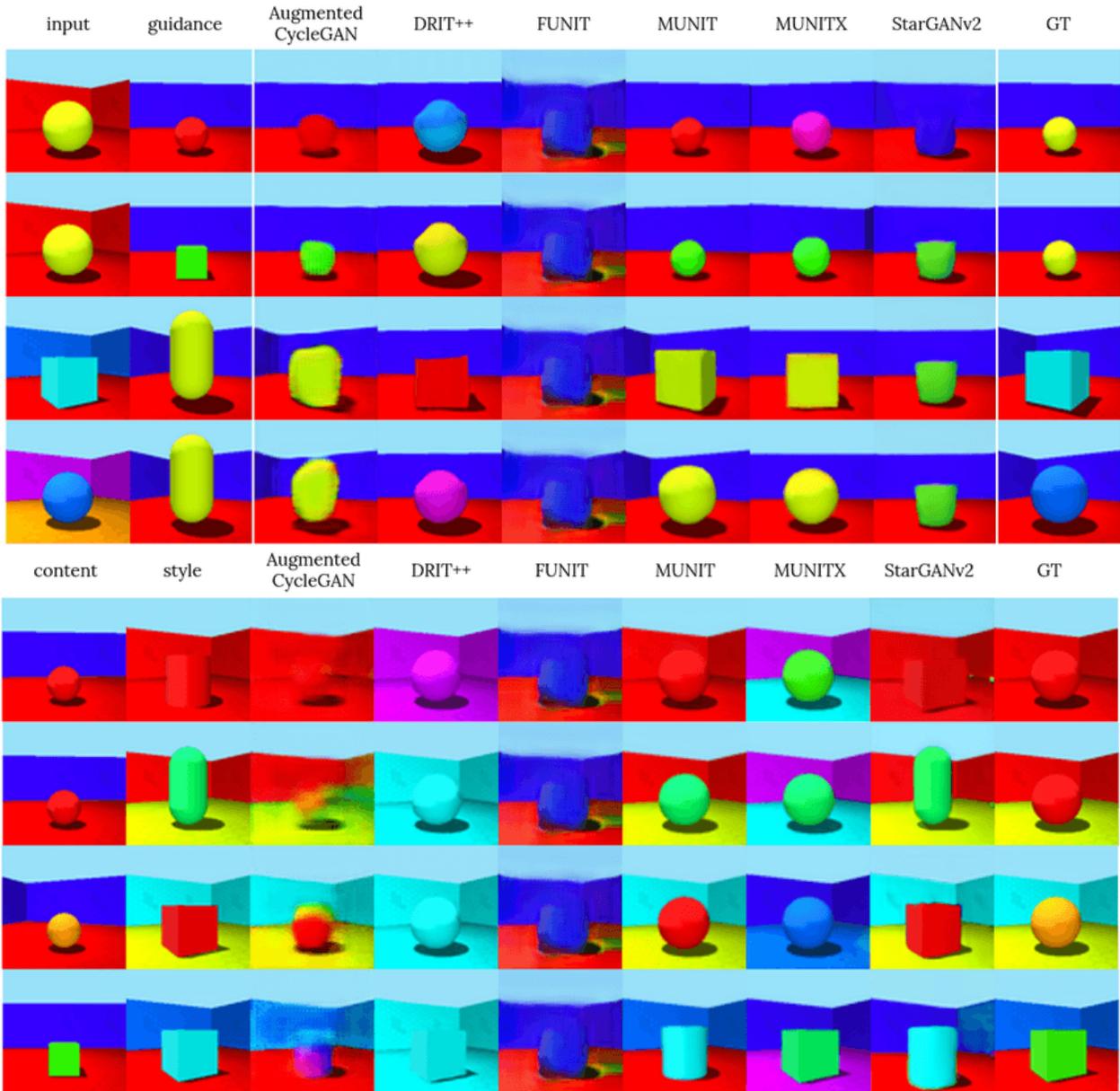


Figure 9. Illustration of many-to-many image translation results on 3D-Shapes subset.



Figure 10. Illustration of many-to-many image translation results on SynAction subset.

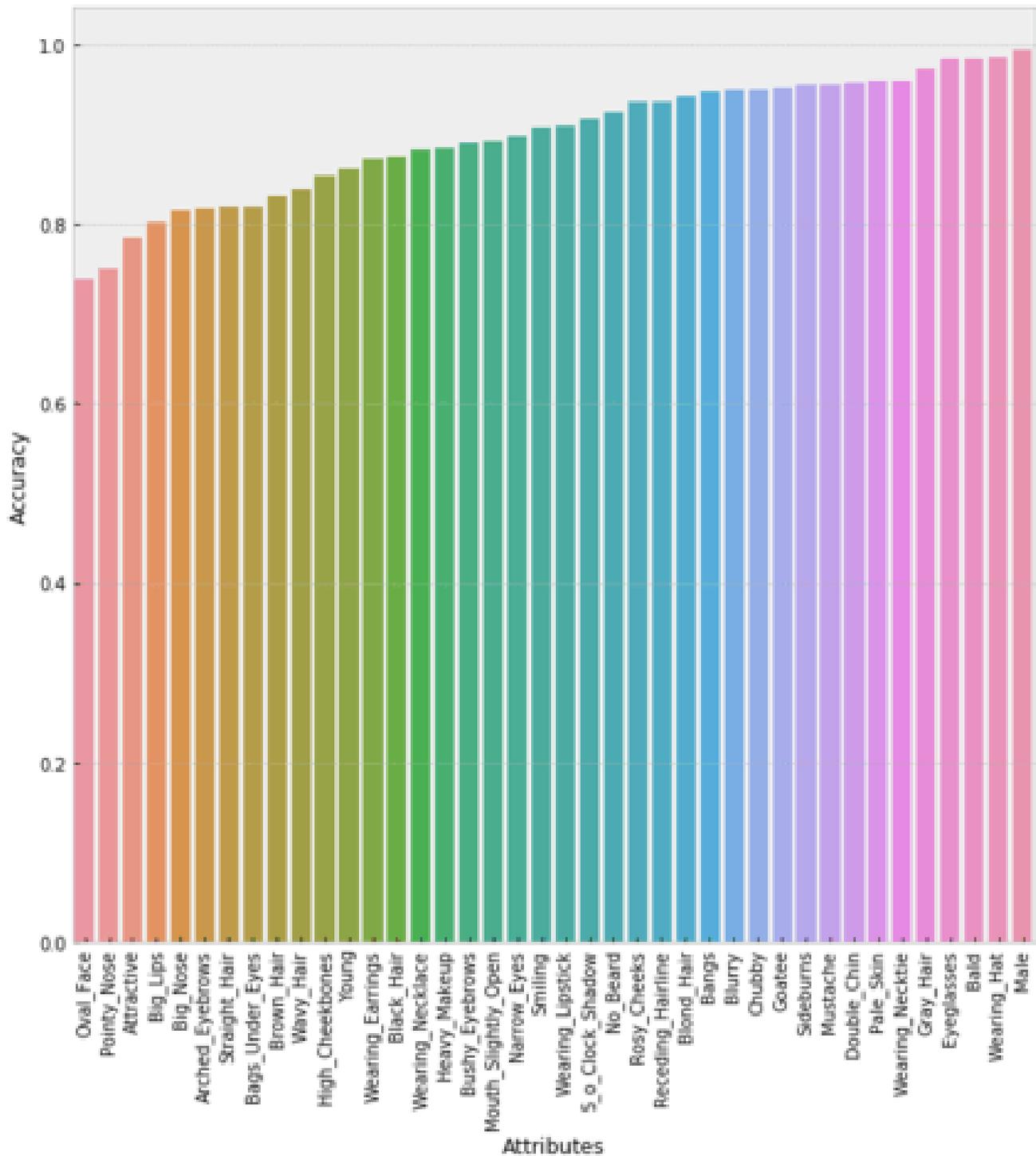


Figure 11. Per-attribute accuracy histogram achieve by our attribute prediction model.

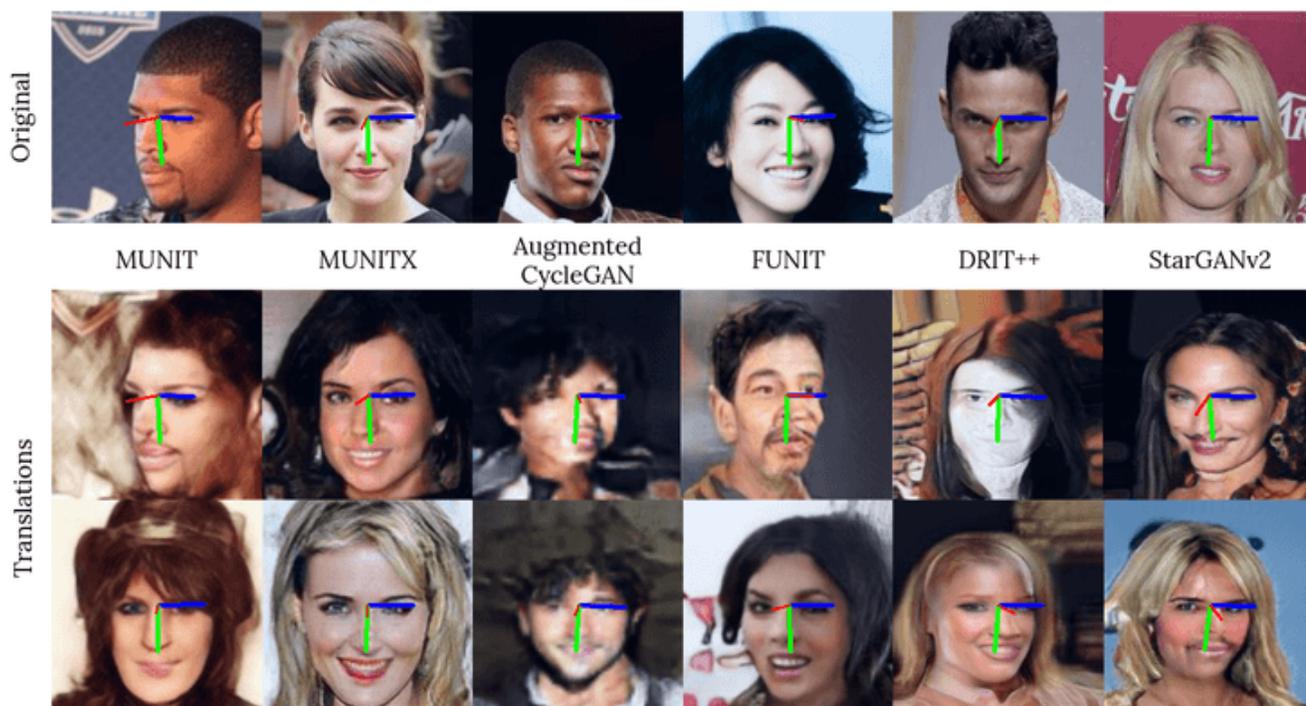


Figure 12. Head pose estimation results on random examples from the original CelebA dataset (**top**) and random translation results (**bottom**). Best viewed in color.

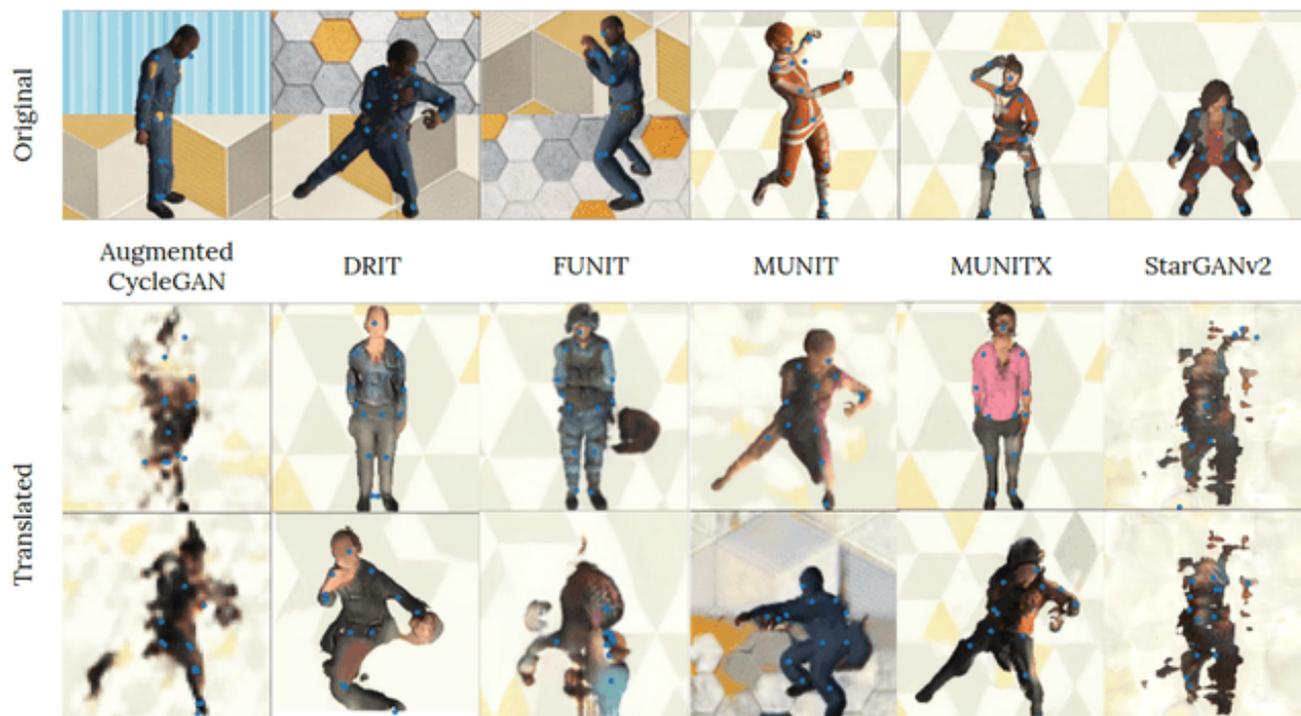


Figure 13. Pose estimation results on random examples from the original SynAction dataset (**top**) and random translation results (**bottom**). Best viewed in color.