JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

3D-Aware Multi-Class Image-to-Image Translation with NeRFs

Senmao Li[1] Joost van de Weijer[2] Yaxing Wang[1*] Fahad Shahbaz Khan[3,4] Meiqin Liu[5] Jian Yang[1]

[1]VCIP,CS,Nankai University, [2]Universitat Auto ǹoma de Barcelona,

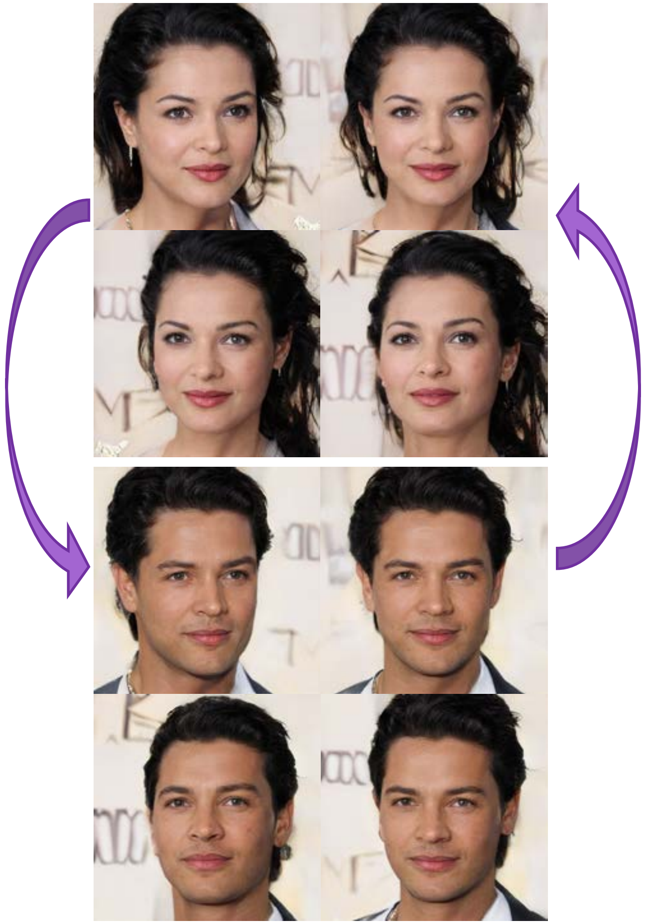
[3]Mohamed bin Zayed University of AI, [4]Linkoping University, [5]Beijing Jiaotong University

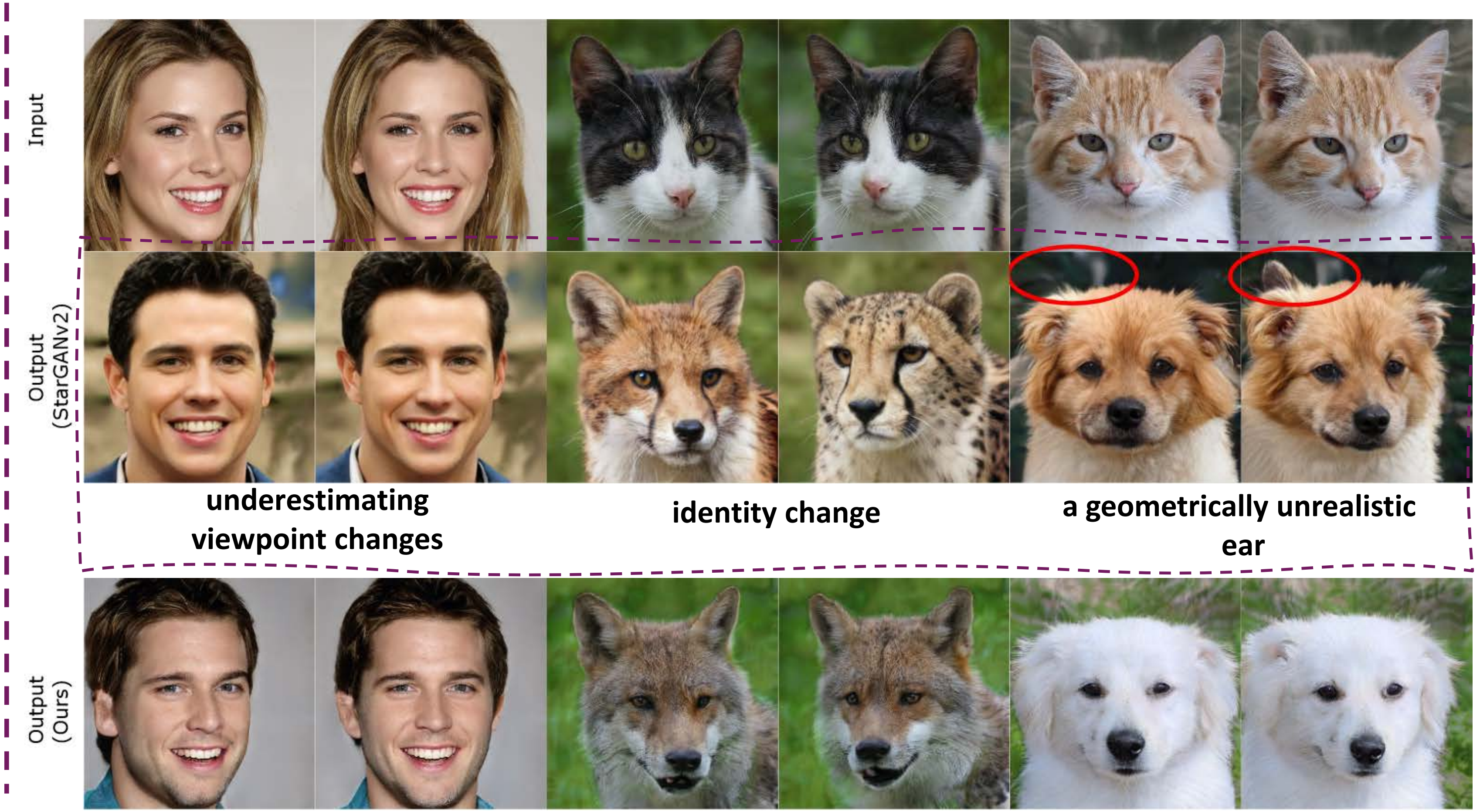
Paper ID 81

Code: https://github.com/sen-mao/3di2i-translation

Problems

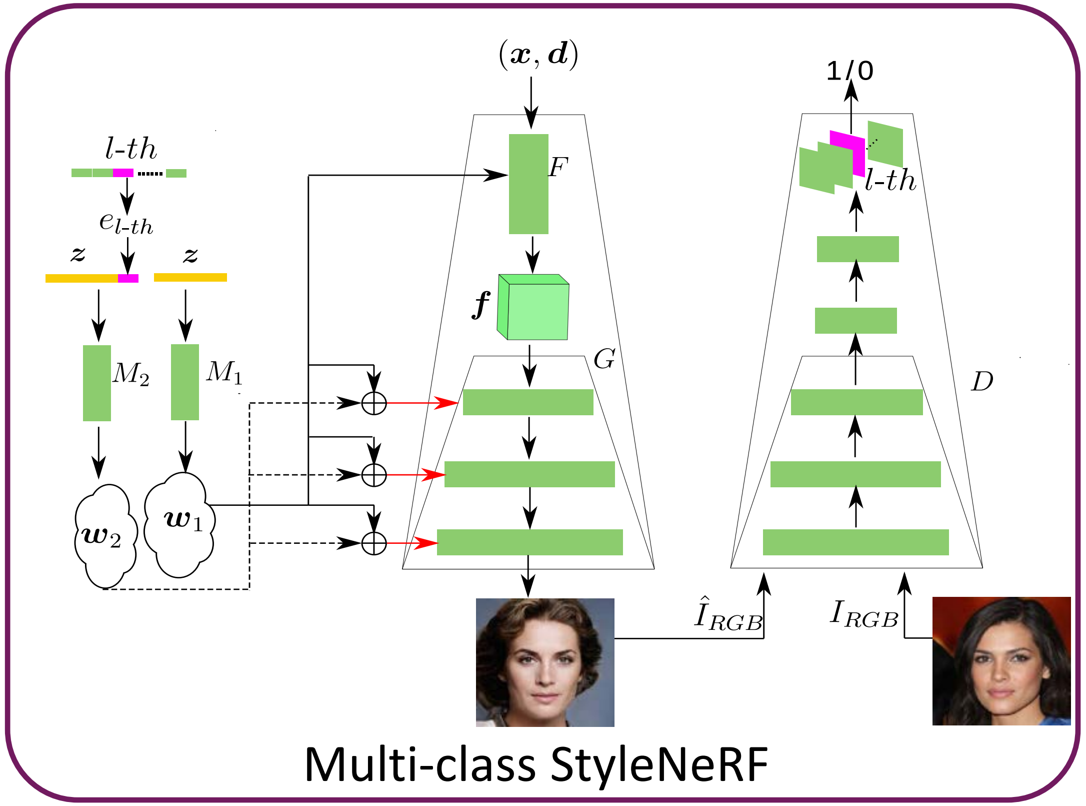
- **No prior works** investigate 3D-aware GANs for 3D consistent multi-class image-to-image (3D-aware I2I) translation.
- 2D-I2I translation methods applied to 3D-I2I translation tasks result in three main challenges (**1. underestimating viewpoint changes, 2. identity change, 3. a geometrically unrealistic ear**) when changing the viewpoint.




Input
Output (StarGANv2)
Output (Ours)
underestimating viewpoint changes
identity change
a geometrically unrealistic ear

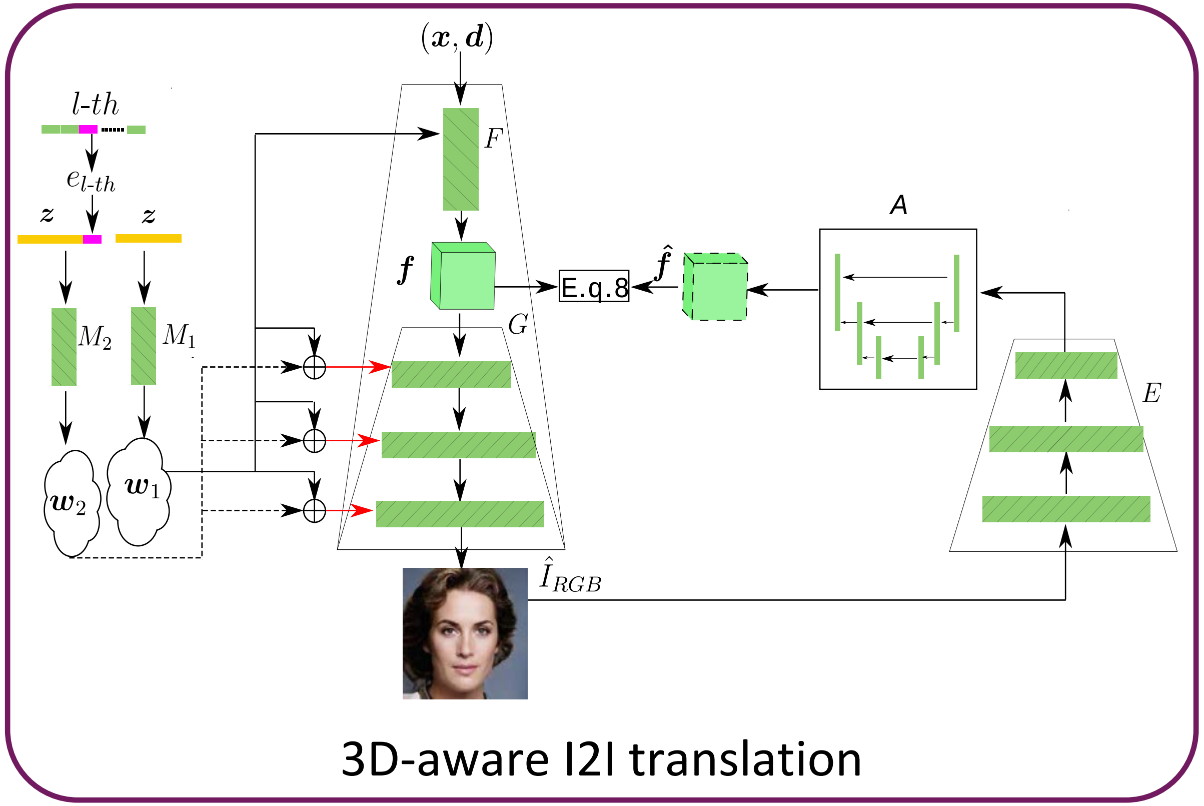
Methods

- We decouple the learning process into **multi-class 3D-aware generation (step1)** and **3D-aware I2I translation (step2)**.



Multi-class StyleNeRF

Step1

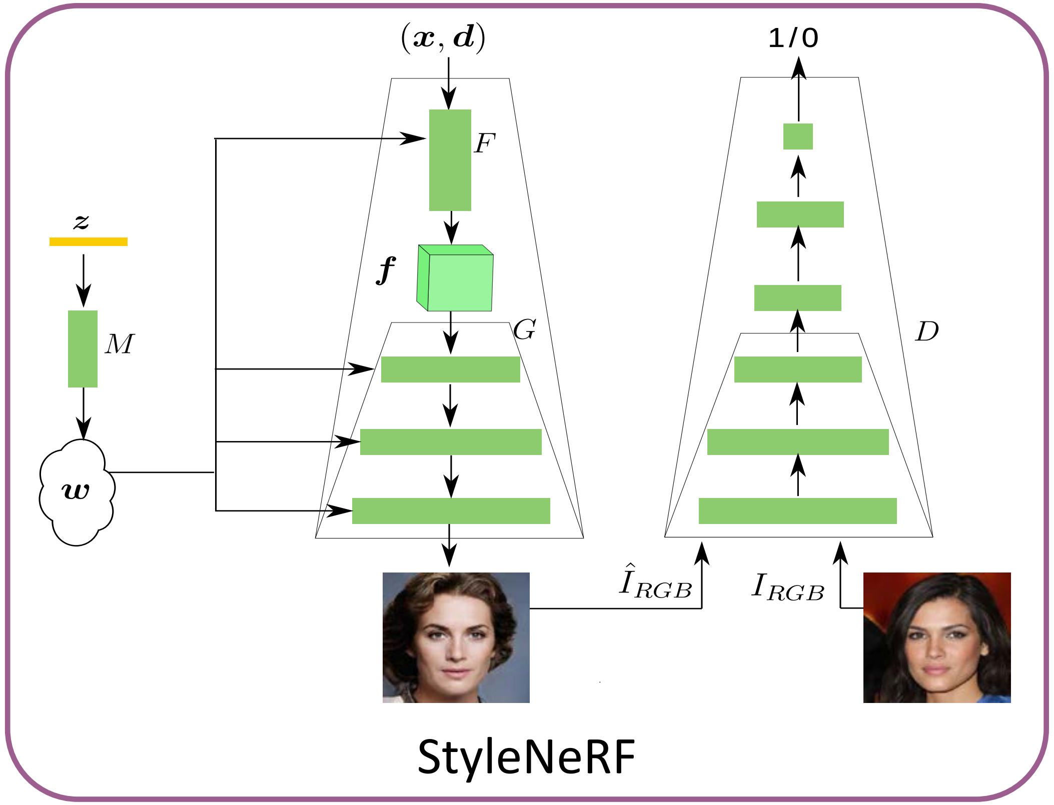


3D-aware I2I translation

Step2

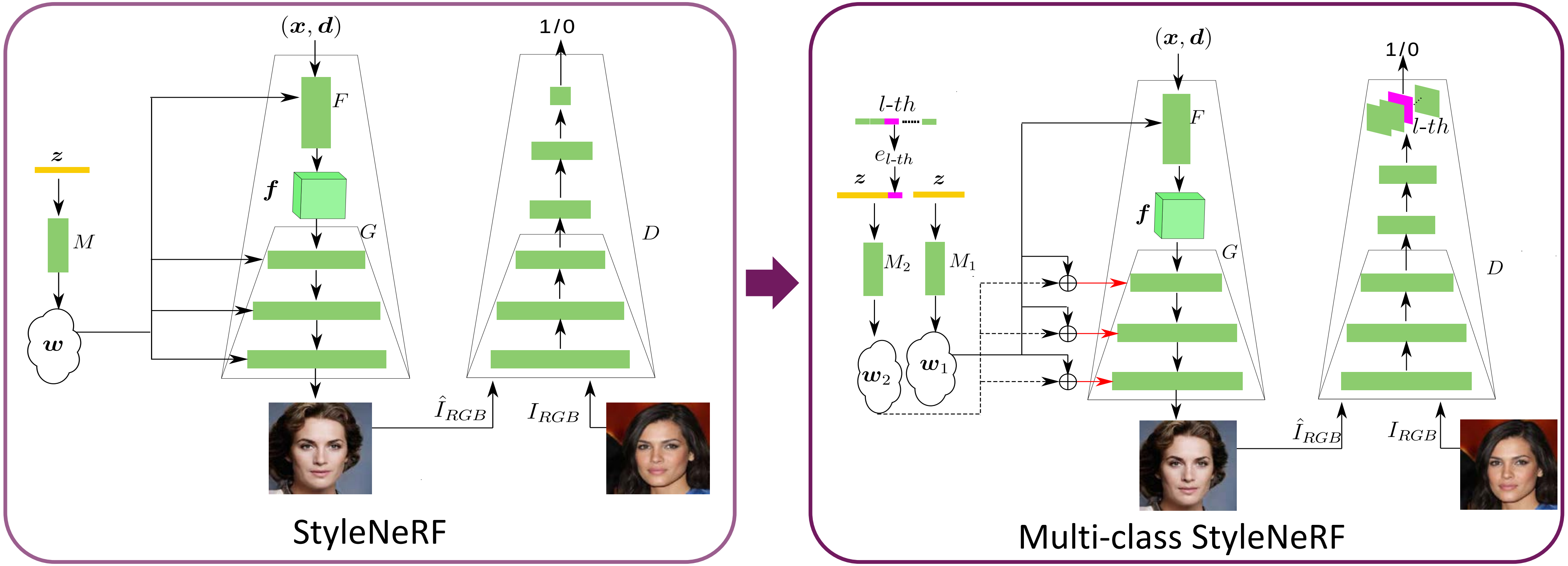
Methods

- **step1**: (1) training an unconditional 3D-aware generative model on datasets (i.e., StyleNeRF) and (2) partially initializing the multi-class 3D-aware generative model (i.e., multi-class StyleNeRF).



Methods

- **step1**: (1) training an unconditional 3D-aware generative model on datasets (i.e., StyleNeRF) and (2) partially initializing the multi-class 3D-aware generative model (i.e., multi-class StyleNeRF).

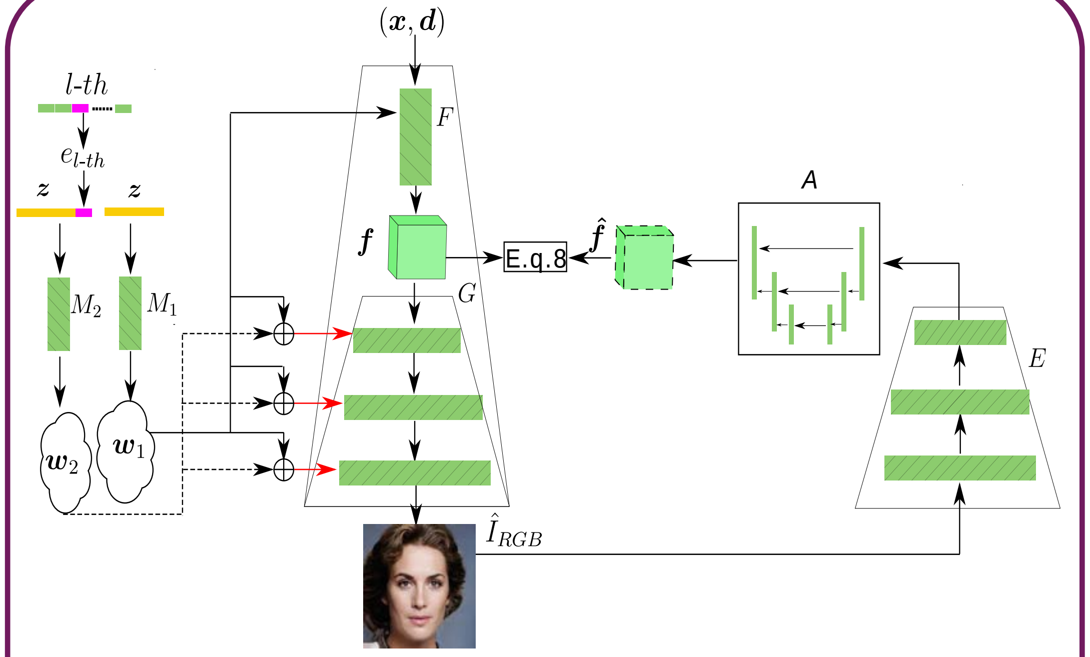


StyleNeRF

Multi-class StyleNeRF

Methods

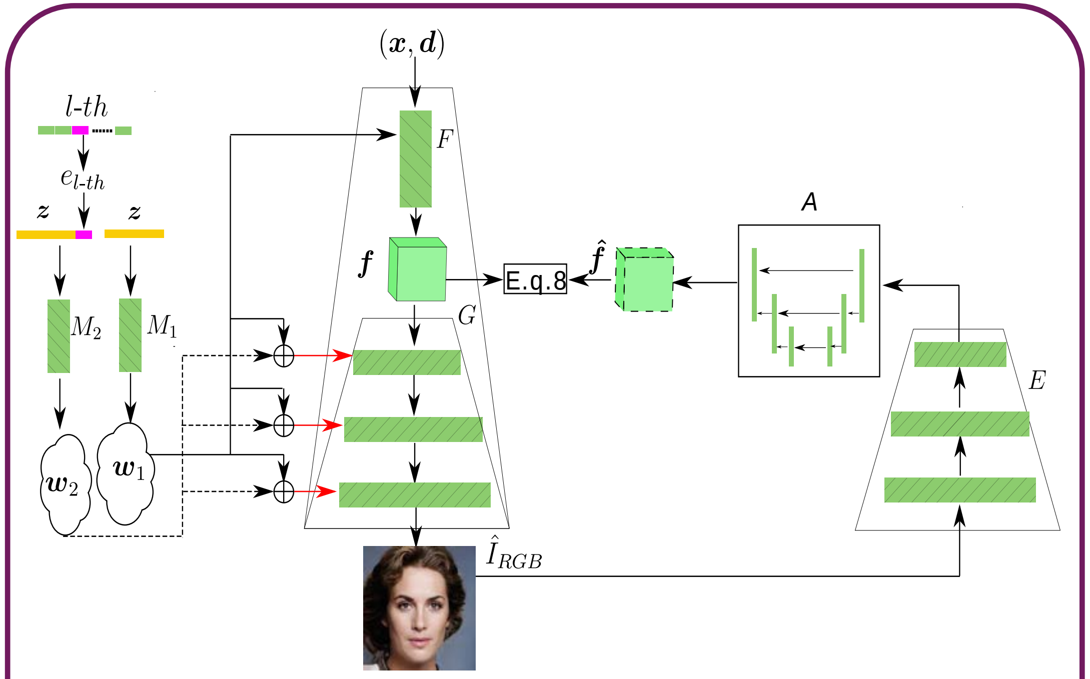
- **step2**: 3D-aware I2I translation architecture adapted from the trained multi-class StyleNeRF (**step1**). This initialization inherits the capacity of being sensitive of view information.



3D-aware I2I translation

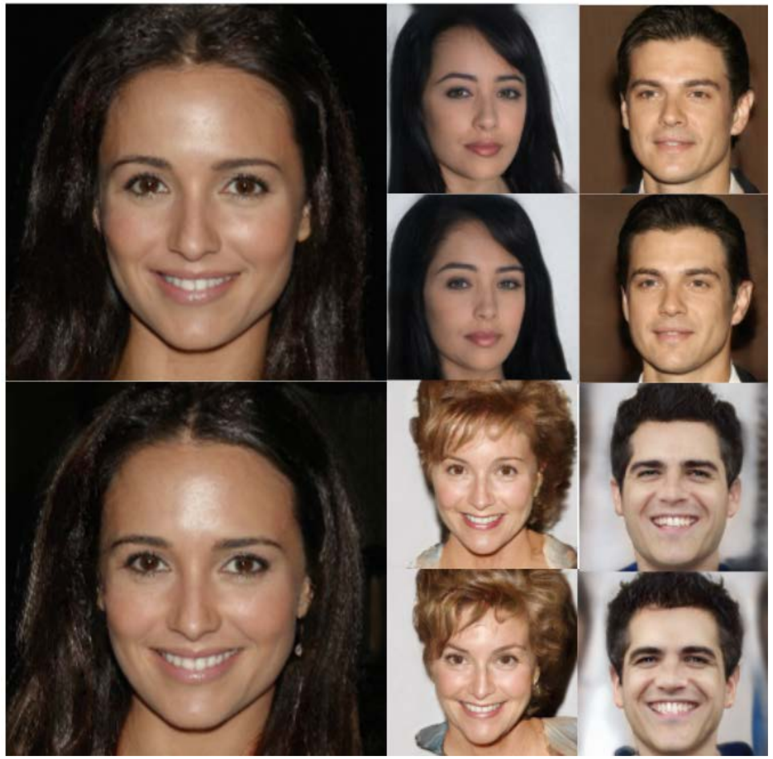
Methods

- **step2**: 3D-aware I2I translation architecture adapted from the trained multi-class StyleNeRF (**step1**). This initialization inherits the capacity of being sensitive of view information.



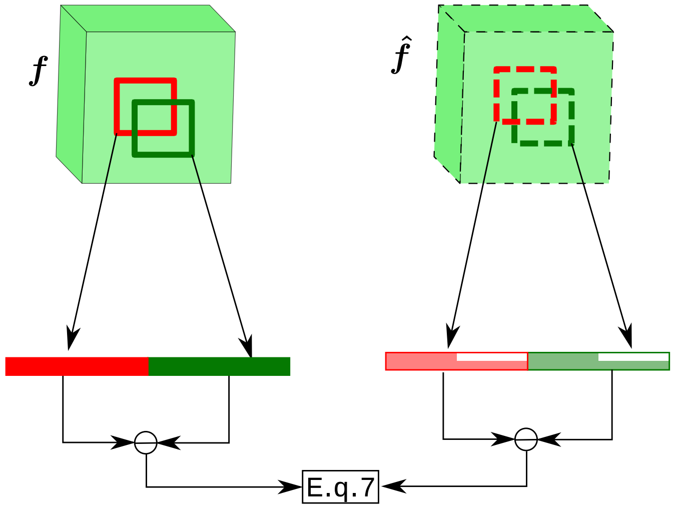
3D-aware I2I translation

- The generated images of step1 (top) and step2 (bottom), which show that we correctly align the outputs of both the NeRF mode F and the adaptor A.

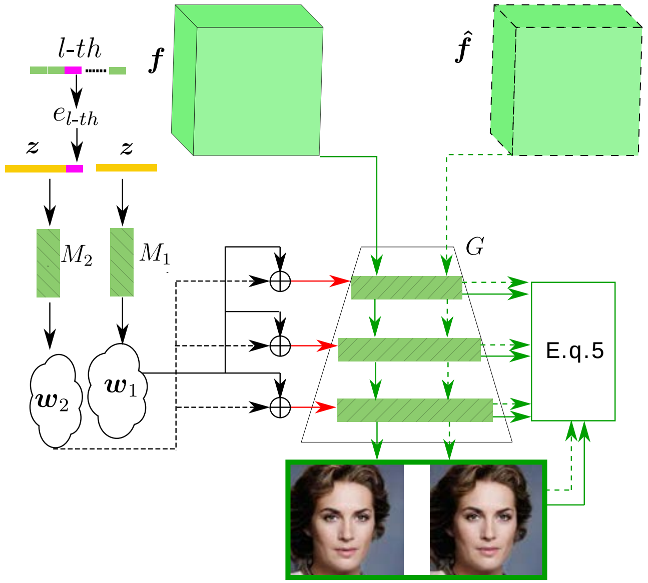


Methods

- several techniques for **step2**: **relative regularization loss** and **hierarchical representation constrain**



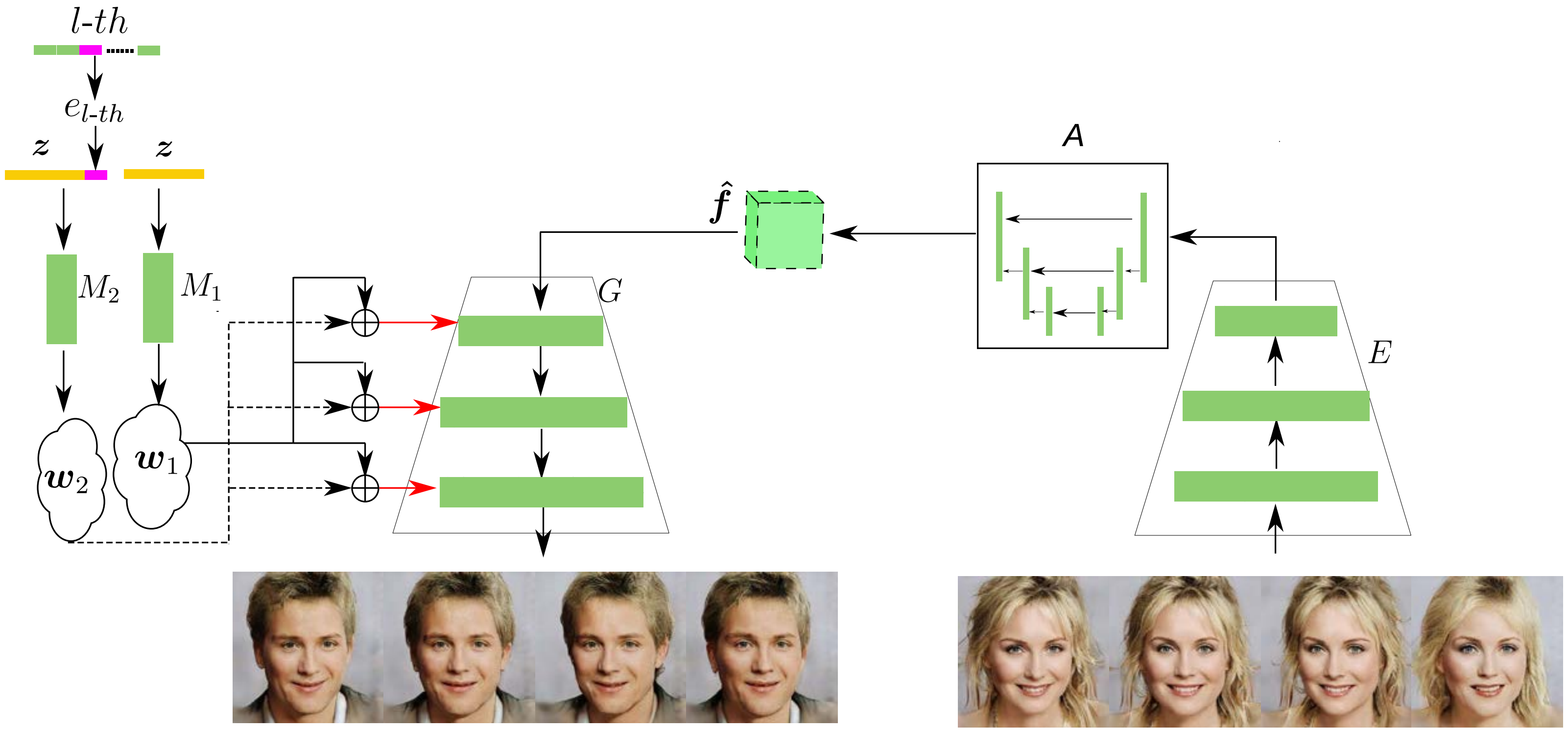
relative regularization loss



hierarchical representation constrain

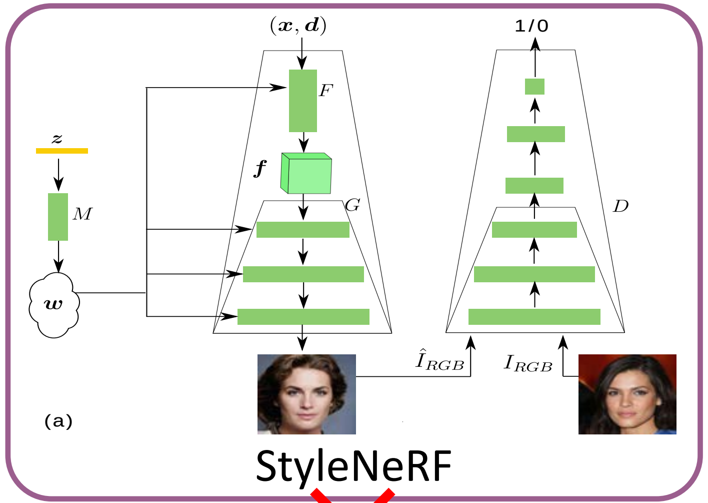
Inference time

- **inference**: the 3D image (e.g. female) is fed into the trained encoder E, and through the adaptor A and generator G , it is eventually translated into other categories of 3D image (e.g. male).



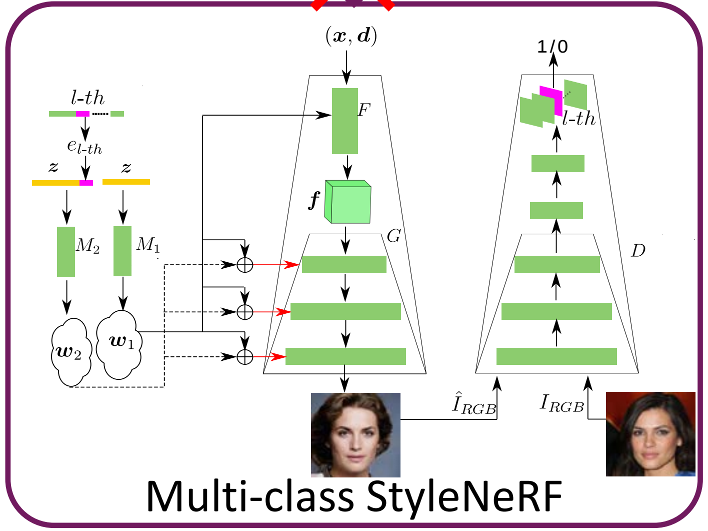
Ablation study

- multi-class StyleNeRF (**step1**) training from scratch (top) causes artifact and mode collapse.

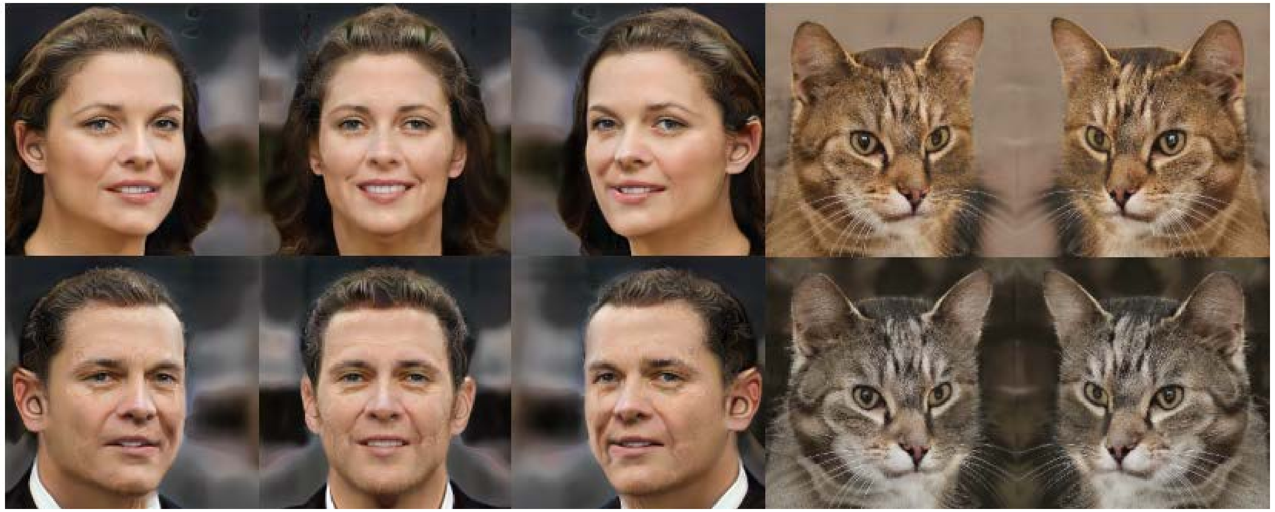


StyleNeRF

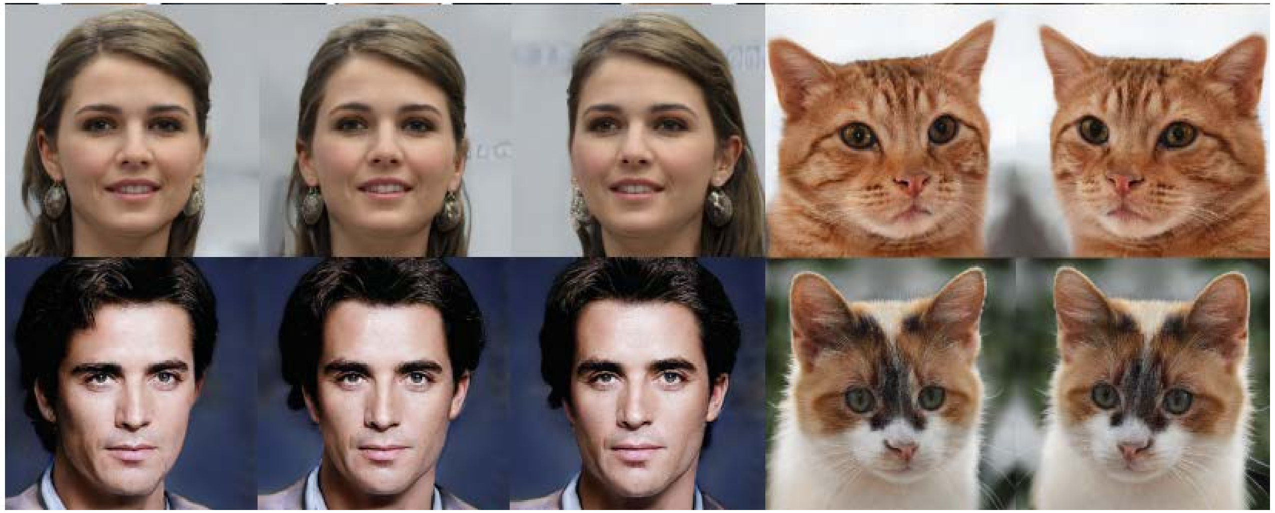
Do not initialize by StyleNeRF weight



Multi-class StyleNeRF



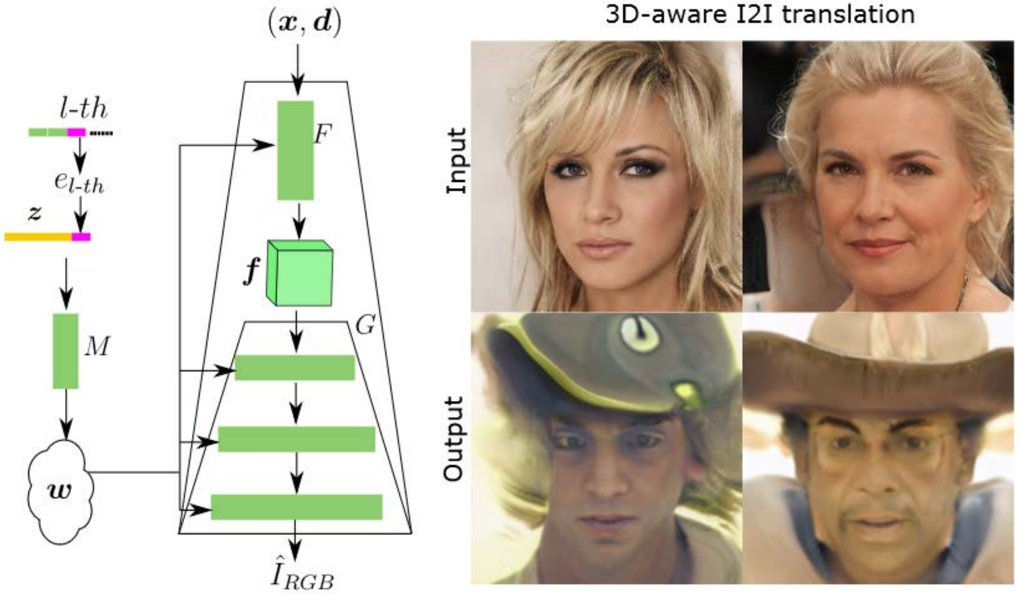
Multi-class StyleNeRF (from scratch)

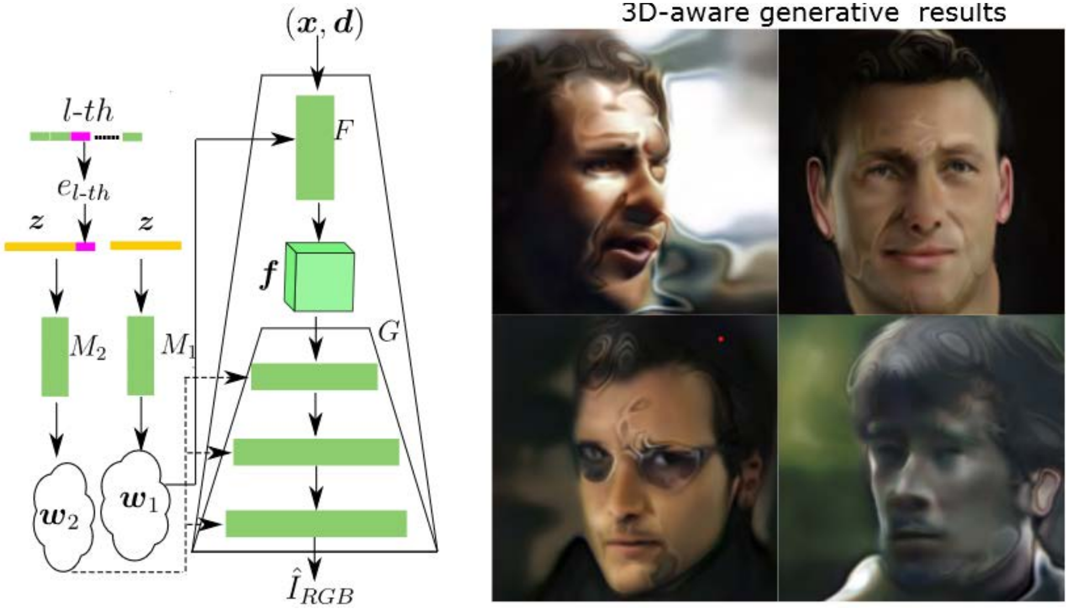


Muti-class StyleNeRF (Ours, initialize by StyleNeRF)

Ablation study

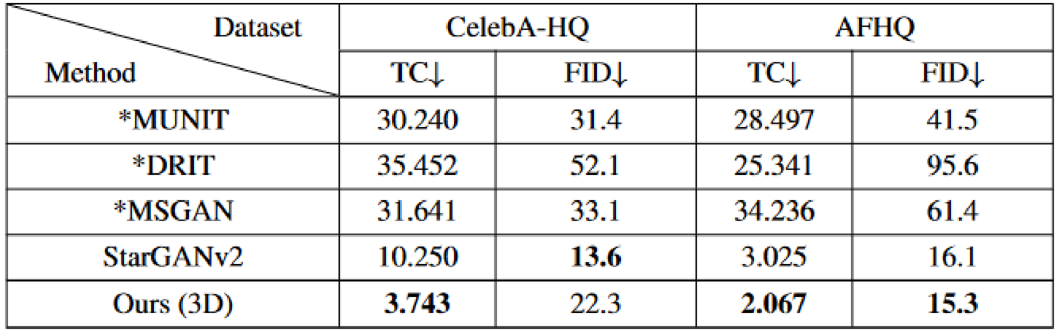
- Both using **a single mapping network (left)** and using two mapping networks **without concatenating (right)** fails to generate satisfactory results.



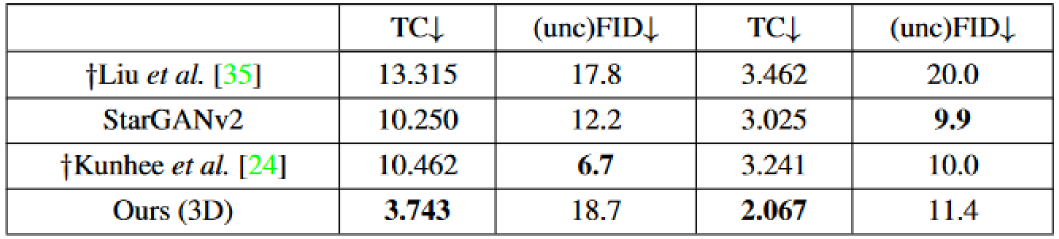


Ablation study

- Comparison with baselines.* denotes that we used the results provided by StarGANv2. † means that we used the pre-trained networks provided by authors.

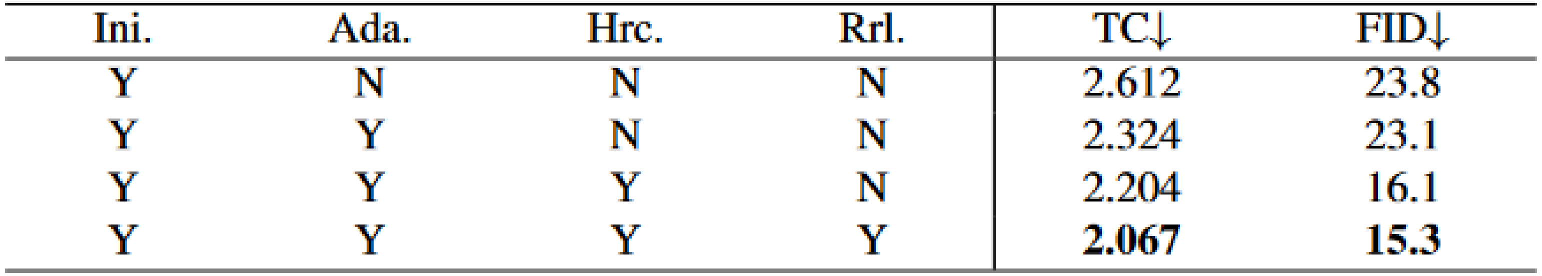


Dataset / Method	CelebA-HQ		AFHQ	
	TC↓	FID↓	TC↓	FID↓
*MUNIT	30.240	31.4	28.497	41.5
*DRIT	35.452	52.1	25.341	95.6
*MSGAN	31.641	33.1	34.236	61.4
StarGANv2	10.250	**13.6**	3.025	16.1
Ours (3D)	**3.743**	22.3	**2.067**	**15.3**



	TC↓	(unc)FID↓	TC↓	(unc)FID↓
†Liu *et al.* [35]	13.315	17.8	3.462	20.0
StarGANv2	10.250	12.2	3.025	**9.9**
†Kunhee *et al.* [24]	10.462	**6.7**	3.241	10.0
Ours (3D)	**3.743**	18.7	**2.067**	11.4

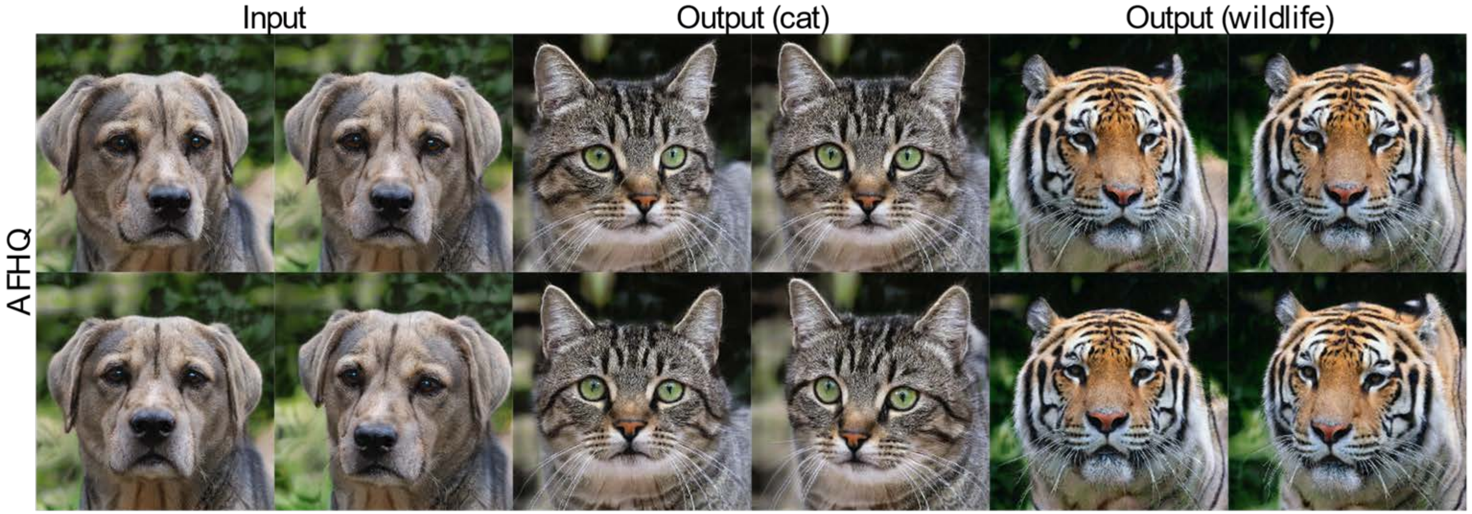
- Impact of several components in the performance on AFHQ. Ini.: initialization method for multi-class StyleNeRF, Ada.: Unet-like adaptor, Hrc.: Hierarchical representation constrain, Rrl.: Relative regularization loss.

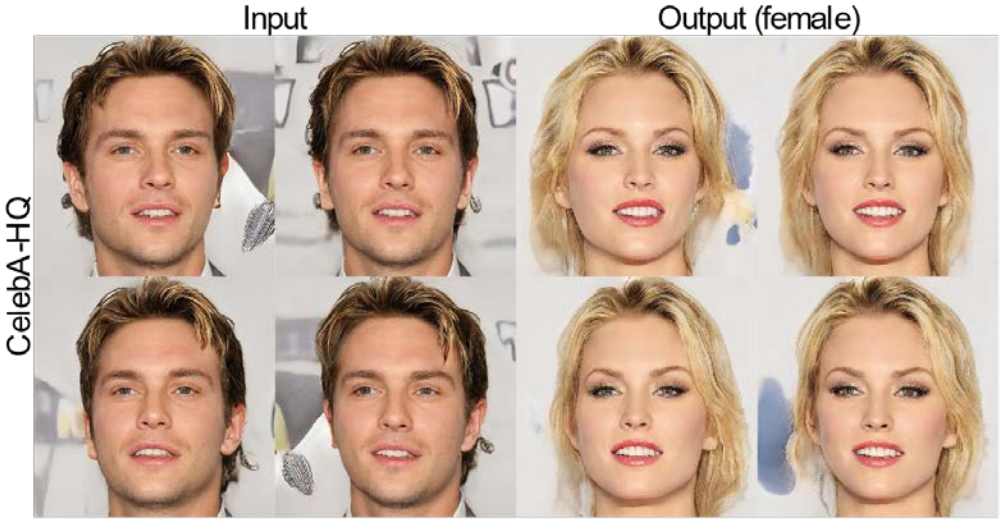


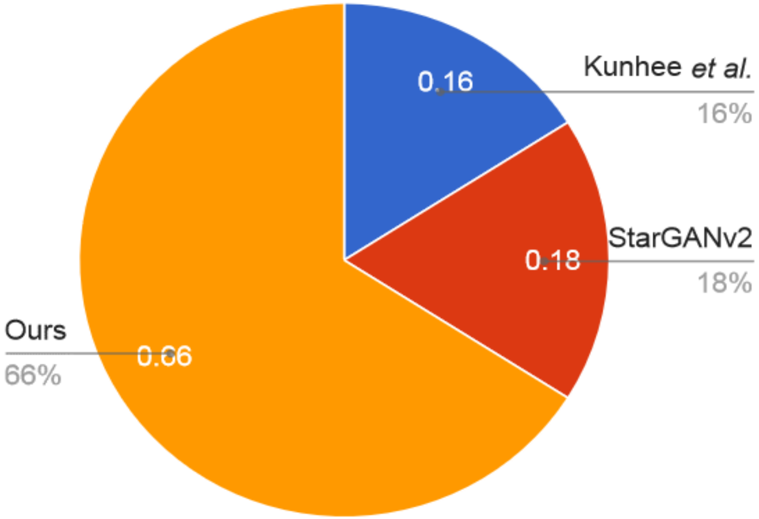
Ini.	Ada.	Hrc.	Rrl.	TC↓	FID↓
Y	N	N	N	2.612	23.8
Y	Y	N	N	2.324	23.1
Y	Y	Y	N	2.204	16.1
Y	Y	Y	Y	**2.067**	**15.3**

Results

- Our approach produces consistent results across viewpoints (up and bottom, left). User study (bottom, right).

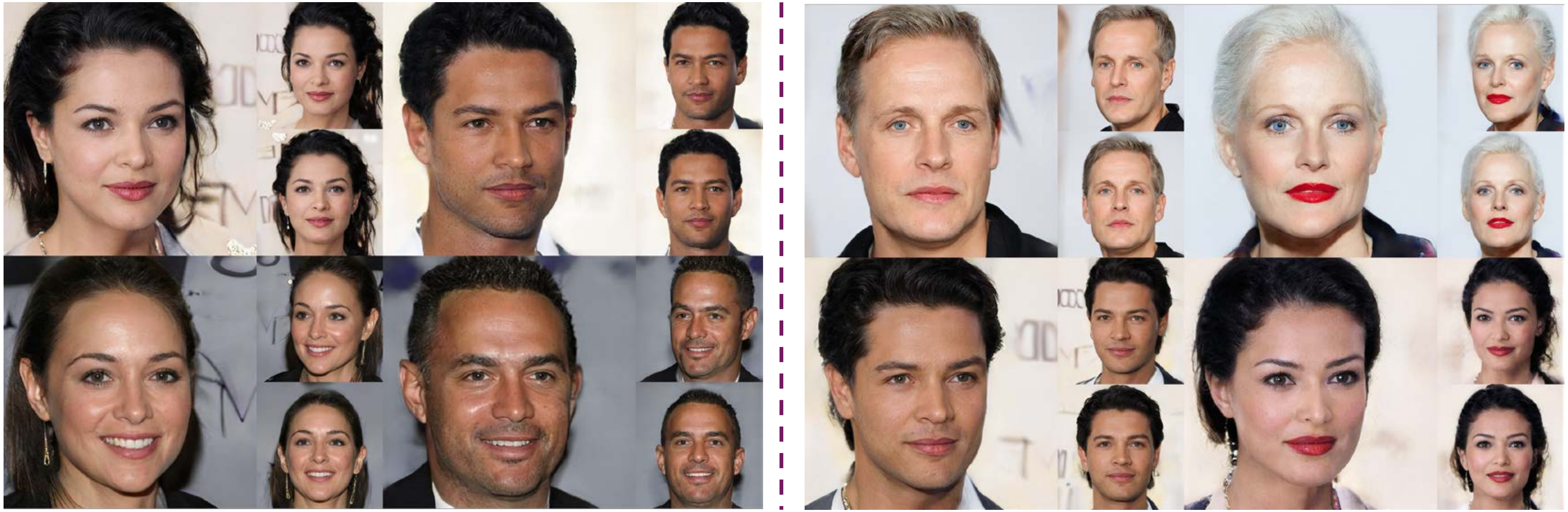






Results

- More results of 3D-aware I2I translation of **female into male (top)** and **male into female (bottom)** on Celeba-HQ 1024×1024



Conclusion

- We are the first to explore 3D-aware multi-class I2I translation, which allows generating 3D consistent videos.
- We decouple 3D-aware I2I translation into two steps. **Step1**: we propose a multi-class StyleNeRF. To train this multi-class StyleNeRF effectively, we provide a new training strategy. **Step2**: we propose a 3D-aware I2I translation architecture.
- To further address the view-inconsistency problem of 3D-aware I2I translation, we propose several techniques: (1) a unet-like adaptor, (2) a hierarchical representation constraint and (3) a relative regularization loss.