Hijack-GAN: Unintended-Use of Pretrained, Black-Box GANs

Hui-Po Wang[1] Ning Yu[2,3] Mario Fritz[1]

[1]CISPA Helmholtz Center for Information Security, Germany [2]University of Maryland, College Park

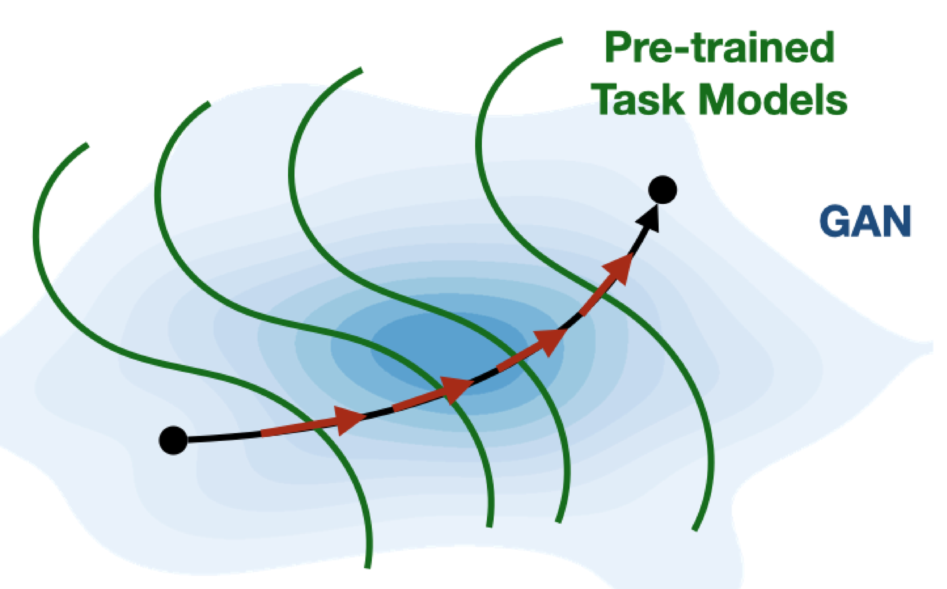
[3]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

Overview

We propose a framework to reuse unconditional, pre-trained, and black-box GANs to achieve novel vision tasks beyond the original intentions. It is important for, e.g.:

- Reusable Green AI Systems
- Understanding Misuse of Released GANs

- **Our Proposed Method (Hijack-GAN)**
  In contrast to prior linear work, HijackGAN edits latent codes in the directions that follows the real manifold and are dynamically decided in each step.



Contributions

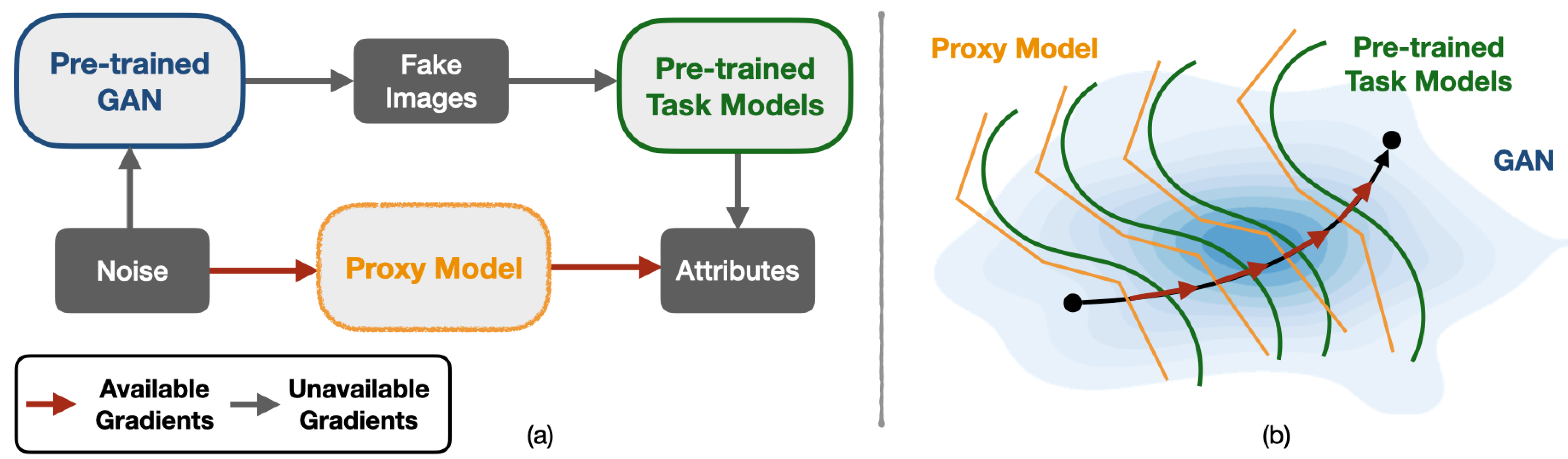
- Propose a framework, HijackGAN, which adapts pre-trained black-box GANs to novel tasks by dynamically traversing the latent.
- Outperform prior work in smoothness, effectiveness, and content preservation.
- Shed light on the potential risks of unintended usage by gaining control over facial attributes, head poses, and landmarks.

Project Page

More results and code can found in our project page `https://a514514772.github.io/hijackgan/`.

Hijack-GAN

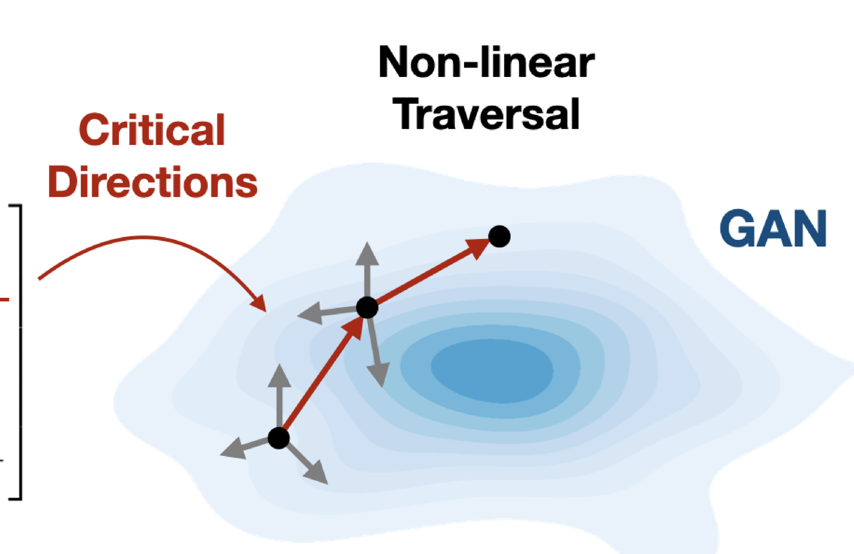
- **Impassable Gradients and the Proxy Model**
  Suppose we can only access a black-box GAN and the desired task model. We circumvent the impassable gradients by training a proxy model. The model is trained with the noise-attribute query pairs. After that, we traverse the latent space under the guidance of gradients from the proxy model.



- **Non-linear Traversal**
  By optimizing toward the critical direction, the target attribute will be activated accordingly.

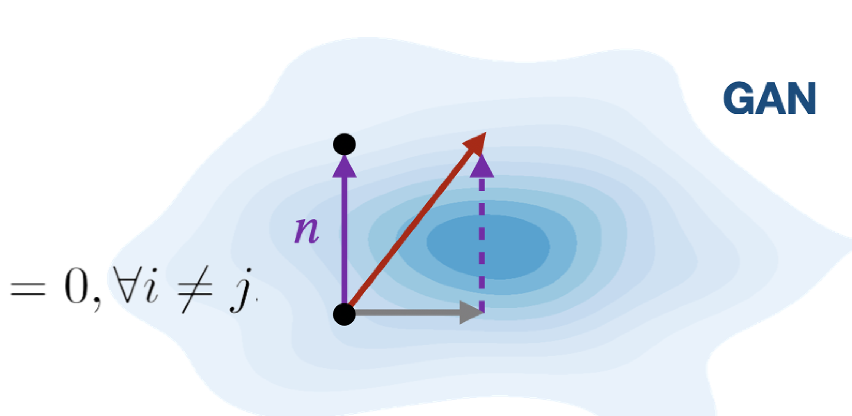
Jacobian Matrix

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathcal{P}_1}{\partial z_1} & \cdots & \frac{\partial \mathcal{P}_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{P}_m}{\partial z_1} & \cdots & \frac{\partial \mathcal{P}_m}{\partial z_n} \end{bmatrix}$$



- **Disentanglement Constraint**
  We derive a direction to reduce the effect on the other non-target attributes to preserve the contents.

$$\underset{n}{\text{maximize}} \quad \mathbf{J}_j^T n$$
$$\text{subject to} \quad \mathbf{J}_i^T n = 0, \forall i \neq j$$



Experimental Results

- **Smoothness** (mPPL)
  Similarity between two random adjacent points on the path.

	Cond.	Eyeglass	Gender	Smile	Age
		PGGAN			
InterfaceGAN	N	**60.69**	65.00	**54.49**	61.16
Ours		64.10	**62.50**	56.55	**60.28**
InterfacGAN	Y	**54.51**	**55.65**	55.07	56.59
Ours		57.40	59.79	**55.00**	**55.30**
		StyleGAN			
InterfaceGAN	N	**99.15**	101.65	**96.90**	96.62
Ours		103.08	**97.93**	97.93	**91.86**
InterfaceGAN	Y	80.46	81.52	92.75	83.51
Ours		**57.23**	**71.67**	**78.18**	**55.41**

- **Function Approximation**
  The more accurate the gradients are, the lower the errors are induced.

	< 1	< 2	< 3	$>= 3$	Avg.
			Eyeglasses		
InterfaceGAN	1.760	2.779	3.644	**1.481**	2.416
Ours	**1.675**	**2.401**	**2.469**	1.557	**2.026**
			Gender		
InterfaceGAN	5.702	4.045	1.798	0.891	3.109
Ours	**4.469**	**3.694**	**1.790**	**0.812**	**2.692**
			Smile		
InterfaceGAN	**1.764**	**1.783**	1.693	0.961	**1.550**
Ours	3.191	2.391	**1.611**	**0.921**	2.028
			Age		
InterfaceGAN	2.350	2.434	2.312	1.354	2.113
Ours	**0.969**	**1.109**	**1.893**	**1.285**	**1.314**

Qualitative Results