Stony Brook University

Computer Vision in SBU: Generative Models and Human Behavior Modeling

Dimitris Samaras
samaras@cs.stonybrook.edu

FAR BEYOND

Computer Vision at Stony Brook

The Computer Vision Lab

Faculty:

Dimitris Samaras

Haibin Ling

Michael Ryoo (Robotics Lab also)

Minh Hoai

40+ PhD students

BioMedical Informatics

Zhaozheng Yin

Chao Chen

Prateek Prasanna

Other Faculty

David Gu

Hong Qin

Arie Kaufman

Strong collaborators in CS and other departments on campus Psychology, Music, Art, BNL, BMI, Ecology, Civil Engineering

Computer Vision at Stony Brook

The Computer Vision Lab

Faculty:

Dimitris Samaras

Haibin Ling

Michael Ryoo (Robotics Lab also)

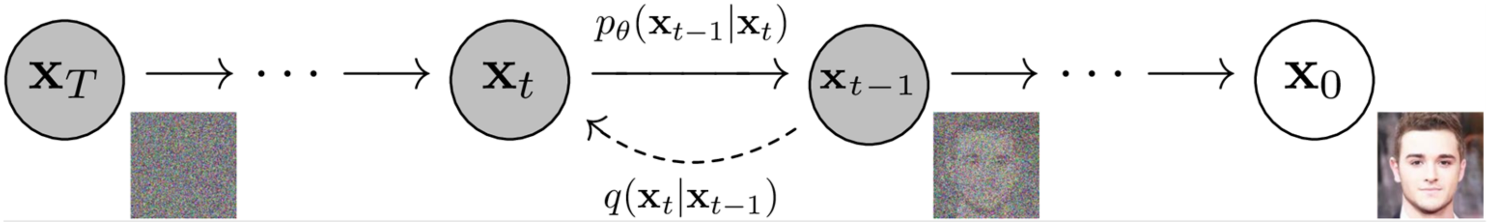
Minh Hoai

40+ PhD students

Diffusion generative models

Setting: Gaussian diffusion models

- Gaussian diffusion models are generative models that learn to reverse a corruption process that adds Gaussian noise
- The forward process (←) is a Markov chain that gradually adds noise to the data
- The reverse process (→) is a Markov chain that gradually denoises the data
  - Denoising diffusion models learn a neural network approximation p_θ to the reverse process, defining the marginal distribution p(**x**)



[Figure from Ho, Jain, and Abbeel, NeurIPS 2020]

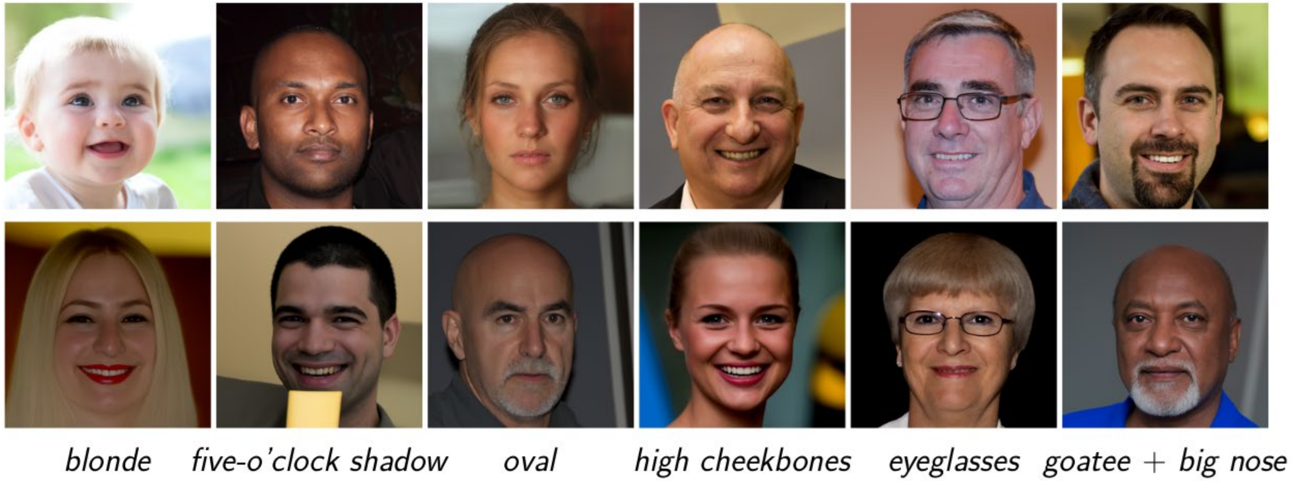
Controllable Generation

- A trained (Gaussian) diffusion model can generate diverse and high-quality unconditional samples from the learned distribution p(**x**)


[Images adapted from Ho, Jain, and Abbeel, NeurIPS 2020]

Controllable Generation - Posterior Inference

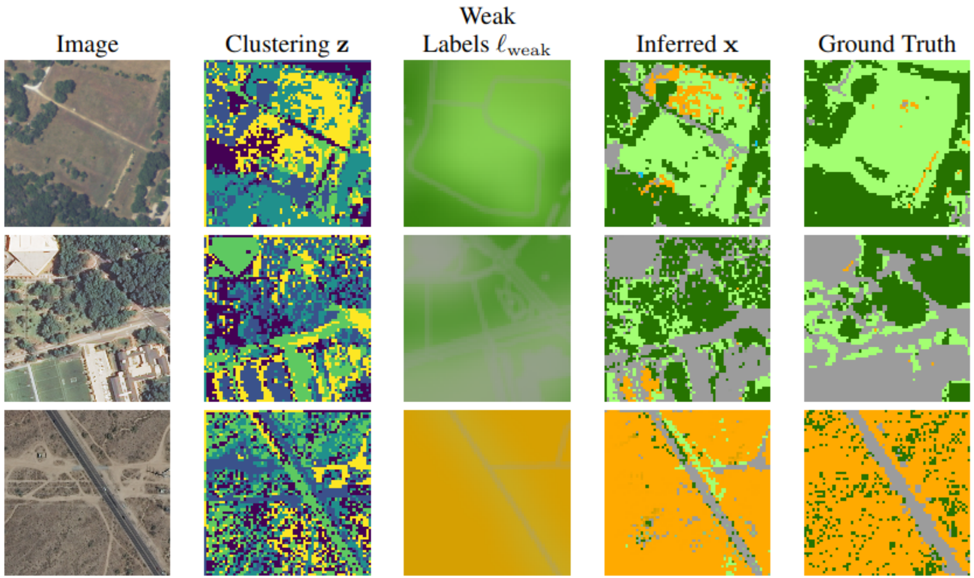
- A trained (Gaussian) diffusion model can generate diverse and high-quality unconditional samples from the learned distribution p(**x**)
- We want to use this trained model with additional constraints c to generate samples that satisfy both p(**x**) and c(**x**, y)
  - c(**x**, y) could be a separately trained attribute classifier, e.g. ***facial attributes***



Diffusion models as plug-and-play priors, Graikos, Malkin, Samaras, Jojic, NeurIPS 2022

Controllable Generation - Segmentation

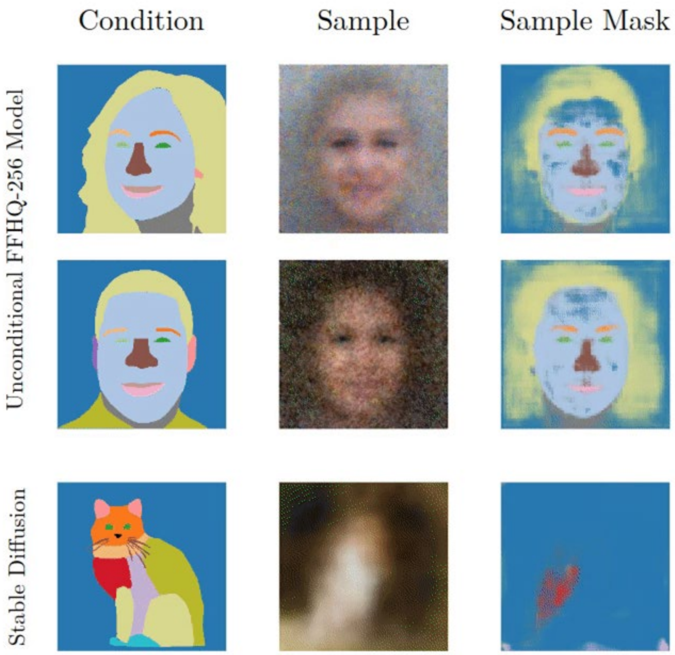
- We also show how a diffusion prior can be used for inferring color-invariant segmentations
  - Using a color clustering of the image we infer the segmentation that matches both a pre-trained diffusion prior and the clustering



Diffusion models as plug-and-play priors, Graikos, Malkin, Samaras, Jojic, NeurIPS 2022

Controllable Generation - Few-shot

- We introduce a method to draw conditional samples from a small set (~10) of condition-image pairs



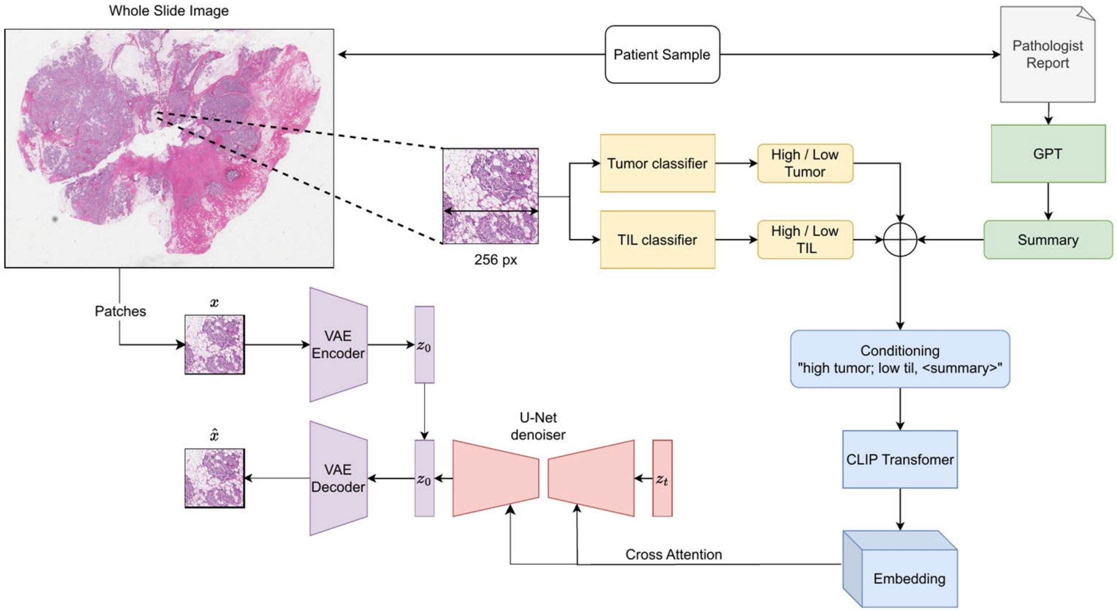
Conditional Generation from Unconditional Diffusion Models using Denoiser Representations, Graikos, Yellapragada, Samaras, BMVC 2023

Diffusion models for Histopathology

- There is a need for generative models in *specialized* domains such as computational pathology
- Recent large-scale generative models depend on training on **vast amounts of data** and providing **per-image conditions** for controllable generation

Diffusion models for Histopathology - Text Conditioning

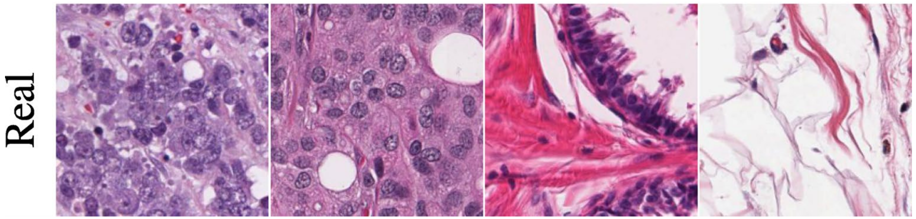
- We utilize recent LLM capabilities to summarize the **unstructured pathology reports** into concise text prompts
- Using these text prompts we train a diffusion model to generate patches of whole-slide histopathology images



PathLDM: Text conditioned Latent Diffusion Model for Histopathology, Yellapragada, Graikos, et al., WACV 2024

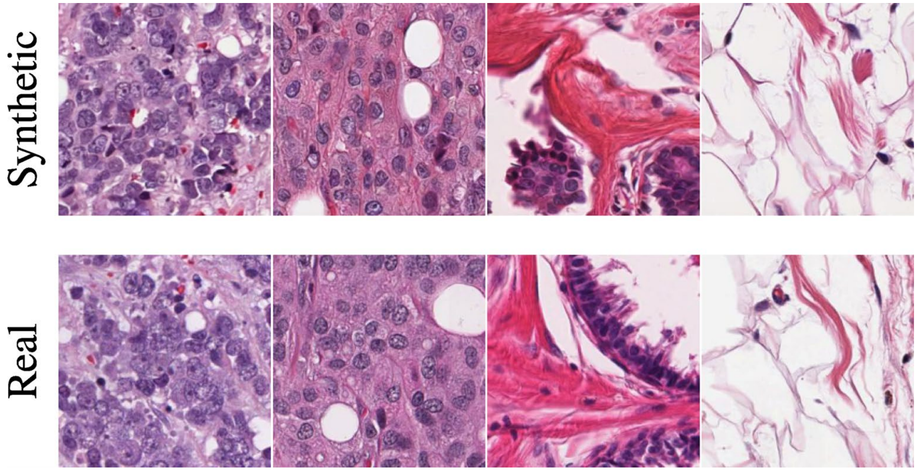
Diffusion models for Histopathology - SSL Conditioning

- Whole-slide text reports fail to describe local details
- Hand-annotating images per-patch is infeasible
  - A dataset of 1000 slides (15M patches) would require **>40.000** expert hours



Diffusion models for Histopathology - SSL Conditioning

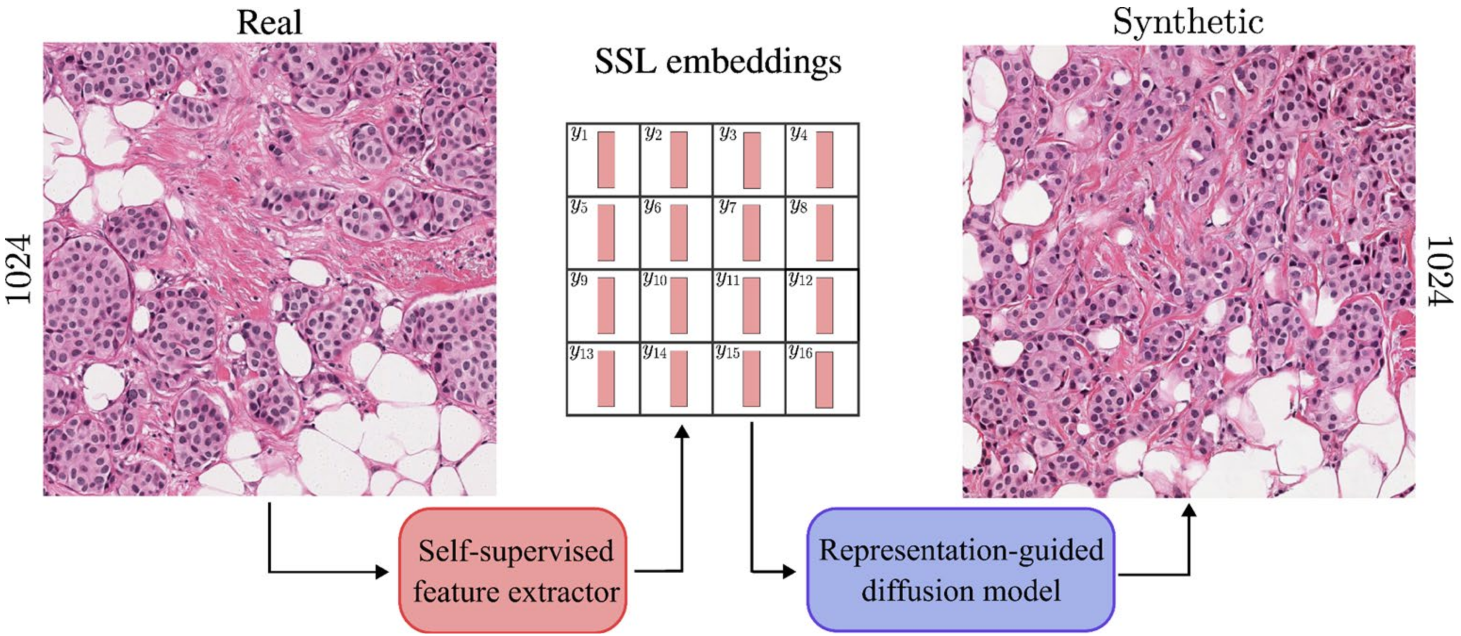
- We propose using **representations** learned with self-supervision **in place of human annotations**
  - We find that SSL representations can accurately describe images allowing us to train large-scale diffusion generative models



Learned representation-guided diffusion models for large-image generation, Graikos et al., CVPR 2024

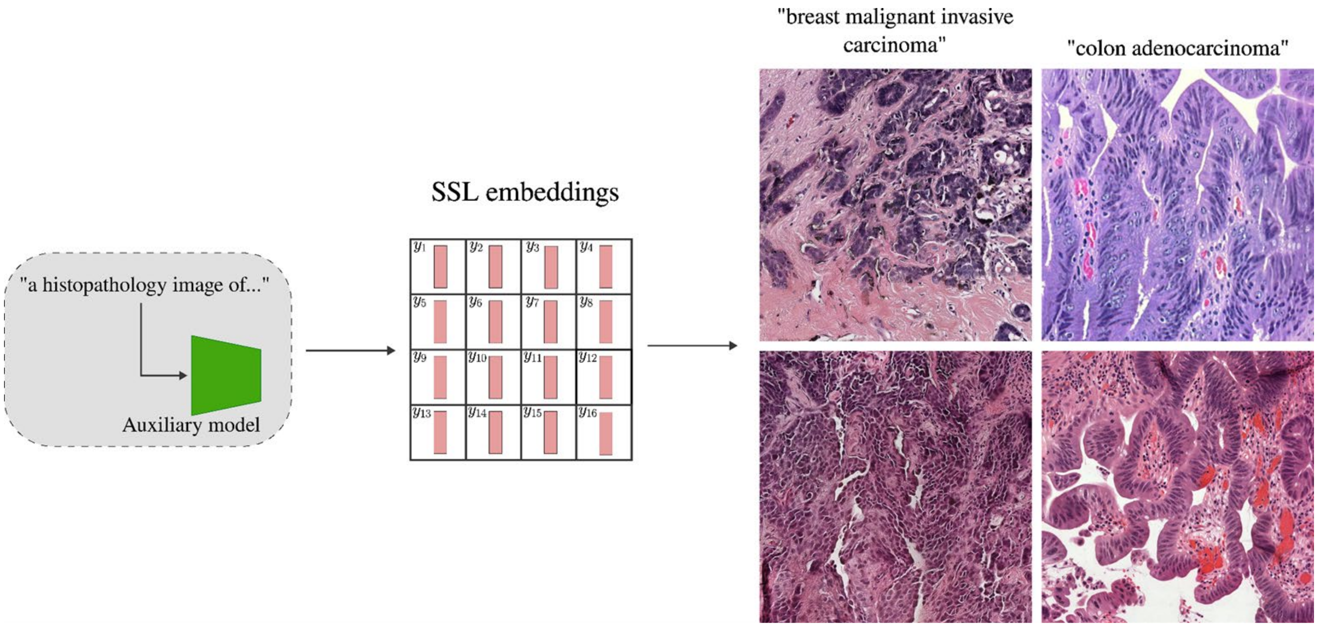
Diffusion Models for Histopathology - Large Images

- Impractical to train directly on the entire digitized slides (32.000 x 32.000 px)
  - We introduce an algorithm to **synthesize large histopathology images** by spatially controlling the local, patch-based model



Diffusion Models for Histopathology - Large Images

- Previous framework constrained to using representations from reference images
  - We train small, auxiliary models that learn to **map any condition** to the self-supervised **representations** and generate new images



JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

AVFace: Towards Detailed Audio-Visual 4D Face Reconstruction

Aggelina Chatziagapi Dimitris Samaras

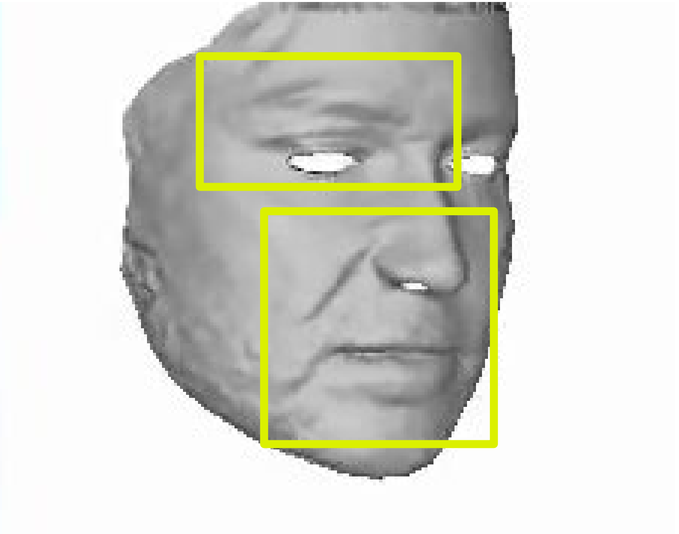
Stony Brook University

Detailed Audio-Visual 4D Face Reconstruction

Input

Reconstruction

Detailed Audio-Visual 4D Face Reconstruction

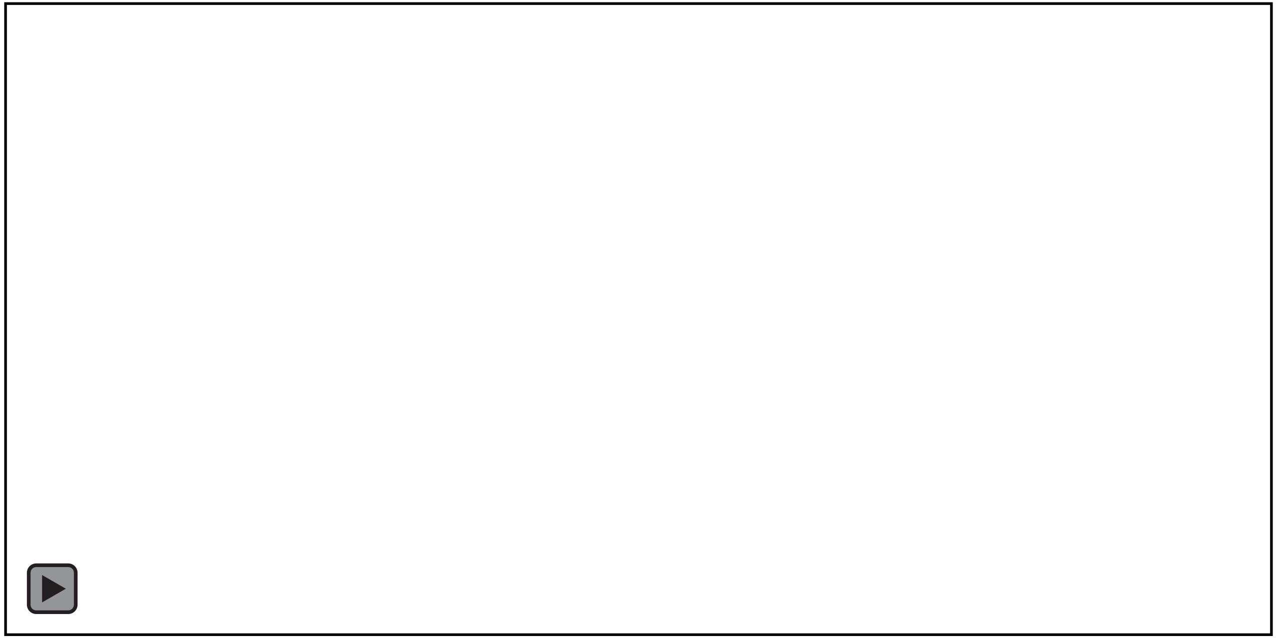


Lip shape & facial details

Input

Reconstructio
n

Detailed Audio-Visual 4D Face Reconstruction



Input

Reconstructio
n

Detailed Audio-Visual 4D Face Reconstruction

construction
clusion

Input

Reconstructio
n

Project page:

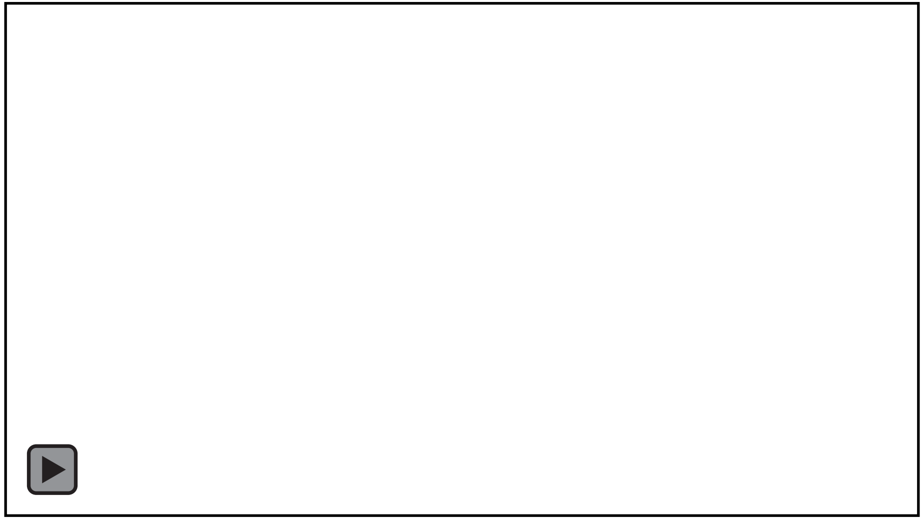
IEEE Biometrics Council
THE 17 TH IEEE CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION
FG 2023 Jan 5-8, Waikoloa, Hawaii

LipNeRF: What is the right feature space to lip-sync a NeRF?

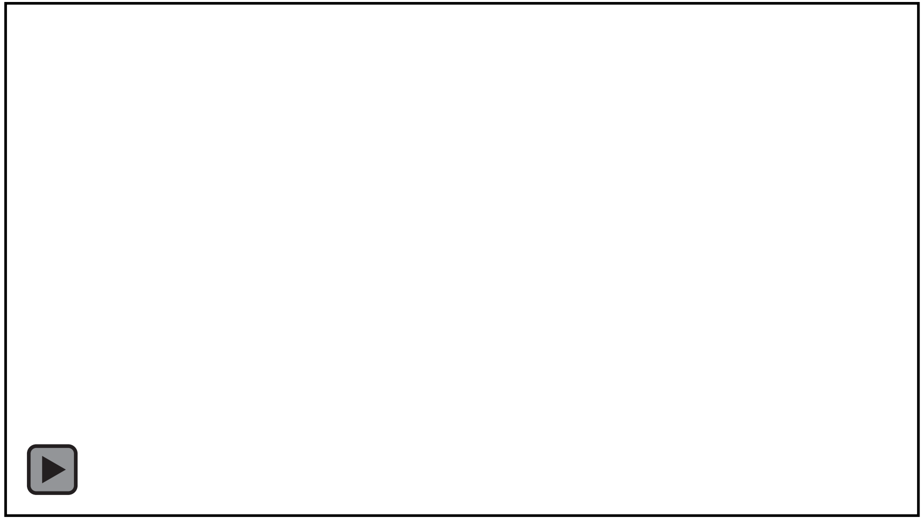
Aggelina Chatziagapi ShahRukh Athar Abhinav Jain
Rohith MV Vimal Bhat Dimitris Samaras

Stony Brook University
digital media engineering prime video

Lip Synchronization with Speech



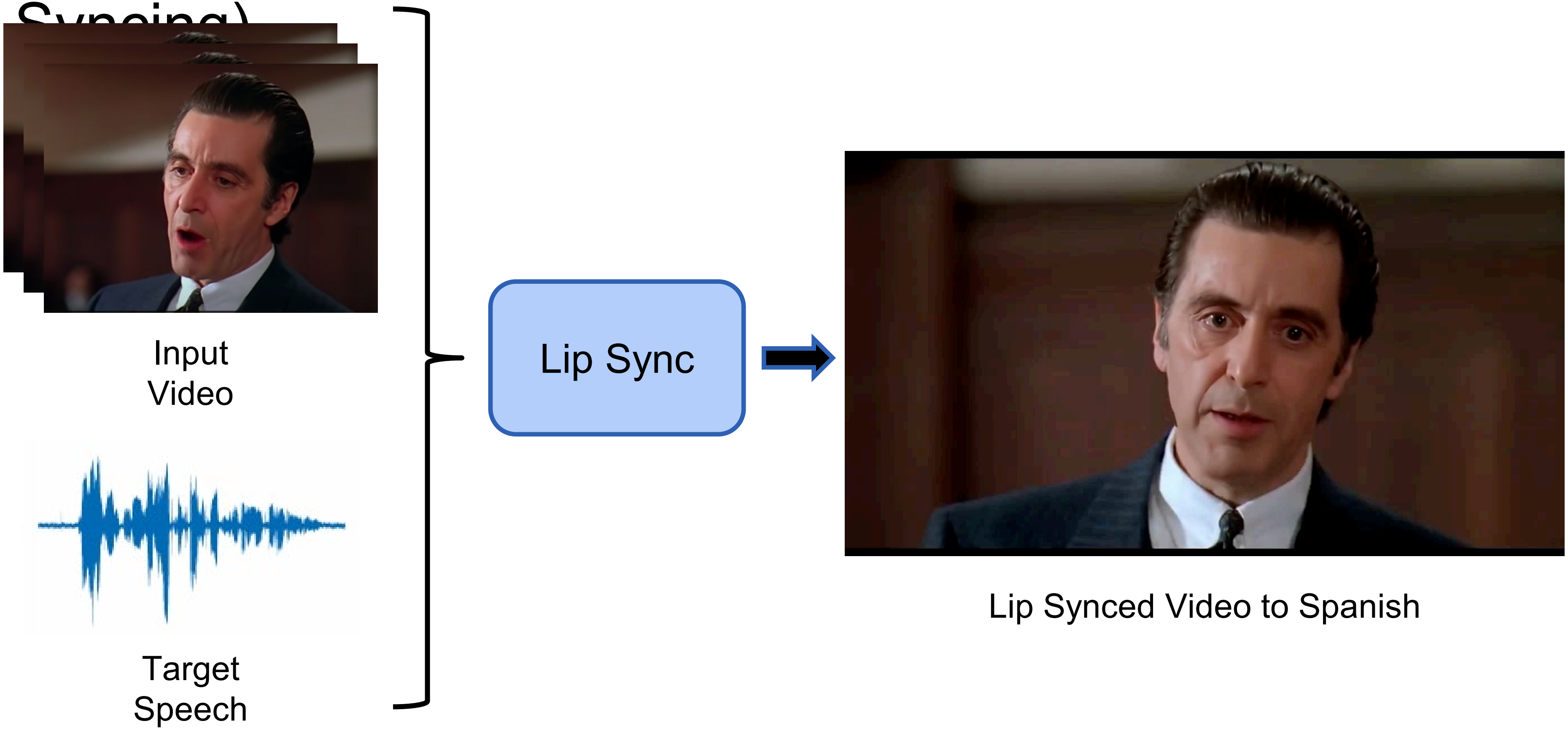
Original Audio & Video



Dubbed Audio & Original Video

Lips are out of sync

Audio-driven Talking Head Video Synthesis (or Lip Syncing)



MI-NeRF: Learning a Single Face NeRF from Multiple Identities

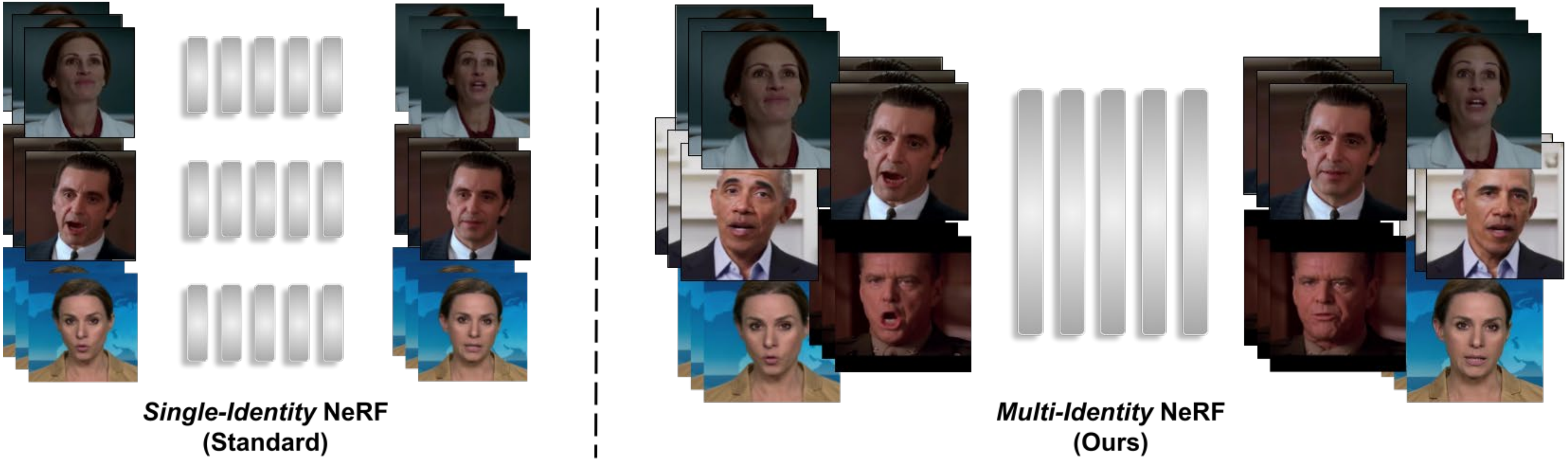
Aggelina Chatziagapi Grigorios G. Chrysos Dimitris Samaras

arXiv 2024

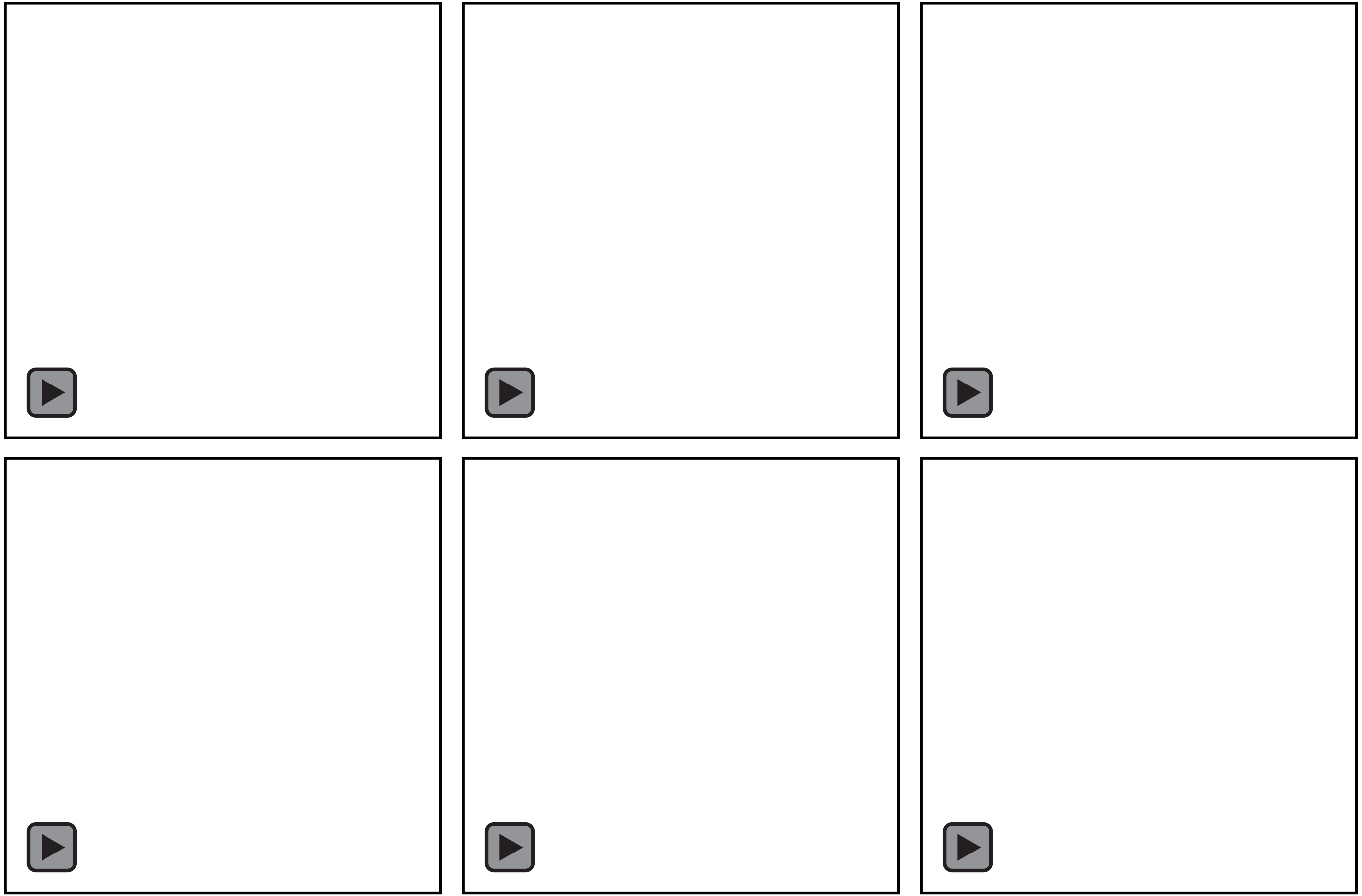
Stony Brook University

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

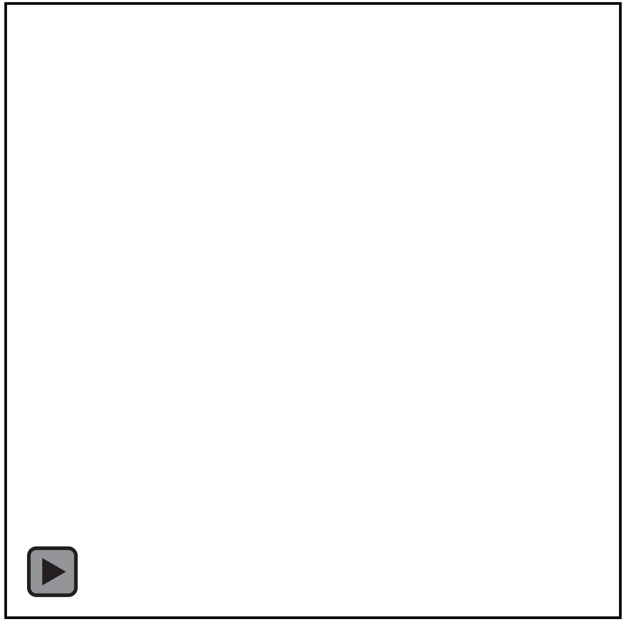
Learning a ***single*** NeRF for ***multiple*** identities



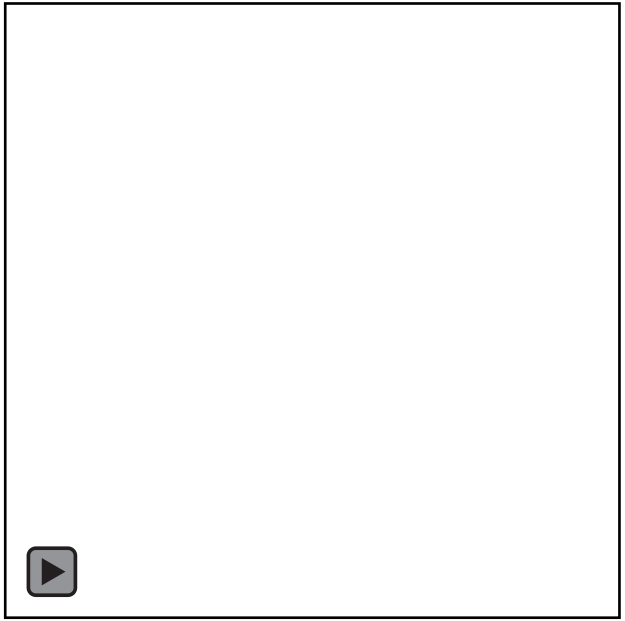
A ***single*** face NeRF
can generate ***multiple*** identities



*Standard **single-identity** NeRFs cannot generalize to challenging novel expressions*

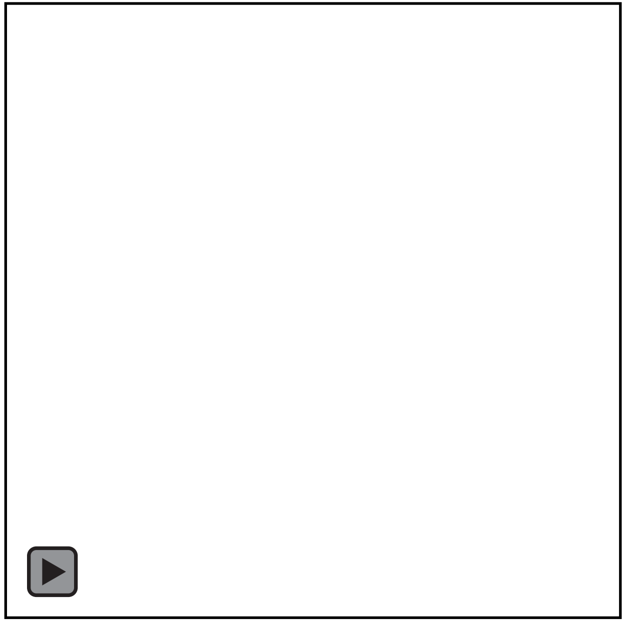


Target Expression

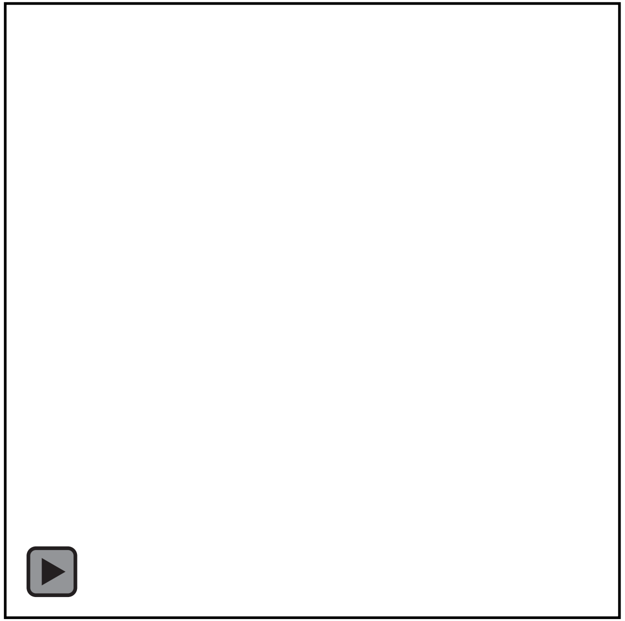


NeRFace
Single-Identity NeRF
(Standard)

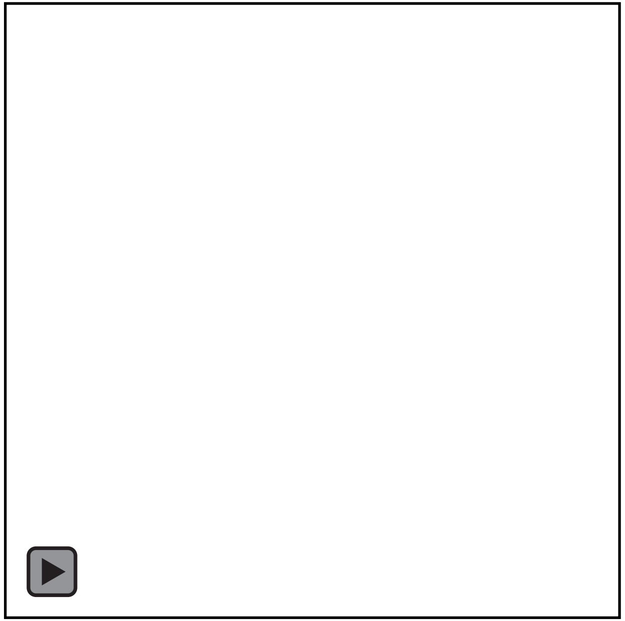
Learning from multiple identities,
*our **multi-identity NeRF (MI-NeRF)** can*
*synthesize **novel** expressions for any input identity*



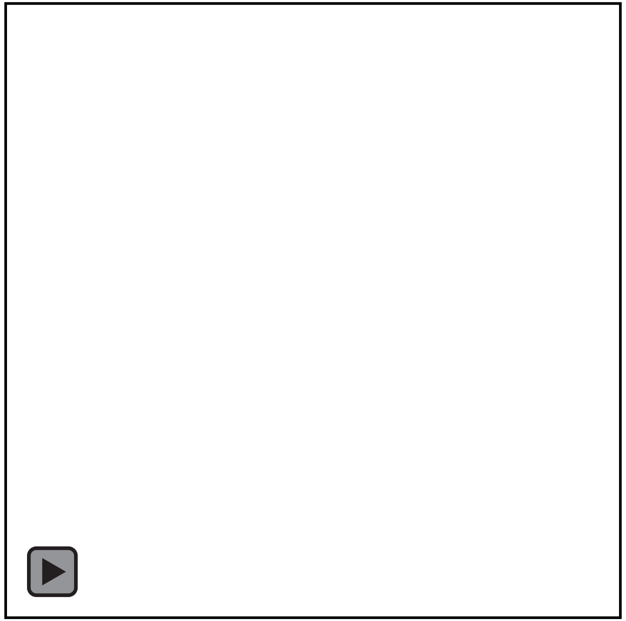
Target Expression



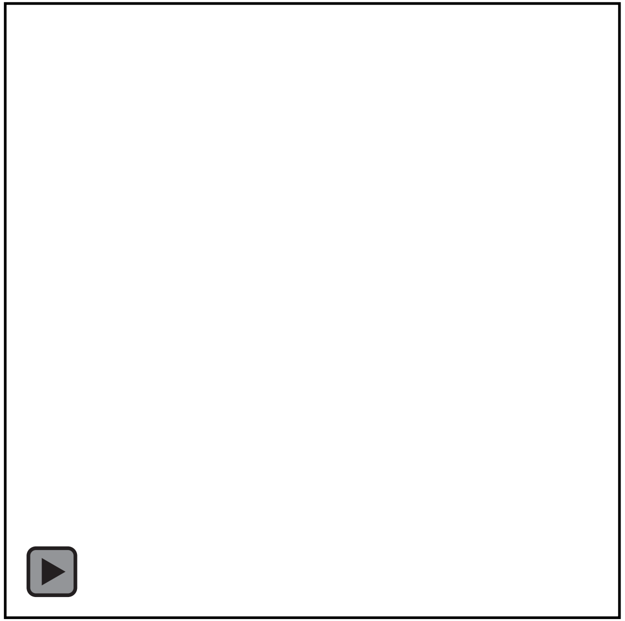
NeRFace
Single-Identity NeRF
(Standard)



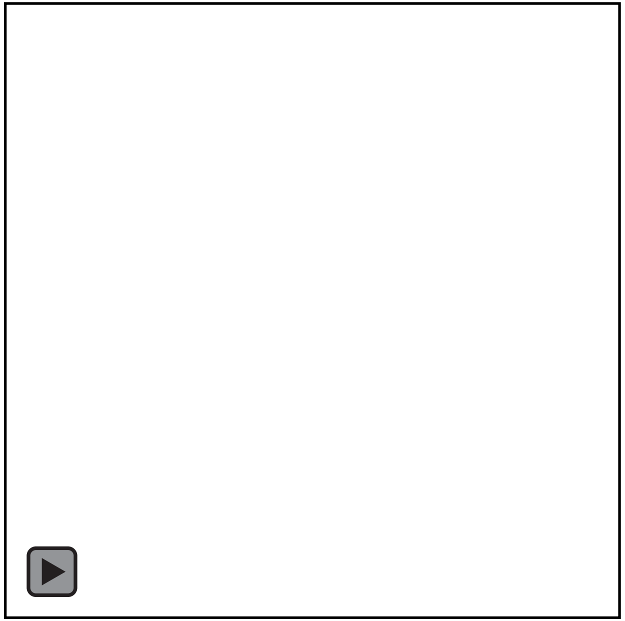
MI-NeRF
Multi-Identity NeRF
(Ours)



Target Expression



NeRFace
Single-Identity NeRF
(Standard)



MI-NeRF
Multi-Identity NeRF
(Ours)

Human Gaze Modeling

Zhibo Yang, Sounak Mondal

Collaborators: Seoyoung Ahn, Yupei Chen, Lihan Huang, Zijun Wei, Ruoyu Xue, Souradeep Chakraborty,
Gregory Zelinsky, Dimitris Samaras and Minh Hoai

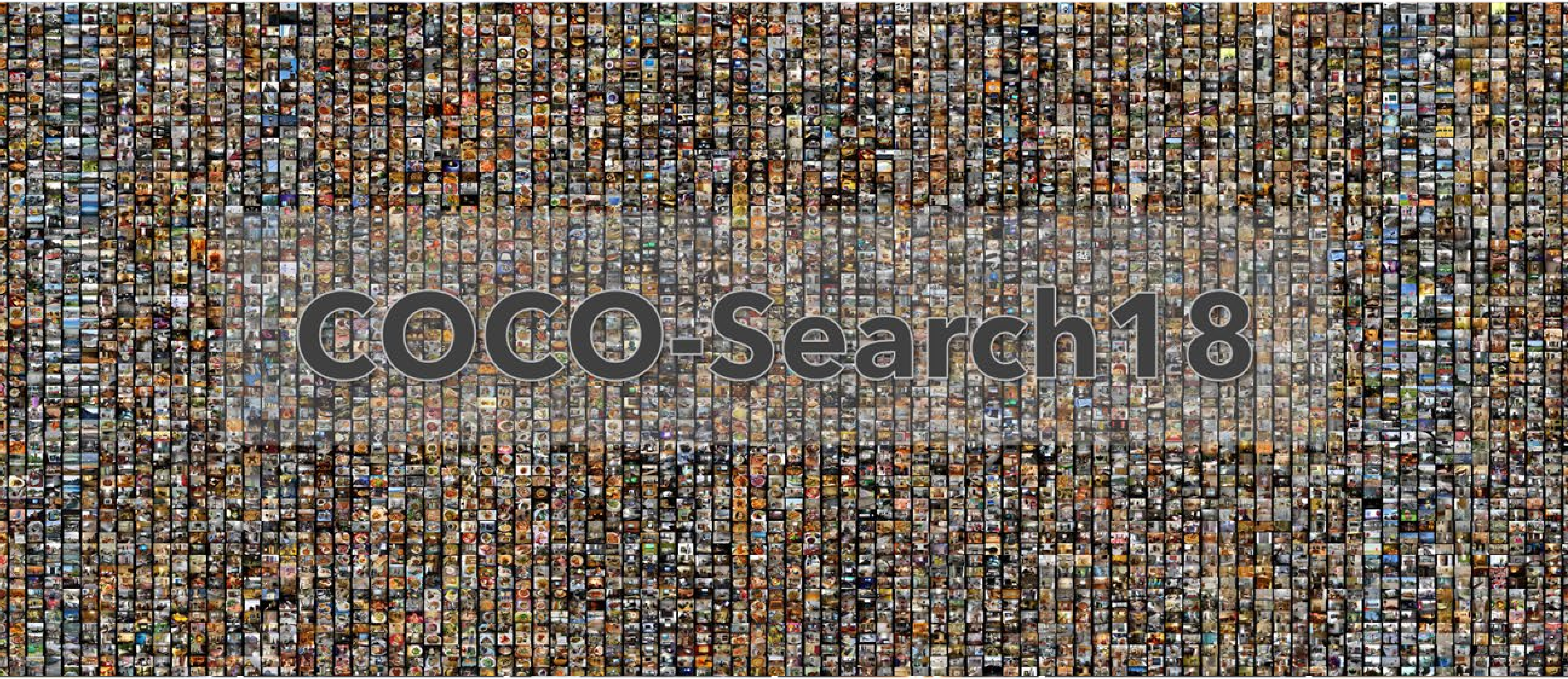
Gaze prediction for Visual Search

- Predict human scanpath for categorical visual search.

Microwave search

Clock search

COCO-Search18



Available at https://github.com/cvlab-stonybrook/Scanpath_Prediction

Predicting Goal-directed Human Attention Using Inverse Reinforcement Learning (CVPR 2020)

Collected behavior data

Dynamic Contextual Beliefs

State
s

Actions **Fixations**

$S_1, R_1, A_2, S_2, R_2, A_3, \ldots, A_{n-1}, S_n, R_n$

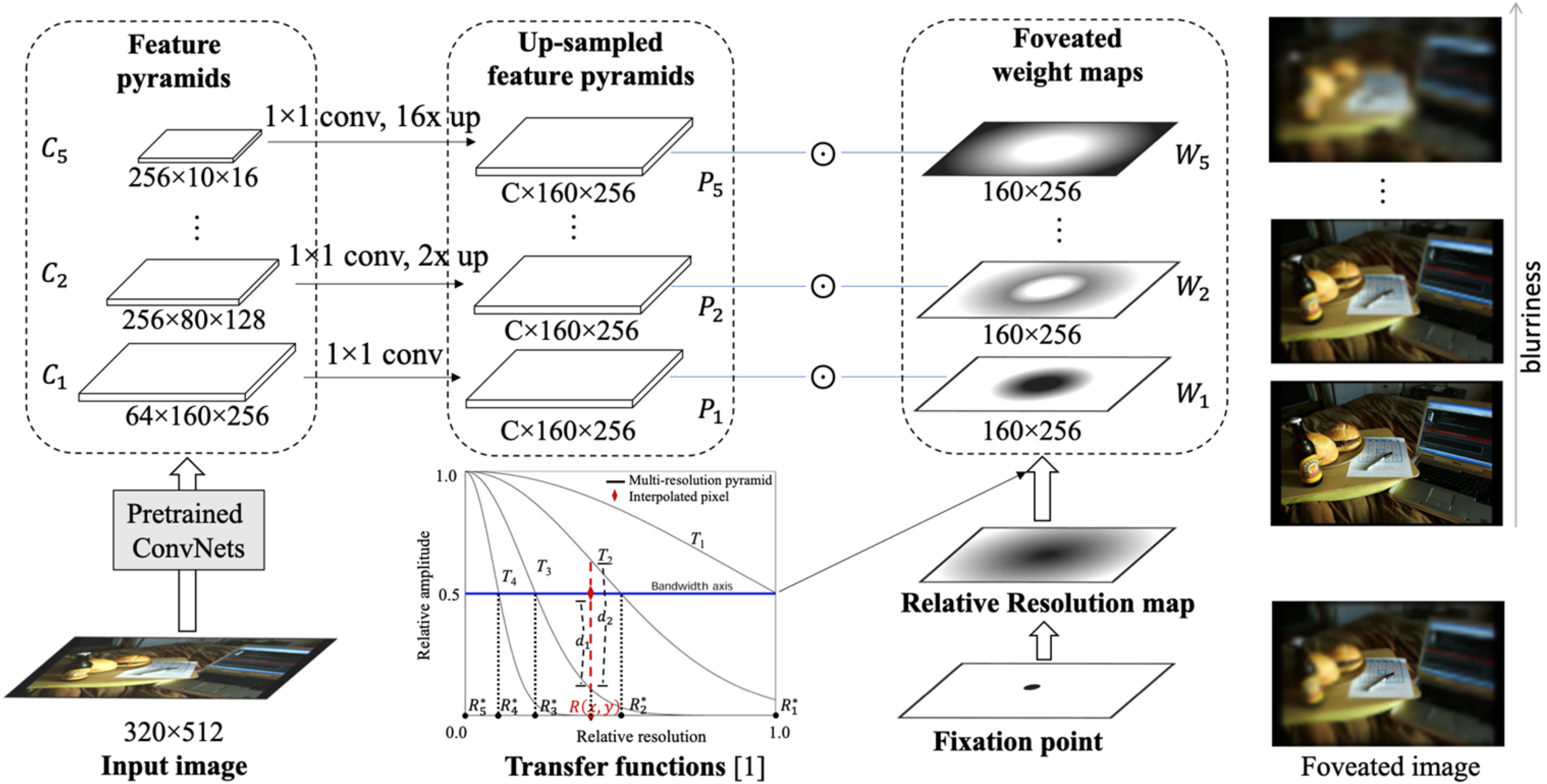
Reward
s

Unknown

Reward can be learned using *inverse reinforcement learning*

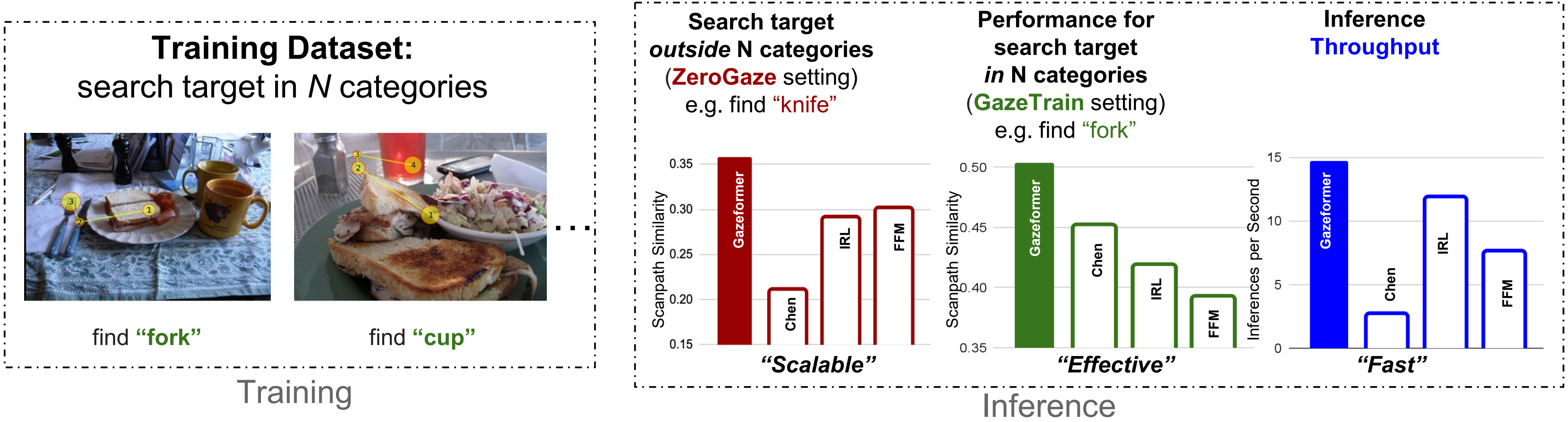
Key assumption: human gaze behaviors are optimal with respect to quickly locating the target (i.e., maximizing the total rewards)

Foveated feature maps (ECCV 2022)



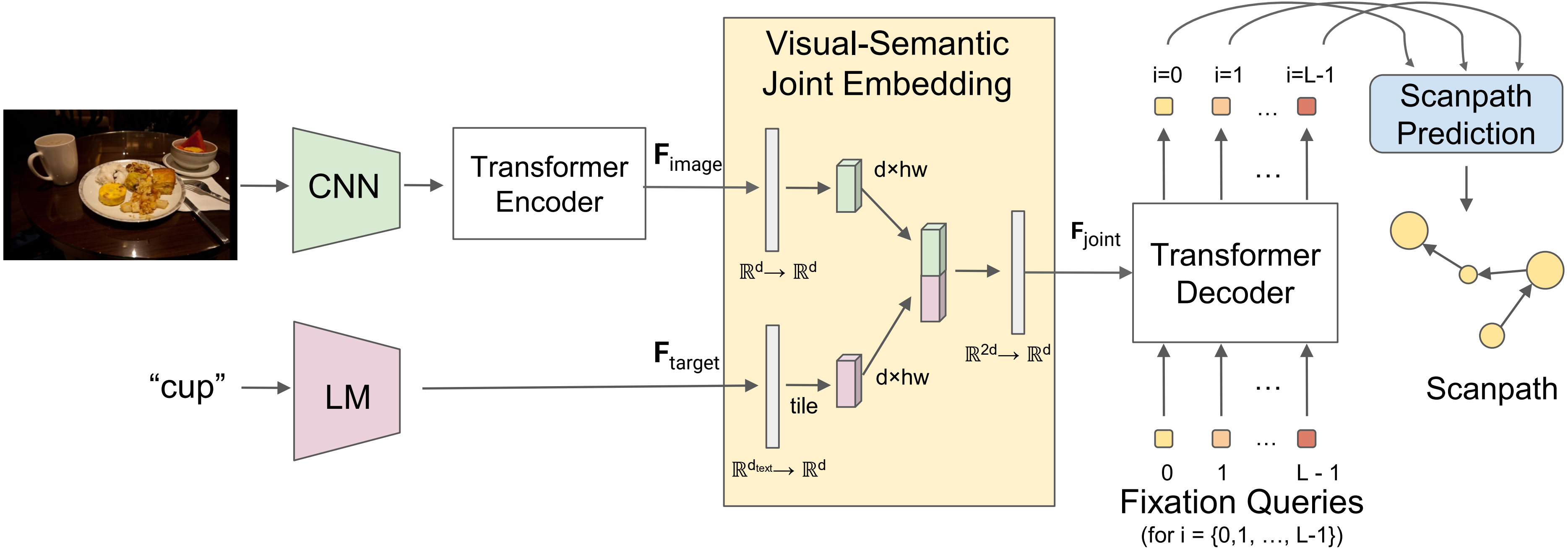
Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention (CVPR 2023)

- We propose a novel ***ZeroGaze*** task to evaluate scalability
- We propose a novel ***Gazeformer*** *model to solve ZeroGaze*
  - *Gazeformer* is more scalable, more effective and faster than previous methods



Gazeformer Architecture

- *Gazeformer* adopts a transformer encoder-decoder architecture
  - Learns interactions between image and target semantics
  - Models spatio-temporal context for scanpath generation



Gazeformer's Extensibility to Uncommon Categories

Hyponyms or synonyms of target names

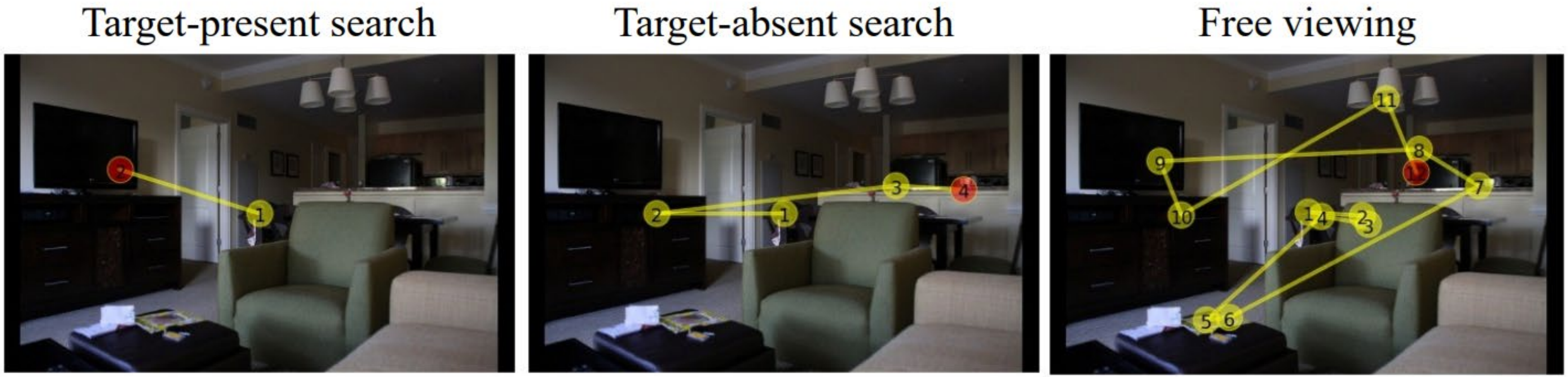
find "hatchback" find "sedan" find "mug"

No annotation in COCO dataset

find "trash can" find "pizza cutter" find "soda can"

- *Gazeformer* extends to **unknown and uncommon targets**

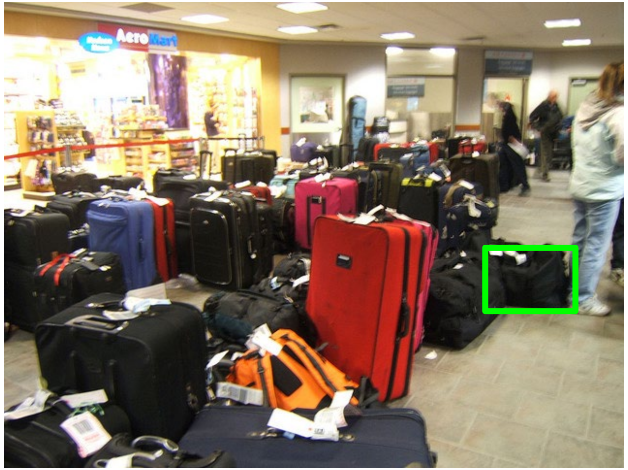
Unifying Prediction of Top-down and Bottom-up attention (CVPR 2024)



- A single model for both top-down (visual search) and bottom-up (free-viewing) attention prediction.
  - TV for target-present (TP), sink for target-absent (TA)
- Human Attention Transformer (HAT)

Current work: Visual Search with Referring Expressions

- In real life,
  - More than one object of same type
  - We use **referring expressions**
    - Instance-level
    - Resolve ambiguity
    - Provide search guidance
  - Visual Grounding of referring expressions
    - Also called object referral
    - Naturalistic visual search



Current Work: RefCOCO-Gaze

- RefCOCO-Gaze dataset
  - Based on RefCOCO dataset
    - MS-COCO training images
    - Referring expressions from RefCOCO
  - ~2000 image-text pairs from RefCOCO
    - Gaze collected *while* listening to the referring expression