INTRODUCTION TO DEEPFAKE

Prepared by: Jet Kan, Yu Tianze
Advised by: Fang Cheng Fang

THE DEEPFAKE BATTLEGROUND

Since its inception in 2016, rapid development of DeepFake in both generation and detection has formed the relationship of battleground, pushing the improvements of each other and inspiring new directions.

manuscript No.
(will be inserted by the editor)

Countering Malicious DeepFakes: Survey, Battleground, and Horizon

Felix Juefei-Xu · Run Wang · Yihao Huang · Qing Guo · Lei Ma · Yang Liu

arXiv:2103.00218v3 [cs.CV] 23 Mar 2022

Received: date / Accepted: date

Abstract The creation or manipulation of facial appearance through deep generative approaches, known as *DeepFake*, have achieved significant progress and promoted a wide range of benign and malicious applications, *e.g.*, visual effect assistance in movie and misinformation generation by faking famous persons. The evil side of this new technique poses another popular study, *i.e.*, *DeepFake detection* aiming to identify the fake faces from the real ones. With the rapid development of the DeepFake-related studies in the community, both sides (*i.e.*, DeepFake generation and detection) have formed the relationship of battleground, pushing the improvements of each other and inspiring new directions, *e.g.*, the evasion of DeepFake detection. Nevertheless, the overview of such battleground and the new direction is unclear and neglected by recent surveys due to the rapid increase of related publications, limiting the in-depth understanding of the tendency and future works.

To fill this gap, in this paper, we provide a comprehensive overview and detailed analysis of the research work on the topic of DeepFake generation, DeepFake detection as well as evasion of DeepFake detection, with more than 318 research papers carefully surveyed. We present the taxonomy of various DeepFake generation methods and the categorization of various DeepFake detection methods, and more importantly, we showcase the battleground between the two parties with detailed interactions between the adversaries (DeepFake generation) and the defenders (DeepFake detection). The battleground allows fresh perspective into the latest landscape of the DeepFake research and can provide valuable analysis towards the research challenges and opportunities as well as research trends and future directions. We also elaborately design interactive diagrams (http://www.xujuefei.com/dfsurvey) to allow researchers to explore their own interests on popular DeepFake generators or detectors.

Keywords DeepFake Generation · DeepFake Detection · Face · Misinformation · Disinformation · DeepFakes

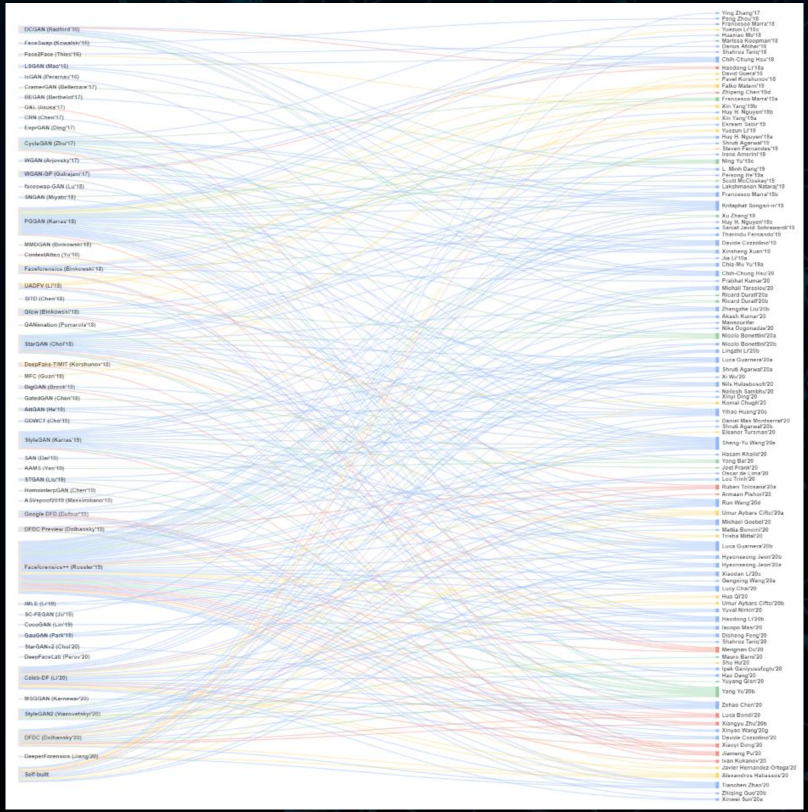
1 Introduction

> If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.
>
> *The Art of War*
> *Sun Tzu*

Ever since digital visual media came along, there has always been a need to manipulate them for various purposes. Usually, such digital media manipulation requires domain expertise and is quite time and effort consuming, such as using professional software like Adobe Photoshop (Adobe, 2021c) for editing a photograph, or Adobe Lightroom (Adobe, 2021b) for retouching it. In the sound and voice domain, similar professional software is available for carrying out various types of signal manipulation such as using Adobe Audition (Adobe, 2021a), or Auto-Tune (Antares Audio Technologies, 2021), *etc.* In the domain of motion pictures, the manipulation of videos oftentimes require very sophisticated theatrical visual effects (VFX) in the post-processing, and when it comes to recreating animated faces with realistic facial muscle movements and expressions, motion capture techniques with the help of

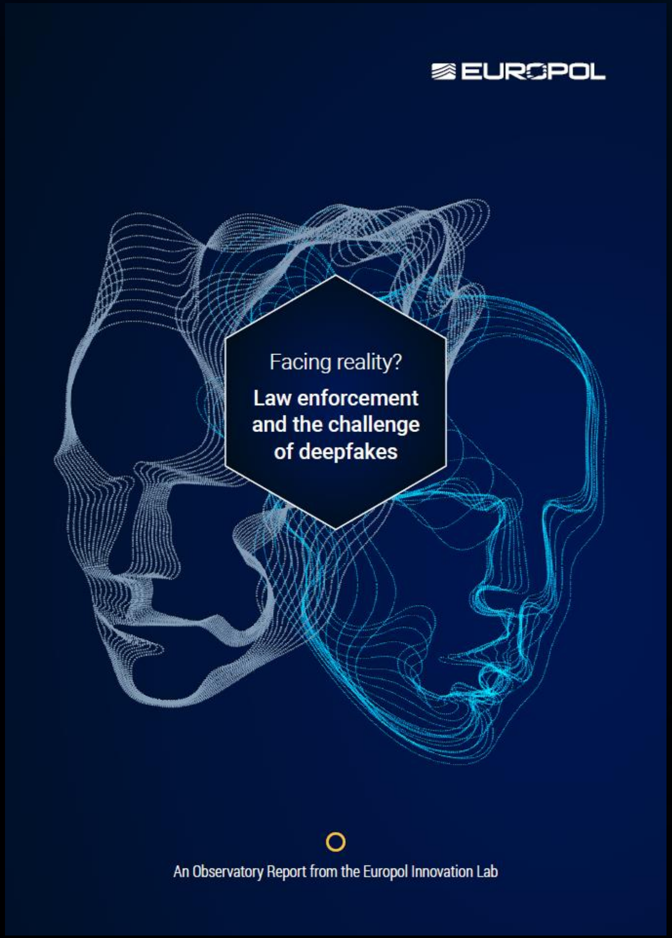
Felix Juefei-Xu is with Alibaba Group, USA · Run Wang is with the Key Laboratory of Aerospace Information Security and Trust Computing, School of Cyber Science and Engineering, Wuhan University, China · Yihao Huang is with East China Normal University, China · Qing Guo is with College of Intelligence and Computing, Tianjin University, China, and Nanyang Technological University, Singapore · Lei Ma is with University of Alberta and Alberta Machine Intelligence Institute (Amii), Canada · Yang Liu is with Zhejiang Sci-Tech University, China and Nanyang Technological University, Singapore · Corresponding authors: Run Wang (email: wangrun@whu.edu.cn) and Qing Guo (email: tsingqguo@ieee.org) · Project lead: Felix Juefei-Xu.

DF survey 2022



Interactions between generation and detection methods

EUROPOL INNOVATION LAB



EUROPOL

Facing reality?

Law enforcement and the challenge of deepfakes

An Observatory Report from the Europol Innovation Lab

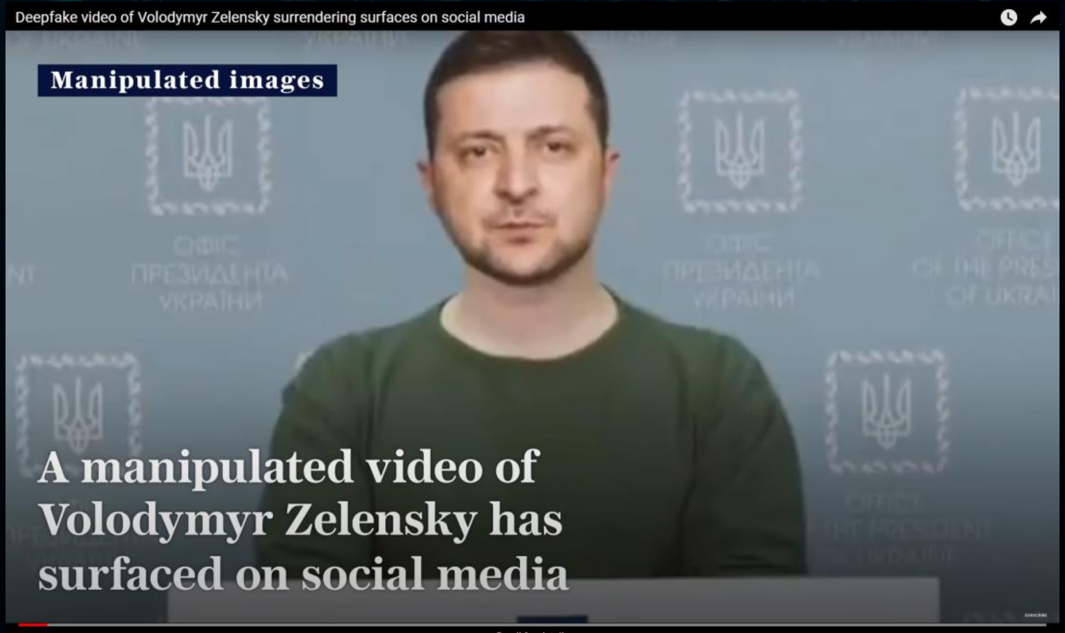
- Europol: mandated by the EU to support the law enforcement in innovation
- Their first report published in 28 April 2022 explores the topic of DeepFake
- Summary:
  - Layman's introduction to DeepFake and Machine Learning
  - Impacts of DeepFake in law, financial and security
  - Current policies to combat fake videos

"Experts estimate that as much as 90% of online content may be synthetically generated by 2026."

- Europol (2022)

MALICIOUS USES OF DEEPFAKE

- Non-consensual pornography
- Document fraud
- Falsifying evidence for criminal justice investigations
- Distributing disinformation and manipulating public opinion
- Supporting the narratives of extremist or terrorist groups



Deepfake video of Volodymyr Zelensky surrendering surfaces on social media

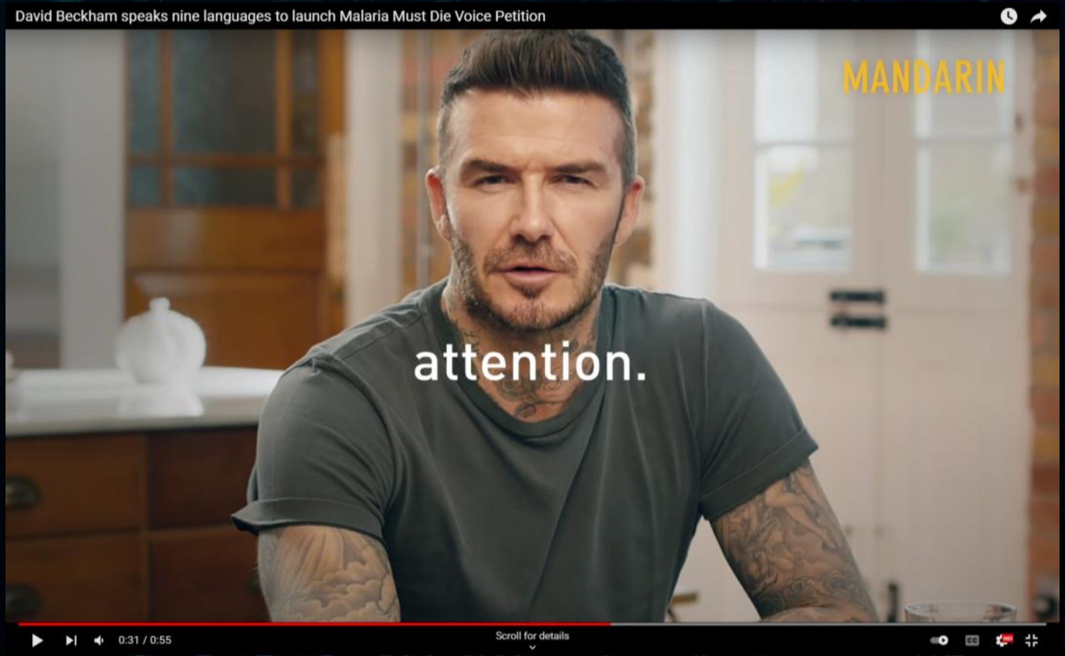
Manipulated images

A manipulated video of Volodymyr Zelensky has surfaced on social media

DeepFakes of President Zelensky telling his soldiers to surrender released by Russian hackers during the Ukraine war

BENIGN USES OF DEEPFAKE

- Anonymize voice and faces for privacy
- Cost-effective entertainment production
- Amplify the reach of public messages through language localization
- Bringing back the loved ones or imagining different stages of life


David Beckham speaks nine languages to launch Malaria Must Die Voice Petition
MANDARIN
attention.
0:31 / 0:55
Scroll for details

Multilingual social campaign video created using DeepFakes

RESPONSES FROM TECH COMPANIES

Company	Policy	Action
Meta	Removes content that has been manipulated in order to mislead users	Developed a detection tool that reverse engineers a single DF image to its generative model, created Deepfake Detection Challenge (DFDC) in 2020 and open-sourced their dataset
Google (YouTube)	Bans manipulated media under scam policies	Released large DeepFake dataset on FaceForensics
Twitter	Removes content that deceptively share synthetic or manipulated media that are likely to cause harm	Introduced a “three-pronged test” (3 human answerable questions) to determine if media violates Twitter’s policy
TikTok	Bans digital forgeries that mislead users by distorting truth and cause harm	Enforces identity check before allowing users to use their face swap filters to prevent unconsented DeepFakes
Reddit	Does not allow content that impersonates entities misleadingly	

RELATED NEWS

1 Sep 2020
Microsoft (Responsible AI) developed Microsoft Video Authenticator which analyses the percentage chance of a photo or video is artificially manipulated

16 Oct 2020
Qualcomm developed a feature which adds digital signature to each photo taken from Qualcomm Snapdragon smartphones to prove its authenticity when posted online

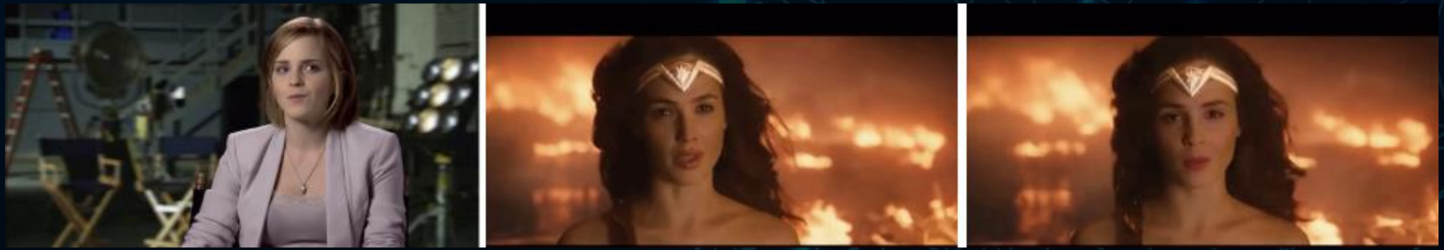
27 Jan 2022
Coalition for Content Provenance and Authenticity (C2PA) had partnered with tech giants including Microsoft, Intel, and Adobe to combat the rapid spread of DeepFake

29 Apr 2022
AI Singapore (AISG) hosted a DeepFake detection competition Trusted Media Challenge. The winner aims to incorporate his AI model into ByteDance's BytePlus platform to make it available to users

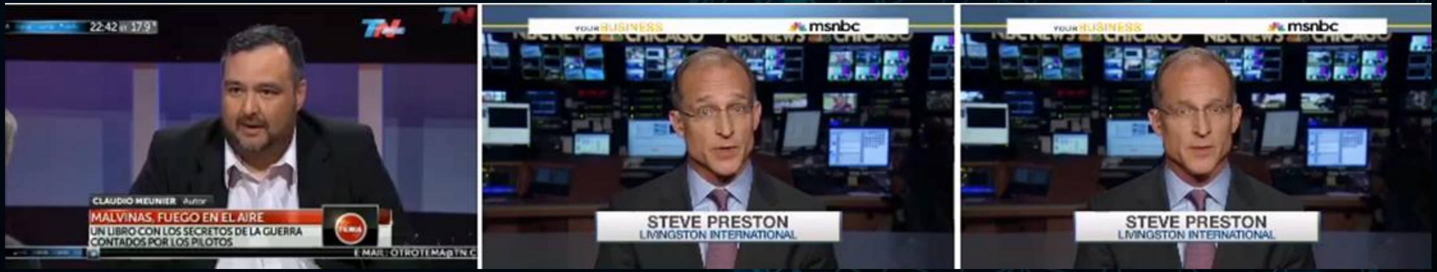
Trusted Media Challenge: https://arxiv.org/pdf/2201.04788v2.pdf

DEEPFAKE GENERATION TYPES

Identity swap

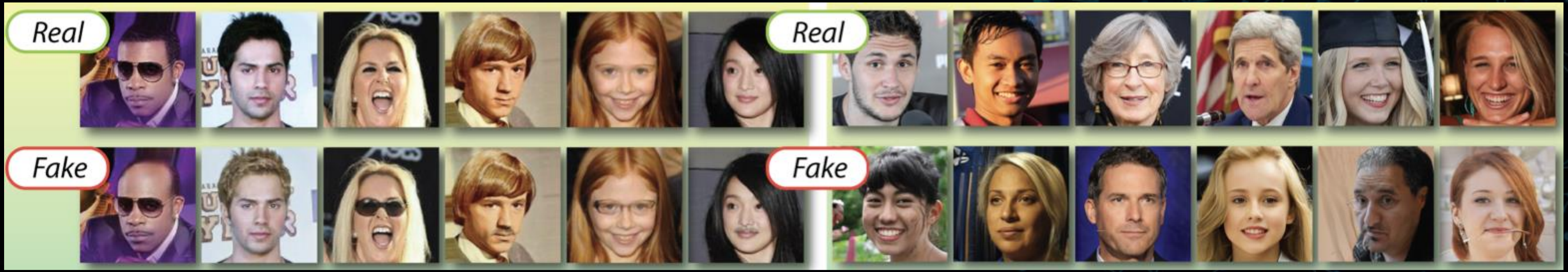


Expression re-enactment



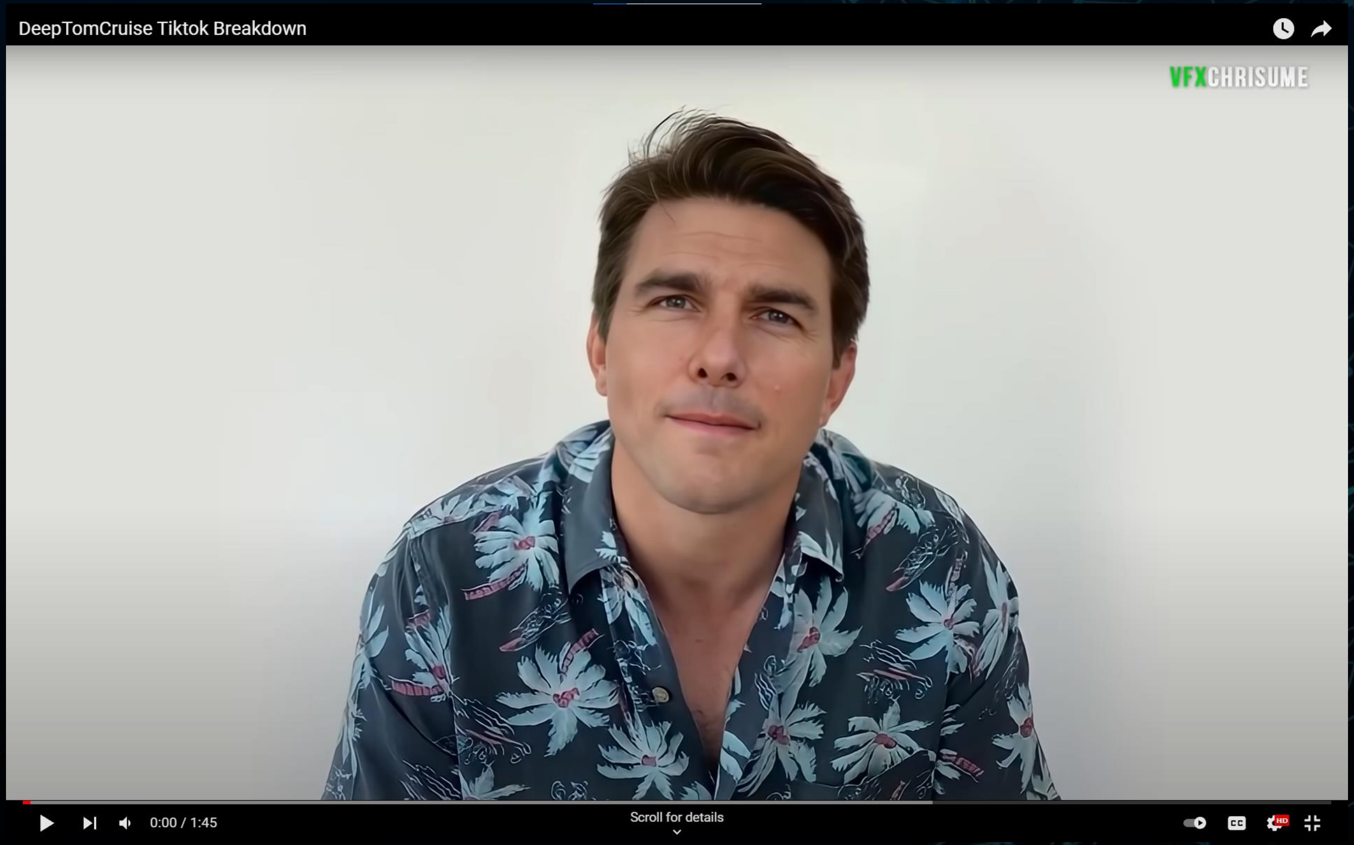
Attribute manipulation

Entire face synthesis



DEEPFACELAB

Most popular DeepFake generation tool



DEEPFAKE WORKFLOW

Data Processing

Curate a database of source & target faces

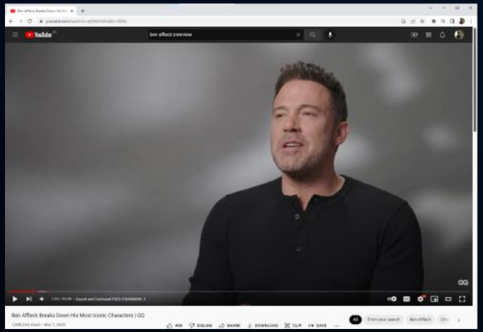
Model Training

Learn the features of both faces

Merging

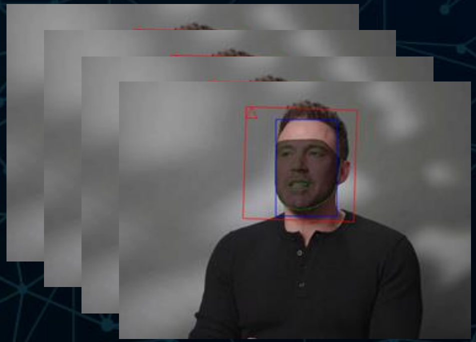
Transfer source features to target face

DATA PROCESSING OVERVIEW

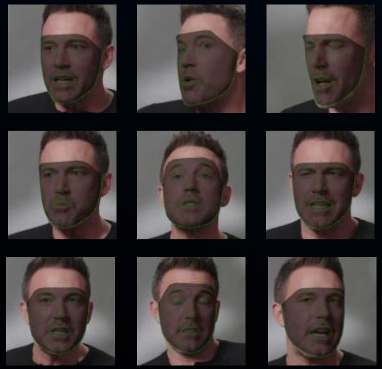


1. Data collection

2. Frame extraction

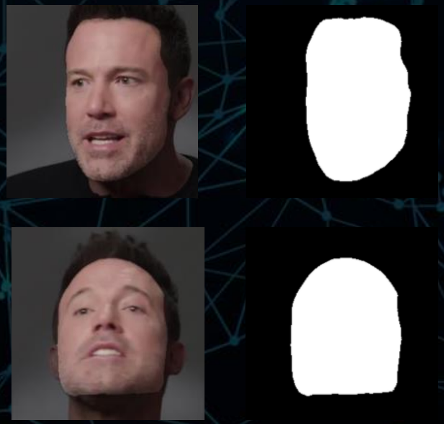


3. Face extraction



4. Landmarking

5. Filtering



6. Masking

3. FACE EXTRACTION

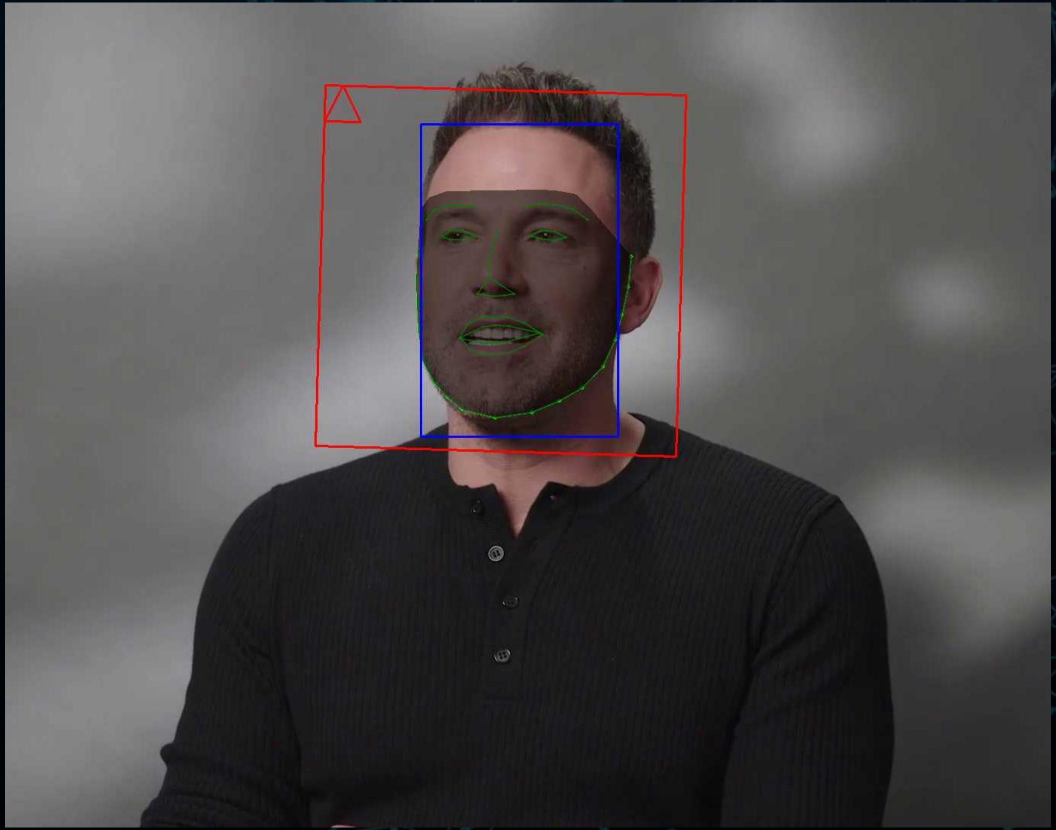
Using Single Shot Scale-invariant Face Detector (S3FD)

Red box = head

Blue box = whole face

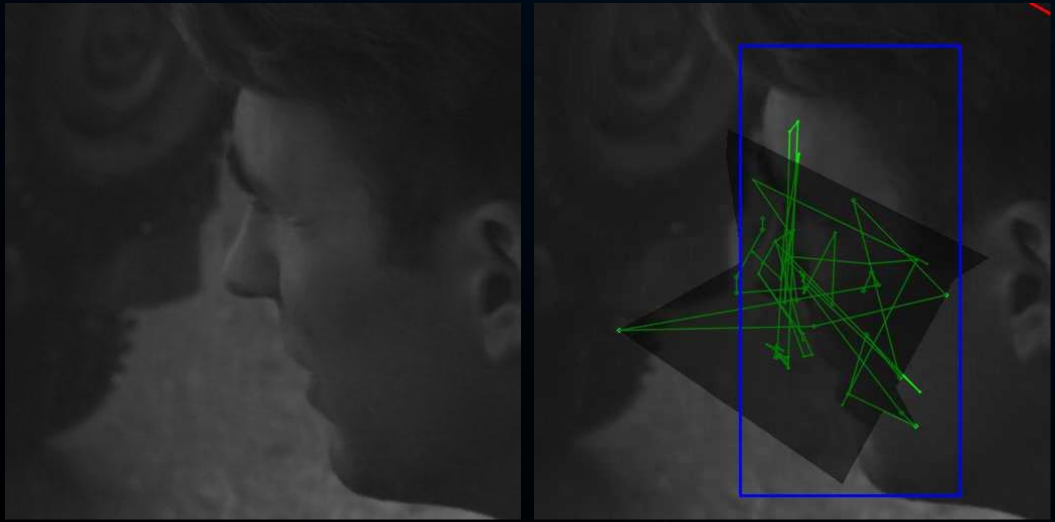
Grey area = full face

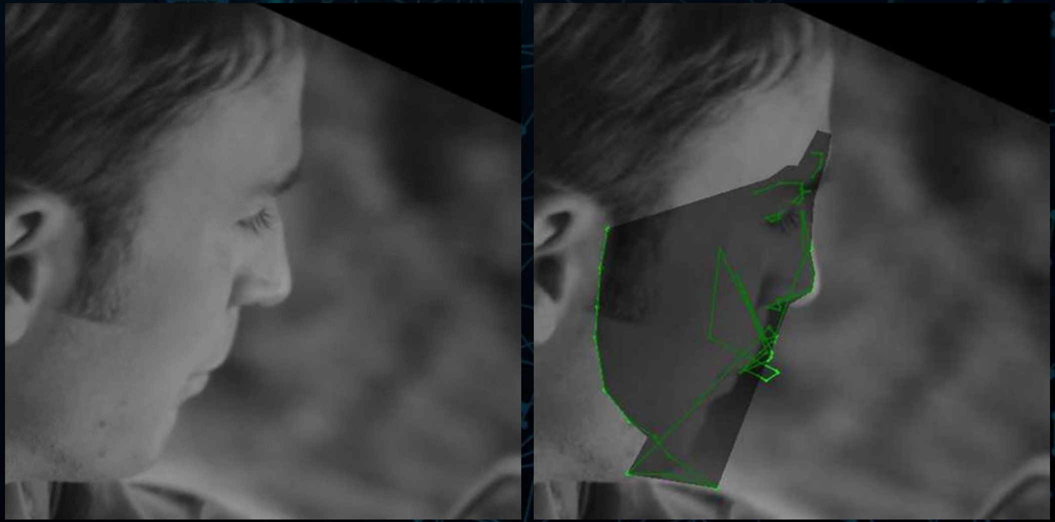
Green lines = landmarks



3. FACE EXTRACTION

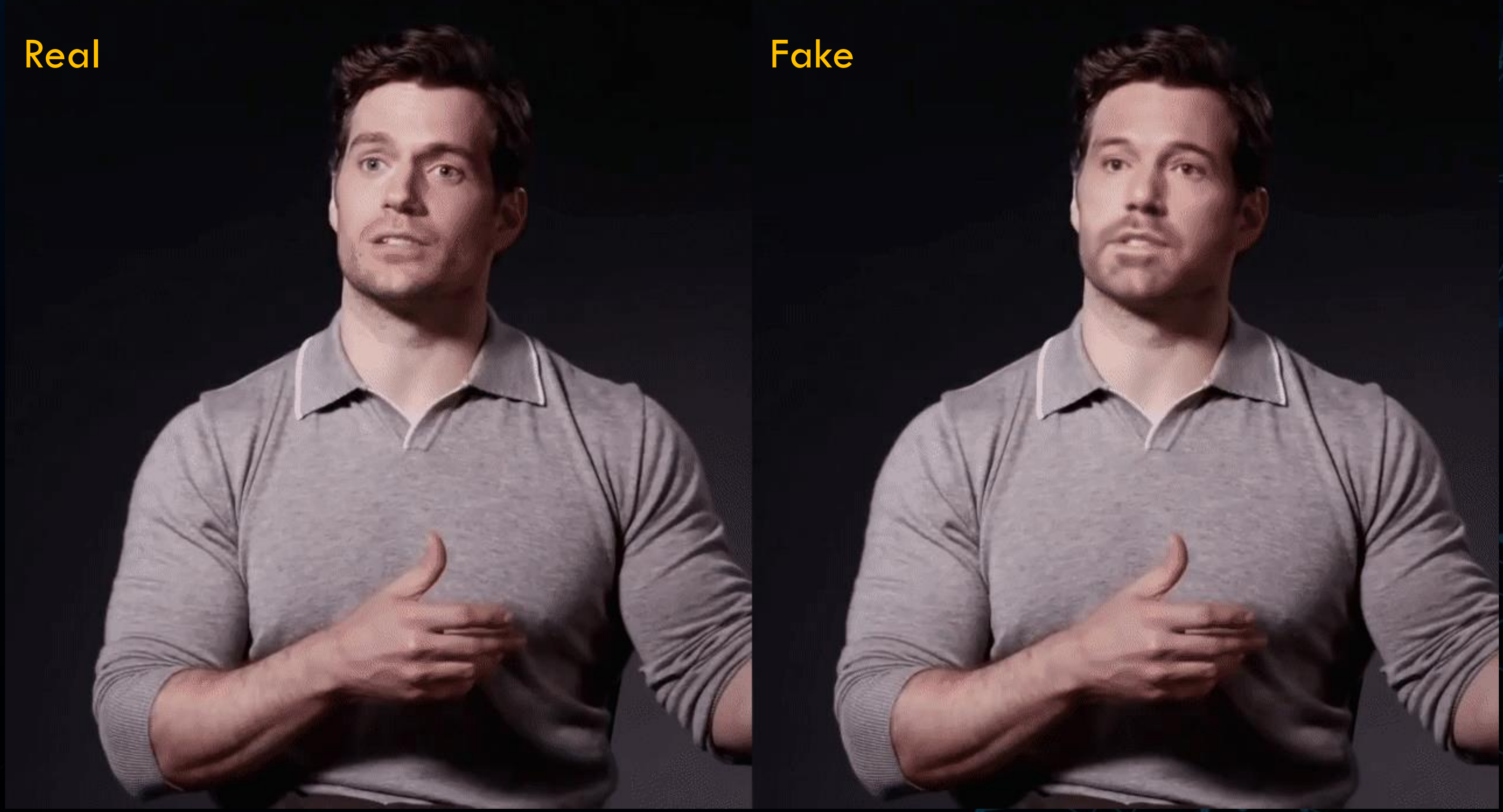
S3FD: Good at detecting frontal faces, weaker at side faces (about 3% error rate)





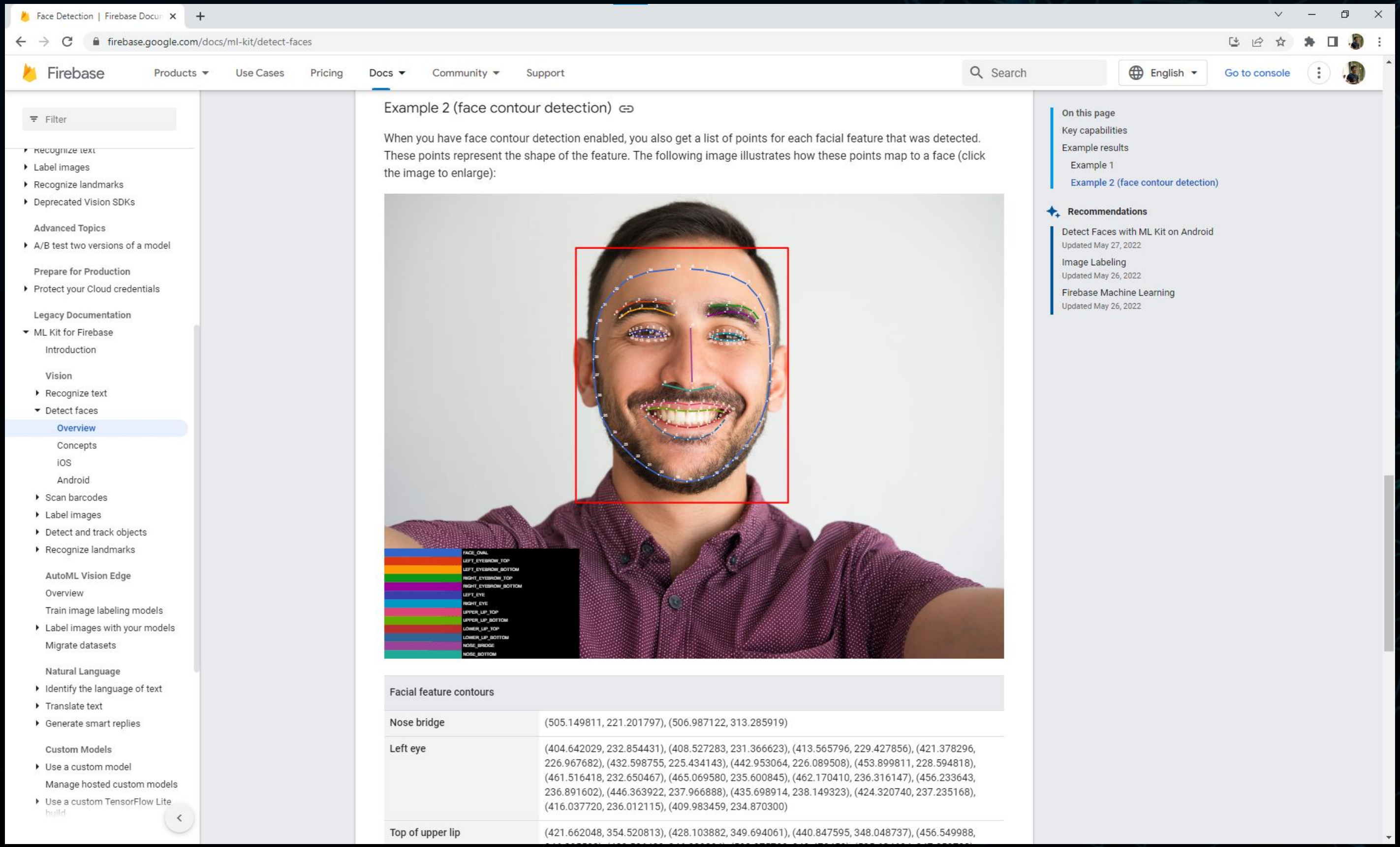
4. LANDMARKING

S3FD: Relatively small amount of landmarks (30~40), unable to capture rich emotions



4. LANDMARKING

Google's Firebase ML Kit with more detailed landmarks



Example 2 (face contour detection)

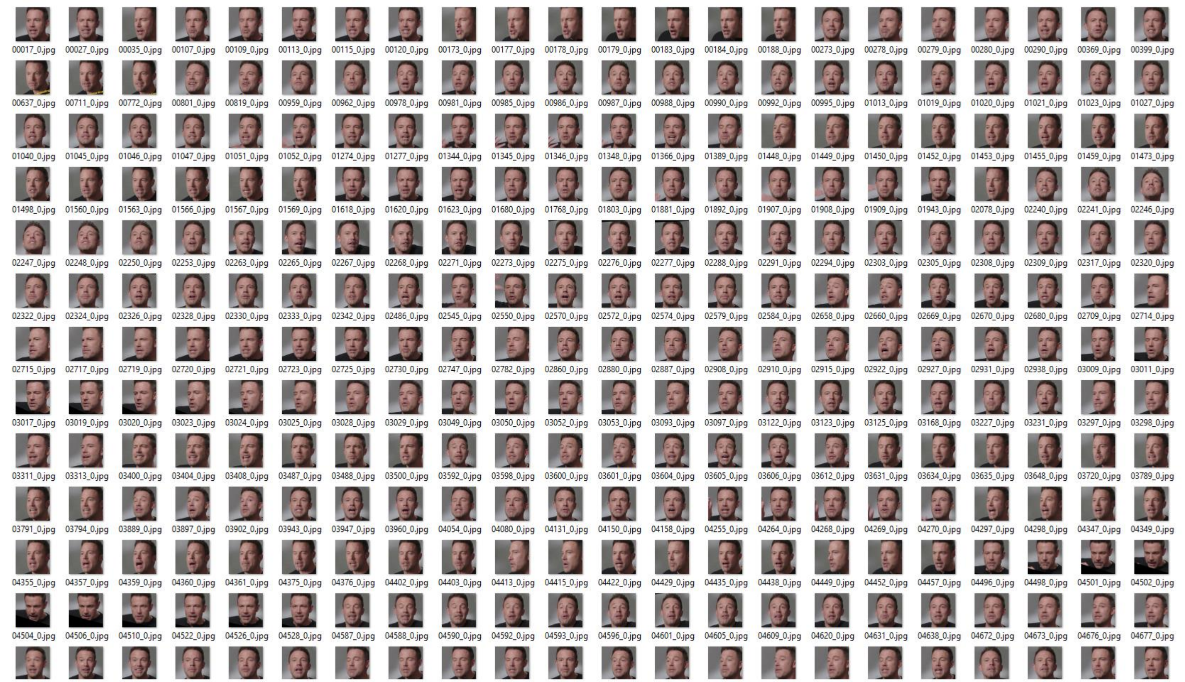
When you have face contour detection enabled, you also get a list of points for each facial feature that was detected. These points represent the shape of the feature. The following image illustrates how these points map to a face (click the image to enlarge):

Facial feature contours	
Nose bridge	(505.149811, 221.201797), (506.987122, 313.285919)
Left eye	(404.642029, 232.854431), (408.527283, 231.366623), (413.565796, 229.427856), (421.378296, 226.967682), (432.598755, 225.434143), (442.953064, 226.089508), (453.899811, 228.594818), (461.516418, 232.650467), (465.069580, 235.600845), (462.170410, 236.316147), (456.233643, 236.891602), (446.363922, 237.966888), (435.698914, 238.149323), (424.320740, 237.235168), (416.037720, 236.012115), (409.983459, 234.870300)
Top of upper lip	(421.662048, 354.520813), (428.103882, 349.694061), (440.847595, 348.048737), (456.549988,

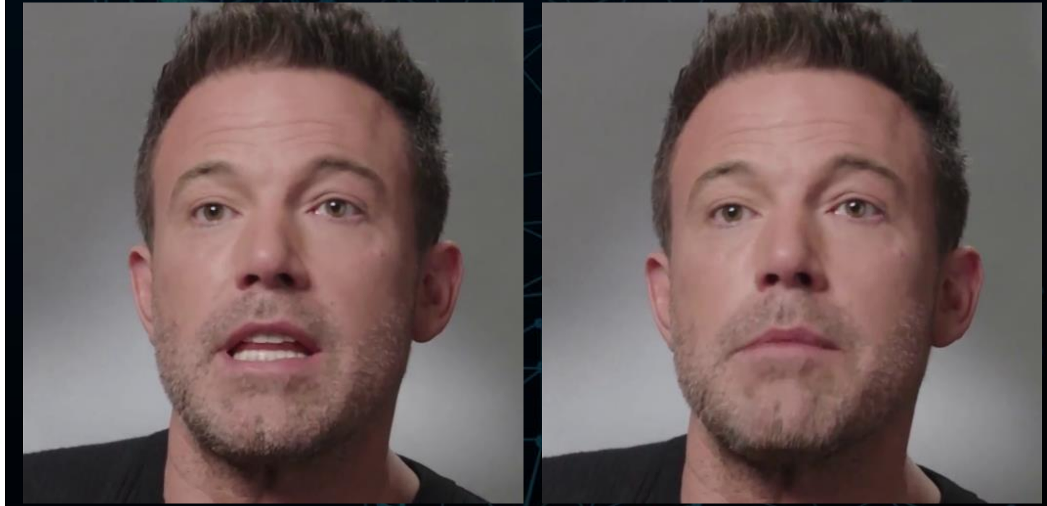
5. SORTING

Goal: cover as many expressions, lighting, and angle as possible

Challenge: remove duplicate faces from a pool of very similar images



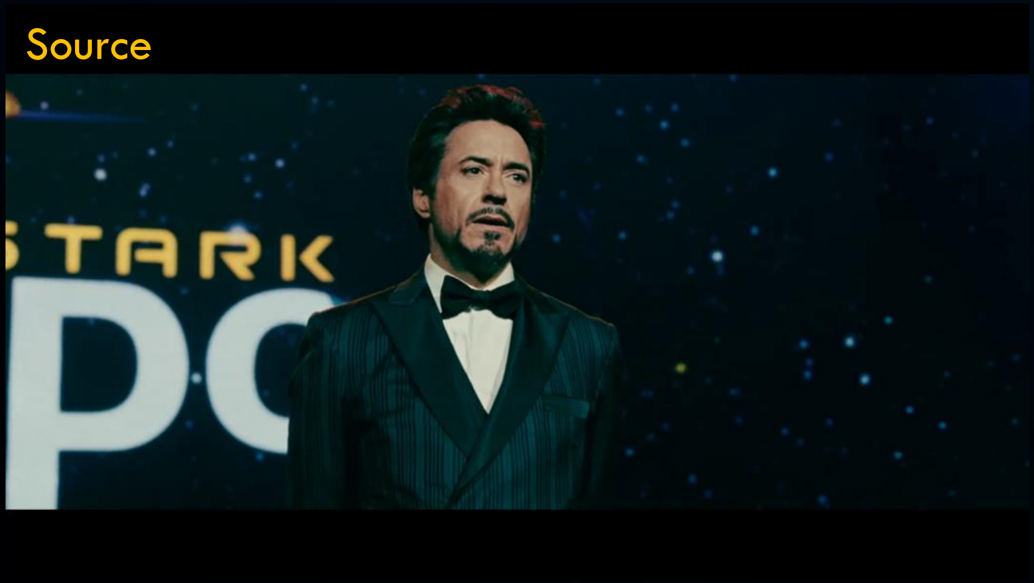
Highly similar faces are removed using Visipics



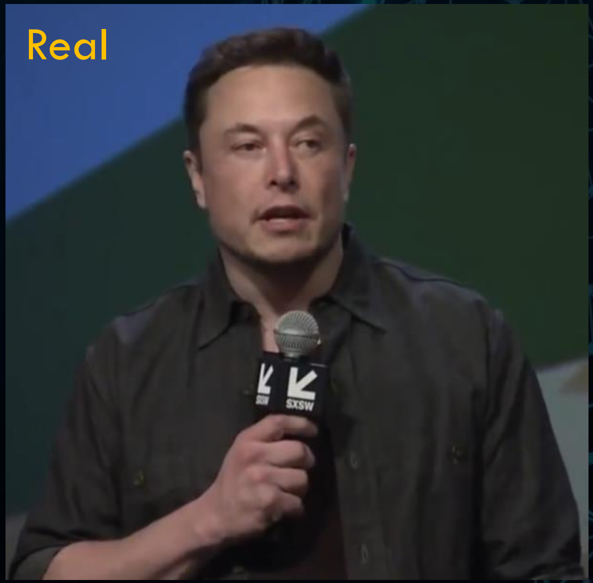
Tip: keep both faces!

5. SORTING

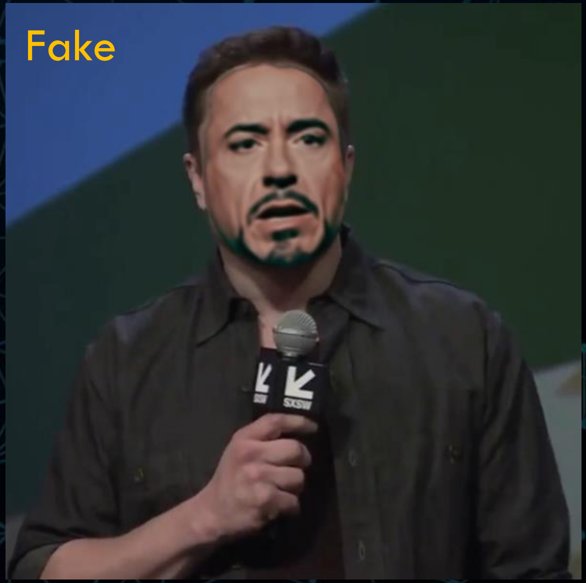
Lack of training data results in pitch black mouth area



Source



Real

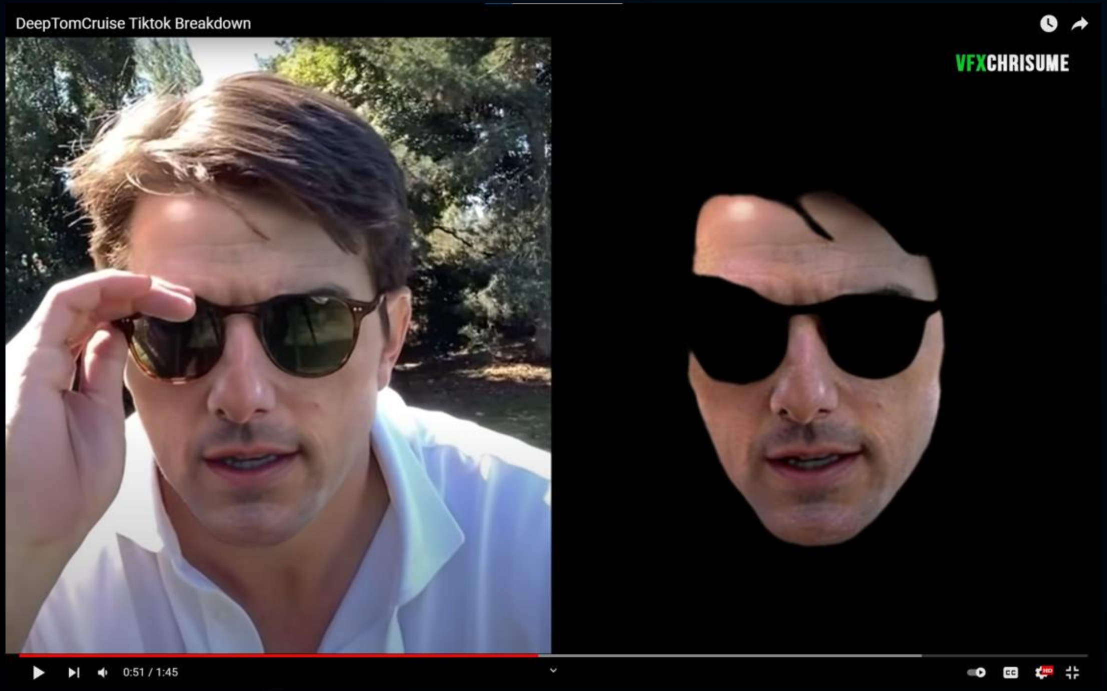


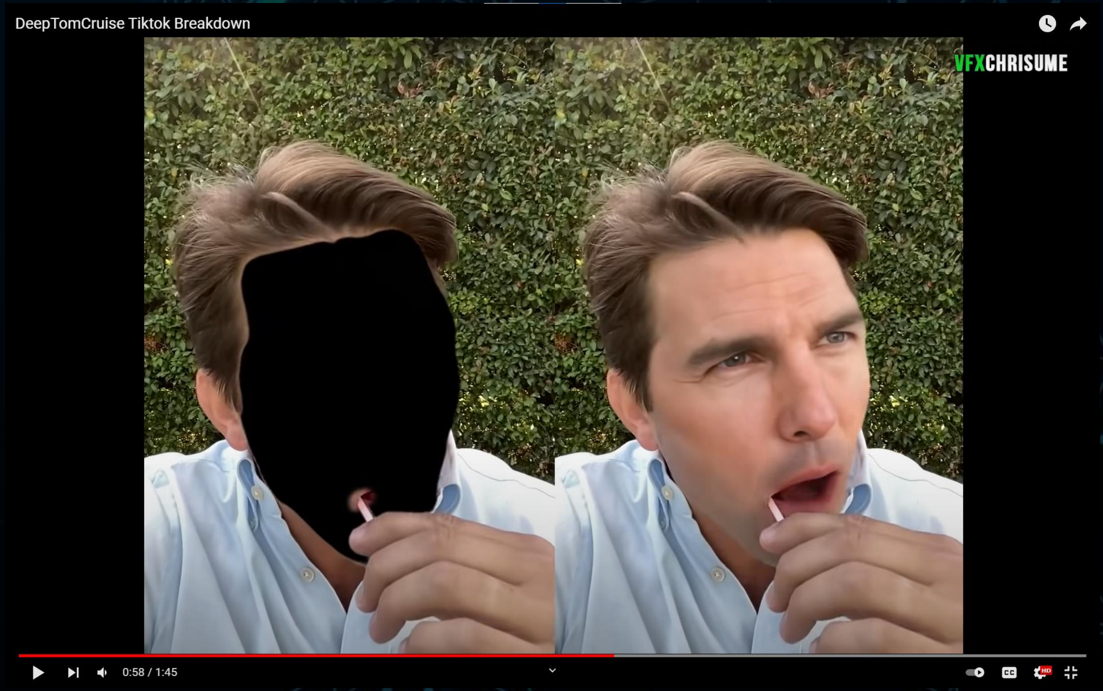
Fake

6. MASKING

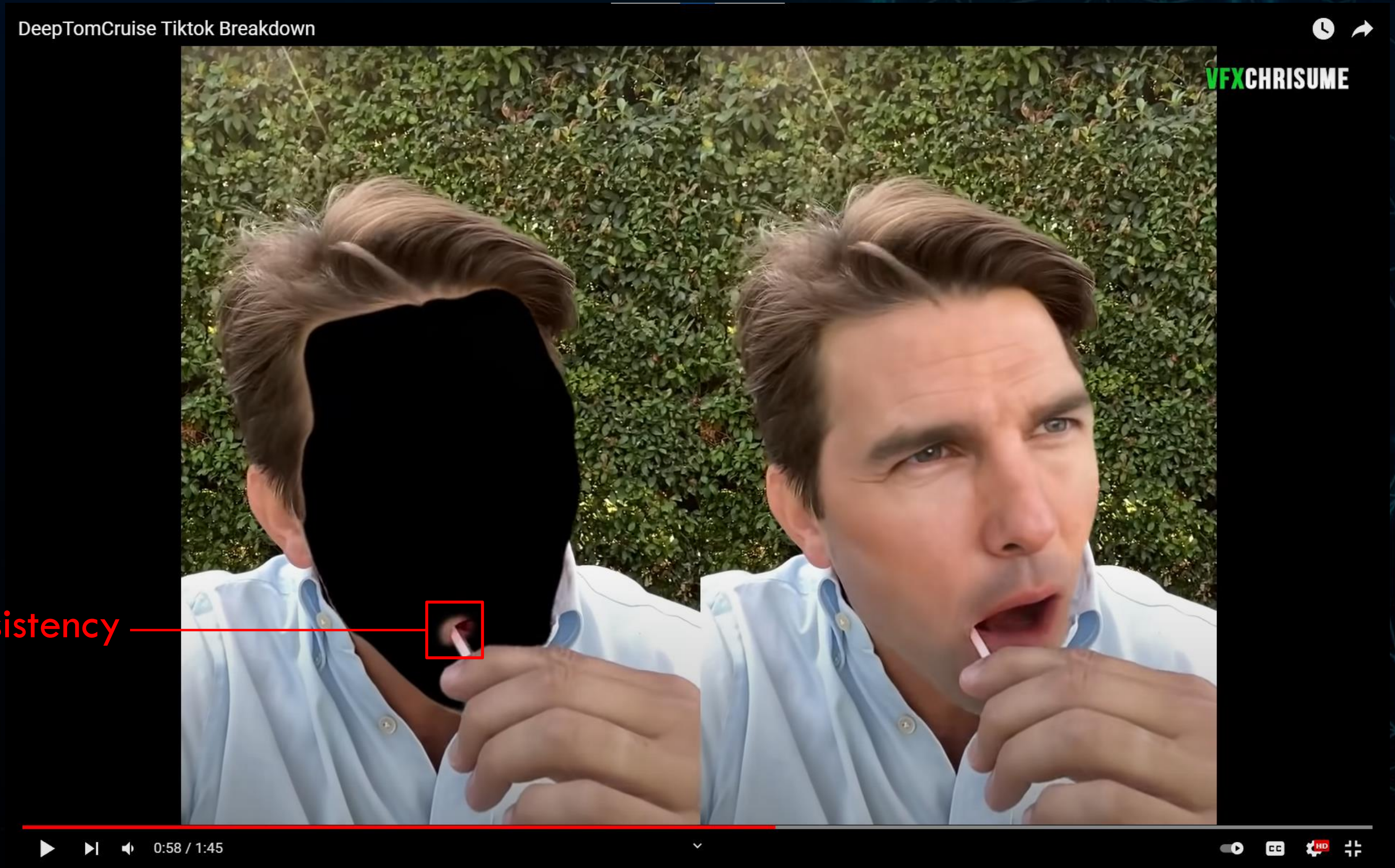
Remove unrelated features such as hair, glasses, hand, etc.

Generic pre-trained XSeg model performs well in most cases



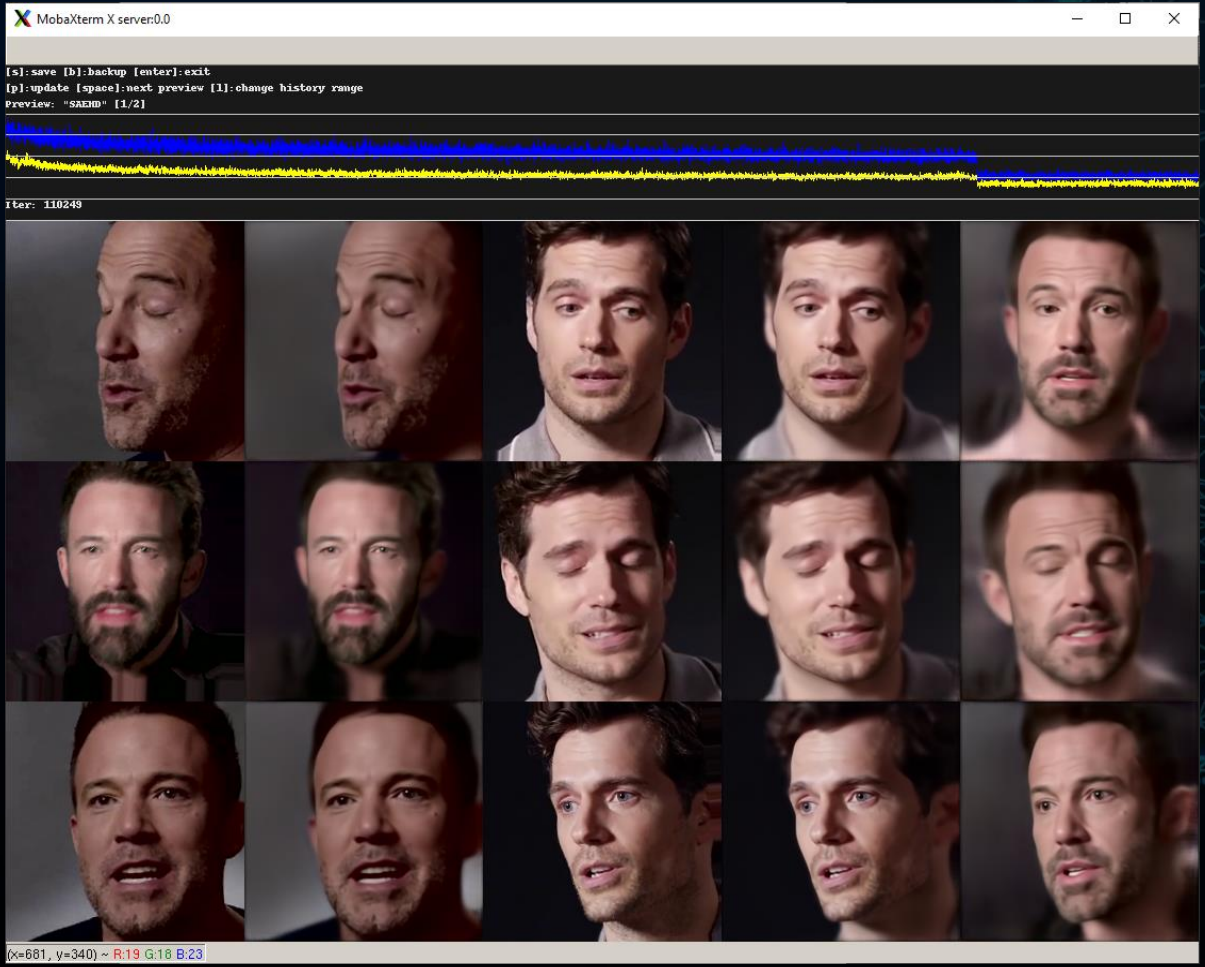


6. MASKING

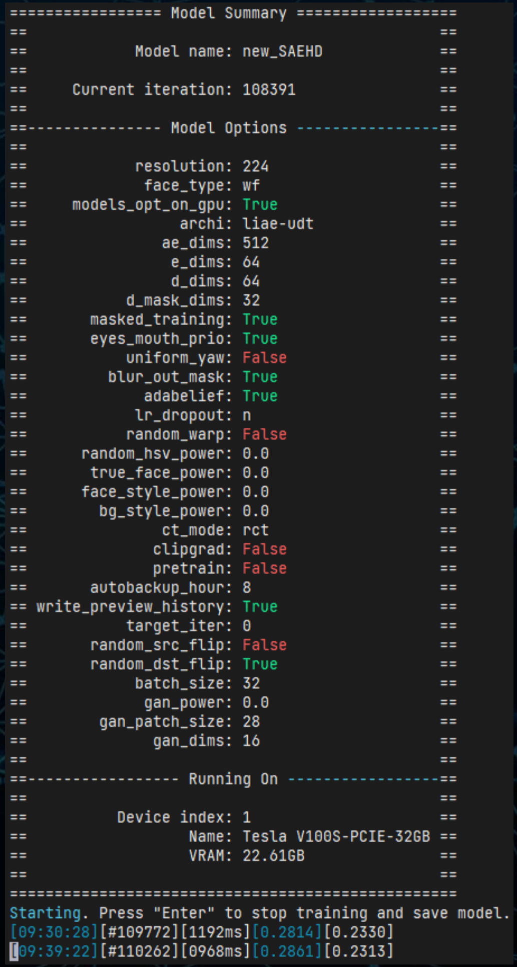


Detect texture inconsistency

MODEL TRAINING

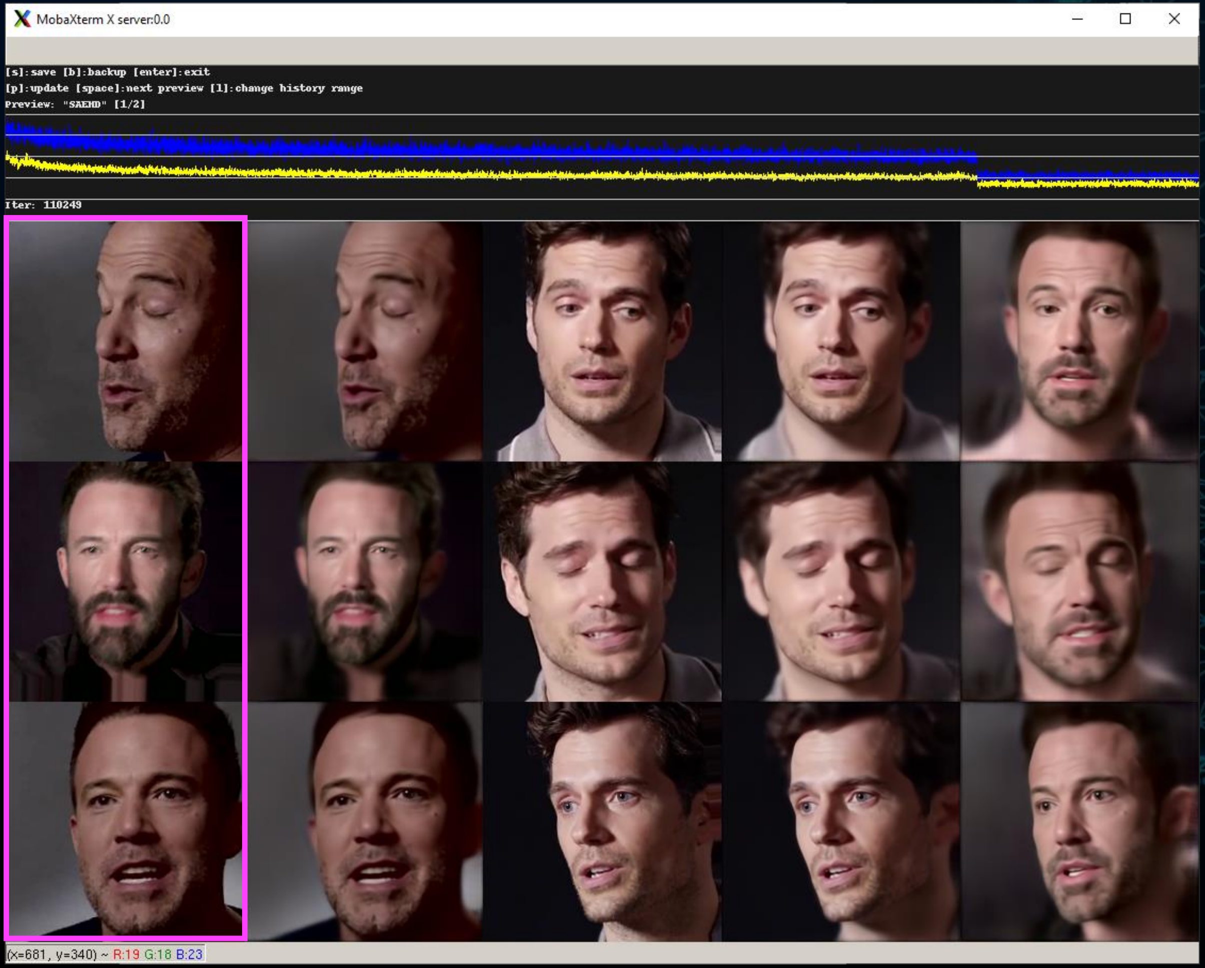


```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]
Iter: 110249
(x=681, y=340) ~ R:19 G:18 B:23
```



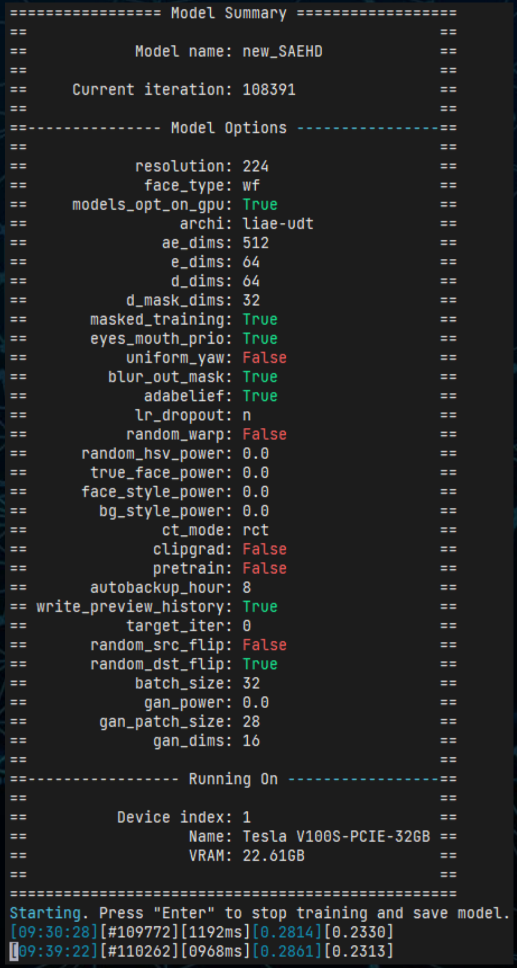
```
================ Model Summary =================
==                                            ==
==            Model name: new_SAEHD           ==
==                                            ==
==     Current iteration: 108391              ==
==                                            ==
==--------------- Model Options --------------==
==                                            ==
==            resolution: 224                 ==
==             face_type: wf                  ==
==     models_opt_on_gpu: True                ==
==                 archi: liae-udt            ==
==               ae_dims: 512                 ==
==                e_dims: 64                  ==
==                d_dims: 64                  ==
==           d_mask_dims: 32                  ==
==       masked_training: True                ==
==       eyes_mouth_prio: True                ==
==           uniform_yaw: False               ==
==         blur_out_mask: True                ==
==             adabelief: True                ==
==            lr_dropout: n                   ==
==           random_warp: False               ==
==      random_hsv_power: 0.0                 ==
==       true_face_power: 0.0                 ==
==      face_style_power: 0.0                 ==
==        bg_style_power: 0.0                 ==
==               ct_mode: rct                 ==
==              clipgrad: False               ==
==              pretrain: False               ==
==       autobackup_hour: 8                   ==
== write_preview_history: True                ==
==           target_iter: 0                   ==
==       random_src_flip: False               ==
==       random_dst_flip: True                ==
==            batch_size: 32                  ==
==             gan_power: 0.0                 ==
==        gan_patch_size: 28                  ==
==              gan_dims: 16                  ==
==                                            ==
==----------------- Running On ---------------==
==                                            ==
==          Device index: 1                   ==
==                  Name: Tesla V100S-PCIE-32GB ==
==                  VRAM: 22.61GB             ==
==                                            ==
================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

MODEL TRAINING



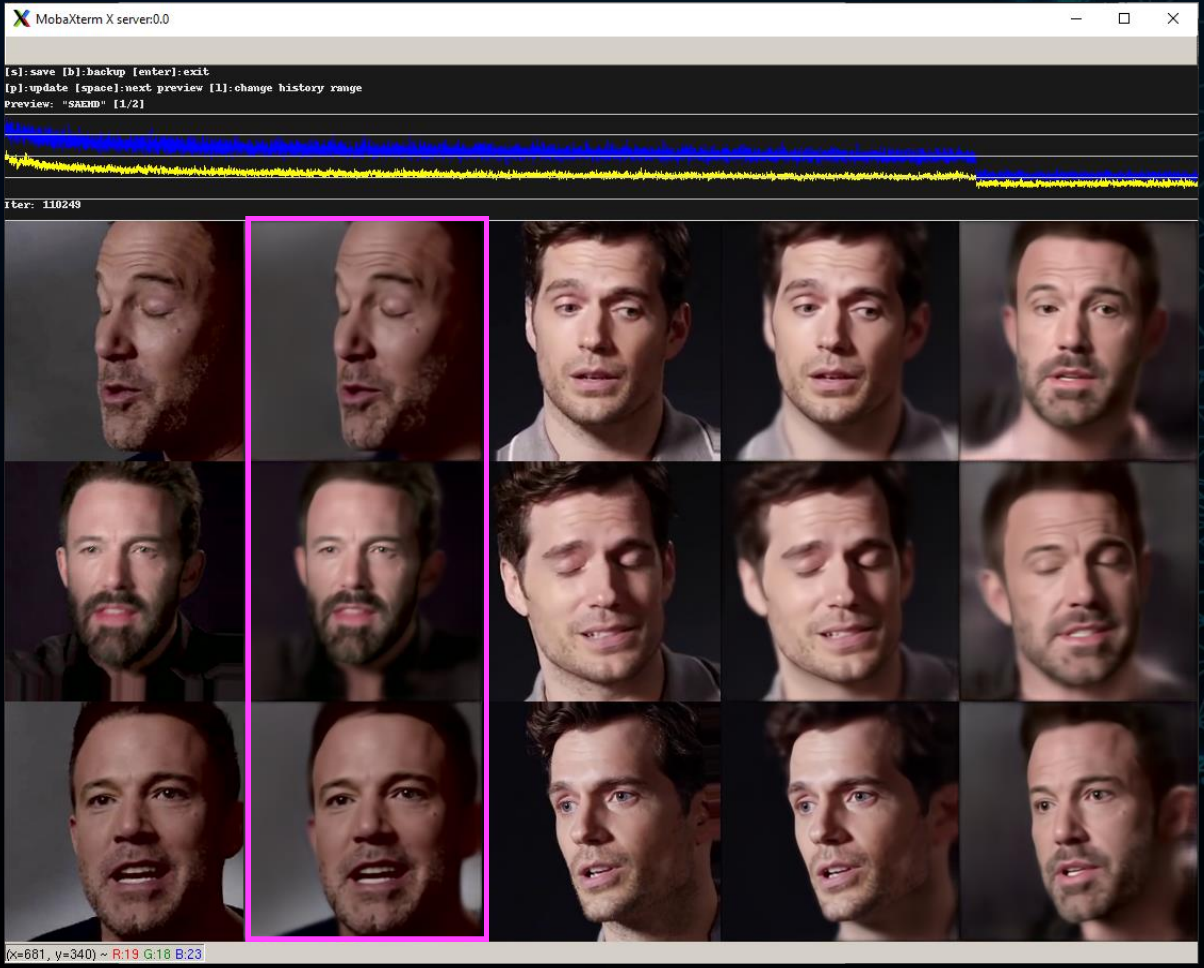
```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]

Iter: 110249
```

source



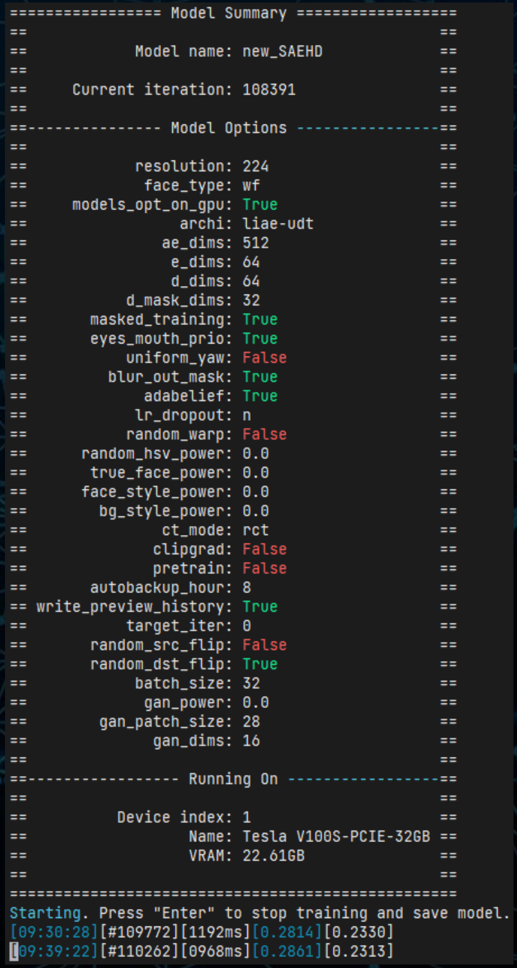
```
================= Model Summary ==================
==                                              ==
==            Model name: new_SAEHD             ==
==                                              ==
==     Current iteration: 108391                ==
==                                              ==
==--------------- Model Options ----------------==
==                                              ==
==            resolution: 224                   ==
==             face_type: wf                    ==
==     models_opt_on_gpu: True                  ==
==                 archi: liae-udt              ==
==               ae_dims: 512                   ==
==                e_dims: 64                    ==
==                d_dims: 64                    ==
==           d_mask_dims: 32                    ==
==       masked_training: True                  ==
==       eyes_mouth_prio: True                  ==
==           uniform_yaw: False                 ==
==         blur_out_mask: True                  ==
==             adabelief: True                  ==
==            lr_dropout: n                     ==
==           random_warp: False                 ==
==      random_hsv_power: 0.0                   ==
==       true_face_power: 0.0                   ==
==      face_style_power: 0.0                   ==
==        bg_style_power: 0.0                   ==
==               ct_mode: rct                   ==
==              clipgrad: False                 ==
==              pretrain: False                 ==
==       autobackup_hour: 8                     ==
== write_preview_history: True                  ==
==           target_iter: 0                     ==
==       random_src_flip: False                 ==
==       random_dst_flip: True                  ==
==            batch_size: 32                    ==
==             gan_power: 0.0                   ==
==        gan_patch_size: 28                    ==
==              gan_dims: 16                    ==
==                                              ==
==----------------- Running On ----------------==
==                                              ==
==          Device index: 1                     ==
==                  Name: Tesla V100S-PCIE-32GB ==
==                  VRAM: 22.61GB               ==
==                                              ==
==================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

MODEL TRAINING



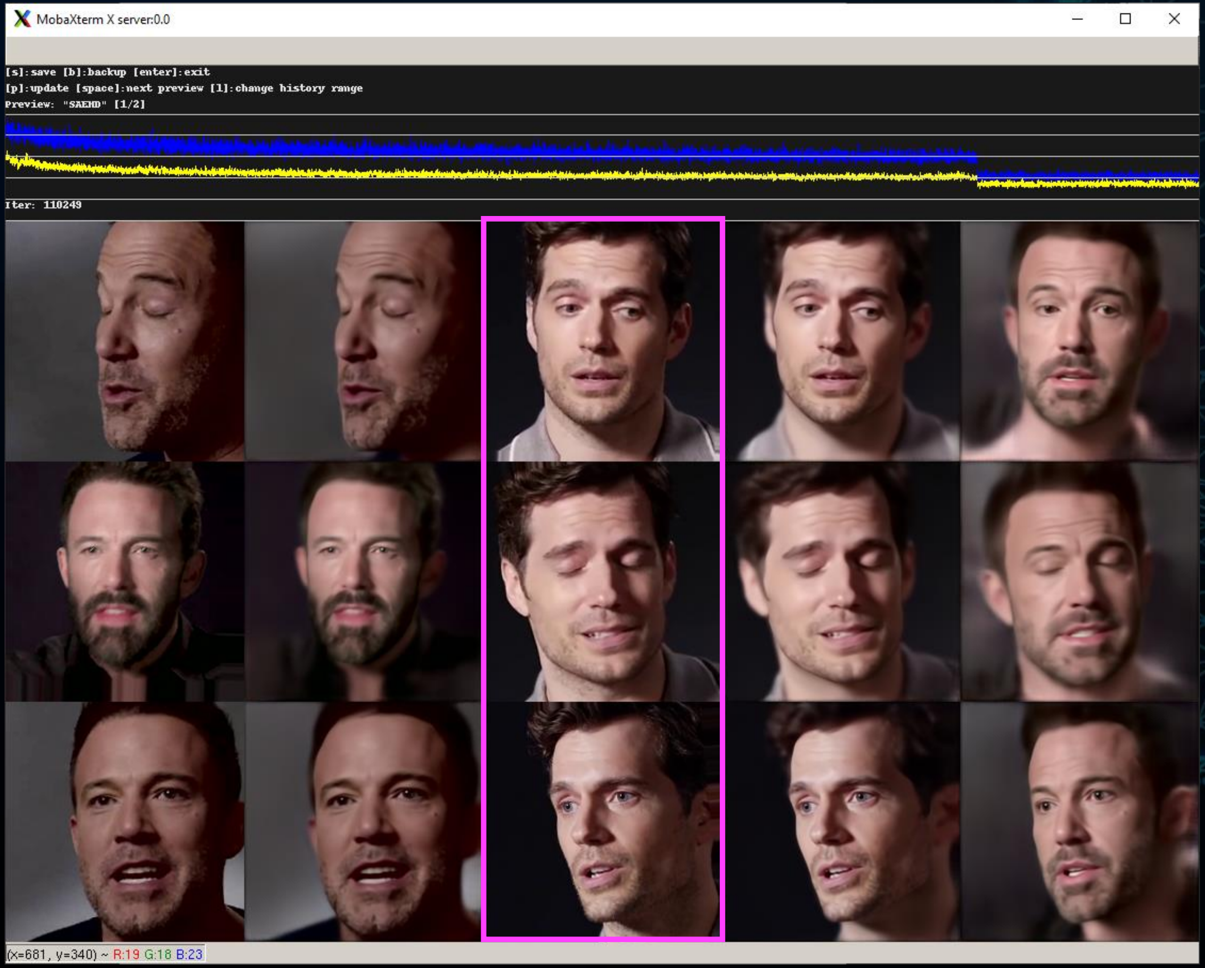
```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]

Iter: 110249
```

fake source



```
================= Model Summary ==================
==                                              ==
==            Model name: new_SAEHD             ==
==                                              ==
==     Current iteration: 108391                ==
==                                              ==
==--------------- Model Options ----------------==
==                                              ==
==            resolution: 224                   ==
==             face_type: wf                    ==
==     models_opt_on_gpu: True                  ==
==                 archi: liae-udt              ==
==               ae_dims: 512                   ==
==                e_dims: 64                    ==
==                d_dims: 64                    ==
==           d_mask_dims: 32                    ==
==       masked_training: True                  ==
==       eyes_mouth_prio: True                  ==
==           uniform_yaw: False                 ==
==         blur_out_mask: True                  ==
==             adabelief: True                  ==
==            lr_dropout: n                     ==
==           random_warp: False                 ==
==      random_hsv_power: 0.0                   ==
==       true_face_power: 0.0                   ==
==      face_style_power: 0.0                   ==
==        bg_style_power: 0.0                   ==
==               ct_mode: rct                   ==
==              clipgrad: False                 ==
==              pretrain: False                 ==
==       autobackup_hour: 8                     ==
== write_preview_history: True                  ==
==           target_iter: 0                     ==
==       random_src_flip: False                 ==
==       random_dst_flip: True                  ==
==            batch_size: 32                    ==
==             gan_power: 0.0                   ==
==        gan_patch_size: 28                    ==
==              gan_dims: 16                    ==
==                                              ==
==----------------- Running On -----------------==
==                                              ==
==          Device index: 1                     ==
==                  Name: Tesla V100S-PCIE-32GB ==
==                  VRAM: 22.61GB               ==
==                                              ==
==================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

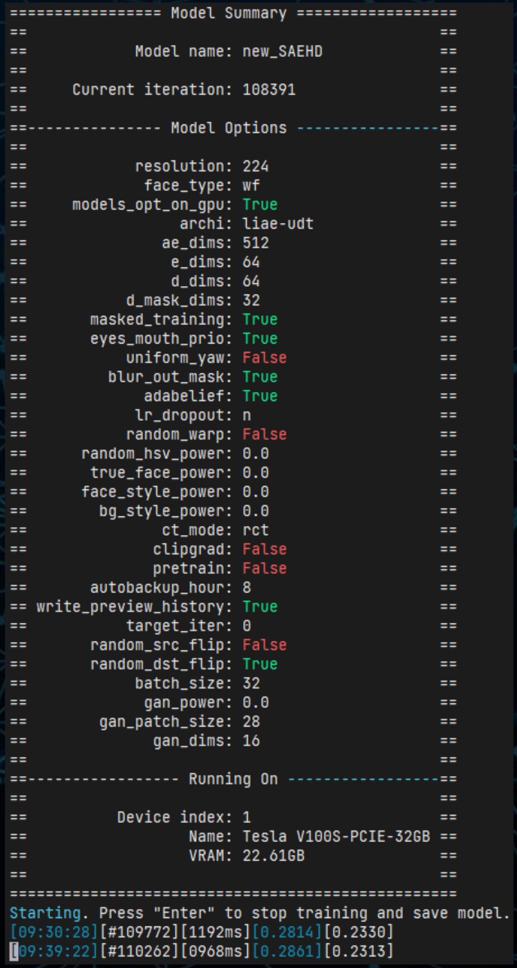
MODEL TRAINING



```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]

Iter: 110249
```

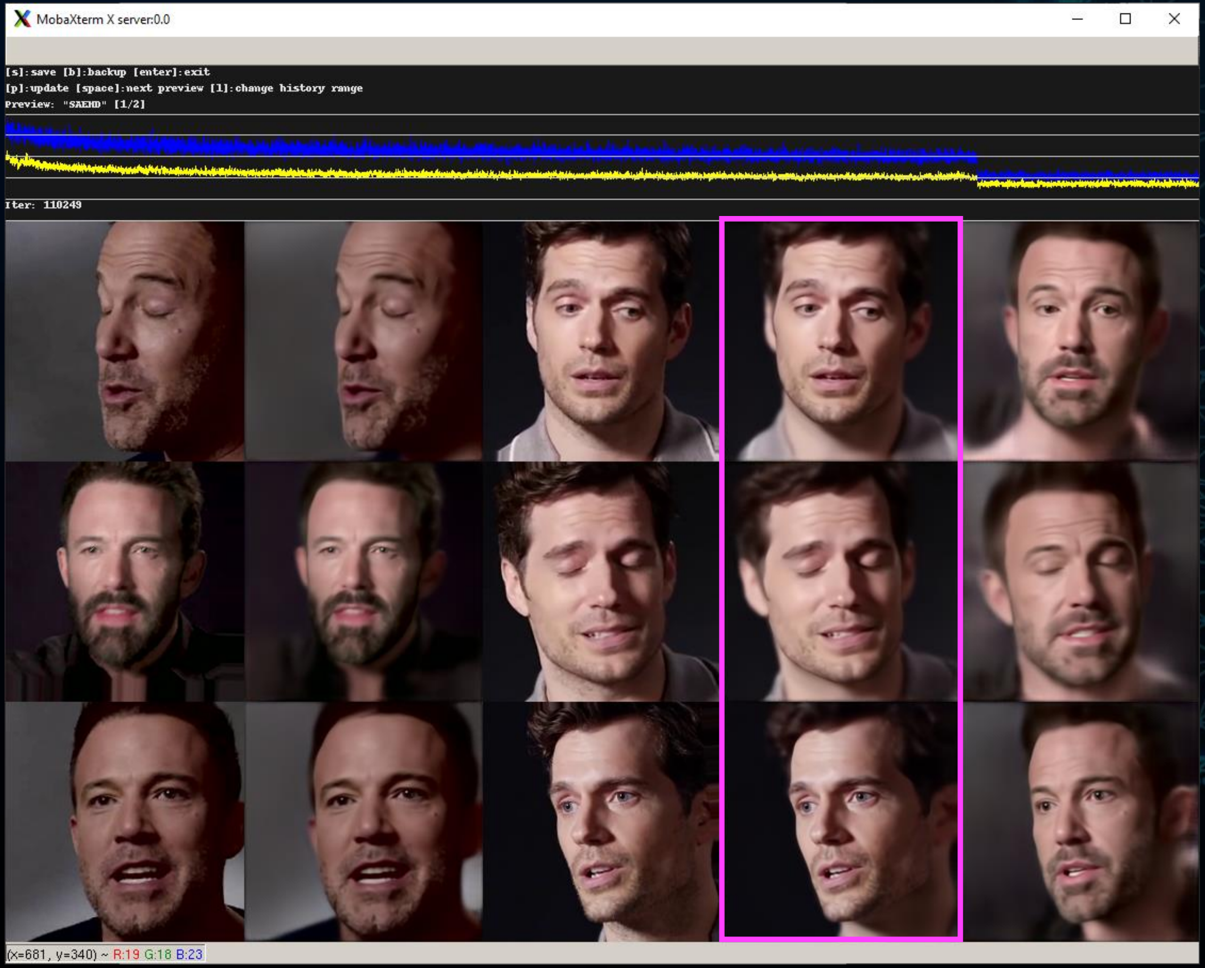
```
(x=681, y=340) ~ R:19 G:18 B:23
```

target



```
================= Model Summary ==================
==                                              ==
==           Model name: new_SAEHD              ==
==                                              ==
==    Current iteration: 108391                 ==
==                                              ==
==--------------- Model Options ----------------==
==                                              ==
==           resolution: 224                    ==
==            face_type: wf                     ==
==    models_opt_on_gpu: True                   ==
==                archi: liae-udt               ==
==              ae_dims: 512                    ==
==               e_dims: 64                     ==
==               d_dims: 64                     ==
==          d_mask_dims: 32                     ==
==      masked_training: True                   ==
==      eyes_mouth_prio: True                   ==
==          uniform_yaw: False                  ==
==        blur_out_mask: True                   ==
==            adabelief: True                   ==
==           lr_dropout: n                      ==
==          random_warp: False                  ==
==     random_hsv_power: 0.0                    ==
==      true_face_power: 0.0                    ==
==     face_style_power: 0.0                    ==
==       bg_style_power: 0.0                    ==
==              ct_mode: rct                    ==
==             clipgrad: False                  ==
==             pretrain: False                  ==
==      autobackup_hour: 8                      ==
== write_preview_history: True                  ==
==          target_iter: 0                      ==
==      random_src_flip: False                  ==
==      random_dst_flip: True                   ==
==           batch_size: 32                     ==
==            gan_power: 0.0                    ==
==       gan_patch_size: 28                     ==
==             gan_dims: 16                     ==
==                                              ==
==----------------- Running On -----------------==
==                                              ==
==         Device index: 1                      ==
==                 Name: Tesla V100S-PCIE-32GB ==
==                 VRAM: 22.61GB                ==
==                                              ==
==================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

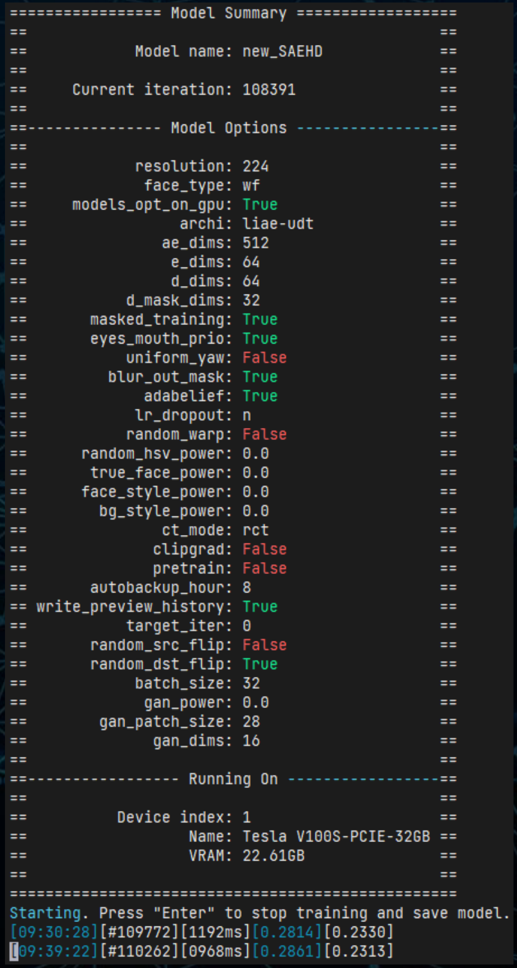
MODEL TRAINING



```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]
Iter: 110249
```

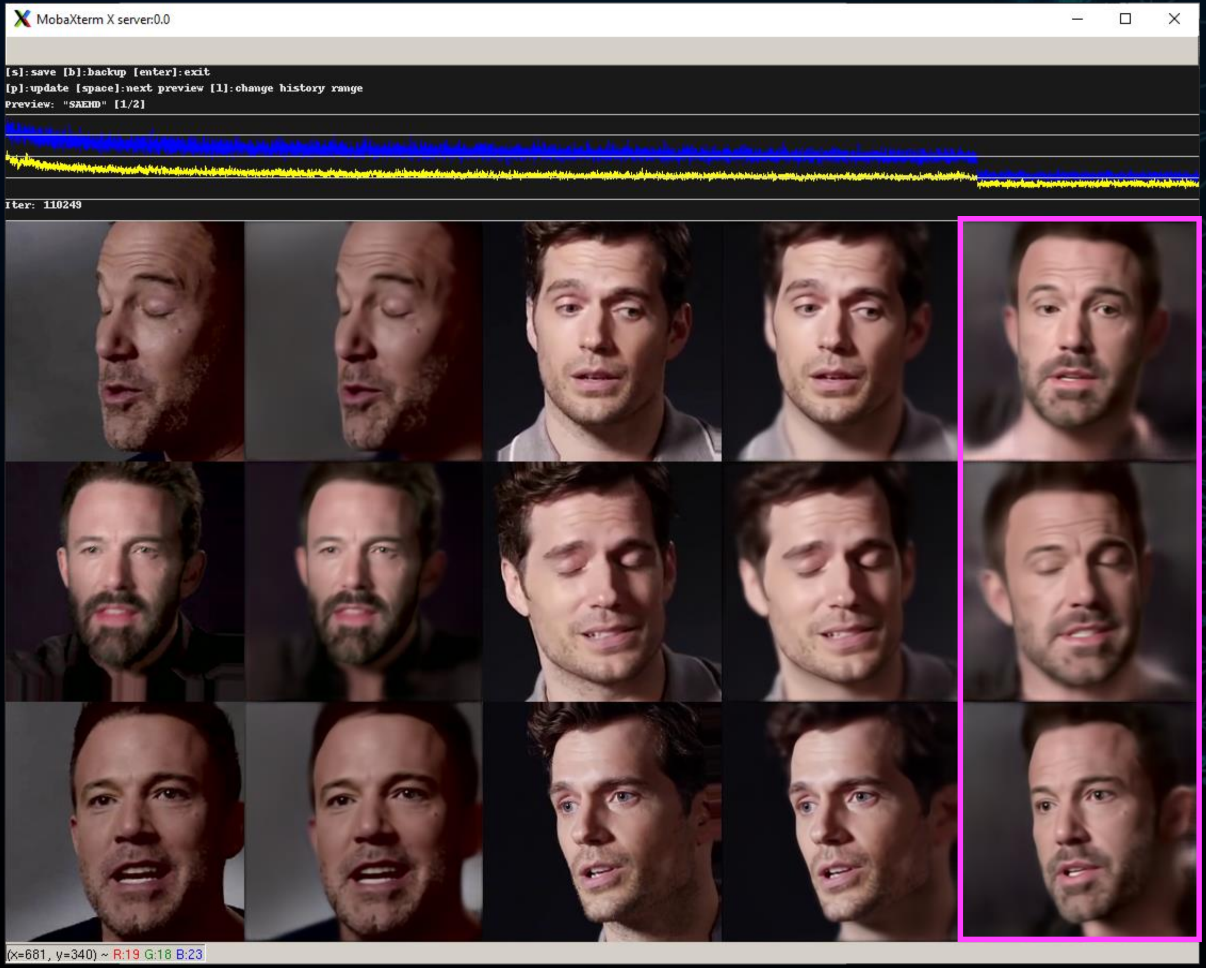
```
(x=681, y=340) ~ R:19 G:18 B:23
```

fake target



```
================ Model Summary =================
==                                            ==
==           Model name: new_SAEHD            ==
==                                            ==
==    Current iteration: 108391               ==
==                                            ==
==-------------- Model Options ---------------==
==                                            ==
==           resolution: 224                  ==
==            face_type: wf                   ==
==     models_opt_on_gpu: True                ==
==                 archi: liae-udt            ==
==               ae_dims: 512                 ==
==                e_dims: 64                  ==
==                d_dims: 64                  ==
==           d_mask_dims: 32                  ==
==       masked_training: True                ==
==       eyes_mouth_prio: True                ==
==           uniform_yaw: False               ==
==         blur_out_mask: True                ==
==             adabelief: True                ==
==            lr_dropout: n                   ==
==           random_warp: False               ==
==      random_hsv_power: 0.0                 ==
==       true_face_power: 0.0                 ==
==      face_style_power: 0.0                 ==
==        bg_style_power: 0.0                 ==
==               ct_mode: rct                 ==
==              clipgrad: False               ==
==              pretrain: False               ==
==       autobackup_hour: 8                   ==
== write_preview_history: True                ==
==           target_iter: 0                   ==
==       random_src_flip: False               ==
==       random_dst_flip: True                ==
==            batch_size: 32                  ==
==             gan_power: 0.0                 ==
==        gan_patch_size: 28                  ==
==              gan_dims: 16                  ==
==                                            ==
==---------------- Running On ----------------==
==                                            ==
==         Device index: 1                    ==
==                 Name: Tesla V100S-PCIE-32GB ==
==                 VRAM: 22.61GB              ==
==                                            ==
================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

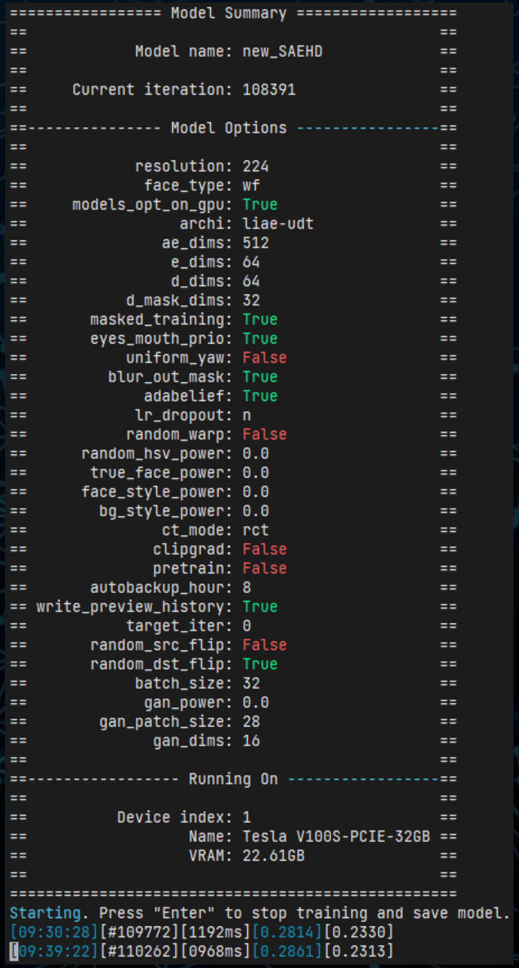
MODEL TRAINING



```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "SAEHD" [1/2]
Iter: 110249
```

(x=681, y=340) ~ R:19 G:18 B:23

fake source + target



```
================ Model Summary =================
==                                            ==
==           Model name: new_SAEHD            ==
==                                            ==
==    Current iteration: 108391               ==
==                                            ==
==--------------- Model Options --------------==
==                                            ==
==           resolution: 224                  ==
==            face_type: wf                   ==
==     models_opt_on_gpu: True                ==
==                archi: liae-udt             ==
==              ae_dims: 512                  ==
==               e_dims: 64                   ==
==               d_dims: 64                   ==
==          d_mask_dims: 32                   ==
==       masked_training: True                ==
==       eyes_mouth_prio: True                ==
==          uniform_yaw: False                ==
==        blur_out_mask: True                 ==
==            adabelief: True                 ==
==           lr_dropout: n                    ==
==          random_warp: False                ==
==      random_hsv_power: 0.0                 ==
==       true_face_power: 0.0                 ==
==      face_style_power: 0.0                 ==
==        bg_style_power: 0.0                 ==
==              ct_mode: rct                  ==
==             clipgrad: False                ==
==             pretrain: False                ==
==       autobackup_hour: 8                   ==
== write_preview_history: True                ==
==          target_iter: 0                    ==
==       random_src_flip: False               ==
==       random_dst_flip: True                ==
==           batch_size: 32                   ==
==            gan_power: 0.0                  ==
==        gan_patch_size: 28                  ==
==             gan_dims: 16                   ==
==                                            ==
==----------------- Running On ---------------==
==                                            ==
==         Device index: 1                    ==
==                 Name: Tesla V100S-PCIE-32GB ==
==                 VRAM: 22.61GB              ==
==                                            ==
================================================
Starting. Press "Enter" to stop training and save model.
[09:30:28][#109772][1192ms][0.2814][0.2330]
[09:39:22][#110262][0968ms][0.2861][0.2313]
```

MODEL TRAINING

Let the model recognize generic human faces first, then specialize on our source & target's faces

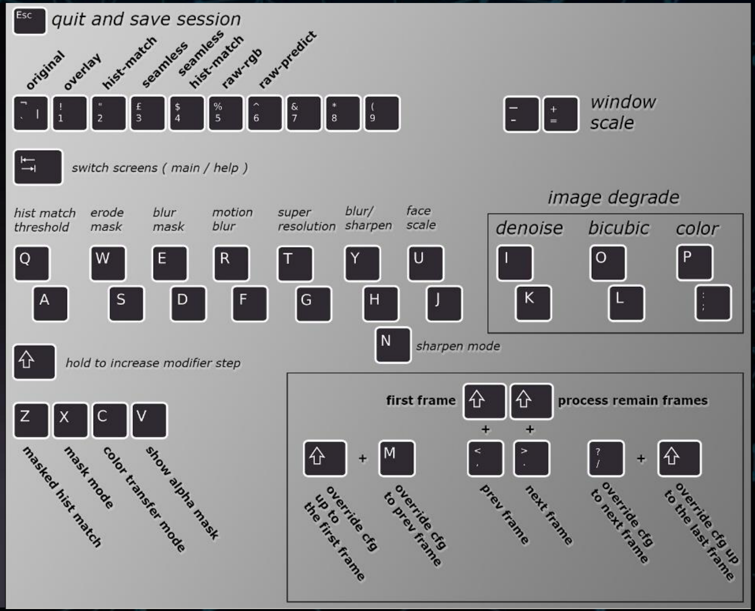
Hyperparameters:

Realistic aspect	Imaginative aspect
Eyes & mouth priority	GAN power
Uniform yaw	Face style power
Learning rate dropout	Background style power
Random warp	
Masked training	

MERGING

Use our trained model to transform face from target to source

Requires manual fine-tune



DEEPFAKE DETECTION

Essentially an evaluation on a DeepFake's production quality

Cryptography concept: look for easy-to-evaluate but hard-to-forge features

Examples:

- Frame level smoothness
- Unnatural movement/expression
- Inconsistencies near masking area
- Reflection in the eyes
- Phoneme-viseme mismatches

STATE-OF-THE-ART DETECTION

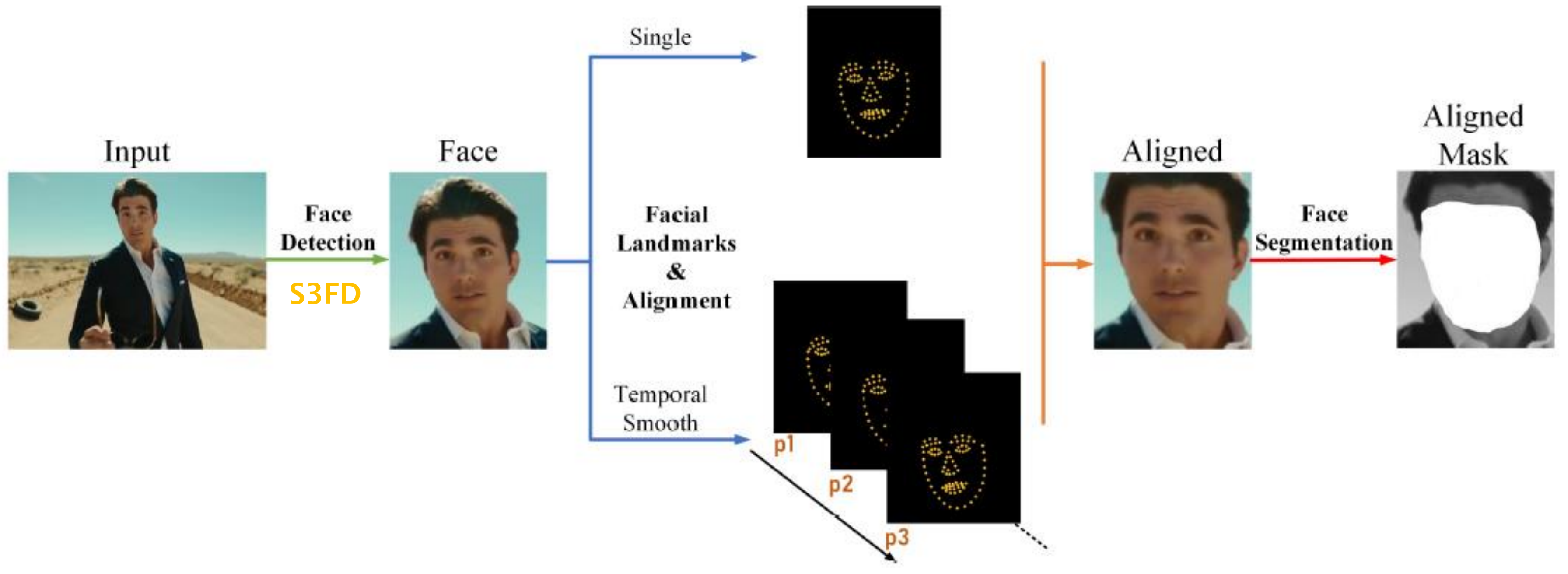
Multimodal approach: predict the authenticity of video, audio, lip sync (video + audio)

Hashing approach: check against hashes of known examples of in-the-wild DeepFakes

THANK YOU

DFL: FACE EXTRACTION

(a) Heatmap-based facial landmark algorithm 2DFAN (for faces with standard pose)
(b) PRNet with 3D face prior information (for faces with large Euler angle – yaw, pitch, roll)



S3FD

Figure 2. Overview of extraction phase in DeepFaceLab (DFL for short).

DFL: MODEL STRUCTURE

Loss:

(1) DSSIM (structural dissimilarity): faster face generalization

(2) MSE: better clarity

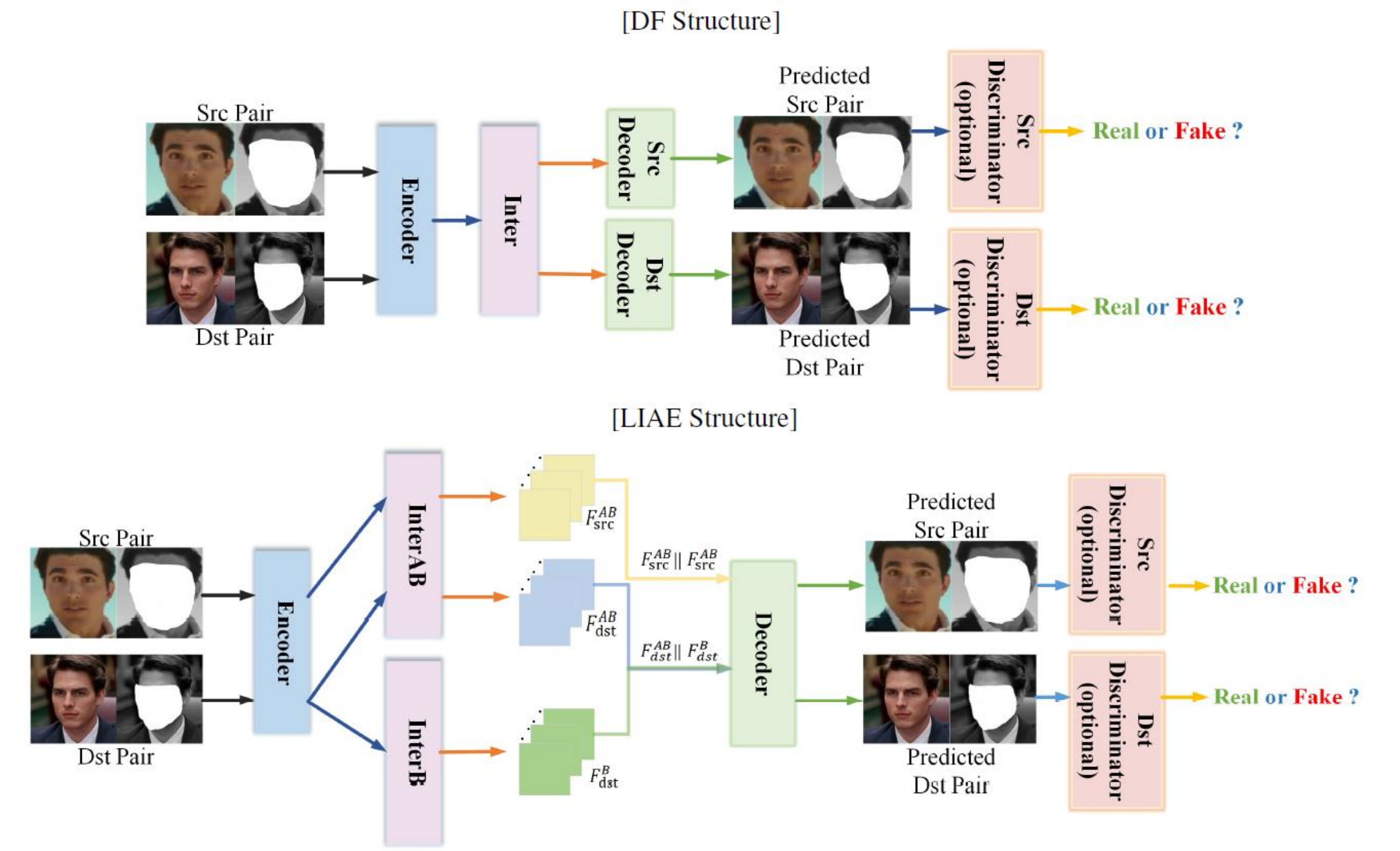


Figure 3. Overview of training phase in DeepFaceLab (DFL). DF structure and LIAE structure are both provided here for illustration, ○||○ represents the concatenation of latent vectors.

DFL: MERGING

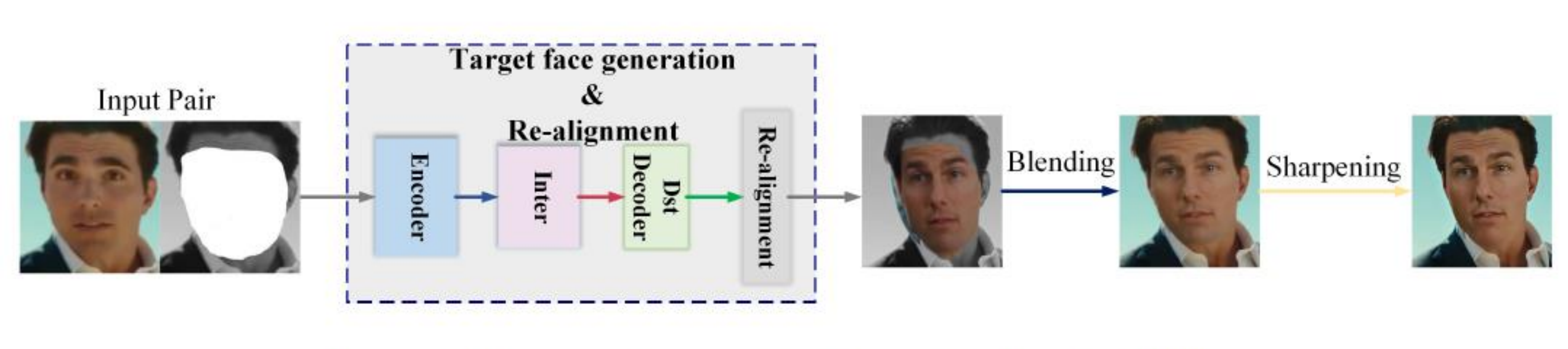


Figure 4. Overview of conversion phase in DeepFaceLab(DFL).