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

Dina Bashkirova
Boston University

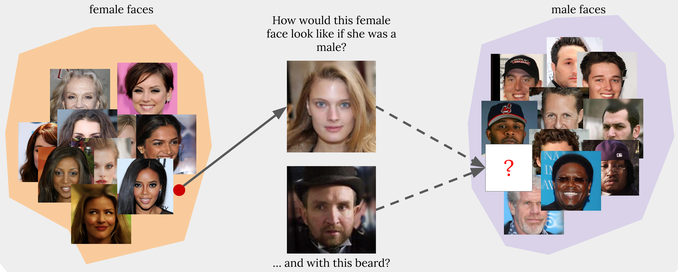
Ben Usman
Boston University

Kate Saenko
Boston University and
MIT-IBM Watson AI Lab

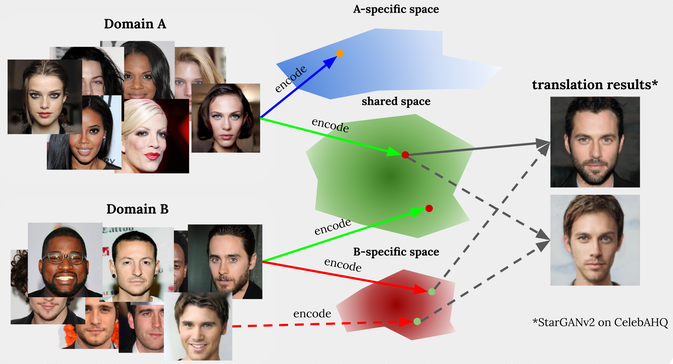
MIT-IBM Watson AI Lab

BOSTON UNIVERSITY

Task setup



Disentanglement in Image-to-Image Translation



UMI2I implicitly disentangles the shared factors varied in both domains domains (green, e.g. pose, facial expression, hair and skin color) from the domain-specific factors (in blue, e.g., makeup, and in red, e.g. facial hair).

Problem: existing metrics do not evaluate semantic correctness

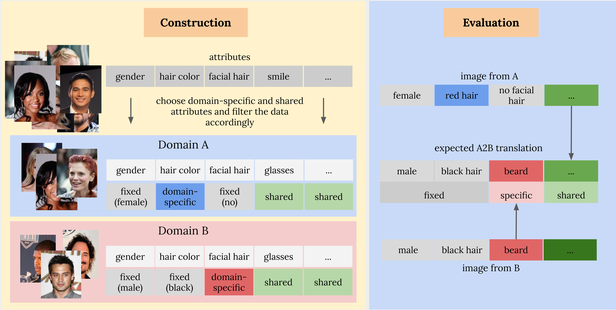
Common image-to-image translation evaluation metrics:

- FID, Inception scores (realism)
- LPIPS score (perceptual similarity, diversity)
- Human evaluation (realism)

None of the existing methods measure semantic correctness of the translation

Evaluation pipeline

1. **Construct** a dataset with known domain-specific and shared factors by utilizing the labeled datasets
2. **Train** UMI2I models on the constructed dataset
3. Using a pre-trained classifier, **predict** the shared and domain-specific attributes of the translation results
4. **Evaluate** how well the domain-specific factors are modified, and how well the shared attributes are preserved

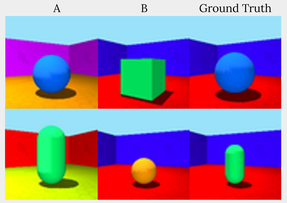


Disentanglement Datasets

CelebA

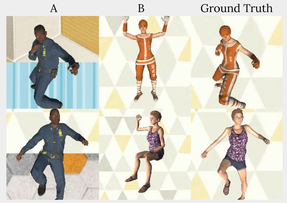
shared: glasses, hats, head pose, etc.
A-specific: facial hair, smile
B-specific: hair color

3D-Shapes



shared: object color and shape
A-specific: wall and floor colors
B-specific: object size and viewpoint

SynAction



shared: pose
A-specific: background
B-specific: identity

Key insights

1. **AdaIN-based methods** (MUNIT, MUNITX, FUNIT and StarGANv2) tend to treat color- and texture- related information as domain-specific factors, even if these factors are actually domain-invariant and should be preserved, and fail to capture spatially localized domain-specific factors.
2. **Non-AdaIN-based methods** (DRIT++, AugCycleGAN) preserved all domain-invariant attributes better across all experiments but performed poorly in transferring any domain-specific information including color- and texture-based attributes.
3. All methods preserved **domain-invariant spatial information** well, such as head orientation in CelebA or body pose in SynAction.
4. On all three datasets **FUNIT** experiments resulted in moderate to severe mode-collapse, resulting in low correctness and high bias scores. Low visual fidelity of images generated by AugCycleGAN prevented reliable automatic evaluation.
5. The **user study** resulted in same trends as the automatic evaluation reported in the main paper (observed human error rate below 10%)

Results

Model	$Q_{\text{tr}}\uparrow$	$D\uparrow$	$D_a^{\text{A2B}}\uparrow$	$D_a^{\text{B2A}}\uparrow$	$D_c\uparrow$	$B\downarrow$
DRIT++ [24]	87.8	45.2	49.1	25.1	53.3	12.8
MUNIT [14]	79.7	44.2	43.0	21.5	**56.2**	**11.1**
MUNITX	81.4	47.4	66.3	20.9	51.2	12.3
FUNIT [26]	92.1	39.1	31.6	**31.4**	46.7	16.0
AugCycleGAN [2]	90.9	42.1	38.5	41.3	44.2	14.8
StarGANv2 [4]	**94.4**	**50.0**	**80.1**	24.7	47.7	12.2
Rand. triplets	72.4	37.5	29.5	20.8	49.9	20.5
Rand. target	96.0	41.6	42.3	37.3	43.4	18.4
Content idt	0.0	50.0	0.0	0.0	100.0	0.0
Guidance idt	96.0	50.0	100.0	100.0	0.0	0.0

Model	3D Shapes						SynAction					
	$Q_{\text{tr}}\uparrow$	$D\uparrow$	$D_a^{\text{A2B}}\uparrow$	$D_a^{\text{B2A}}\uparrow$	$D_c\uparrow$	$B\downarrow$	$Q_{\text{tr}}\uparrow$	$D\uparrow$	$D_a^{\text{A2B}}\uparrow$	$D_a^{\text{B2A}}\uparrow$	$D_c\uparrow$	$B\downarrow$
DRIT++	93.5	32.7	10.5	14.9	**52.6**	43.746	89.6	55.7	27.3	12.4	91.6	25.7
MUNIT	94.9	**68.1**	**76.3**	**99.7**	48.1	**5.12**	86.8	66.9	52.2	19.0	**98.2**	23.8
MUNITX	92.9	33.2	18.6	10.2	52.0	38.2	92.1	**69.4**	**58.3**	**23.7**	97.9	20.5
FUNIT	25.0	14.6	18.2	5.39	17.4	66.5	37.4	27.1	6.79	0.5	50.9	**15.2**
AugCycleGAN	50.2	16.5	14.8	10.0	20.6	63.6	99.9	37.6	14.2	36.2	50.0	33.3
StarGANv2	**95.9**	29.0	10.3	89.2	8.28	26.0	**96.8**	31.3	17.6	8.69	49.6	33.1
Rand. triplets	54.7	11.7	7.77	4.55	17.4	59.4	49.9	28.0	5.85	6.31	50.0	36.5
Rand. target	99.9	14.9	13.5	10.1	18.0	49.5	99.9	30.9	11.1	12.7	50.0	18.0
Content idt	0.0	50.0	0.0	0.0	100.0	0.0	0.0	50.0	0.0	0.0	100	0.0
Guidance idt	99.8	50.0	100.0	100.0	0.0	0.0	99.8	50.0	100	100	0.0	0.0

UMMI2I disentanglement metrics on CelebA, 3DShapes and SynAction datasets. The bottom part contains the results on the "naive" baselines (random triplets of images, random images from the correct domains, content and guidance image as translation. Gray indicates low confidence in the corresponding disentanglement scores.

Translation results