Towards Real-World Blind Face Restoration with Generative Facial Prior

Medical **I**maging & **I**ntelligent **R**eality **L**ab.
Convergence Medicine/Radiology

Git: https://github.com/TencentARC/GFPGAN

Paper: https://arxiv.org/pdf/2101.04061.pdf

서울아산병원 의공학연구소

발표자: 경성구

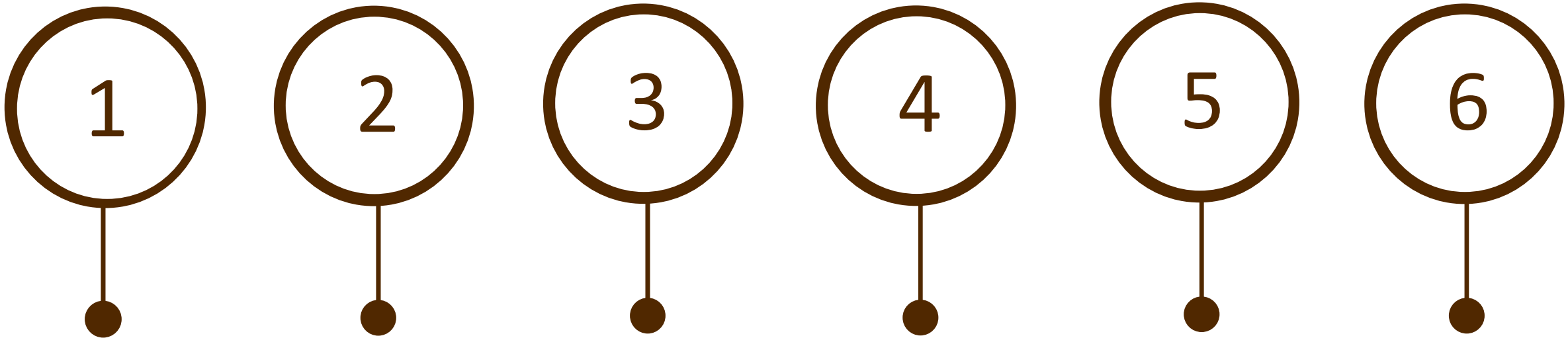
Towards Real-World Blind Face Restoration with Generative Facial Prior

Xintao Wang Yu Li Honglun Zhang Ying Shan
Applied Research Center (ARC), Tencent PCG
{xintaowang, ianyli, honlanzhang, yingsshan}@tencent.com

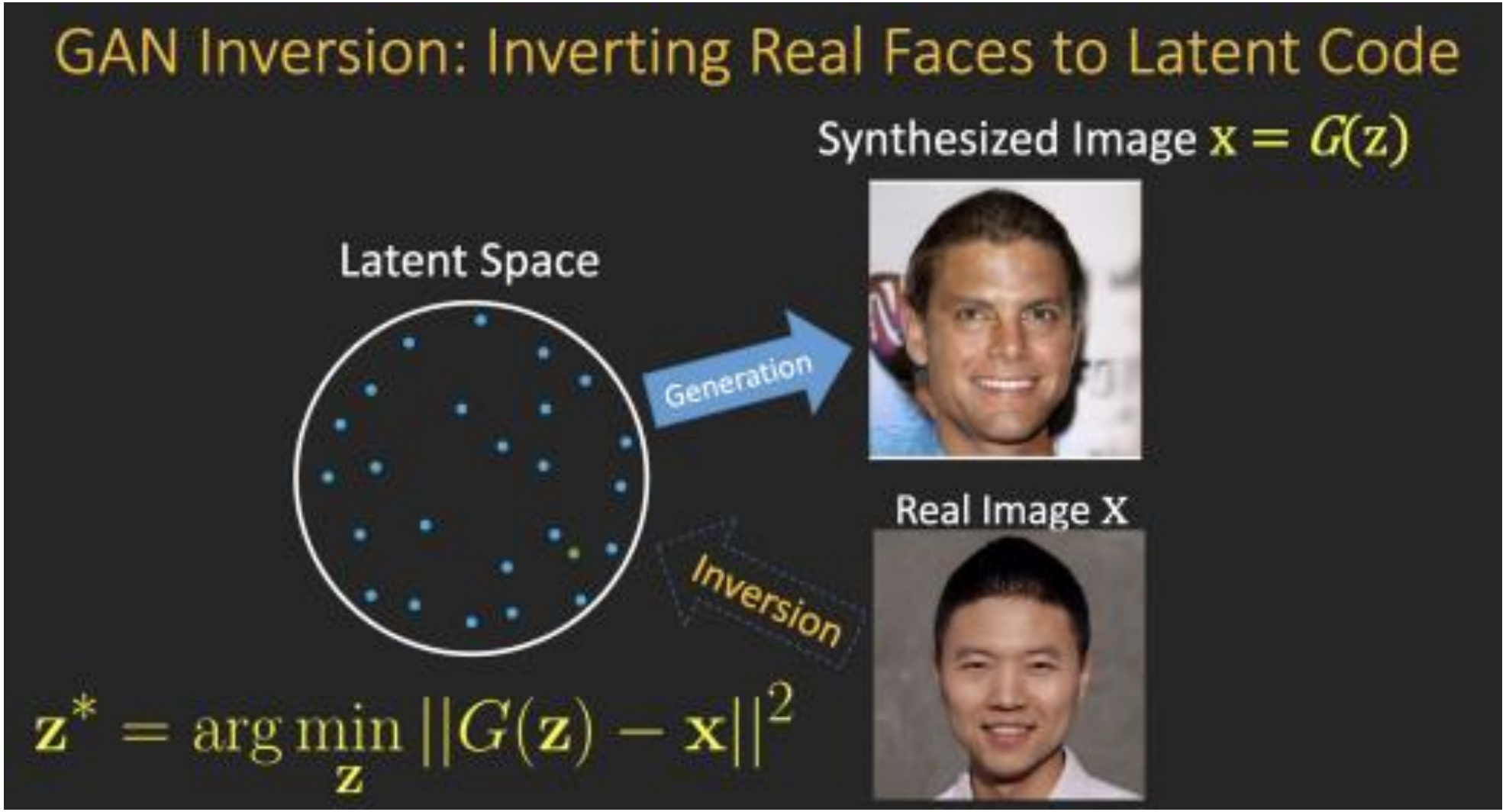
https://github.com/TencentARC/GFPGAN

Figure 1: Comparisons with state-of-the-art face restoration methods: HiFaceGAN [67], DFDNet [44], Wan *et al*. [61] and PULSE [52] on the real-world low-quality images. While previous methods struggle to restore faithful facial details or retain face identity, our proposed GFP-GAN achieves a good balance of realness and fidelity with much fewer artifacts. In addition, the powerful generative facial prior allows us to perform restoration and color enhancement jointly. (**Zoom in for best view**)

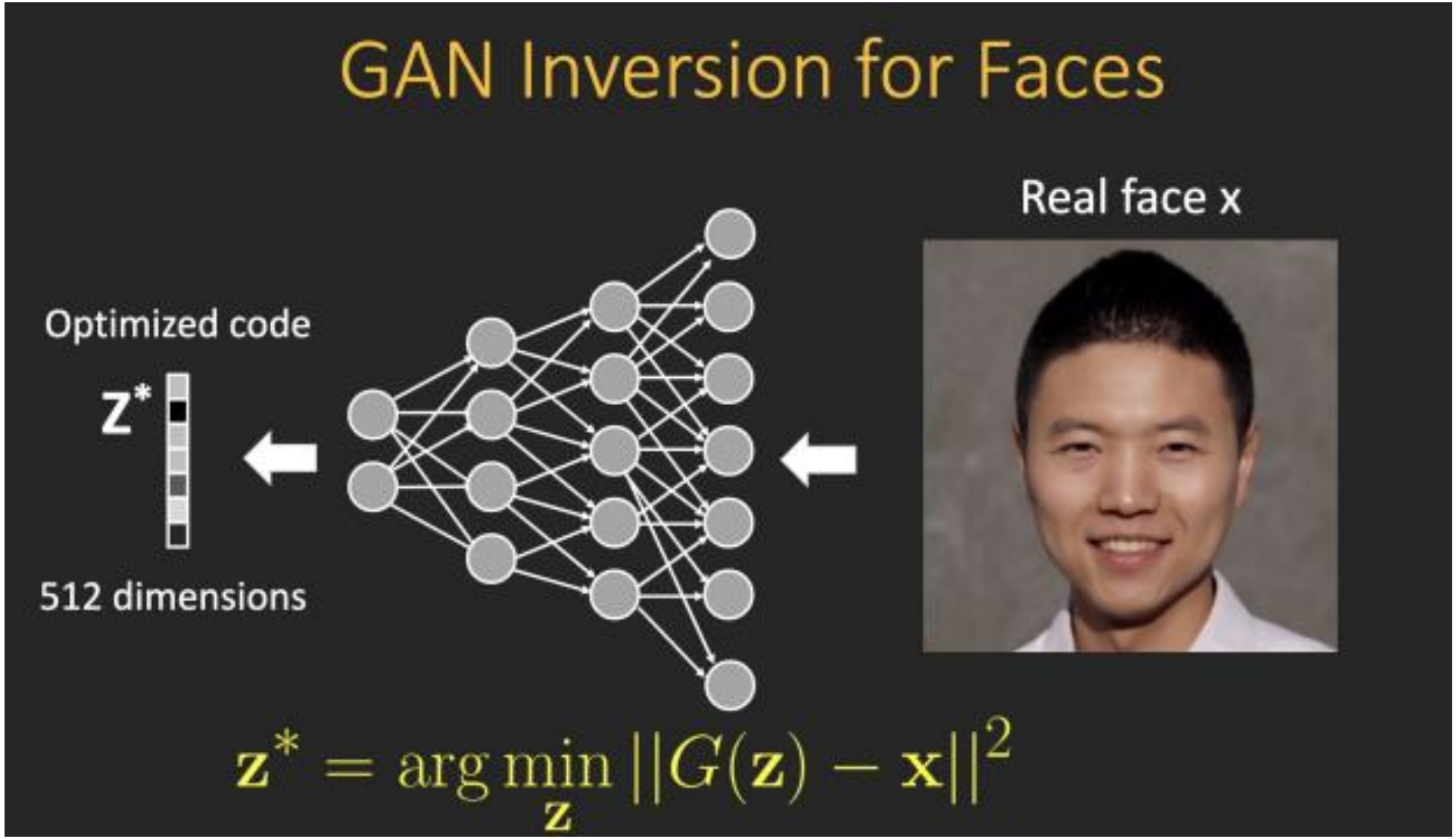
Paper Contents



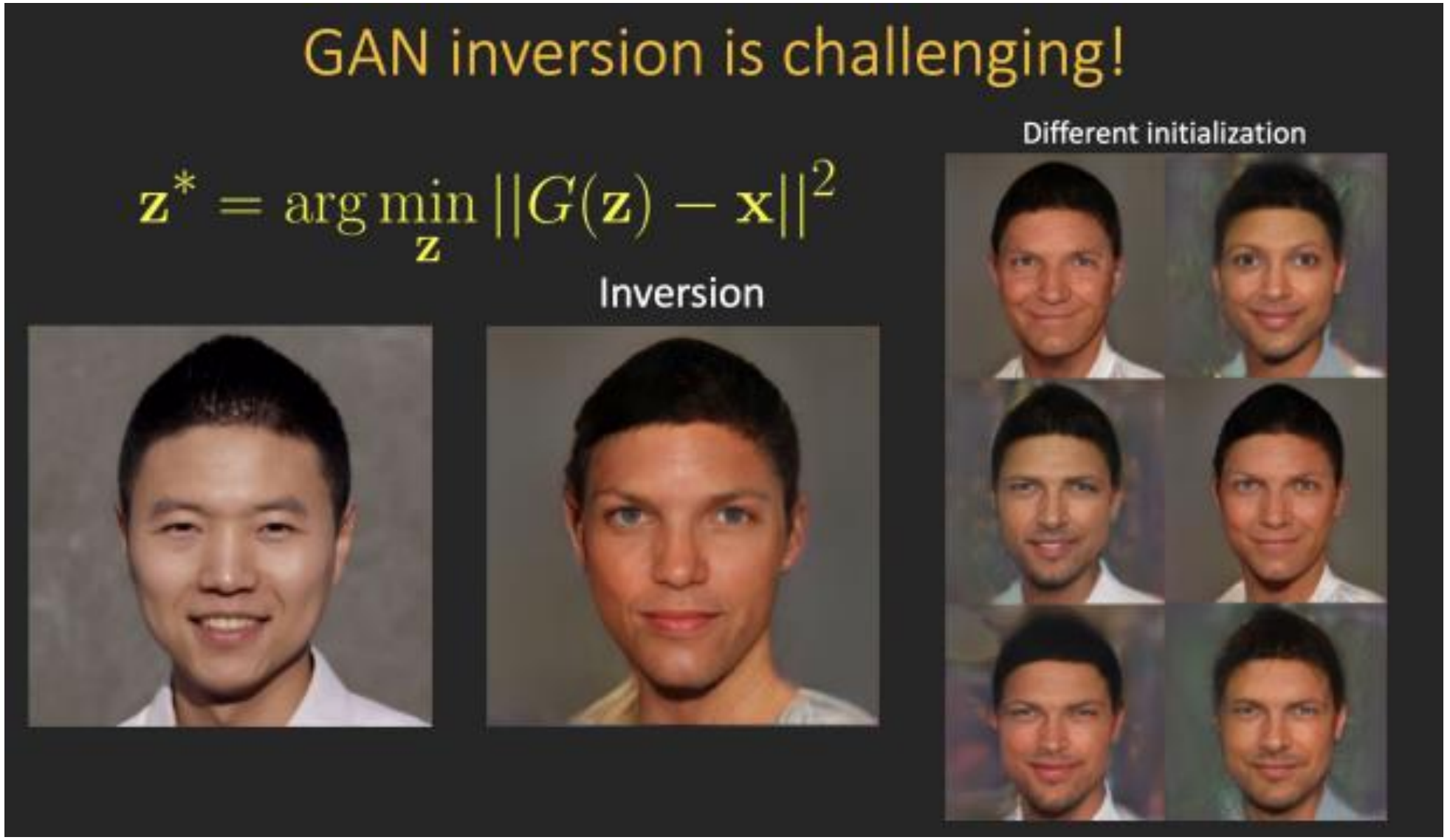
Background – GAN inversion



Background – GAN inversion



Background – GAN inversion



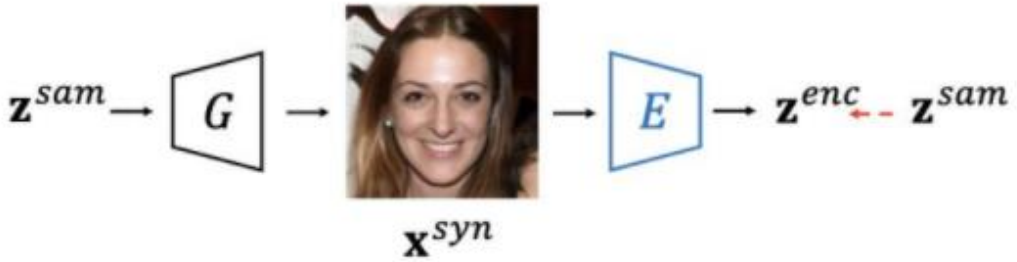
Conventional GAN Inversion Approaches

- Learning-based
  - first synthesizes a collection of images with randomly sampled latent codes
  - then uses the images and codes as inputs and supervisions respectively to train a deterministic model (encoder)
- Optimization-based
  - deals with a single instance at one time
  - directly optimizing the latent code to minimize the pixel-wise reconstruction loss
- Some work combines these two ideas by using the encoder to generate an initialization for optimization
  - In-Domain GAN Inversion is also one of these

Background – GAN inversion

Conventional GAN Inversion Encoder

$$\min_{\Theta_E} \mathcal{L}_E = ||\mathbf{z}^{sam} - E(G(\mathbf{z}^{sam}))||_2,$$



- Training an encoder is commonly used for GAN inversion problem.
  - Existing methods simply learn a deterministic model with no regard to whether the codes produced by the encoder align with the semantic knowledge learned by G(·)
  - A collection of latent codes z^{sam} are randomly sampled and fed into G(·) to get the corresponding synthesis x^{syn}

Towards Real-World Blind Face Restoration with Generative Facial Prior

Xintao Wang Yu Li Honglun Zhang Ying Shan
Applied Research Center (ARC), Tencent PCG
{xintaowang, ianyli, honlanzhang, yingsshan}@tencent.com

https://github.com/TencentARC/GFPGAN

Figure 1: Comparisons with state-of-the-art face restoration methods: HiFaceGAN [67], DFDNet [44], Wan *et al*. [61] and PULSE [52] on the real-world low-quality images. While previous methods struggle to restore faithful facial details or retain face identity, our proposed GFP-GAN achieves a good balance of realness and fidelity with much fewer artifacts. In addition, the powerful generative facial prior allows us to perform restoration and color enhancement jointly. (**Zoom in for best view**)

Abstract

- Blind face restoration usually relies on facial priors, to restore realistic and faithful details. However, **very low-quality inputs cannot offer accurate geometric prior**, limiting the applicability in real-world scenarios.

- In this work, we propose GFP-GAN that leverages rich and diverse priors encapsulated in **a pre-trained face GAN** for blind face restoration.

- This Generative Facial Prior (GFP) is incorporated into the face restoration process via **U-Net**, which allow our method to achieve a good balance of **realness** and **fidelity**.

- Thanks to the powerful generative facial prior and delicate designs, our GFP-GAN could jointly restore **facial details** and **enhance colors** with just **a single forward pass**, while **GAN inversion methods** require image-specific optimization at inference.

- Extensive experiments show that our method achieves superior performance to prior art on **both synthetic and real-world datasets.**

Introduction - issue

- Blind face restoration aims at recovering high-quality faces from the low-quality counterparts suffering from unknown degradation (**low-resolution, noise, blur, compression artifacts**).

<blind face restoration example>

- *face-specific priors*: facial landmarks, parsing maps, facial component heatmaps
  However, those priors are usually estimated from input images and **inevitably degrades with very low-quality inputs** in the real world. (E.g., eye pupil)

- *reference priors*: high-quality guided faces or facial component dictionaries
  However, **the inaccessibility of high-resolution references** limits its practical applicability, while **the limited capacity of dictionaries** restricts its diversity and richness of facial details

Introduction - contribution

- (1) We leverage rich and diverse **generative facial priors** for blind face restoration. Those priors contain sufficient facial textures and color information, allowing us to jointly perform face restoration and color enhancement.

- (2) We propose **the GFP-GAN framework** with delicate designs of architectures **and losses** to incorporate generative facial prior. Our GFP-GAN with CS-SFT layers achieves a good balance of fidelity and texture faithfulness in **a single forward pass**.

- (3) Extensive experiments show that our method achieves superior performance to prior art on **both synthetic and real-world datasets**.

Related work

- **Image Restoration**
  - Image Restoration typically includes super-resolution, denoising, deblurring and compression removal

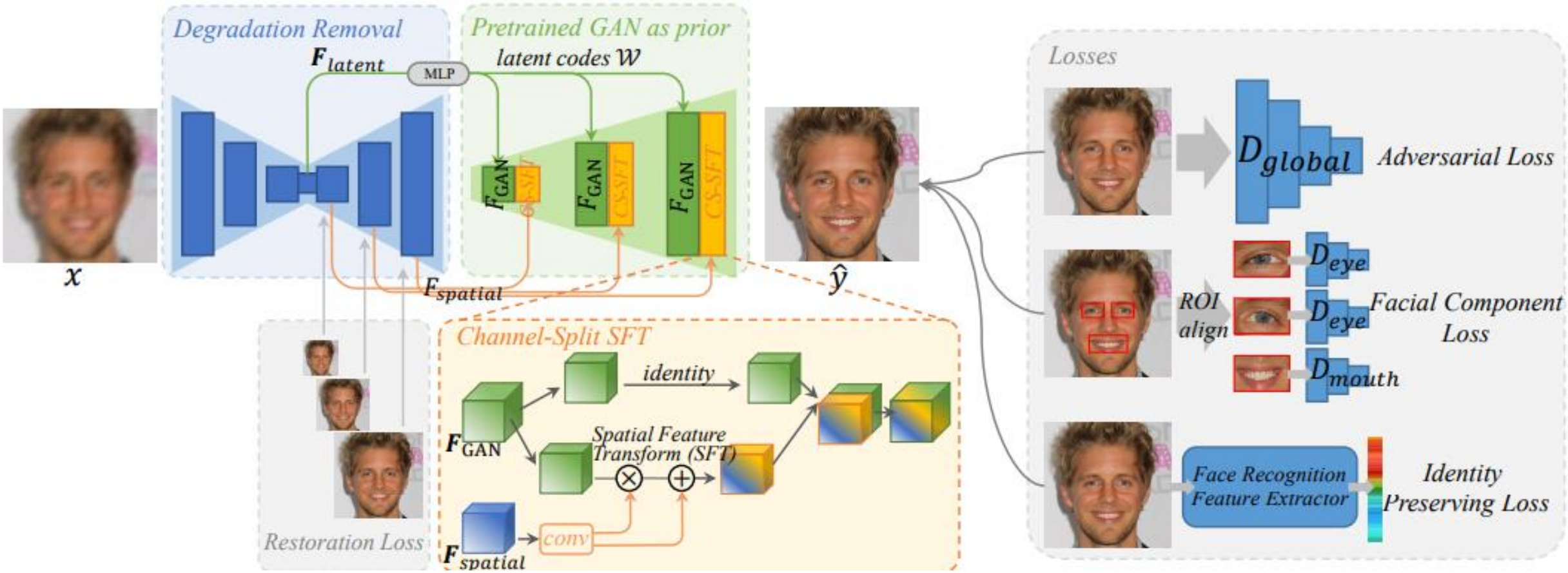
- **Face Restoration**
  - geometry priors and reference priors

- **Generative Priors**
  - pre-trained GANs is previously exploited by GAN inversion, whose primary aim is to find the closest latent codes given an input image

- **Channel Split Operation**
  - It is used to design compact models and improve model representation ability (ex. MobileNet, GhostNet)

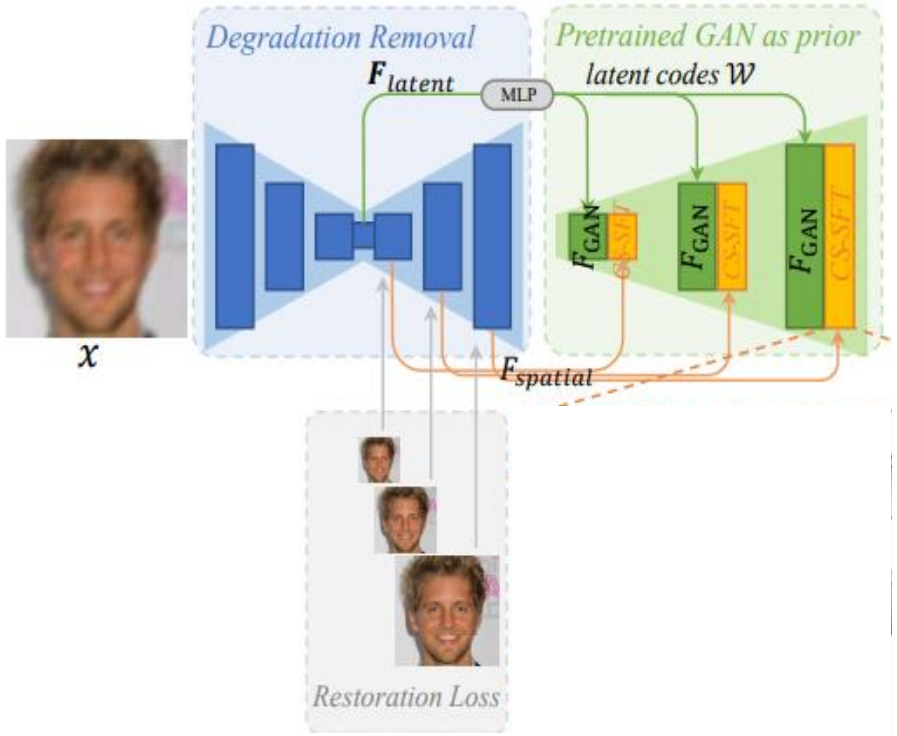
- **Local Component Discriminators**
  - Local discriminator is proposed to focus on local patch distributions

Methods



1) intermediate restoration losses to remove complex degradation
2) Facial component loss with discriminators to enhance facial details
3) identity preserving loss to retain face identity.

Methods - Degradation Removal Module



- The degradation removal module is designed to explicitly **remove the above degradation** and **extract 'clean' features F_latent** and **F_spatial, alleviating the burden of subsequent modules.**

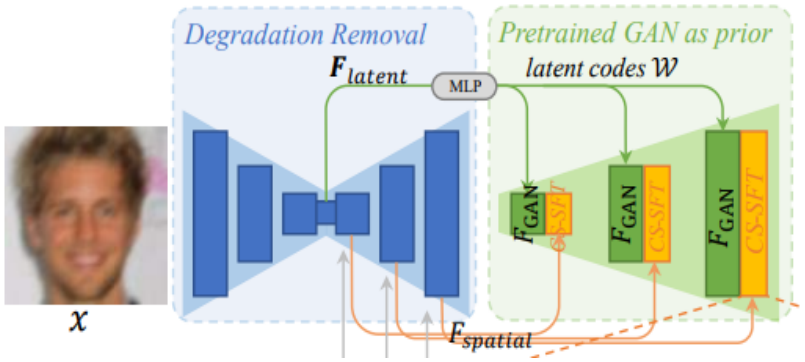
- U-Net structure as our degradation remove module
  1) increase receptive field for large blur elimination.
  2) generate multi-resolution features.

(*F_latent is used to map the input image to the closest latent code in StyleGAN2.)
(*F_spatial are used to modulate the StyleGAN2 features.)

- In order to have intermediate supervision for removing degradation, we employ the **L1 restoration loss** in each resolution scale in the early stage of training.

- Specifically, we also output images for each resolution scale of the U-Net decoder, and **then restrict these outputs to be close to the pyramid** of the ground-truth image.

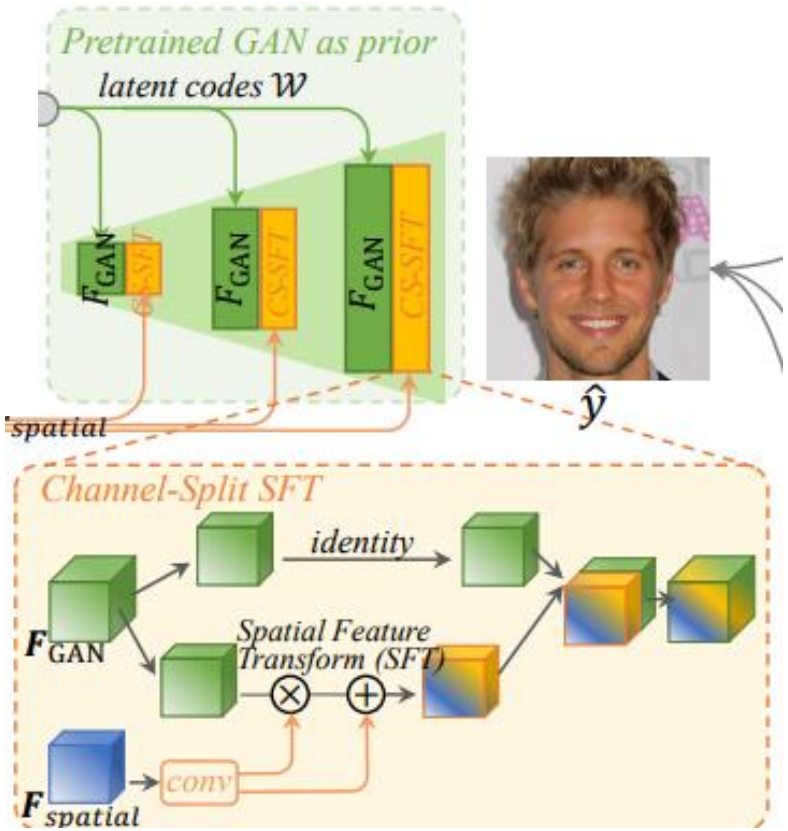
Methods - Generative Facial Prior and Latent Code Mapping



$$\mathcal{W} = \texttt{MLP}(\boldsymbol{F}_{latent}),$$
$$\boldsymbol{F}_{\text{GAN}} = \texttt{StyleGAN}(\mathcal{W}).$$

- A typical way of deploying generative priors is to map the input image to its closest latent codes **Z**, and then generate the corresponding output by a pre-trained GAN.
- However, these methods usually require time-consuming iterative optimization for preserving fidelity.
- Instead of producing a final image directly, we generate **intermediate convolutional features F_GAN of the closest face**, as it contains more details and could be further modulated by input features for better fidelity.
- Specifically, given the encoded vector F_latent of the input image, **we first map it to intermediate latent codes W** for better preserving semantic property i.e., the intermediate space transformed from **Z** with several MLP.
- The latent codes **W** then pass through each convolution layer in the pre-trained GAN, and generate F_GAN features for each resolution scale

Methods - Channel-Split Spatial Feature Transform



- In order to **better preserve fidelity**, we further use the input **spatial features F_spatial** to modulate the GAN features F_GAN.

- Preserving spatial information from inputs is crucial for face restoration, as it usually requires **local characteristics** for fidelity preservation, and adaptive restoration at **different spatial locations** of a face.

- Therefore, we employ Spatial Feature Transform (SFT), which generates affine transformation parameters for **spatial-wise feature modulation**:

$$\alpha, \beta = \texttt{Conv}(\boldsymbol{F}_{spatial}),$$
$$\boldsymbol{F}_{output} = \texttt{SFT}(\boldsymbol{F}_{\text{GAN}}|\alpha, \beta) = \alpha \odot \boldsymbol{F}_{\text{GAN}} + \beta.$$

- To achieve a better balance of realness and fidelity, we further propose **channel-split** spatial feature transform (CSSFT) layers, which perform spatial modulation on part of the GAN features by input features F_spatial (**contributing to fidelity**) and leave the left GAN features (**contributing to realness**) to directly pass through :

$$\begin{aligned}\boldsymbol{F}_{output} &= \texttt{CS-SFT}(\boldsymbol{F}_{\text{GAN}}|\alpha, \beta) \\ &= \texttt{Concat}[\texttt{Identity}(\boldsymbol{F}_{\text{GAN}}^{split0}), \alpha \odot \boldsymbol{F}_{\text{GAN}}^{split1} + \beta],\end{aligned} \tag{4}$$

where $\boldsymbol{F}_{\text{GAN}}^{split0}$ and $\boldsymbol{F}_{\text{GAN}}^{split1}$ are split features from $\boldsymbol{F}_{\text{GAN}}$ in channel dimension, and $\texttt{Concat}[\cdot,\cdot]$ denotes the concatenation operation.

- As a result, CS-SFT enjoys the benefits of directly **incorporating prior information and effective modulating by input images**, thereby achieving a good balance between texture faithfulness and fidelity. Besides, CS-SFT could **also reduce complexity** as it requires fewer channels for modulation. We conduct channel-split SFT layers at each resolution scale, and finally generate a restored face.

Methods - Channel-Split Spatial Feature Transform

Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform

Xintao Wang [1] Ke Yu[1] Chao Dong[2] Chen Change Loy[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong, [2]SenseTime Research
{wx016, yk017, ccloy}@ie.cuhk.edu.hk dongchao@sensetime.com

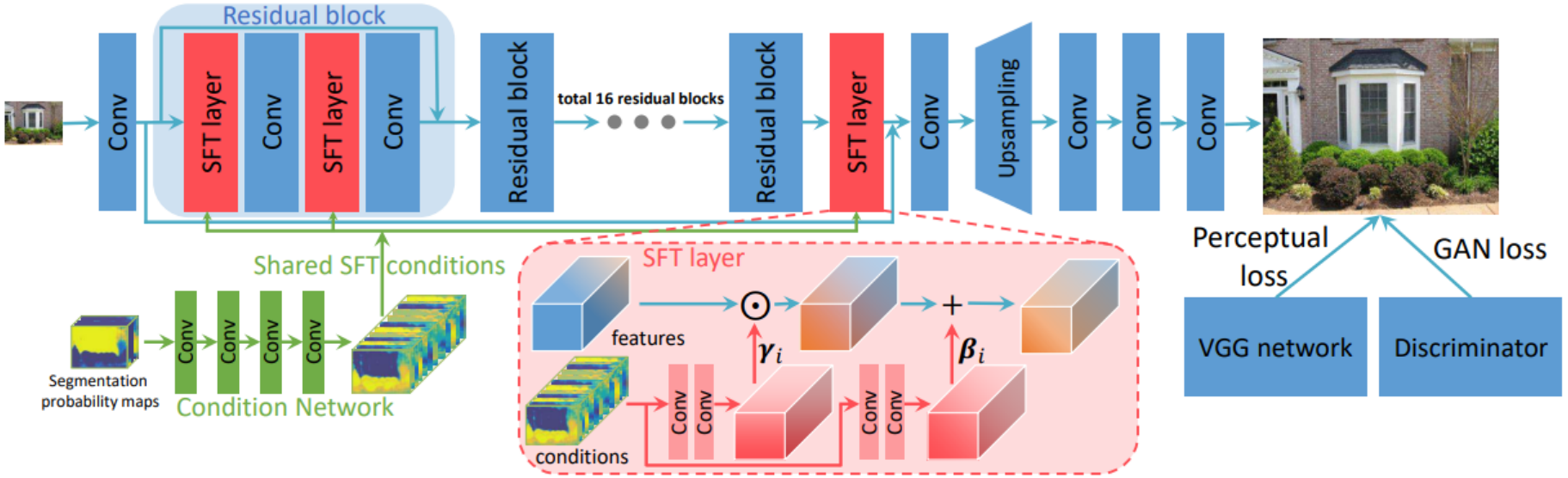
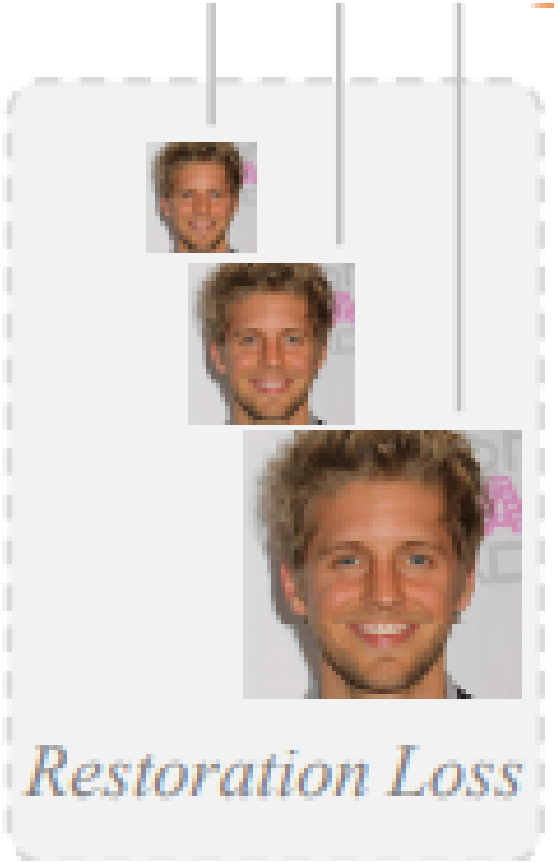
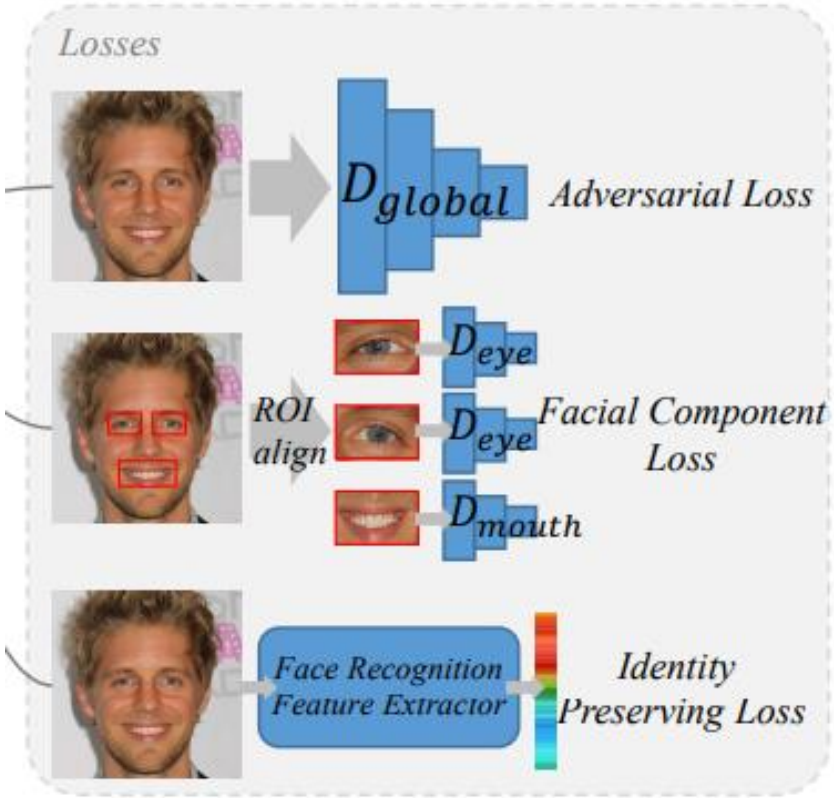


Figure 3. The proposed SFT layers can be conveniently applied to existing SR networks. All SFT layers share a condition network. The role of the condition network is to generate intermediate conditions from the prior, and broadcast the conditions to all SFT layers for further generation of modulation parameters.

Methods – model Objectives





Methods – model Objectives

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{comp} + \mathcal{L}_{id}. \quad (9)$$

The loss hyper-parameters are set as follows: $\lambda_{l1} = 0.1$, $\lambda_{per} = 1$, $\lambda_{adv} = 0.1$, $\lambda_{local} = 1$, $\lambda_{fs} = 200$ and $\lambda_{id} = 10$.

(**y_hat**: the outputs, **y**: close to the ground-truth)

Reconstruction Loss

$$\mathcal{L}_{rec} = \lambda_{l1}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_1 + \lambda_{per}\|\phi(\hat{\boldsymbol{y}}) - \phi(\boldsymbol{y})\|_1, \quad (5)$$

where ϕ is the pretrained VGG-19 network [59] and we use the $\{\texttt{conv1}, \cdots, \texttt{conv5}\}$ feature maps before activation [64]. λ_{l1} and λ_{per} denote the loss weights of L1 and perceptual loss, respectively.

Adversarial Loss

$$\mathcal{L}_{adv} = -\lambda_{adv}\mathbb{E}_{\hat{\boldsymbol{y}}}\,\texttt{softplus}(D(\hat{\boldsymbol{y}})) \quad (6)$$

where D denotes the discriminator and λ_{adv} represents the adversarial loss weight.

Facial Component Loss

$$\mathcal{L}_{comp} = \sum_{\texttt{ROI}} \lambda_{local}\,\mathbb{E}_{\hat{\boldsymbol{y}}_{\texttt{ROI}}}[\log(1 - D_{\texttt{ROI}}(\hat{\boldsymbol{y}}_{\texttt{ROI}}))] + \lambda_{fs}\,\|\texttt{Gram}(\psi(\hat{\boldsymbol{y}}_{\texttt{ROI}})) - \texttt{Gram}(\psi(\boldsymbol{y}_{\texttt{ROI}}))\|_1 \quad (7)$$

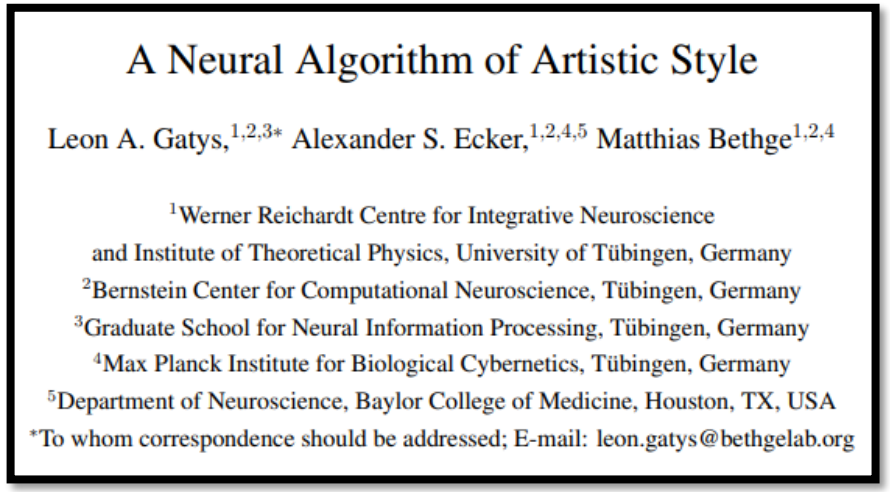
where ROI is region of interest from the component collection $\{\texttt{left\_eye}, \texttt{right\_eye}, \texttt{mouth}\}$. $D_{\texttt{ROI}}$ is the local discriminator for each region. ψ denotes the multi-resolution features from the learned discriminators. λ_{local} and λ_{fs} represent the loss weights of local discriminative loss and feature style loss, respectively.

Identity Preserving Loss

$$\mathcal{L}_{id} = \lambda_{id}\|\eta(\hat{\boldsymbol{y}}) - \eta(\boldsymbol{y})\|_1, \quad (8)$$

where η represents face feature extractor, *i.e.* ArcFace [10] in our implementation. λ_{id} denotes the weight of identity preserving loss.

Methods – model Objectives



A Neural Algorithm of Artistic Style

Leon A. Gatys,[1,2,3*] Alexander S. Ecker,[1,2,4,5] Matthias Bethge[1,2,4]

[1]Werner Reichardt Centre for Integrative Neuroscience
and Institute of Theoretical Physics, University of Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, Tübingen, Germany
[3]Graduate School for Neural Information Processing, Tübingen, Germany
[4]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[5]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
*To whom correspondence should be addressed; E-mail: leon.gatys@bethgelab.org

- **Style Loss**

다음으로, style에 대한 loss를 정의해보자. 이 논문에서 style이라는 것은 같은 layer의 서로 다른 filter들끼리의 correlation으로 정의한다. 즉, filter가 N_l개 있으므로 이것들의 correlation은 $G^l \in \mathcal{R}^{N_l \times N_l}$이 될 것이다. 이때, correlation을 계산하기 위하여 각각의 filter의 expectation 값을 사용하여 correlation matrix를 계산한다고 한다. 즉, l번째 layer에서 필터가 100개 있고, 각 필터별로 output이 400개 있다면, 각각의 100개의 필터마다 400개의 output들을 평균내어 값을 100개 뽑아내고, 그 100개의 값들의 correlation을 계산했다는 것이다. 이렇게 계산한 matrix를 Gram matrix라고 하며 G^l_{ij}라고 적으며 다음과 같이 계산할 수 있다.

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{kj}.$$

Experiments - Dataset

Train dataset

- Train our GFP-GAN on the **FFHQ dataset**, which consists of 70,000 high-quality images and resize all the images to 5,122 during training.

- Our GFP-GAN is trained on **synthetic data** that approximate to the real low-quality images and generalize to real-world images during inference. We adopt the following degradation model to synthesize training data:

$$\boldsymbol{x} = [(\boldsymbol{y} \circledast \boldsymbol{k}_\sigma) \downarrow_r + \boldsymbol{n}_\delta]_{\text{JPEG}_q}. \qquad (10)$$

- The high quality image $\mathbf{y}$ is first convolved with Gaussian blur kernel $\mathbf{k}$ followed by a down-sampling operation with a scale factor γ. After that, additive white Gaussian noise $\boldsymbol{n}$ is added to the image and finally it is compressed by JPEG with quality factor $\boldsymbol{q}$. for each training pair.
  (randomly sample $\boldsymbol{\sigma}$, γ , δ and $\boldsymbol{q}$ from {0:2 : 10, 1 : 8, 0 : 15, 60 : 100}, respectively)

- add color jittering during training for color enhancement.

Test dataset - one synthetic dataset and three different real datasets

- CelebA-Test, LFW-Test, CelebChild-Test, WebPhoto-Test

Experiments - Implementation

- We adopt the **pre-trained StyleGAN2** with 5,122 outputs as our generative facial prior.
- The U-Net for degradation removal consists of **seven** down-samples and **seven** up-samples, each with a residual block.
- For each CS-SFT layer, we use **two** convolutional layers to generate **the affine parameters** and respectively.
- The training mini-batch size is set to 12.
- We augment the training data with **horizontal flip** and **color jittering**.
- We consider three components: **left eye, right eye, mouth** for face component loss as they are perceptually significant.
- **Each component is cropped by ROI align with face landmarks** provided in the origin training dataset.
- The learning rate was set to 2×10^{-3} and then decayed by a factor of 2 at the 700k-th, 750k-th iterations.
- We implement our models with the PyTorch framework and train them using four NVIDIA Tesla P40 GPUs.

Results

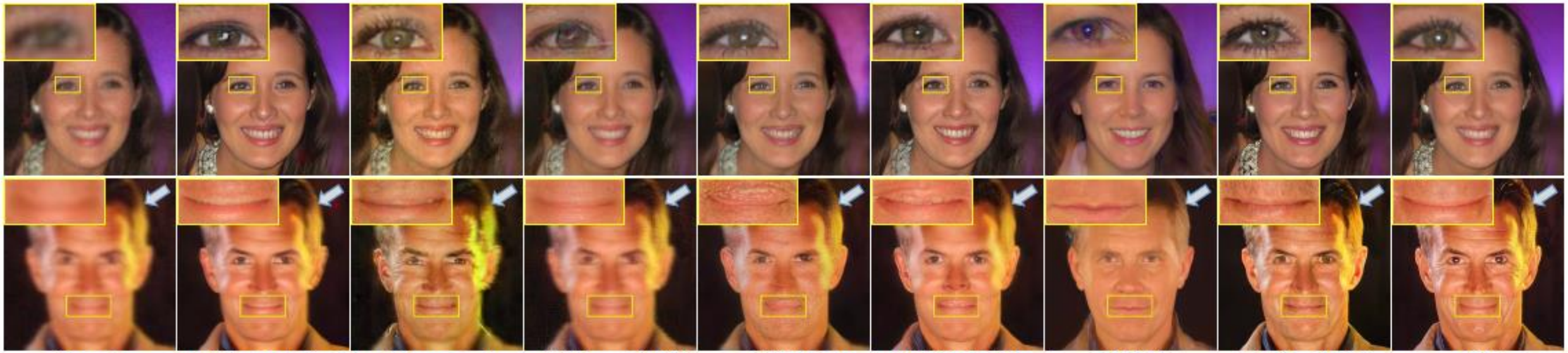


Figure 3: Qualitative comparison on the **CelebA-Test** for blind face restoration. Our GFP-GAN produces faithful details in eyes, mouth and hair. **Zoom in for best view.**

Results

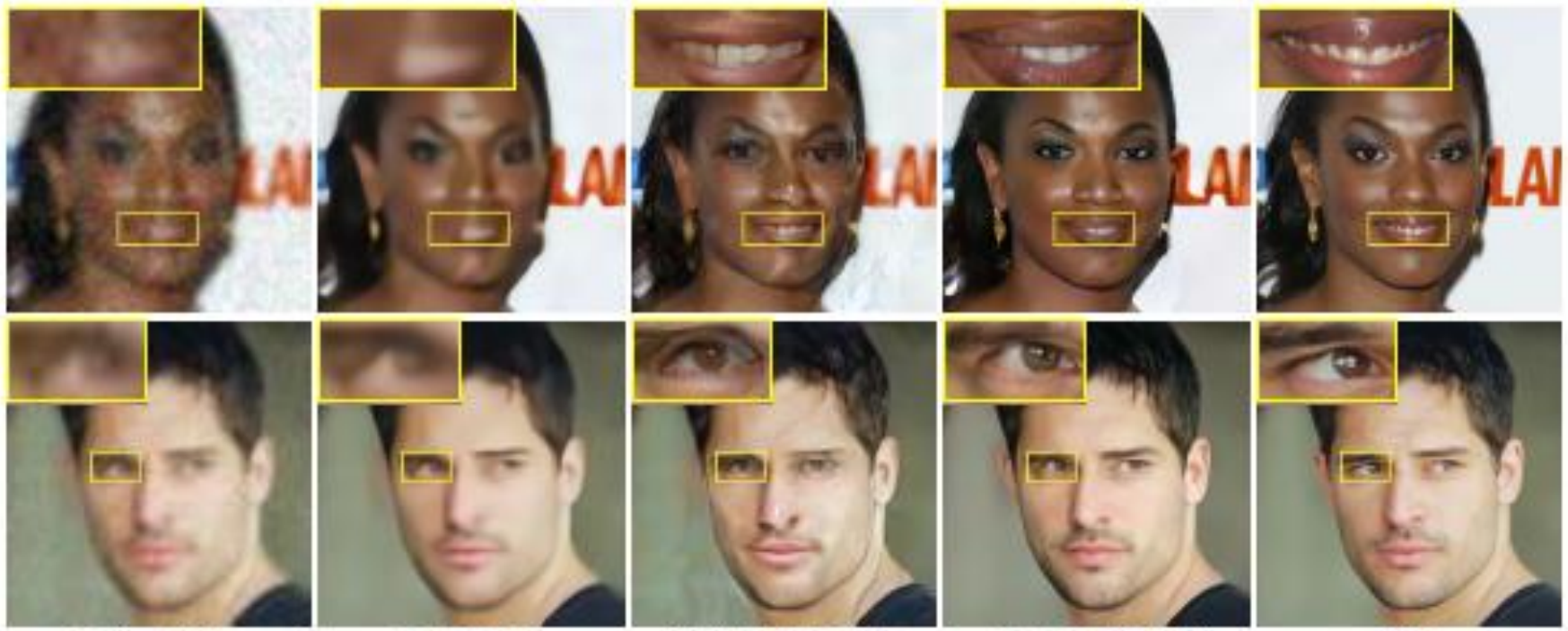


Figure 4: Comparison on the **CelebA-Test** for $\times 4$ face super-resolution. Our GFP-GAN restores realistic teeth and faithful eye gaze direction. **Zoom in for best view.**

Results

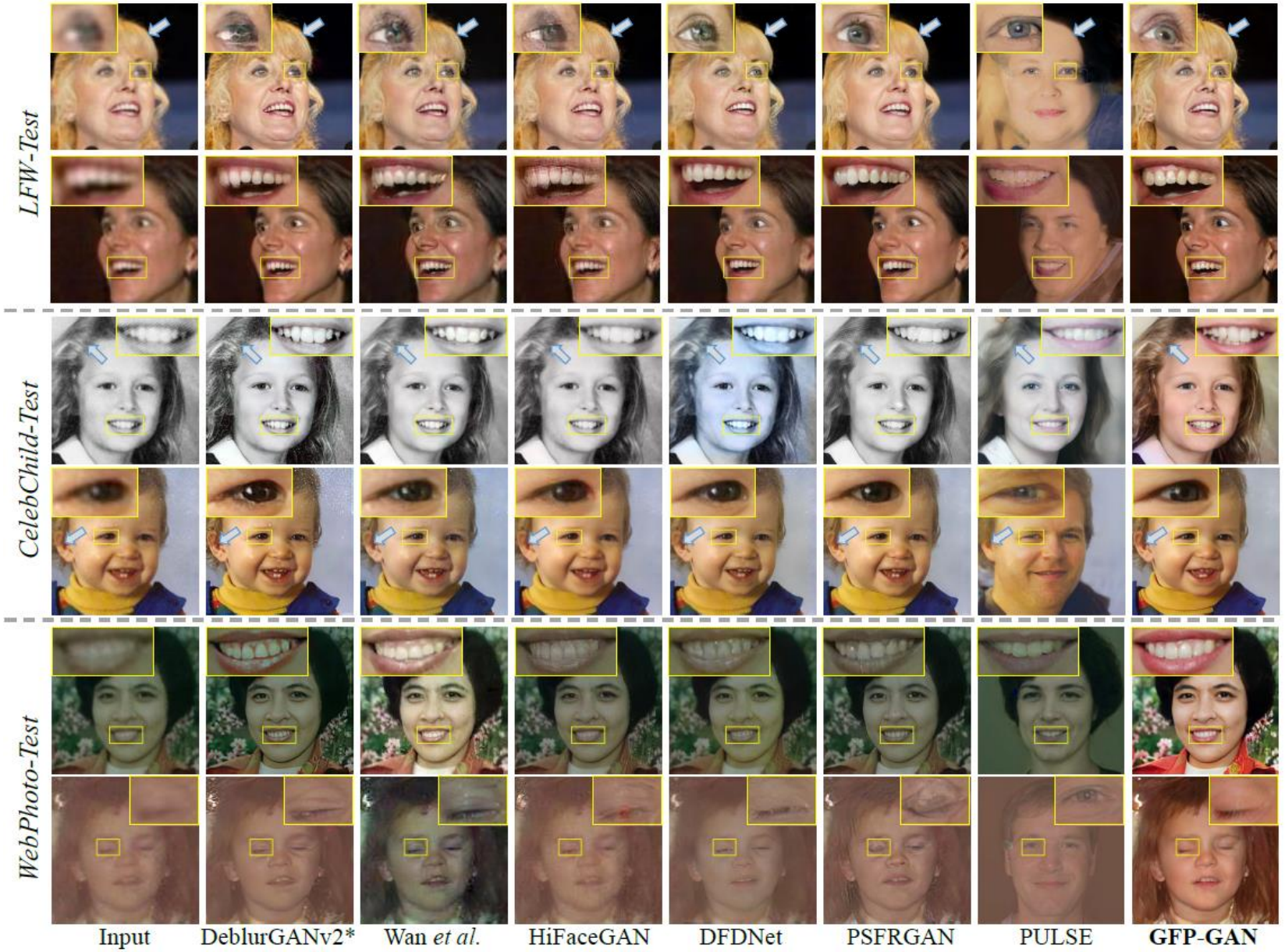


Figure 5: Qualitative comparisons on three **real-world** datasets. **Zoom in for best view.**

Results

Table 1: Quantitative comparison on **CelebA-Test** for blind face restoration. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Methods	LPIPS↓	FID↓	NIQE ↓	Deg.↓	PSNR↑	SSIM↑
Input	0.4866	143.98	13.440	47.94	25.35	0.6848
DeblurGANv2* [40]	0.4001	52.69	4.917	39.64	**25.91**	**0.6952**
Wan *et al.* [61]	0.4826	67.58	5.356	43.00	24.71	0.6320
HiFaceGAN [67]	0.4770	66.09	4.916	42.18	24.92	0.6195
DFDNet [44]	0.4341	59.08	4.341	40.31	23.68	0.6622
PSFRGAN [6]	0.4240	47.59	5.123	39.69	24.71	0.6557
mGANprior [19]	0.4584	82.27	6.422	55.45	24.30	0.6758
PULSE [52]	0.4851	67.56	5.305	69.55	21.61	0.6200
GFP-GAN (ours)	**0.3646**	**42.62**	**4.077**	**34.60**	25.08	0.6777
GT	0	43.43	4.292	0	∞	1

Table 2: Quantitative comparison on **CelebA-Test** for 4× face super-resolution. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Methods	LPIPS↓	FID↓	NIQE ↓	Deg.↓	PSNR↑	SSIM↑
Bicubic	0.4834	148.87	10.767	49.60	25.377	0.6985
RCAN* [74]	0.4159	93.66	9.907	38.45	**27.24**	**0.7533**
ESRGAN* [64]	0.4127	49.20	4.099	51.21	23.74	0.6319
Super-FAN [4]	0.4791	139.49	10.828	49.14	25.28	0.7033
GFP-GAN (ours)	**0.3653**	**42.36**	**4.078**	**34.67**	25.04	0.6744
GT	0	43.43	4.292	0	∞	1

Table 3: Quantitative comparison on the *real-world* **LFW**, **CelebChild**, **WebPhoto**. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Dataset	**LFW-Test**		**CelebChild**		**WebPhoto**	
Methods	FID↓	NIQE ↓	FID↓	NIQE ↓	FID↓	NIQE ↓
Input	137.56	11.214	144.42	9.170	170.11	12.755
DeblurGANv2* [40]	57.28	4.309	110.51	4.453	100.58	4.666
Wan *et al.* [61]	73.19	5.034	115.70	4.849	100.40	5.705
HiFaceGAN [67]	64.50	4.510	113.00	4.855	116.12	4.885
DFDNet [44]	62.57	4.026	111.55	4.414	100.68	5.293
PSFRGAN [6]	51.89	5.096	107.40	4.804	88.45	5.582
mGANprior [19]	73.00	6.051	126.54	6.841	120.75	7.226
PULSE [52]	64.86	5.097	**102.74**	5.225	**86.45**	5.146
GFP-GAN (ours)	**49.96**	**3.882**	111.78	**4.349**	87.35	**4.144**

Results – ablation study

Table 4: Ablation study results on **CelebA-Test** under blind face restoration.

Configuration	LPIPS↓	FID↓	NIQE ↓	Deg.↓
Our GFP-GAN with SC-SFT	**0.3646**	**42.62**	**4.077**	**34.60**
a) No spatial modulation	0.550 (↑)	60.44 (↑)	4.183 (↑)	74.76 (↑)
b) Use SFT	0.387 (↑)	47.65 (↑)	4.146(↑)	34.38 (↓)
c) w/o GFP	0.379 (↑)	48.47 (↑)	4.153 (↑)	35.04 (↑)
d) – Pyramid Restoration Loss	0.369 (↑)	45.17 (↑)	4.284 (↑)	35.50 (↑)

Results – ablation study

GFP-GAN a) No SC-SFT b) SFT c) No GFP d) No Pyramid

Figure 6: Ablation studies on CS-SFT layers, GFP prior and pyramid restoration loss. **Zoom in for best view.**

Input No *D* + *D* + *D* + *fm* + *D* + *fs*

Figure 7: Ablation studies on facial component loss. In the figure, *D*, *fm*, *fs* denotes component discriminator, feature

Discussion and Limitations

Figure 8: Results on dark-skinned faces.

Figure 9: Limitations of our model. The results of PSFR-GAN [6] are also presented.

- **Training bias**.

when input images are gray-scale, the face color may have a bias (last example in Fig. 8), as the inputs do not contain sufficient color information. Thus, a diverse and balanced dataset is in need.

- **Limitations**.

As shown in Fig. 9, when the degradation of real images is severe, the restored facial details by GFPGAN are twisted with artifacts. Our method also produces unnatural results for very large poses. This is because the synthetic degradation and training data distribution are different from those in real-world. One possible way is to learn those distributions from real data instead of merely using synthetic data, which is left as future work.

Conclusion

- We have proposed the GFP-GAN framework that leverages the rich and diverse **generative facial prior** for the challenging blind face restoration task.

- This prior is incorporated into the restoration process with **channel-split spatial feature transform layers**, allowing us to achieve a good balance of realness and fidelity.

- Extensive comparisons demonstrate the superior capability of GFP-GAN in joint face restoration and color enhancement for **real-world images**, outperforming prior art.

Reference

- **Reference**
- PR-293: In-Domain GAN Inversion for Real Image Editing

Thank you for your Attention….!