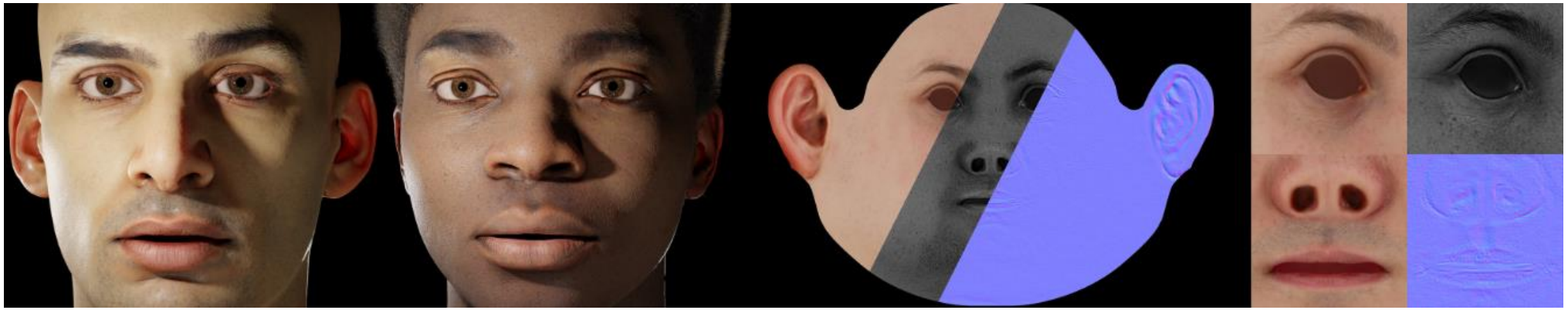


GANtlitz: Ultra High Resolution Generative Model for Multi-ModalFace Textures

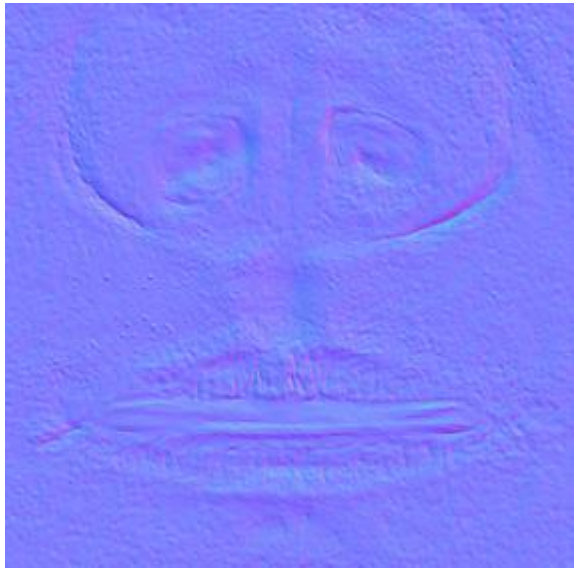
A. Gruber, E. Collins, A. Meka, F. Mueller, K. Sarkar, S. Orts-Escolano, L. Prasso, J. Busch, M. Gross T. Beeler

Abstract

- To render photoreal digital humans - High-resolution texture maps is essential
- GANtlitz – makes multi-modal ultra-high-resolution face appearance maps
- Solves three challenges:
  - Large data for training generative models
  - Limitations of training a GAN at ultra-high resolutions
  - Improves consistency of appearance features across different modalities.

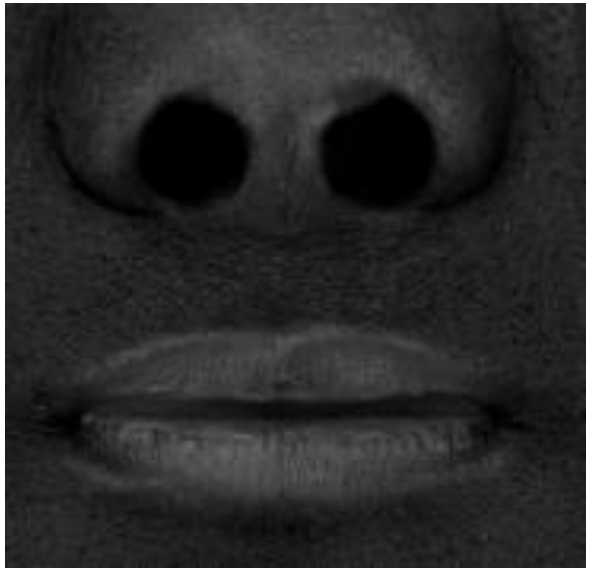
Introduction

- Capturing high-quality multi-model texture maps from real humans is difficult
  - Involves multi-view camera setups, designed illumination, geometry and appearance reconstruction
- Large scale dataset is therefore very expensive and difficult.
- GANtlitz can synthesize texture maps at high resolution (6144x4096) from few samples (<100)



Normals

Albedo



Specularity

- Advances in GANs, allow for generating visually rich imagery in specific domains.
- But it comes with two unique challenges.
  - Lack of high-resolution texture maps -> fewer images lead to either divergent training or mode collapse
  - Representing intricate appearance details requires extremely high resolution.

Generative Models of Human Faces

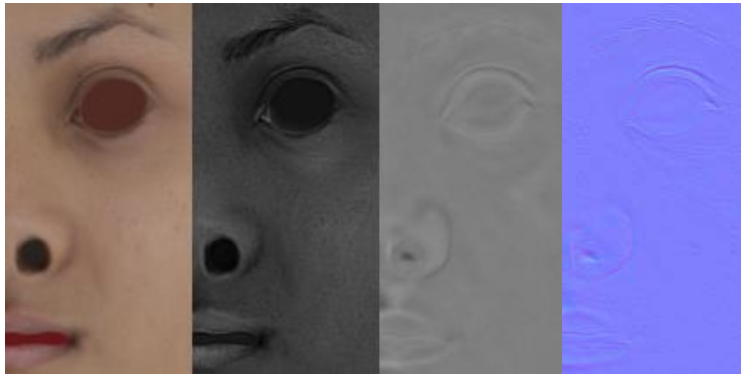
- Morphable models of human faces can generate both the geometry and appearance.
  - It has limitations however in both resolution and high-frequency details
- Several Autoencoder based systems can struggle with produce realistic high frequency detail
  - Some learning a separate super-resolution step, another can demonstrate high quality, but only to a single subject
- StyleGAN models- Excelled at learning a distribution of real human faces
- Combined StyleGAN and traditional morphable mode
  - Many different models - but everyone is limited to the resolution of StyleGAN

Training GANs with Limited Data

- The training data of GANtlitz consists of only 97 high resolution face textures
- Some recent methods use regularization to improve the generative capacity
- Loss regularization has been suggested for specifically targeting the data scarcity issue

Dataset

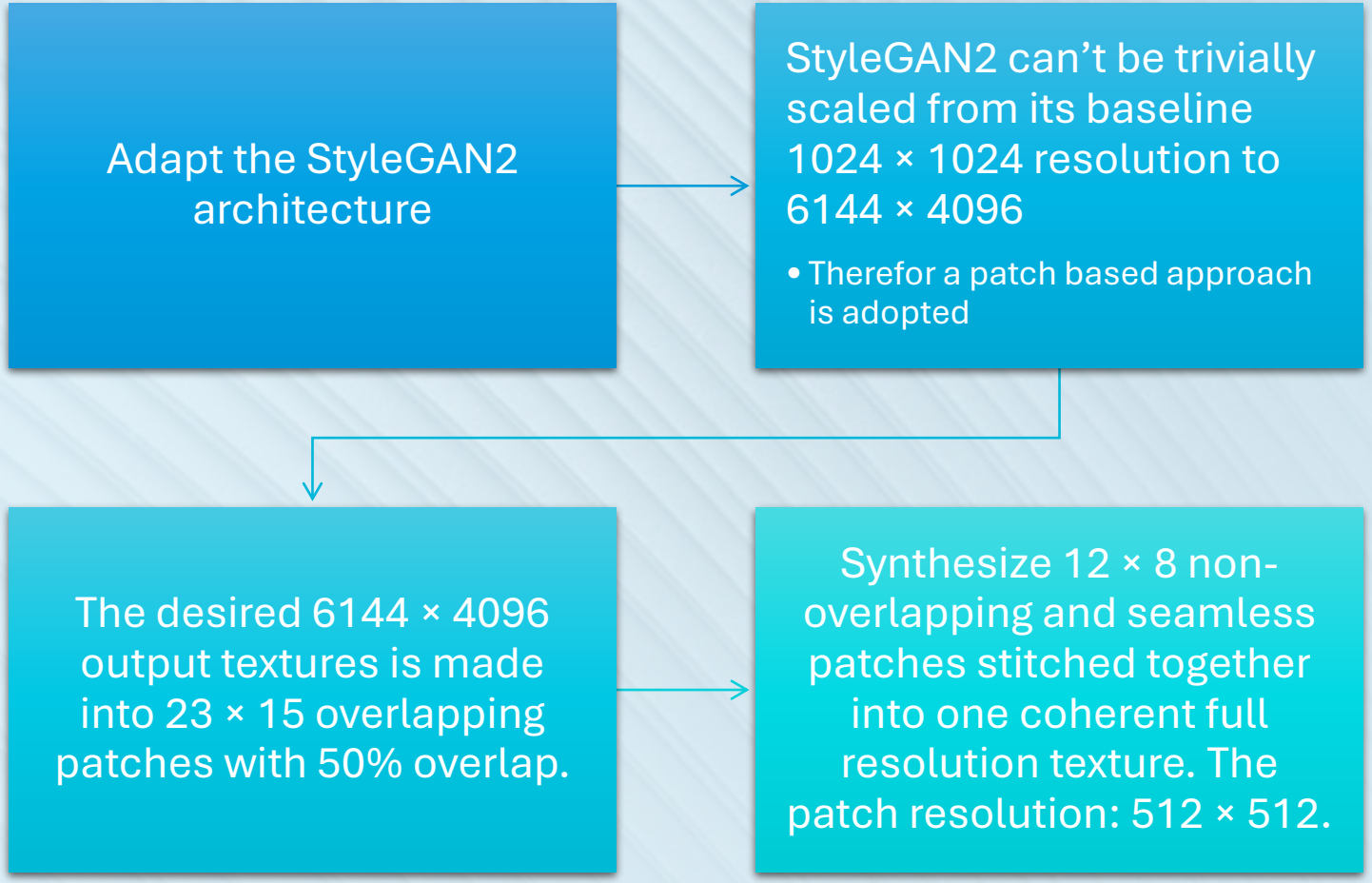
- High-resolution face textures that capture intricate appearance details are challenging to acquire for technological, logistical and privacy-related reasons
- In the paper they use commercially available texture maps from www.3dscanstore.com,
- For each subject, multiple modalities are provided
- To ensure optimal training, following pre-processing has been applied:
  - Cropping and masking
  - Generating Displacement Maps.
  - Exploiting Symmetry.



Method

- The method builds uppon the adversarial architecture of StyleGAN2
- StyleGAN – while it has ecellent image quality, does not reach our target high resolution
- The primary challenges:
  - hardware memory constraints
  - limited training set size

Generator


Adapt the StyleGAN2 architecture
StyleGAN2 can’t be trivially scaled from its baseline 1024 × 1024 resolution to 6144 × 4096
• Therefor a patch based approach is adopted
The desired 6144 × 4096 output textures is made into 23 × 15 overlapping patches with 50% overlap.
Synthesize 12 × 8 non-overlapping and seamless patches stitched together into one coherent full resolution texture. The patch resolution: 512 × 512.

Discriminator Augmentation Scheme

- With only 97 samples, they operate on severe data scarcity
  - This is addressed by employing discriminator augmentations, which increase the variety of data the model sees, which stabilizes training when the dataset is small.
- Augmentations are differentiable transformations applied to both the real and generated data before being fed to the discriminator.
  - Each transformation is employed probabilistically at a rate p.
  - A probability at less than 85% prevents the "leakage“ into the generator.
  - If leakage occurs, the generator starts to mimic the Augmentations rather than producing realistic outputs
  - Color Transformations Are Avoided as this led to undesirable effects or "leakage"

- Not all augmentations are applied with varying probability
  - mid-band filter is used at **constant rate** of p=0.85
- Mid-Band Filter forces the discriminator to evaluate the generated images across the entire frequency spectrum
  - Leads to a more frequency-diverse generator, which is noticeable in high-frequency features, such as wrinkles and stubbles

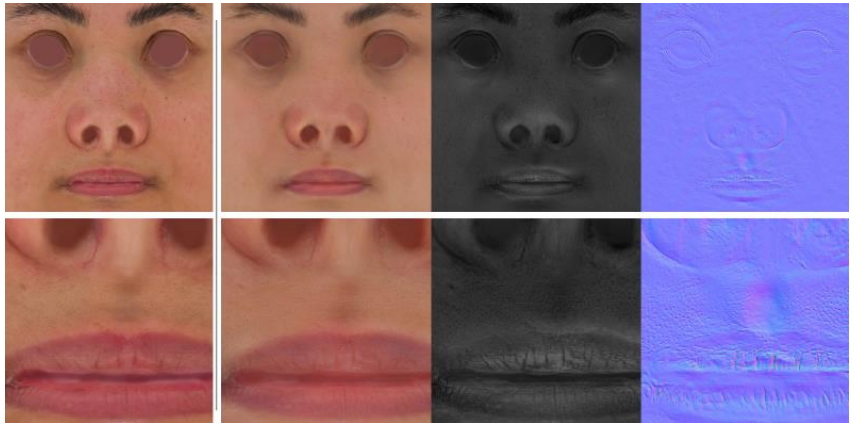
Memory Constraints and Modality Dropout

- Training on A100 GPUs = memory budget is sufficient to train a large version GANtlitz High Capacity Model (GANtlitz-L).
- With V100 GPUs GANtlitz Standard Capacity Model (GANtlitz-S)
- Total memory requirements for a single training iteration depend on:
  - model size and number of modalities to be generated

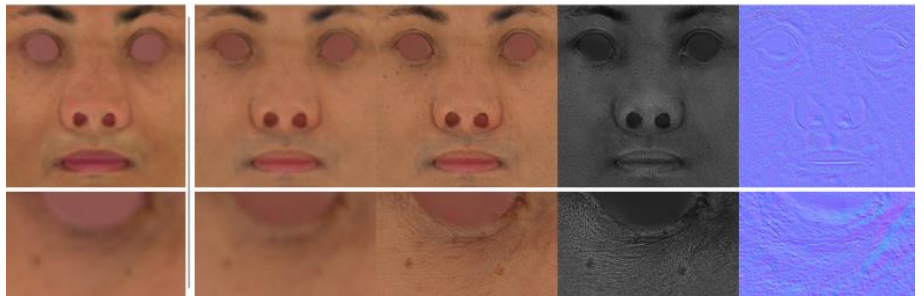
Qualitative Analysis

- Facial details are rendered with high resolution
- Due to the limited training dataset, low-frequency variation is limited.
- Low frequency structure are largely unchanged, while mid- and high-frequency details show good variation

The inversion process



- The inversion process involves projecting textures into the latent space of the generator through a structured two-stage optimization.
  - Optimizes the latent codes first while updating noise maps, followed by fixing the latent codes and fine-tuning the noise maps.
- The inversion process enables two specific applications:
  - Modality Completion: the capability to fill in or enhance missing modalities (like texture, color, or other features)
  - Super Resolution: increasing the resolution of the generated textures.

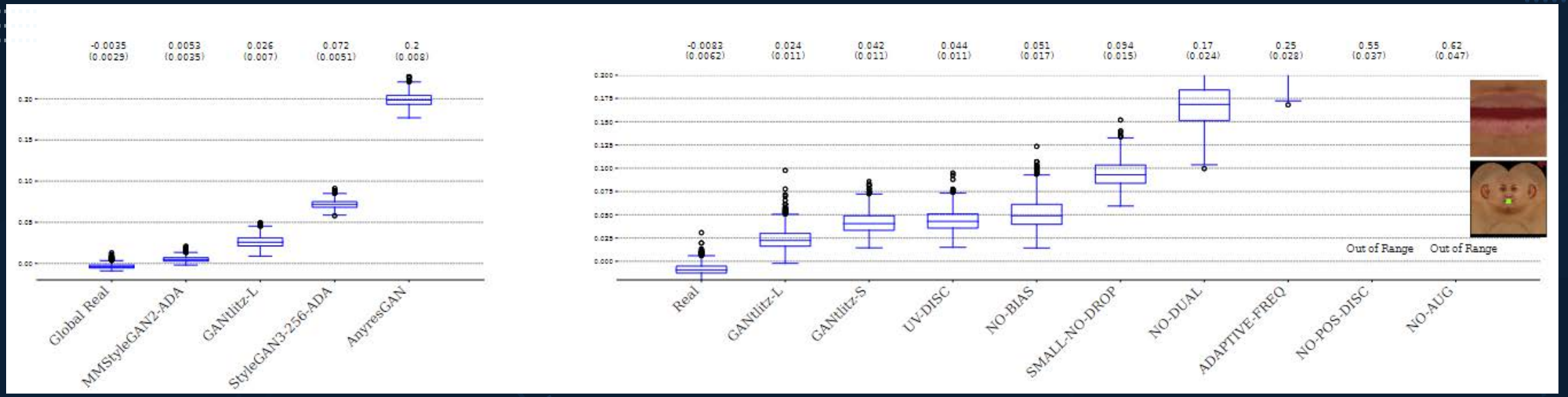


Baseline Comparisons

- StyleGAN2 – for a meaningful comparison, its adapted to produce all three modalities and the generator output is multiplied with the UV layout mask
- AnyresGAN – due to multiple challenges during training. multi-modal experiments with AnyresGAN was abandoned

Quantitative Evaluation

- Goal is to analyze how well the method captures the data distribution
- Kernel Inception Distance (KID) is chosen as the quantitative metric for this analysis, due to:
  - Better Convergence Characteristics
  - Unbiased Measurement
- Difficulties due to the limited training data (only 97 subjects)



- Despite being trained with only patch-level supervision, the proposed GANtlitz-L model demonstrates good performance at the global scale,
- At the patch level, the authors note that their model outperforms other variants

- **AnyresGAN:** The teacher model shows weak performance, and the subsequent training phase does not significantly improve results.
- **MMStyleGAN2-ADA** (1K) shows a good match with the real distribution, benefiting from training on entire images rather than patches. However, it suffers from overfitting, meaning it performs well on training data but poorly on unseen data.

Ablation Study

- No Augmentations (NO-AUG) - data augmentations are critical for model convergence under data scarcity. Without them the model fails to converge.
- Adaptive frequency – Use frequency-band augmentation at const. 85%. With rate based on discriminator performance, we see a reduction in learned high frequency detail.
- SMALL-NO-DROP: modality dropout is a constant augmentation that also helps reduce memory usage
- NO-POS-DISC: Convergence of our patch-wise training scheme isn’t possible without providing positional encodings to the discriminator.
- UV-DISC: Instead of sinusoidal positional encodings, we feed two-dimensional UV coordinates as positional information to the discriminator. In this case it is observed a degradation in performance

Sources

- “GANtlitz: Ultra High Resolution Generative Model for Multi-ModalFace Textures” https://onlinelibrary.wiley.com/doi/epdf/10.1111/cgf.15039 (last visited 24/09/2024)