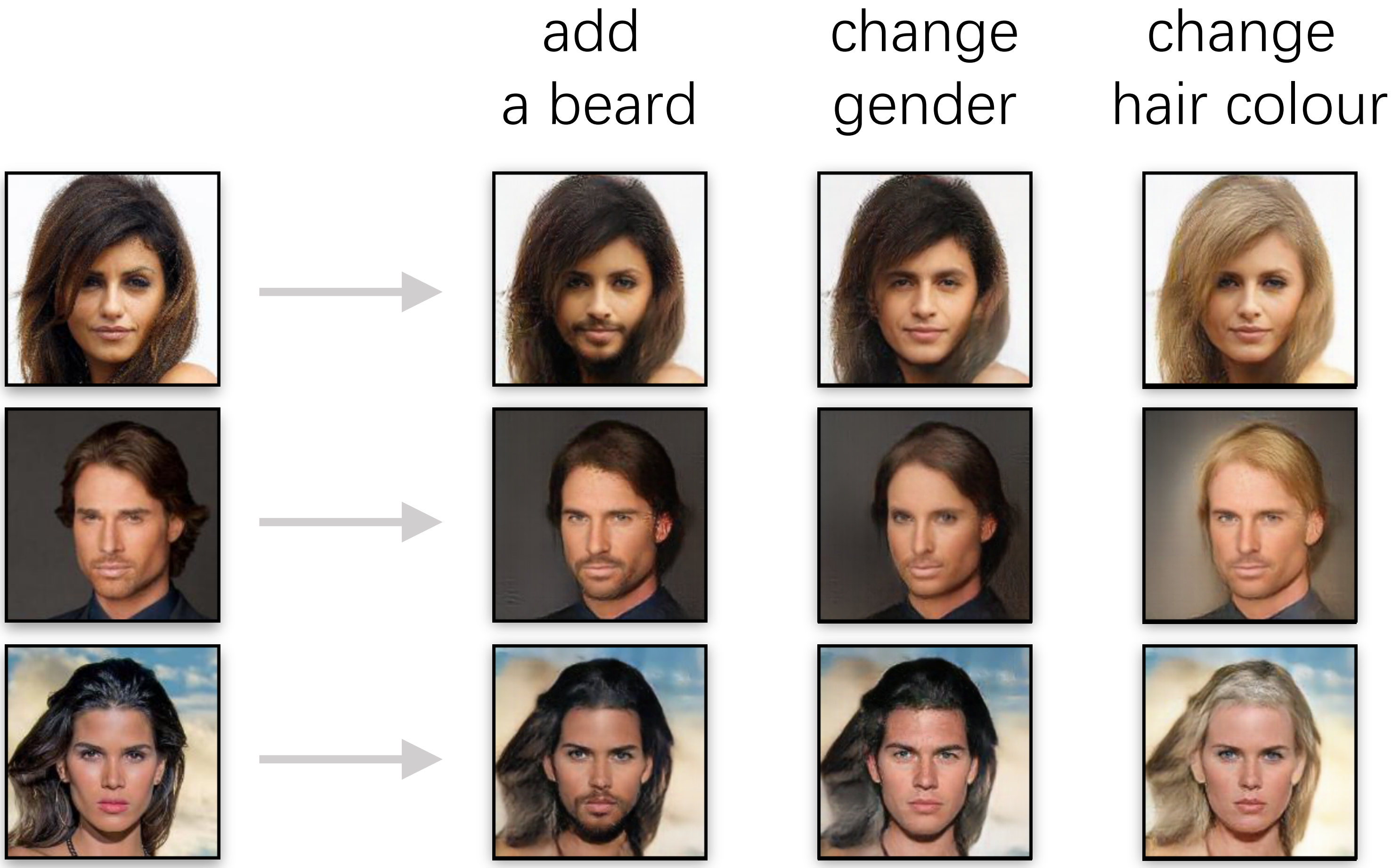
Latent Space Factorisation and Manipulation via **M**atrix **S**ubspace **P**rojection

Xiao Li   Chenghua Lin   Ruizhe Li
Chaozheng Wang   Frank Guerin

University of Aberdeen   The University of Sheffield   University of Surrey

The website of the paper: https://xiao.ac/proj/msp

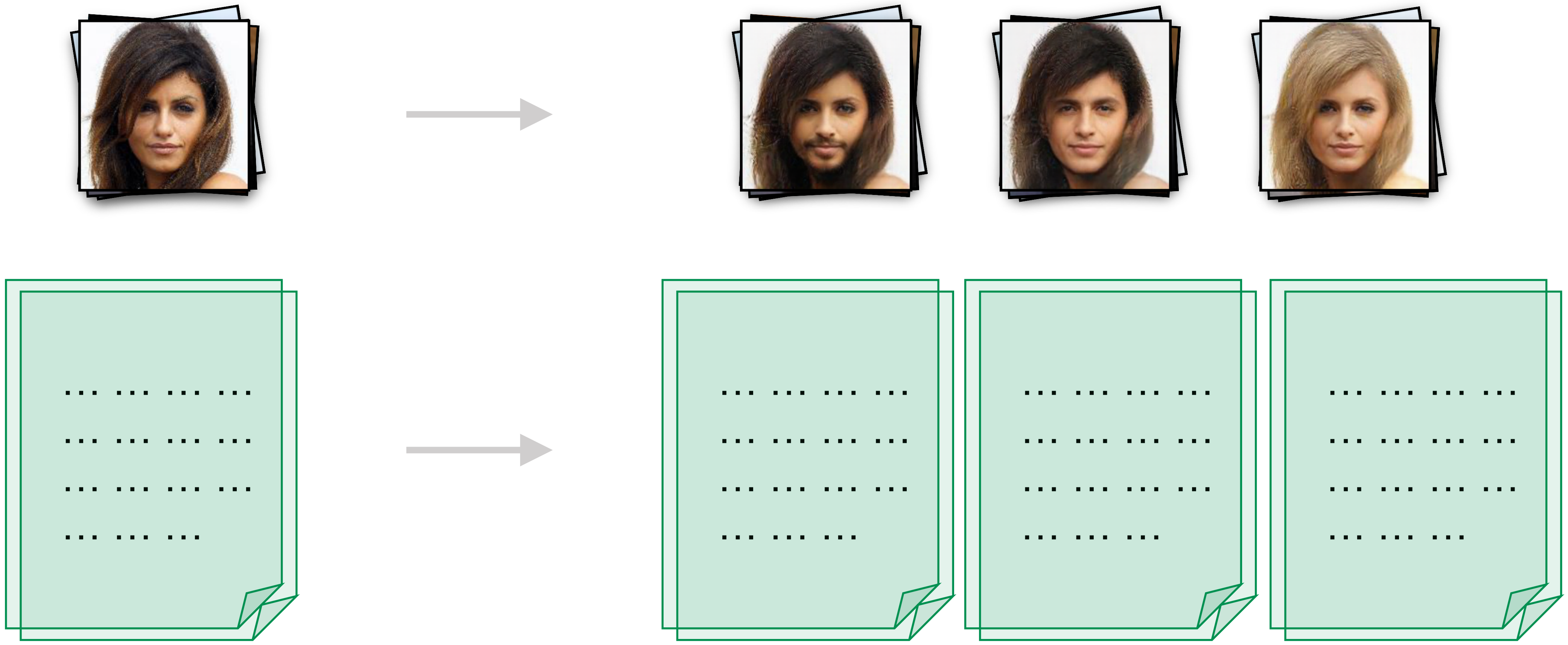
Manipulating the Attributes of Data
(image example)



Manipulating the Attributes of Data
(text example)

	set `EatType = coffee shop`	set `Near = Avalon`
the blue spice pub , near crowne plaza hotel ,has a customer rating of 5 out of 5 .	the blue spice **coffee shop** , near crowne plaza hotel ,has a customer rating of 5 out of 5 .	the blue spice pub , near **avalon** ,has a customer rating of 5 out of 5 .

Manipulating the Attributes of Data
(text example)



Training Dataset (CelebA)

Examples

Label:

`gender=female` `beard=false` `glasses=true`
`hair=blond` `smiling=true` …

Label:

`gender=male` `beard=true` `glasses=true`
`hair=black` `smiling=true` …

Label:

`gender=female` `beard=false` `glasses=false`
`hair=brown` `smiling=false` …

**Each pic has 40 labels*

Training Dataset (CelebA.)

Examples

Label: `gender=female and earring=true`

Lot of

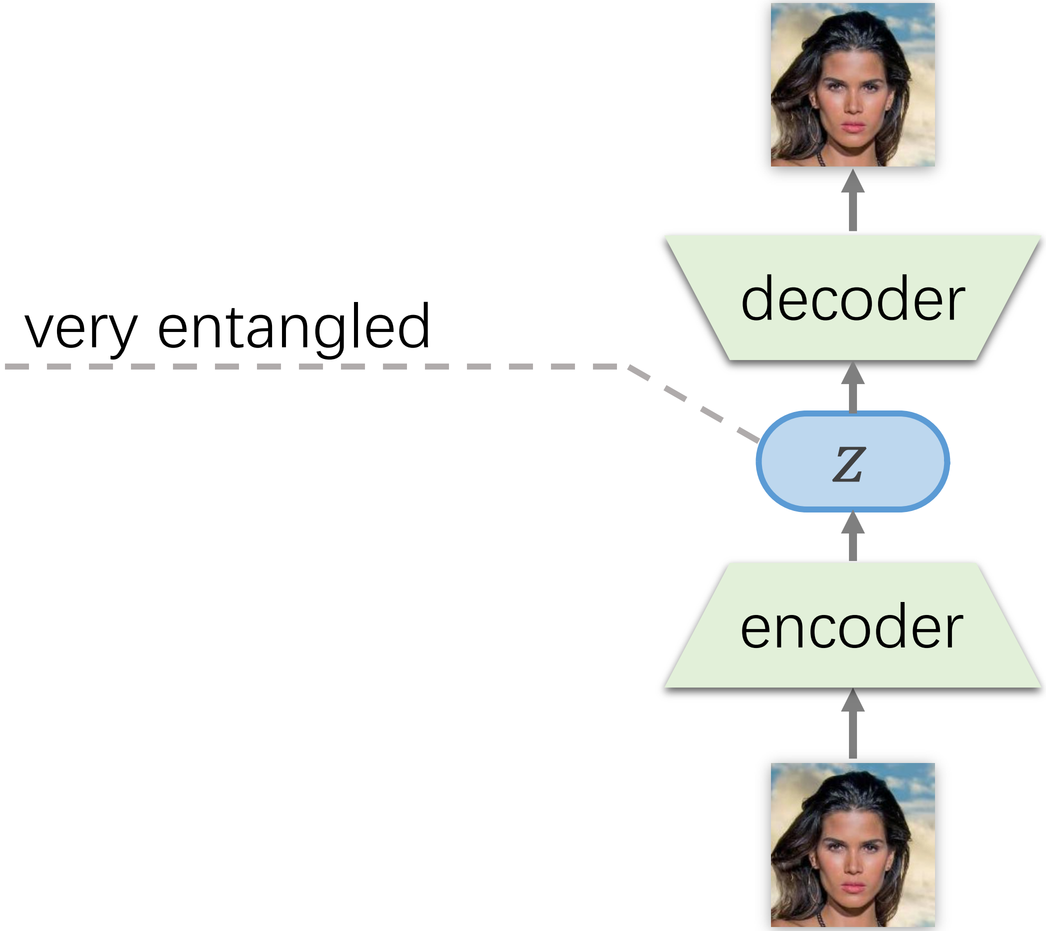
Label: `gender=male and earring=true`

Rare

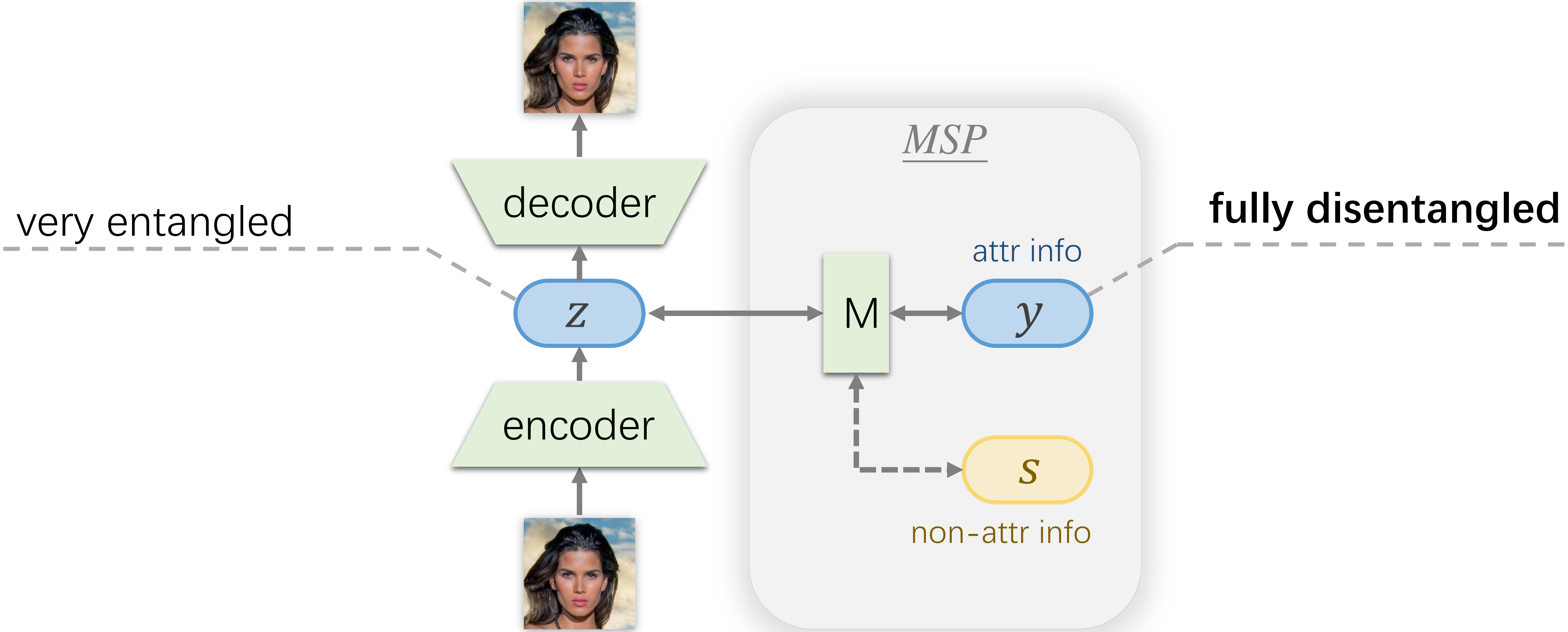
Label: `gender=female and beard=true`

Never seen!

A Typical Autoencoder (without MSP)

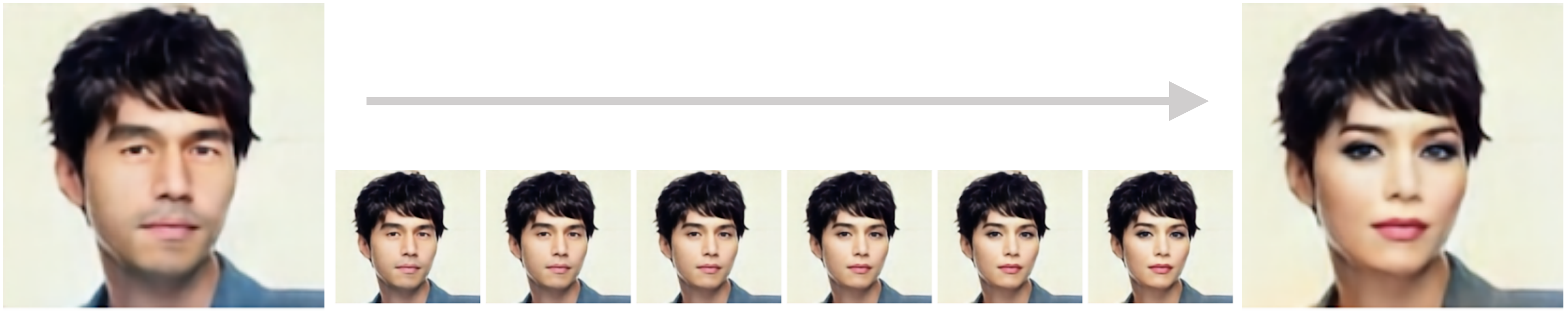


Autoencoder with MSP



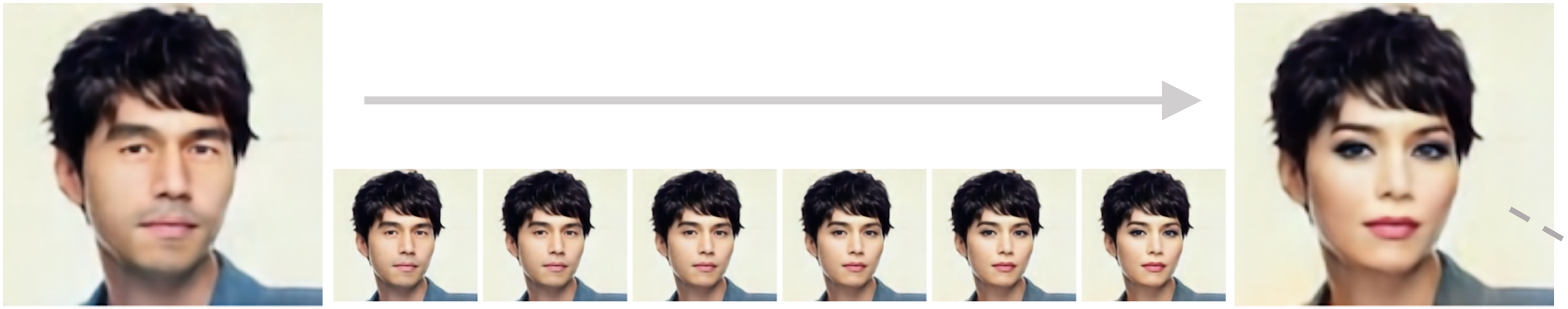
Related Work

Related Work



Lampe, G., Zeghidour, N., Usunier, N., Bordes, A., De-noyer, L., et al. Fader networks: Manipulating imagesby sliding attributes. InAdvances in Neural InformationProcessing Systems, pp. 5967–5976, 2017.

Related Work



`5 O'clock shadow` is removed and `make-up` is added!

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., De-noyer, L., et al. **Fader networks**: Manipulating imagesby sliding attributes. InAdvances in Neural InformationProcessing Systems, pp. 5967–5976, 2017.

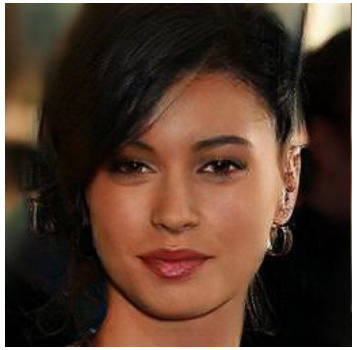
> “ Although Fader Networks is capable for multiple attribute editing with one model, in practice, multiple attribute setting makes the results blurry. ”
>
> -- He et al. (2019)

Related Work

Lampe, G., Zeghidour, N., Usunier, N., Bordes, A., De-noyer, L., et al. **Fader networks**: Manipulating imagesby sliding attributes. InAdvances in Neural InformationProcessing Systems, 2017.

5 O'clock shadow is removed and make-up is added!

Wu, P.-W., Lin, Y.-J., Chang, C.-H., Chang, E. Y., and Liao,S.-W. **Relgan**: Multi-domain image-to-image transla-tion via relative attributes. InThe IEEE InternationalConference on Computer Vision (ICCV), October 2019.



significant changes in skin colour, eyebrows, eyes, and lips

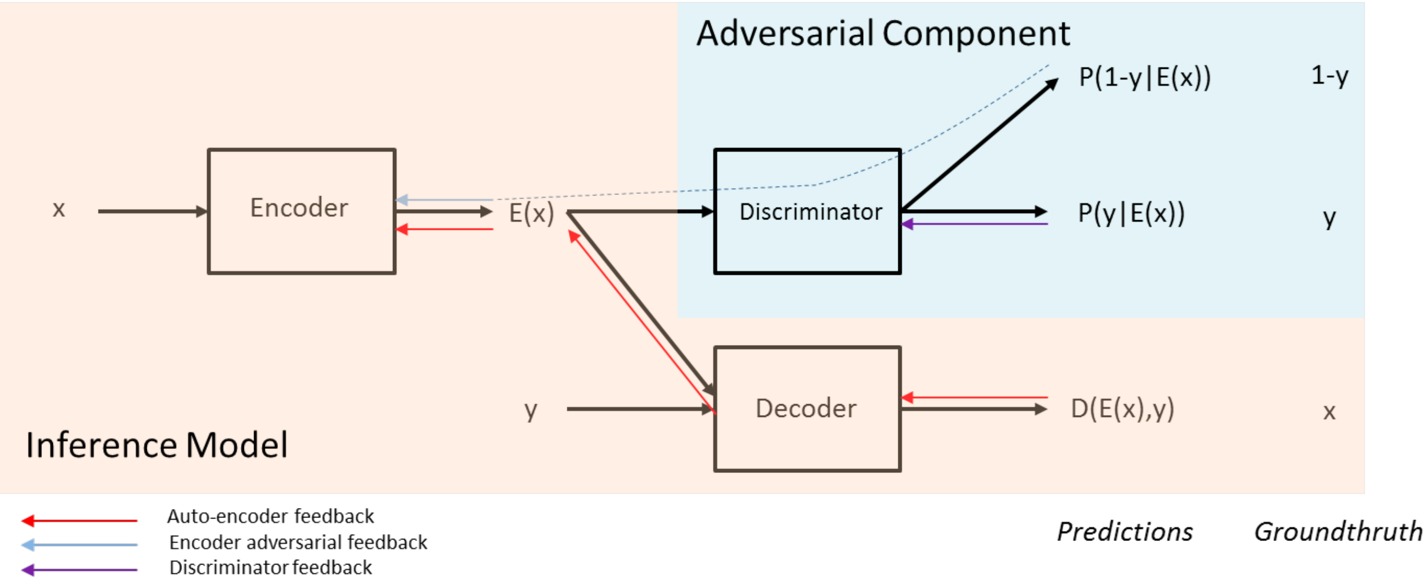
Related Work

the face has four eyebrows!

A further problem with many other works that use skip connections.

The face was changed from female to male. New bushy male eyebrows were added, but the skip connections also preserved the original feminine eyebrows in their original position.

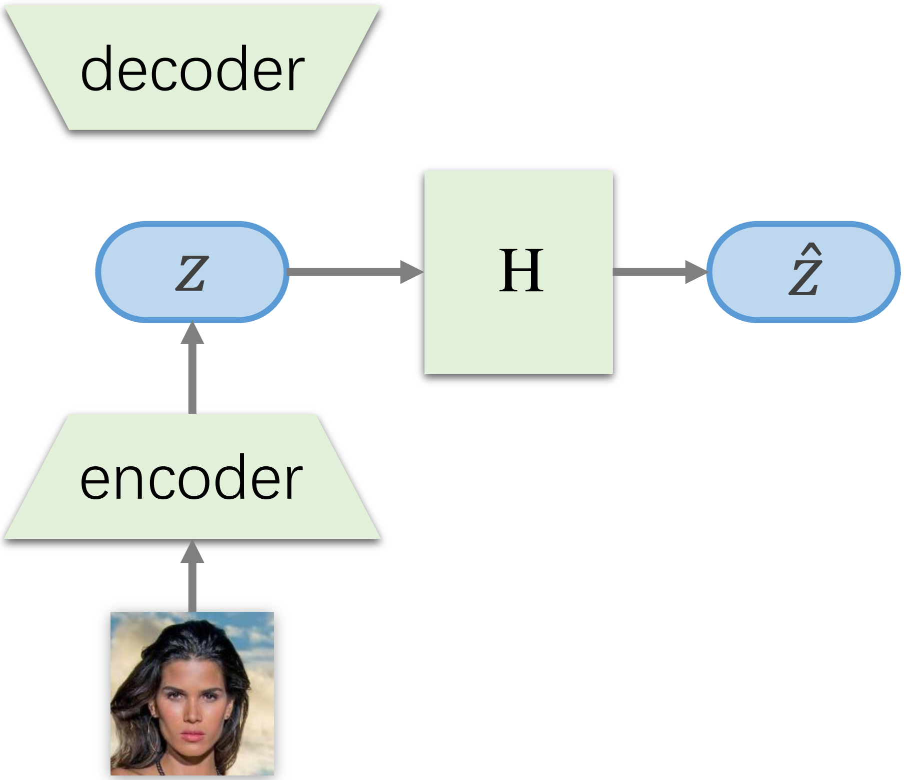
Related Work



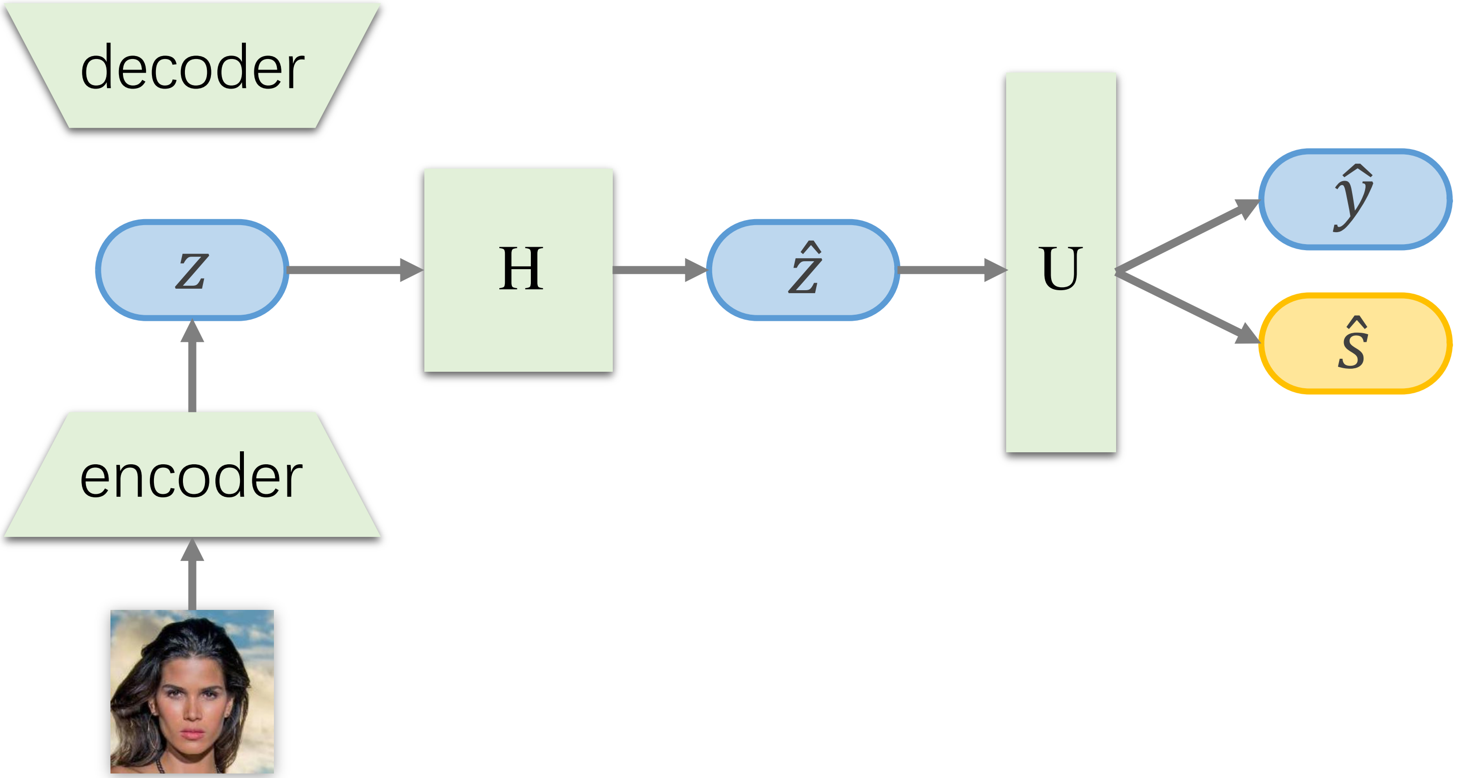
Lample, G., Zeghidour, N., Usunier, N., Bordes, A., De-noyer, L., et al. Fader networks: Manipulating imagesby sliding attributes. InAdvances in Neural InformationProcessing Systems, 2017.

Methodology

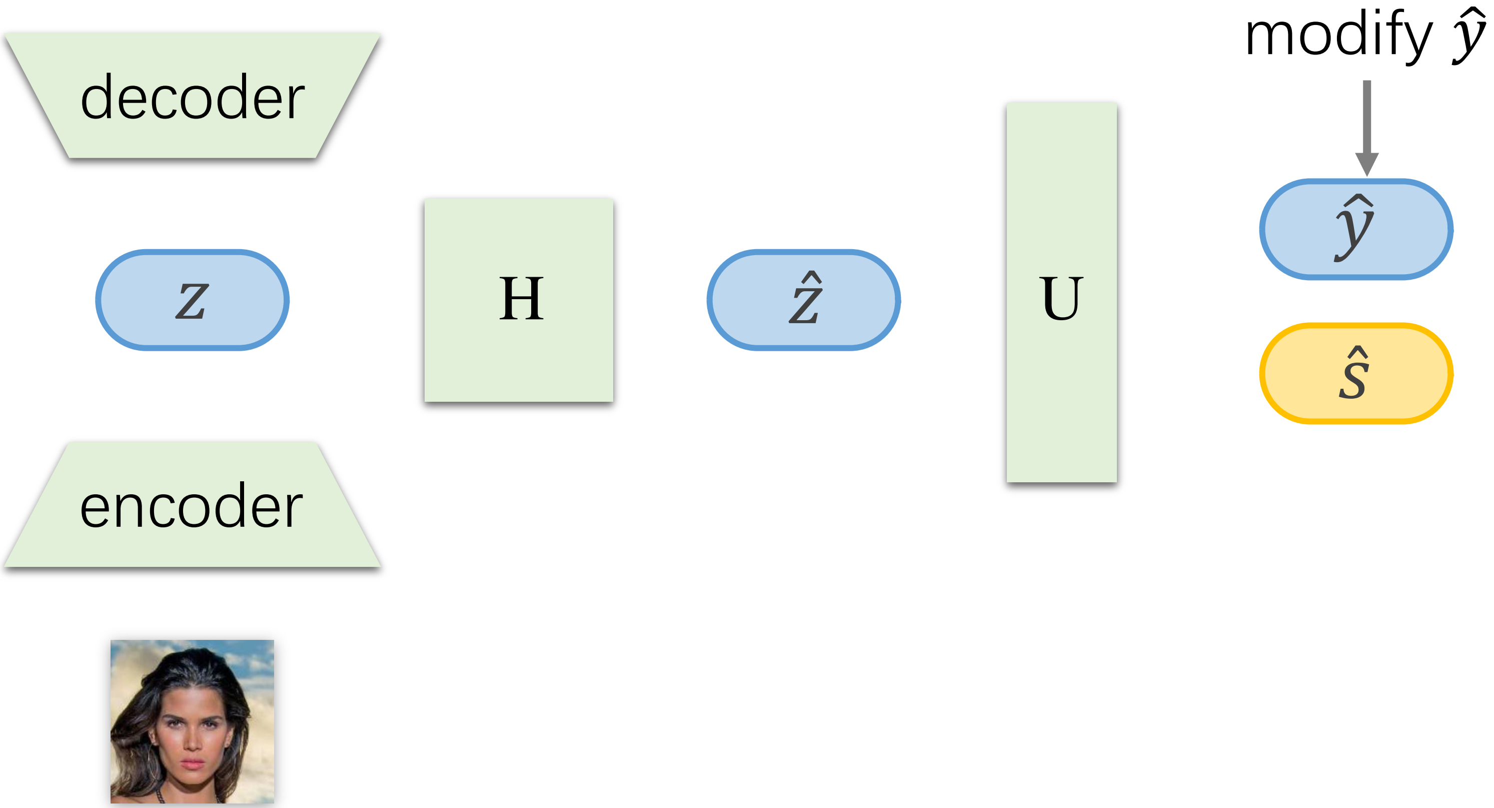
The Overall Workflow of MSP



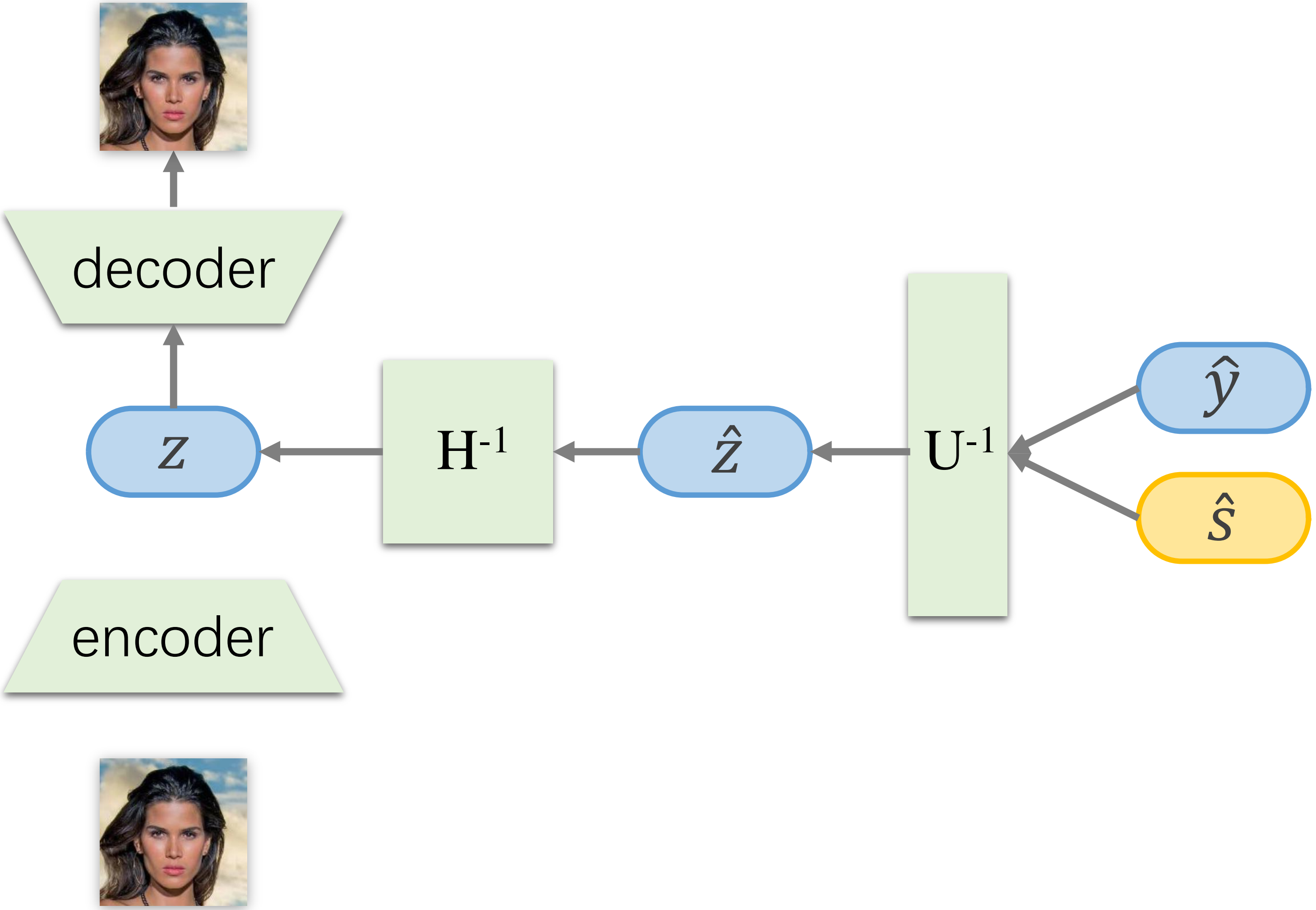
The Overall Workflow of MSP



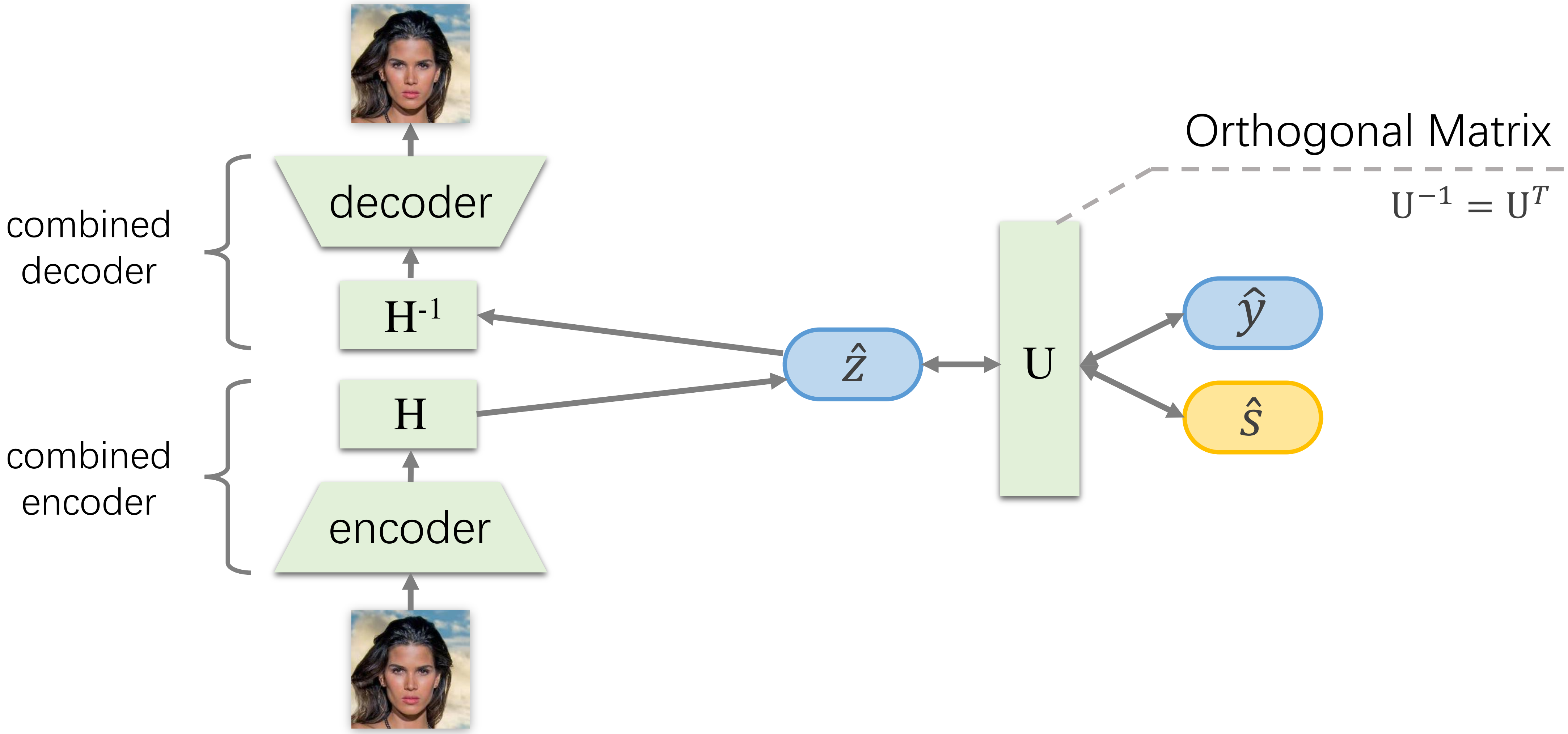
The Overall Workflow of MSP



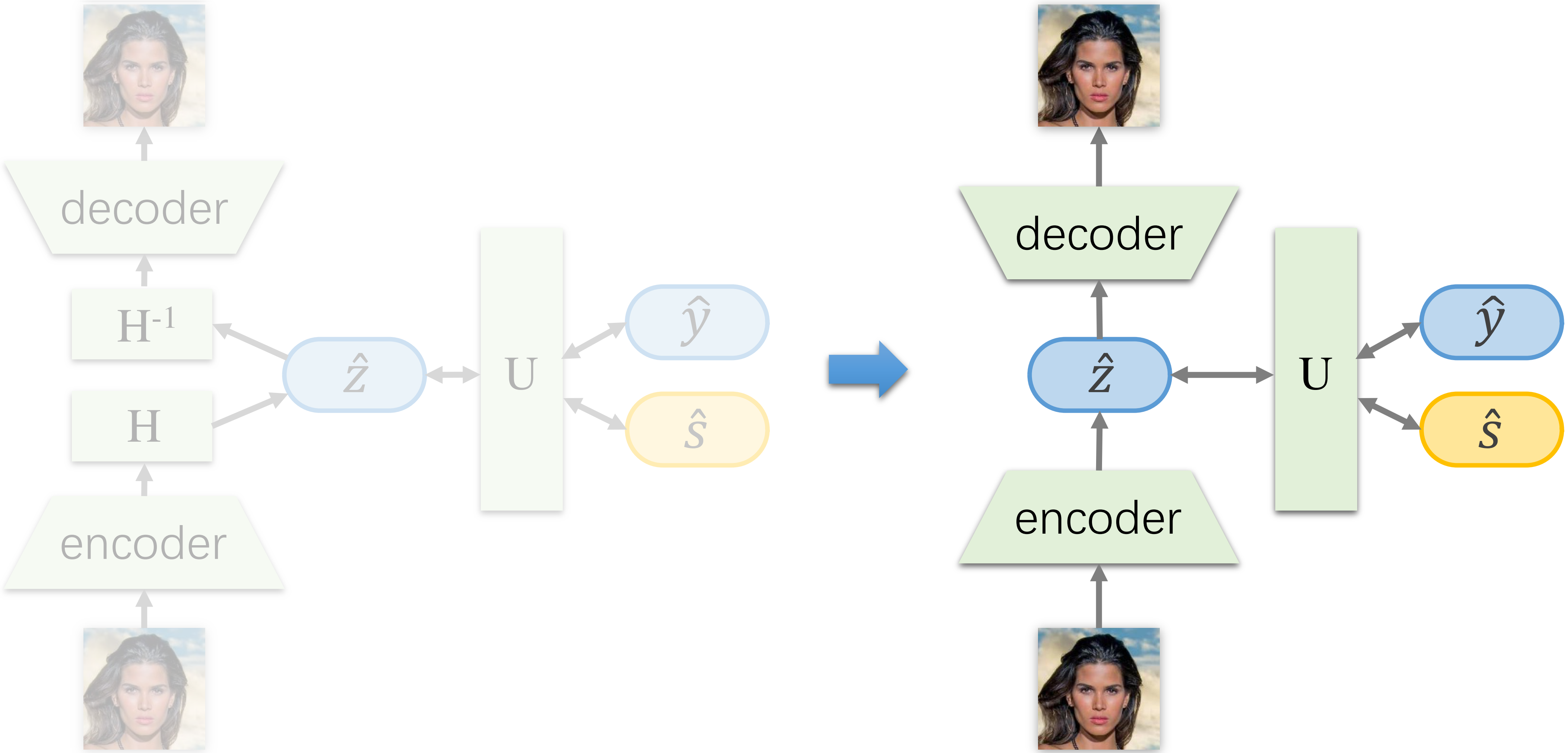
The Overall Workflow of MSP



Combined Encoder and Decoder



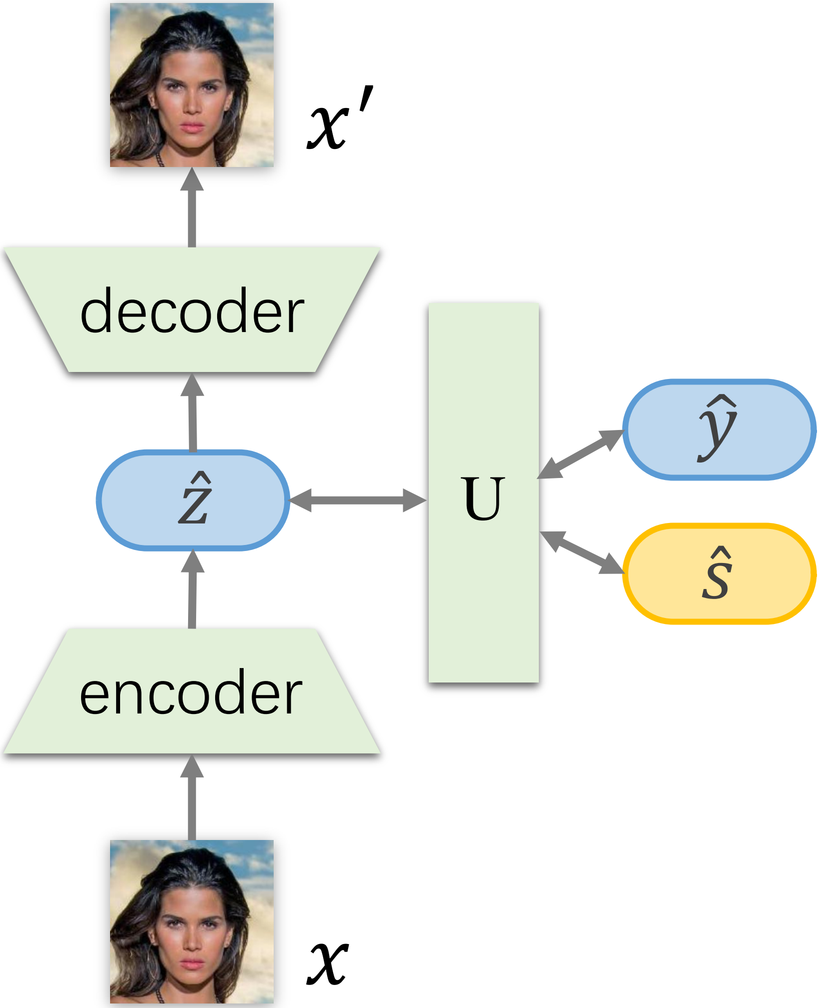
Combined Encoder and Decoder



Loss Function

- The produced image should be close to the input as much as possible:

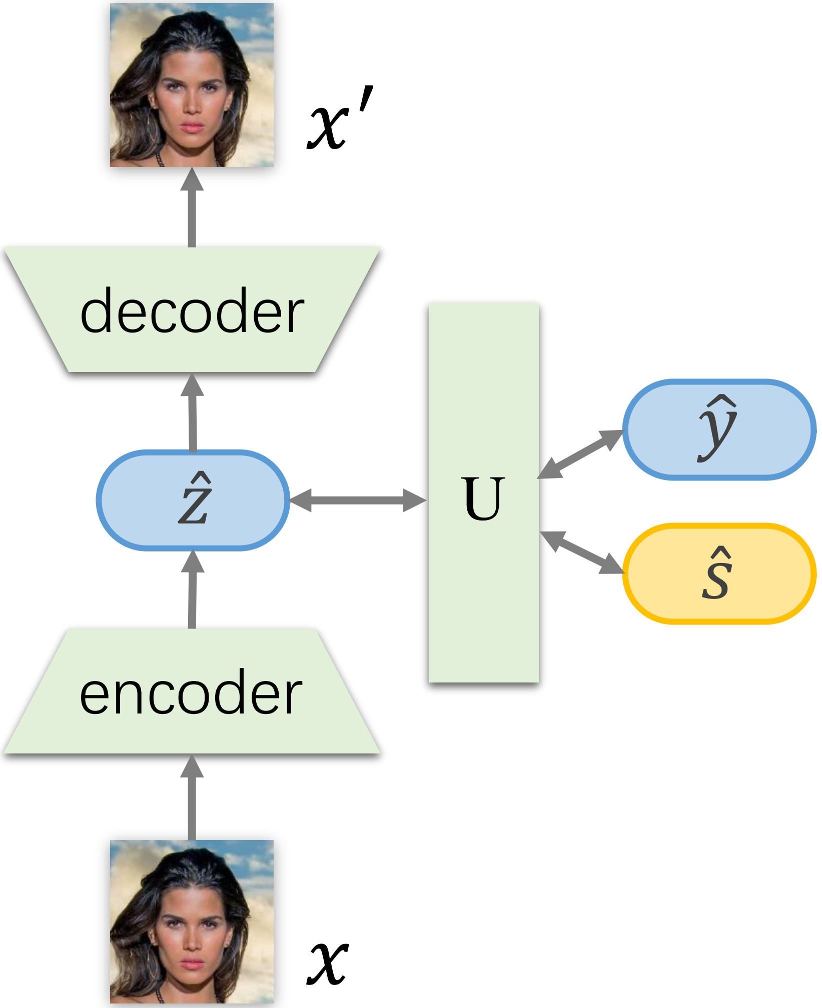
$$\mathcal{L}_{AE} = \|x' - x\|^2$$



Loss Function

- The predicted attributes should be close to the given attributes:

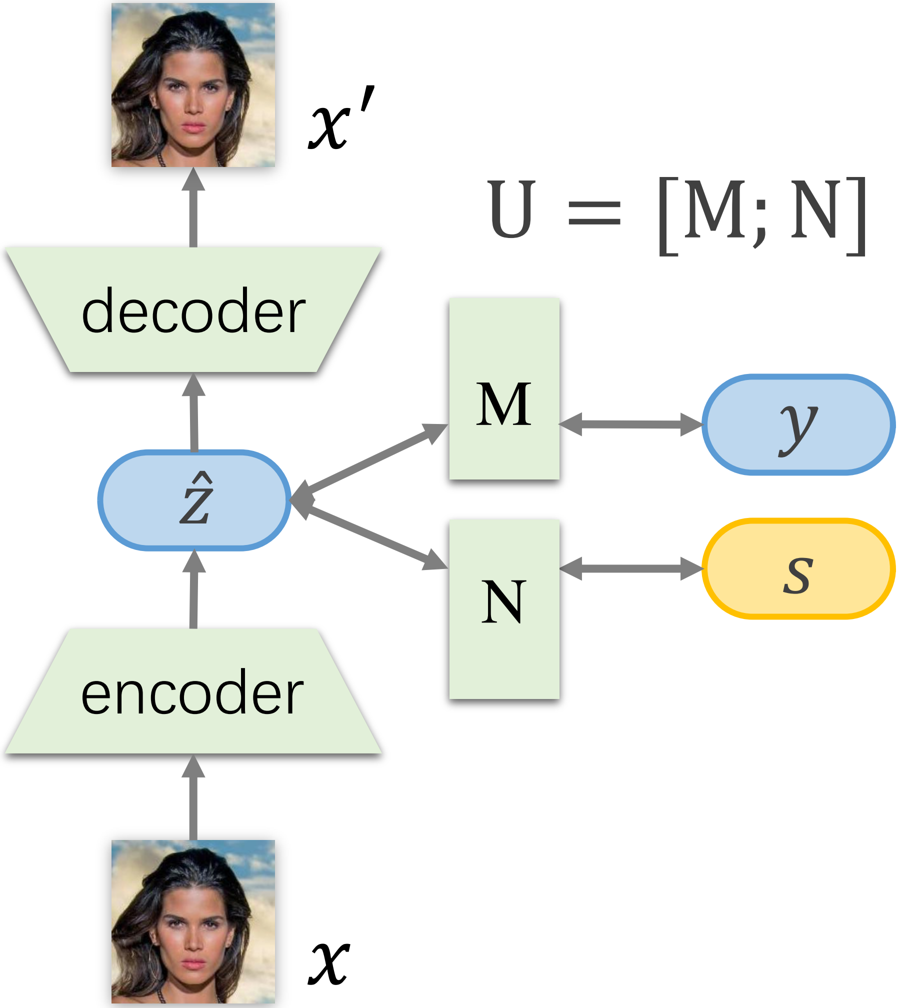
$$\mathcal{L}_1 = ||\hat{y} - y||^2$$



Loss Function

- The predicted attributes should be close to the given attributes:

$$\mathcal{L}_1 = ||\hat{y} - y||^2$$
$$= ||\mathrm{M} \cdot \hat{z} - y||^2$$



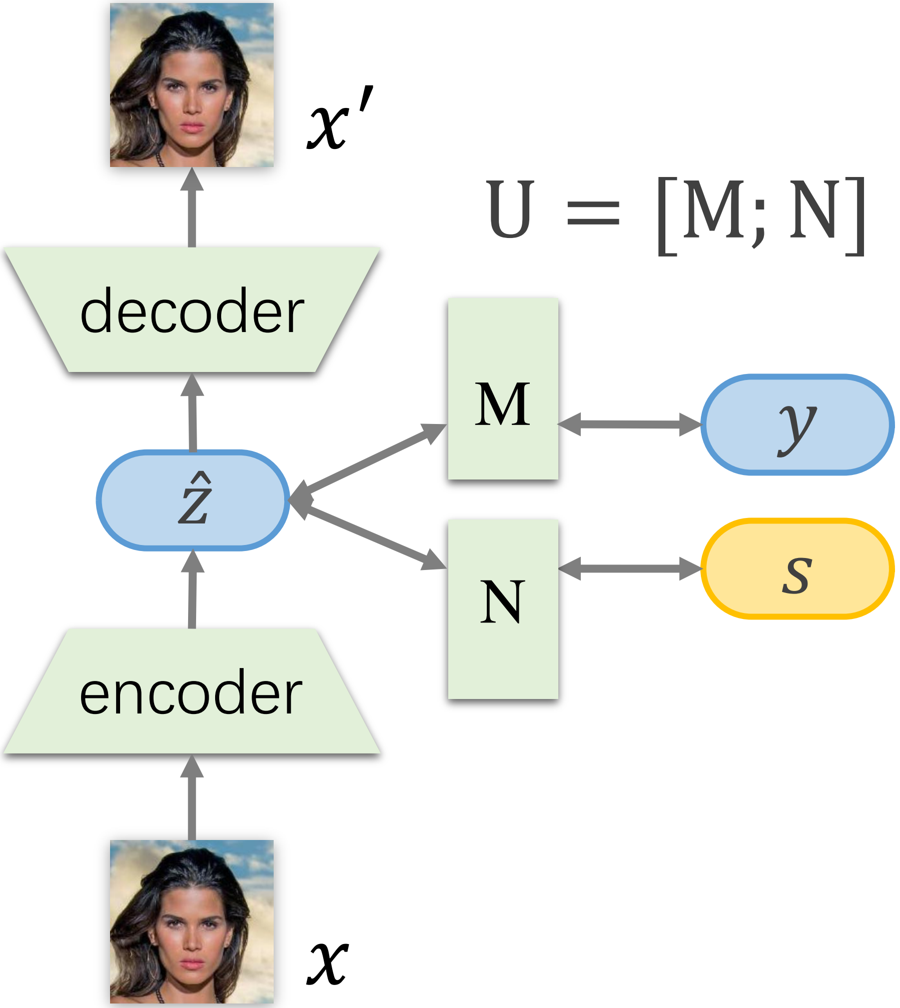
Loss Function

- $\hat{s}$ contains as little information from $\hat{z}$ as possible:

$$\mathcal{L}_2 = \|\hat{s}\|^2$$
$$= \|\hat{s} - 0\|^2$$
$$= \|[\hat{y}; \hat{s}] - [\hat{y}; 0]\|^2$$
$$= \|\mathrm{U} \cdot \hat{z} - [\hat{y}; 0]\|^2$$

- When U is orthogonal (more later):

$$\mathcal{L}_2 = \left\|z - \mathrm{M}^{\mathrm{T}} \cdot \hat{y}\right\|^2 \approx \left\|z - \mathrm{M}^{\mathrm{T}} \cdot y\right\|^2$$



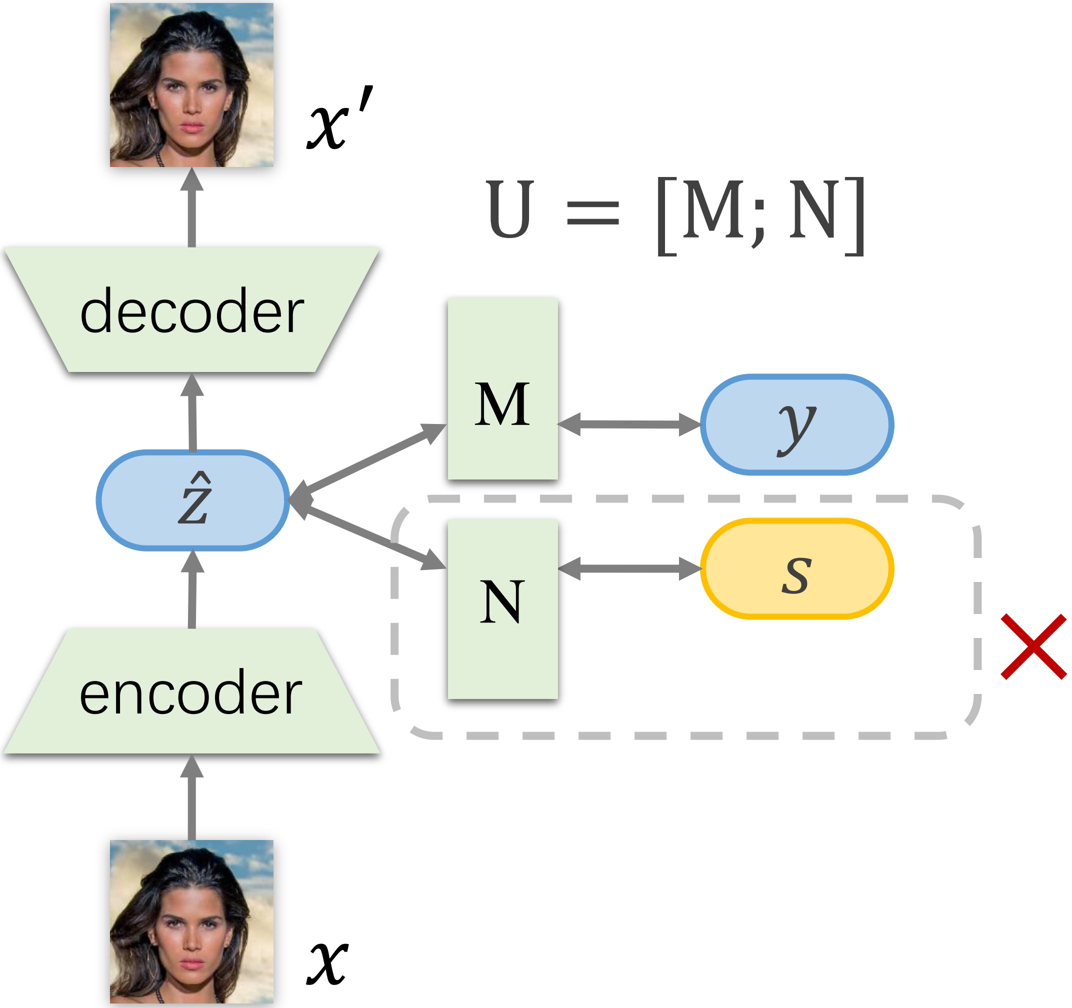
Loss Function

- The total loss is:

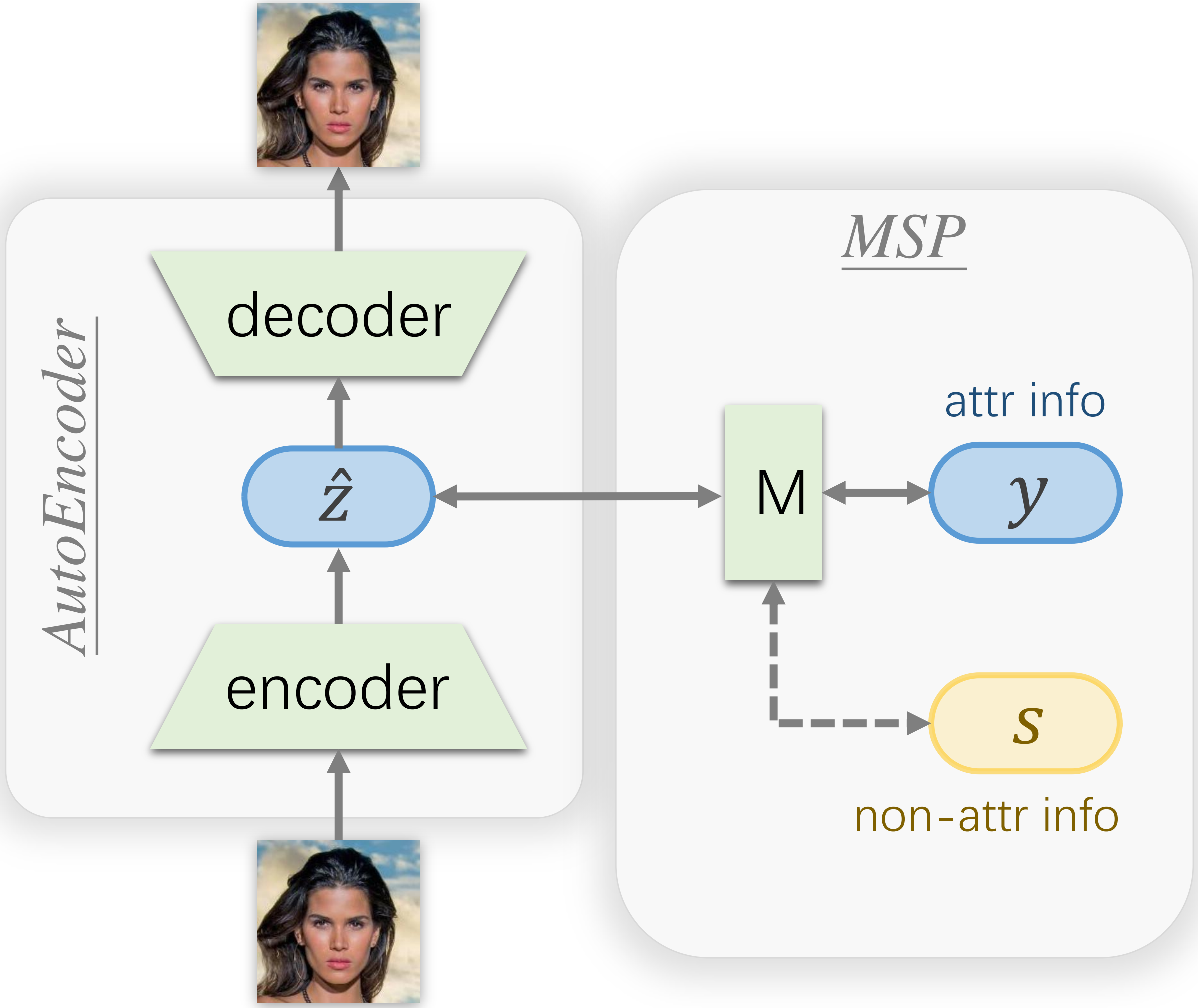
$$\mathcal{L} = \mathcal{L}_{AE} + \mathcal{L}_1 + \mathcal{L}_2$$

$$= \|x' - x\|^2$$
$$+ \|\mathrm{M} \cdot \hat{z} - y\|^2$$
$$+ \left\|z - \mathrm{M}^{\mathrm{T}} \cdot \hat{y}\right\|^2$$

$\hat{s}$ does not appear in the total loss, so we don't need to learn N!



Autoencoder with MSP



Evaluation

Picture Interpolation Generated by MSP

- Gender - Beard interpolations

female without beard

↓

female with beard

↓

male with beard

↓

male without beard

↓

female without beard

Picture Interpolation Generated by MSP

- Mouth open - Smiles interpolations

mouth closed and no smile

↓

mouth open and no smile

↓

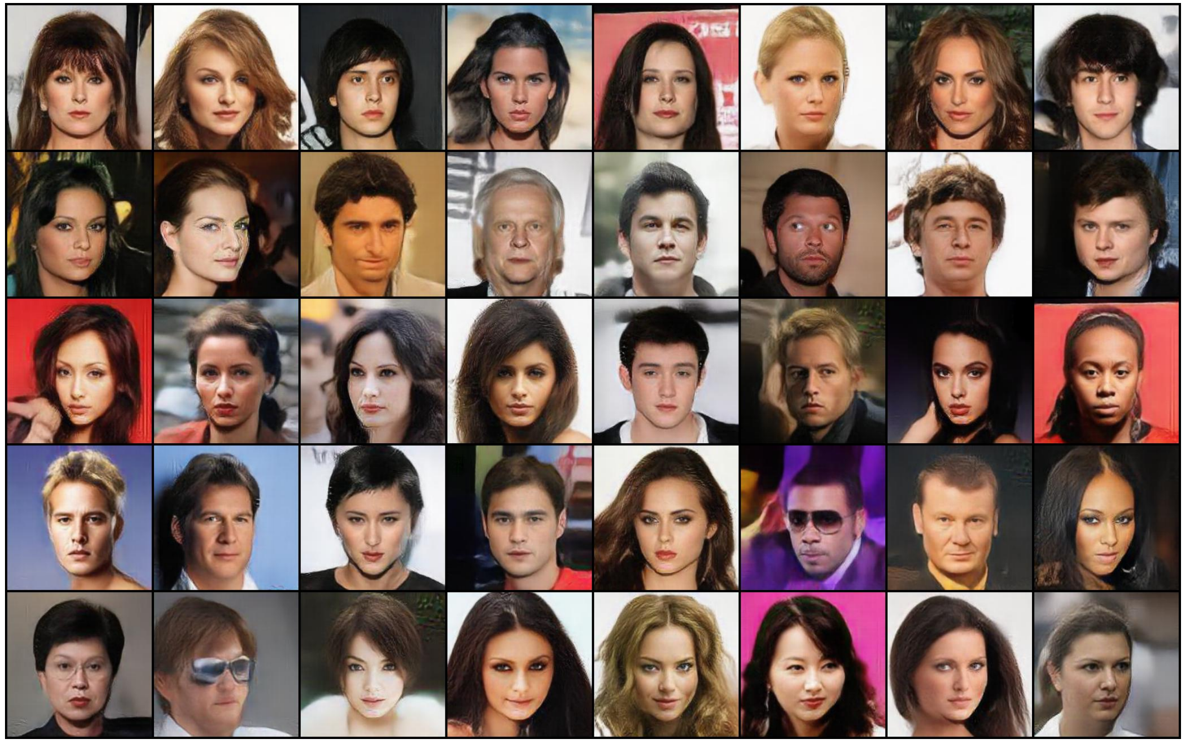
mouth open and smile

↓

mouth closed and smile

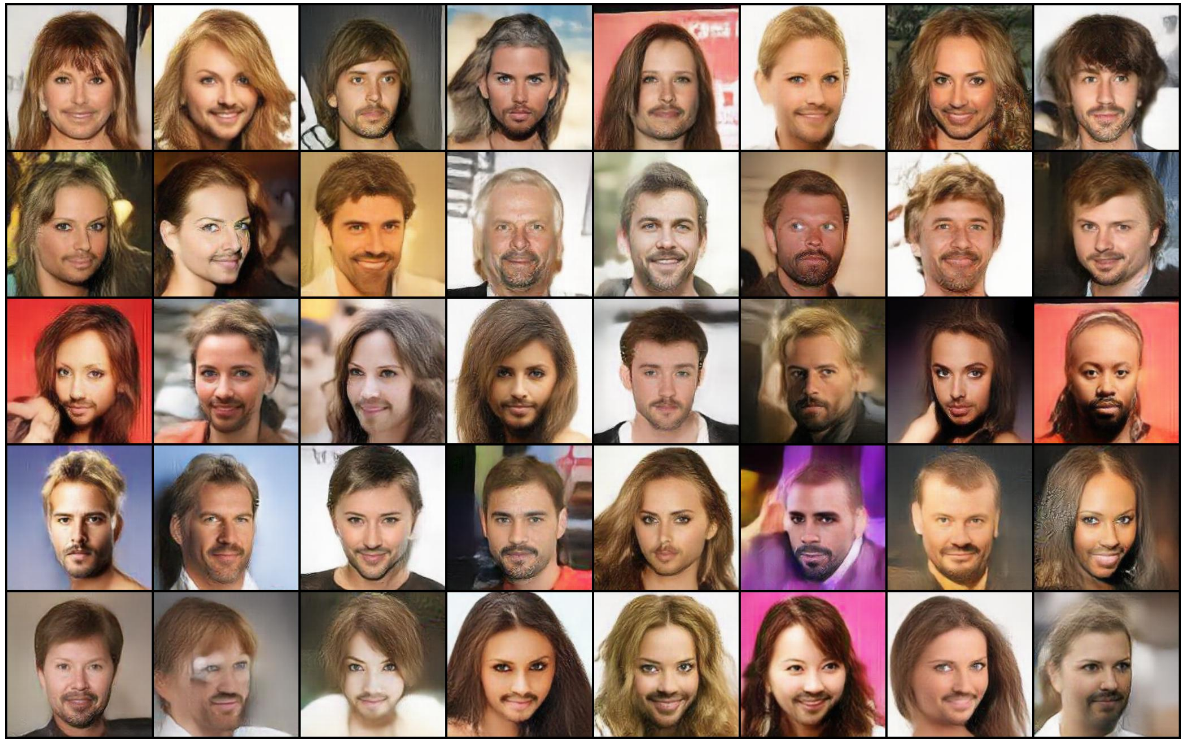
↓

mouth closed and no smile

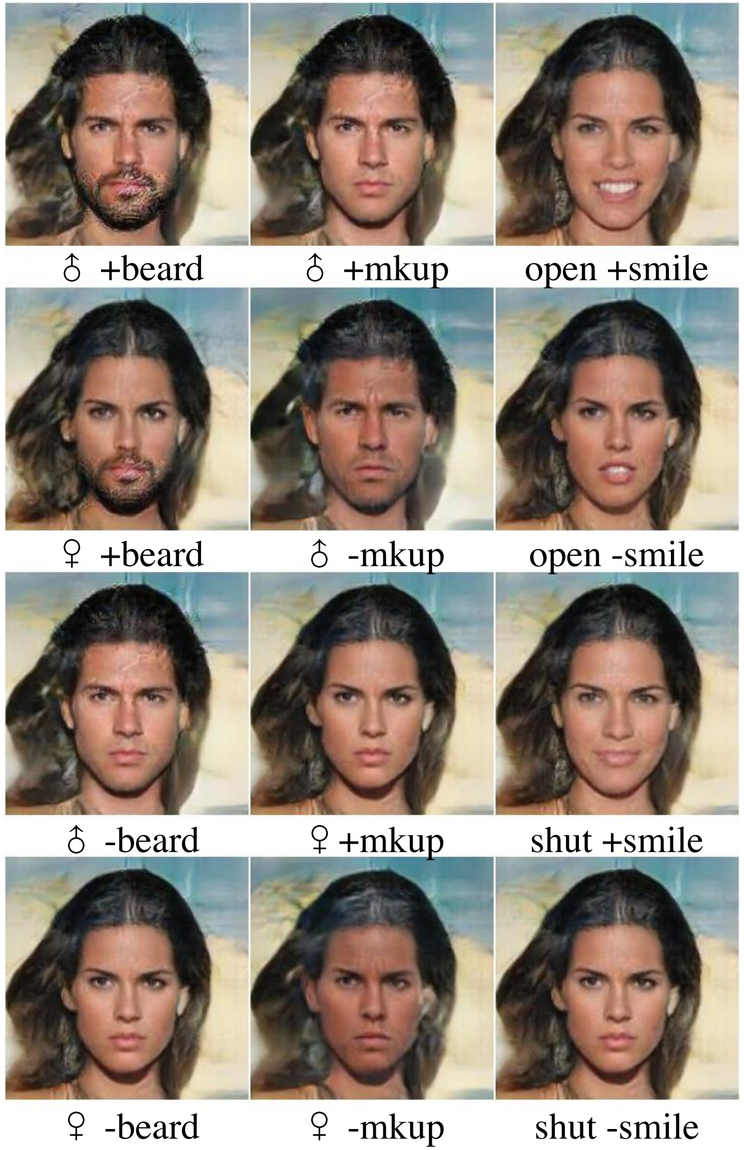


Picture Interpolation Generated by MSP

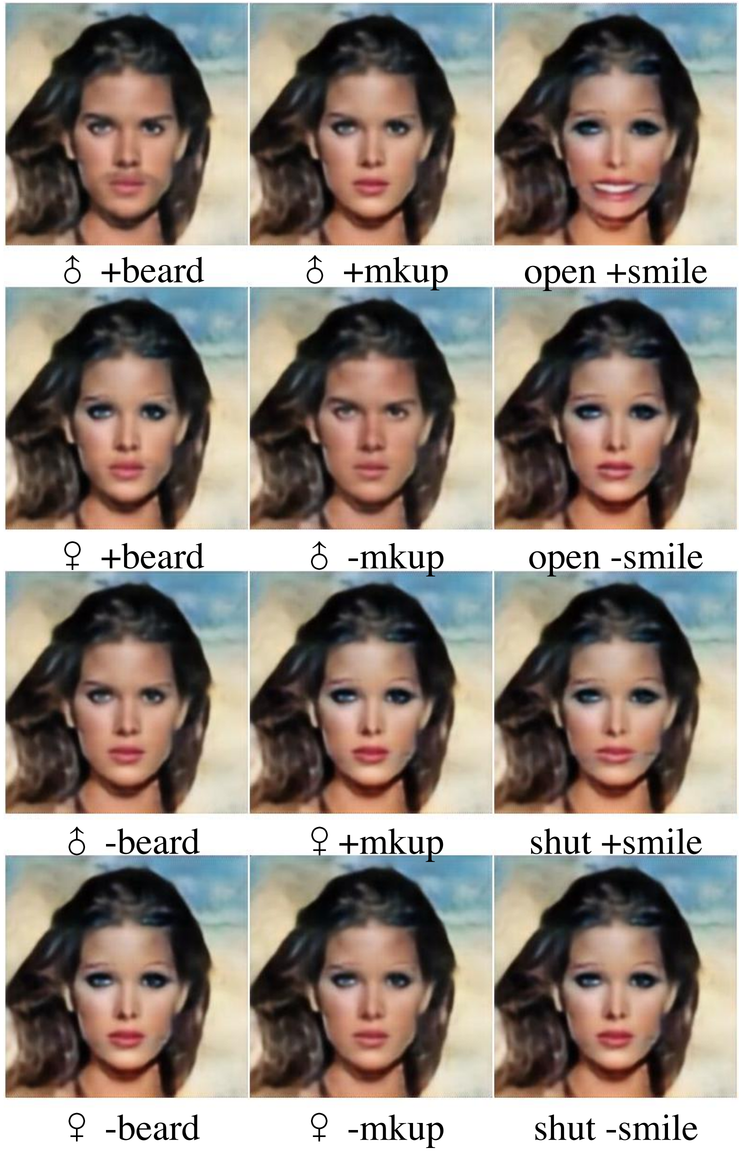
- Multiple attribute interpolations including:
  - glasses
  - beard
  - hair colour
  - narrow eyes
  - mouth open



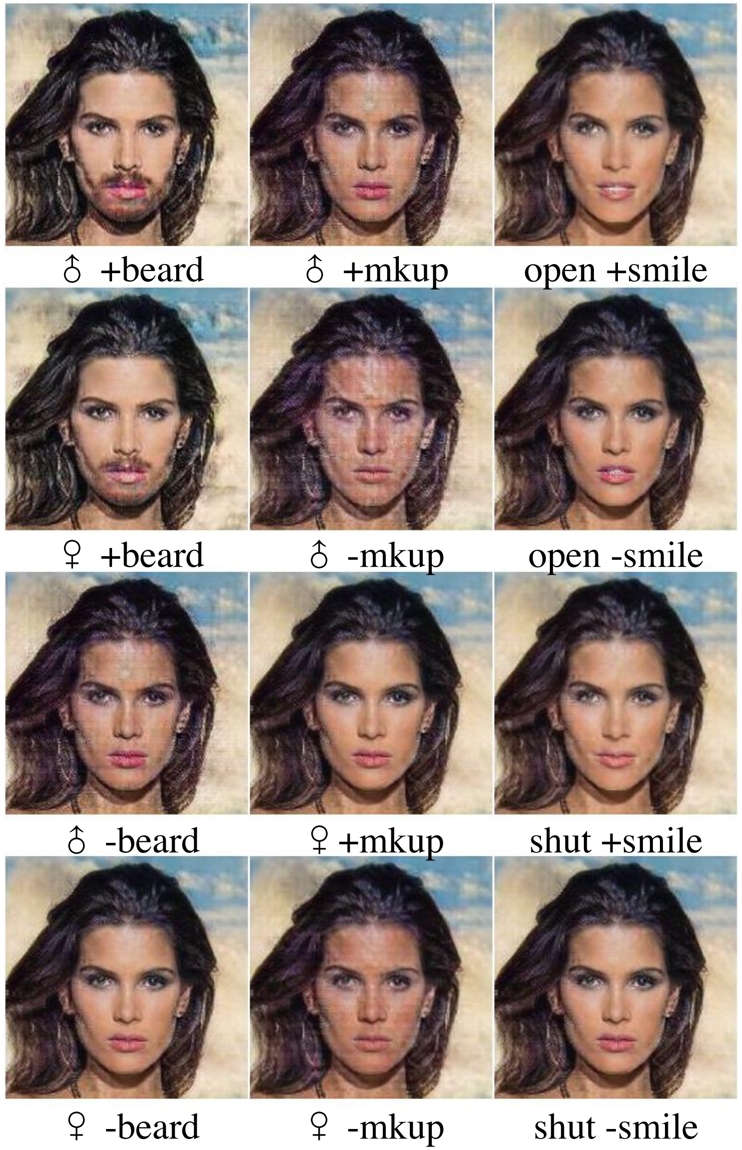
MSP



baseline
Fader Networks



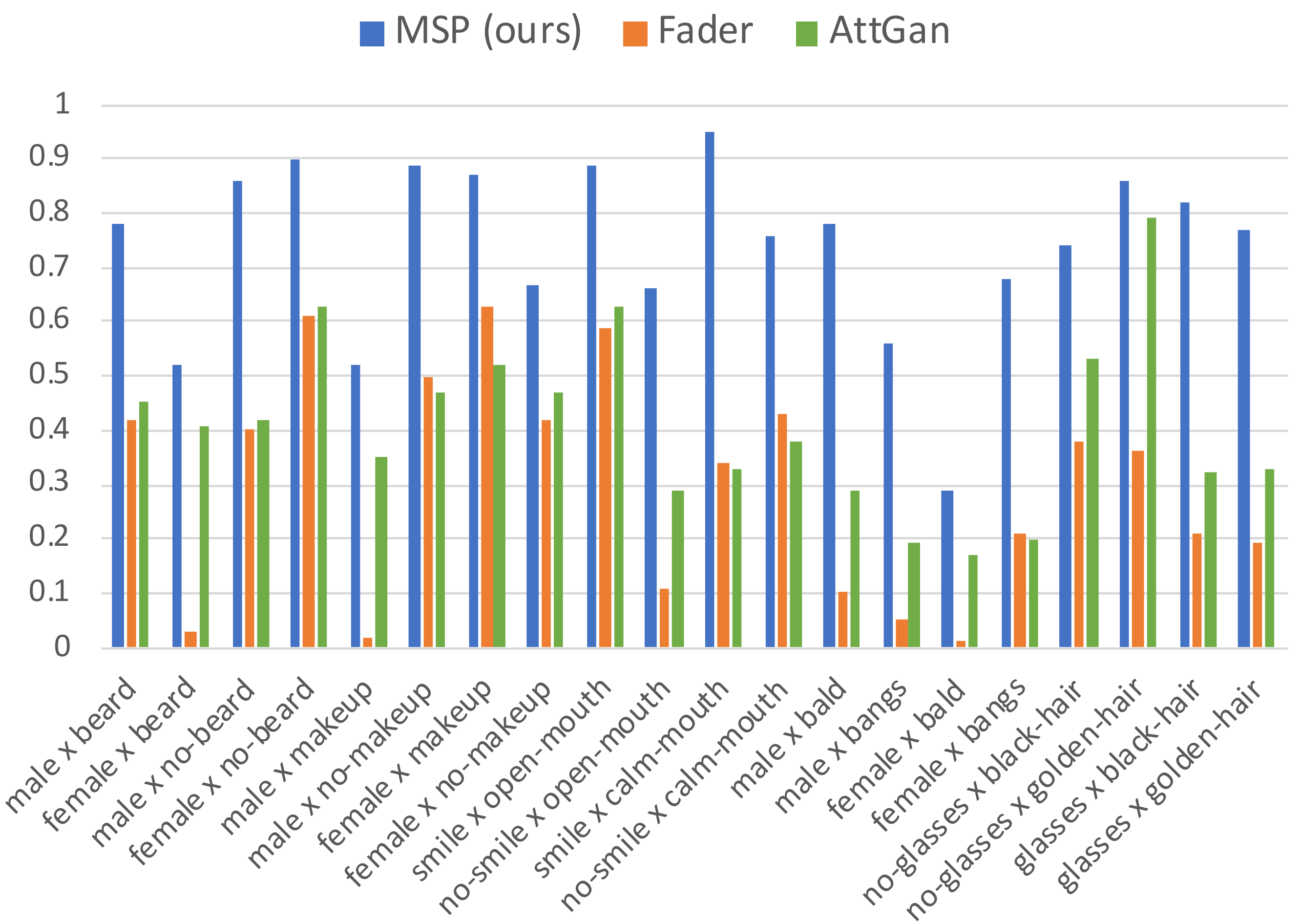
baseline
AttGan



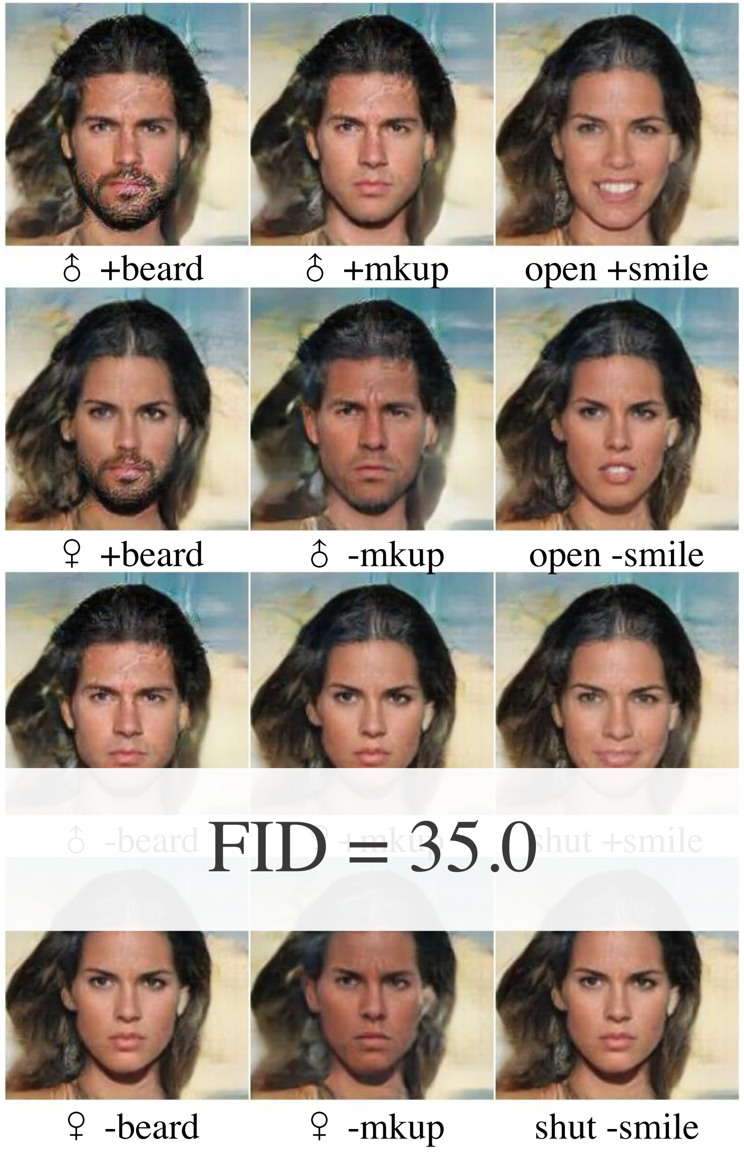
Quantitative Evaluation

	MSP(ours)	Fader	AttGAN
male x beard	**0.78**	0.42	0.45
female x beard	**0.52**	0.03	0.41
male x no-beard	**0.86**	0.40	0.42
female x no-beard	**0.90**	0.61	0.63
male x makeup	**0.52**	0.02	0.35
male x no-makeup	**0.89**	0.50	0.47
female x makeup	**0.87**	0.63	0.52
female x no-makeup	**0.67**	0.42	0.47
smile x open-mouth	**0.89**	0.59	0.63
no-smile x open-mouth	**0.66**	0.11	0.29
smile x calm-mouth	**0.95**	0.34	0.33
no-smile x calm-mouth	**0.76**	0.43	0.38
male x bald	**0.78**	0.10	0.29
male x bangs	**0.56**	0.05	0.19
female x bald	**0.29**	0.01	0.17
female x bangs	**0.68**	0.21	0.20
no-glasses x black-hair	**0.74**	0.38	0.53
no-glasses x golden-hair	**0.86**	0.36	0.79
glasses x black-hair	**0.82**	0.21	0.32
glasses x golden-hair	**0.77**	0.19	0.33

Table 2. The classification accuracy of generated images using MSP, Fader Networks and AttGAN.



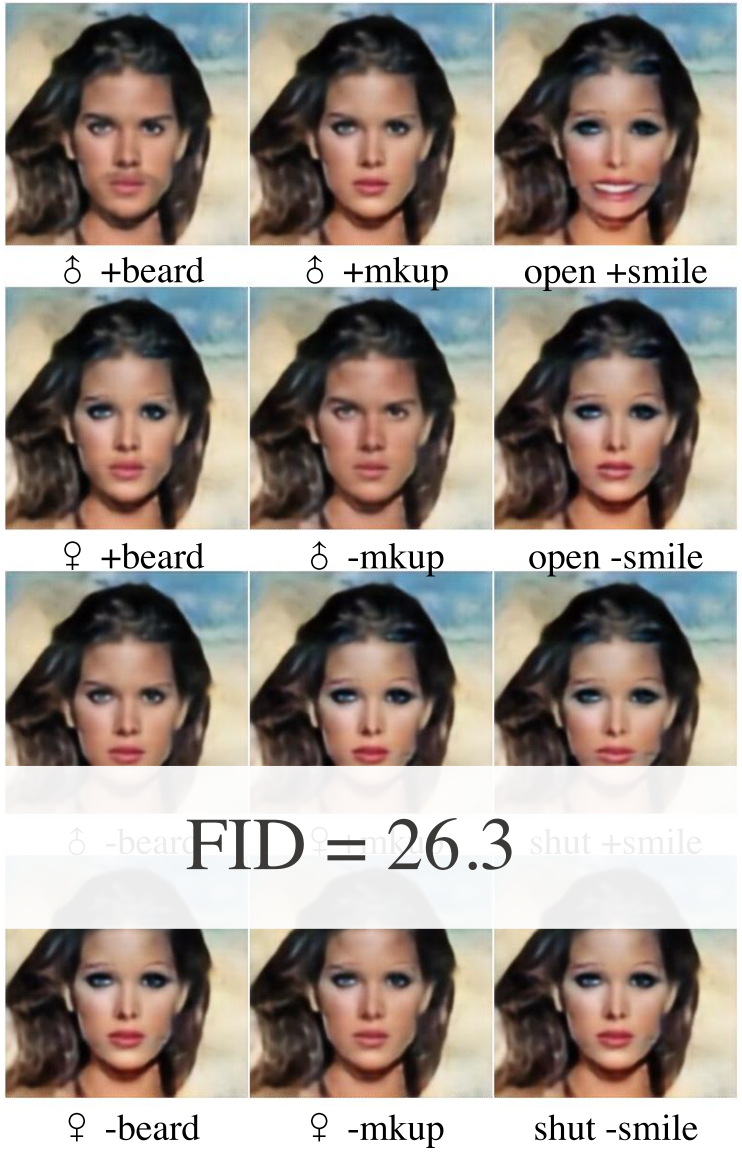
MSP



FID = 35.0

baseline

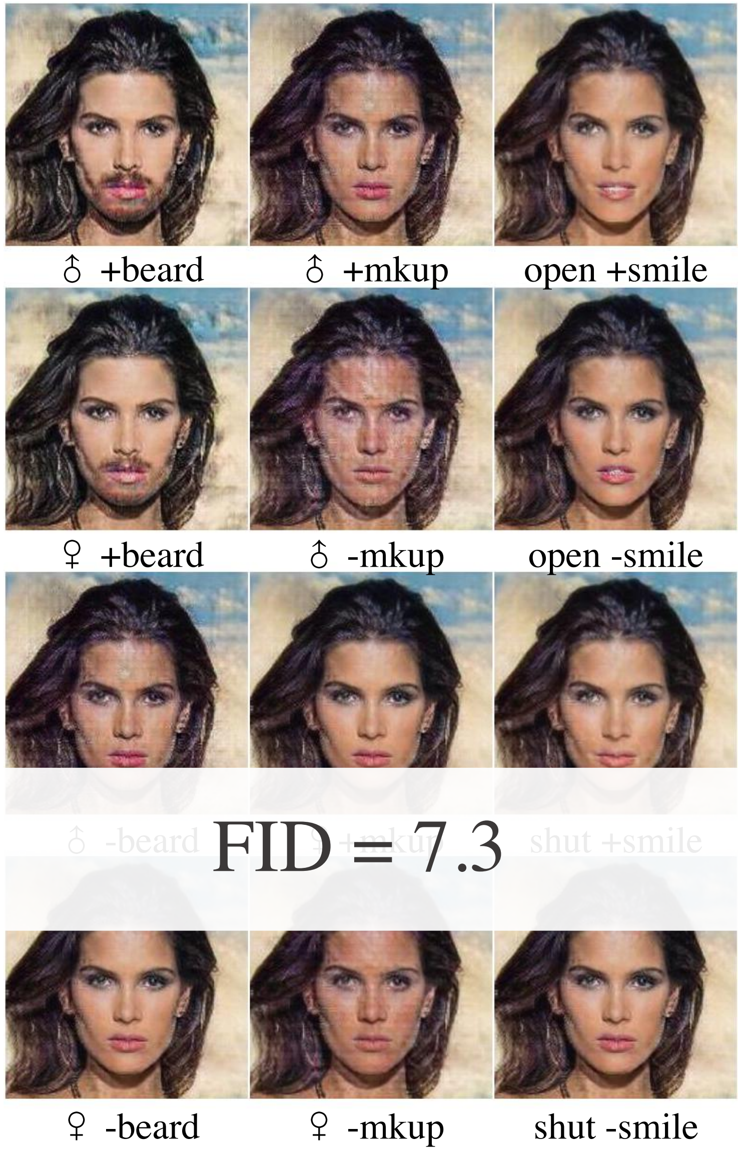
Fader Networks



FID = 26.3

baseline

AttGan



FID = 7.3

Human Evaluation

	MSP(ours)	Fader	AttGAN
male x beard	**0.78**	0.42	0.45
female x beard	**0.52**	0.03	0.41
male x no-beard	**0.86**	0.40	0.42
female x no-beard	**0.90**	0.61	0.63
male x makeup	**0.52**	0.02	0.35
male x no-makeup	**0.89**	0.50	0.47
female x makeup	**0.87**	0.63	0.52
female x no-makeup	**0.67**	0.42	0.47
smile x open-mouth	**0.89**	0.59	0.63
no-smile x open-mouth	**0.66**	0.11	0.29
smile x calm-mouth	**0.95**	0.34	0.33
no-smile x calm-mouth	**0.76**	0.43	0.38
male x bald	**0.78**	0.10	0.29
male x bangs	**0.56**	0.05	0.19
female x bald	**0.29**	0.01	0.17
female x bangs	**0.68**	0.21	0.20
no-glasses x black-hair	**0.74**	0.38	0.53
no-glasses x golden-hair	**0.86**	0.36	0.79
glasses x black-hair	**0.82**	0.21	0.32
glasses x golden-hair	**0.77**	0.19	0.33

Table 2. The classification accuracy of generated images using MSP, Fader Networks and AttGAN.

male / beard attributes morphing			
	Fader Network	AttGAN	VAE+GAN MSP
perfect	38.3%	55.9%	74.4%
recognizable	8.3%	11.2%	11.6%
unreco/unchang	53.3%	32.9%	14.0%

mouth open / smiling attributes morphing			
	Fader Network	AttGAN	VAE+GAN MSP
perfect	36.7%	47.5%	68.3%
recognizable	20.8%	15.3%	4.9%
unreco/unchang	42.5%	37.2%	26.8%

Table 4. Manual valuation results of disentanglement. Numbers in the table denote percentage of participants under the column heading who felt the images represented the specified attribute (e.g. smiling) in a way that was perfect, recognisable, or unrecognisable/unchanged.

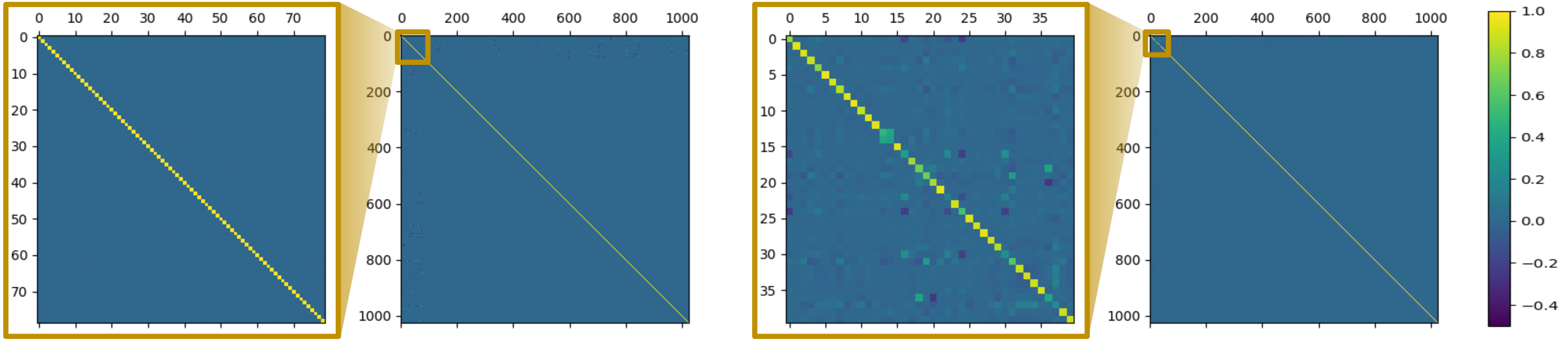
Evaluation (textural task)

- E2E dataset.
- A classic seq2seq autoencoder (lstm-lstm) + MSP

	Example 1	Example 2
Orig-attribute	eatType[pub], customer-rating[5-out-of-5], name[Blue-Spice], near[Crowne-Plaza-Hotel]	familyFriendly[yes], area[city-centre], eatType[pub], food[Japanese], near[Express-by-Holiday-Inn], name[Green-Man]
Orig-text	the blue spice pub , near crowne plaza hotel , has a customer rating of 5 out of 5 .	near the express by holiday inn in the city centre is green man . it is a japanese pub that is family-friendly .
New-attribute	eatType[coffee-shop], customer-rating[5-out-of-5], name[Blue-Spice], near[Avalon]	familyFriendly[no], area[riverside], eatType[coffee-shop], food[French], near[The-Six-Bells], name[Green-Man]
New-text	the blue spice coffee shop , near avalon has a customer rating of 5 out of 5 .	near the six bells in the riverside area is a green man . it is a french coffee shop that is not family-friendly .

Orthogonality of U

- Since $\mathcal{L}_2$ uses the orthogonality of U, do we train U as an orthogonal matrix? Almost!
  (MSP only learns M, but can obtain U by calculating the *null space* of M)



$\mathrm{M} \cdot \mathrm{M}^T$ $\mathrm{U} \cdot \mathrm{U}^T$ $\mathrm{M} \cdot \mathrm{M}^T$ $\mathrm{U} \cdot \mathrm{U}^T$

for the E2E dataset　　for the CelebA dataset

Conclusion

Conclusion

- We proposed MSP, which fully disentangles the latent space of an autoencoder to manipulate the multiple attributes in the latent space.
- Our model is a plug-in, which in principle can be attached to any type of autoencoder (e.g. for images or text), and we have a principled weighting strategy for combining the loss terms for training.
- MSP shows strong performance on learning disentangled latent representations of multiple attributes.
- We also suggested a way to train a matrix to be orthogonal.

Thanks!

The code of MSP and relative data is here:
https://xiao.ac/proj/msp