REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
22-12-2021	Final Report	1-May-2020 - 30-Apr-2021

4. TITLE AND SUBTITLE
Final Report: Research for Generalizable, Explainable and Robust Machine Learning

5a. CONTRACT NUMBER
W911NF-20-1-0046

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Rutgers, The State University of New Jersey
33 Knightsbridge Road
2nd Floor, East Wing
Piscataway, NJ 08854 -3925

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75746-CS-RIP.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ms. Lana Metayer
UU	UU	UU	UU		19b. TELEPHONE NUMBER 802-656-3360

RPPR Final Report
as of 18-Aug-2022

Agency Code: 21XD

Proposal Number: 75746CSRIP **Agreement Number: W911NF-20-1-0046**
INVESTIGATOR(S):

Name: Ms. Lana Metayer
Email: spa@uvm.edu
Phone Number: 8026563360
Principal: Y

Organization: **Rutgers, The State University of New Jersey - New Brunswick**
Address: 33 Knightsbridge Road, Piscataway, NJ 088543925
Country: USA

DUNS Number: 001912864 EIN: 226001086
Report Date: 31-Jul-2021 Date Received: 22-Dec-2021
Final Report for Period Beginning 01-May-2020 and Ending 30-Apr-2021
Title: Research for Generalizable, Explainable and Robust Machine Learning
Begin Performance Period: 01-May-2020 **End Performance Period:** 30-Apr-2021
Report Term: 0-Other

Submitted By: Ms. Lana Metayer Email: spa@uvm.edu
Phone: (802) 656-3360

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: We are requesting instrumentation at Rutgers University's CBIM Center which is directed by Dr. Metaxas and focuses on research in novel machine learning, physics-based modeling, computer vision, AI, human behavior analytics, visualization and medical image analytics. The requested instrumentation includes 6 GPU servers with a total of 1TB of main memory and 8 GPU TESLA V100 processors per server, related networking switches, cables and power supplies. The proposed instruments will enable novel challenging ongoing and future research at CBIM on robust machine learning to data variability, coupling of machine learning and physics-based modeling, explainable and interpretable machine learning and AI for scene understanding and generalizability to unseen objects and scene classes, real time data analytics, dynamic event recognition, human behavior analytics and natural language understanding. This basic research requires the use of large amounts of data and is computationally intensive and CBIM currently does not have this level of computational resources. The requested instrumentation will support several existing and planned new projects at CBIM in machine learning and their applications and will enable CBIM researchers to produce state of the art research and educated graduate and undergraduate students. CBIM projects include sponsorship from AFOSR, ONR, ARO, DARPA, NSF and NIH.

The requested instrumentation will allow us to enhance the current and move towards the next level of educational activities at CBIM. The proposed equipment will be used by the current 65 PhD, 15 MSc and 10 Undergraduate students at CBIM to improve their research and education in machine learning and related applications through the use of the best GPU systems. The requested instrumentation will be installed, administered and maintained for CBIM through systems experts from the Laboratory for Computer Science at Rutgers University.

Accomplishments: 1) We analyzed facial expressions as a predictor of trust, dominance and liking. Players who were more dominant, likable and trusted had more intense facial expressions.
2) We proposed a transformer-based approach for joint generative and discriminative training for deception detection. We use video self-supervision to alleviate overfitting to small datasets.
3) We analyzed deceptive behaviors from self-attention visualization and generated videos. Specific patterns of deceptive behavior emerge that are consistent with findings in communication theory.

For a detailed description of the accomplished goals please see the attached presentation and technical report.

RPPR Final Report
as of 18-Aug-2022

Training Opportunities: The following students were directly involved in the MURI SCAN project, which gave them a lot of opportunities to hone their research skills and publish numerous papers. The students are listed as follows:

- Ligong Han (PhD)
- Anastasios Stathopoulos (PhD)
- Long Zhao (PhD – graduated) Works at Google Brain as of Jan 2022
- Lezi Wang (PhD – graduated) Works at Facebook as of Sept 2021
- Yu Tian (PhD – graduated) Works at SNAP as of June 2021
- Fangzheng Wu (MS) Just graduated.
- Sri Musunuri (MS)

Results Dissemination: 1) Keynote and Distinguished lectures

1. IEEE DACI 2021 Workshop on AI, Big Data and IoT, July 14, 2021
2. Chalearn American Sign Language Recognition Workshop, organized in conjunction with CVPR 2021, June 25, 2021
3. ChaLearn Fair Face Recognition and Analysis Workshop, organized in conjunction with ECCV 2020, August 28, 2020.
4. International Conference on Computer Science and Education (ICCSE), Toronto August 20, 2020. Title: Scalable and Explainable Analytics for Computer Vision and Medical Applications
5. International Conference on Movement and Computing, Jersey City, NJ, July 15, 2020
6. Computer Science and Engineering, University of Nebraska – Lincoln, January 18, 2020
7. PSEG Power Lunch Series, 21st Century Innovation; How does it affect you? March 23, 2021

2) Papers Published:

- Wang, L., Bai, C., Bolonkin, M., Burgoon, J.K., Dunbar, N.E., Subrahmanian, V.S. and Metaxas, D., "Attention-based facial behavior analytics in social communication". In Detecting Trust and Deception in Group Interaction (pp. 123-137). Springer, Cham, 2021
- Burgoon, J.K., Metaxas, D., Nunamaker, J.F. and Ge, S.T., "Cultural Influence on Deceptive Communication". In Detecting Trust and Deception in Group Interaction (pp. 197-222). Springer, Cham, 2021
- Stathopoulos, A., Han, L., Dunbar, N., Burgoon, J.K. and Metaxas, D., "Deception Detection in Videos Using Robust Facial Features". In Proceedings of the Future Technologies Conference (FTC), 2020 (Best Student Paper Award)
- Han, L., Musunuri, S.H., Min, M.R., Gao, R., Tian, Y. and Metaxas, D., "AE-StyleGAN: Improved Training of Style-Based Auto-Encoders". In IEEE Winter Conf. on Applications of Computer Vision (WACV), 2022
- Zhao, L., Zhang, Z., Chen, T., Metaxas, D. and Zhang, H., "Improved Transformer for High-Resolution GANs". In Advances in Neural Information Processing Systems (NeurIPS), 2021
- Han, L., Min, M.R., Stathopoulos, A., Tian, Y., Gao, R., Kadav, A. and Metaxas, D.N., "Dual Projection Generative Adversarial Networks for Conditional Image Generation". In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
- Zhao, L., Wang, Y., Zhao, J., Yuan, L., Sun, J.J., Schroff, F., Adam, H., Peng, X., Metaxas, D. and Liu, T., 2021. Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 (Oral)
- Metaxas, D.N., Zhao, L. and Peng, X., "Disentangled Representation Learning and Its Application to Face Analytics". In Deep Learning-Based Face Analytics (pp. 45-72). Springer, Cham, 2021
- Zhao, L., Peng, X., Tian, Y., Kapadia, M. and Metaxas, D.N., "Towards Image-to-Video Translation: A Structure-Aware Approach via Multi-stage Generative Adversarial Networks". In International Journal of Computer Vision (IJCV), 2020

Honors and Awards: - Best Student Paper Award in FTC 2020

Protocol Activity Status:

Technology Transfer: 1. International Patent Filed (PCT/US21/34414) "DISTRIBUTED GENERATIVE ADVERSARIAL NETWORKS SUITABLE FOR PRIVACY-RESTRICTED DATA" Patent Filed: 27-MAY-2021 (D. Metaxas, Q. Chang, H. Qu and Y. Zhang).

RPPR Final Report
as of 18-Aug-2022

PARTICIPANTS:

Participant Type: PD/PI
Participant: Dimitris Metaxas
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yuxiao Chen
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Ligong Han
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Anastasios Stathopoulos
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Tinfeng Li
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Kostantinos Dafnis
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Shiyu Zhao
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Song Wen
Person Months Worked: 10.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Zhuowei Li
Person Months Worked: 10.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Zhixing Zhang
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Evgenia Chroni
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Zhaoyang Xia
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Qilong Zhangli
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Qi Chang
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Ananya Jana

Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yunhe Gao
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Meng Ye
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Bingyu Xin
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Di Liu
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Xiaoxiao He
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Long Zhao
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Lezi Wang
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:

National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yu Tian
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Zhiqiang Tang
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Pemgxiang Wu
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Jingru Yi
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Fangzheng Wu
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Ryhan Moghe
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Sri Musunuri
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 18-Aug-2022

Participant Type: Faculty
Participant: Valdimir Pavlovic
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Faculty
Participant: Mubbasir Kapadia
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Dimitris Metaxas
Signature Date: 12/22/21 2:27PM

Analysis and Discovery of Deceptive Behavior using Discrete Neural Representations: Current and Future Plans

Dimitris N. Metaxas

CBIM, Rutgers University

Collaborators: MURI Team

Rutgers PhD Students: Anastasis Stathopoulos, Ligong Han

Goal

We mainly focus on analysis of the visual cues from faces (non-verbal) for *deception detection*

Research Activities Over the Past Year

Part1: Analysis of *trust*, *dominance* and *liking* using the framework we developed last year for deception detection in collaboration with MURI Team

Part2: Method with Embedded Explainability and Discovery of Complex FAUs for Deception Detection

- We adapt a method, proposed for scene classification by our group, to predict deceptive behavior from raw input clips and constrain prediction on a learned Facial Action Unit (FAU) basis using MURI data

Research Activities Over the Past Year (cont.)

Part3: Analysis by Synthesis, Understanding and Robust Deception for Deception

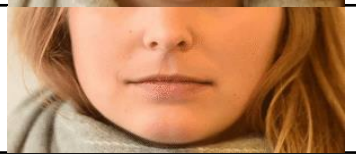
- Unified architecture for generative and discriminative learning
- Transformer-based models for detection of deceptive behavior
- VQ-GAN to learn a condensed representation of the input video sequences using video self-supervision
- Video generation conditioned on a specific behavior (deceptive/non-deceptive) using the same model
- Enables us to analyze specific patterns present in deceptive behavior and also understand the FAUs that contribute

Deception Behavior Analysis from Communication Theory

- In the latest deception theory, deception is represented by the combination of facial Action Units (AUs), including:
  - More blinks (AU45) with emotional responding and masking, fewer blinks with cognitively loaded responses and efforts at neutralization
  - Sneer (AU9 + AU10) while feigning sadness
  - Lip adaptors (AU18, AU19, AU23, AU24)
  - etc.
- Sources for the above come from various articles and include:
  - DePaulo (2003) (but this is seriously outdated)
  - Cohn, Zlochower, Lien & Kanade (1999)
  - Porter & ten Brinke (2008)
  - Waller, Cray, & Burrows (2008)
  - Kessous Castellano & Caridakis (2009)
  - Matsumoto, Willingham & Olide (2009)
  - Hurley & Frank (2011)
  - ten Brinke & Porter (2012)
  - ten Brinke, Porter & Baker (2012)
  - Matsumoto & Hwang (2017)

Deception Behavior Analysis from Communication Theory

- Samples of Action Units are considered as deception:

Action Unit	Description	Facial Muscle	Example (Hover to Play)
AU45	Blink	Relaxation of *Levator Palpebrae and Contraction of Orbicularis Oculi, Pars Palpebralis.*	
Sneer **AU9** + AU10	Nose Wrinkler	*Levator labii superioris alaquae nasi*	
Sneer AU9 + **AU10**	Upper Lip Raiser	*Levator Labii Superioris, Caput infraorbitalis*	
Lip adaptors (**AU24**)	*Lip Pressor*	*Orbicularis oris*	
Faked happiness (**AU12**, but missing AU6)	Lip Corner Puller	*Zygomatic Major*	

Previous Years

- We have developed attention methods that indicate features responsible for deception that were consistent to what we know based on communication theory

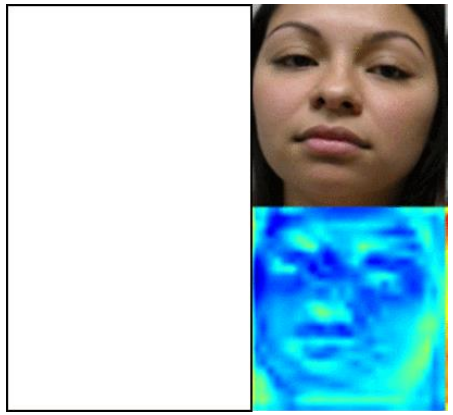
- We applied a 3D convolutional neural (C3D) [1] network to classify a player as the Deceiver or Truth-Teller

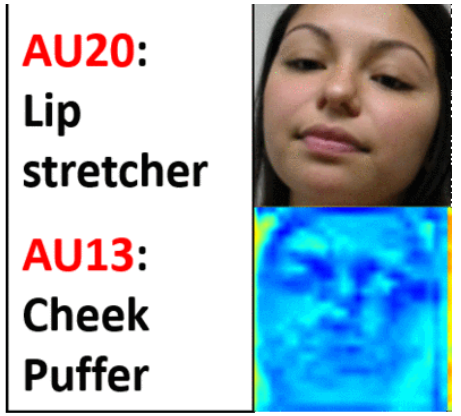
- We used attention mechanism **at the pixel level** to perform retrospective model analysis and discover AUs that relate to deceptive behavior

[1] Tran et al., **Learning spatiotemporal features with 3d convolutional network**, ICCV 2015

Method 1: Use Attention on C3D model to Discover AUs

- A video is broken into different short clips
- The clips with high Deceiver/Truth-Teller probabilities from the C3D model are used for deeper investigation
- We compare what the attention model predicts to known deception cues provided by our group experts
  - The facial cues are converted to facial action units (AU)
- Based on the game **spies** are more often **deceptive** than **villagers** who are more often **Truth-Tellers**. But as we observe from the analysis roles can reverse
- We do not know the locations and duration of when deception occurs





Limitations

- The previous model works at the pixel level and therefore cannot separate well facial geometry from expressions and behaviors

- The previous model is hard to train and is prone to overfitting to the identity of a person due to the small number of training samples

- Retrospective analysis of deceptive behavior is done at the pixel level
  - We need higher-level analysis (AUs and facial expressions) with explainability and understanding of how deception is manifested visually

Previous Year: Method 2

To further improve our model:

- We proposed an approach for deception detection that infers deceptive behavior based on facial features and person invariant
- We proposed a new non-pixel level attention mechanism to enable analysis of deceptive behavior by directly studying correlations of AUs

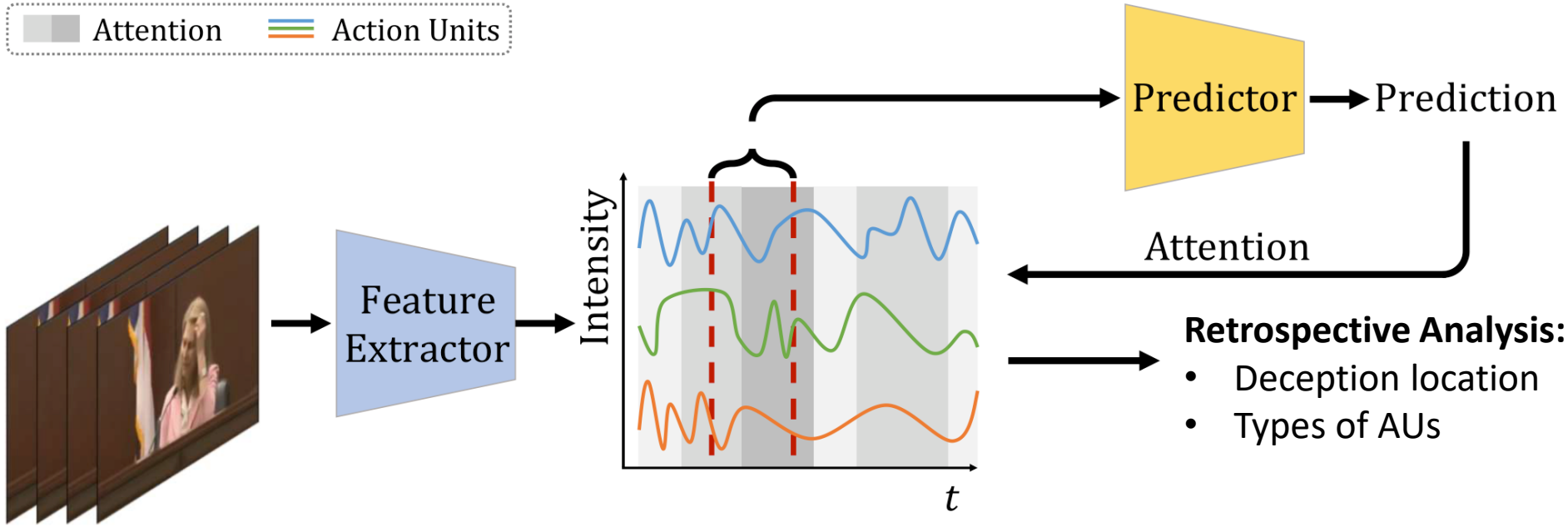
Previously: Deceiver vs Truth-Teller Classification

- **Method 2: Two-stage approach**
  1. For each video frame, extract identity invariant and robust facial features (17 AUs intensities and gaze angles) and concatenate them channel-wise
  2. Feed facial features to a Temporal Convolution Network (TCN) [2] and train the model for classification at the video level (no frame-wise ground-truth)

[2] Lea et al., **Temporal convolutional networks: A unified approach to action segmentation**, ECCV 2016

Previously: Deceiver vs Truth-Teller Classification

- Pipeline overview
  - 1-D FAU signals are extracted from video sequences
  - A predictor temporal convolutional network is then trained on the extracted waveforms
  - AU based Attention: In this ML framework attentions are computed by backpropagating the trained predictor model to discover type of AUs and deception location



Previously: Evaluation and Results

- Results on Resistance Game (ours)

Method	ACC (%)
LBP [1]	49.6
TSN [2]	51.2
Ours	**71.1**

- For LBP [3], we reimplemented the baseline. LBP serves as a naïve baseline.
- For TSN [4], we use the official implementation and run on our Resistance Game dataset.

Best Student Paper Award in Future Technology Conference (FTC) 2020

[3] Ojala et al., **Multiresolution gray-scale and rotation invariant texture classification with local binary patterns**, PAMI 2002
[4] Wang et al., **Temporal Segment Networks,** ECCV 2016

Previously: Localizing and Interpreting Deceptive Behavior

- Adapt Grad-CAM [5] to find the find the attention of the model in the time domain (this way AUs and their durations/locations can possibly be detected)
- For positive (deception) samples we can compute the key time-steps for the decision of the detection model
- Utilize the gradient of the model w.r.t. a feature layer and AUs
- Interpret Deception based on presence of AUs

[5] Selvaraju et al., **Grad-Cam: Visual explanations from deep networks via gradient-based localization,** ICCV 2017

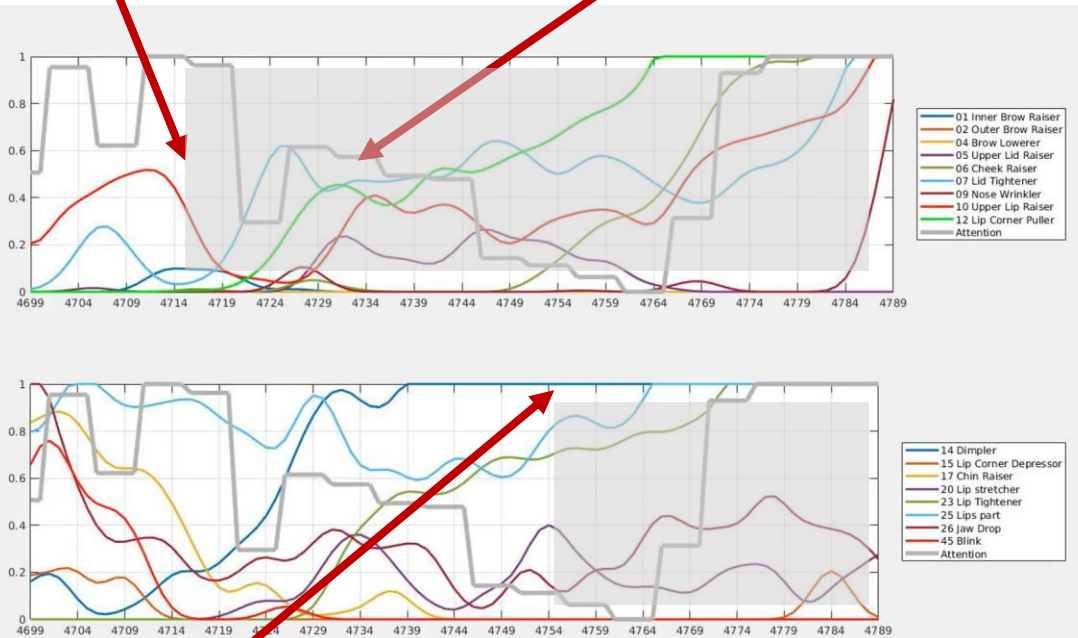
Attention based on our Data. Location: US, AZ

- AZ-005 1

Gray curve indicates attention score

AU09 Nose Wrinkler

AU10 Upper Lip Raiser



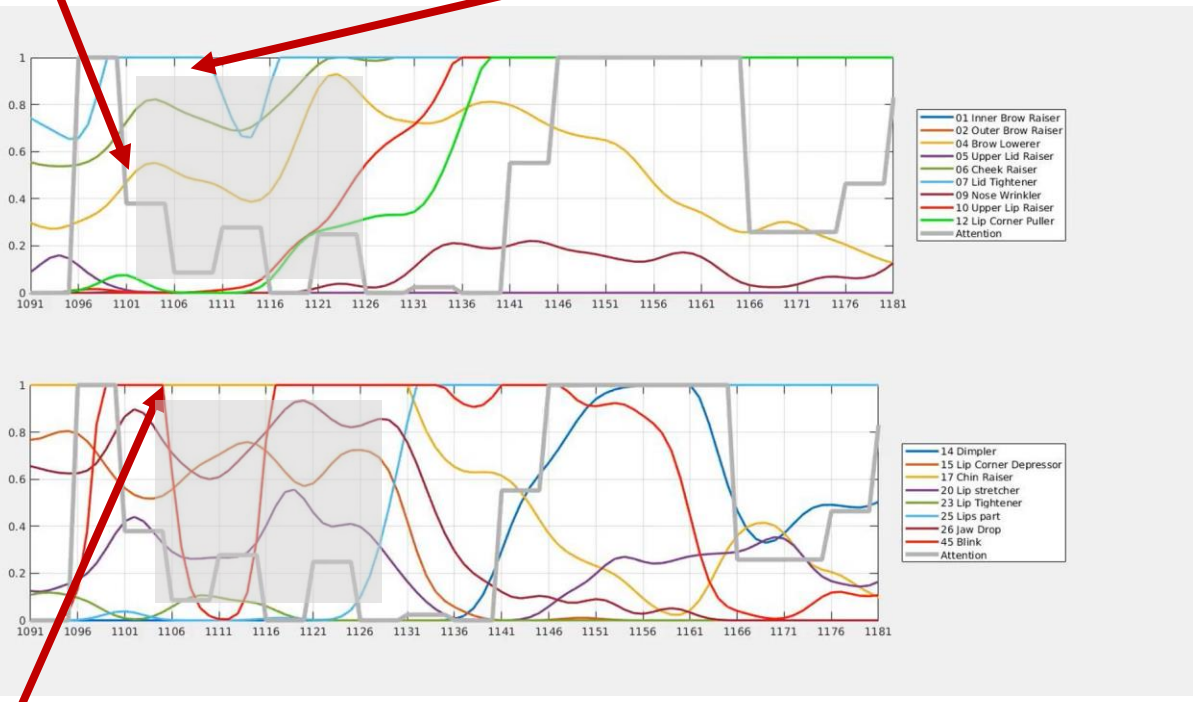
AU45 Blinks

Attention based on our Data. Location: Israel

- ISR-001 2

Gray curve indicates attention score

AU12 Lip Corner Puller



AU45 Blinks

Limitations of Previous Method

- Attention analysis is *retrospective*
- The model performance is upper-bounded by the accuracy of the off-the-shelf FAU detector
- The framework is not easily extendable to multi-modal input signals (such as raw videos, text)

This year

1. **Part 1:** We use our previous general framework for **trust**, **dominance** and **liking**

2. **Part 2:** For **explainability** and model **robustness** we adapt our ScenarioNet [6] for deception detection

3. **Part 3:** We propose a novel unified architecture for generative and discriminative learning (VQ-GAN + Transformer). We use it for deception detection and video generation conditioned on a specific behavior (deceptive/non-deceptive), which lets us analyze specific **higher-level patterns** present in **deceptive behavior**.

[6] Daniels and Metaxas, **ScenarioNet: An Interpretable Data-Driven Model for Scene Understanding**, IJCAIW 2018

Part 1: Trust, Dominance and Liking

- We used the same framework for analysis of **trust**, **dominance** and **liking**
- We used 250 videos and trained our model to regress dominance, liking and trust using an MSE loss function
- Given a trained model, we predicted dominance, liking and trust on a test set and retained 25 subjects with the lowest error for further analysis
- By using the proposed attention mechanism, we identified the key frames in the input video and performed retrospective analysis on the facial features that are most prevalent in the model's prediction

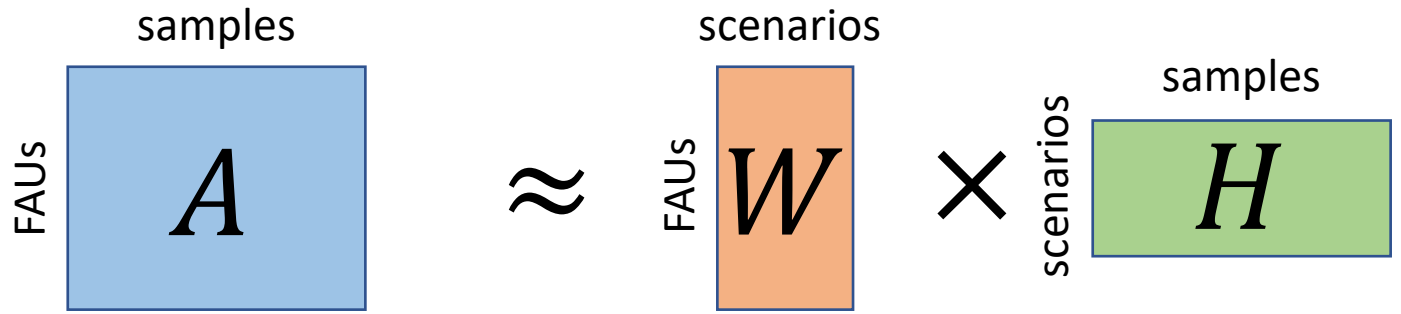
Part 1: Trust, Dominance and Liking

- Analysis of the players' facial behaviors suggests that subjects who were more dominant, likable and trusted had more intense facial expressions
- It seems that players more involved in the game gained the trust of their peers
- Facial expressions such as ***smiling*** and ***eyebrow raising*** emerged more than others
- ***No noticeable difference*** when examining the AUs of players across different countries, which supports our hypothesis that expressions of trust are ***culture-invariant***

Book chapter "The Psychology of Trust from Relational Messages" with UA and USB

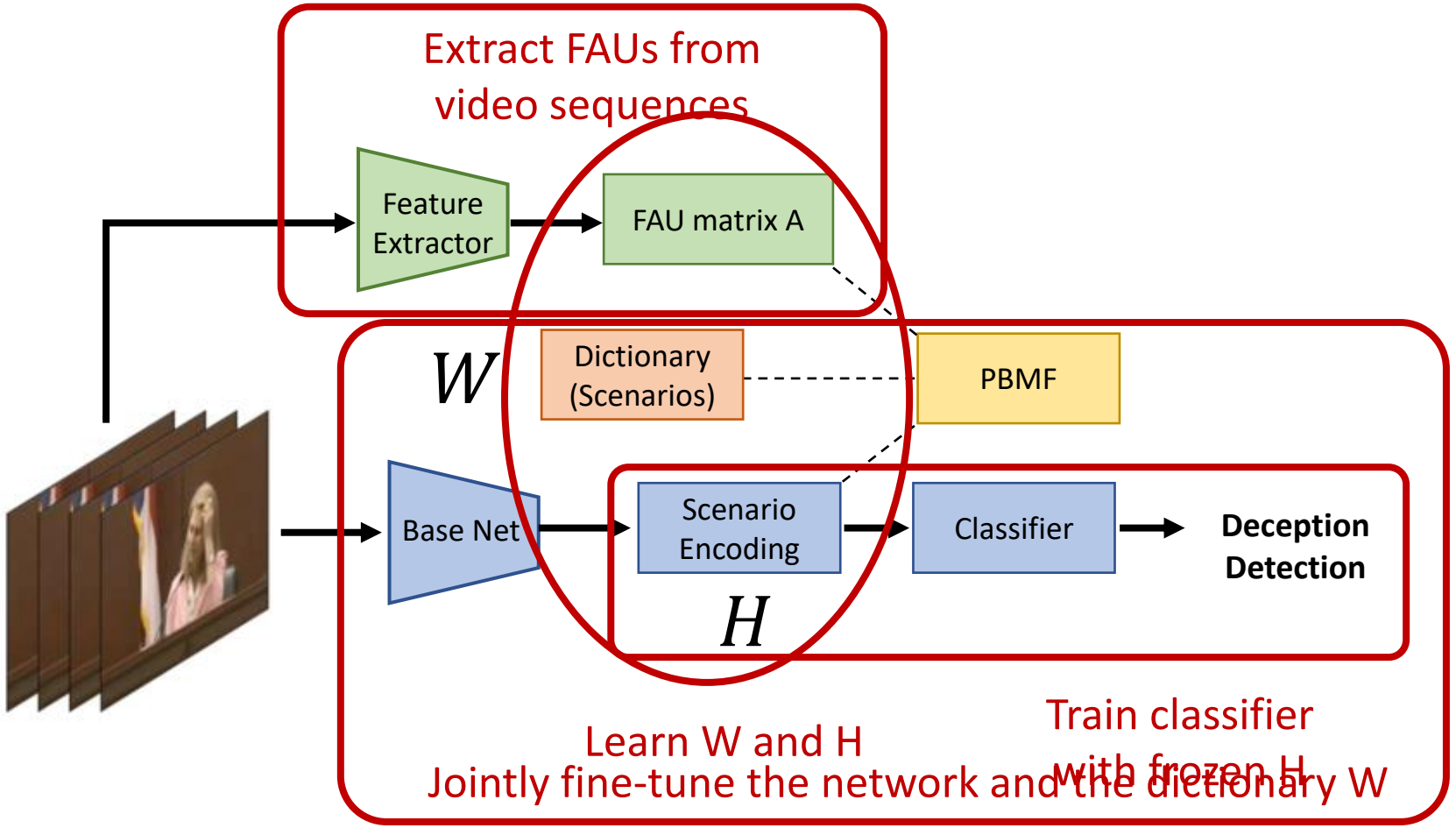
Part 2: ScenarioNet

- We adapt ScenarioNet [6] for deception detection
- **Scenarios:** learned groupings of FAUs



[6] Daniels and Metaxas, **ScenarioNet: An Interpretable Data-Driven Model for Scene Understanding**, IJCAIW 2018

Part 2: ScenarioNet



Part 2: Pseudo-Boolean Matrix Factorization (PBMF)

- We create a binary matrix A that contains extracted FAUs from input videos
- A can be factorized to H and W

$$\min_{W,H} ||(A - W \circ H)||_1 \; s.t. \; W \in \{0,1\}, H \in \{0,1\}$$

- To solve the optimization problem we use a gradient-based approach and optimize

$$\begin{aligned} &\min_{W,H} ||(A - \min(WH, 1 + 0.01WH))||_F^2 \\ &+ \alpha_1 ||W^\intercal W - diag(W^\intercal W)||_F^2 + \alpha_2 ||W||_1 + \alpha_3 ||H||_1 \\ &s.t. \; W \in [0,1], H \in [0,1], \end{aligned}$$

where we approximate $W \circ H \approx \min(WH, 1 + 0.01WH)$.

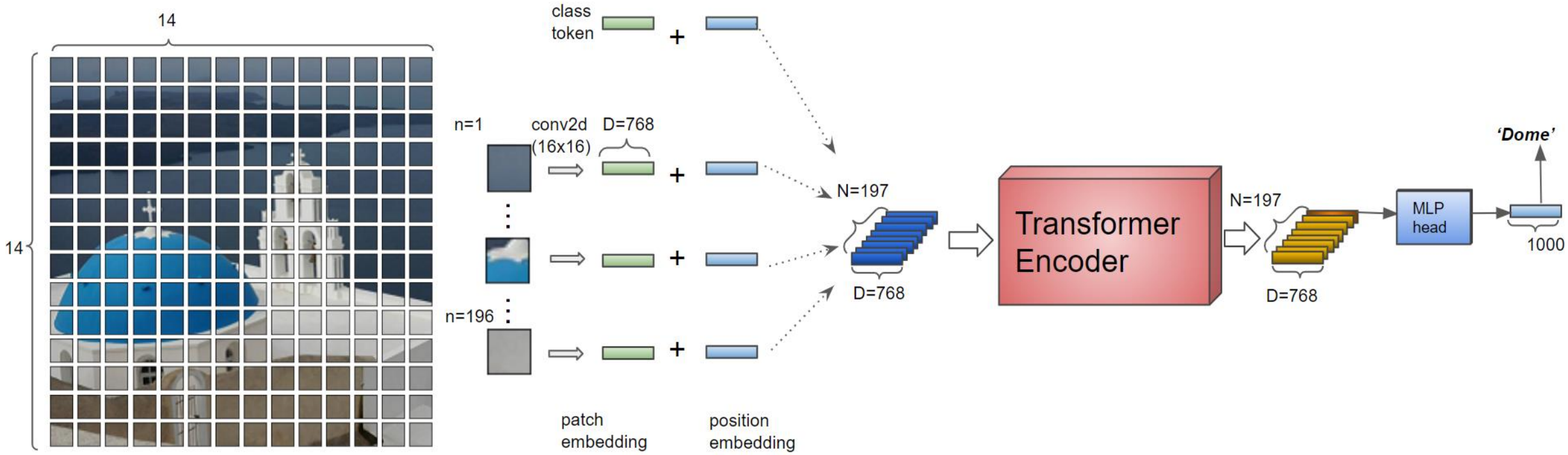
Part 2: Results

Improvement due to nonlinear basis based on PBMF which results in basis discovery and nonlinear combinations of FAUs

- Predicting from nonlinear FAU feature combinations makes the model **interpretable** and **robust**
- Can also augment the vocabulary when we see new cases which we can recognize if their combination of FAUs is not representable
- This approach results in 73% classification accuracy performance
- To push the envelope in deception detection we augment the approach for multimodality explainable prediction with Transformers [7,8].

[7] Vaswani et al., **Attention Is All You Need**, NeurIPS 2017
[8] Dosovitskiy et al., **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**, ICLR 2021

Part 3: Vision Transformer (ViT)



[8] Dosovitskiy et al., **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**, ICLR 2021

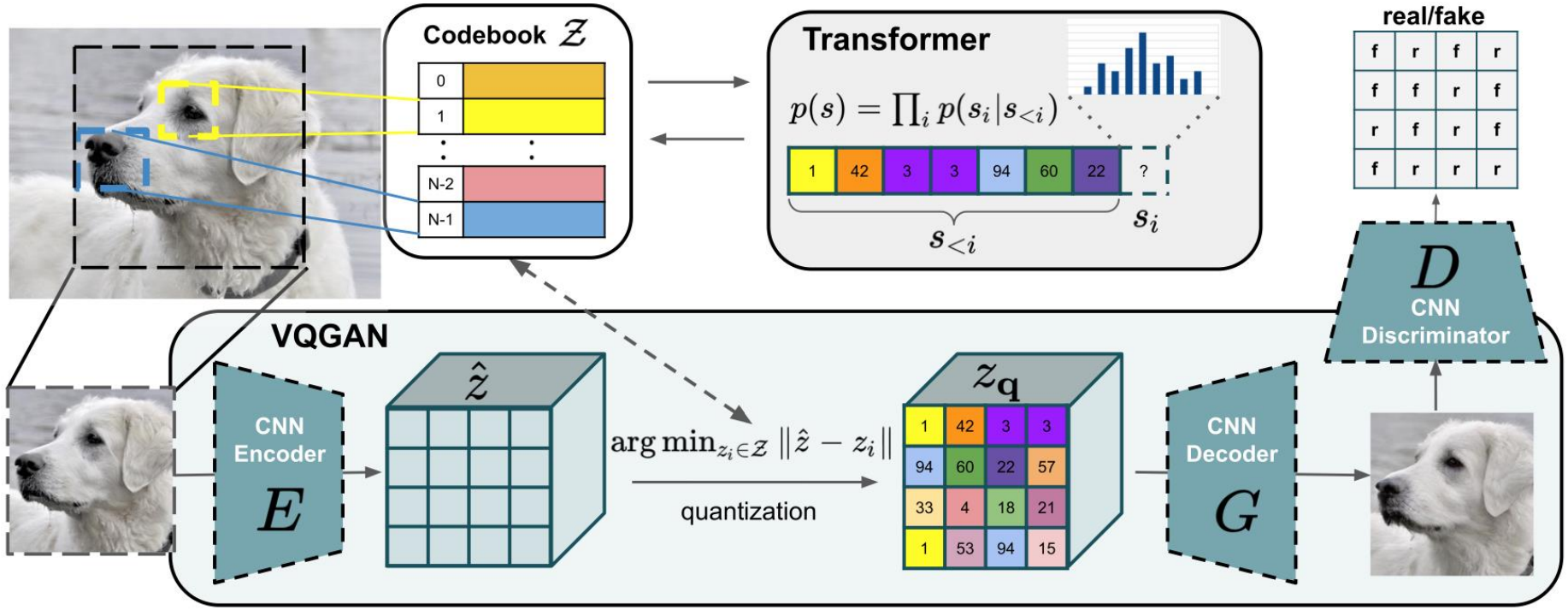
Part 3: VQ-GAN + Transformer

- We propose to use transformer-based models for detection of deceptive behavior
- We propose to use video self-supervision to learn a condensed representation of the input video sequences
- We can also use the same model for video generation conditioned on a specific behavior (deceptive/non-deceptive), which lets us analyze specific patterns present in deceptive behavior
- Our goal is to make the prediction of the model more robust through generative joint training [12]

[9] Esser et al., **Taming Transformers for High-Resolution Image Synthesis**, CVPR 2021
[12] Grathwohl et al., **Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One,** ICLR 2020

Part 3: VQ-GAN

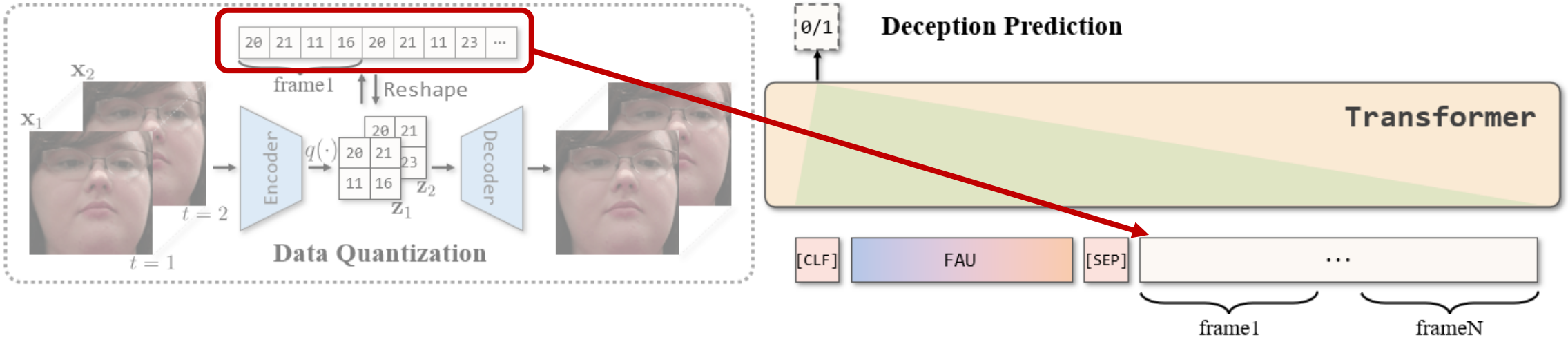
VQ-GAN [9] improves upon VQ-VAE [10] by using an adversarial and a perceptual loss



[9] Esser et al., **Taming Transformers for High-Resolution Image Synthesis**, CVPR 2021
[10] Oord et al., **Neural Discrete Representation Learning** , NeurIPS 2017

Deception Detection using (VQ-GAN + Transformer)

- We quantize the input frames
- We incorporate FAUs as an extra modality



Part 3: Results

- *Preliminary results:* we use the dataset used before in our publication [11].
- *Ongoing*: we are processing new videos extracted from critical decision points annotated by UCSB team

Method	Accuracy (%)
C3D [11]	65.4
VQ-GAN + Transformer	74.4

[11] Wang et al., Attention-based Facial Behavior Analytics in Social Communication, BMVC 2019

Part 3: Self-Attention Visualization from Transformer

- We visualize the self-attention maps of the model for players exhibiting deceptive behavior
- The model mostly focuses on specific facial micro-expressions such as **rapid eye movement** and **blinking**, which is consistent with current communication theory for deceptive behavior

Part 3: Self-Attention Visualization from Transformer

- The model mostly focuses on

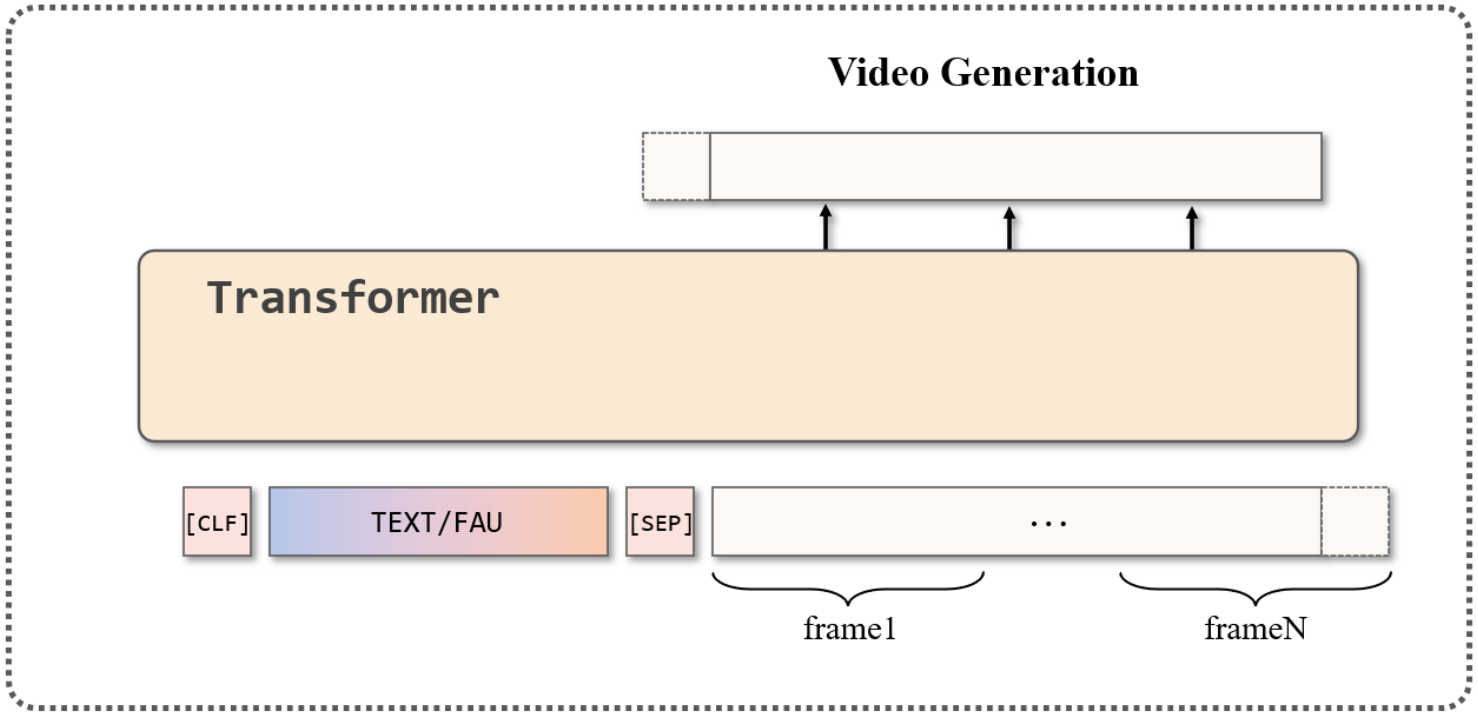
Lip Corner Pulling
Forehead Movement

Fake Smile

Blinking

Part 3: Video Generation

- With the same model we can perform video generation conditioned on a specific behavior
- We can also condition on other modalities such as detected FAUs



Part 3: Video Generation

- We generate video clips using the following text prompt:
  - *"This person is making a deceptive expression"*

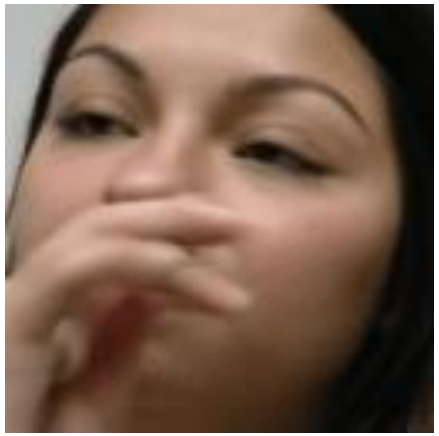
- We analyze the generated video clips: e.g., we can study several patterns of deception that are captured by our model

- Next, we visualize input as well as reconstructed and generated clips

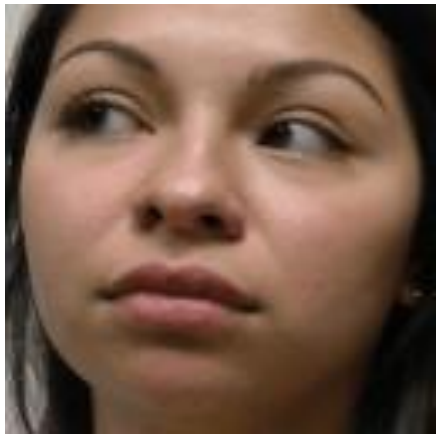
Part 3: Generated Videos

Input clip

Reconstructed clip



Generated clip 1



Generated clip 2

Part 3: Generated Videos

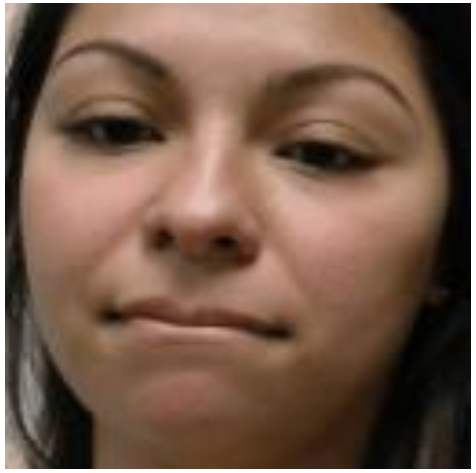
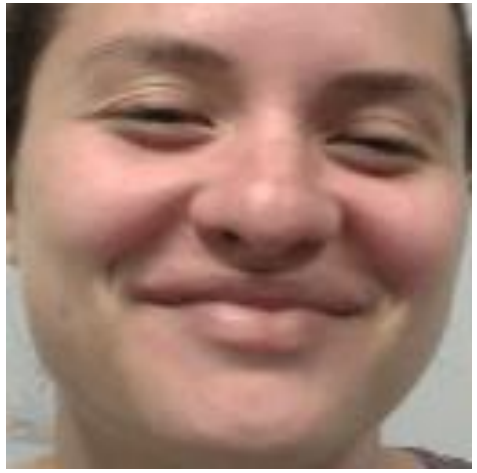
Input clip

Reconstructed clip

Generated clip 1

Generated clip 2

Additional Generated Videos

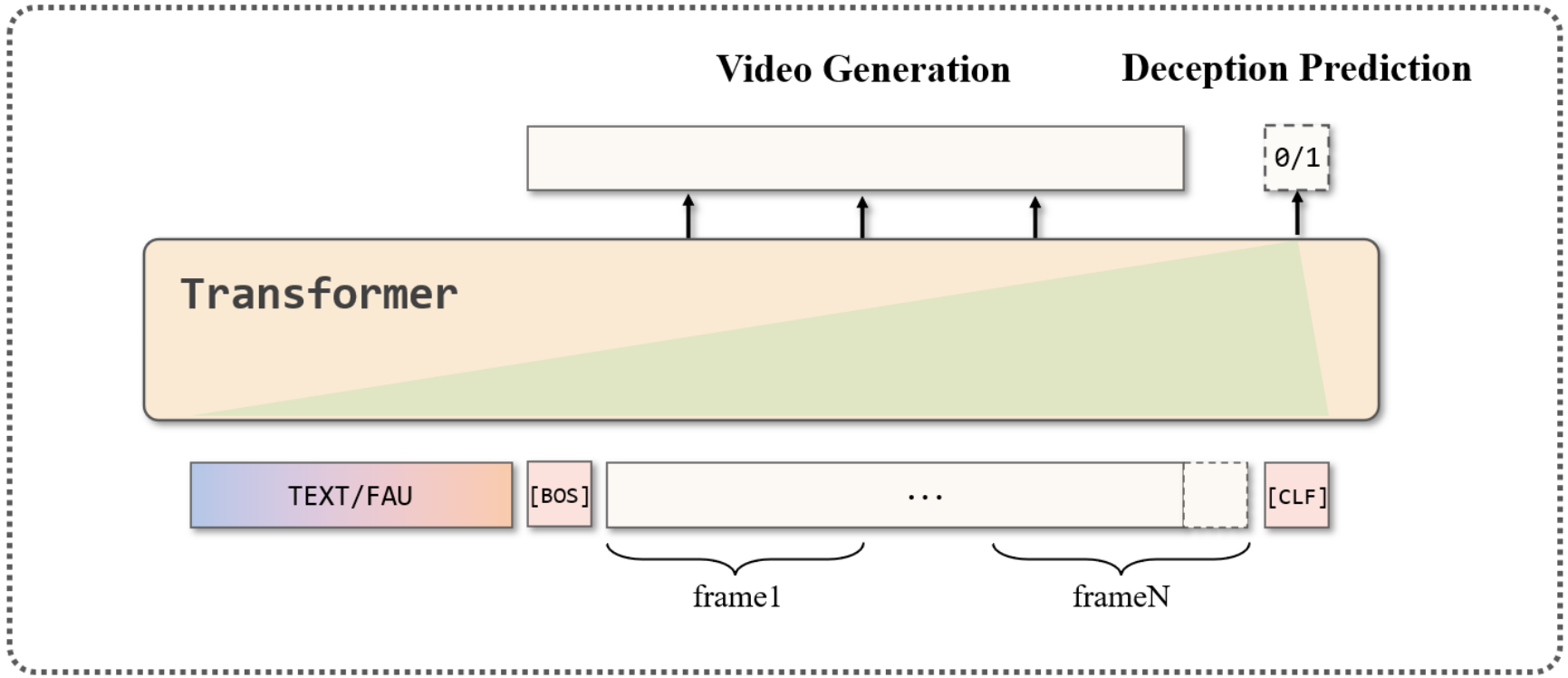

We have a *submitted* manuscript to CVPR'22 about video generation:
"Show me what and tell me how: Video synthesis via multi-modal conditioning"

Ongoing Work

A. Joint Generative and Discriminative Learning

- We investigate the benefit of jointly training the model for generative (clip generation) and discriminative learning (deception detection)

B. Rigorous testing on Deception, Trust and Liking Collaboratively with MURI Team



Summary

- We analyzed facial expressions as a predictor of trust, dominance and liking
  - Players who were more dominant, likable and trusted had more intense facial expressions

- We proposed a transformer-based approach for joint generative and discriminative training for deception detection
  - We use video self-supervision to alleviate overfitting to small datasets

- We analyzed deceptive behaviors from self-attention visualization and generated videos
  - Specific patterns of deceptive behavior emerge that are consistent with findings in communication theory

- Our approach emphasizes :
  - Explainability and Understanding
  - Representations are motivated from communication theory

PhD Students

- Anastasis Stathopoulos
- Ligong Han
- Yu Tian (ByteDance Research)
- Long Zhao (Google Research)

Publications

- In collaboration with the MURI team:
  - Wang, L., Bai, C., Bolonkin, M., Burgoon, J.K., Dunbar, N.E., Subrahmanian, V.S. and Metaxas, D., "Attention-based facial behavior analytics in social communication". In *Detecting Trust and Deception in Group Interaction* (pp. 123-137). Springer, Cham, 2021
  - Burgoon, J.K., Metaxas, D., Nunamaker, J.F. and Ge, S.T., "Cultural Influence on Deceptive Communication". In *Detecting Trust and Deception in Group Interaction* (pp. 197-222). Springer, Cham, 2021
  - Stathopoulos, A., Han, L., Dunbar, N., Burgoon, J.K. and Metaxas, D., "Deception Detection in Videos Using Robust Facial Features". In *Proceedings of the Future Technologies Conference (FTC),* 2020 ***(Best Student Paper Award)***
- Other publications:
  - Han, L., Musunuri, S.H., Min, M.R., Gao, R., Tian, Y. and Metaxas, D., "AE-StyleGAN: Improved Training of Style-Based Auto-Encoders". In *IEEE Winter Conf. on Applications of Computer Vision (WACV),* 2022
  - Zhao, L., Zhang, Z., Chen, T., Metaxas, D. and Zhang, H., "Improved Transformer for High-Resolution GANs". In *Advances in Neural Information Processing Systems (NeurIPS),* 2021
  - Han, L., Min, M.R., Stathopoulos, A., Tian, Y., Gao, R., Kadav, A. and Metaxas, D.N., "Dual Projection Generative Adversarial Networks for Conditional Image Generation". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 2021
  - Zhao, L., Wang, Y., Zhao, J., Yuan, L., Sun, J.J., Schroff, F., Adam, H., Peng, X., Metaxas, D. and Liu, T., 2021. Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021 ***(Oral)***
  - Metaxas, D.N., Zhao, L. and Peng, X., "Disentangled Representation Learning and Its Application to Face Analytics". In *Deep Learning-Based Face Analytics* (pp. 45-72). Springer, Cham, 2021
  - Zhao, L., Peng, X., Tian, Y., Kapadia, M. and Metaxas, D.N., 2020. "Towards Image-to-Video Translation: A Structure-Aware Approach via Multi-stage Generative Adversarial Networks". In *International Journal of Computer Vision (IJCV),* 2020

References

[1] Tran et al., **Learning spatiotemporal features with 3d convolutional network**, ICCV 2015

[2] Lea et al., **Temporal convolutional networks: A unified approach to action segmentation**, ECCV 2016

[3] Ojala et al., **Multiresolution gray-scale and rotation invariant texture classification with local binary patterns**, PAMI 2002

[4] Wang et al., **Temporal Segment Networks,** ECCV 2016

[5] Selvaraju et al., **Grad-Cam: Visual explanations from deep networks via gradient-based localization,** ICCV 2017

[6] Daniels and Metaxas, **ScenarioNet: An Interpretable Data-Driven Model for Scene Understanding**, IJCAIW 2018

[7] Vaswani et al., **Attention Is All You Need**, NeurIPS 2017

[8] Dosovitskiy et al., **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**, ICLR 2021

[9] Esser et al., **Taming Transformers for High-Resolution Image Synthesis**, CVPR 2021

[10] Oord et al., **Neural Discrete Representation Learning** , NeurIPS 2017

[11] Wang et al., **Attention-based Facial Behavior Analytics in Social Communication**, BMVC 2019

[12] Grathwohl et al., **Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One,** ICLR 2020

Contact Information

Dimitris N. Metaxas
Director CBIM Center
Department of Computer Science
Rutgers University
dnm@cs.rutgers.edu
www.cs.rutgers.edu/~dnm
Tel: 848-445-2914

Analysis and Discovery of Deceptive Behavior

Abstract

In this work, we focus on the problem of deception detection from non-verbal facial cues. We propose two new methods for discovery and analysis of deceptive behavior, improving on previously proposed approaches. First, we adapt ScenarioNet [2] for deception detection. ScenarioNet is a method proposed by our group for scene understanding using compositional and interpretable representations. Second, we propose a transformer-based model. The input to the transformer model is a condensed representation of the input video sequences that we learn using video self-supervision. In addition, we can use the same model for video generation conditioned on a specific behavior (deceptive/non-deceptive), which lets us analyze specific patterns present in deceptive behavior.

1 Introduction

Deception is a dynamic process that can be manifested in various forms and is hard to detect even by domain experts. Relevant studies [11, 1, 8] suggest that during deception there is leakage of certain emotions. Leakage in deception is manifested most overtly in nonverbal signals and mainly from one's facial expressions. For this reason, we propose to build systems for automatic deception detection based on facial cues.

Our previous works on deception detection [10, 7] approach the problem in a two-stage manner. First, we train a model for Deceiver vs Truth-Teller classification using an input video. Then, we use a gradient-based localization method, similar to Grad-CAM [6], to localize deception in time as suggested by our model. We study the facial micro-expressions in the localized frames to analyze deceptive behavior.

This year we go beyond those methods by proposing two approaches that do not need any post-processing for localizing the frames where deception is manifested. We adapt ScenarioNet [2] – a method proposed by our group for interpretable scene understanding – for deception detection. Additionally, we propose a transformer-based model [9], which can incorporate input from different modalities (e.g. RGB frames, FAUs) with no architectural modification.

2 Deception Detection with ScenarioNet

Our goal is to learn FAU groupings – which we call scenarios – from the data that will help us for detecting deceptive behavior. For each training instance (input video sequence), we create a vector with FAU presence over time. We concatenate these vectors to form a matrix A where each row corresponds to a specific FAU at a specific time step and each column is a training instance. After specifying the number k of desired scenarios, we decompose A into a dictionary matrix W representing a set of scenarios and an encoding matrix H that expresses input videos as combinations of scenarios.

We identify scenarios using an approximation of Boolean Matrix Factorization (BMF) [5] using Eq. 1

$$\min_{W,H} ||(A - W \circ H)||_1 \; s.t. \; W \in \{0,1\}, H \in \{0,1\} \tag{1}$$

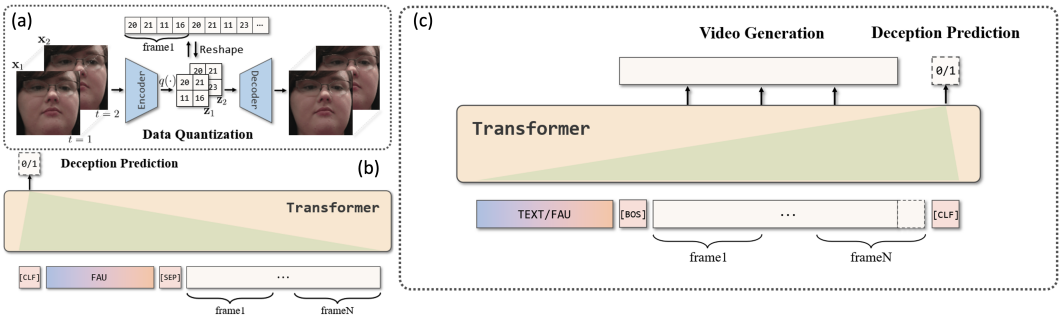


Figure 1: **Proposed Framework of Learning Discrete Neural Representations for Deception Detection.** (a) Video quantization with VQGAN [3] as Stage I, we finetune a VQGAN on all frames given a video dataset and quantize the video clips framewise and serialize the codes via `Reshape`; (b) A baseline model for BERT-like discriminative task, i.e., we prepend a `CLF` token to the input sequence and train a bidirectional transformer with binary cross-entropy; (c) The proposed discriminative and generative joint training framework, where `CLF` is appended at the end of input sequence since next-token-prediction is used for generative joint training and causal attention masks are applied to the transformer at every block.

where A, W, and H are binary matrices and $\circ$ denotes the boolean matrix multiplication operation. We solve the optimization problem with gradient descent, so we approximate $W \circ H \approx \min(WH, 1) \approx \min(WH, 1 + 0.01WH)$. Our final objective becomes

$$\begin{aligned} &\min_{W,H} ||A - \min(WH, 1 + 0.01WH)||_F^2 + \alpha_1 ||W^\intercal W - diag(W^\intercal W)||_F^2 \\ &+ \alpha_2 ||W||_1 + \alpha_3 ||H||_1 \qquad s.t.\ W \in [0, 1], H \in [0, 1] \end{aligned} \tag{2}$$

where orthogonality and sparse penalties are included in the optimization.

The overall procedure of training ScenarioNet can be summarized as follows. First, the scenario dictionary is learned using extracted FAUs from input videos. Then, a neural network is trained to predict the scenario encodings while the dictionary is fine-tuned. Next, we train a classifier on top of a frozen network. Finally, we jointly fine-tune the network for scenario recognition while once again fine-tuning the dictionary.

3 Deception Detection using Discrete Neural Representaitons

To further improve the expressiveness and extendability of the model, in this section, we propose to use transformer-based models for detection of deceptive behavior. Specifically, we propose to use video self-supervision to learn a condensed representation of the input video sequences. With the help of the recent development of autoregressive modeling, we can also use the same model for video generation conditioned on a specific behavior, e.g., deceptive or non-deceptive, which lets us analyze specific patterns present in deceptive behavior. As such, we hope to make the prediction of the model more robust through generative joint training [4].

The proposed frameworks are shown in Figure 1, (a) illustrates the data quantization pipline: we finetune a VQGAN [3] on all frames given a video dataset and quantize the video clips framewise, and serialize the codes via `Reshape` operation. Figure (b) shows a baseline model for BERT-like discriminative task, i.e., we prepend a `CLF` token to the input sequence and train a bidirectional transformer with binary cross-entropy. In order to regularize the discriminative task with joint generative training, we append the `CLF` token at the end of input sequence (as shown in Figure 1 (c)). This is because a simple next-token-prediction task is used for the generative training, and causal attention masks are applied to the transformer at every block. For the joint model, bidirectional transformers can also be used (in this case causal masking is disabled) if we replace next-token-prediction with masked-language-modeling task as in BERT (in this case the generation process will be a mask-predict-like algorithm). We leave this for future work.

4 Conclusion

To conclude, we proposed two new methods for deception detection. ScenarioNet which is an interpretable approach to deception detection that results in basis discovery and non-linear combinations of FAUS. We further proposed a transformer-based approach for joint generative and discriminative training for deception detection, where we use video self-supervision to alleviate overfitting to small datasets.

References

[1] David B Buller and Judee K Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.

[2] Zachary A Daniels and Dimitris Metaxas. Scenarionet: An interpretable data-driven model for scene understanding. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2018*, 2018.

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[4] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[5] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, and Heikki Mannila. The discrete basis problem. *IEEE transactions on knowledge and data engineering*, 20(10):1348–1362, 2008.

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[7] Anastasis Stathopoulos, Ligong Han, Norah Dunbar, Judee K Burgoon, and Dimitris Metaxas. Deception detection in videos using robust facial features. In *Proceedings of the Future Technologies Conference*, pages 668–682. Springer, 2020.

[8] Leanne Ten Brinke and Stephen Porter. Cry me a river: identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior*, 36(6):469, 2012.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[10] Lezi Wang, Chongyang Bai, Maksim Bolonkin, Judee Burgoon, Norah Dunbar, VS Subrahmanian, and Dimitris N Metaxas. Attention-based facial behavior analytics in social communication. In *30th British Machine Vision Conference, BMVC 2019*, 2019.

[11] Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier, 1981.