8. Supplementary material

8.1. More details on the datasets

We did not split the data into train and test subsets to train disentanglement methods in order to exclude the effect of possible differences in the train/test distributions on the disentanglement results. It does not lead to overfitting because the task is unsupervised and we never observe any ground truth during training.

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 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 [33] 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 3 for a short description and Section 8.1 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.

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 3. Short description of the domain attribute splitting used assemble the CelebA-D dataset. *The full list of content attributes can be found in Section 8.1 of the supplementary material.

To remove inconsistency in the original labels of CelebA, we removed the examples for which the hair color is not annotated. Such choice of style attributes is dictated by the aim to leave as many examples from the original dataset as possible, i.e. to filter out the smallest number of examples when the opposite domain style attributes are being fixed, while using the most visible attributes as style attributes.

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".

8.2. 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 3×3 filters followed by a ReLU activation function and max pooling layer with pooling stride 2×2, another convolution layer with 32 3×3 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 3×3 filters followed by a ReLU activation function and max pooling layer with pooling stride 2×2, another convolution layer with 16 3×3 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 Figure 15).

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 14. 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 7.

8.3. Additional Tables

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

8.4. Examples of images from generated datasets

For illustrations of generated datasets (CelebA, 3D-Shapes and SynAction) described in details in Section 8.1 see Figures 6-8.

8.5. More translation examples

For more illustrations of the UMMI2I translation on CelebA, 3D-Shapes and SynAction, please see Figures 9-13. Our findings are summarized in the Section 6 of the main paper and are backed by metrics introduced and reported in this paper.

8.6. Pose estimation examples

Please see the illustration of pose estimation results on Figures 15 and 16. The pose estimation network succeeded in estimating poses even with severe generation artifacts.

8.7. User study

Also Table 9 reports results of the human study illustrated and described in Figure 17. Subjects were explicitly asked to label the images as having the specific attribute matching attribute values other images. These responses were used to compute human evaluation metrics reported in Table 9 and show same trends as automatic evaluation reported in Table 2.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
DRIT	42	40	52	**61**	49	54	44	**61**	**59**	82	37	69	38	48	46	**55**	46	**81**
MUNIT	50	50	51	58	51	**68**	48	54	53	**89**	**55**	66	**45**	46	**64**	51	**51**	65
MUNITX	**54**	**53**	**55**	49	55	55	**50**	54	54	54	35	55	44	**49**	53	**55**	43	57
FUNIT	52	35	51	51	41	52	41	53	55	52	34	51	38	**49**	39	52	44	54
AugCycleGAN	50	38	53	46	47	46	49	52	53	40	21	43	47	49	38	47	32	46
StarGANv2	36	40	**54**	50	44	54	44	55	57	78	43	**71**	36	35	49	41	21	51

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.

	Content		**A-specific**		**B-specific**	
Method	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 6. Per-attribute results on 3D-Shapes subset.

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 7. 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.

	Content	**A-specific**	**B-specific**
Method	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	$\boldsymbol{Q}_{\text{tr}}$ ↑	$\boldsymbol{D}$ ↑	$\boldsymbol{D_s}$ ↑	$\boldsymbol{D_c}$ ↑	$\boldsymbol{B}$ ↓
DRIT++	90.17	34.95	17.41	52.49	34.18
MUNIT	89.86	69.16	88.44	49.88	2.12
MUNITX	89.19	41.36	28.11	54.62	39.09
FUNIT	49.26	13.08	12.36	13.79	54.46
AugCycleGAN	72.81	32.33	20.97	43.69	36.73
StarGANv2	97.18	39.07	67.67	10.46	31.37

Table 9. Human evaluation of disentanglement on 3D-Shapes. It shows same trends as Table 2 with automatic evaluation results in the main paper.

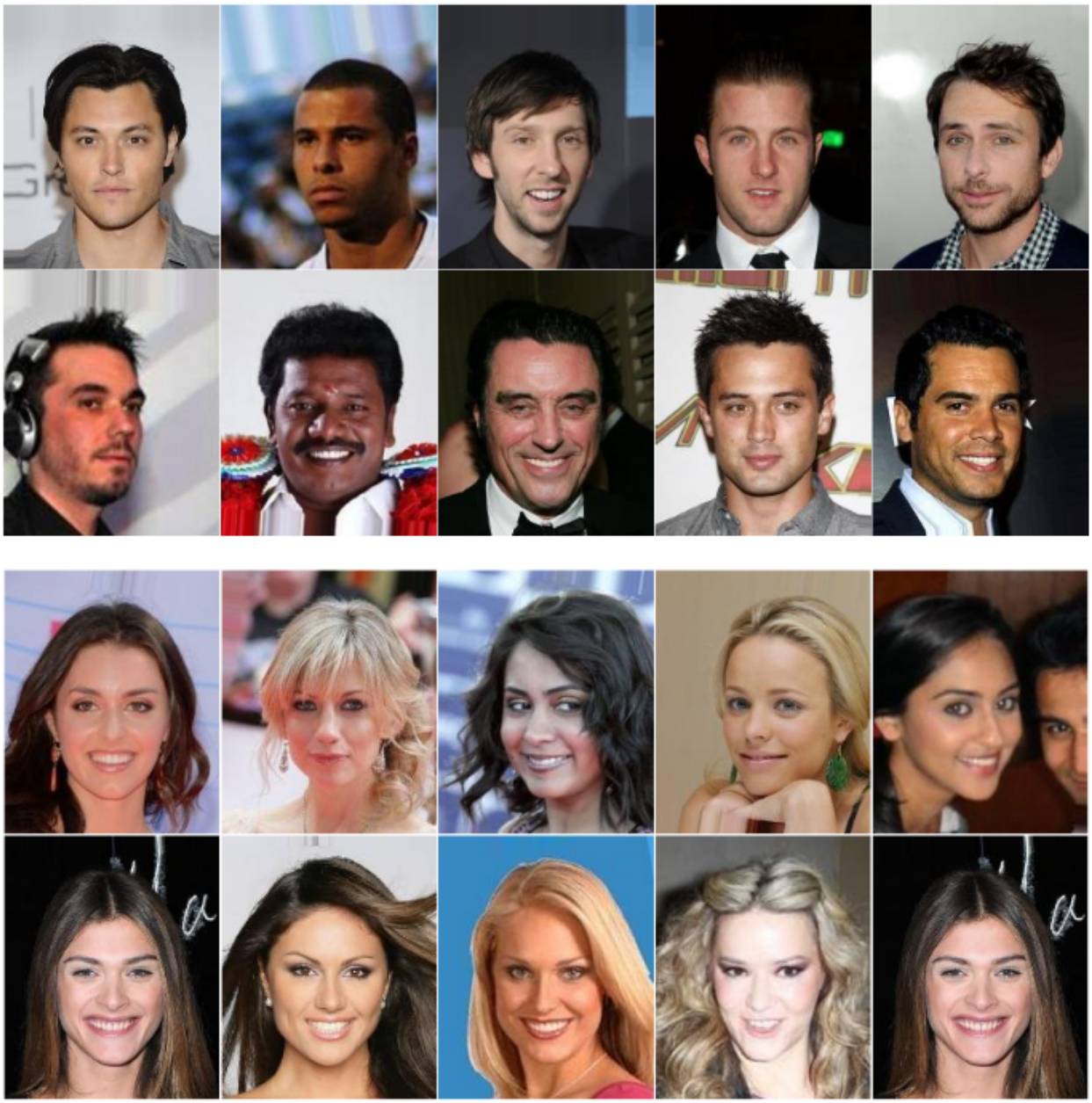

Figure 6. Random examples images from the proposed CelebA split: human faces with variable orientation of 1) males with black hair and variable amount of facial hair, amount of smile and age (top), and 2) smiling young females with variable hair color (bottom). We discuss the motivation behind this specific split in Section 5 of the main paper.

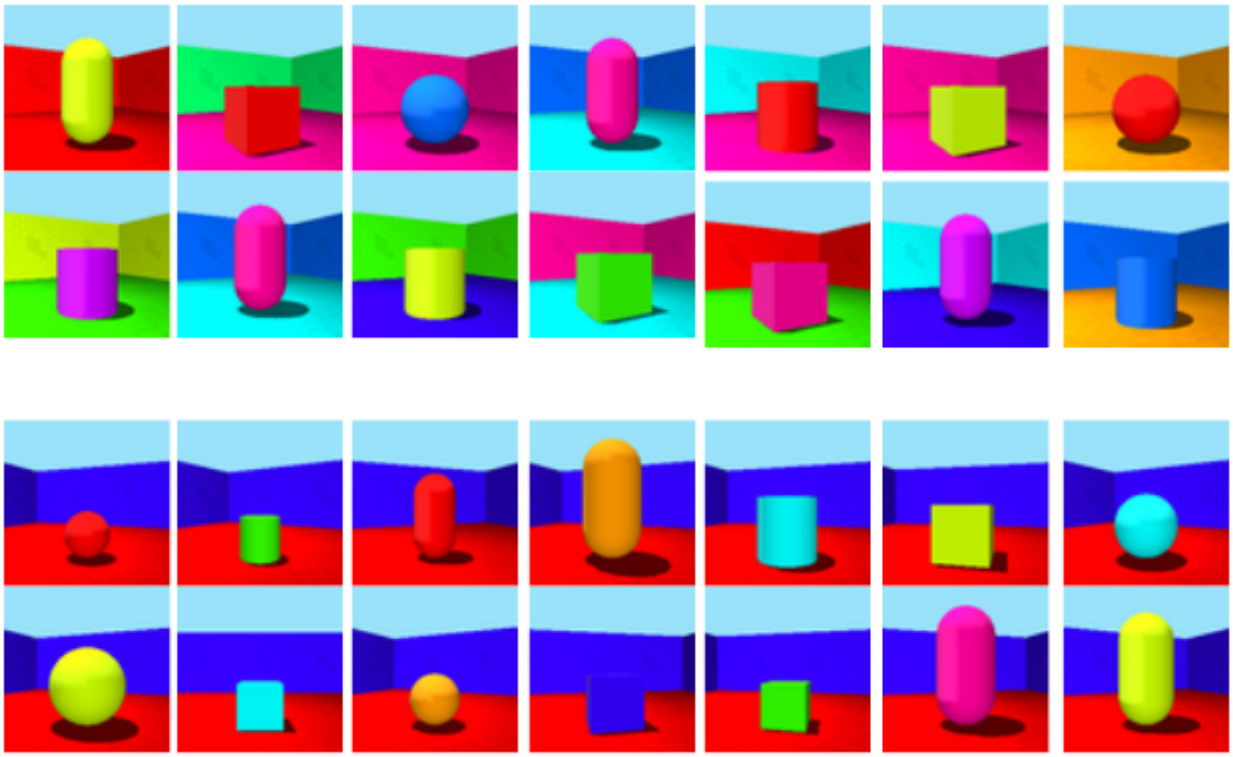

Figure 7. Random examples images from the proposed Shapes-3D split: 3D renders of one of four shapes in one of ten colors in both dataset, with floor and wall color variations in one domain, and size and view angle variation in the other domain

Figure 8. Random examples images from the proposed SynAction split: 3D renders of humans in different poses with background texture variations in one domain (top) and clothing and identity variations in the other (bottom).

Figure 9. Illustration of many-to-many image translation results on Celeba-D subset. A correct translation should have domain-specific attributes of the guidance image (hair color in the top four lines; facial hair, smile and age in the bottom four lines), and the rest of attributes (facial features, orientation, etc.) from the input image.

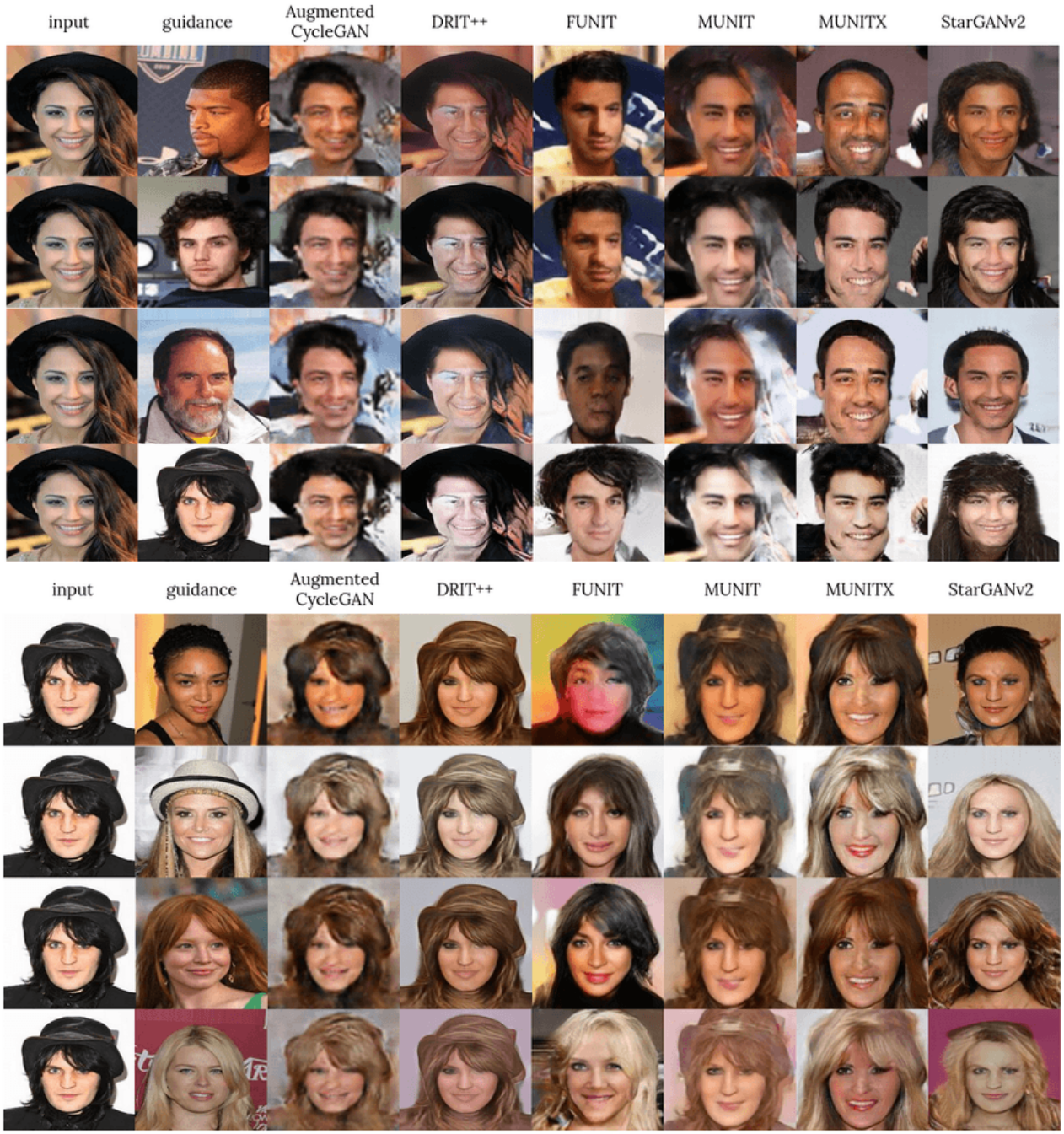

Figure 10. Illustration of many-to-many image translation results on Celeba-D subset. A correct translation should have domain-specific attributes of the guidance image (hair color in the top four lines; facial hair, smile and age in the bottom four lines), and the rest of attributes (facial features, orientation, etc.) from the input image.

Figure 11. Illustration of many-to-many image translation results on Celeba-D subset. A correct translation should have domain-specific attributes of the guidance image (hair color in the top four lines; facial hair, smile and age in the bottom four lines), and the rest of attributes (facial features, orientation, etc.) from the input image.

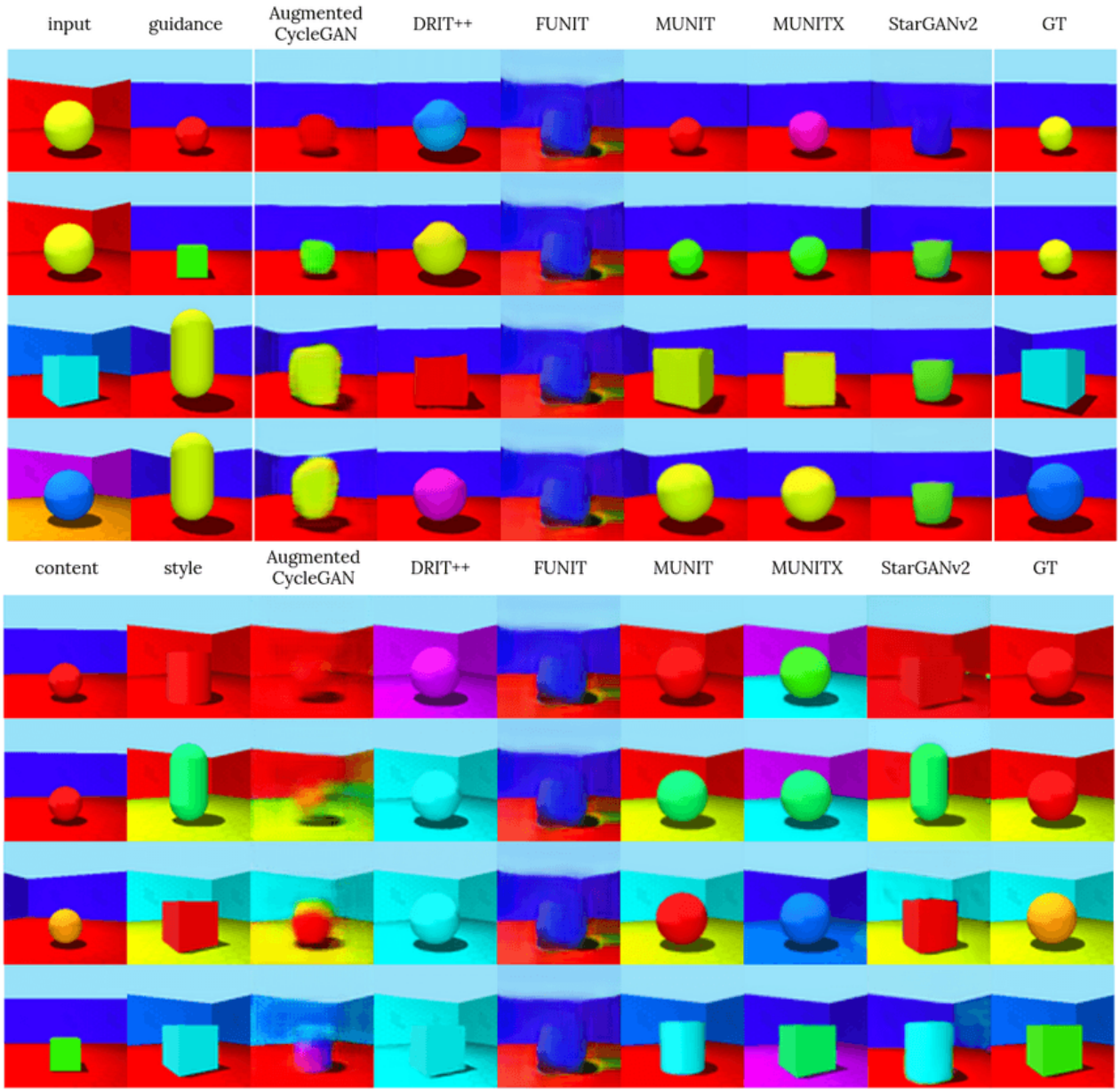

Figure 12. Illustration of many-to-many image translation results on 3D-Shapes subset. A correct translation should have domain-specific attributes of the guidance image (orientation and size in top four lines; wall and floor color in the bottom four lines), and the rest of attributes (shape type and shape color) from the input content image.

Figure 13. Illustration of many-to-many image translation results on SynAction subset. A correct translation should have domain-specific attributes of the guidance image (background texture in top four lines; clothing and identity in the bottom four lines), and the pose from the input content image.

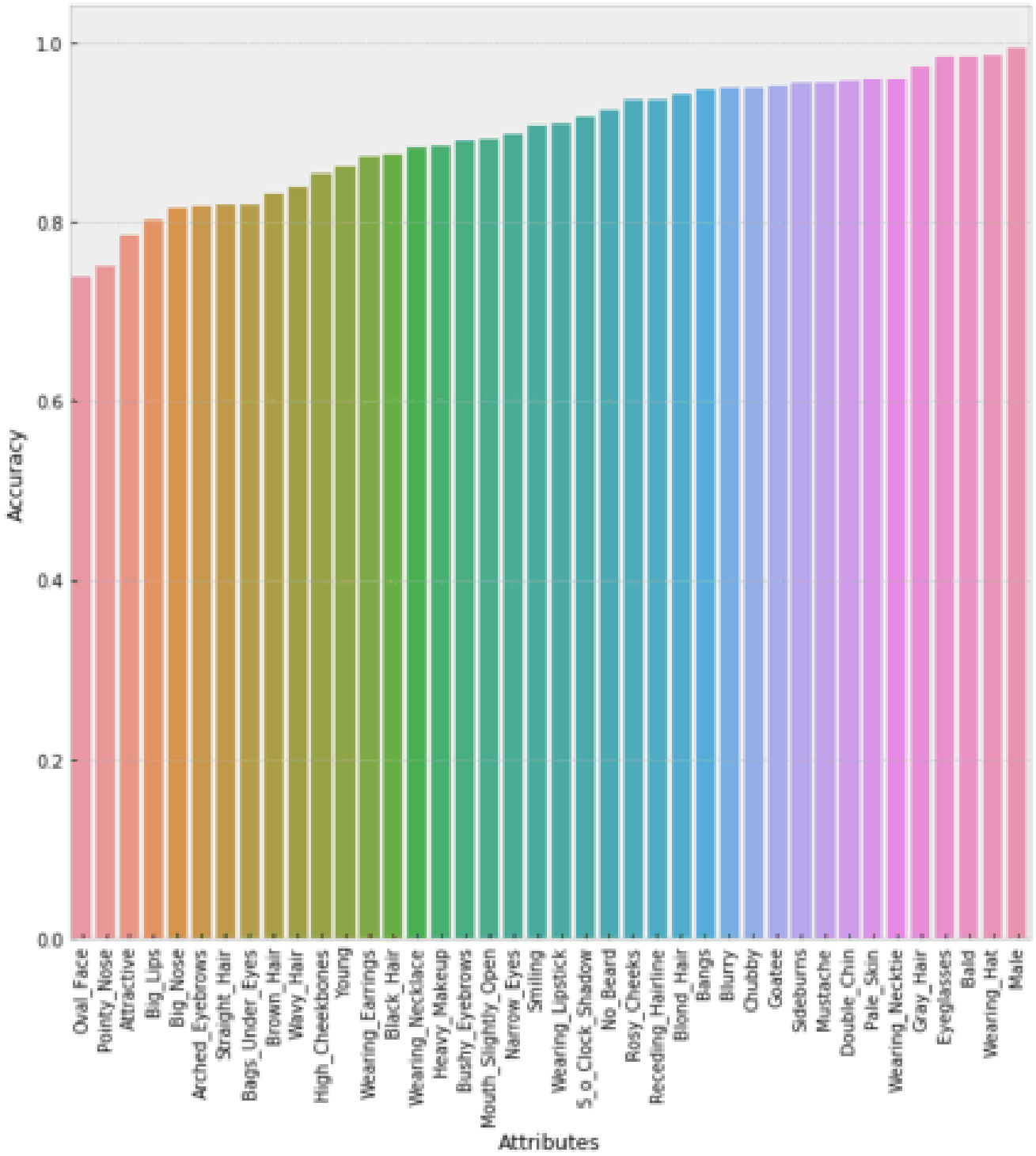

Figure 14. Per-attribute accuracy histogram achieve by our attribute prediction model on **CelebA** validation split.

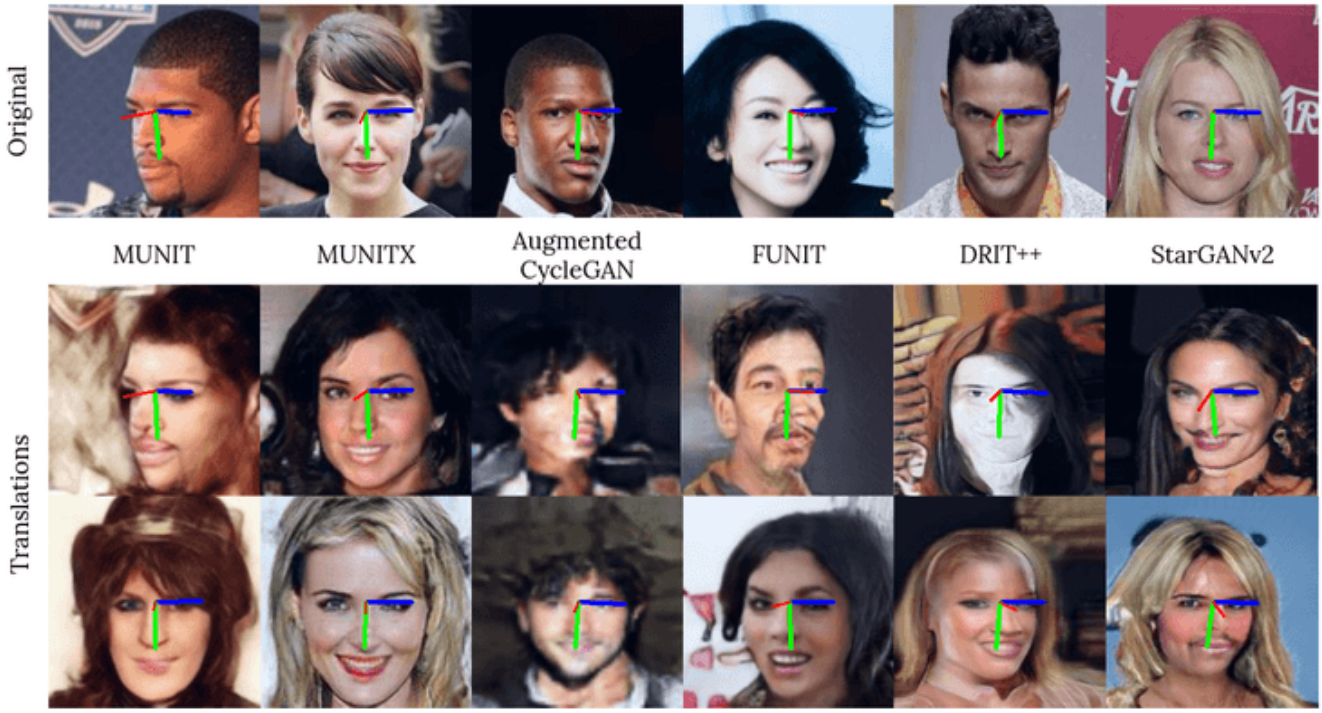

Figure 15. Head pose estimation results on random examples from the original CelebA dataset (**top**) and random translation results (**bottom**). Best viewed in color.

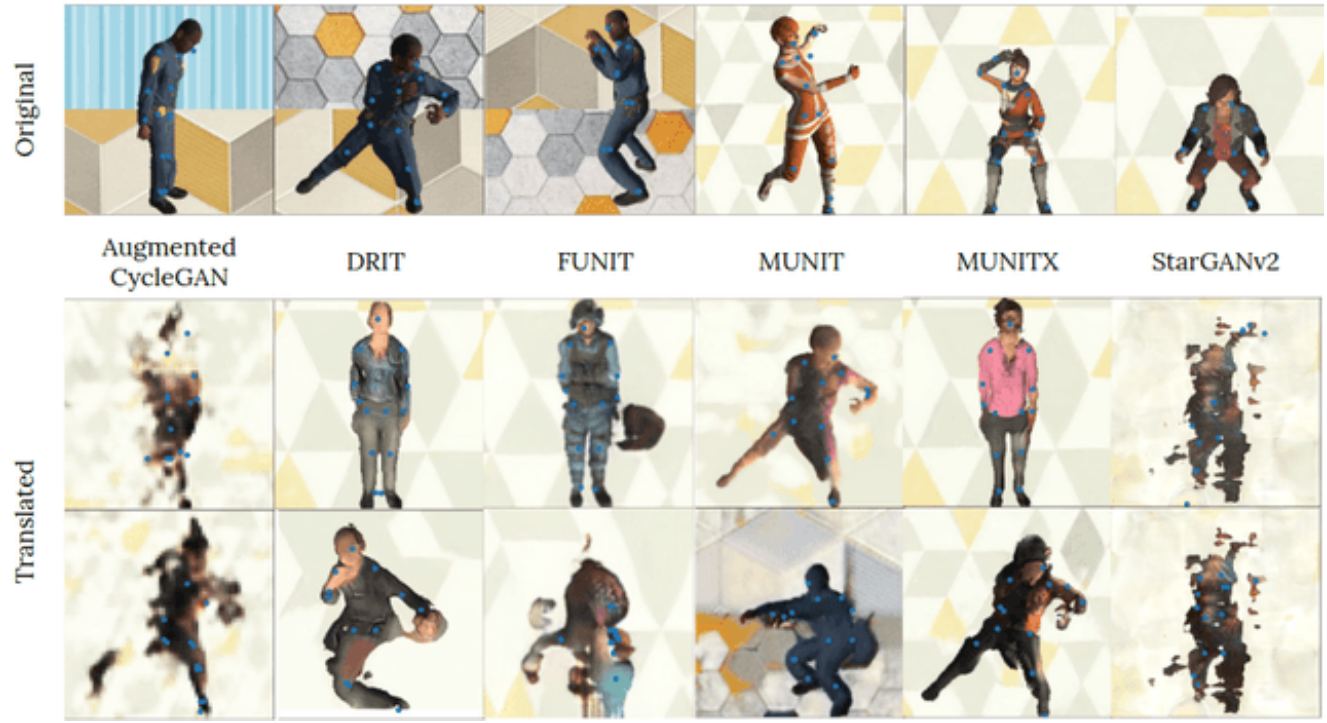

Figure 16. Pose estimation results on random examples from the original SynAction dataset (**top**) and random translation results (**bottom**). The pose estimation network succeeded in estimating poses even with severe generation artifacts. Best viewed in color.

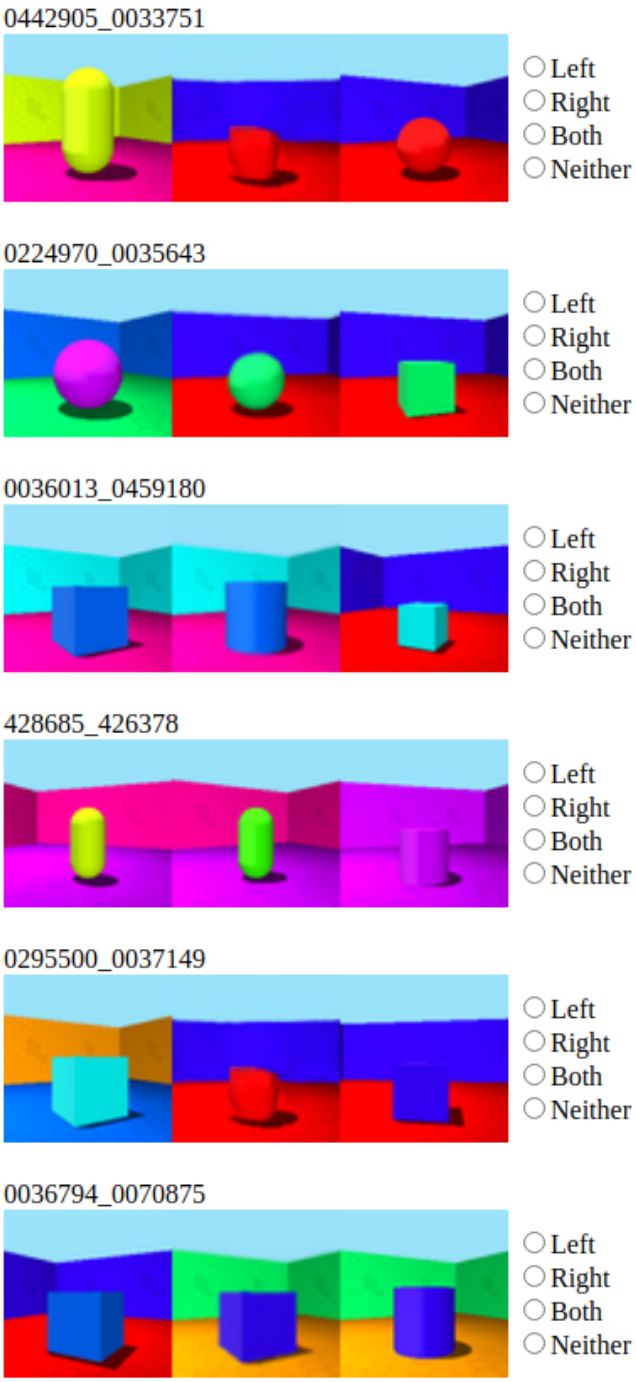

Figure 17. User study screen example. Subjects are explicitly asked to label the **center** images as having the specific attribute (e.g. shape color or view angle) as coming from either left, right, both images (if both have the same matching attribute value) or neither (if both images have attribute value not matching the center image). These responses were used to compute human evaluation metrics reported in Table 9 and show same trends as automatic evaluation reported in Table 2.