AI-Synthesized Media and How to Detect Them

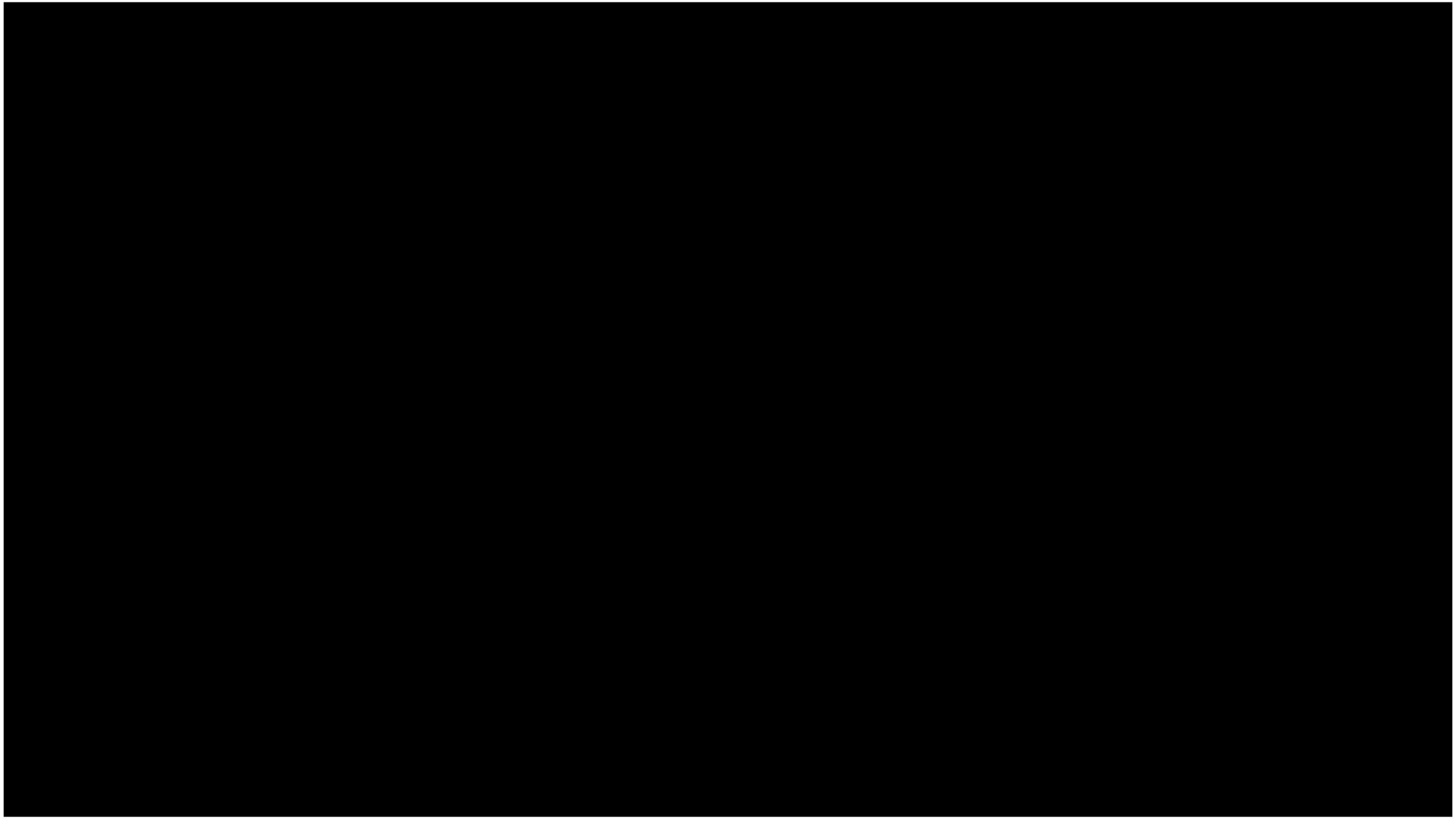
Ziwei Liu

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

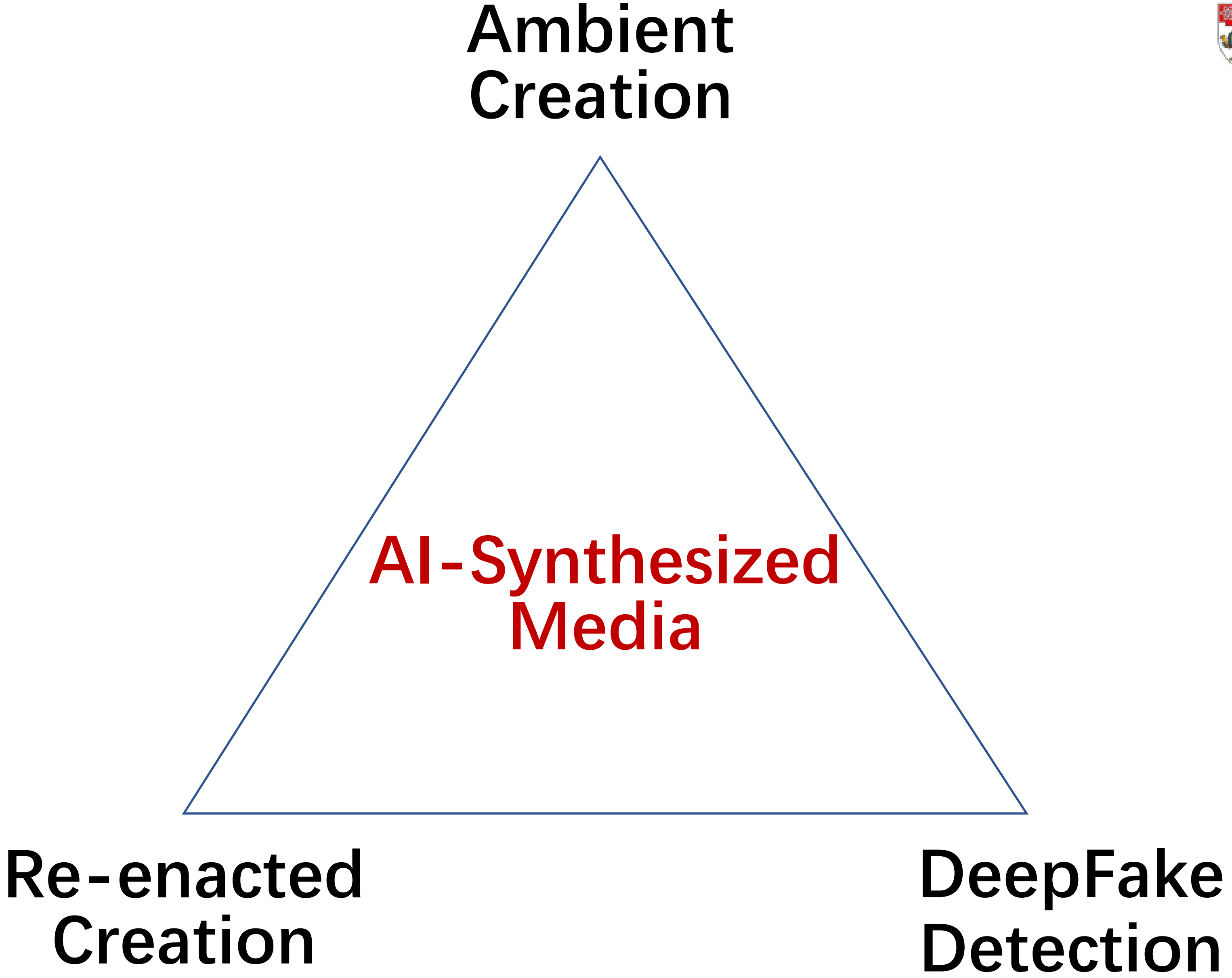
Visual Illusion

AI is Good at Creating Illusioins



AI is Good at Creating Illusioins

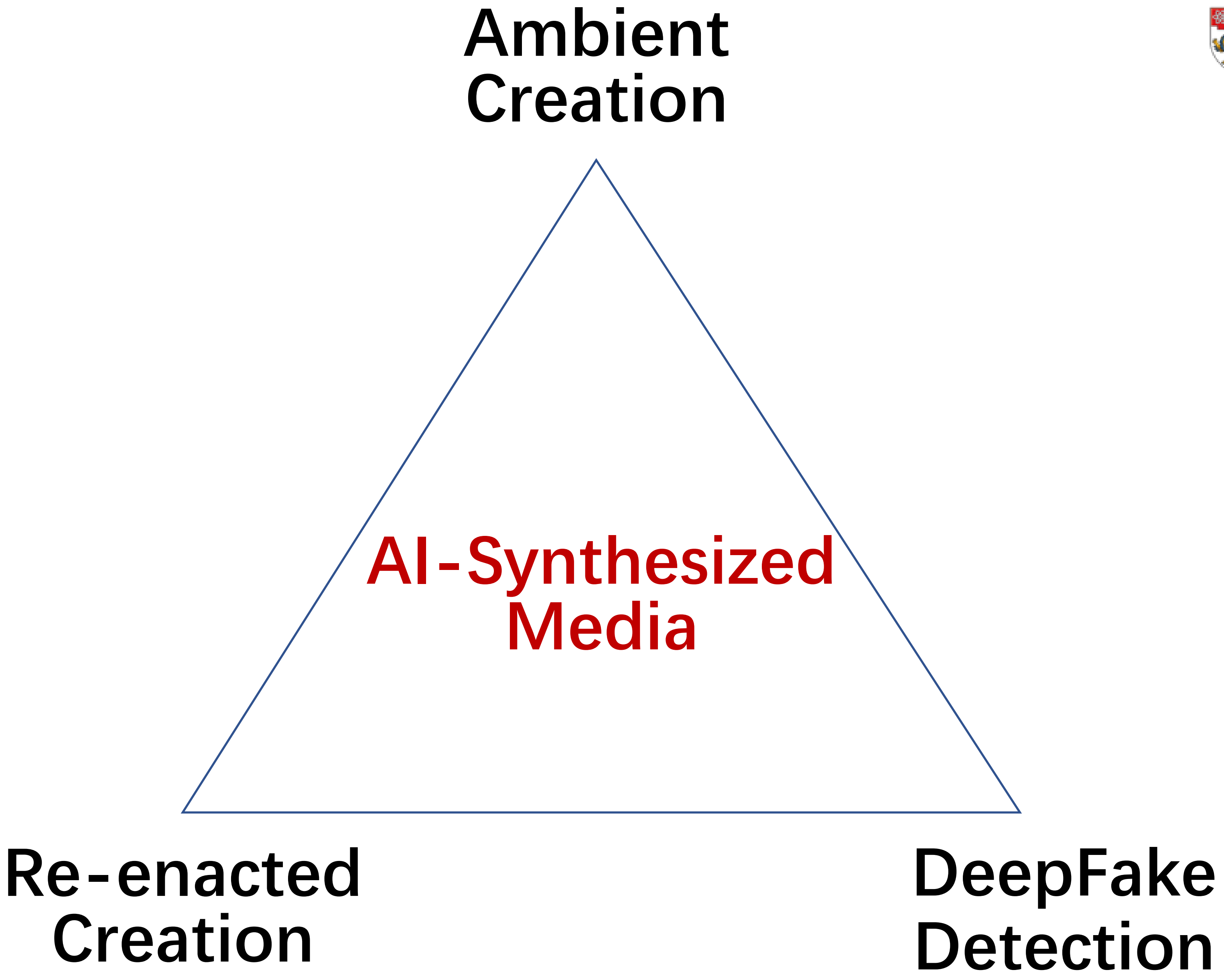

We are in the Metaverse


Ambient
Creation
AI-Synthesized
Media
Re-enacted
Creation
DeepFake
Detection

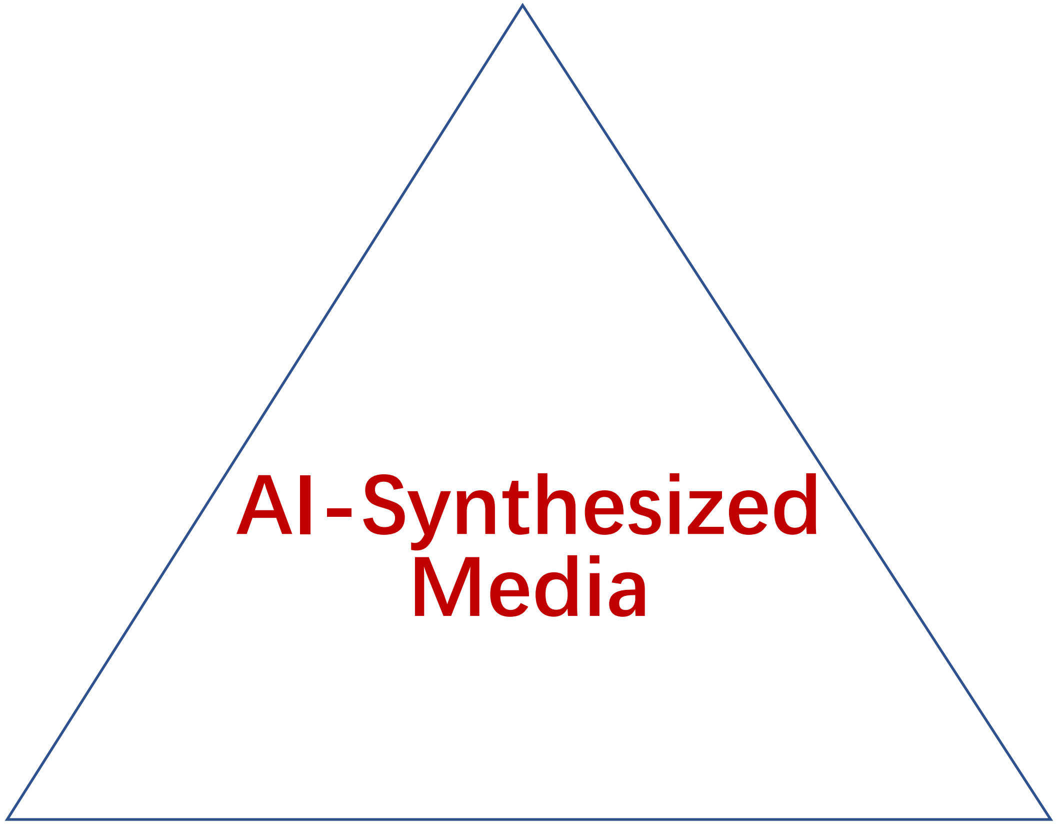
Ambient Creation

Re-enacted Creation

DeepFake Detection


Ambient Creation
AI-Synthesized Media
Re-enacted Creation
DeepFake Detection

Ambient Creation

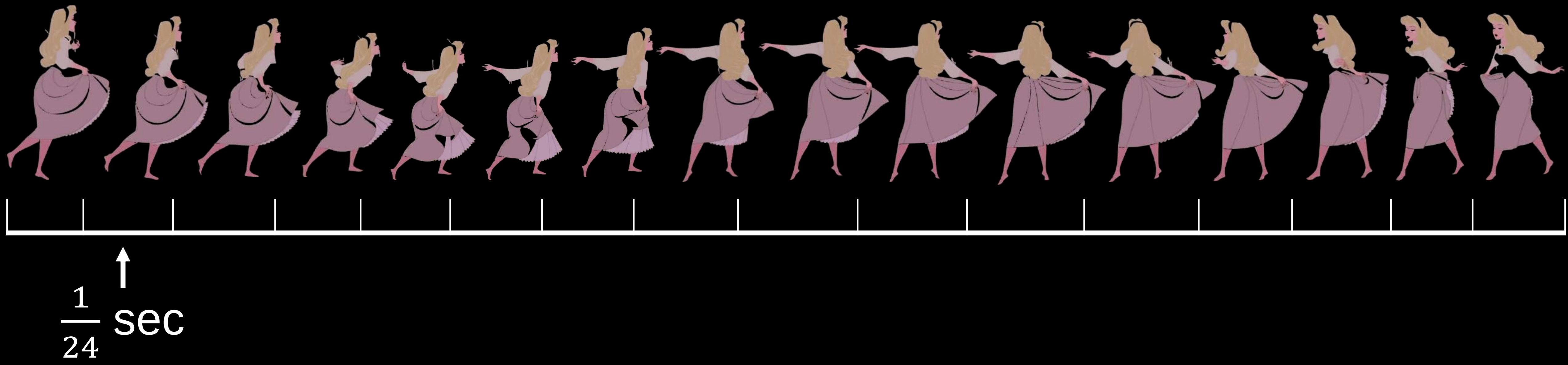


Deep Animation Video Interpolation in the Wild

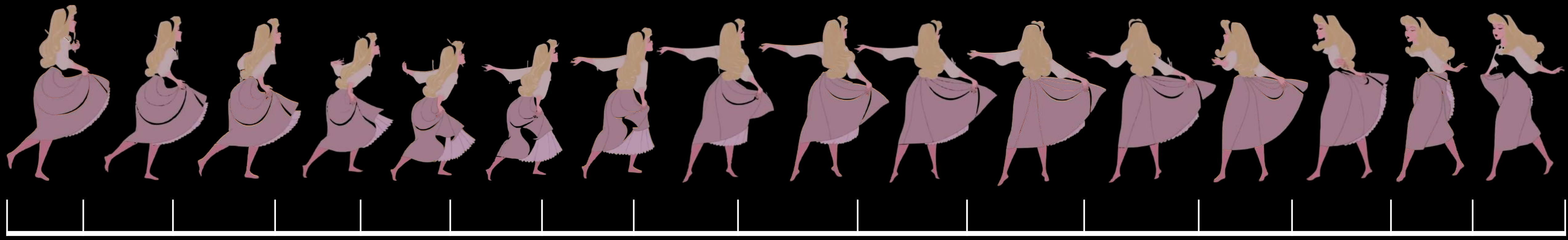
Li Siyao*, Shiyu Zhao*, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, Ziwei Liu

SenseTime Research, Rutgers University, Sun Yat-sen University, Shanghai AI lab, Nanyang Technological University

¥ ¥ ¥



full frame rate 24 fps

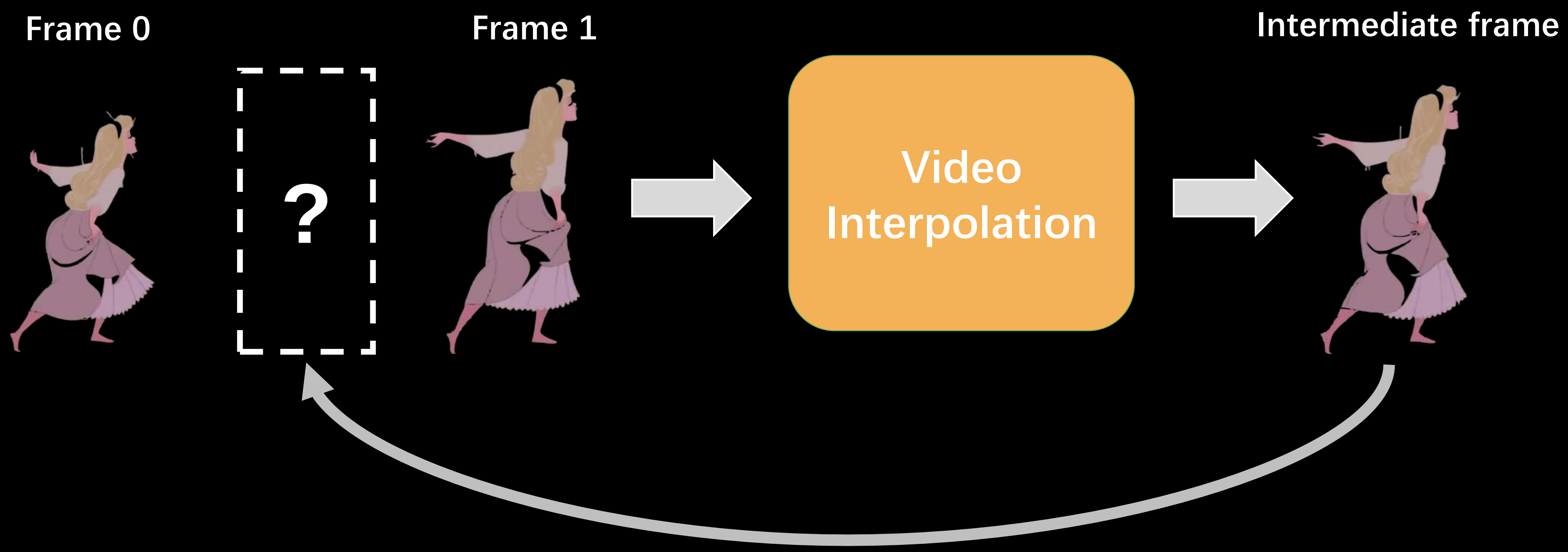


“on twos” 24 fps → 12 fps

“on threes” 24 fps → 8 fps

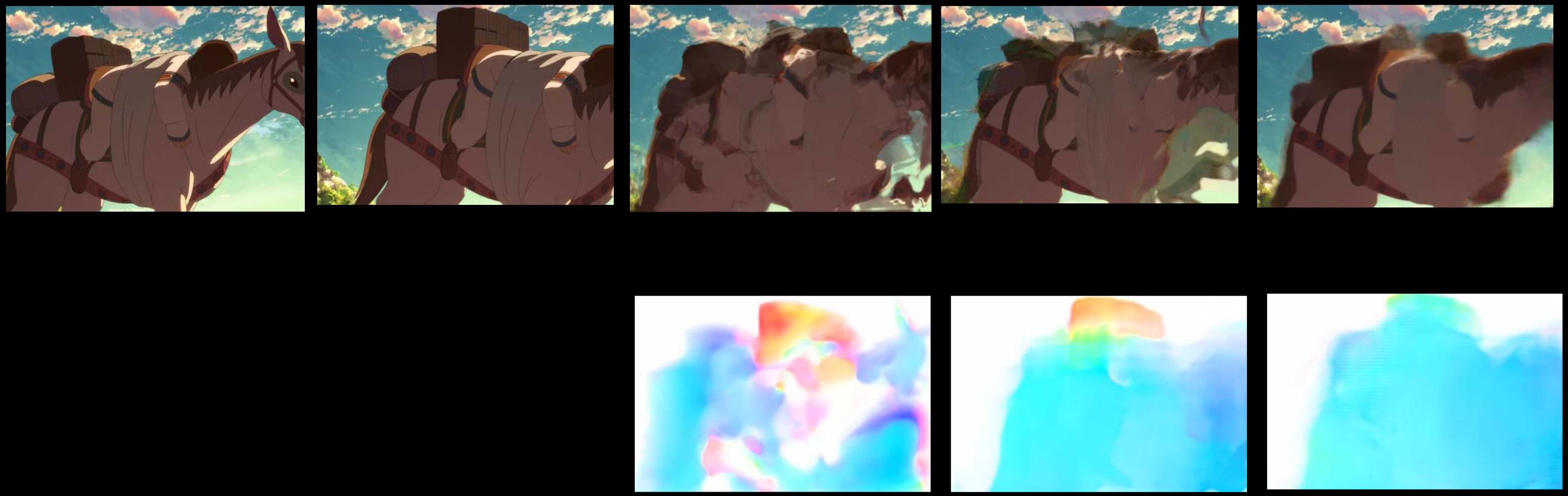
24 fps

8 fps


Frame 0
?
Frame 1
Video Interpolation
Intermediate frame

Problems

- 1. Existing methods do not perform well on animation
- 2. No animation dataset for training/testing of video interpolation



Animation Triplet Dataset (ATD-12K)

Training set 10K

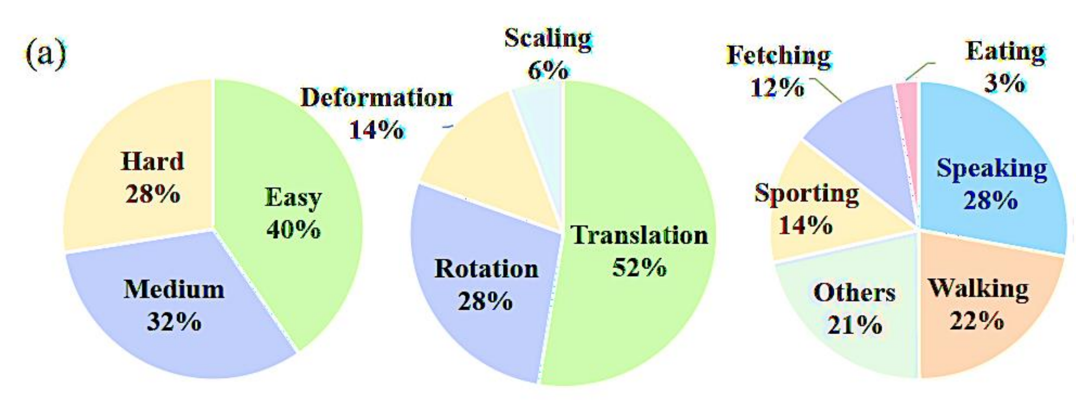
Test set 2K

Rich Annotations:

- Difficulty level
- Movement tags
- Salient Motion Region

Rich annotations

- Hardness level
- Motion type
- Movement categories
- ROI for salient movement



Difficulties on animation video interpolation

- Animations are made of color pieces and lack of texture

- Motion between anime frames are non-linear and extremely large

Segment-Guided Matching

image

Contour

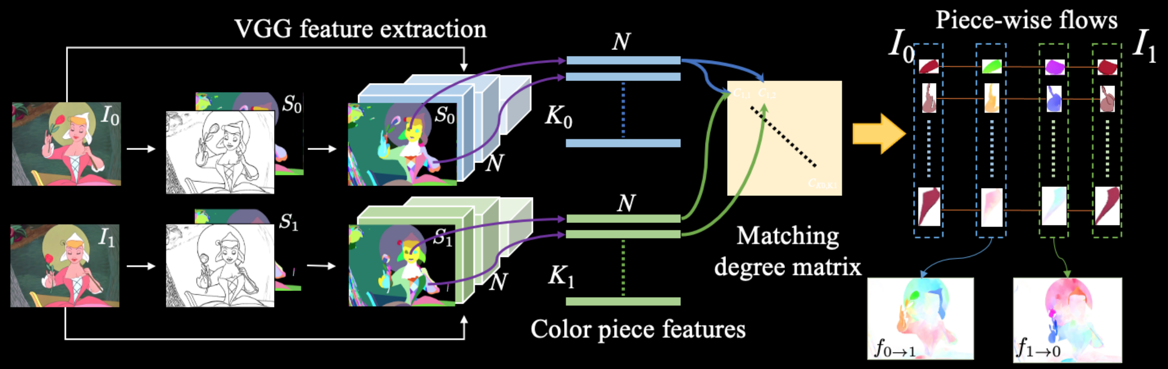
Segmentation

AnimeInterp

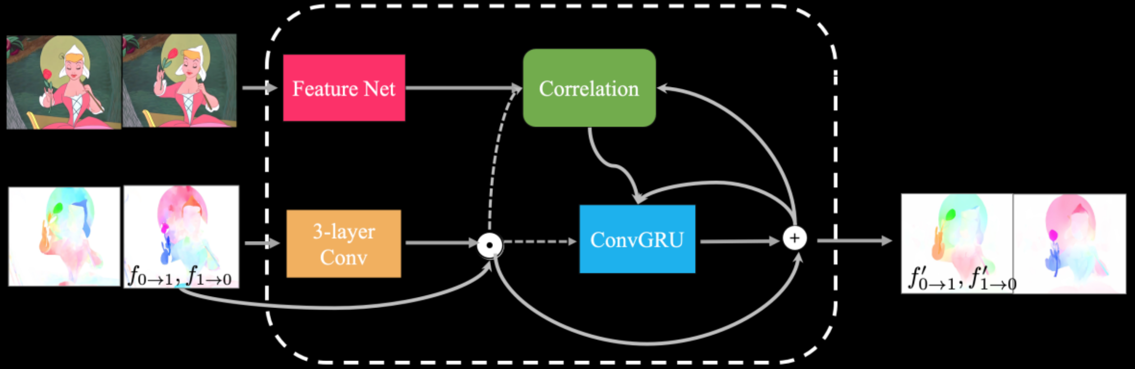
Segment-Guided Matching

Recurrent Flow Refinement

Warping and Synthesis



SGM computes coarse piece-wise flows



RFR refines pixel-wise flows

Experimental results

Table 1: **Quantitative results on the test set of ATD-12K.** The best and runner-up values are bold and underlined, respectively.

	Whole		RoI		Easy		Medium		Hard	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Super SloMo w/o. ft.	27.88	0.946	24.56	0.886	30.66	0.966	27.29	0.948	24.63	0.917
Super SloMo [9]	28.19	0.949	24.83	0.892	30.86	0.967	27.63	0.950	25.02	0.922
DAIN w/o. ft.	28.84	0.953	25.43	0.897	31.40	<u>0.970</u>	28.38	0.955	25.77	0.927
DAIN [1]	29.19	0.956	25.78	0.902	31.67	**0.971**	28.74	0.957	26.22	0.932
QVI w/o. ft.	28.80	0.953	25.54	0.900	31.14	0.969	28.44	0.955	25.93	0.929
QVI [33]	29.04	0.955	25.65	0.901	31.46	<u>0.970</u>	28.63	0.956	26.11	0.931
AdaCoF w/o. ft.	28.10	0.947	24.72	0.886	31.09	0.968	27.43	0.948	24.65	0.916
AdaCoF [12]	28.29	0.951	24.89	0.894	31.10	0.969	27.63	0.951	25.10	0.925
SoftSplat w/o. ft.	29.15	0.955	25.75	0.904	31.50	<u>0.970</u>	28.75	0.956	26.29	0.934
SoftSplat [18]	29.34	<u>0.957</u>	25.95	<u>0.907</u>	31.60	<u>0.970</u>	28.96	<u>0.958</u>	26.59	<u>0.938</u>
Ours w/o. SGM	<u>29.54</u>	**0.958**	<u>26.15</u>	**0.910**	<u>31.80</u>	**0.971**	<u>29.15</u>	**0.959**	<u>26.78</u>	**0.939**
Ours w/o. RFR	27.62	0.944	24.43	0.887	29.78	0.959	27.29	0.946	24.94	0.920
Ours	**29.68**	**0.958**	**26.27**	**0.910**	**31.86**	**0.971**	**29.26**	**0.959**	**27.07**	**0.939**

original

interpolated

original

interpolated

original

interpolated

original

interpolated

original

interpolated

original

interpolated

original

interpolated

x8 slower

original

interpolated


Super SloMo

Ours

DAIN

Ours

SoftSplat

Ours

New task

Study animation VI for the first time

New dataset

A large-scale dataset for training and test

New method

An animation-specific model making progress in this task

AI-Synthesized
Media

Re-enacted
Creation

CVPR VIRTUAL JUNE 19-25

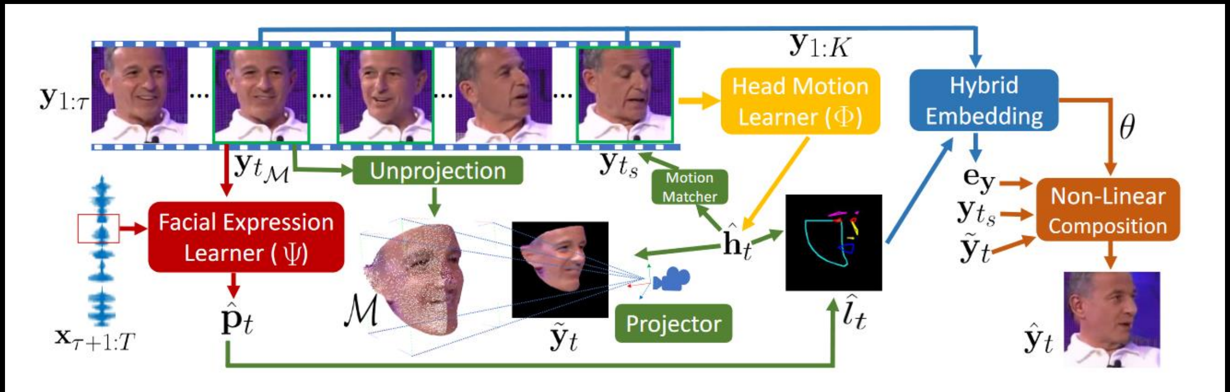
Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation

Hang Zhou,[1] Yasheng Sun,[2, 4] Wayne Wu,[3, 4] Chen Change Loy,[3] Xiaogang Wang,[1] and Ziwei Liu[3]

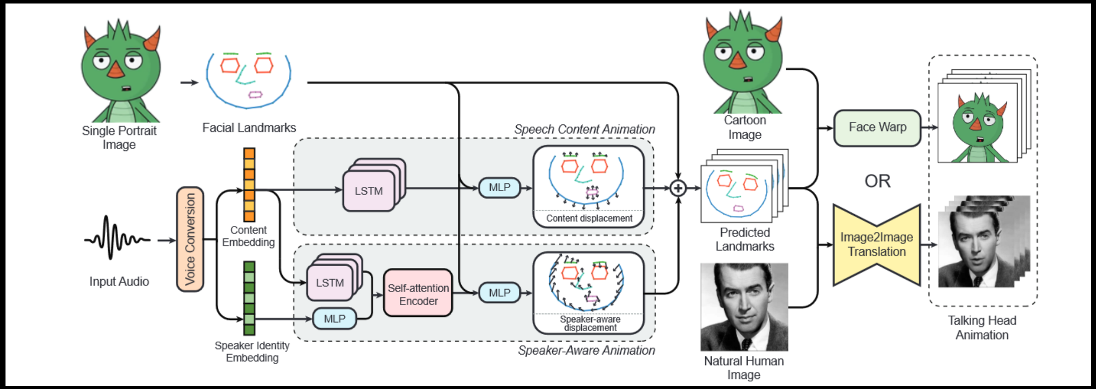
1. The Chinese University of Hong Kong
2. Tokyo Institute of Technology
3. Nanyang Technological University
4. SenseTime Research

Previous Methods

- Rely on intermediate representations (2D/3D landmarks, 3D face reconstruction). These representations are not accurate under extreme cases.
- Pure reconstruction-based methods by latent feature learning cannot change pose.
- No method has shown the results of free pose control with large views in this area.



Talking-head Generation with Rhythmic Head Motion. (ECCV 2020)



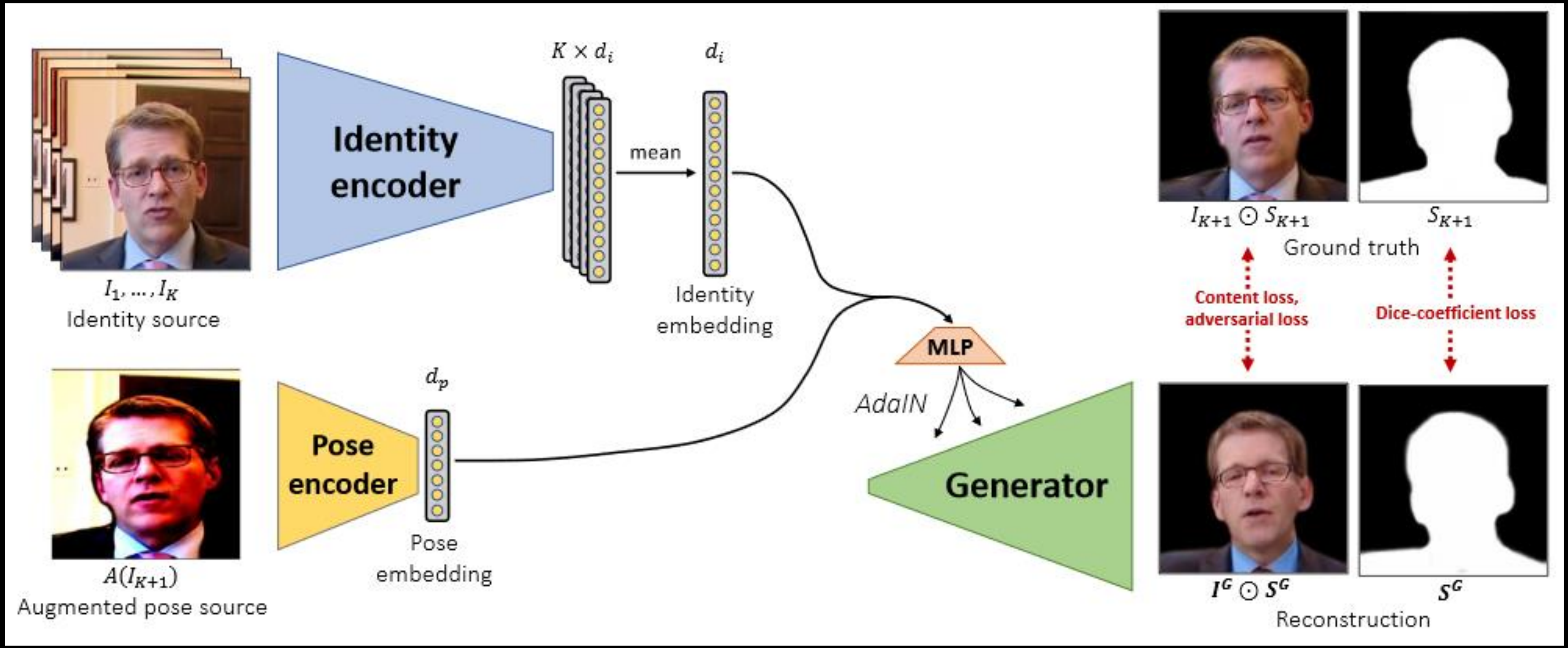
MakeItTalk: Speaker-Aware Talking-Head Animation (TOG 2020)

Core Ideas

- Without structural intermediate representation.
- Identify a non-identity space with data augmentation.
- Leverage contrastive audio-visual learning for lip sync.
- Devise an implicit pose code using 3D prior.
- Style-based generator for information balancing.

Inspiration: Face Reenactment

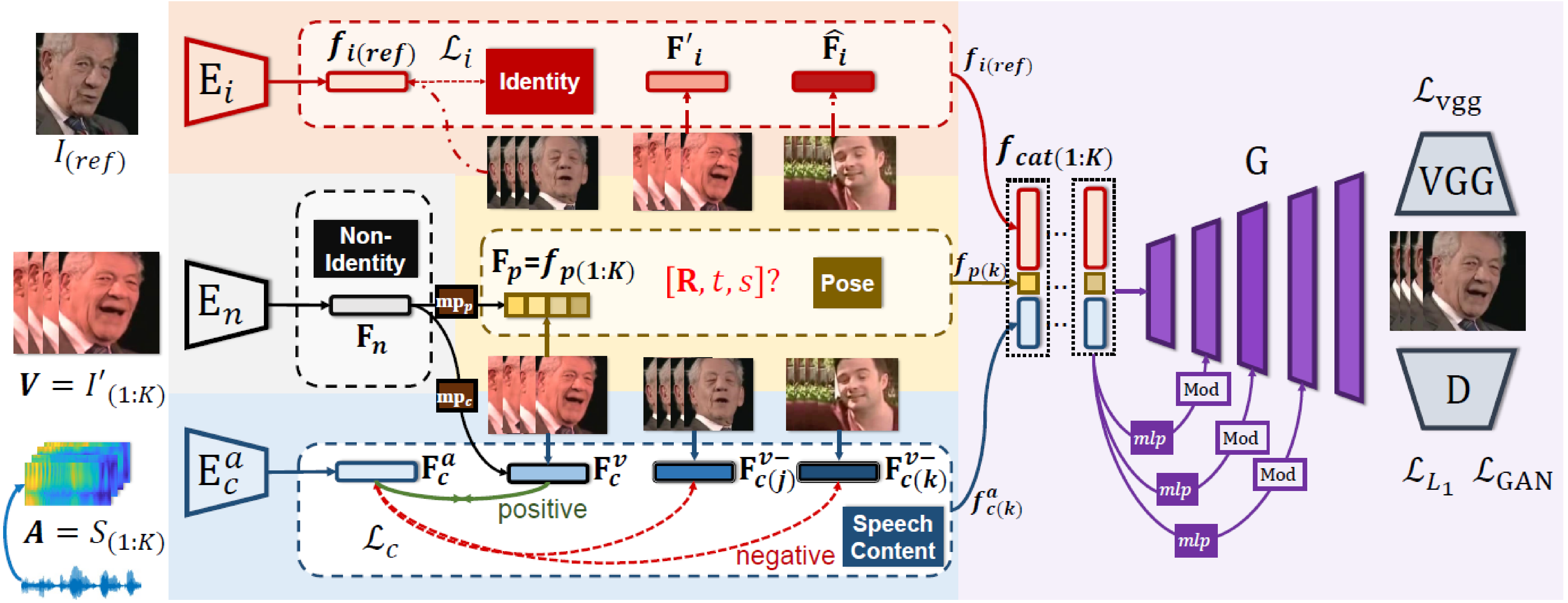
- Identity information can be repelled by frame augmentation.
- Style-based generator can automatically balance identity and identity-irrelevant information.



Neural head reenactment with latent pose descriptors. (CVPR 2020)

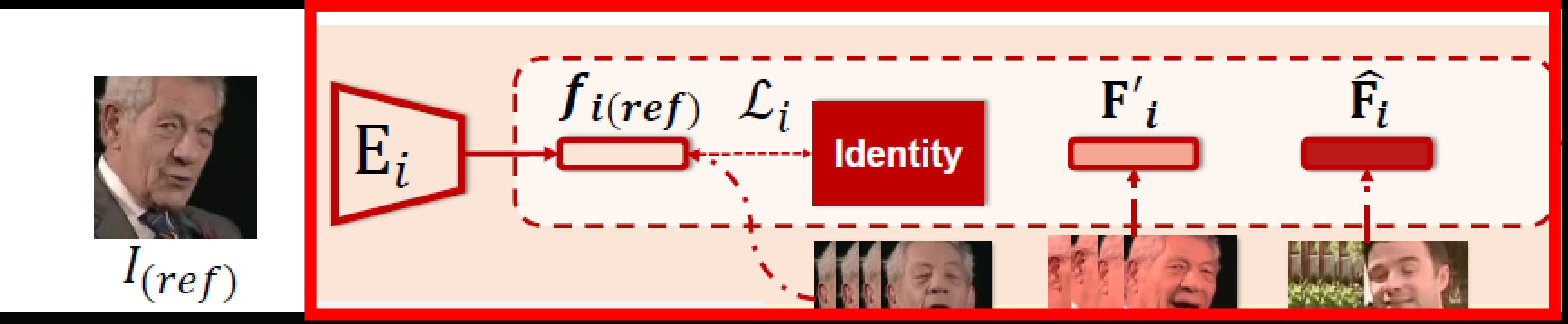
Pipeline: Pose-Controllable Audio-Visual System

- Modularize 3 spaces, including identity, speech content and pose.



Pipeline

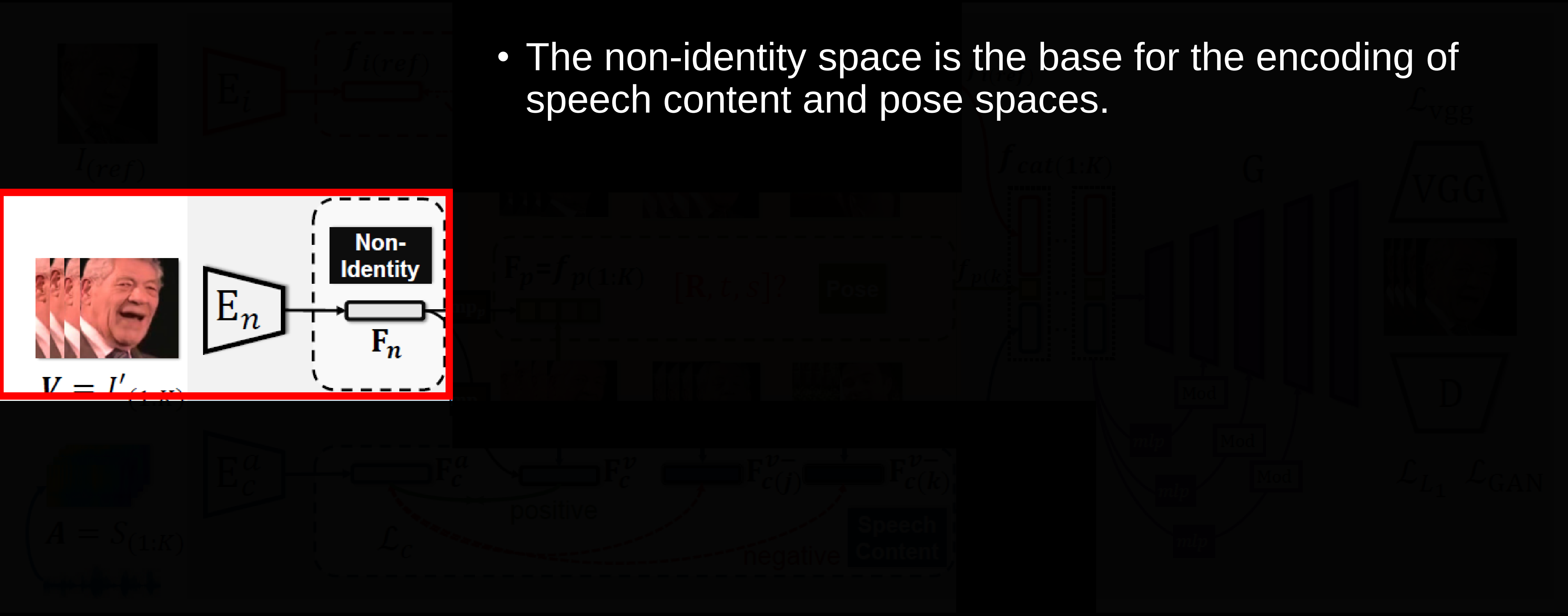
- Identity space encoding



  - Identity space can be easily encoded with ID supervision.

Pipeline

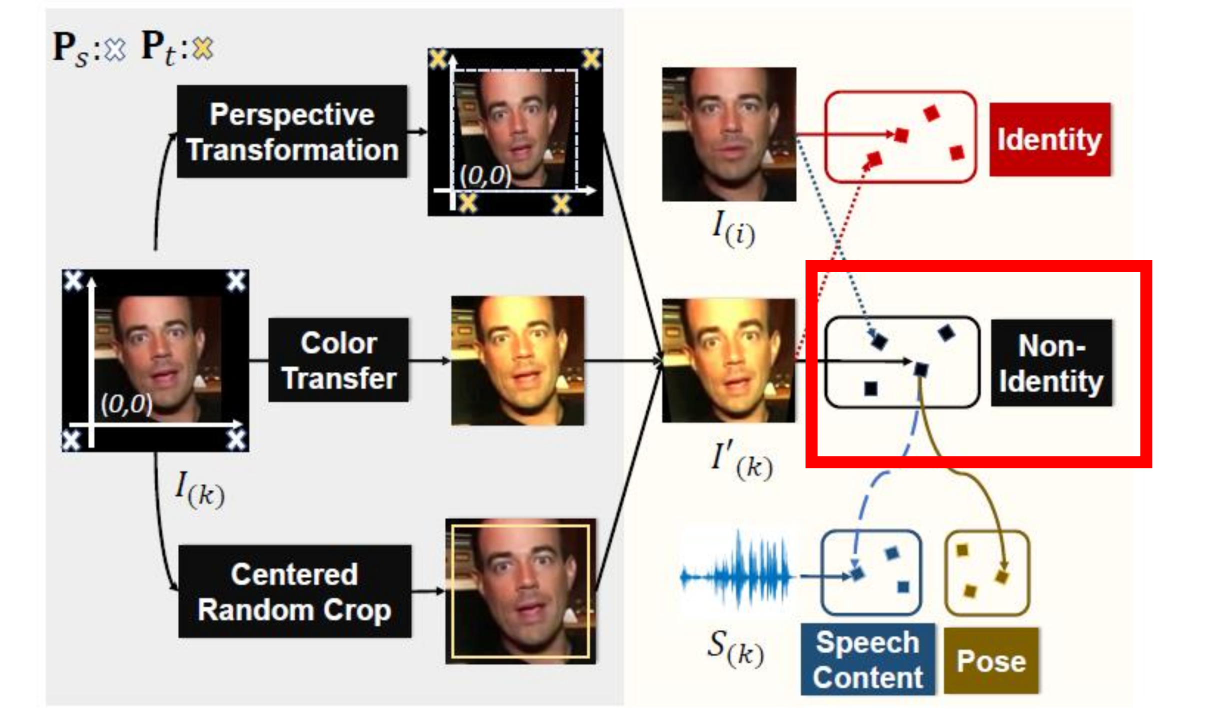
- Encode non-identity space.



- The non-identity space is the base for the encoding of speech content and pose spaces.

Non-Identity Space Encoding

- Target Frame Augmentation.
  - Perspective transform for shape.
  - Color transfer for texture.
  - Centered crop for scale shift.

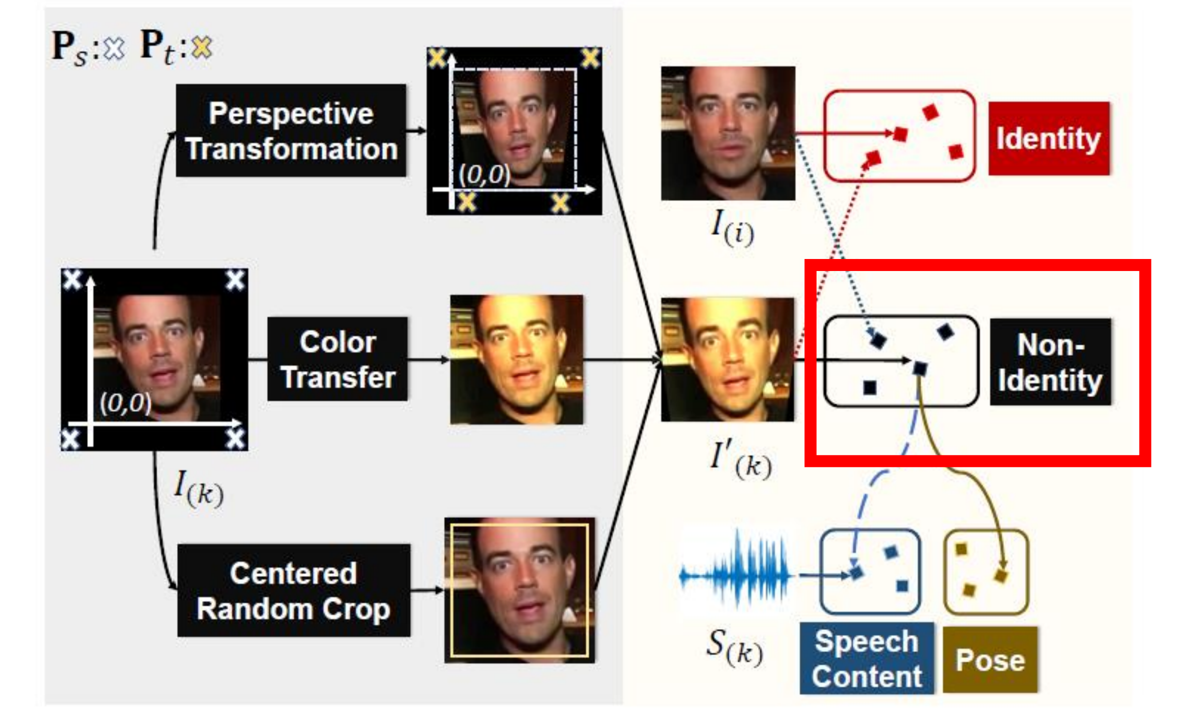


(1) Target Frame Augmentation (2) Feature Space Encoding

- Intuition: Source for Desired Information
  - Speech content and pose information should originate from this latent space.

Non-Identity Space Encoding

- Target Frame Augmentation.
  - Perspective transform for shape.
  - Color transfer for texture.
  - Centered crop for scale shift.



(1) Target Frame Augmentation (2) Feature Space Encoding

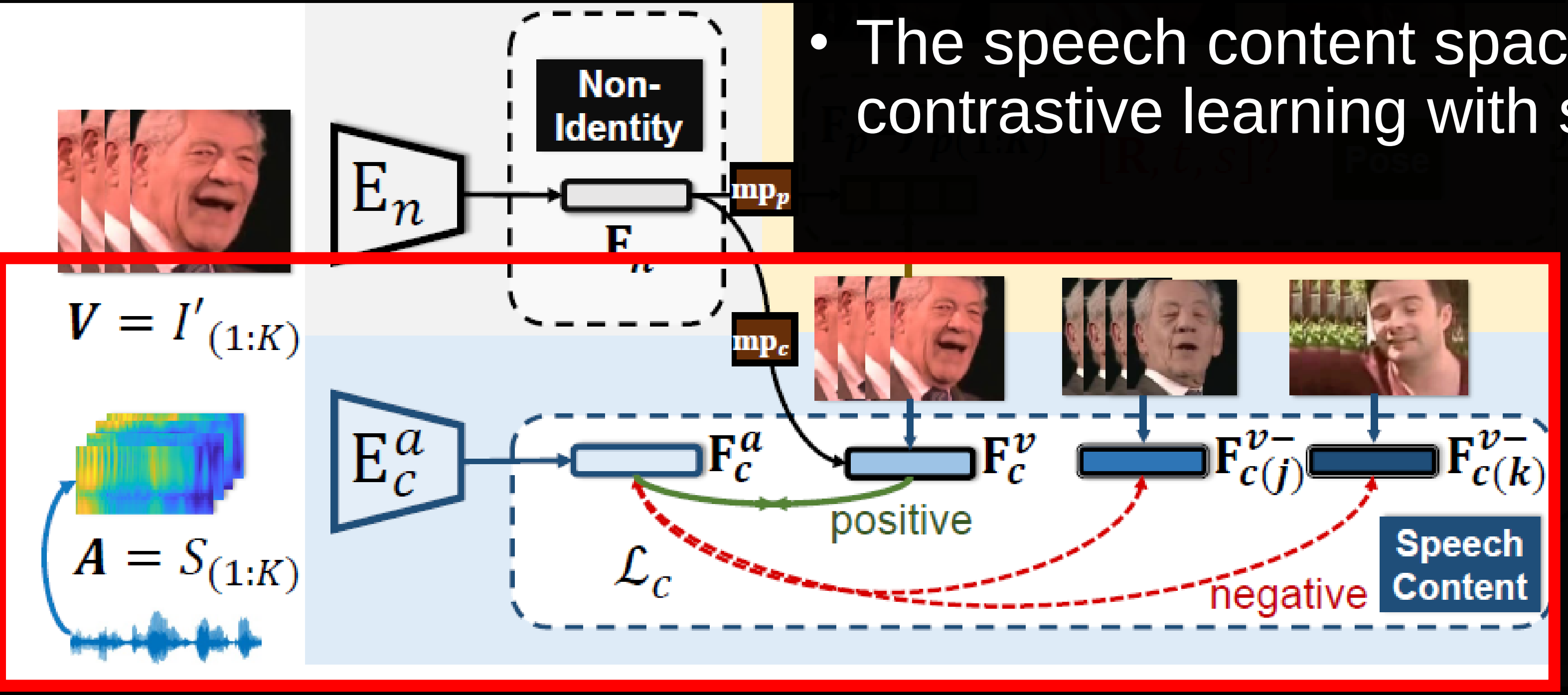
- Intuition: Source for Desired Information
  - Speech content and pose information should originate from this latent space.

Pipeline

- Speech content space

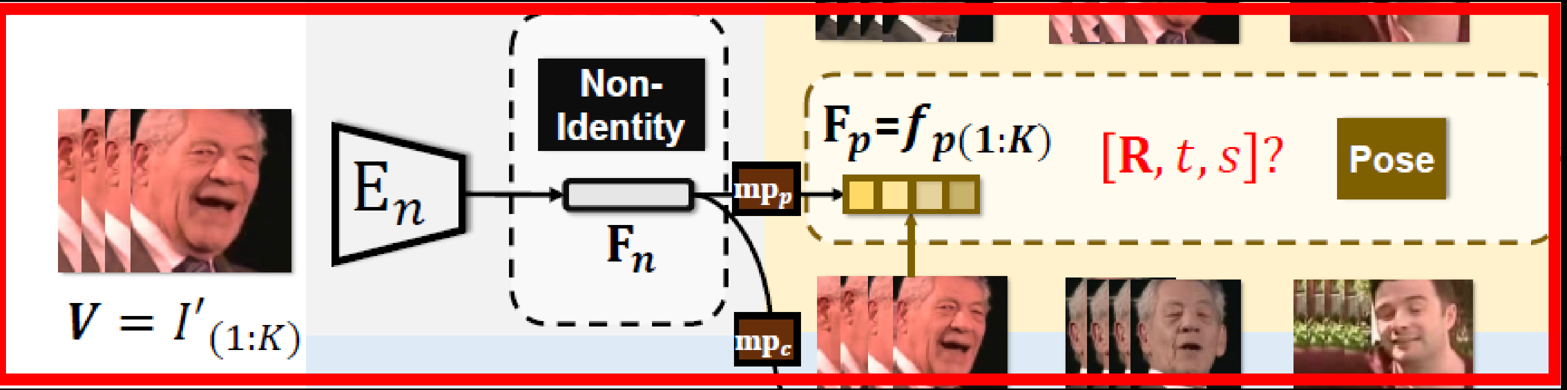
- The non-identity space is the base for the encoding of two spaces.

- The speech content space is encoded through contrastive learning with softmax contrastive loss.



Pipeline

- Pose space encoding.

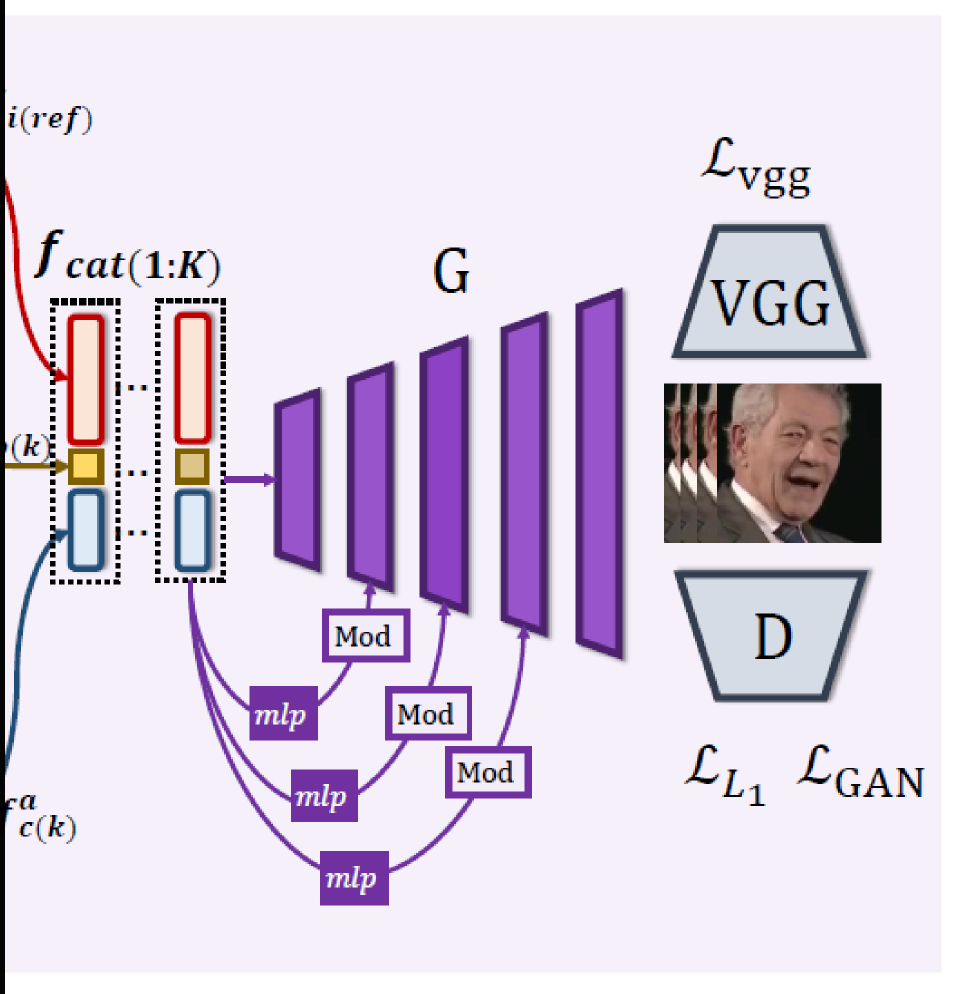


- The non-identity space is the base for the encoding of two spaces.
- The pose space is implicitly devised to a length of 12.

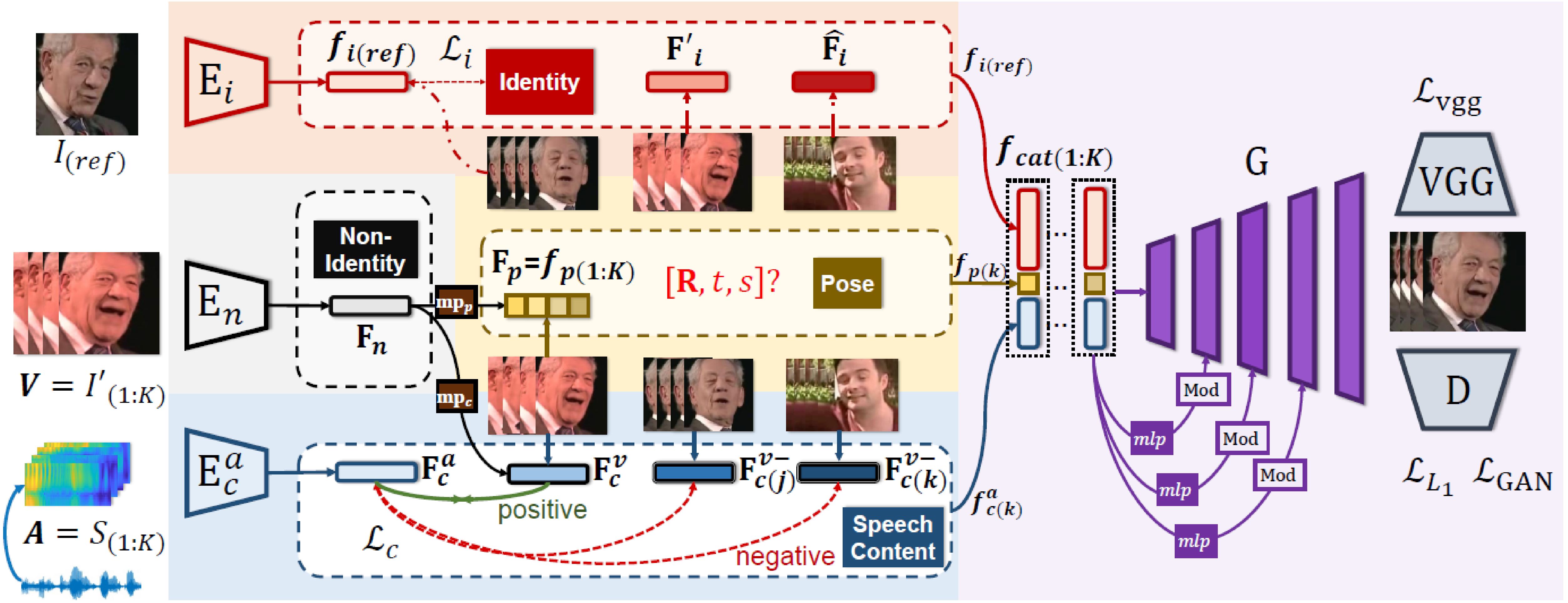
Pipeline

- Features from the three spaces are concatenated and sent to a StyleGAN2-based generator.

- The reconstruction loss implicitly enforces the pose code to learn the desired information.



Pipeline



$I_{(ref)}$
$V = I'_{(1:K)}$
$A = S_{(1:K)}$
E_i
$f_{i(ref)}$
$\mathcal{L}_i$
Identity
F'_i
$\hat{F}_i$
E_n
Non-Identity
F_n
mp_p
mp_c
$F_p = f_{p(1:K)}$
$[R, t, s]?$
Pose
E_c^a
F_c^a
F_c^v
$F_{c(j)}^{v-}$
$F_{c(k)}^{v-}$
positive
negative
$\mathcal{L}_c$
Speech Content
$f_{i(ref)}$
$f_{p(k)}$
$f_{c(k)}^a$
$f_{cat(1:K)}$
G
Mod
mlp
$\mathcal{L}_{vgg}$
VGG
D
$\mathcal{L}_{L_1}$ $\mathcal{L}_{GAN}$

Evaluation

Table 1: **The quantitative results on LRW [16] and VoxCeleb2 [15].** All methods are compared under the four metrics. For LMD the lower the better, and the higher the better for other metrics. †Note that we directly evaluate the authors' generated samples on VoxCeleb2 under their setting. They have not provided examples on LRW.

Method	LRW [16]				VoxCeleb2 [15]			
	SSIM ↑	CPBD ↑	LMD ↓	$Sync_{conf}$ ↑	SSIM ↑	CPBD ↑	LMD ↓	$Sync_{conf}$ ↑
ATVG [10]	0.810	0.102	5.25	4.1	0.826	0.061	**6.49**	4.3
Wav2Lip [44]	**0.862**	0.152	5.73	**6.9**	0.846	0.078	12.26	4.5
MakeitTalk [75]	0.796	0.161	7.13	3.1	0.817	0.068	31.44	2.8
Rhythmic Head† [8]	-	-	-	-	0.779	0.802	14.76	3.8
Ground Truth	1.000	0.173	0.00	6.5	1.000	0.090	0.00	5.9
Ours-Fix Pose	0.815	0.180	6.14	6.3	0.820	**0.084**	7.68	5.8
PC-AVS (Ours)	**0.861**	**0.185**	**3.93**	6.4	**0.886**	**0.083**	6.88	**5.9**

- Structured similarity SSIM.
- Cumulative probability blur detection (CPBD).
- Landmarks Distance (LMD) around the mouths.
- Confidence score (Sync conf) proposed in SyncNet.

Comparison with Previous Methods

- ATVG (Chen et al. 2019) (2D Landmark-based)
- Wav2Lip (Prajwal et al. 2020) (Reconstruction-based)
- MakeitTalk (Zhou et al. 2020) (3D Landmark-based)
- Rhythmic Head (Chen et al. 2020) (3D model-based)
- Ours (Reconstruction-based) (Poses are retrieved from 50 random pose source videos in the test set)

Conclusion

- Synchronization between audio and visual information is the basic self-supervision and is beneficial for cross-modal synthesis.
- Pose and possibly other information can be implicitly disentangled through learning the speech content within audio-visual synchronization
- Style-based generator is capable of information balancing through reconstruction training.

Code and models: https://github.com/Hangz-nju-cuhk/Talking-Face_PC-AVS

AI-Synthesized Media

DeepFake Detection

ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis

Yinan He[1,2*] Bei Gan[1,3*] Siyu Chen[1,3*] Yichun Zhou[1,4*]
Guojun Yin[1,3] Luchuan Song [5†] Lu Sheng[4] Jing Shao[1,3‡] Ziwei Liu[6]

商汤 sensetime

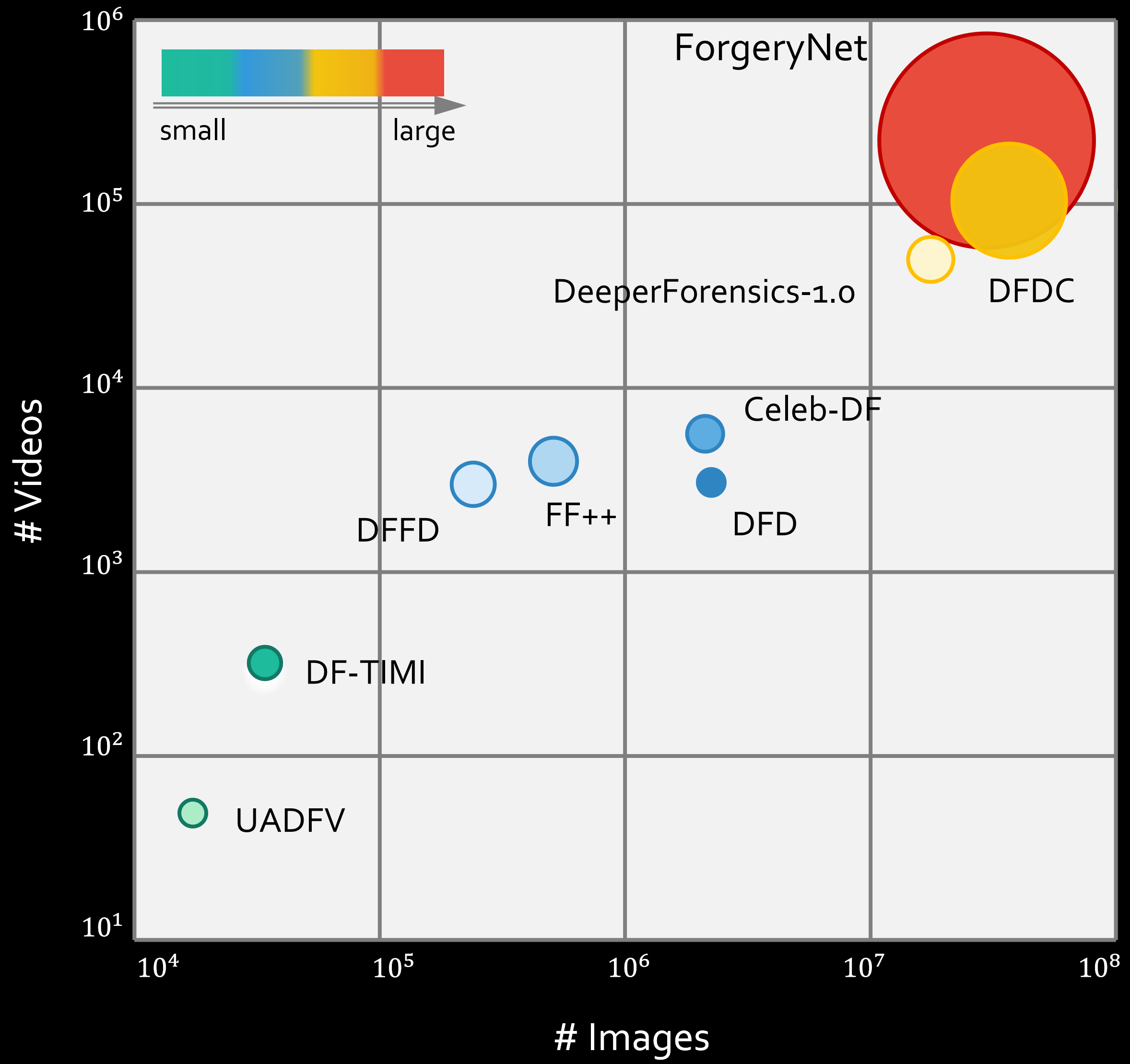
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory


Real or Fake?
NERDGEIST
what is 3.45 pounds expressed in grams?
pounds → grams
For more videos, check out:
EOS 7D
OCDE
3. add relevant detail to explain or answers
4. connect your sentence connectors
YTN사이언스24
"북, 신형 ICBM도 첫 시험발사 가능성"


NERDGEIST
What is 3.45 pounds expressed in grams?
pounds → grams
For more videos, check out:
www.video
chemistry
textbook.com
Good News in Oakland
OCDE
Coherence is the ability to:
1. expand your answers
2. answer the questions directly
YTN사이언스24
“북, 신형 ICBM도 첫 시험발사 가능성”

NERDGEIST
PART WOLF
What is 3.45 pounds expressed in grams?
pounds → grams
For more videos, check out:
www.video chemistry textbook.com
Good News in Oakland
ATHLETE
OCDE
Coherence is the ability to:
1. expand your answers
2. answer the questions directly
3. add relevant detail to explain or
answers
4. connect your senten
connectors
YTN사이언스24
"북, 신형 ICBM도 첫 시험발사 가능성"

Current Forgery Dataset



Limited scales

Limited diversity

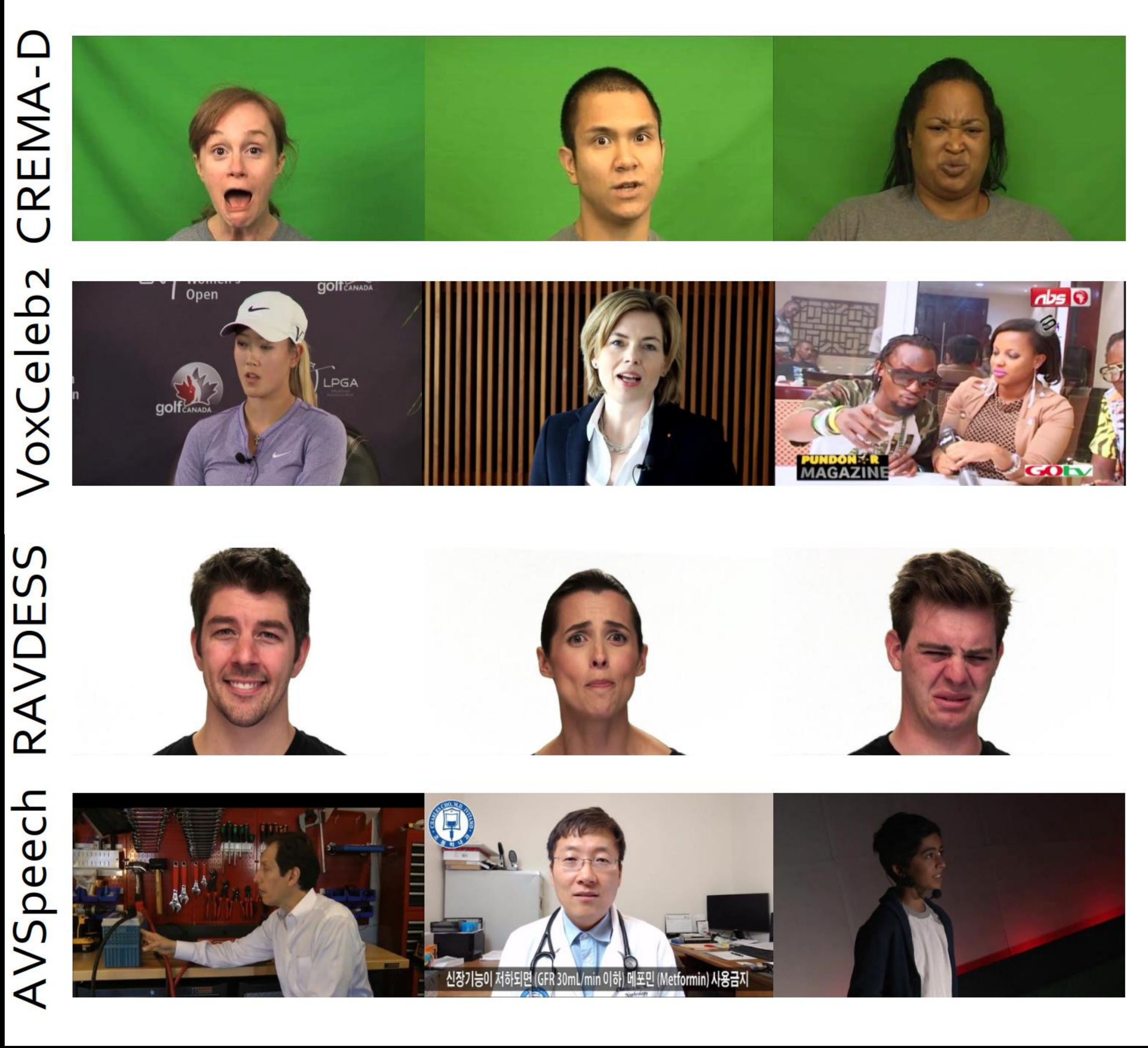
Only binary label

How about ForgeryNet ?

ForgeryNet: **Wild Original Data**

diversified dimensions

- Angle
- Expression
- Identity
- Lighting
- Scenario
- ……



ForgeryNet: **Various Forgery Approaches**

15 approaches

variety of learning-based models

2.9M still images

220k video clips

Real **Fake**

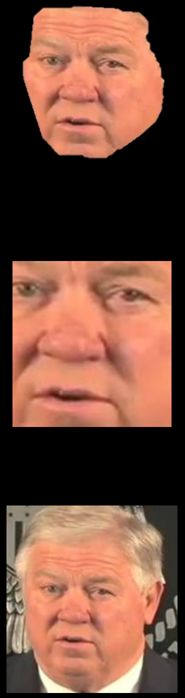
Swap Identity

DeepFakes
FSGAN
MMReplacement
FaceShifter
BlendFace
...

Modify Attribute

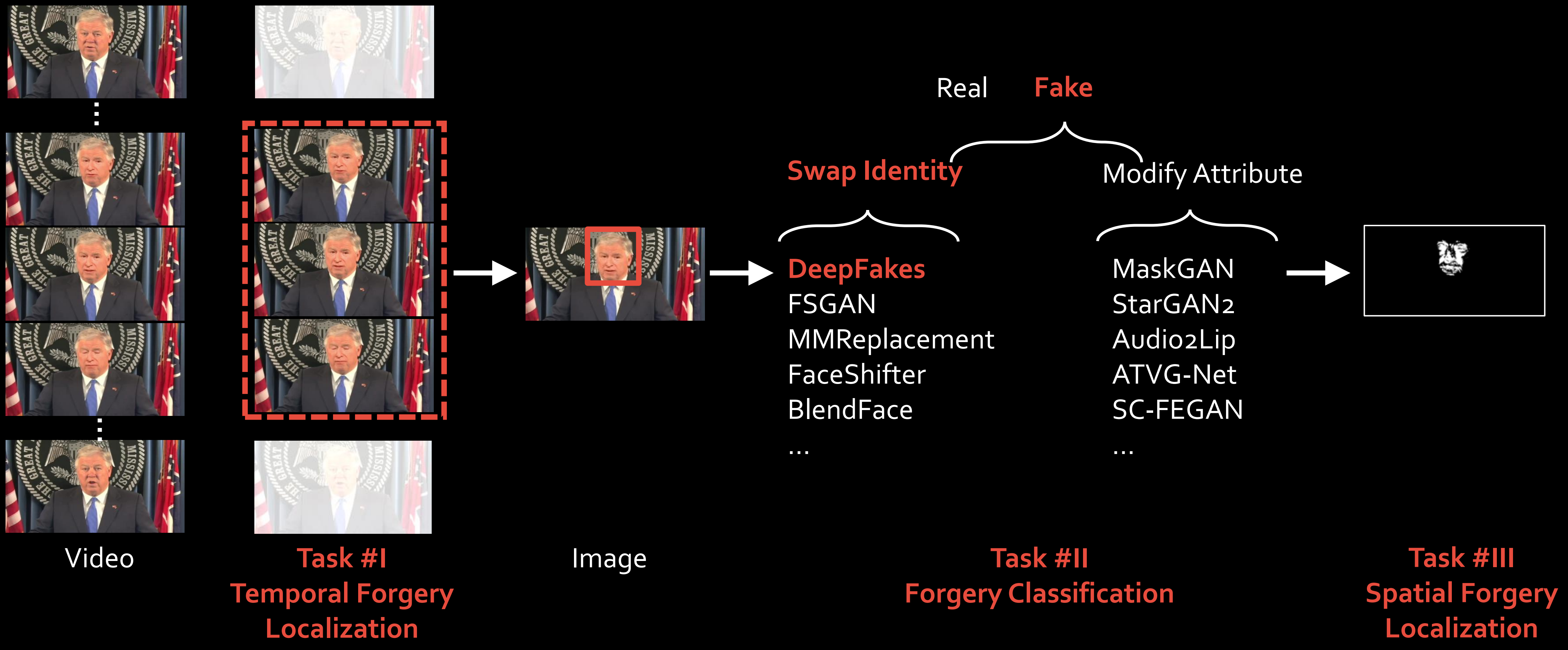
MaskGAN
StarGAN2
Audio2Lip
ATVG-Net
SC-FEGAN
...

ForgeryNet: **Diverse Re-rendering Process**



more than 36 mix-perturbations

ForgeryNet: **Comprehensive Annotations and Tasks**

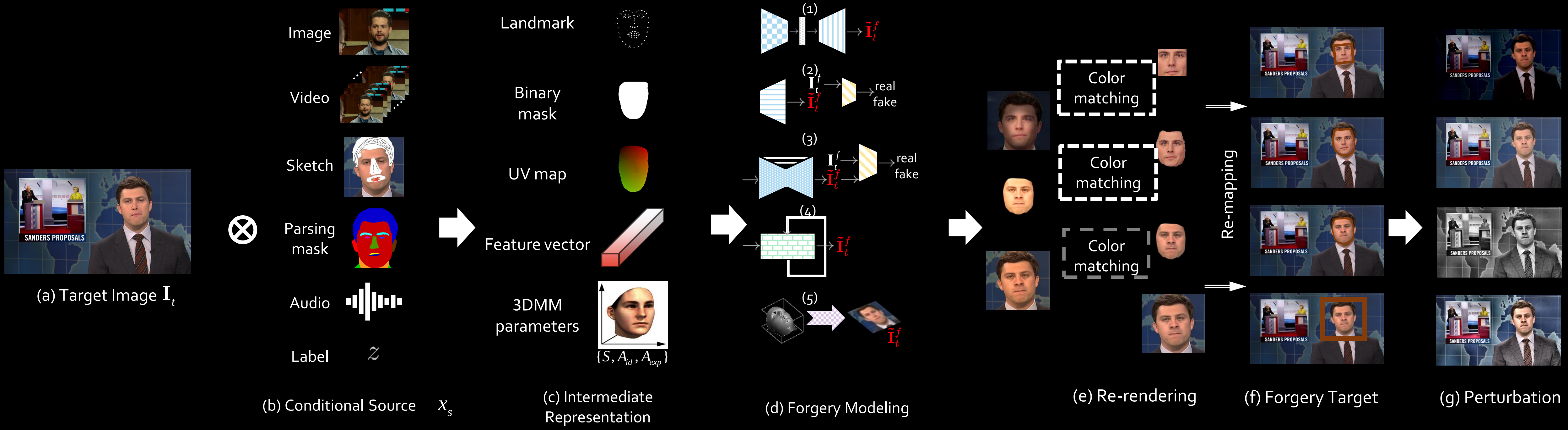


4 Tasks

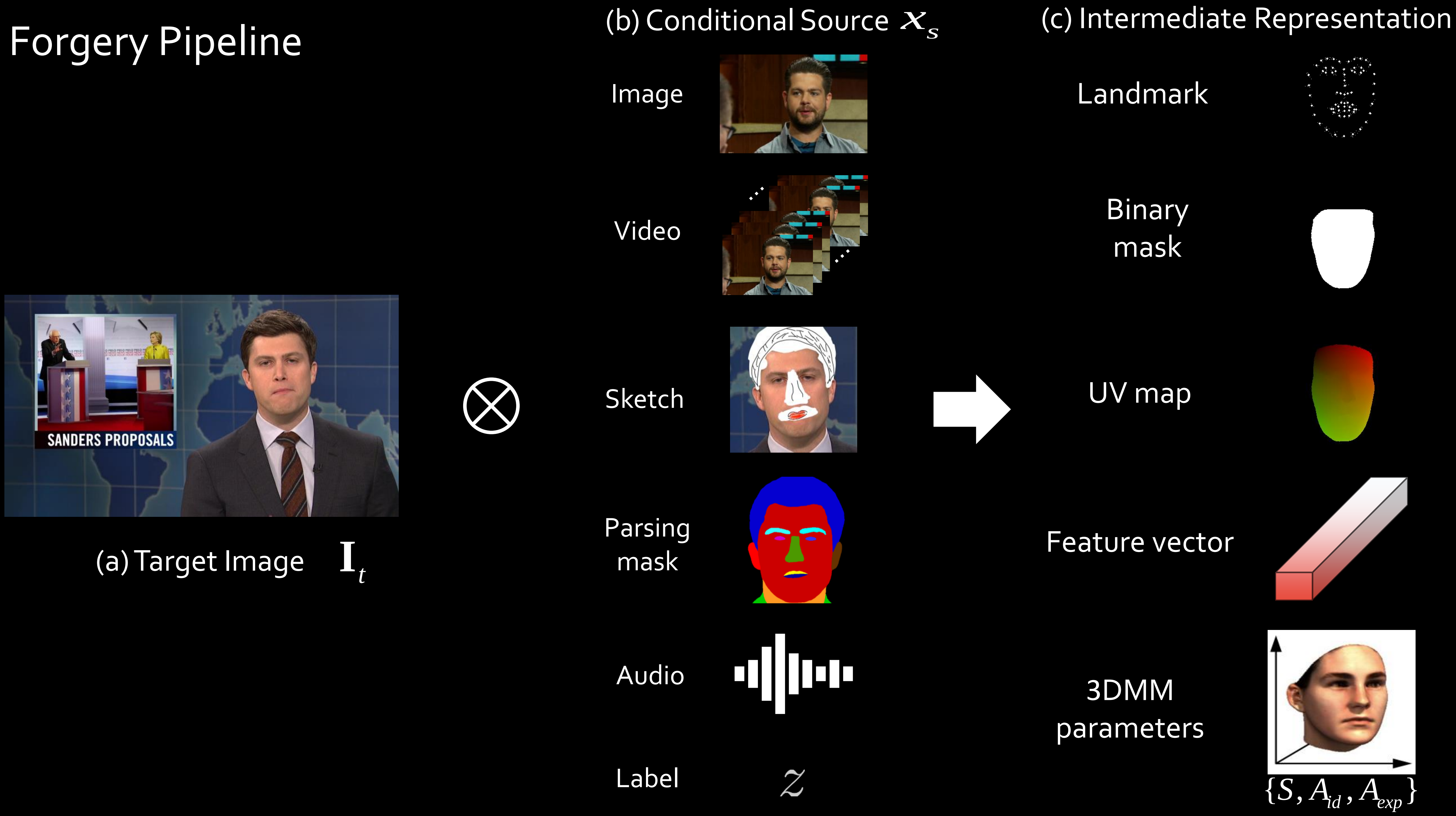
9.4M annotations

Forgery Pipeline

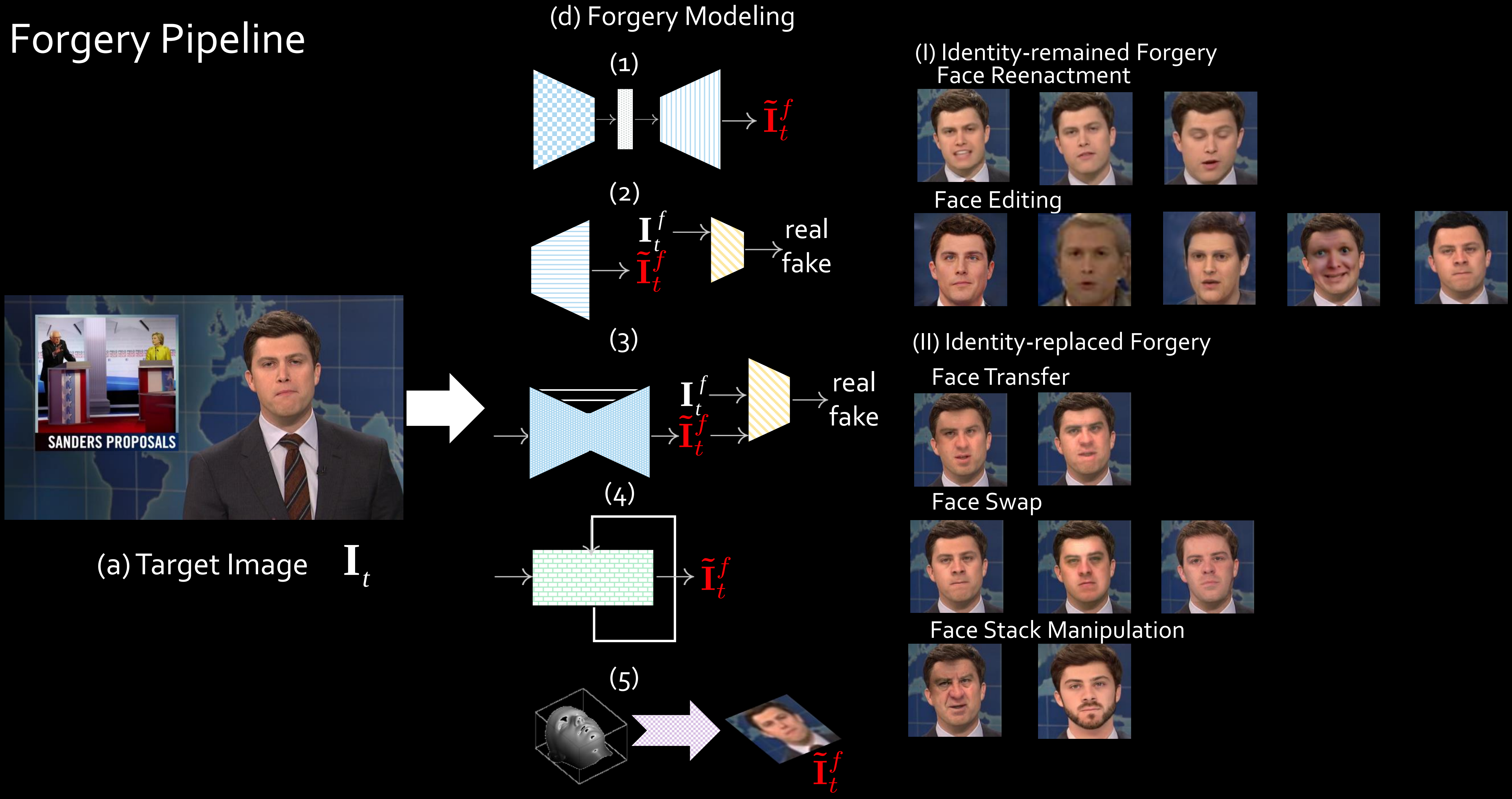
Forgery Pipeline



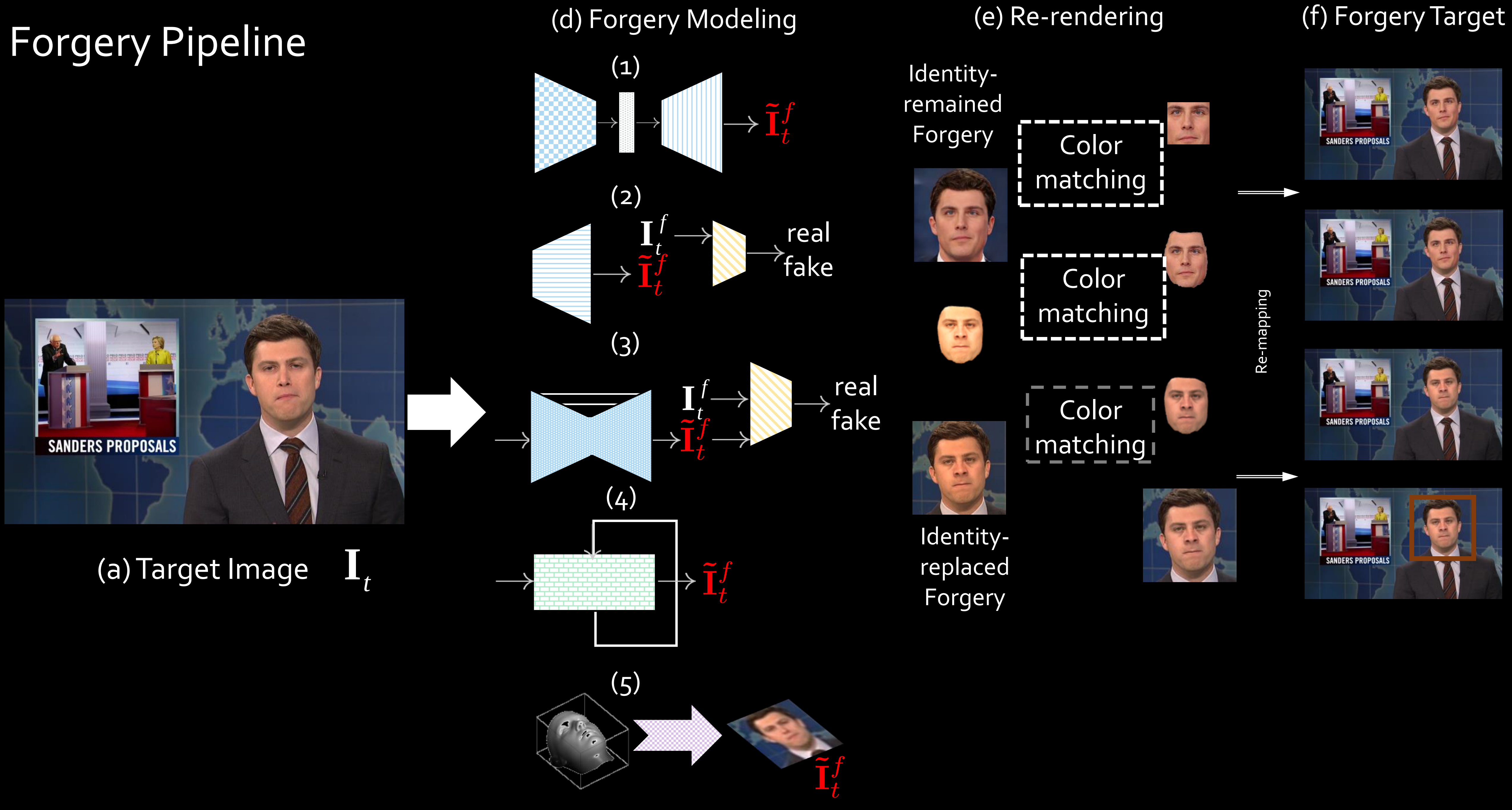
Forgery Pipeline



Forgery Pipeline



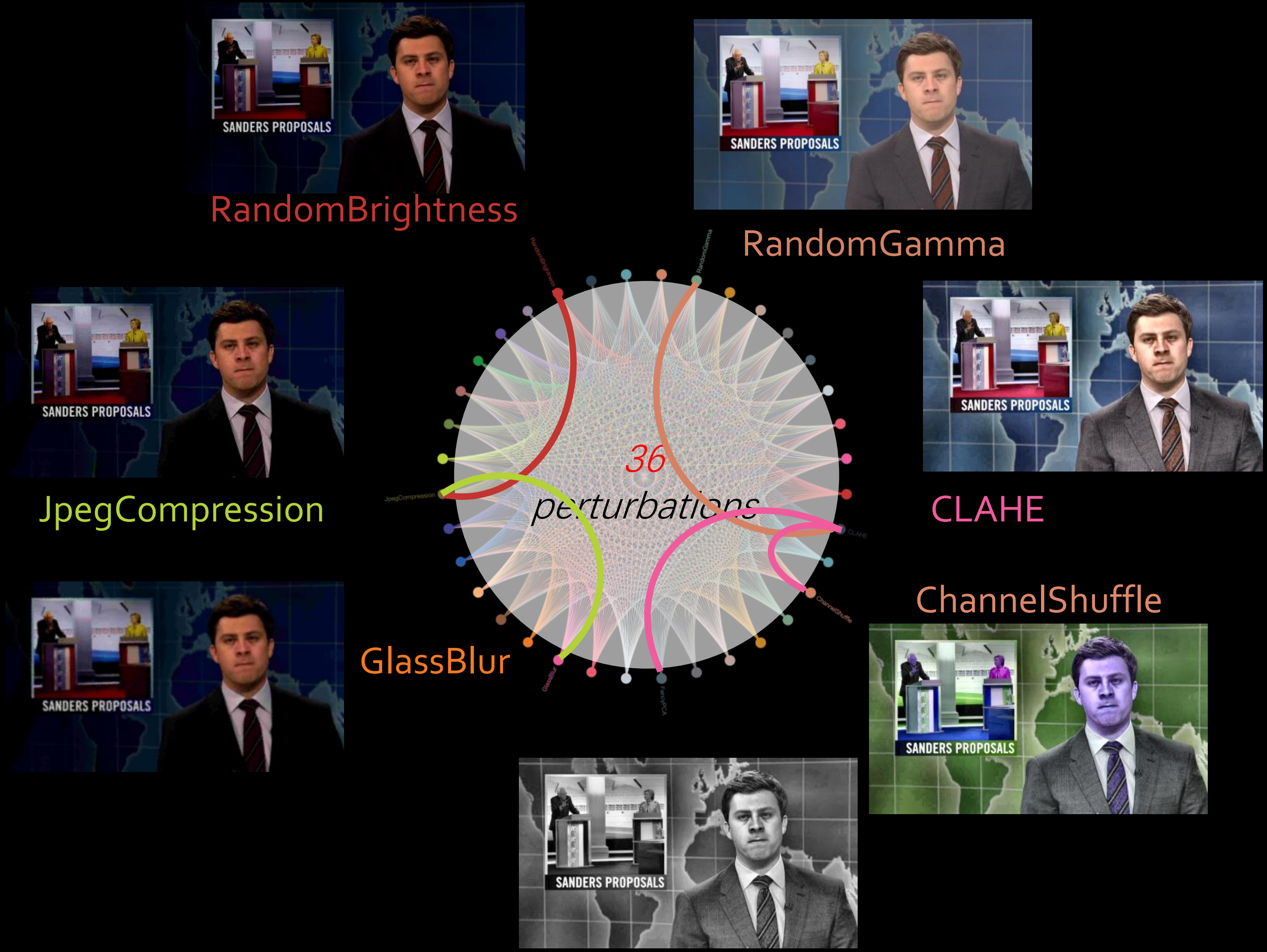
Forgery Pipeline



Forgery Pipeline

(f) Forgery Target($\tilde{\mathbf{I}}_t$)

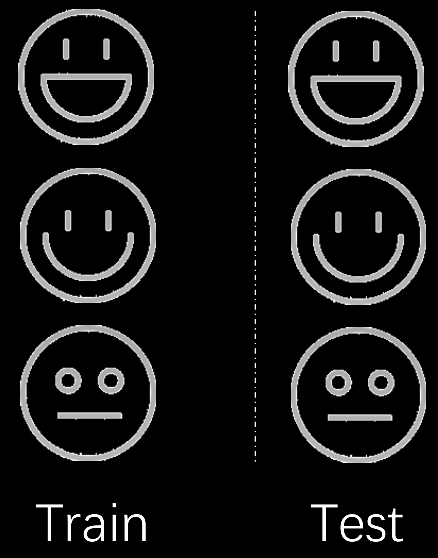
(g) Perturbation

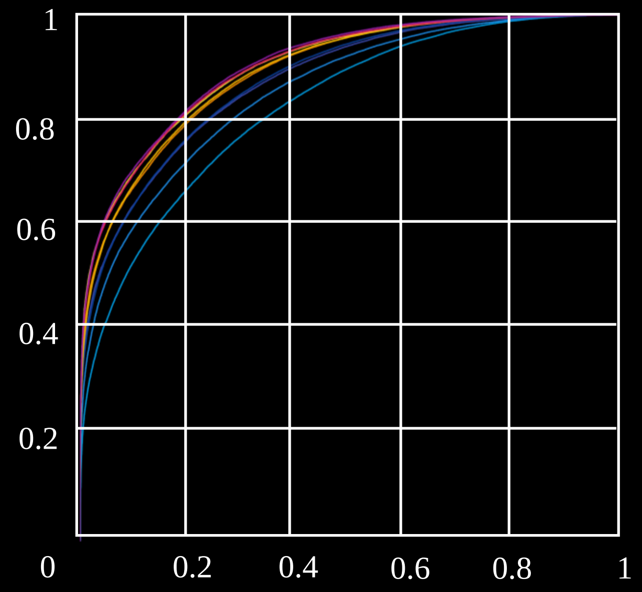


4 Task Benchmark

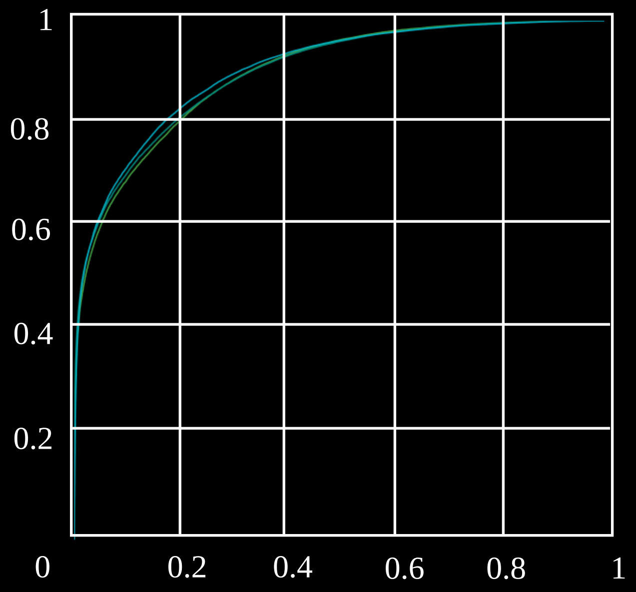
Task I: Image Forgery Classification

Intra-forgery Evaluation

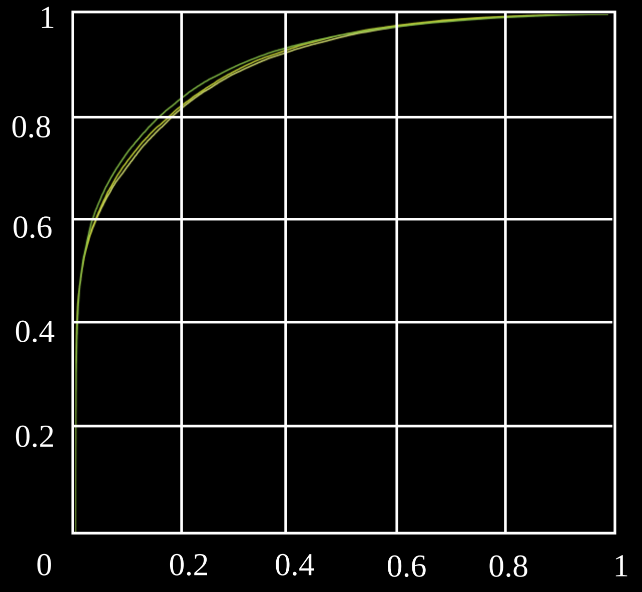




Binary Classification



3-way Classification

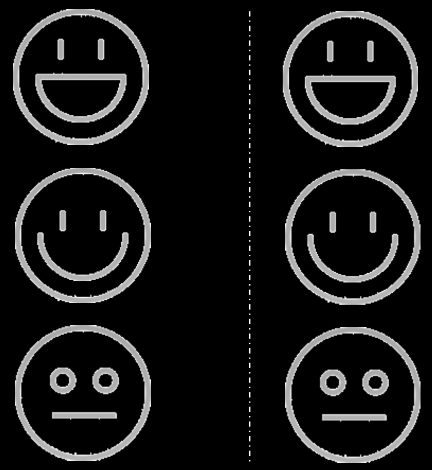


n-way Classification

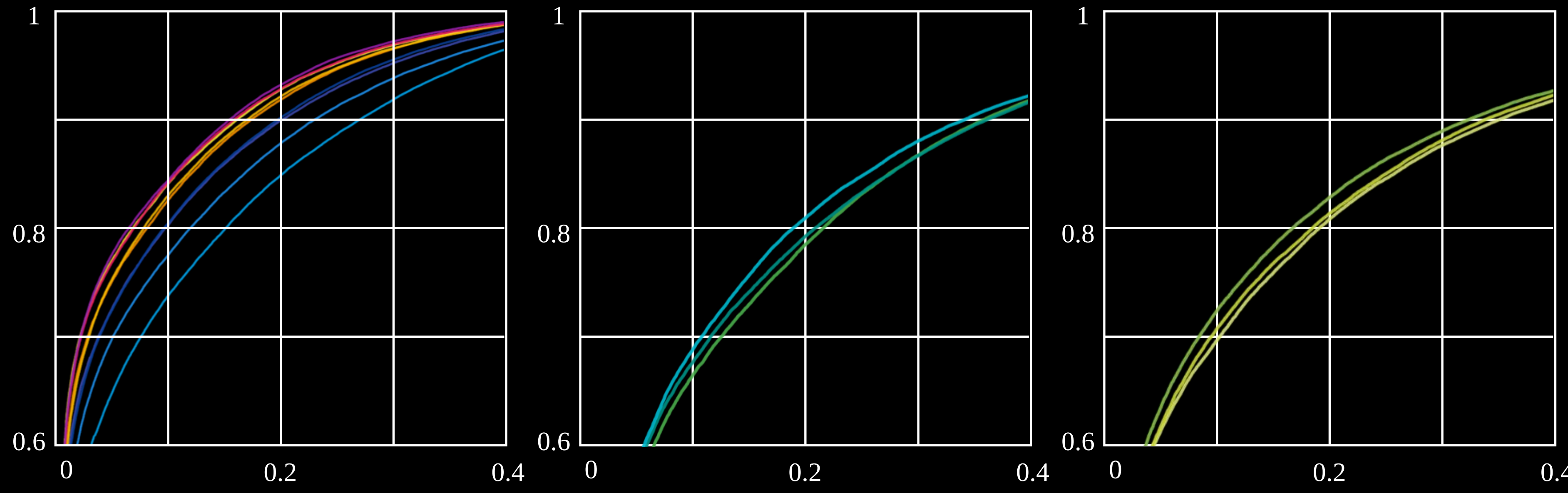
- ELA-Xception
- MobileNet-v3 Small
- MobileNet-v3 Large
- ResNet-18
- EfficientNet b0
- SAN
- Xception
- F3-Net
- Xception-GramNet
- SNR-Xception
- Xception-3class
- GramNet-3class
- F3-Net-3class
- Xception-nclass
- GramNet-nclass
- F3-Net-nclass

Task I: Image Forgery Classification

Intra-forgery Evaluation



Train   Test



Binary Classification   3-way Classification   n-way Classification

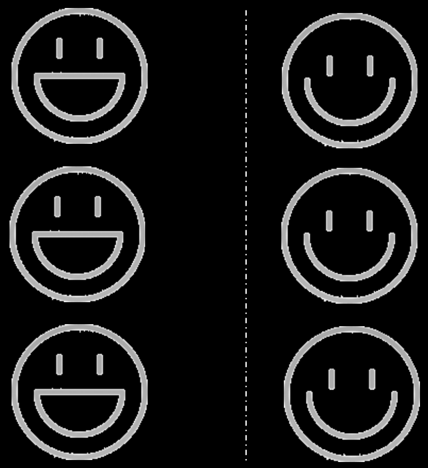
- ELA-Xception
- MobileNet-v3 Small
- MobileNet-v3 Large
- ResNet-18
- EfficientNet b0
- SAN
- Xception
- F3-Net
- Xception-GramNet
- SNR-Xception
- Xception-3class
- GramNet-3class
- F3-Net-3class
- Xception-nclass
- GramNet-nclass
- F3-Net-nclass

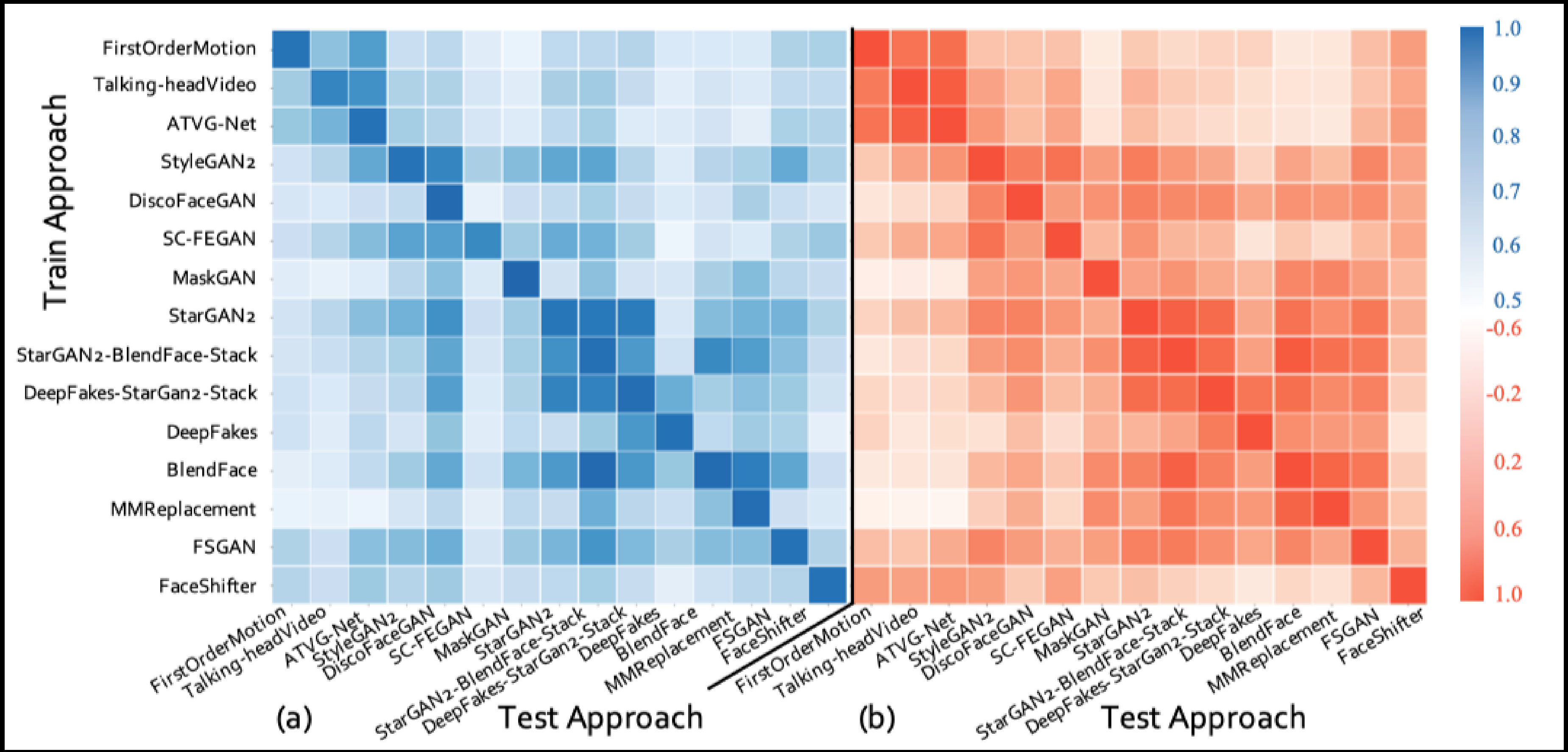
Classification becomes more difficult when the number of categories increases.

More auxiliary information potentially makes the forensics model more discriminative.

Task I: Image Forgery Classification

Cross-forgery Evaluation





AUC score map

Correlation map

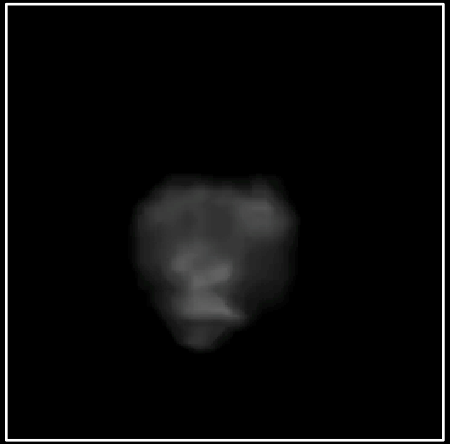
Forgery approaches belonging to the same meta-category usually have higher correlations mutually.

The generalization ability of forensics methods across forgery approaches.

Task II: Spatial Forgery Localization

Video

Ground-truth



Predicted map

Images along with forgery masks are used to train the localization model

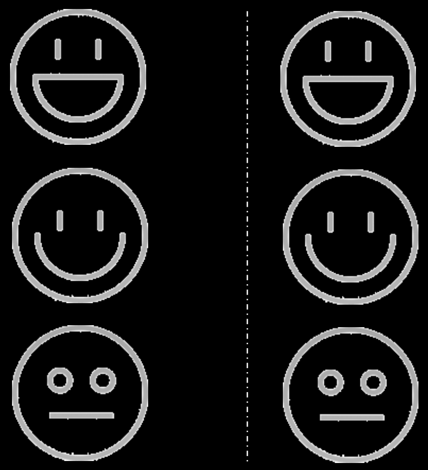
aims to specify manipulated regions

Method	IoU		IoU_{diff}			$Loss_{l1}$
	0.1	0.2	0.01	0.05	0.1	
Xception+Reg.	89.55	93.70	67.57	83.25	89.22	0.0131
Xeption+Unet [37]	95.99	98.76	79.71	92.70	97.13	0.0134
HRNet [42]	96.27	98.78	88.73	92.99	96.27	0.0114

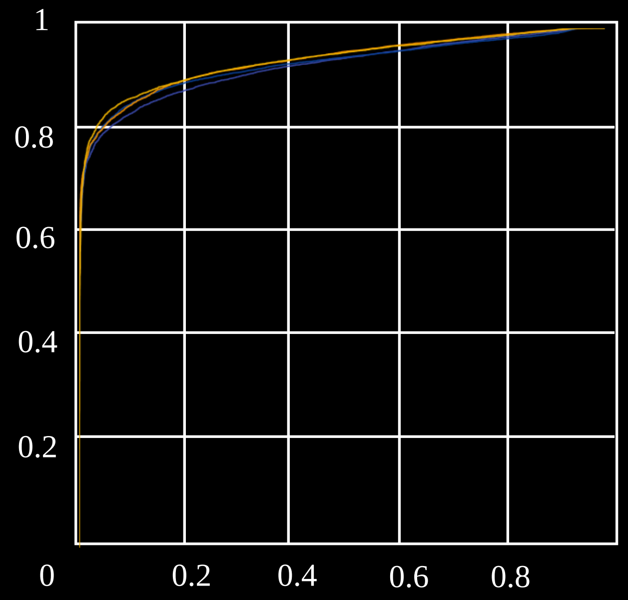
results with IOU, IOUdiff and L1 distance.

Task III: Video Forgery Classification

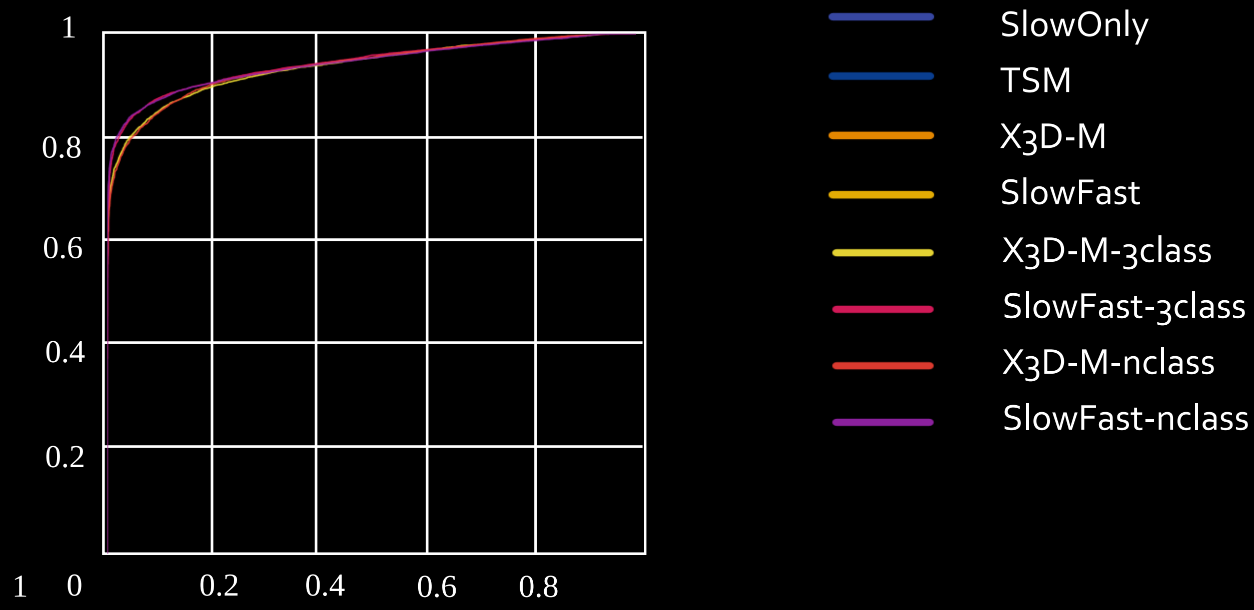
Intra-forgery Evaluation



Train Test

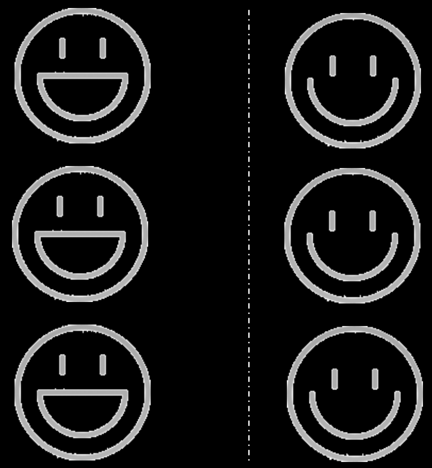


Binary Classification



n-way Classification

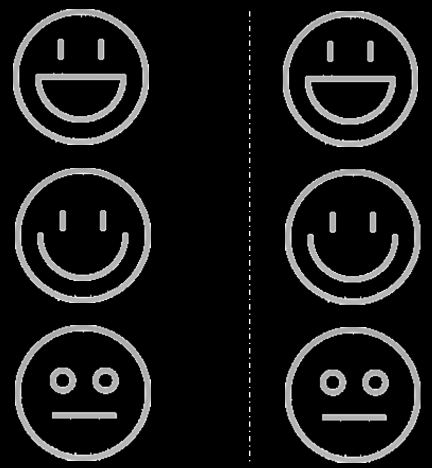
Cross-forgery Evaluation



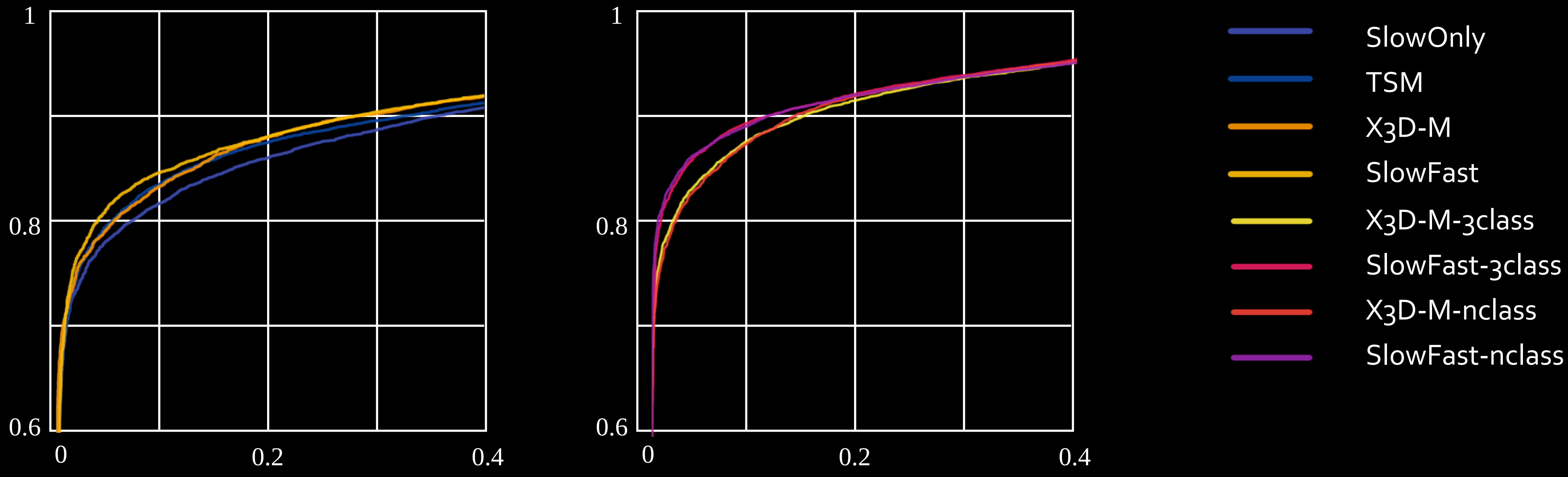
Train Test

Task III: Video Forgery Classification

Intra-forgery Evaluation

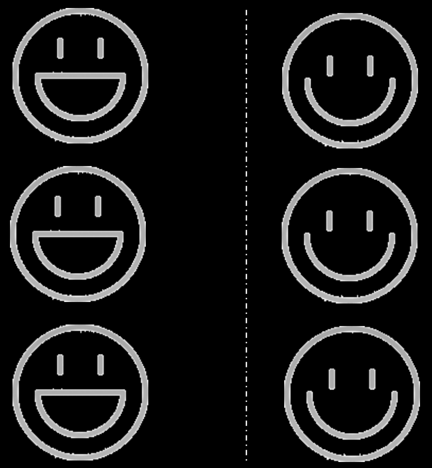


Train Test



Binary Classification   n-way Classification

Cross-forgery Evaluation

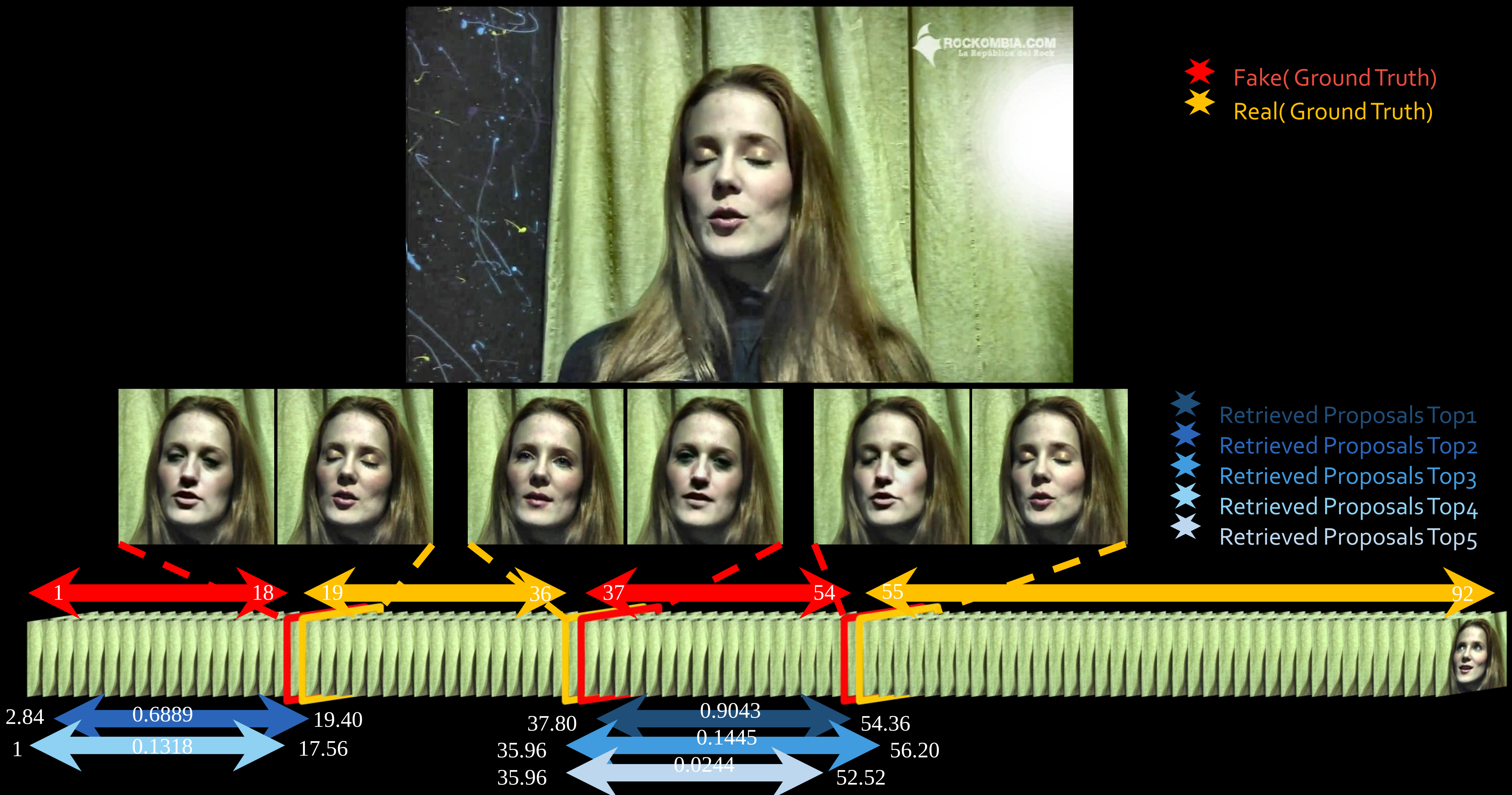


Train Test

		ID-replaced		ID-remained	
		Acc.	AUC	Acc.	AUC
X3D-M	ID-replaced	87.92	92.91	55.25	65.59
	ID-remained	55.93	62.87	88.85	95.40
SlowFast	ID-replaced	88.26	92.88	52.64	64.83
	ID-remained	52.70	61.50	87.96	95.47

Video Forgery Classification (Protocol 2)

Task IV: Temporal Forgery Localization



provide temporal boundaries of forgery segments and the corresponding confidence values


104.6 RTL
Berlins Hit-Radio
Es ist gut wieder hier zu sein. Es ist eine Weile her.
FOX NEWS SUNDAY
ABADI: NOW UNDERSTANDING THAT IRAQ IS AN ALLY
IRAQI PRIME MINISTER EXCLUSIVE
zammuTV
NEW
OLD
Liz Gannes
Video: The Greatest Pre-Launch Start-Up Pitch Ever?
LU 800
M T D
www.mtdcnc.com
That all peoples may know Him
SALVATION
Provincial Response
Global NEWS

Summary

(1)Wild Original Data

(2) Various Forgery Approaches

(3)Diverse Re-rendering Process.

(4)Comprehensive Annotations and Tasks.

Scan to download ForgeryNet

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Ambient Creation

AI-Synthesized Media

Re-enacted Creation

DeepFake Detection