Pattern Recognition Letters 137 (2020) 99–107

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep video-to-video transformations for accessibility with an application to photosensitivity

Andrei Barbu[1,*], Dalitso Banda, Boris Katz

Massachusetts Institute of Technology, Center for Brains Minds and Machines, 32 Vassar Street, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Available online 27 June 2019

MSC:
68T45
68T05
68N01
68U10

Keywords:
Photosensitivity
Accessibility
Computer vision
Video-to-video transformation

ABSTRACT

We demonstrate how to construct a new class of visual assistive technologies that, rather than extract symbolic information, learn to transform the visual environment to make it more accessible. We do so without engineering which transformations are useful allowing for arbitrary modifications of the visual input. As an instantiation of this idea we tackle a problem that affects and hurts millions worldwide: photosensitivity. Any time an affected person opens a website, video, or some other medium that contains an adverse visual stimulus, either intended or unintended, they might experience a seizure with potentially significant consequences. We show how a deep network can learn a video-to-video transformation rendering such stimuli harmless while otherwise preserving the video. This approach uses a specification of the adverse phenomena, the forward transformation, to learn the inverse transformation. We show how such a network generalizes to real-world videos that have triggered numerous seizures, both by mistake and in politically-motivated attacks. A number of complimentary approaches are demonstrated including using a hand-crafted generator and a GAN using a differentiable perceptual metric. Such technology can be deployed offline to protect videos before they are shown or online with assistive glasses or real-time post processing. Other applications of this general technique include helping those with limited vision, attention deficit hyperactivity disorder, and autism.

1. Introduction

The visual world is not equally accessible to everyone. Even with perfect eyesight you may be limited in your ability to perceive the environment around you. For example, those with autism spectrum disorder often have difficulty perceiving facial expressions while those with photosensitive seizure disorders can have adverse reactions to certain kinds of flashes and patterns. The effects of this can range from causing feelings of isolation, losing access to important information about the physical and social environment, loss of quality of life, all the way to life-threatening seizures. Much prior work has been symbolic, focusing on communicating the state of the world to a listener by extracting needed information and presenting it to them. Instead, we demonstrate how, with little supervision, one can automatically learn to manipulate the visual environment in order to make it safer and more accessible by learning video-to-video transformations. While we primarily focus on photosensitive seizure disorders due to their high impact on viewers, the techniques presented can be adapted for other applications[2]. Preliminary work on enhancing face perception is presented in the discussion.

Photosensitive seizure disorders are often, but not always, a form of epilepsy where certain kinds of visual stimuli, primarily repeating patterns and flashes, can trigger seizures [1]. They impact millions worldwide, roughly 1 in every 4000 people, with many having abnormal EEGs without necessarily experiencing seizures but sometimes showing other symptoms such as migraines [2]. Children are more likely to be affected than adults, for reasons that are unclear at present. Such stimuli have been used in attacks, such as defacing the American epilepsy foundation website with videos crafted to trigger seizures. Recently, politically-motivated attacks have attempted to target reporters [3,4] using Tweets which resulted in seizures. Put simply, if there was a visual filter that would inoculate stimuli, these disorders would be significantly mitigated.

This is no easy task as the range of harmful stimuli is broad. This problem follows the general pattern seen in computer vision where the forward instance of the task – creating adverse stimuli – is easy, while the inverse instance – fixing adverse stimuli – is far

* Corresponding author.
E-mail addresses: abarbu@mit.edu, andrei@0xab.com (A. Barbu).

[1] Barbu developed the idea and drafted the manuscript. Banda implemented the methods. The authors jointly edited the submission.

[2] Source code for this submission is available online at http://github.com/abarbu/photosensitivity-video-to-video.

https://doi.org/10.1016/j.patrec.2019.01.019
0167-8655/

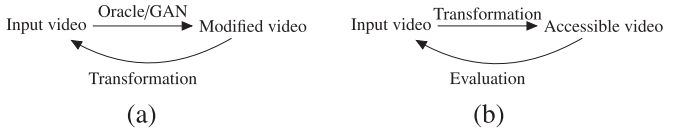

Fig. 1. Two general ways to learn transformations that create more accessible visual worlds. (a) A hand-crafted oracle or a GAN can learn to insert harmful stimuli in videos. The accessibility transformation can then be automatically learned to invert this operation. Enough untransformed data ensures that innocuous stimuli are not affected. This is the approach taken here for photosensitive seizure disorder. (b) An automated means to evaluate the output of a transformation, for example using humans on Mechanical Turk or using neural networks as models of the visual system, can directly supervise the transformation. This is the approach taken in the preliminary results on enhancing face recognition. Potentially, neuroimaging could provide this feedback directly, although we do not explore that alternative here.

more difficult and even ill-posed. In addition, corrections must be applied with a light touch to not degrade the experience for others who may not have photosensitive seizure disorders. This is particularly important since many do not know that they are sensitive to flashes or patterns, and may want to be preemptively protected. Upon exposure, one can become more sensitive to such stimuli underscoring the importance of being proactive. Wide-spread photosensitivity testing is not conducted for this reason, along with the fact that most of the time most content providers avoid such flashes as they are generally annoying to users. We demonstrate neural networks which learn transformations of videos that mitigate flashes and patterns while preserving the structure and quality of videos, without feature engineering, to make the visual environment safer for all.

Guidelines to mitigate photosensitivity have existed since the 1990s, in part, due to an incident in Japan. Children watching an episode of Pokemon were exposed to around four seconds of full-screen bright red-blue flashes at 12 Hz. This caused 685 hospital visits with around 200 hospital admissions, primarily children, most of whom had never experienced a seizure before [5]. The guidelines consist of hand-crafted rules, such as disallowing large high-luminance red-blue flashes between 3 Hz and 49 Hz. Such guidelines are imperfect, must be followed by each content producer (periodically major movies accidentally include such flicker), and do not meet the needs of every user. They are far less effective with new online media where even errors like incorrectly encoded videos can create flashes. We envision that in the future users will run full-screen accessibility transformations to improve and safeguard their media.

Hand-crafted filters that attenuate flashes have existed for several decades. Most significantly, Nomura et al. [6] created an adaptive temporal filter which reduces flicker at 10–30 Hz. These devices are imperfect; it is difficult to capture notions such as the fact that red-blue and white-black flickers have different rages, that flicker may occur in many different ways, and that patterns can be just as disruptive as flicker.

We propose a method to automatically learn filters that are robust with respect to different kinds of patterns, frequencies, and sizes in the field of view; see Fig. 1. We show three methods to train such video-to-video transformations without specifying what the transformation should be. First, using an oracle which can take videos and insert harmful stimuli. Second, using a GAN that learns to insert harmful stimuli. And third, using a model for the human visual system which can predict which stimuli are harmful. A fourth method is in principle possible, but we do not implement or discuss it in detail: using neuroimaging one could gather additional training data for the transformation and even perhaps train without any user responses at all.

In each case, a neural network takes as input a short video and learns to transform it into an innocuous video. For the first two cases, the network is trained on videos taken from YouTube, likely not to be particularly problematic, and asked to recover those original videos after they have been augmented to contain problematic stimuli. For the third case, we take as input problematic stimuli and learn to transform them into stimuli that the visual system will consider acceptable. Regardless of the overall structure, a network and objective function are specified (rather than the precise transformation) allowing one to discover new features and patterns. With more explainable networks this might contribute to understanding how an adverse stimulus causes problematic responses. Additionally, the methods are adaptable to other accessibility issues.

This paper makes several contributions: 1) a neural network that takes as input videos and produces versions that mitigate flashes rendering them safer for those with photosensitive epilepsy, 2) a means to train such a network without specifying the precise transformation the network should learn, 3) two approaches which augment videos either using a hand-coded oracle or a GAN that learns such transformations and allows the automated use of unlimited amounts of video data from the web without any annotations, 4) an approach that uses differentiable approximations of the human visual system to recover this transformation, 5) a novel architecture that builds on U-Net using stacked convolutional LSTMs to transform videos, 6) the general approach of transforming videos to enhance or suppress features in videos that are detrimental or difficult to recognize. This approach is applicable to other domains such as improving emotion recognition for those with autism or learning to suppress distractors for those with ADHD.

2. Related work

Sequence-to-sequence models have been employed to caption videos [11–13], and to recognize actions in videos [14,15], although this is a video-to-text transformation. Increasing the resolution of videos by fusing together multiple video sources has also benefited from such approaches [16]. Optical flow transforms videos into flow fields, a transformation that can be automatically learned [17]. Networks can be used to predict future frames [18] and predict how physics will affect objects [19]. Text-to-audio transformations have enabled speech synthesis [20]. Video-to-video transformations can increase spatial and temporal resolution [16].

Since single-image transformations have seen significant popularity, we discuss only a few related publications. Style transfer networks learn to apply a visual style provided one or more reference images [21]. Image segmentation transforms images into densely labeled counterparts [22,23]. Similar ideas have been applied in robotics to enhance the realism of simulated scenes and enable transfer to real-world settings [24,25]. More generally, techniques such as CycleGANs have enabled arbitrary image-to-image transformations [26] for many domains.

Leo et al. [27] provide a thorough and recent review of computer vision for accessibility. A few hand-crafted video-to-video transformations have been considered in the past, see Table 1 for an overview. Farringdon and Oni [28] record which objects were seen in prior interactions and cue a user that they have already seen those objects thereby enhancing their memory. Damen et al. [7] highlight important objects allowing users to discover on their own which objects are useful for which tasks. Betancourt et al. [29] review a rich first person view literature with a focus on wearable devices.

Tools have been created to detect segments of a video which can be problematic for those with photosensitivity [1,30,31]. Signal processing has been used to create filters that can eliminate some forms of flashing [32]. Such filters are difficult to create as making problematic stimuli is much easier than detecting and fixing them.

Table 1
Related work on creating visual transformations for accessibility. These take as input images or videos, are carefully hand-crafted to help certain populations overcome particular disabilities, or perform particular tasks. We present a generalization of these methods where the transformation is automatically learned rather than hand-crafted for a specific disability or individual.

Task/Disability	*Publication*	*Approach*
Photosensitivity	Nomura et al. [6]	Filter attenuates 10–30 Hz flashes
Memory loss	Damen et al. [7]	Object detector and reminders
Navigation	Stoll et al. [8]	Render depth maps as audio
Dyslexia	Rello and Baeza-Yates [9]	New font shapes and sizes
Color blindness	Simon-Liedtke and Farup [10]	Many daltonization transformations

Blue-tinted lenses in many cases increase the threshold at which a stimulus is problematic [33].

As Leo et al. [27] point out, computer vision has so far been used to replace particular human abilities by developing algorithms that perform those tasks and provide users with symbolic information about the environment. For example, commercial systems such as OrCam primarily extract text features, object properties, and face identities from videos and present them in an interactive manner to a user. A few notable exceptions exist. We provide a short summary of these in Table 1. For example, approaches which convert depth maps into auditory cues to aid navigation [8,34]. While not explicitly a video-to-video transformation, Rello and Baeza-Yates [9] present fonts that help those with dyslexia. Daltonization [10] is an image-to-image transformation that makes images easier to understand for those with limited color vision. Such hand-crafted transformations show the value of the general approach of transforming the visual world. To our knowledge, ours is the first work that builds on these successes to learn arbitrary transformations that make videos accessible thereby opening up this line of work to helping many more people with varied accessibility issues.

3. Method

We separate our approach into two parts: a forward/synthesis pass which automatically creates a dataset consisting of pairs of videos – original videos and videos that are transformed to hide or highlight some feature – and an inverse pass which learns a video-to-video mapping. In the discussion section, we will describe several other potential methods, including preliminary results of a method which can dispense with the forward pass, replacing it with a differentiable computational model for part of the human visual system. We explore two complimentary ways to create the forward pass: specifying the transformations to be applied manually and training a GAN [35] that learns to apply the transformation. We employ three types of networks which can learn video-to-video transformations: a bidirectional convolutional LSTM that learns embeddings of the video and then decodes a modified video [15], a novel architecture that combines a U-Net with a bidirectional stacked LSTM [23], and an adapted spatio-temporal autoencoder [36]. Our main contribution is not the specifics of any one network, although each of the networks is well-adapted for flash mitigation; rather the approach of learning video-to-video transformations for accessibility.

Next, we will describe each of the approaches and how they have been adapted to safeguarding videos. In each case, we take fixed-size snippets of a video as input, 64x64 pixel volumes that consist of 101 frames, transform them, then learn the inverse transformation. These snippets are extracted from random videos downloaded from YouTube. While it is unlikely that any one video from YouTube will have an adverse stimulus, it is still possible, and we rely on the robustness of the neural networks to learn despite this minor source of error in the training data.

3.1. Forward pass

This operation is the equivalent of video synthesis or rendering in computer vision – creating videos which contain adverse stimuli or, in the more general case, which hide useful information. Since the forward pass is easier to specify and construct, we provide both a manual and automatic method of doing so. Note that if enough data could be collected naturally, this dataset creation pass might be unnecessary, or might be augmented with real-world data. Being able to generate data automatically means that we merely specify the desired behavior which then provides unlimited data for a network to learn to replicate that behavior.

The purpose of the forward pass is only to enable the acquisition of the inverse transformation. It is not an end in and of itself meaning that it can be noisy and approximate as long as it provides enough constraints and data to eliminate undesirable transformations. This is unlike prior work in many other areas of assistive computer vision where a noisy or approximate approach would lead to misreporting events to a user.

3.1.1. Manual transformations

Much research has explored the precise conditions under which photosensitive disorders are triggered [6,37–39]. The likelihood that a flash will be problematic is a function of its frequency, wavelength, intensity, and size in the field of view. Flashes above 3 Hz are widely problematic although 4–5% of those who are affected can be triggered by flashes as low as 1–2 Hz [38]. Wavelength and intensity interact in a way that is not yet well characterized although it is known that flashes of equivalent intensity are more problematic if they are red/blue flashes, wavelengths of 660–720 nm. While the size of an event is important, flashes of only 1–2° in the visual field can trigger an adverse reaction; moreover such flashes can do so from the periphery even if subjects do not fixate on them. Flashes need not be unitary; they can aggregate from small flashes distributed throughout the visual field. Moreover, while usually only flashes are discussed, patterns such as moving gratings can be equally problematic. Here we use flashes as a shorthand but also include moving gratings and other patterns.

We used this body of knowledge to create a generator that takes as input video clips and modifies them to include flashing and problematic patterns. It encodes the above guidelines and randomly inserts such features separately in the RGB channels, intensity channel, the HSV color space, and by modifying the distribution of wavelengths in a video. Crucially, we insert stimuli both above the threshold at which they are problematic and below that threshold. A combination of above- and below-threshold examples enables learning a transformation that preserves the detail in innocuous videos and corrects problematic videos. This step could in

principle be tuned to a specific user, a critical need given the large population that is significantly more sensitive to such flashes.

3.1.2. Generative adverserial networks

Separately from creating a generator encoding our prior knowledge about how problematic stimuli are created, we can measure how problematic a stimulus is regardless of how it was created. We take advantage of this to create a GAN that generates new problematic stimuli in ways that a generator constructed by a human would not, thus adding robustness to our approach. To build this network, we adopt the approach of Wu et al. [40] and Johnson et al. [41] who introduce a perceptual loss. This is a network which measures the quality of an image or video whose output can be added to the loss function of a GAN to ensure that the resulting images are more visually appealing. In our case, we construct a detector, much like those of Harding and Jeavons [1], Vanderheiden [31] and Park et al. [30], which takes as input a video and produces a score—the likelihood that the video is problematic. Unlike that previous work, we construct this detector from differentiable operations. In essence, this is a CNN operating over videos which provides a perceptual loss for photosensitivity.

This allows us to employ a GAN to produce videos. A GAN consists of a generator and a discriminator. The generator and discriminator are convolutional networks that take as input video volumes. The generator transforms its input video while the discriminator attempts to determine if the result has some desirable property, for example being hard to distinguish from some collection of videos. As described above, we create a linear combination between the loss of the discriminator and a perceptual loss for photosensitivity. Crucially, our loss is differentiable (we will discuss potential differentiable losses for other applications in Section 5), and it allows for selecting for videos that are both above and below the threshold at which flashing becomes problematic. This means that the generated videos are a mixture of videos that should be preserved as they are, and videos that should be transformed and made safe by the inverse pass.

3.2. Inverse pass

The inverse pass is the focus of our efforts. It transforms unsafe videos into safe videos with minimal loss in quality and without compromising videos when no unsafe stimuli are present. We first create a large corpus of videos which is repeatedly transformed by the forward pass as described above. These triples of original videos, modified videos, and metrics about how determinetal the applied modifications were, are used to acquire the inverse transformation. We present three methods for learning this transformation: a bidirectional LSTM in Section 3.2.1, a novel stacked LSTM U-Net in Section 3.2.2, and a spatio-temporal autoencoder in Section 3.2.3. The range of possible sequence-to-sequence transformations that could be employed is vast.

3.2.1. Residual bidirectional LSTMs

This first network is by design shallow and relatively simple, serving to demonstrate how well such methods work even with few parameters. The following two networks are significantly larger and more complex. Long short term memory, LSTM, [42] networks are recurrent neural networks that integrate reasoning across time with a gated memory unit. Input videos are processed frame-by-frame by a CNN with 4×4 convolutions having just 4 filters followed by batch normalization and RELU activation. A bidirectional convolutional LSTM [43] takes as input this embedding for each frame of a video, updates its internal state, and predicts a new frame feature vector. Bidirectional LSTMs have the ability to integrate information from both the future and the past potentially allowing them to more accurately detect stimuli such as flashes and changing patterns. A final layer consists of a 3×3 convolution with 3 filters, one for each color channel, that produces a residual added to the original input. Residuals have been shown to be far easier to learn, requiring the network to learn and represent only the difference between the input and the desired output rather than needing to embed the entire high-dimensional input. Because of its small size, this network learns quickly from few examples but then has difficulty adapting to the more complex forms of flashing and patterns produced by the GAN.

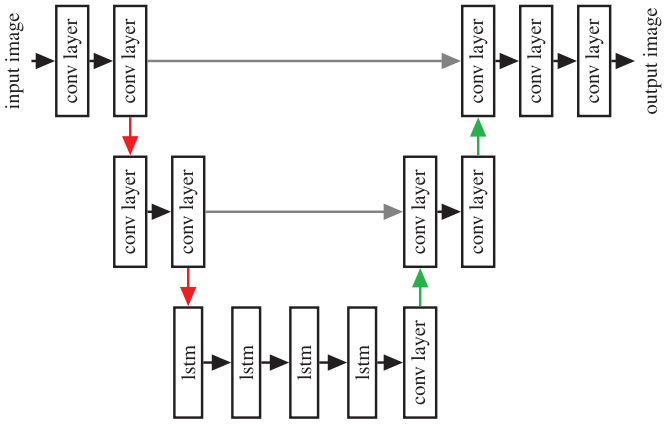

Fig. 2. An overview of the stacked LSTM U-Net architecture. Each convolutional layer uses 3×3 convolutions while each LSTM layer has 64 units. Convolutional LSTMs have 3×3 filters. We perform 2×2 pooling and 2×2 upsampling while forwarding the feature map from the contracting path into the expanding path. The contracting path retains the information necessary to analyze a scene, the LSTMs process and remember that information over time, while the expanding path along with the forward connections reconstruct the image.

3.2.2. Stacked LSTM U-Net

Next we adapt the U-Net architecture [23] to video-to-video transformations; see Fig. 2. U-Net is a popular approach to image segmentation which finds low-dimensional embeddings for an input image and then decodes those embeddings into a segmentation of the image. It consists of two paths, a contracting path that lowers the dimensionality of the input and an expanding path that raises it. The outputs of the contracting path are forwarded to the expanding path at each step, much like the residual was used in the network described in the previous section but doing so between internal layers not just at the input and output layers. The contracting path and expanding path contain the same number of convolutional layers, in this case two 3×3 convolutions, followed by downsampling in the contracting path and upsampling in the expanding path. Rather than predicting a segmentation map, we predict an RGB image per frame, and add a 3×3 convolutional layer after the contracting path to predict this image. This would be a 1×1 convolution if segmenting the image.

Between the contracting and the expanding path, at the point where each frame has been reduced to a low-dimensional embedding, we use four stacked convolutional LSTMs [43] to process that embedding. These used 3×3 convolutions with 64 units to provide long-term memory. Intuitively, the contracting path discovers important features, the LSTMs combine these with features from earlier frames, and then the expanding path creates a new frame. This network is shallower than the original U-Net network, in part because the convolutional LSTMs can replace some of the operations of the original network, and because the input is smaller.

This network is significantly larger than the one described in Section 3.2.1 and is better designed to progressively eliminate undesired portions of the input or enhance the input through multiple processing stages. As will be shown in the experimental results section, having access to the temporal structure of the video

is crucial as a per-frame U-Net does not learn to perform the required transformations. The significant amount of information being stored between frames provides the capacity to learn structures at different temporal scales, important for recognizing non-flashing but problematic patterns such as rotating gratings. This is the first application combining stacked LSTMs and a U-Net architecture to video-to-video transformations.

3.2.3. Spatio-temporal autoencoder

The third architecture is repurposed from an existing video-to-video transformation: computing the optical flow in videos. Spatio-temporal autoencoders [36], referred to in results as STA, take each frame as input and predict an output frame based on the sequence of frames observed up to that point. Unlike the previous networks they are neither bidirectional nor predict a residual. They embed the input into a low-dimensional space, in our case using a 4-layer 32-filter 3×3 convolutional network with RELU activation. This embedding is used by two stacked convolutional LSTMs with 32 filters to predict the optical flow. The network then reconstructs its beliefs about the next frame of the video by combining the embedding of the current frame with its predicted optical flow and decoding that prediction using the inverse of the embedding network. We use this approach for flash mitigation. Intuitively, since to compute optical flow we rely on the ability to predict the next frame based on previous frames, we can use this ability to predict the behavior of a flashing stimulus or pattern to determine if we should suppress it.

4. Experiments

We demonstrate the efficacy of our approach in several ways including qualitative evaluations of the corrected videos (Section 4.1), quantitative evaluations of the forward video generation pass (Section 4.2.1), of the inverse video correction pass (Section 4.2.2), demonstrating baselines and ablations (Section 4.2.3), as well as performance on held out real-world videos that have caused seizures (Section 4.2.4). We show both objective metrics and subjective human judgments rating the presence of adverse stimuli in the resulting videos as well as their overall quality and fidelity to the original.

We collected a training set consisting of 20,000 random popular videos, without manual filtering, from YouTube. From each video, we cropped 10 clips consisting of 100 frames in a 64×64 window—resulting in 200,000 training clips. An additional 2000 videos were collected and similarly post-processed to be used as a test set. The training and test sets were disjoint, not only in the clips used but in the original source videos as well, ensuring that no information can leak between the two.

4.1. Qualitative

In Fig. 3 we show original and transformed frames from several videos. Note that flashing is suppressed while the quality of the videos is largely preserved both for the frames which have flashing and those that do not. The networks attempt to infer what the frames with the adverse stimuli should contain to replace flashes and problematic patterns. Fig. 3(c) and (d) show frames from videos which are known to have caused hundreds of preventable hospital visits.

4.2. Quantitative

We collect quantitative metrics using human subject experiments to determine if the video transformations have successfully removed flashing and maintained video quality. Human subjects

Table 2
To determine how well the forward pass—which inserts flashes—operates, we modified each video in two ways: using the manual method, in red, and using the GAN, in blue. Humans compared each of the modified videos against the original and decided which of the two had more flashing, shown in the *chose modified* column. Overwhelmingly they identified the modified videos as having more flashing. The degree to which subjects believed that the modified videos contain more flashing was established using a slider: where -1 represents the original containing far more flashing and 1 represents the modified containing far more flashing. Note that the optimal values are not -1 and 1, as the flashing inserted is by design subtle to probe the boundary between below-threshold flashing which is innocuous and above-threshold flashing which is problematic. Subjects determined that on average the added flashes were of moderate intensity for both cases.

Modification method	Chose modified	Degree of flashing in original	Degree of flashing in modified
Manual	91%	−0.68	+0.34
GAN	83%	−0.16	+0.34

experiments are necessary since no accurate metrics for determining the quality of the videos exist at present [44]. In all cases, subjects are shown two repeating 64×64 video clips side-by-side, taking up no more than 10% of the screen, and complying with World Wide Web Consortium, W3C, criteria for preventing seizures due to photosensitivity [37]. After viewing the videos for at least 5 s, subjects are prompted to answer one of two questions: which clip contains more flashing or which clip has higher visual quality. They answer these questions using a slider to note the magnitude of their preferences. The same video was shown in three different ways: in its unmodified original form, with the forward pass modification that inserts the adverse stimulus (flashing), and the transformed version where that adverse stimulus is suppressed by a trained model.

Two kinds of values are reported from the human responses: thresholded and averaged results. Thresholded results determine the likelihood that a user would prefer one type of video over another, however slight that preference might be. Averaged results determine the preference for a particular condition and complement the thresholded results to show the magnitude of the effect. Results are reported on a scale of -1 to 1, -1 is an unequivocal preference for one stimulus while 1 is an unequivocal preference for the other stimulus. Presentation order was randomized at each trail. Note that it is unusual for subjects to completely prefer one condition over another; even a video without any flashing compared to one which has substantial and annoying flashing often does not get selected in its entirety.

For an application of the approach described here to be useful it must meet a few criteria. First, when the forward pass inserts flashes they must be clearly detectable to human subjects. Second, when the inverse pass removes flashes humans should agree that they have been removed and the quality should be high. Third, the transformation should transfer to real videos, not just videos which we have artificially added flashes to. Next, we discuss these three key issues along with several ablations.

4.2.1. How good is the forward, video generation, pass?

We first compare the videos before and after our generators has inserted problematic flashes and patterns. As shown in Table 2, subjects decided that the modified videos contain far more flashing with significant magnitude regardless of the method, 80–90% of the time. In the common case, where the modified video was chosen as having more flashing, the preference was the same for the hand-crafted generator and the learned GAN generator (+0.34). In the rare case, where the original was chosen as having more flashing, GANs produced videos which were selected against less strongly than the original videos (-0.16 vs -0.68). Videos can have significant amounts of flashing to begin with, and the hand-crafted

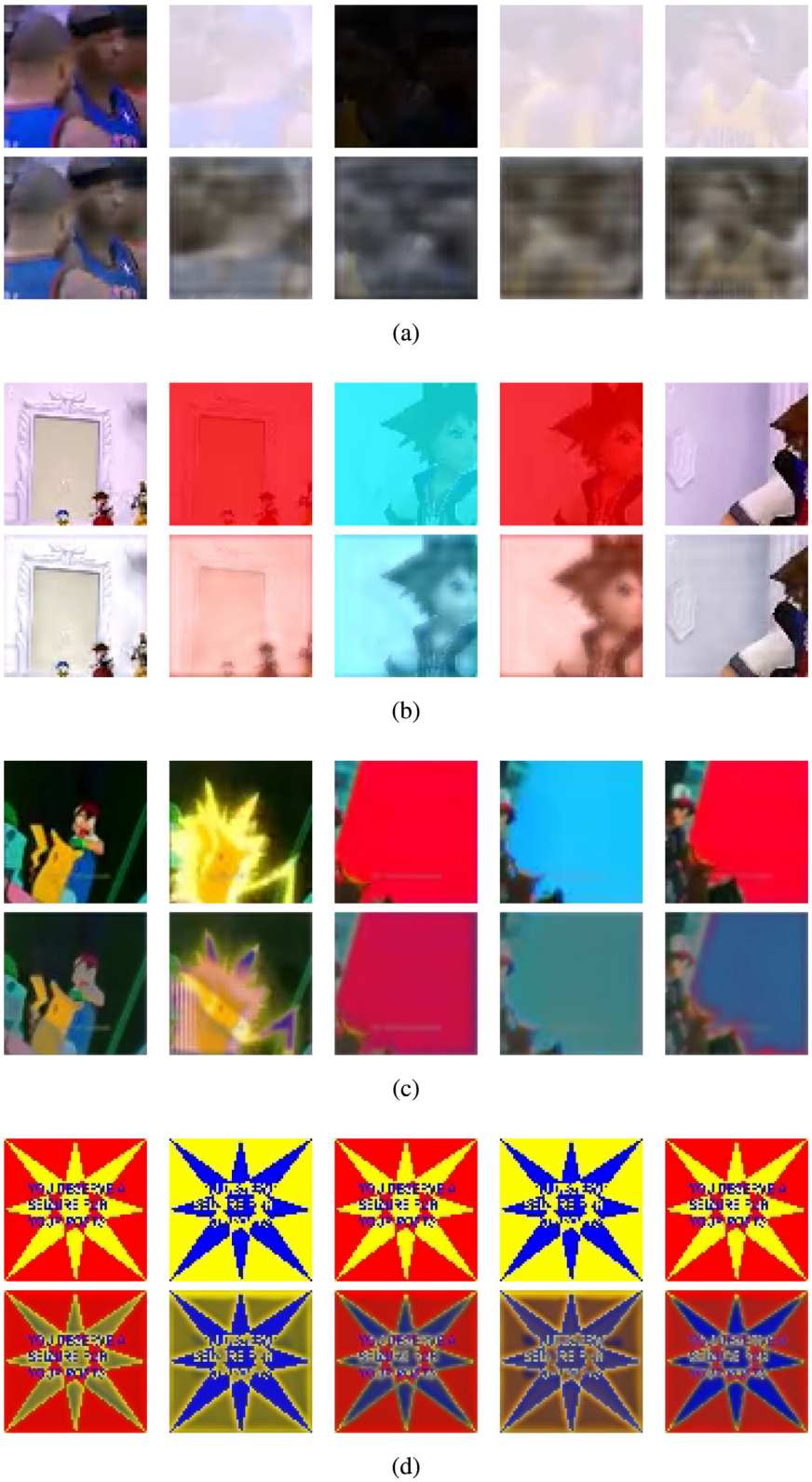

Fig. 3. Five frames from 4 videos with their corresponding transformed versions. Note that frames which are not involved in flashing are preserved while flashes are suppressed in such a way as to not significantly disturb the underlying contents. Videos (c) and (d) are *not* synthetic examples created by our generator, (c) having triggered hundreds of attacks in children [2] and (d) having been used to cause seizures in a politically-motivated attack.

Table 3
Comparing the networks on three types of inputs; conditions denoted by X. The first condition, O, compares against original unmodified inputs where the priority is to keep the stimulus intact meaning we do not desire a preference either way. Similarly, for the next condition, − , where videos were modified to add flashing but with properties that put it below the threshold at which it is likely to cause harm. Finally, + where we see the networks undoing the transformation applied while producing higher quality output and where users prefer their output over the modified videos. Quality is measured as the likelihood of preferring the network output to the output of the condition that is being compared against. Subjects generally judge the quality of the two to be the same, meaning our networks preserve the structure of the videos.

Network	X	Chose video	σ^2	Pref. for X	Pref. for transformed	Quality
LSTM	O	47.8%	7.38	−0.06	+0.01	44.4%
	−	35.0%	16.08	−0.14	+0.01	53.1%
	+	81.8%	17.50	−0.30	+0.56	56.6%
U-Net	O	51.3%	3.68	−0.13	+0.16	27.3%
	−	56.2%	18.85	−0.09	+0.35	37.0%
	+	76.8%	6.15	−0.27	+0.51	55.1%
STA	O	49.2%	6.51	−0.04	+0.06	50.1%
	−	46.5%	18.91	−0.10	+0.09	50.5%
	+	74.2%	12.48	−0.32	+0.58	55.7%

generator may be reducing the total amount of flashing in such cases. The GAN seems to insert subtler flashing but without hiding any flashes or problematic patterns in the original videos. The preference against the original video was smaller when using GANs, indicating that they are inserting more subtle flashes.

While we do want subjects to choose the generated videos that have more flashing, we do not expect, or desire, that subjects mark the modified videos as entirely consisting of flashes. In other words, the optimal results are not −1 and 1 for the preference values, because very often the flashes that are problematic are brief and have fairly low intensity. Note that, qualitatively the GAN-generated videos look very different from the ones produced by the hand-crafted generator making these two methods complimentary. In the results that follow we equally mixed videos generated by both.

4.2.2. How good are the video-to-video transformations?

Next we compare how the networks perform at applying the inverse transformation, i.e., removing the problematic stimuli. We tested each network against a collection of videos. Videos were modified to contain flashing by an equal mixture of the hand generator and the GAN. Subjects judged how effective networks were at mitigating flashes and preserving video quality.

Results are summarized in Table 3. The unmodified video was transformed in the O condition to test how networks affect non-flashing video. Each of the three methods (LSTM, U-Net, STA) had little impact on the original video with chance-level performance (47%, 51%, 49.2%) when deciding which video has less flashing. LSTM and STA were judged to have left the quality of the original videos intact with users preferring them over the modified ones 44.4% and 50% of the time. On the other hand, U-Net, displayed a weak but marked decrease in quality of the original video with only 27% choosing the modified video as having equal or higher quality compared to the original. Since the U-Net used contains many more parameters than the other models, it might be overfitting and inserting minor but noticeable artifacts. Such cases can often be resolved by additional training with far more data.

To further test non-flashing behavior, we explored adding below-threshold flashes in the - condition. LSTM and STA preserved the original content in this case, with quality ratings being around 44% and 50%. U-Net seems to have slightly degraded quality, with subjects having a weak preference (−0.09) against the transformed videos with a significant drop in quality, 37%. While the LSTMs were conservative and less frequently removed flashes, with 35% of videos being selected as having less flashing, the U-Net and STA are more aggressive with 56% and 46% of the videos containing fewer flashes. This is consistent with the above results in the O condition.

When significant flashes were added, the + case, we explore how well problematic changes are undone. In each case, the networks undo significant portions of the transformations applied while maintaining the quality of the original (56%, 55%, 57%) All methods were effective in removing or reducing flashes (81.8%, 76.8%, and 74.2% at the time). Overall the variance of the U-Net is much smaller than that of other networks while still maintaining their positive attributes. A modified U-Net trained on several orders of magnitude more data is likely to be the best approach. With the current amount of data, STA appears to be the best approach, showing balanced results with respect to reducing flashing and keeping the quality of the video high when no flashes are present.

4.2.3. Baselines and ablations

We want to ensure that these results are not due to some artificial correlations in the input videos, despite our efforts to randomize every aspect of the corpus and video transformation creation. We chose one well-performing network, STA, and ablated it. We provided the network with a single frame rather than an entire video, with shuffled frames, 10 video frames, 50 video frames, and 50 frames but only 50 past frames with no future lookahead. These conditions, as expected, significantly degraded performance. When comparing the full model against the ablated models in each case users preferred the full model: 94.4% of the time in the single frame case, 90% of the time in the shuffled frame case, 86.6% of the time in the 10 frame case, 88.5% of the time in the 50 frame case, and 75% of the time in the 50 frame lookahead-only case. When networks receive less data, shuffled videos or single frames in the extreme, they become ineffective. As they gain access to more and more of the video, their performance increases significantly.

4.2.4. Real videos

In Fig. 3(c) and (d) we show frames from videos which are known to have caused seizures. Note that the transformed frames are significantly clearer and easier to understand. We refer users to our website to see the results more clearly as videos, in addition to having a range of results on other videos which have caused similar problems. Additionally, we evaluated these transformed videos with existing tools [30] which attempt to detect flashing in videos. These tools report that while the originals are extremely likely to cause harm, our transformed videos are very unlikely to do so. We do not report precise numbers as these tools are not calibrated with human psychophysics, their numbers are not probabilities, and they are not validated with large-scale studies and thus are only useful qualitatively.

5. Discussion

We have introduced the notion that one should think about accessibility in a sub-symbolic way and learn to transform a visual scene into a form that is safer and clearer for a particular user. We connected this notion to long-standing ideas in vision and general perception about forward and inverse transformations. In general, this is a more flexible view of accessibility than the task-replacement paradigm which dominates prior efforts, as described by Leo et al. [27]. Next, we describe a preliminary effort to applying these principles to another problem: unlocking the content of faces for those who cannot easily perceive emotions. In doing so, we intend to demonstrate that these notions are more flexible than a particular instantiation for photosensitive seizure disorders.

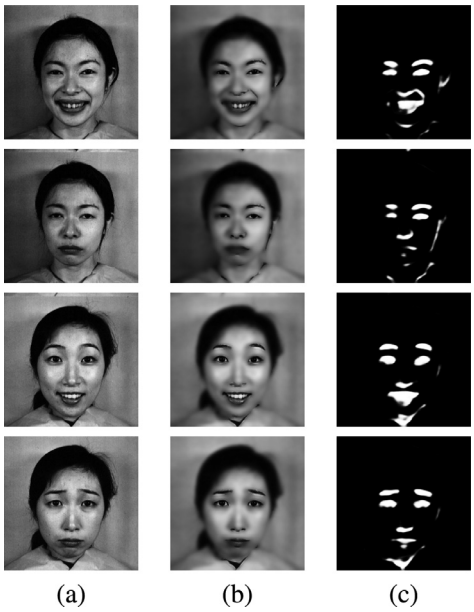

Fig. 4. Preliminary results showing four (a) input faces, (b) faces where important features are retained while the rest of the image is blurred, and (c) masks which show the location of the retained unblurred areas. We use an emotion recognition network as a proxy for the human visual system and optimize its performance while selectively blurring an image of a face. Some face features, like the mouth, are not necessary for all emotions, while features around the eyes are almost always critical.

5.1. Applications to face processing

Deficiencies in face processing are associated with a number of disorders such as autism. Often the underlying issue is not visual impairment but instead the neural mechanisms behind face perception. The Stanford Autism Glass Project [45] builds on this fact and demonstrates that highlighting facial expressions can be extremely useful. It can even lead to learning to better recognize facial expressions in the wild.

In addition to manually annotating faces and their emotional valence, one could learn to highlight the features which are important for most faces. Since those features change with each face, each emotion, and as each emotion is being enacted, this is a dynamic process. This is very similar to the photosensitivty application discussed early, except that rather than removing flashes to enable one to see the rest of a video, we highlight features to encourage a user to pay attention to them. This approach and the one being employed by the Stanford Autism Glass project are complimentary: theirs provides guidance and training to focus on certain facial features and associate them with mental states while the one presented here actively transforms all faces to always guide someone to the most important facial features for a particular emotion.

Crucially, in this application we use a proxy for the visual system rather than generating a dataset directly. That proxy is a convolutional network that recognizes emotions. We motivate this by the fact that convolutional networks seem to approximate part of the human visual system with some fidelity [46], so one might expect that the results of transforming an image to make it easier for a network to understand will similarly make it easier for humans to understand.

We pick random images of human faces from the Japanese female facial expression [47] dataset, and post-process them to create a distribution over the emotional content of each face using a convolutional network [48]. This network is pretrained on the CASIA WebFace [49] dataset and fined-tuned on the Emotion Recognition in the Wild Challenge [50]. Then an image-to-image network, analogous to the video ones described earlier, is trained to transform such faces. The network is a U-Net, as described in Section 3.2.2, which takes as input an image and learns to predict a sparse mask which selectively blurs that image. Its objective function is to produce a peaked unimodal response in the distribution of the output of the emotion recognition network. In Fig. 4 we show four input faces, transformed faces, and the masks showing which areas were not blurred. Significant work remains to refine such transformations, apply them to real-time videos, and test the efficacy of this approach.

5.2. Conclusions

Learning to transform the visual environment is a new tool for accessibility building on prior experience with hand-crafted transformations like daltonization. Just like hand-crafted features have yielded to automatically learned features for tasks like object detection, here we repurpose those techniques to automatically learn new visual transformations with the goal of increasing accessibility. The space of possible transformations for different populations is immense; for example, one could learn to transform text to help those with dyslexia or to subtly manage visual attention for someone with ADHD. Such technology may benefit everyone by slightly altering the visual environment making it more pleasant, more readily accessible, and less distracting – an effect known as the *curb-cut effect* where accessibility technologies can end up helping everyone. In future work, we intend to explore a wide range of applications for this general idea as well as clinically validating the work presented here on photosensitivity and deploy it in real-time to screens everywhere.

Transformations could be customized to different individuals, not just to disabilities or populations. This is particularly important because disabilities and impairments are heterogeneous: they are often not total, they differ in how they affect each individual, and are sometimes associated with other disorders. Rather than providing a small number of preprogrammed customization options, we could rely on human in the loop learning. An individual might be given an automatically generated test to fine-tune the transformation to their particular needs and preferences. The fact that object detectors are easily fine-tuned to new datasets with few examples indicates that such a test would likely be short; moreover one need not go through an entire test to improve the transformation or even stop at any predetermined point. Looking further into the future, one might record the neural activity of users, for example by using EEG, and subtly adjust the transformation to enable them to read more quickly, be less confused, or have fewer EEG artifacts even when no overt reaction is apparent. Since this does not necessarily require user feedback, transformations trained in this way may be of use to those who cannot easily communicate their needs and preferences. In principle, such networks can continue to be trained through a lifetime and adapt to users as they change. We are excited to see the development of new kinds of transformations, not just visual but perhaps involving other modalities, which can create a safer and more accessible world for all.

Conflict of interest

As corresponding author I, Andrei Barbu, hereby confirm on behalf of all authors that:

1. We have no conflicts of interest and no financial support that would influence its outcome.
2. All funding sources have been identified and acknowledged.
3. No intellectual property concerns are present in this publication
4. No conflicting personal relationships exist.

5. All authors have contributed to, reviewed, and approved this submission.

Acknowledgments

This work was funded, in part, by the Center for Brains, Minds and Machines (CBMM), NSF STC award CCF-1231216. We would like to thank Judy Brewer, the director of the Web Accessibility Initiative at the W3C, for her assistance and insight.

References

[1] G.F.A. Harding, P.M. Jeavons, Photosensitive Epilepsy, 133, Cambridge University Press, 1994.
[2] J. Parra, S.N. Kalitzin, F.H.L. da Silva, Photosensitivity and visually induced seizures, Curr. Opin. Neurol. 18 (2) (2005) 155–159.
[3] K. Poulsen, Hackers assault epilepsy patients via computer, Wired News 28 (2008).
[4] Q. Do, B. Martini, K.-K.R. Choo, Cyber-physical systems information gathering: a smart home case study, Comput. Netw. (2018).
[5] G.R. Harding, Tv can be bad for your health, Nat. Med. 4 (3) (1998) 265.
[6] M. Nomura, T. Takahashi, K.-I. Kamijo, T. Yamazaki, A new adaptive temporal filter: application to photosensitive seizure patients, Psychiatry Clin. Neurosci. 54 (6) (2000) 685–690.
[7] D. Damen, T. Leelasawassuk, W. Mayol-Cuevas, You-do, i-learn: egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance, Comput. Vis. Image Underst. 149 (2016) 98–112.
[8] C. Stoll, R. Palluel-Germain, V. Fristot, D. Pellerin, D. Alleysson, C. Graff, Navigating from a depth image converted into sound, Appl. Bionics Biomech. 2015 (2015).
[9] L. Rello, R. Baeza-Yates, Good fonts for dyslexia, in: Proceedings of ACM SIGACCESS, ACM, 2013, p. 14.
[10] J.T. Simon-Liedtke, I. Farup, Evaluating color vision deficiency daltonization methods using a behavioral visual-search method, J. Vis. Commun. Image Represent. 35 (2016) 236–247.
[11] P. Pan, Z. Xu, Y. Yang, F. Wu, Y. Zhuang, Hierarchical recurrent neural encoder for video representation with application to captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1029–1038.
[12] Y. Pan, T. Mei, T. Yao, H. Li, Y. Rui, Jointly modeling embedding and translation to bridge video and language (2016b).
[13] N. Siddharth, A. Barbu, J. Mark Siskind, Seeing what you're told: sentence-guided activity recognition in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 732–739.
[14] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, C. Schmid, Action tubelet detector for spatio-temporal action localization, in: ICCV-IEEE International Conference on Computer Vision, 2017.
[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Action classification in soccer videos with long short-term memory recurrent neural networks, in: International Conference on Artificial Neural Networks, Springer, 2010, pp. 154–159.
[16] O. Shahar, A. Faktor, M. Irani, Space-time super-resolution from a single video, in: Proceedings of the Conference on Computer Vision and Patter Recognition, IEEE, 2011, pp. 3353–3360.
[17] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[18] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, J.C. Niebles, Learning to decompose and disentangle representations for video prediction, arXiv:1806.04166 (2018).
[19] P. Battaglia, R. Pascanu, M. Lai, D.J. Rezende, Interaction networks for learning about objects, relations and physics, in: Advances in neural information processing systems, 2016, pp. 4502–4510.
[20] D.L.W. Hall, D. Klein, D. Roth, L. Gillick, A. Maas, S. Wegmann, Sequence to sequence transformations for speech synthesis via recurrent neural networks, 2018, US Patent App. 15/792,236.
[21] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: CVPR, IEEE, 2016, pp. 2414–2423.
[22] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[23] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[24] A. Piergiovanni, A. Wu, M.S. Ryoo, Learning real-world robot policies by dreaming, arXiv:1805.07813 (2018).
[25] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, et al., Using simulation and domain adaptation to improve efficiency of deep robotic grasping, arXiv:1709.07857 (2017).
[26] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2242–2251.
[27] M. Leo, G. Medioni, M. Trivedi, T. Kanade, G.M. Farinella, Computer vision for assistive technologies, Comput. Vis. Image Underst. 154 (2017) 1–15.
[28] J. Farringdon, V. Oni, Visual augmented memory (VAM), Digest of Papers. Fourth International Symposium on Wearable Computers, 2000.
[29] A. Betancourt, P. Morerio, C.S. Regazzoni, M. Rauterberg, The evolution of first person vision methods: a survey, IEEE Trans. Circt. Syst. Video Technol. 25 (5) (2015) 744–760.
[30] S.J. Park, Y.M. Kang, H.R. Choi, S.G. Hong, Y.S. Kang, Development of image and color evaluation algorithm for the web accessibility evaluation tools, in: Asia-Pacific Conference on Computer Human Interaction (ACCV), Springer, 2008, pp. 389–395.
[31] G.C. Vanderheiden, Quantification of accessibility: guidance for more objective access guidelines, in: International Conference on Universal Access in Human–Computer Interaction, Springer, 2009, pp. 636–643.
[32] Y. Takahashi, T. Sato, K. Goto, M. Fujino, T. Fujiwara, M. Yamaga, H. Isono, N. Kondo, Optical filters inhibiting television-induced photosensitive seizures, Neurology 57 (10) (2001) 1767–1773.
[33] T. Takahashi, Y. Tsukahara, Usefulness of blue sunglasses in photosensitive epilepsy, Epilepsia 33 (3) (1992) 517–521.
[34] S. Blessenohl, C. Morrison, A. Criminisi, J. Shotton, Improving indoor mobility of the visually impaired with depth-based spatial sound, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 26–34.
[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[36] V. Patraucean, A. Handa, R. Cipolla, Spatio-temporal video autoencoder with differentiable memory, arXiv:1511.06309 (2015).
[37] W.W.W. Consortium, et al., Web content accessibility guidelines (WCAG) 2.0(2008).
[38] R.S. Fisher, G. Harding, G. Erba, G.L. Barkley, A. Wilkins, Photic-and pattern-induced seizures: a review for the epilepsy foundation of america working group, Epilepsia 46 (9) (2005) 1426–1441.
[39] T. Takahashi, Y. Tsukahara, Photoparoxysmal response elicited by flickering dot pattern stimulation and its optimal spatial frequency of provocation, Clin. Neurophysiol. 106 (1) (1998) 40–43.
[40] B. Wu, H. Duan, Z. Liu, G. Sun, Srpgan: perceptual generative adversarial network for single image super resolution, arXiv:1712.05927 (2017).
[41] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[42] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[43] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems (NIPS), 2015, pp. 802–810.
[44] S. Winkler, Issues in vision modeling for perceptual video quality assessment, Signal Process. 78 (2) (1999) 231–252.
[45] P. Washington, C. Voss, N. Haber, S. Tanaka, J. Daniels, C. Feinstein, T. Winograd, D. Wall, A wearable social interaction aid for children with autism, in: Proceedings of the Conference Extended Abstracts on Human Factors in Computing Systems (CHI), ACM, 2016, pp. 2348–2354.
[46] D.L. Yamins, J.J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex, Nat. Neurosci. 19 (3) (2016) 356.
[47] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, J. Budynek, The Japanese female facial expression (jaffe) database, in: Proceedings of Third International Conference on Automatic Face and Gesture Recognition, 1998, pp. 14–16.
[48] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 503–510.
[49] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, arXiv:1411.7923 (2014).
[50] A. Dhall, R. Goecke, J. Joshi, K. Sikka, T. Gedeon, Emotion recognition in the wild challenge 2014: Baseline, data and protocol, in: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, 2014, pp. 461–466.