Building Secure and Engaging Video Communication by Using Monitor Illumination

Jun Myeong Choi, Johnathan Leung, Noah Frahm, Max Christman, Gedas Bertasius, Roni Sengupta
University of North Carolina at Chapel Hill.

Abstract

In this paper, we develop a neural network that can detect a mismatch between the light emitted from a monitor and the light reflected from the face of a user sitting in front of a monitor-webcam setup. This can be useful to detect the presence of a deep fake virtual avatar or an inattentive attendee to create a secure and engaging virtual communication platform, e.g. a student in a virtual education environment. We can perform this detection passively, without requiring the authenticator to project any specific patterns intermittently on the screen, hence it does not disrupt the meeting flow or alert the bad actors. We develop a personalized model, where the authenticator requires each team member to watch ∼30 minutes of video content, only once, on their monitor as their faces are captured with a webcam. We then train a neural network that learns to predict the monitor content from their facial image and compares it with the intended monitor content to detect 'on-task' (real) vs 'off-task' (fake). This personalized network can then detect 'off-task' scenarios, where monitor lighting does not match the face, for any unseen user appearances. Our method produces a binary classification accuracy of 70%, surpassing a baseline that always predicts 'on-task' with 58% accuracy.

1. Introduction

In recent years we have observed a strong continued growth in virtual communication for conducting business, educational, and personal activities across different geographical locations through video calls and live streaming. Virtual communication offers accessibility, cost-effectiveness, flexibility, productivity, collaboration, global reach, environmental sustainability, improved work-life balance, remote learning opportunities, and innovation. Yet, there has been a growing concern in recent years about the insufficiency of security measures [6, 8] and the diminished levels of engagement [20, 37], posing obstacles to opting for virtual communication over in-person interactions, despite its numerous advantages.

A new threat concerning the security of video calls is the emergence of powerful deep fake avatars, where a malicious actor can manipulate video content to deceive or impersonate individuals. For example, consider this recent incident, where a finance worker at a multinational firm was tricked into paying $25 million to a fraudster posing as the company's chief financial officer in a video conference call using deepfake technology [3]. In the future, the emergence of powerful virtual talking head avatars (e.g. recent tools like github.com/iperov/DeepFaceLive and github.com/alievk/avatarify-python) will likely create many such catastrophic incidents.

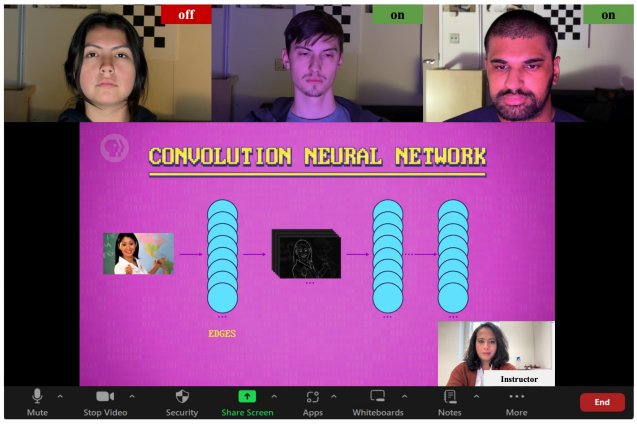


Figure 1. We utilize the mismatch between the lighting emitted from the monitor and reflected from the face to detect a deep fake avatar or inattentive participant. We train a personalized neural network that can predict the content of the monitor for each participant and compare it with the intended content presented by the authenticator, e.g. lecture slides presented by the instructor, to predict whether they are on-task (in green), i.e. real humans watching the presentation, or off-task (in red), i.e. deep fake avatars or inattentive participant watching something else.

Similarly, online education in the form of virtual lectures and group discussions suffers from a lack of engagement. Understanding student engagement helps instructors design better lectures and makes it easier to identify inattentive students. While an instructor can easily infer student engagement and attentiveness when teaching a class in person, it is extremely difficult when teaching online [20, 37]. Even with stricter enforcement like turning on their webcam videos, students may often be disengaged, watching

YouTube videos, playing games, browsing social media, etc., while fixing their gaze on the monitor and sitting in an attentive posture.

Both deep fake virtual avatars and students disengaging from class content often share one underlying physical phenomenon, i.e. the lighting on their faces does not match the lighting emitted from the monitor. In this paper, we aim to create a secure and engaging virtual communication platform by detecting deep fake actors and disengaged participants with the mismatch between screen illumination and facial lighting. We can authenticate real and engaged users by simply using the content being currently shared on the screen as a part of the virtual meeting, without projecting any specific patterns on the monitor that disrupt the meeting flow and alert malicious actors. We consider a scenario, where an authenticator, e.g. an instructor or a team leader, is sharing specific contents during the virtual meeting that every attendee is supposed to watch, e.g. lecture slides or monthly project report. If an attendee is a real human who is paying attention and watching the presentation, the lighting emitted from the monitor will reflect from their face, we call this scenario 'on-task'. If they are a deep fake avatar or an inattentive attendee, e.g. a student browsing social media instead of watching the virtual lecture, the lighting on their face will not match the content shared by the authenticator, we call this 'off-task'.

We design a neural network that takes a portrait image of the user captured by their webcam as input and generates the 'predicted monitor' content in low resolution, compares it to the 'intended content' shared by the authenticator, and predicts a binary classification of 'on-task' vs 'off-task'. The neural network consists of a pre-trained backbone (ResNet or ViT), a generator (convolution or pre-trained StyleGAN) that outputs a predicted monitor, a mapping network that maps backbone features to the generator, and a binary classifier. The binary classifier predicts attention, on-task or off-task, from the 'predicted monitor' and the 'intended monitor'.

In this paper, we train personalized detection models for each user separately. We assume that the authenticator, i.e. manager or instructor, requires their team or class to watch specific videos (YouTube videos) on their monitor for $\sim$30 minutes and record them with a webcam. This can be done once when a new employee joins a team or at the beginning of each term for the students. Then personalized AI models will be trained on these captured videos and will be applied to unseen appearances of the same user. We argue that training a personalized model is easier than developing a generalized model that requires capturing hundreds of participants and can pose more privacy concerns.

Our work is closely related to [11], which also uses illumination difference between monitor and face to detect a deep fake. However, the difference is that our approach uses passive illumination in contrast to active illumination used in [11], which requires the authenticator to share specific contents, i.e. short video of distinct time-varying patterns, intermittently. Our method uses whatever content the authenticator shares on screen as a part of the meeting, e.g. lecture slides, hence does not disrupt the meeting flow and alert malicious actors.

We test our idea by developing personalized detection models for 4 participants of different skin types (Types I through IV) captured under different illumination conditions, head motion and expression variations. From detailed qualitative and quantitative evaluations, our key observations are: (i) Our best model – which uses multiple layers of features from a pre-trained ViT encoder and a StyleGAN generator – obtains an average F1 score of 75% and an accuracy of 69% across all participants in different environments. (ii) Multi-task learning, where we jointly predict actual monitor content and perform binary classification, improves by $\sim$3% (accuracy) over only performing binary classification. (iii) The performance of our best model deteriorates from small head motion to large head motion by 4.84% (accuracy) but is not impacted by ambient room lighting. (iv) Our approach does not work when a strong source of illumination, e.g. a bright lamp, is placed very close to the participant's face nullifying the effect of lighting emitted by the monitor.

In summary, the contributions of this paper are:

- We show that light emitted from the monitor and reflected from the face is an effective signal to detect the presence of deep fake avatars or inattentive attendees in video conference calls. We show that this detection can be done passively without requiring the authenticator to share specific patterns intermittently, which disrupts the meeting flow and alerts the bad actors.
- We introduce a general neural network framework where we extract features from the face and use them to predict the student's monitor content. We then use the predicted monitor and the intended monitor to perform binary classification for 'on-task' (real) vs 'off-task' (fake). We show that jointly predicting monitor and classifying student attention improves accuracy by 3% over only classifying attention.
- We perform an extensive quantitative and qualitative evaluation on four users with varying skin tones (Type I to IV) analyzing the impact of different architectural choices, head motion, and background ambient lighting.

2. Related work

Exploiting Screen Illumination for Relighting. Prior research in computer vision has often considered using monitors [5, 39] or projector screen [29] as light sources to enable relighting of static objects. Recently this has been extended to facial portrait relighting for video calls with lighting emitted from the monitor [4, 30]. Our method is also

inspired by these approaches, but instead of relighting we use screen illumination for deep fake detection. We further show in Sec. 5.2 how existing relighting techniques [4, 30, 32] can be used by students to prevent the instructor from monitoring engagement using monitor lighting.

Deepfake detection. The rise of generative image and video editing tools led to the proliferation of deepfakes raising concerns regarding misinformation, privacy breaches, and the potential to deceive or manipulate audiences on a large scale. Hence, researchers have developed various AI algorithms that focus on detecting deep fakes, a survey of these techniques are presented in [21, 35]. However, most existing papers focus on detecting deep fakes in images, videos, and audio offline, i.e. the generated content already circulating in the media. Only recently, live deep fakes, synthesized in real-time by a virtual camera has become popular, e.g. tools like github.com/iperov/DeepFaceLive and github.com/alievk/avatarify-python). Soon in the future, these tools will become even more powerful posing a real threat to live video communication platforms that were traditionally considered 'believable' to humans.

Recently researchers have explored active illumination projected on the user's screen to detect deep fakes. LiveScreen [22] uses screen illumination reflected from the skin to detect liveness. Similarly, researchers used active illumination probing on corneal reflection [12] to detect deep fakes. Both Gerstner *et al.* [11] and Shang *et al.* [31] use active screen illumination to project-specific patterns for authentication. While [31] uses a simple change in the brightness of the screen, it is not robust to real webcams with built-in auto exposure. In contrast [11] relies on hue variation that is independent of the auto-exposure. Our method does not require any active illumination and simply utilizes natural change in intensity and chromaticity of the content the user is viewing for authentication.

Active vs Passive Screen Illumination. Our work is closely related to [11, 31], where the author uses active monitor illumination to authenticate videos as real or 'deepfaked'. The authors assume that a call participant can project a distinct temporally varying illumination pattern on a shared screen and authenticate other user videos based on the reflection of this pattern on their skin. The active illumination approach can be used at the beginning of the meeting by an instructor to validate whether a participant's video is real or fake. However, using active illumination during the meeting disrupts the meeting flow and alerts the bad actors. Our proposed approach does not require any active illumination, i.e. creating specific content for authentication, and can passively authenticate with the current content being presented on the screen.

Online student engagement prediction Researcher have used various different signals for predicting student engagement in virtual learning platforms, e.g. physiological signals from sensor readings [9, 23, 24, 33], browsing patterns and mouse movements [2, 18, 19]. These techniques often require active participation by requiring students to put on wearable devices and consent to enable remote access to their devices.

Thus often researchers have relied on eye gaze [1, 15, 17, 25–27, 34, 38] as a passive measure of engagement in virtual learning environments. However, uncalibrated eye gaze tracking techniques based on deep learning often lack fine-grained angular precision and are primarily used to tell if a student is paying attention to the monitor. In many practical scenarios, inattentive students simply perform different tasks on their monitor, e.g. browsing social media or watching YouTube videos, while faking an attentive posture and gaze. Our proposed approach is particularly focused on these scenarios where gaze estimation does not indicate student engagement. Instead, we exploit the reflections of the monitor light from the person's face as a visual cue for predicting engagement.

Another visual cue often used for predicting engagement involves detecting facial expressions [13, 36] or full-body pose [10]; However, like eye gaze, this prediction method can be easily fooled with faked facial expressions.

3. Problem formulation

In this work, we detect whether the screen illumination matches the face lighting or not during live video calls. We assume we can detect this passively from the already present content on the screen in contrast to existing approaches [11, 31] that require the authenticator to share specific patterns, i.e. active illumination, which disrupts the meeting flow and alerts 'bad actors'. We develop an AI algorithm that can analyze the light reflected from the face and infer the nature of the monitor content. This allows us to predict whether the 'predicted monitor' content matches the 'intended monitor', which is currently on-screen, to detect whether the face is a deep fake 'bad actor' in live video calls or a student not engaged with the class content.

On-task vs. Off-task classification. We assume the authenticator, i.e. either the person in charge of conducting the meeting or an instructor in online education, is sharing content for everyone to view, e.g. slide deck or meeting notes. We term this content as 'intended monitor' content which should be viewed by everyone else in the meeting. We define 'on-task' as a situation where the light emitted from the screen content, i.e. 'intended monitor' matches the light being reflected from the face of the person. This indicates that the person is viewing the 'intended' content and is a real live person and not a deepfake. Similarly, we define 'off-task' as a situation where the light emitted from the monitor does not match the light being reflected from the face, indicating either the presence of a deep fake avatar or a person disengaged with the 'intended monitor' content, e.g. a student browsing social media instead of watching

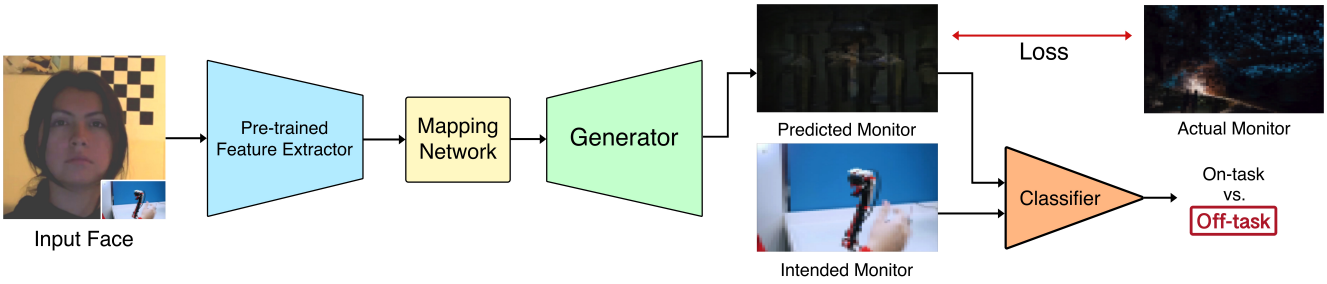


Figure 2. We introduce a neural architecture that extracts features from the face and then maps them to a latent space using a mapping network. We then use a generator to generate predicted monitor contents from the latent space. Finally, we use a classifier that takes the predicted monitor and the intended monitor as inputs and predicts a binary classification of either on-task or off-task. We train this neural architecture jointly with binary classification loss and a loss between the predicted and actual monitor. In the given example, the participant is 'off-task', i.e. lighting on the face does not match monitor lighting. This diagram represents models with task classifier TC2. For models with TC1, the output is passed directly from the mapping network to the task classifier and the generator is not present.

the lecture. Our method predicts whether each user is 'on-task' or 'off-task' for every frame of their webcam video in real-time (at 30 fps).

Personalized models for predicting engagement. In this paper we focus on developing a personalized algorithm that detects on or off task conditions, by training a neural network only on the images of a user and testing on unseen appearances of the same user. We build personalized models instead of generalized models since the latter would require capturing a few hundred individuals in front of a monitor, which takes significantly more time and resources. Thus we will rely on personalized models to prove the effectiveness of passive monitor lighting in deep fake detection. We train separate neural networks for four individuals of different genders, ethnicities, and skin colors.

To train a personalized model, we only require users to watch ∼30 minutes of arbitrary video content on their monitor while capturing their faces with a webcam. We argue that this is accessible enough to make the development of personalized models practical for virtual meetings. For example, before the start of a semester, each student could be required to watch 30 minutes of video content and send their webcam recordings to the instructor where they could build personalized models for each student. Similarly, a manager of a small organization can also build a personalized model for their employees when they join the organization to authenticate their presence in live video calls.

Collecting training data for a personalized model. For 'on-task' we capture real humans watching YouTube videos on a monitor screen. For each user, we record them watching 4 different videos at different times. Each video is roughly 8 minutes in length. We use the monitor content as both an 'intended monitor' and an 'actual monitor' of what they are watching. This capture data can be collected by a manager of a team or an instructor of a class at the beginning of the year. In our formulation, they do not need to capture separate videos for 'off-task' but simply simulate 'off-task' scenarios from 'on-task' captured data. This makes it significantly easier to train the models without require any separate capture for 'off-task'.

For 'off task', we need to create a scenario where the screen lighting and face lighting do not match, i.e. the 'intended monitor' content does not match the content on the screen being watched by the user 'actual monitor'. This allows us to simulate both live deep fakes and disengaged participants scenarios. We first divide the 8-minute captured video into short random chunks. For each chunk, we consider them either to be 'on-task' by using the 'actual monitor' content = 'intended monitor' content or we will create 'off-task' labels by simply using a random video as the 'intended monitor'. We ensure that each chunk is used as both 'on-task' and 'off-task' during the training and validation process. Figure 3 explains this data collection process.

Since, in most practical scenarios slide contents are presented during virtual meetings we aim to mimic that by considering the 'intended monitor' as slideshows in our test videos. We created an 8-minute test video by randomly replacing short chunks of a lecture slide video with slideshows with a YouTube video. Here, the original lecture video will be the 'intended monitor' content and the test video (a mix of lecture slides and YouTube clips) will be the 'actual monitor' content that the user watches.

Capture details. We use 5 monitor videos to capture 4 users of varying skin tones. We capture these 4 users in 4 distinct conditions to simulate different real-world scenarios: (A0-M0) a dimly lit room, small pose and expression changes; (A0-M1) a dimly lit room, larger pose and expression changes; (A1-M0) a room with ambient lighting, small pose and expression changes; (A1-M1) a room with ambient lighting, larger pose and expression change.

4. Method

We aim to extract lighting representation from a person's portrait image and determine if that matches the lighting emitted from the screen. Our model consists of: a **feature extractor**, which is a pre-trained model that extracts

Model	Backbone	Backbone Output	Generator	Mapping Net	Classifier	Total Loss
RX	ResNet	Final layer	–	–	TC1	L_{on}
RC	ResNet	Final layer	Conv	–	TC2	$L_{\text{on}} + L_{\text{moni}}$
RS1	ResNet	Final layer	StyleGAN	✓	TC2	$L_{\text{on}} + L_{\text{moni}} + L_w$
RS2	ResNet	Multiple layer	StyleGAN	✓	TC2	$L_{\text{on}} + L_{\text{moni}} + L_w$
VX	ViT	Final layer	–	–	TC1	L_{on}
VC	ViT	Final layer	Conv	–	TC2	$L_{\text{on}} + L_{\text{moni}}$
VS1	ViT	Final layer	StyleGAN	✓	TC2	$L_{\text{on}} + L_{\text{moni}} + L_w$
VS2	ViT	Multiple layer	StyleGAN	✓	TC2	$L_{\text{on}} + L_{\text{moni}} + L_w$

Table 1. We illustrate the configuration of eight different models used for evaluation in terms of architecture modules and training losses.

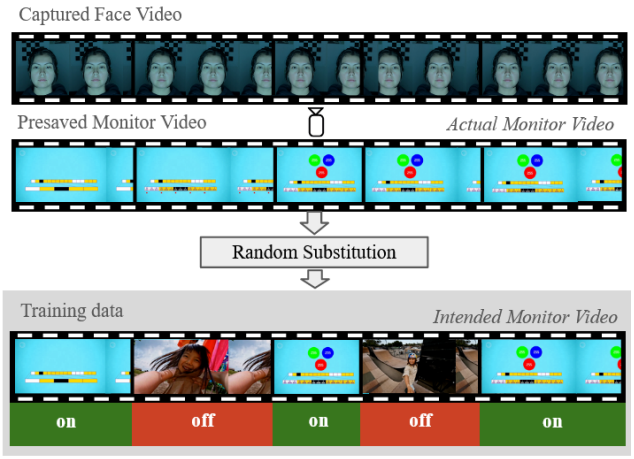


Figure 3. We require every authorized participant to record themselves watching ∼30 minutes of video content on their monitor only once and label this as 'on-task' scenario, e.g. making an introductory presentation when joining a team. We then randomly replace different segments of varying lengths of the recorded slideshow with random YouTube video clips, and change the label of these segments to 'off-task'. This augmented video clip with random chunks of slideshow and YouTube videos will be considered as the 'intended video' and the original slideshow video will be considered the 'actual video' the user watches with corresponding 'on-task' and 'off-task' labels.

facial features from the portrait image, a **mapping network** that transforms backbone features into input features for the generator, a **generator** designed to reconstruct the predicted monitor content, and a **task classifier** that uses the predicted monitor and the input intended monitor for binary classification of on-task vs. off-task.

The **feature extractor** operates by utilizing a pre-trained network to extract features from a cropped portrait, focusing exclusively on the facial region of a 480×480 resolution image. We use either a ResNet18[14] or ViT Base with 16×16 image patches [7] as a feature extractor backbone.

The ResNet feature extractor has two possible configurations. In the case of ResNet18 with a final layer output, a feature map of dimensions 512x7x7 is produced from the final convolutional layer. In the ResNet18 with Multiple Layer output, four intermediate features are extracted from the convolutional layer. The ViT backbone with a final layer output extracts a feature map of dimensions 3x16x16, specifically targeting the classifier token. The ViT backbone with a multiple layer output, takes the 12 intermediate classifier tokens from the output of all 12 Transformer layers.

Generator network uses the output of the mapping network or that of the feature extractor itself to estimate the 'actual monitor' content as the predicted monitor. We consider two different design choices for the generator: (i) a *convolutional generator*, where we use a learnable convolution network to generate an $18 \times 32 \times 3$ resolution monitor image directly from the output of the feature extractor without using any mapping network. (ii) a *StyleGAN generator*[16], where we use the pre-trained StyleGAN-XL[28], trained on a large dataset of landscape images, to generate a $64 \times 64 \times 3$ monitor image. The StyleGAN-XL network uses the Mapping Network to convert the output of the Backbone Feature Extractor to StyleGAN latent space.

Mapping network is exclusively used when employing StyleGAN-XL [28] for the Generator. This network maps the output of the backbone to the $\mathcal{W}$ space of dimension 512. We use a convolutional network if the feature extractor backbone is a ResNet, and a multi-layer perceptron if the feature extractor is ViT.

Task classifier calculates the on-task probability using the intended monitor and the predicted monitor. There are two types of task classifiers TC1 and TC2. Models with TC1 extract the features for the input face directly from the mapping network. Models with the TC2 classifier are similar but extract features from the predicted monitor.

Loss functions. Our model is trained by minimizing a weighted combination of three loss functions: on-task classification loss, monitor loss, and latent code loss. First, we utilize the binary cross-entropy loss for the on-task loss to calculate the probability of being on-task (equal to 1).

$$L_{\text{on}} = \text{BCELoss}(\text{on-task}, \hat{\text{on-task}}) \tag{1}$$

Next, we take the monitor loss between the intended monitor and predicted monitor. We employ a combination

of L1 loss and perceptual loss, with λ_{L1} and λ_P set to 0.2 and 0.8, respectively. Also $M_{on-task}$ and $\hat{M}$ denotes on-task monitor and predicted monitor:

$$L_{\text{moni}} = \lambda_{L1} L_1(M_{\text{on-task}}, \hat{M}) + \lambda_P L_P(M_{\text{on-task}}, \hat{M}) \quad (2)$$

Finally, the latent code loss L_w was computed using the L2 loss. We first conducted latent space optimization on every monitor image to find the latent code $\mathbf{w}$ of the actual monitor by gradient descent on $\mathbf{w}$. The loss L_w is defined as follows

$$L_w = L_2(\hat{w}_{\text{optimized}}, \hat{w}) \quad (3)$$

where $\hat{w}_{\text{optimized}}$ represents the optimized latent code obtained through latent space optimization, and $\hat{w}$ denotes the output obtained using the mapping network.

5. Experiments

In this section, we outline our comprehensive approach to experimental setup and evaluation. We gather both training data and testing data following the specifications outlined in Sec. 3. Subsequently, from the pool of 5 training videos, 1 is randomly selected as validation data, with the remaining 4 videos constituting the training data. The model is then trained using this data, and after identifying the best threshold through validation, evaluation is performed on the testing data. In Sec. 5.1, we elucidate the metrics employed for binary classification and predicted monitoring. Experimental outcomes for classification and monitoring prediction are detailed in Sec. 5.2.

5.1. Evaluation metrics

We apply four metrics for evaluating classification: accuracy, precision, recall, and F1 score. We employ recall, indicating the ratio of true positives to actual positives; precision, measuring the ratio of true positives to predicted positives; and F1 score, which offers a balanced assessment of precision and recall. We also use RMSE (Root Mean Square Error) and PSNR (Peak Signal-to-Noise Ratio) to evaluate the quality of the predicted monitor.

In Tab. 3 we compare the average performance of different approaches across all four users under different lighting and head motion. We consider a baseline that *always* predicts each frame as 'on-task' or 'off-task'. Since our test data is biased towards 'on-task', the accuracy and F1 score are higher for the 'always on-task' baseline.

5.2. Observation

The optimal model (VS2) utilizes a combination of features from a ViT encoder and a StyleGAN generator. In Tab. 3, we observe that VS2 demonstrated superior performance in terms of accuracy, F1 score, and recall metrics across four diverse capturing environments and four users. Notably compared to the 'Always-On' baseline VS2 performs 11% better in accuracy, 13.5% better precision while having slightly better F1-score by 1.5%

ViT feature extractor and StyleGAN as a monitor predictor contribute to performance improvement. In Tab. 3, the average accuracy and F1 score of models using ViT backbone are 66.71% and 73.74, respectively, whereas for ResNet backbone, they are 65.98% and 71.69. This suggests that using ViT is more favorable than using ResNet. Among ViT backbone models predicting monitors (VC, VS1, VS2), those employing StyleGAN, namely VS1 and VS2, demonstrate approximately 11.2% higher accuracy and 3% higher F1 compared to the convolution-based generator approach of VC.

Multi-task learning for monitor prediction proves to be more effective in enhancing performance compared to solely performing binary classification. Particularly, VS2 exhibits a 4% improvement in accuracy and a 2.3% increase in F1 score over VX. Additionally, as shown in Fig. 6, it aids in understanding the contents of the user's screen.

Different skin types have minimal impact in predicting student engagement. We compared F1 and accuracy measures averaged across different models and users based on their skin types. We observe that, on average, all models perform roughly similarly across all users without any specific trend. Individual variations can be attributed to randomness in data augmentation and over-exposure during the capture process.

Effects of room light and head pose motion is investigated in Tab. 2. We note that with small head motion, the accuracy of our top-performing model VS2 remains almost the same for both dimly lit ('A0-M0') and ambient lighting ('A1-M0'), indicating our method can handle ambient room lighting. For dimly lit conditions, large head motion only slightly deteriorates the performance of VS2 by 3.5% ('A0-M1' vs 'A0-M0'). However, in ambient light conditions, accuracy decreases by an average of 11.8% with head motion ('A1-M1' vs 'A1-M0'). We argue that significantly large head motion, with more than 30-degree head pose variation on average between frames, is rare in virtual meetings.

Effect of color differences between intended and actual monitor in predicting 'off-task' accuracy is studied in Fig. 5, by depicting variations of accuracy with respect to differences in hue, saturation, and value. We consider our best-performing model, VS2, and evaluate its accuracy for one user in a dimly lit room with a small head motion. We note that our model performs best with large differences in value or intensity between the actual and intended monitor, as expected. The accuracy of our model remains roughly the same for hue or color differences between the intended and actual monitor, showing that our model can differentiate small variations in hue.

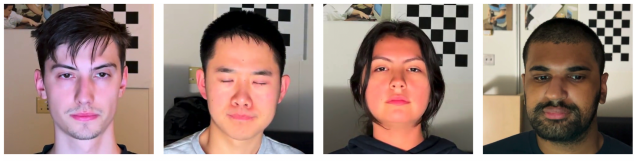
Predicting monitor content. We evaluate the error be-

	A0-M0		A0-M1		A1-M0		A1-M1	
Model	**Acc. (%) ↑**	**F1 score ↑**	**Acc. (%) ↑**	**F1 score ↑**	**Acc. (%) ↑**	**F1 score ↑**	**Acc. (%) ↑**	**F1 score ↑**
RX	67.96	0.7662	**70.34**	**0.7543**	61.92	0.6808	58.55	0.7365
VX	70.87	0.7780	67.17	0.7203	64.07	0.7029	59.49	0.7322
VS1	67.39	0.7594	68.97	0.7308	**71.42**	0.7460	59.77	0.7385
VS2	**71.37**	**0.7809**	67.92	0.7121	71.04	**0.7618**	**62.29**	**0.7444**

Table 2. We evaluate the average performance (higher the better) of our leading algorithms across 4 different users of varying skin tones on different pose and lighting conditions. 'A0' and 'A1' indicate dimly lit room and room with ambient lighting, while 'M0' and 'M1' indicate small head motion and large head motion respectively.

Model	**Acc. (%)**	**F1 score**	**Precision**	**Recall**
Always-On	58.32	0.7367	0.5832	1.0
Always-Off	41.67	0	0	0
RX	66.11	0.7340	0.6816	0.8137
RC	67.94	0.7389	0.7071	0.7948
RS1	64.59	0.6653	0.6447	0.6976
RS2	65.29	0.7295	0.6791	**0.8172**
VX	66.76	0.7336	0.6895	0.8048
VC	62.05	0.7203	0.6621	0.7398
VS1	68.53	0.7449	0.7132	0.7989
VS2	**69.51**	**0.7510**	**0.7188**	0.8069

Table 3. We evaluate the average performance (higher the better) of our leading algorithms across 4 different users in different pose and lighting conditions. We consider two baselines where we always predict each frame as on-task or off-task.



User 1 User 2 User 3 User 4

	Accuracy (%)			
Model	**User1**	**User2**	**User3**	**User4**
VS1	68.65	70.19	66.14	72.05
VS2	69.44	73.19	69.25	68.56

Figure 4. We evaluate all eight proposed models separately on 4 users of different skin types in a dimly lit room with minimal head motion (top). Our top two best-performing models VS1 and VS2 (bottom) produce similar results for all users. Small variations could be attributed to randomness in data augmentation and overexposure during capture.

tween the predicted and the actual monitor for both dimly lit (A0-M0) and ambient room lighting (A1-M0) with small head motion. The error between the predicted monitor compared to the actual monitor is similar for all models, as observed in Tab. 4. Our best-performing model for monitor prediction is VS2 in both settings. In Fig. 6, we observe that our model can faithfully recover the color tone of the original monitor for both 'on-task' and 'off-task'. Note that

	A0-M0		A1-M0	
Model	**RMSE ↓**	**PSNR ↑**	**RMSE ↓**	**PSNR ↑**
RC	10.22	27.94	10.21	27.95
RS1	10.26	27.90	10.26	27.91
RS2	10.14	<u>28.01</u>	10.22	27.95
VC	10.16	27.99	10.258	27.92
VS1	<u>10.15</u>	28.00	<u>10.19</u>	<u>27.98</u>
VS2	**10.12**	**28.03**	**10.16**	**28.00**

Table 4. We evaluate the error between the predicted monitor and the 'actual monitor' for a small head motion with dimly lit ('A0-M0') and ambient lit ('A1-M0') data.

when the actual monitor has two dominant colors (white and yellow in row 4), the predicted monitor can also recover these colors. However, since we are using StyleGAN-XL, pre-trained to generate landscape images, we often hallucinate structures in the middle of the image. We note that we can mainly recover the color of the monitor content and fail to reconstruct the basic layout or recognizable objects

Relighting algorithms virtually change facial lighting and thus has the potential to defeat our system which relies on the correlation between monitor lighting and facial lighting. We evaluate the performance of two existing relighting algorithms, Relight-Net1[30] and Relight-Net2[32] in Fig. 7 on all four conditions for one user. We observe that after applying relighting the accuracy of our system drops significantly, especially for small head motion in dimly lit rooms (A0-M0) by roughly 15%. However, we notice that with large head motion and ambient room light (A1-M1) the relighting algorithms are highly inaccurate and thus fail to defeat our approach by any significant margin.

6. Conclusion

We present a new technique that leverages screen lighting reflected from a participant's face to determine whether they are real humans or deep fakes and whether they are paying attention or not to the intended screen content, e.g. lecture slides. We can do so passively, without requiring any specific patterns to be projected on the screen intermittently, thus not disrupting the meeting flow and alerting the bad actors. We envision that such a technique can be ex-

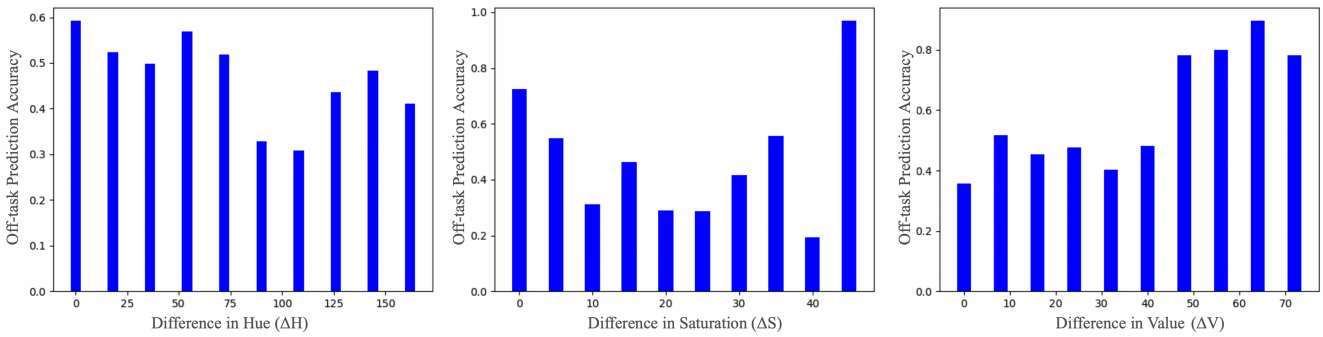


Figure 5. We show that the off-task prediction accuracy of our best-performing model, VS2, is independent of the hue, saturation, and value differences between the actual and intended monitor.

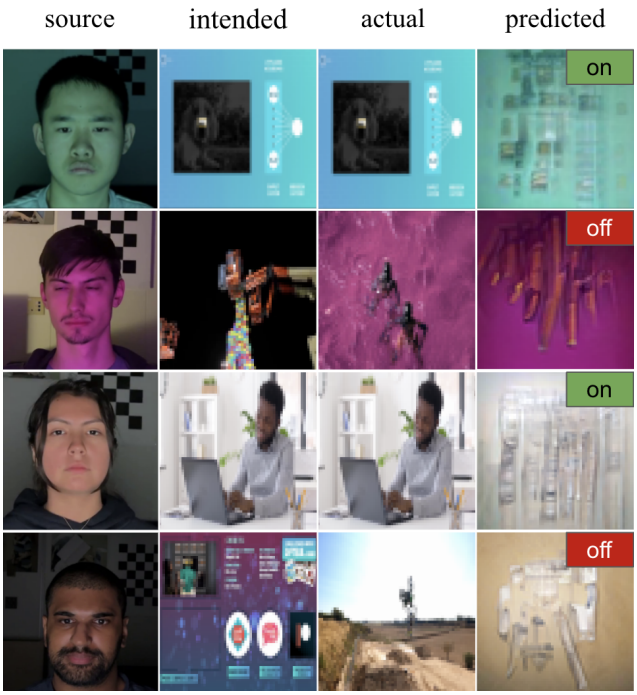


Figure 6. Our best-performing model, VS2, can generate a predicted monitor which has similar color tones as the 'actual monitor'. Predictions of our model are shown in the inset.

tremely useful in identifying deep fake avatars during video conference calls, thereby mitigating security risks and preventing fraudulent activities. This can be also beneficial in online education systems enabling instructors to better understand student engagement and attention, which can then be used to refine their own lecture materials or group discussion plans. We train a personalized network architecture that takes the face video and intended monitor image, e.g. lecture slides, as input and predicts the actual monitor content, finally classifying the user as on-task or off-task. We present a detailed analysis to understand the impact of varying environmental conditions like head motion, ambient room lighting, color variations in monitor, and network architecture designs. Our research introduces a new problem to the computer vision community that can be further explored to build better models that can also generalize across many users.

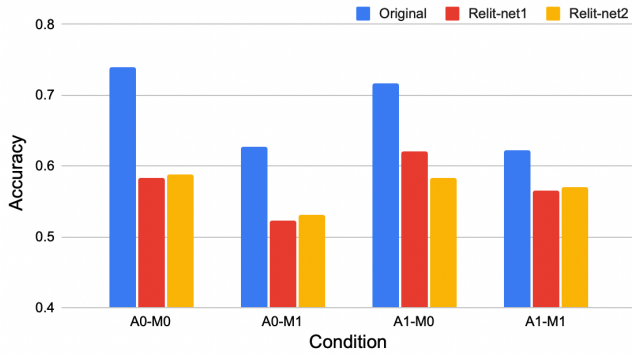


Figure 7. We show that applying existing relighting algorithms, Relight-Net1[30] and Relight-Net2[32], on face videos can reduce the performance of our system significantly.

Ethical considerations. While the goal of this paper is to create a tool that will enable instructors and managers to create a secure and engaging virtual communication environment, we recognize that this technique can be potentially used to violate the privacy of participants. In this current paper, we assume that all meeting participants consent to the authenticator building their personalized detection models, but in the future, a more generalized system can be used to predict what any user is watching on their monitor without their consent. However, it is extremely easy to defeat this system by using a strong light in front of the face, e.g. a lamp or sunlight coming through the window. Even when these additional light sources are not available, our system can be counteracted by virtual relighting algorithms. While current relighting algorithms could only reduce the F1 score to ∼65%, we believe future research on virtual relighting can also utilize our problem setup to validate their efficacy in defeating our system.

References

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments, 2022. 3

[2] Ioannis Arapakis, Mounia Lalmas, and George Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, page 1439–1448, New York, NY, USA, 2014. Association for Computing Machinery. 3

[3] Heather Chen and Kathleen Magramo. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. *CNN*. 1

[4] Jun Myeong Choi, Max Christman, and Roni Sengupta. Personalized video relighting using casual light stage. *arXiv preprint arXiv:2311.08843*, 2023. 2, 3

[5] Yung-Yu Chuang, Douglas E Zongker, Joel Hindorff, Brian Curless, David H Salesin, and Richard Szeliski. Environment matting extensions: Towards higher accuracy and real-time capture. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 121–130, 2000. 2

[6] Roberto Di Pietro and Stefano Cresci. Metaverse: Security and privacy issues. In *2021 third IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pages 281–288. IEEE, 2021. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 5

[8] Pardis Emami-Naeini, Tiona Francisco, Tadayoshi Kohno, and Franziska Roesner. Understanding privacy attitudes and concerns towards remote communications during the {COVID-19} pandemic. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 695–714, 2021. 1

[9] Stephen H. Fairclough and Louise Venables. Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology*, 71(1):100–110, 2006. 3

[10] Maria Frank, Ghassem Tofighi, Haisong Gu, and Renate Fruchter. Engagement detection in meetings. *CoRR*, abs/1608.08711, 2016. 3

[11] Candice R Gerstner and Hany Farid. Detecting real-time deep-fake videos using active illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–60, 2022. 2, 3

[12] Hui Guo, Xin Wang, and Siwei Lyu. Detection of real-time deepfakes in video conferencing with active probing and corneal reflection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[13] Abhay Gupta, Richik Jaiswal, Sagar Adhikari, and Vineeth Balasubramanian. DAISEE: dataset for affective states in e-learning environments. *CoRR*, abs/1609.01885, 2016. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[15] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2021. 3

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 5

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[18] Dmitry Lagun and Mounia Lalmas. Understanding user attention and engagement in online news reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, page 113–122, New York, NY, USA, 2016. Association for Computing Machinery. 3

[19] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 113–122, New York, NY, USA, 2014. Association for Computing Machinery. 3

[20] Shui-fong Lam, Shane Jimerson, Bernard Wong, Eve Kikas, Hyeonsook Shin, Feliciano Veiga, Chryse Hatzichristou, Fotini Polychroni, Carmel Cefai, Valeria Negovan, Elena Stănculescu, Hongfei Yang, Yi Liu, Julie Basnett, Robert Duck, Peter Farrell, Brett Nelson, and Josef Zollneritsch. Understanding and measuring student engagement in school: The results of an international study from 12 countries. *School Psychology Quarterly*, 29:213–232, 2014. 1

[21] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024. 3

[22] Hongbo Liu, Zhihua Li, Yucheng Xie, Ruizhe Jiang, Yan Wang, Xiaonan Guo, and Yingying Chen. Livescreen: Video chat liveness detection leveraging skin reflection. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1083–1092. IEEE, 2020. 3

[23] Ning-Han Liu, Cheng-Yu Chiang, and Hsuan-Chin Chu. Recognizing the degree of human attention using eeg signals from mobile sensors. *Sensors*, 13(8):10273–10286, 2013. 3

[24] Yu Lu, Sen Zhang, Zhiqiang Zhang, Wendong Xiao, and Shengquan Yu. A framework for learning analytics using commodity wearable devices. *Sensors*, 17(6), 2017. 3

[25] Ko Nishino and S.K. Nayar. The world in an eye [eye image interpretation]. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I, 2004. 3

[26] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[27] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9367–9376, 2019. 3

[28] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *CoRR*, abs/2202.00273, 2022. 5

[29] Yoav Y Schechner, Shree K Nayar, and Peter N Belhumeur. Multiplexing for optimal lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 29(8):1339–1354, 2007. 2

[30] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. A light stage on every desk. *CoRR*, abs/2105.08051, 2021. 2, 3, 7, 8

[31] Jiacheng Shang and Jie Wu. Protecting real-time video chat against fake facial videos generated by face reenactment. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 689–699. IEEE, 2020. 3

[32] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *CoRR*, abs/1905.00824, 2019. 3, 7, 8

[33] Daniel Szafir and Bilge Mutlu. Pay attention! designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 11–20, New York, NY, USA, 2012. Association for Computing Machinery. 3

[34] Kang Wang, Shen Wang, and Qiang Ji. Deep eye fixation map learning for calibration-free eye gaze tracking. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, page 47–55, New York, NY, USA, 2016. Association for Computing Machinery. 3

[35] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. Security and privacy on generative data in aigc: A survey. *arXiv preprint arXiv:2309.09435*, 2023. 3

[36] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1): 86–98, 2014. 3

[37] Karen Wilson and James H. Korn. Attention during lectures: Beyond ten minutes. *Teaching of Psychology*, 34(2):85–89, 2007. 1

[38] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[39] Douglas E Zongker, Dawn M Werner, Brian Curless, and David H Salesin. Environment matting and compositing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 537–546. 2023. 2