Latent Space Manipulation

Vadim Titov, MIPT

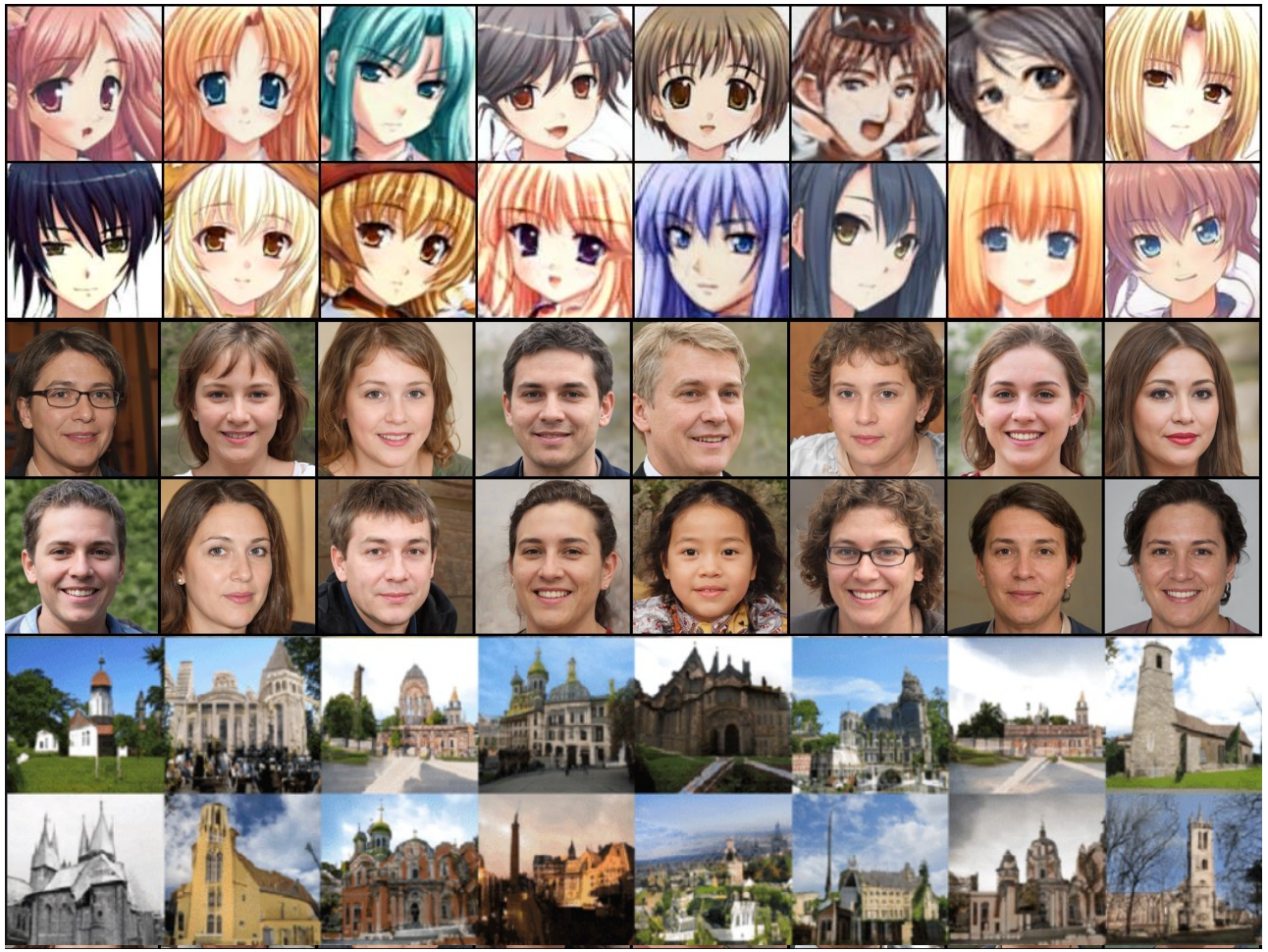
Tg: @MACderRU

Content

- Introduction, StyleGAN2 and Latent Space semantic.
- Supervised methods.
- Unsupervised methods.
- Text-driven latent space manipulation.

Introduction.
Image semantic. Datasets.

- Anime Faces
- Flickr Faces
- Celeba
- Animal Faces
- Churches

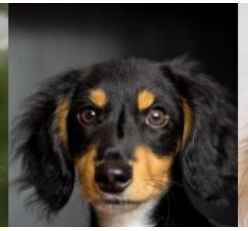
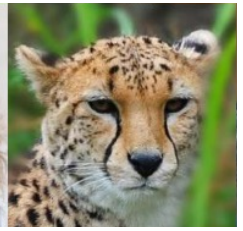


Introduction.
Image semantic. Datasets.

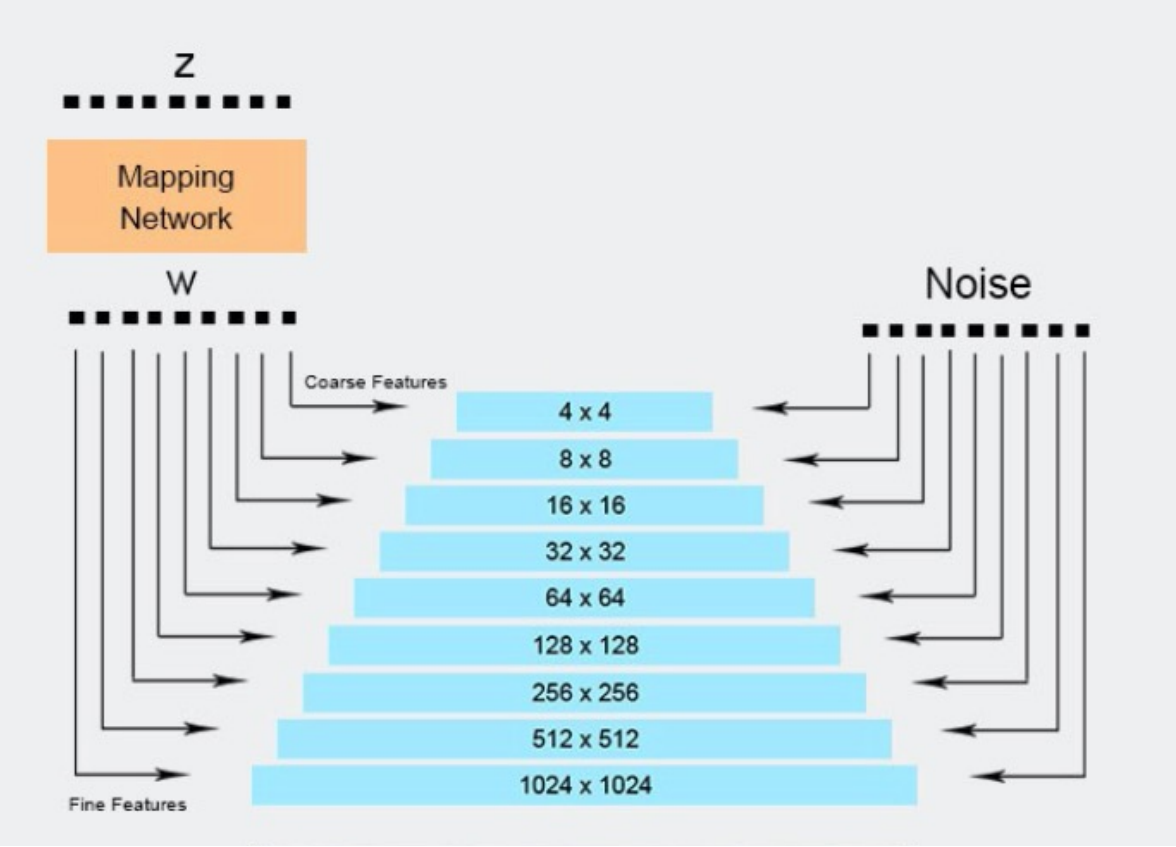
Celeba Dataset

Attributes			
Bald	❌	❌	❌
Bangs	❌	❌	❌
Brown hair	✅	❌	✅
Black hair	❌	✅	❌
Blond hair	❌	❌	❌
Big lips	❌	✅	✅

Animal Faces Dataset

Class				
Cat	✅	❌	❌	❌
Dog	❌	✅	✅	❌
Wildlife	❌	❌	❌	✅

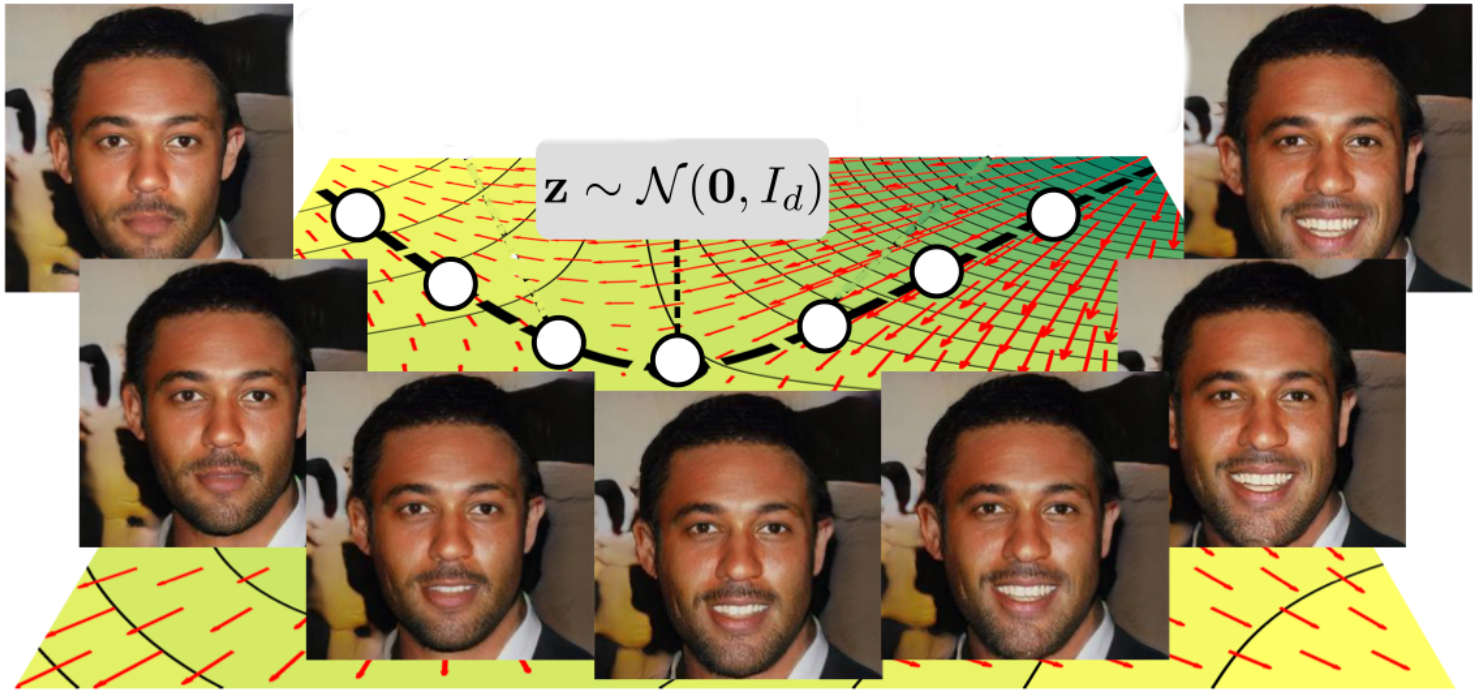
Introduction.
Generative Adversarial Networks. High-reality synthesis.



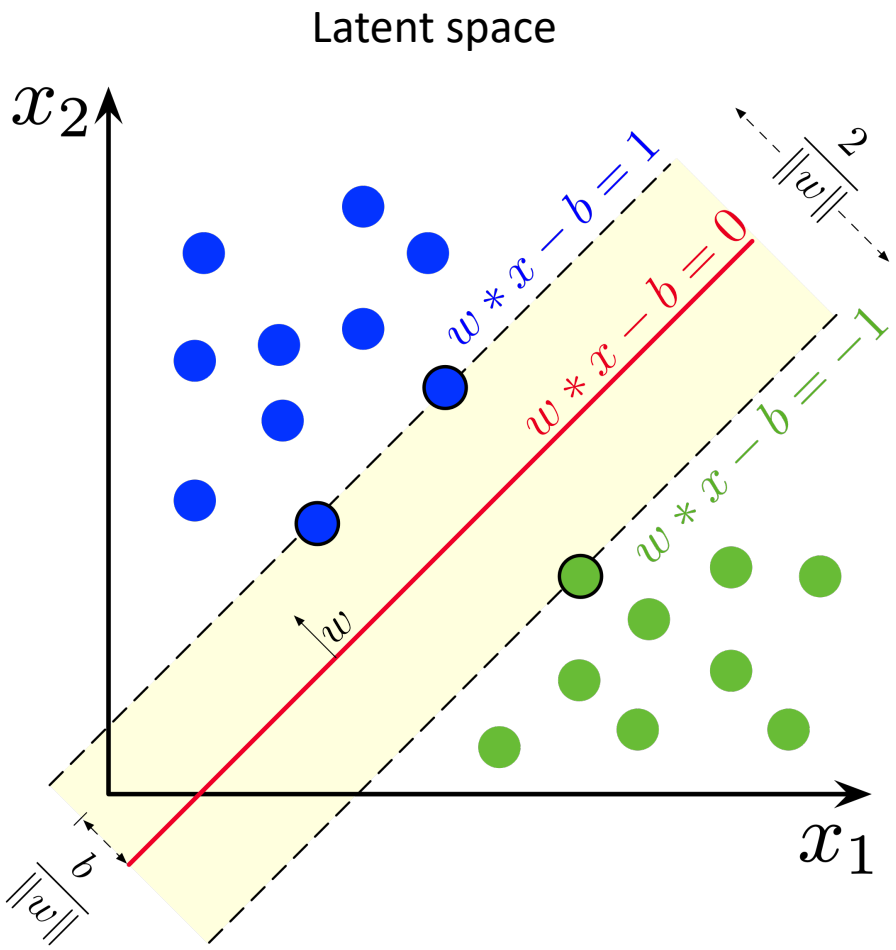
Introduction. Subject.
Image editing via latent code manipulation.

Z – initial latent code
G – pretrained generator

Initial Image: $G(z)$
Edited image: $G(z + shift)$



Supervised approach. Baseline



- Inverted image to latent space
- Positive semantic label
- Negative semantic label

Supervised approach. Baseline. Results

Original

Supervised approach. Baseline

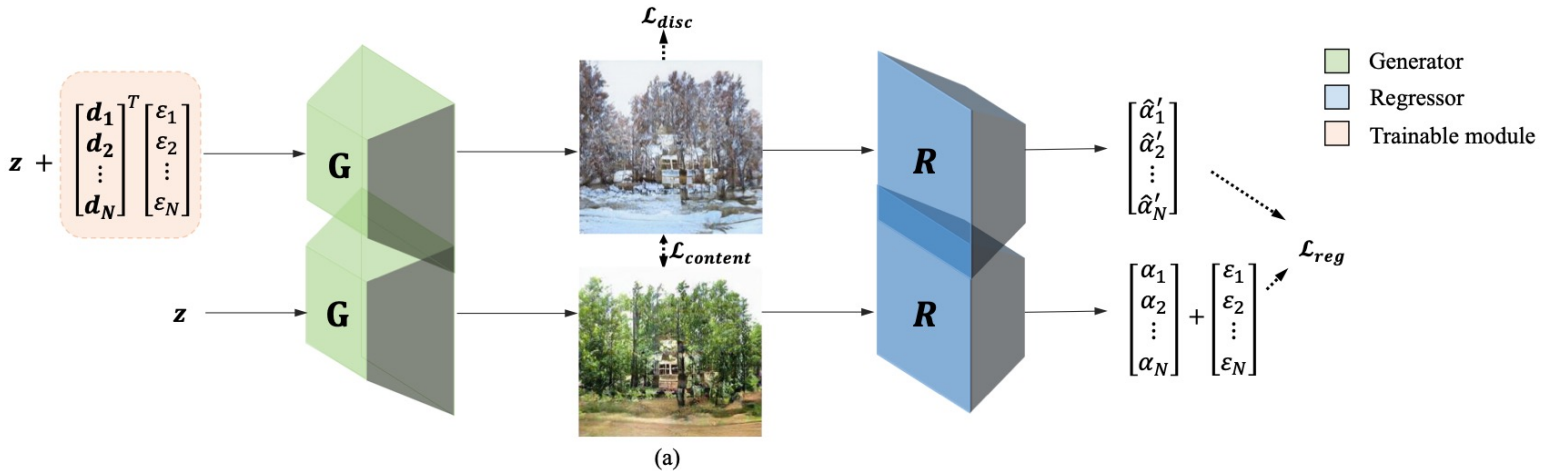
- Advantages:
  - Simple idea

- Disadvantages:
  - Dependency on pretrained models
  - Computationally inefficient
  - Correlation between semantic directions

Supervised approach. Improved Framework.

New framework proposes to optimize shift parameter with respect to improved set of losses:

1. Discriminator consistent loss
2. Regressor consistent loss
3. Perceptual content loss



Supervised approach. Improved Framework.

$\boldsymbol{T} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N\}$ - set of shift vectors in latent space

Framework assumption:

Given image I = G(z), $\alpha = R(I)$
$\alpha + \epsilon = R(G(z + \sum_{k=0}^{n} \epsilon_k d_k))$

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z},\boldsymbol{\varepsilon}\sim\mathcal{D}_\varepsilon}[-\hat{\boldsymbol{\alpha}}' \log \boldsymbol{\alpha}' - (\mathbf{1} - \hat{\boldsymbol{\alpha}}') \log (\mathbf{1} - \boldsymbol{\alpha}')].$$

$$\mathcal{L}_{\text{disc}} = \mathbb{E}_{\boldsymbol{z}'\sim\mathcal{Z}'|\boldsymbol{z}} \left[\log(1 - D(G(\boldsymbol{z}')))\right].$$

$$\mathcal{L}_{\text{content}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z},\boldsymbol{z}'\sim\mathcal{Z}'|\boldsymbol{z}} \sum_{i\in\mathcal{D}_{\text{content}}} \|F_i(G(\boldsymbol{z}')) - F_i(G(\boldsymbol{z}))\|_2^2,$$

$$\min_{\boldsymbol{T}} \mathcal{L} \triangleq \lambda_1 \mathcal{L}_{\text{reg}} + \lambda_2 \mathcal{L}_{\text{disc}} + \lambda_3 \mathcal{L}_{\text{content}}.$$

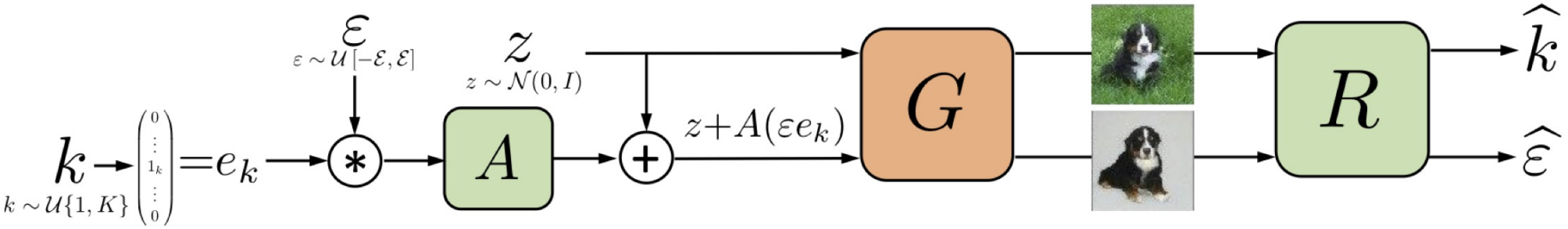
Supervised approach. Improved Framework. Results.

Supervised approach. Improved Framework

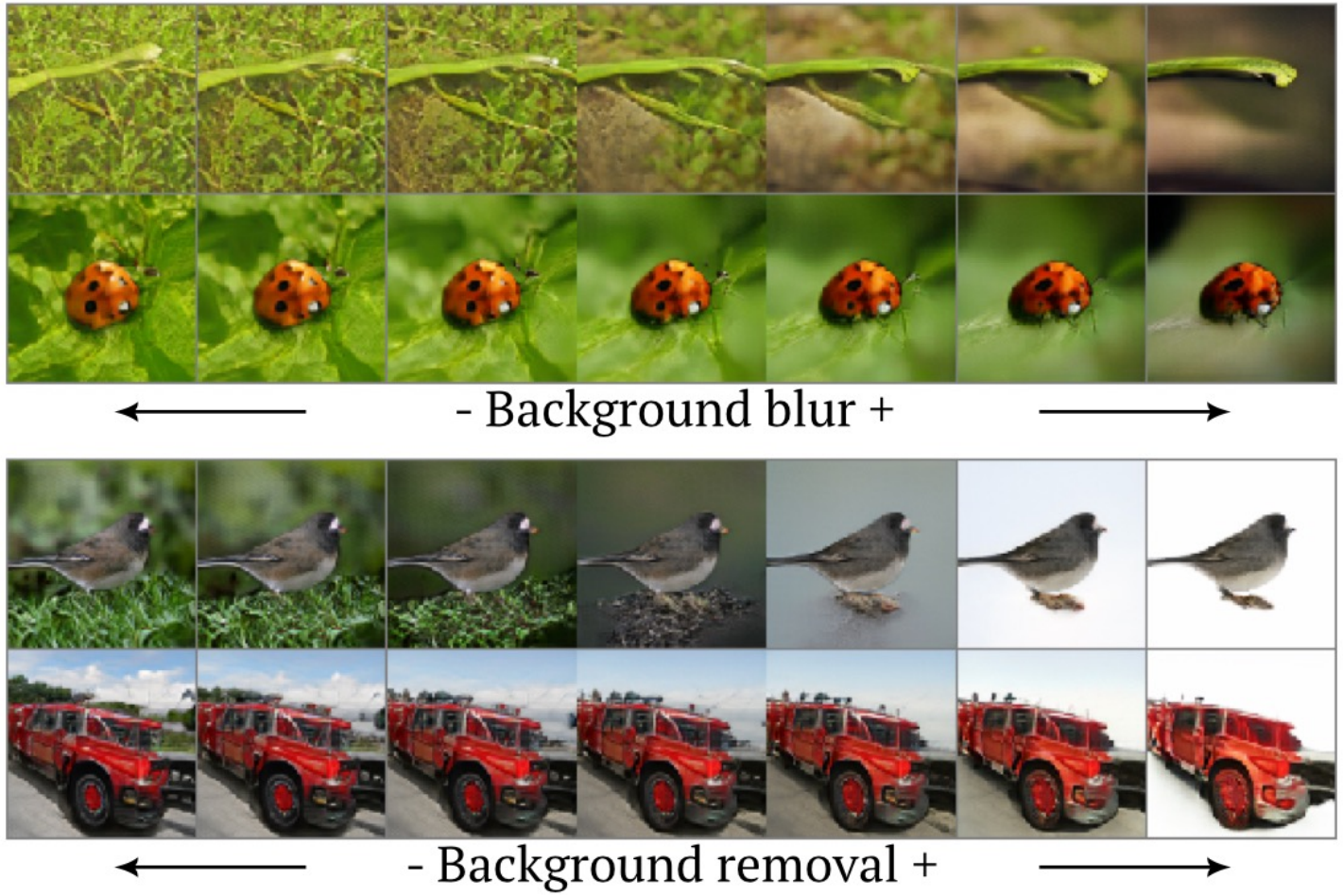
- Advantages:
  - End-to-End method
  - Scallable for non-linear trajectories

- Disadvantages:
  - Dependency on pretrained models

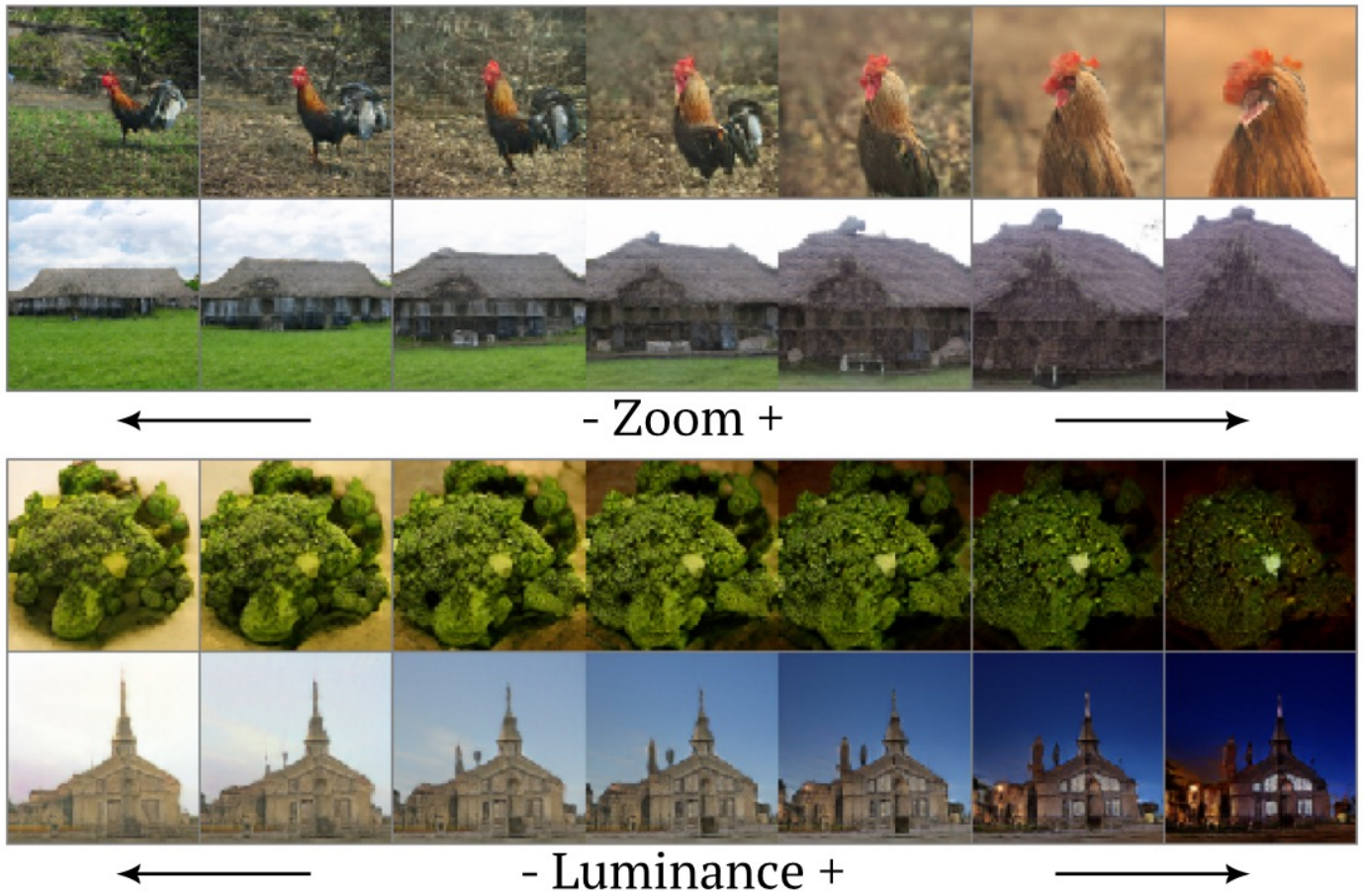
Unsupervised approach. Core framework.



Unsupervised approach. Core framework. Results.



Unsupervised approach. Core framework. Results.

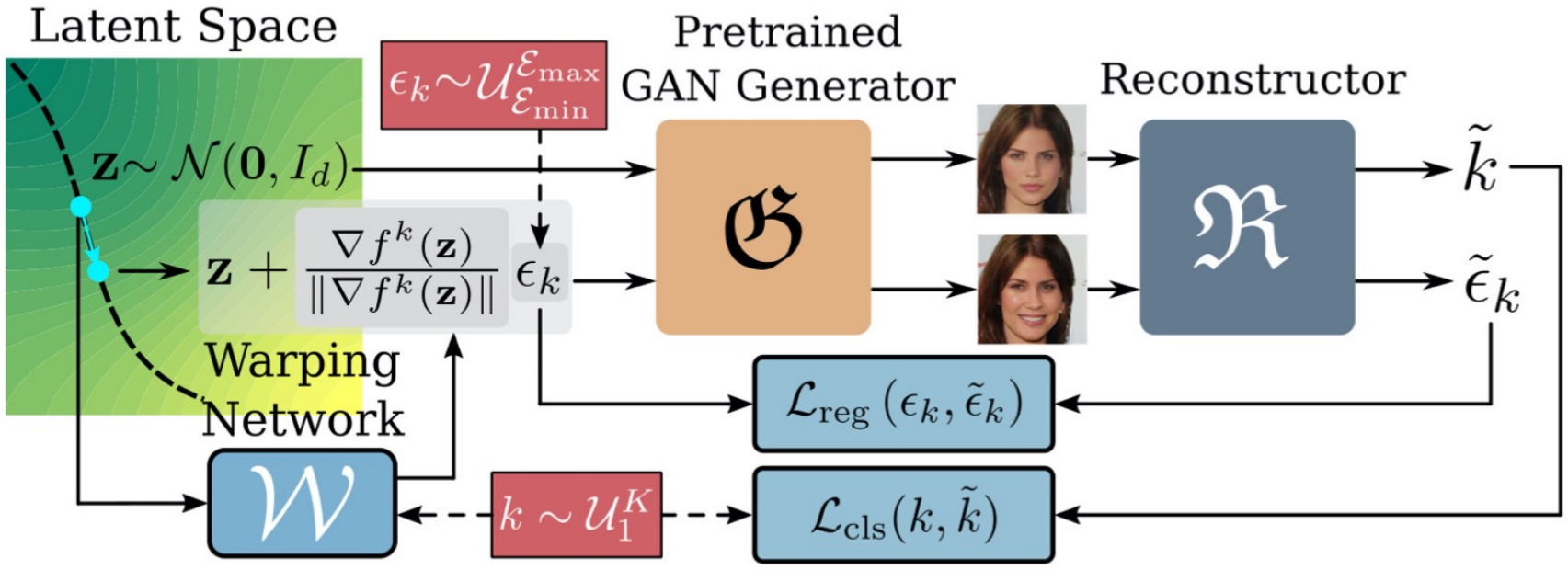


Unsupervised approach. Core framework.

- Advantages:
  - Only pretrained generator is needed
  - Scalable for various direction parameterizations

- Disadvantages:
  - Directions with same semantic properties
  - Need help from assessors to determine semantic property

Unsupervised approach. Non linear manipulation.

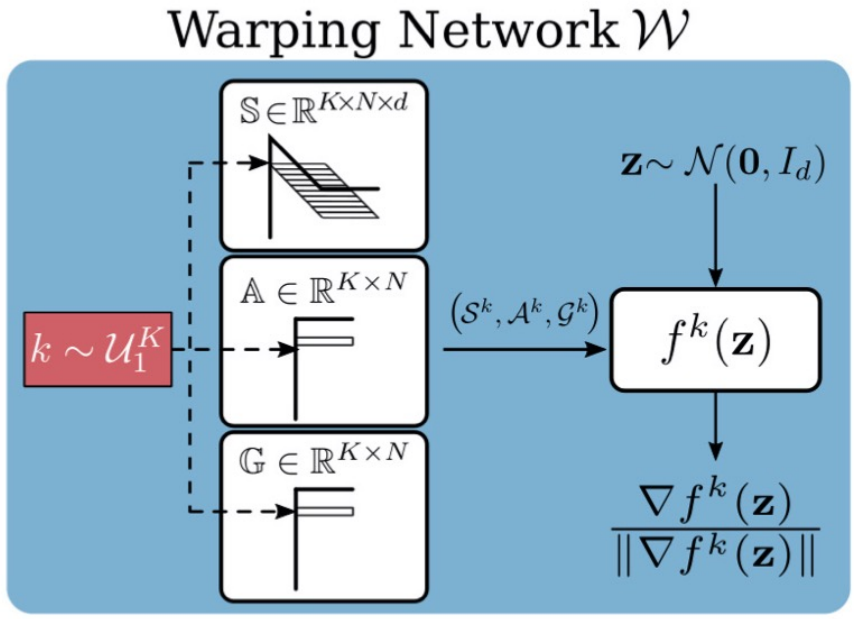


Unsupervised approach. Non linear manipulation.

- WarpedGANSpace

$$f(\mathbf{z}) = \sum_{i=1}^{N} \alpha_i \exp\left(-\gamma_i \|\mathbf{z} - \mathbf{s}_i\|^2\right)$$

$$\nabla f(\mathbf{z}) = -2 \sum_{i=1}^{N} \alpha_i \gamma_i \exp\left(-\gamma_i \|\mathbf{z} - \mathbf{s}_i\|^2\right) (\mathbf{z} - \mathbf{s}_i)$$



Unsupervised approach. Non linear manipulation.Results.

Unsupervised approach. Non linear manipulation.Results.

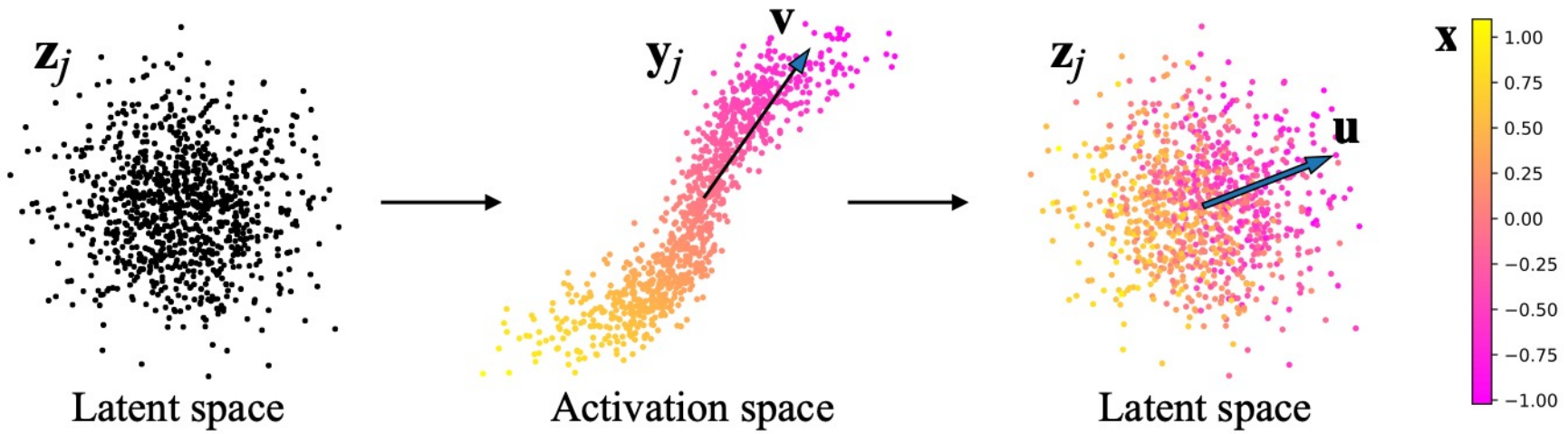
Unsupervised approach. Non linear manipulation.Results.

Unsupervised approach. Non linear manipulation.

- Advantages:
  - Only pretrained generator is needed
  - Universal framework
  - In case of semantic dependent warping network no assessor help is needed

- Disadvantages:
  - Directions with same semantic properties aswell
  - Need help from assessors to determine semantic property

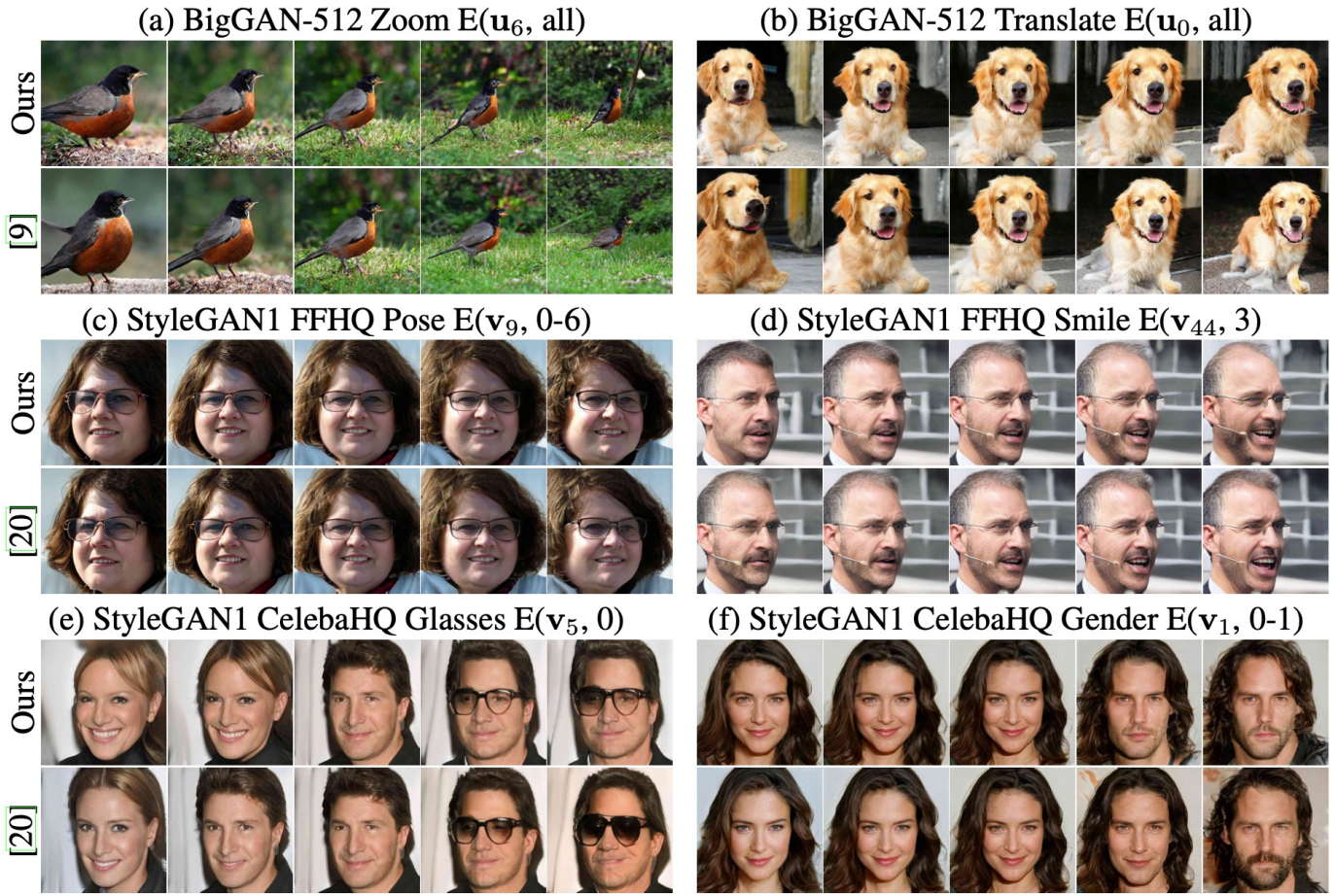
Unsupervised approach. Principal Components.



Unsupervised approach. Principal Components.

- Advantages:
  - Only pretrained generator is needed
- Disadvantages:
  - Semantic edits correlate due to correlation in domain data
  - Need help from assessors to determine semantic property

Unsupervised approach. Principal Components. Results.



Metrics. Unsupervised methods.

	Yaw	Pitch	Smile	Race	Hair
Yaw	**0.32**	0.05	0.01	0.07	0.03
Pitch	0.04	**0.38**	0.13	0.03	0.01
Smile	0.03	0.07	**0.61**	0.03	0.03
Race	0.03	0.12	0.08	**0.29**	0.17
Hair	0.02	0.11	0.13	0.02	**0.49**

(a) Non-linear paths (Ours).

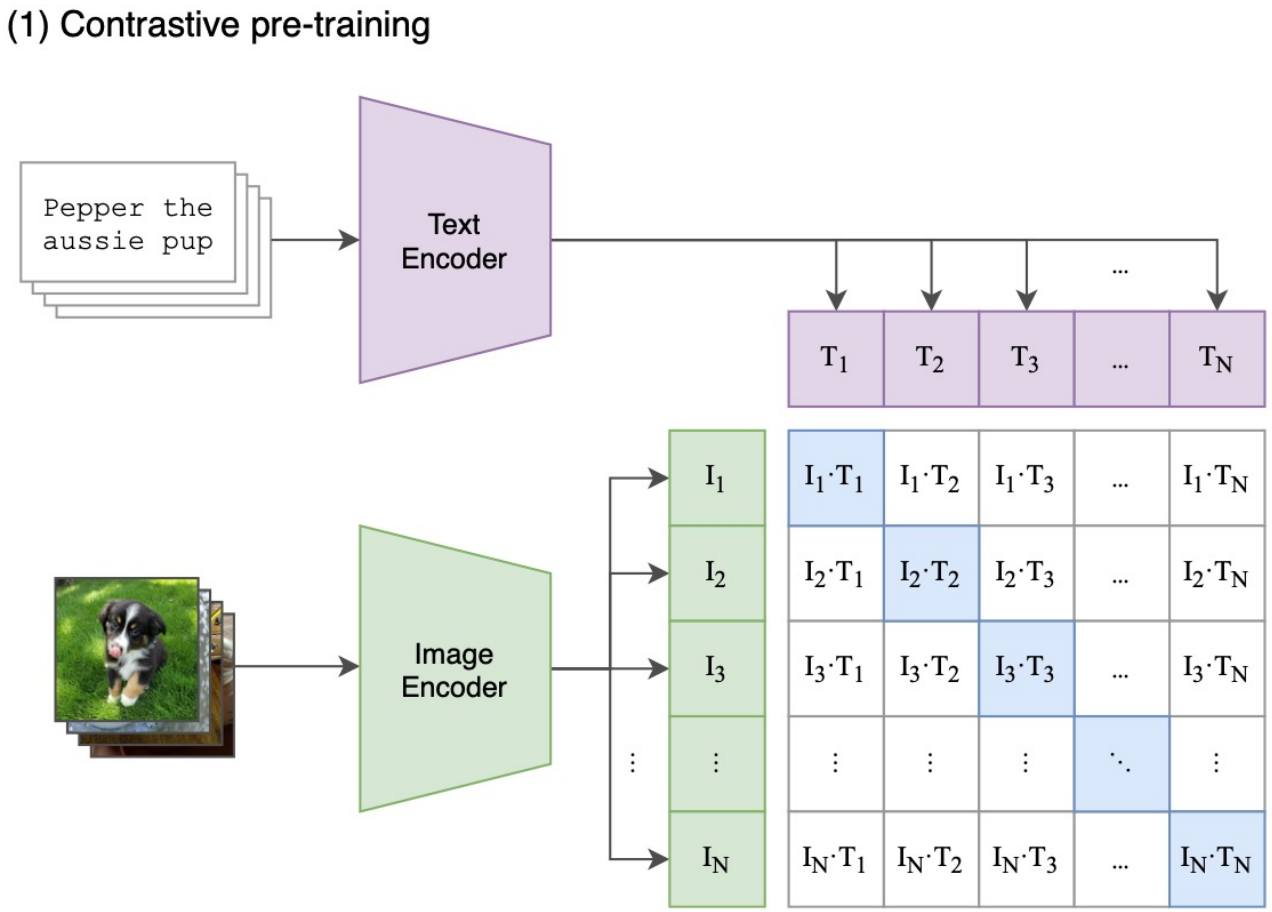
	Yaw	Pitch	Smile	Race	Hair
Yaw	**0.24**	0.21	0.01	0.02	0.01
Pitch	0.01	**0.25**	0.04	0.00	0.22
Smile	0.01	0.04	**0.57**	0.05	0.09
Race	0.05	0.02	0.10	**0.31**	0.01
Hair	0.00	0.10	0.06	0.04	**0.36**

(b) Linear directions (Voynov and Babenko [34]).

	Yaw	Pitch	Smile	Race	Hair
Yaw	**0.27**	0.04	0.13	0.03	0.06
Pitch	0.05	**0.38**	0.09	0.02	0.01
Smile	0.00	0.07	**0.55**	0.08	0.08
Race	0.11	0.02	0.12	**0.27**	0.12
Hair	0.05	0.06	0.03	0.08	**0.34**

(c) Linear PCA directions (GANSpace [11]).

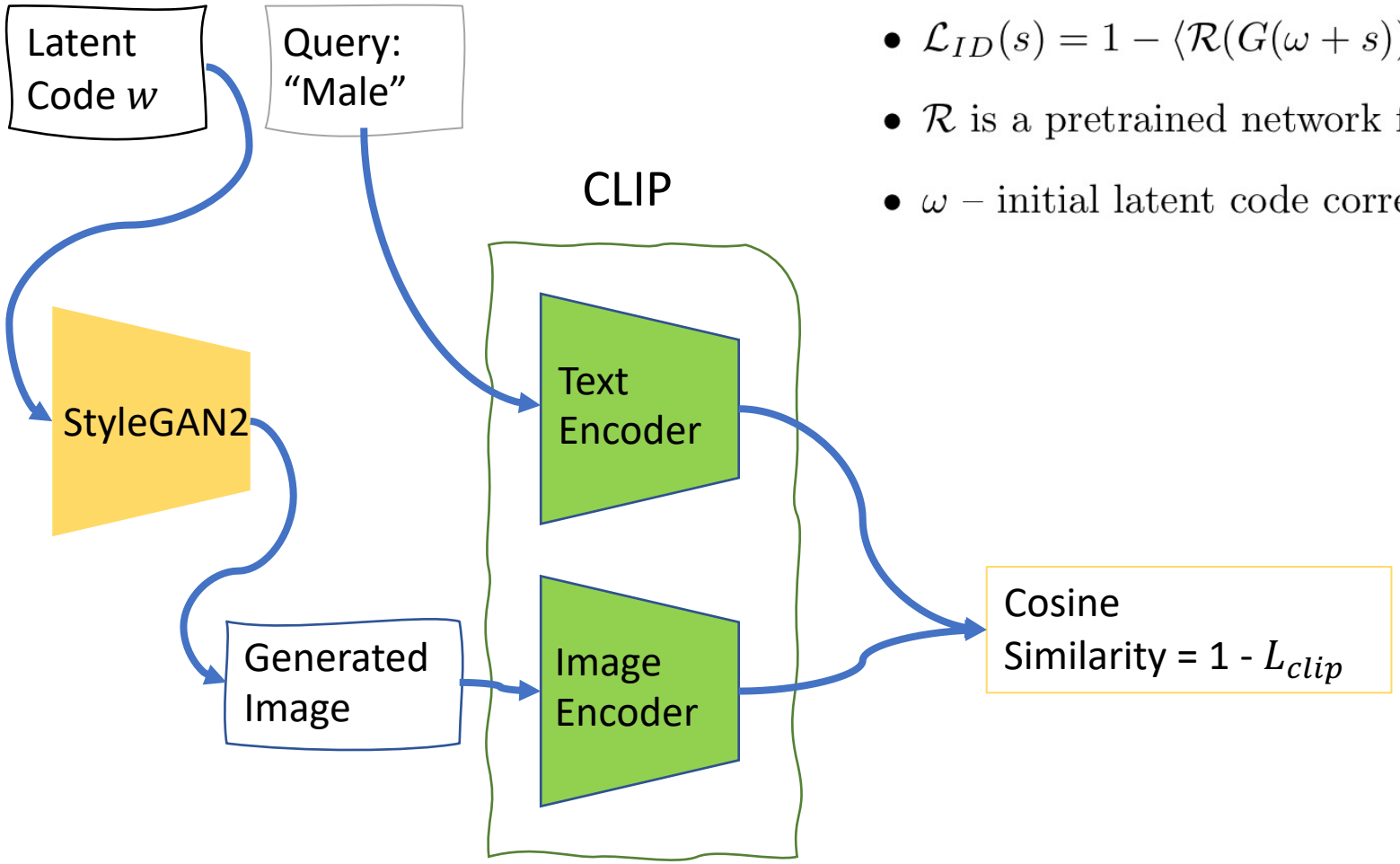
Text-driven manipulation. CLIP



Text-driven manipulation. CLIP supervision

$$shift = \arg\min_{s} \quad \mathcal{L}_{clip}(G(\omega + s), "query") + \lambda_{l2}||s||_2 + \lambda_{ID}\mathcal{L}_{ID}(s)$$

- $\mathcal{L}_{ID}(s) = 1 - \langle \mathcal{R}(G(\omega + s)), \mathcal{R}(G(\omega)) \rangle$
- $\mathcal{R}$ is a pretrained network for face recognition
- ω – initial latent code corresponding to image



Text-driven manipulation. CLIP supervision. Initial results.

Query: “A person with makeup”

Query: “Red hairs”

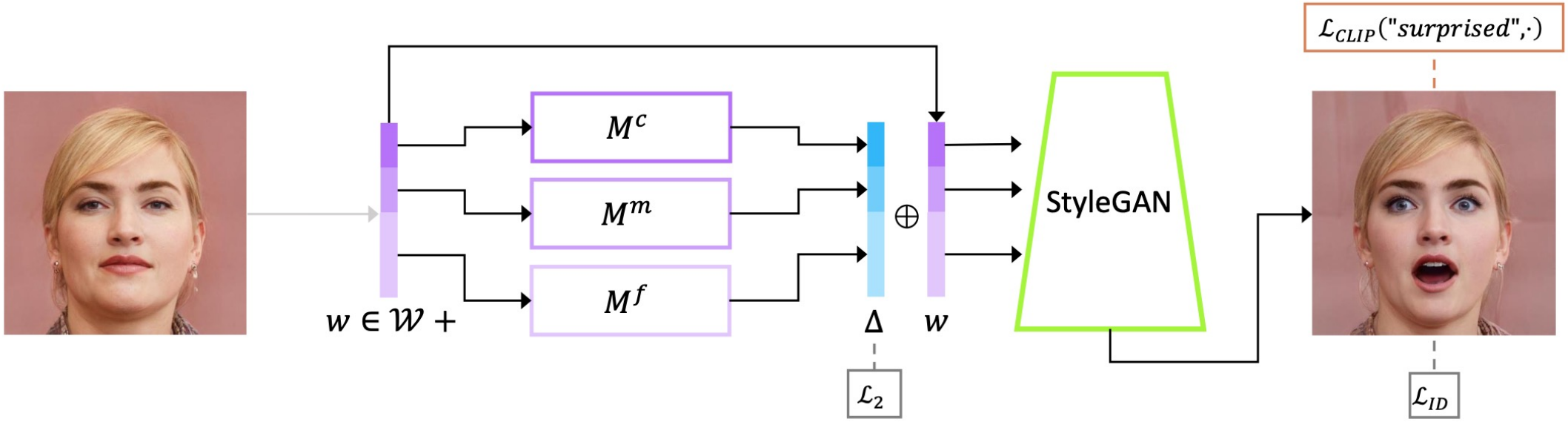
Query: “Woman with curly hairs”

Query: “Afro haircut”

Query: “Young person”

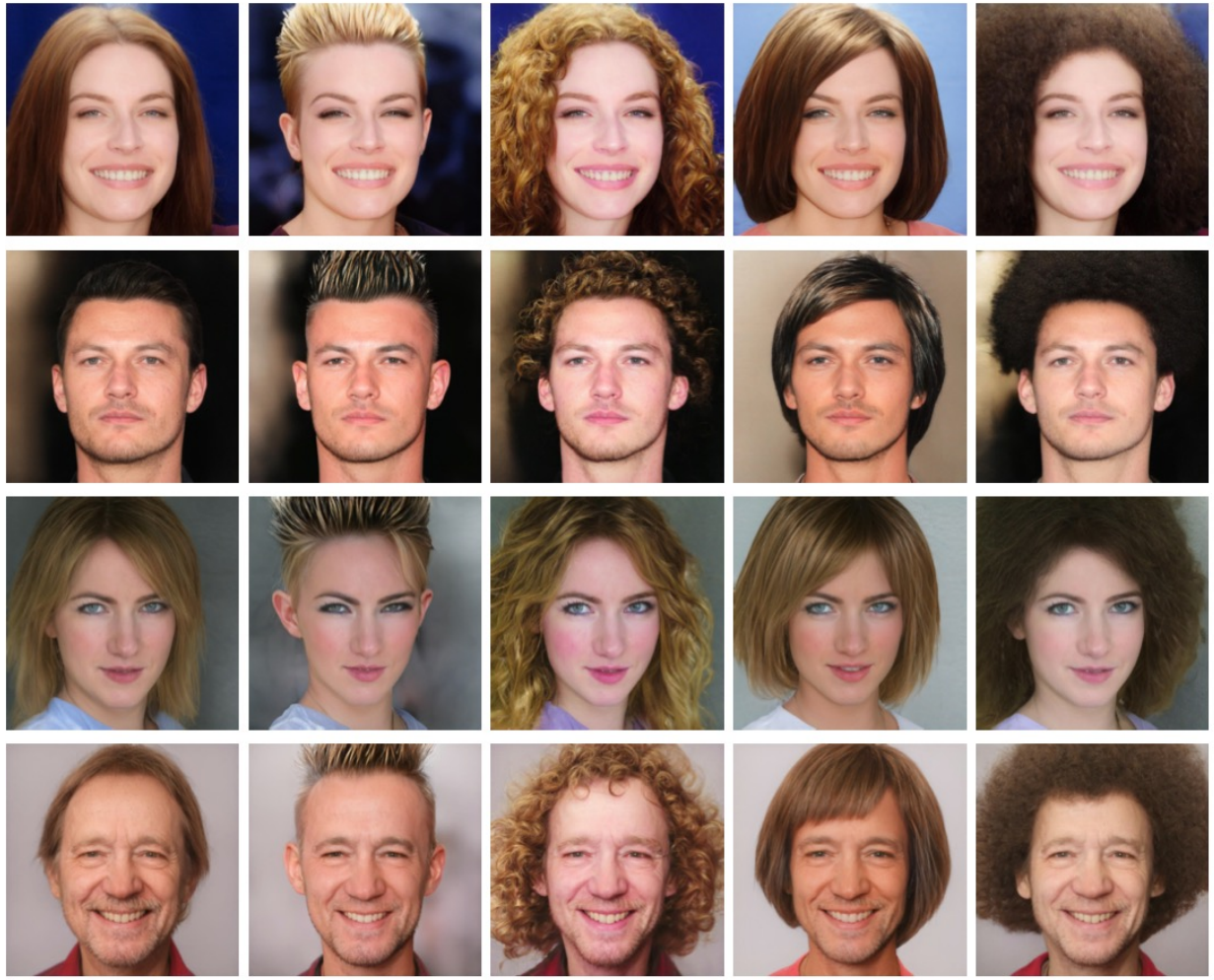
Query: “Sad face”

Text-driven manipulation. StyleCLIP.



$$\theta = \arg\min \ \mathbb{E}_{\omega \in W} \mathcal{L}_{clip}(G(\omega + M_\theta(\omega)), "query") + \lambda_{l2}||M_\theta(\omega)||_2 + \lambda_{ID}\mathcal{L}_{ID}(M_\theta(\omega))$$

Text-driven manipulation. StyleCLIP. Results



Text-driven manipulation. StyleCLIP. Results.