Artificial Intelligence for Privacy Conservation in Remote Learning

HONGBO ZHANG; LEI MIAO; JIA-XING ZHONG; AND AIMIN YAN

COVID-19 resulted in a significant impact on academic learning when schools shut down across the world. Globally, over 1.2 billion children were out of the classroom during the pandemic (Li & Lalani, 2020). Remote learning does offer an alternative learning mode when in-person instruction is infeasible; however, learning outcomes of remote learning are mixed. The effectiveness of remote learning has been compromised due to the lack of in-person interactions between students and instructors. In addition, student privacy is difficult to maintain in front of a camera. This leads to hesitation to show the student's complete profile, including body language and environmental contexts, which therefore reduces the engagement, effusiveness, and immersion of the overall learning process (Yang et al., 2020).

There are numerous methods available for preserving user privacy during remote video calls. Among them, a change of background is often used. Through revision of the background, students can disable their actual background which can contain sensitive and private information. Changing a remote video call's background is made feasible through several different techniques. When replacing the actual background with a virtual background, image matting is one major technique used for this task. This chapter will review different image matting techniques for how image matting is implemented.

For preserving privacy, the suppression of background noise is also important. Background noise is one primary source of privacy concerns. Suppressing background noise and only exposing the student's intended sound to others will improve privacy. Background noise removal is different from complete noise removal. As such, simple noise removal will not work. This chapter will discuss different techniques for recognition of the speaker's sound features and suppression of others. Different techniques involving recurrent neural network, conditional GAN network, and WaveNet will be discussed.

The blurring of shared screen content dynamically and selectively is also an important technique for preserving privacy. For example, a student or teacher may not want other people to visualize the websites that they have visited, their desktop background, or the apps they have downloaded or been using. The current method for achieving this is through a selective sharing mechanism, where students or the teacher share a particular window. However, it can be inconvenient for students to share only a particular window since they may need to switch to different apps dynamically through the process. Therefore, it will be helpful for computers to smartly blur the privacy related content while still showing other content to participants. For this, deep learning-based natural language processing models will be reviewed for how to recognize content specific to the meeting while blurring others.

Methods

Virtual Background for Privacy

Overview of Virtual Background for Privacy

Image matting is one primary technique used for replacing the actual background with a virtual background during remote meetings. Image matting is the process of estimating foreground objects in images and videos. The process starts by estimating the dimensions of the foreground images and then extracting the foreground from the background. In cases of transparent objects, the situation would become complex since the pixels would belong to the foreground and background at the same time. Such objects include human hair and animal fur, which require estimating the transparency values of the object.

The major differences between image matting versus image segmentation is the introduction of the alpha channel. When the alpha value is one, it is pure foreground. Conversely, when the alpha value is zero, it is pure background. When the alpha value is between zero and one, it is part of both the foreground and background, therefore it is a mixed value. However, there are limitations of such a matting-based method. The method is based on the color differences for differentiating the foreground and background. Color can be very different based on different lighting and environmental conditions, even for the same object, hence it is not reliable. Second, the generation of the ground truth for matting is known to be very difficult. It involves labor extensive human-image interaction work for generation of the ground truth. Due to this, the current available datasets are small. Most of them contain small ground truth (around 100 or 1000 images or videos). This further makes the training of image matting quite challenging (Xu et al., 2017).

Methods of Virtual Background for Privacy

Different methods are used for image matting, including sampling-based methods such as ray casting, searching the entire boundary, and sampling from color clusters (Feng et al., 2016; Gastal & Oliveira, 2010). Another method also includes the measurement of the distance of samples from a known pixel to conclude the similarity of the foreground and background (Levin et al., 2007). Similarly, KL divergence and sparse coding approaches are also used for such sampling-based methods (Bu et al., 2018). The average of the foreground and background are assumed to calculate the foreground, background, and alpha value (Levin et al., 2007). More statistically meaningful research has modeled the foreground and background using Gaussian distribution and therefore uses statistical learning methods for image matting (Chuang et al., 2001). Likewise, the matting problem can also be formulated as a Poisson equation form between the foreground and background. The Poisson equation models the matting gradient field and Dirichlet boundary conditions of the foreground and background, hence numerical methods are used to solve the Poisson equation (Sun et al., 2004). Other approaches attempt to formulate the matting problem as an optimization problem. The approach assigns a regularization term to a pixel and then optimizes the belonging of the pixel to either the foreground, background, or both (Levin et al., 2007). Furthermore, deep learning methods have gained ground in image matting. Different methods rely on different mechanisms for image matting. One approach is to learn the TriMap method through the matting process.

GAN networks have also started advancing to generate virtual backgrounds, hence GAN techniques for generation of virtual background will also be discussed. Among them, conditional Generative Adversarial Networks (cGAN) have

been successfully used for background subtraction. Within this method, the input for the generator is the image and background, and the output is the foreground mask. The discriminator learns to differentiate the real from the fake foreground. Studies show that with CDnet 2014 and BMC databases, the proposed cGAN method can achieve appreciated performance for background subtraction (Bakkay et al., 2018). Other cGAN methods seek a different training pipeline for background removal. In this method, the generator is trained to generate images without background and the discriminator is trained to remove background, which is different from traditional background removal where an image-to-semantics process is taken (Wang et al., 2020). For removal of dynamic objects from the scene, a more dynamic approach such as moving object segmentation should be used. With this method, the generator produces dynamic backgrounds similar to the test sequences to increase the generation fidelity. In particular, the training considers the loss function in image space and feature space such that the generated images have superior performance in both these two spaces (Sultana et al., 2022). Compared to convolutional neural network (CNN) methods, the GAN network method is appealing since it can generate the foreground image instead of being primarily limited to segmentation as compared to CNN methods. Free of such limitations, the method is promising for remote learning where obtaining large scale semantic segmentation is a challenge.

For practical background removal for remote learning, real time background removal is critical. Real time background removal requires a simple computation pipeline. For this end, numerous methods have been proposed. Among them, a TriMap-free method requiring less annotation effort is appealing (Sengupta et al., 2020). This method asks the user to first show the background image without users in the scene. It is followed by the generation of the alpha matte with the deep network, along with input of the pre-captured background image. Then, an image of the user is created in the scene as a soft segmentation image, and where motion cues will be considered. Specific consideration of motion cues will make the method practical for remote learning, where the scene is dynamic rather than static. To further improve the fidelity of the background subtraction for immersive remote learning, a self-supervised GAN training pipeline is used. In comparison to a non-GAN-based training pipeline, the discriminator does not need labeled images during the training phase, therefore it is feasible to train on large scale customized data.

Another practical concern of the privacy conservation for remote learning is the dynamic scene of the learning environment. As such, methods for removal of background through dynamic scenes are critical. Moving object segmentation based dynamic background subtraction is therefore proposed for this purpose (Patil et al., 2021). Within this method, multiple frames of the remote learning video are explicitly considered. Hence, the motion cues are built through the consecutive frames of the video. The motion cues are useful in tracking highly dynamic motion objects to achieve ideal background subtraction. Furthermore, the dilated convolution is used to extract multiple scale features from the scene to achieve better feature extraction for learning foreground and background. Then a GAN based generator and discriminator are used to further enhance the background subtraction quality to reduce the subtraction artifact. Results show that the method can remove background objects from dynamic scenes such as walking humans and driving cars. This method is therefore potentially useful for outdoor-based learning environment privacy conservation.

Challenges and Problems of Virtual Background for Privacy

The current background subtraction methods still present a few challenges. Primarily, no background subtraction methods have considered the removal of unexpected objects, such as a child suddenly appearing in the remote learning scene. The current background subtraction methods have only considered the static scene. As such, the abrupt appearance of a child in the remote learning scene is a challenge for the background subtraction to remove. Methods able to effectively track dynamic scenes are better at dealing with this abrupt appearance of an object (Patil et al., 2021). However, the current dynamic scene background subtraction methods have mostly only considered the tracking of foreground objects rather than the background objects; hence, modification of the existing methods is necessary.

The second major challenge is the use consideration of the outdoor environment. At the present, most methods are designed to work for static scenes such as an indoor environment. For outdoor environments, the scene is rather complex. Frequently, moving objects are present in the background, and most existing methods are not ready to handle this condition. The last major challenge is to selectively replace background elements. Often, learners will not want to replace the whole background scene, but the existing methods mostly do not consider this. Therefore, selective background subtraction is difficult to achieve at the present.

Specific to remote learning, students are required to turn on the video camera to expose their test environment for teachers to monitor cheating during test-taking. The desired outcome is for students to take tests without cheating; however, student privacy is invaded through the process. Under this condition, selective exposure of the student's background is critical to ensure teachers can visualize the student's environment while also protecting their privacy. However, in practice, it is extremely challenging to achieve this balanced goal. It is difficult to know which objects are privacy-sensitive in the student's environment. It is also possible that what students believe should be private may actually allow students to cheat. In order to achieve such a balanced goal, a large-scale database needs to be built to understand privacy-critical objects and the objects that allow students to cheat through the remote test. Similarly, once the database is built, deep learning-based methods could be used to train a background subtraction model for selective background subtraction for ideal privacy conservation.

It is also common that teachers may record the remote learning sessions. Through the recording, information related to the student's environment persists in either a local or remote database. Most often, the database is stored in a cloud learning platform such as Blackboard or Canvas. The recorded remote learning session is available to be viewed by all other students. A practical concern is that the recorded video might contain private information. Therefore, it is desirable to remove such *privacy-sensitive* background information. However, until present, it is largely an open research question as to how to define the *privacy-sensitive* information and consequently to remove it. Fortunately, there are a few video background removal models for us to use, which includes a real time video background subtraction method, the segmentation method (Cioppa et al., 2020). Within the segmentation method, a real time object classifier is introduced through consideration of inter frame information cues. It can achieve better semantic segmentation of the foreground objects. In order to achieve real time subtraction of the video background, a simple Manhattan distance (the distance between two points in a grid (Black, 2019)) between the current pixel's color and ground truth is used for making the alpha matting map. Such a simple threshold-based rule enables real time background subtraction. Video background removal is limited to videos seen by computers, which are restricted by the number and types of training videos. For the general-purpose type of videos, an unseen video background removal is proposed (Tezcan et al., 2021). The method takes both spatial and temporal information of the video into consideration and uses data augmentation such as spatial-temporal crop and spatially aligned crop techniques to generalize the types of videos for background subtraction. It is expected that the video background removal will enable long-term privacy conservation for remote learning, thus easing students' and parents' concerns about participating and engaging in remote learning.

Suppression of Background Sound for Privacy

As one of the main potential causes of privacy disclosure, background noise leakage in online calls is increasingly emphasized by both parties in the remote learning process. For the purpose of background noise suppression, speech enhancement (SE) technology is adopted ubiquitously to online meeting software (such as Zoom, Teams, and Skype). For this, the chapter will describe the status quo of AI technology in background noise suppression from the following two perspectives: methods and their performances.

Methods of Suppression of Background Sound

The existing methods can be roughly categorized into two groups: traditional statistical models and deep learning models. The statistical models usually hypothesize that the noisy observations are based on stationary background noises, which makes it highly difficult to deal with real-world scenarios with non-stationary noises. Owing to the strong modeling capacity of deep neural networks (DNNs), it is feasible to apply deep learning to background sound suppression in the non-stationary setting.

Traditional Statistical Models. Ephraim and Malah (1985) proposed a short-time spectral amplitude (STSA) estimator and examined it while enhancing noisy speech. By finding the minimum of the log-spectral mean square error between the original STSA in the speech signal and its estimate, this estimator showed effectiveness in improving the quality of noisy speech. To provide simpler alternatives to the STSA rule, Wolfe and Godsill (2001) presented the Bayesian approaches. Under the same modeling assumptions, these approaches exhibited almost identical behavior to STSA. Compared with the unmodified STSA, they were efficient to implement and yielded intuitive interpretation. Lotter and Vary (2005) devised two spectral amplitude estimators for acoustical background sound suppression. The two estimators were based on the maximum a posteriori estimation and a super-Gaussian statistical model, respectively. These estimators were able to optimally fit the distribution of the speech spectrum for a background sound reduction system. Srinivasan et al. (2007) trained codebooks of speech and noise linear predictive coefficients. Furthermore, they developed both memoryless and memory-based estimators to obtain the minimum mean squared error estimate of the clean speech signal. In this manner, their proposed scheme performed well in a noisy background. For single microphone SE, Reddy et al. (2017) derived a gain function based on super-Gaussian joint maximum a posterior (SGJMAP). In the SGJMAP-based function, a tradeoff parameter is further introduced to customize the listening preference. Experimental results reflected the usefulness of this SGJMAP-based application in real-world noisy backgrounds.

Deep Learning Models. In contrast to conventional researchers on background noise suppression who focus on reducing the minimum mean square error (MMSE), Xu et al. (2014) attempted to find a mapping function between noisy and clean speech signals based on DNNs. Xu et al. regarded SE as a supervised learning task, in which case clean speech is provided as the fitting target on training datasets. This supervised learning paradigm has been adopted by many follow-up works as illustrated hereunder. Luo and Mesgarani (2019) developed a fully convolutional time-domain audio separation network (Conv-TasNet) for end-to-end time-domain speech separation. To separate individual speakers, Conv-TasNet encodes a representation of the speech waveform spectrum and inverts it back to the waveforms via a linear decoder. Likewise, Pandey and Wang (2019) put forward another fully convolutional neural network for real-time SE, which is dubbed a Temporal Convolutional Neural Network (TCNN). Under the supervision in a speaker- and noise-independent way, TCNN encodes a low-dimensional representation of a noisy input frame and decodes the representation to reconstruct clean speech. In addition to amplitude prediction, Yin et al. (2020) address the problem of phase prediction by putting forward a Phase-and-Harmonics-Aware Speech Enhancement Network (PHASEN). As an innovative framework, PHASEN captures long-range correlations along the frequency axis and does well in time-frequency spectrogram reconstruction. Ephrat et al. (2018) introduce a joint audio-visual model to separate a single speech signal from a mixture of audio such as other speakers' voices and background sound. This method shows superiority in audio-only speech separation in cases of mixed speech, and it is a speaker-independent solution (trained once, applicable to any speaker). To enable isolated control over the importance of speech distortion versus noise reduction, Xia et al. (2020) devise two mean–squared error-based loss functions as the learning objectives. By optimizing these two objectives, the model achieves high performance in real-time single-channel speech enhancement. Koyama et al. (2020) propose a STFT-based method and a loss function with problem-agnostic speech encoder (PASE) features. By doing this, their model achieves excellent performance in the task of deep noise suppression. Westhausen and Meyer (2020) combine a short-time Fourier transform (STFT) with a learned analysis and synthesis basis in a stacked-network approach. By training a dual-signal transformation network on 500-hour noisy speech, the STFT-based method can

suppress real-time background noise. Recently, Watcharasupat et al. (2022) have exploited the offset-compensating property of complex time-frequency masks and presented an end-to-end complex-valued neural network architecture. The presented architecture further utilizes a dual-mask technique, thereby simultaneously suppressing background sound and canceling acoustic echo.

Performance of Background Suppression

To evaluate the performance of noise suppression, researchers have proposed both objective and subjective metrics. The former aims to consider the sound quality not influenced by personal feelings, while the latter intends to correlate well with the testing results of human subjectivity.

Objective Metrics. There exist quite a few objective measures, e.g., Speech to Distortion Ratio (SDR) (Nocerino et al., 1985), Signal to Noise Ratio (SNR) (Johnson, 2006), Perceptual Objective Listening Quality Analysis (POLQA) (Beerends et al., 2013), Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), and Virtual Speech Quality Objective Listener (ViSQOL) (Hines et al., 2015). Due to the convenience in definition and calculation, objective speech quality metrics are widely reported by an overwhelming majority of the literature.

Subjective Metrics. However, as pointed out by Reddy et al. (2019), the objective metrics may deviate from the experimental results in the subjective tests conducted by human beings. Therefore, it is necessary to introduce subjective metrics to better reflect the speech quality of human subjectivity. For example, the Deep Noise Suppression Challenge in 2022 (Dubey et al., 2022) is ranked according to the crowd-sourcing measure of the ITU-T Recommendation P.835 with Validation (ITU-T P.835) (Naderi & Cutler, 2020), comprised of three scores for each audio clip: speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). It is likely, however, that *privacy-sensitive* content needs to be evaluated by human subjects in order to fully understand the extent of privacy needs.

Dynamic Blurring or Hiding of Shared Screen for Privacy

Overview of Dynamic Blurring or Hiding Shared Screen

Screen sharing is a typical operation during remote learning. Screen sharing includes the sharing of websites, documents, and videos. Sharing this content is useful to increase the effectiveness of remote learning. One challenge, however, is that the shared screen can include sensitive information (Lieberman, 2020). It is especially challenging to share only the content useful to the remote learning sessions. Because of this, it is common that during remote learning sessions, students sometimes are not willing to share their screens, which compromises the effectiveness of learning. In addition, on the teacher's side, not being able to selectively share a screen is also inefficient for learning because it takes time for a teacher to identify the content that they want to share. Meanwhile, a teacher's personal information may be exposed to all students during the trial-and-error process in finding the right content. As such, selective sharing of the needed content is quite crucial for successful remote learning in terms of both learning effectiveness and privacy conservation.

An effective strategy for a selective sharing of the needed content can be achieved through dynamic blurring or hiding of the shared screen. This process would blur or hide the privacy-critical information while keeping the needed content for others to view. Unfortunately, this concept is rather new, and there are no available applications able to achieve this goal yet. The following sections will outline the steps and fundamental techniques essential to implement dynamic blur and hiding of content for remote learning privacy.

Methods of Dynamic Blurring or Hiding a Shared Screen

The first necessary step to achieve dynamic blurring or hiding the shared screen is to recognize the text of the shared screen. "The text" refers to text anywhere on the computer screen. Through recognition of the text, it is feasible to blur or hide the privacy-sensitive content. For recognition of the screen text, optical character recognition (OCR) techniques can be used. OCR relies on machine learning to automatically recognize the natural language of a scene. Tesseract, originally developed by HP and became open source in 2006, is a well-established OCR app for this purpose (Smith, 2013). Tesseract relies on a series of steps in recognition of text. First, an image is put through an adaptive thresholding to remove non-interested features such as the background. It is followed by page layout analysis and word recognition passes. Typically, two passes are needed for recognition of the text. Post processing is immediately followed to correct the recognized text height and fuzzy text correction, as well as word bigram correction. Word bigram correction is a measure of word sequence in a sentence (Srinidhi, 2019). For example, in the sentence, "I need your help," "help" is the word most likely to follow the word "your." If the OCR gave some word other than "help," Tesseract would automatically correct it. As a commercial ready application, Tesseract is most convenient to be integrated to the remote learning environment to automatically recognize the shared screen text. Once the privacy-sensitive content is recognized, blurring or hiding operation of the shared screen can be executed for preserving privacy.

Recognition of text in the regular scene is not a challenging task for Tesseract. However, recognition of text under a natural scene such as the text on a wall or text on a coffee cup in the shared screen is rather challenging for Tesseract. It is because of this that text does not have regular shapes, sizes, and orientations. It is therefore difficult for Tesseract to recognize it. The situation becomes worse if the lighting conditions vary, such as in a dark environment. To tackle this condition, a more powerful OCR application needs to be created. Recent deep learning-based OCR research has started to tackle this problem. Among them, TextOCR has achieved success for this goal (Singh et al., 2021). TextOCR is a large-scale arbitrary shape text recognition application that has created a 900K large scale database with various sizes, shapes, and orientation texts. It uses faster region-based convolution neural network (RCNN) to localize the text (Girshick, 2015). A text extractor uses a segmentation proposal network to extract the text. The text extractor can extract arbitrary shape and orientation, therefore making it suitable for natural scene OCR text recognition (Liao et al., 2020). It then obtains the OCR text and embeds the text into vector (Hu et al., 2020). Consequently, a pointer-based network is used to organize the sequence of the words to ensure the semantic meaning of the recognized text (Singh et al., 2021). With a large database (900K) for training, also benefiting from the rigorous data processing pipeline, TextOCR is superior for recognition of text in natural scenes. It can achieve the goal of recognizing the shared screen text for preserving privacy in remote learning.

Recognition of the meaning of an image is also important for preserving privacy. For example, a student who has visited a gaming website may not want other students or the teacher to know the games that they have played. It is therefore expected that the gaming image needs to be recognized. For this purpose, classification of an image into *privacy-sensitive* and *non-privacy sensitive* is important. Classification of an image requires the collection and labeling of large-scale images. Image classification is a relatively simple task and well-studied. Therefore, the use of image classification for verification of image sensitivity is practical. The recent deep-learning revolution has largely improved image classification accuracy. State-of-the-art image classifiers have achieved over 85-90% accuracy and (Lu & Weng, 2007) . For CNN-based image classification, EfficientNet is considered superior in its performance and accuracy. EfficientNet uses a neural network search method to search for an ideal network structure. It balances the network depth, width, and feature space resolution. The results of the optimal search have made EfficientNet yield optimal classification. It has achieved 84% top-1 accuracy on the large-scale image database, ImageNet. Meanwhile, the model is seven times smaller and five times faster than Resnet-152 (Tan & Le, 2019). The great improvement of efficiency and small size of the model mean that EfficientNet offers practical real-world use for classification.

CNN-based image classification has been facing significant bottlenecks for further improvement of accuracy. Recent research on deep learning transformer image processing has further revolutionized image classification (Chen et al.,

2021; Dosovitskiy et al., 2020; Zhai et al., 2022; Zhuang et al., 2021). In contrast to CNN-based classification, transformer-based image classification treats images as a series of patches. These patches are further tokenized with position encoding, followed by a transformer encoder to perform normalization and multiple head attention on these patches. Through the process, it is crucial that the multiple head attention extracts global features precisely to ensure the network attends to correct features of the image for image classification. It is proven that such global attention mechanisms are critical for classification accuracy (Zhai et al., 2022; Zhuang et al., 2021). Vision Transformer has shown its excellent performance in model scaling. Trained on the large-scale image database ImageNet, Vision Transformer shows excellent top-1 image classification accuracy of over 90%. Similar performance is also attained by training on a very large image database, where 3 billion images are used for image classification. Similarly, Vision Transformer shows a better performance than convolutional neural network-based classification (Zhai et al., 2022). Such improvement of classification accuracy demonstrates that classification of images for remote learning privacy preservation has become feasible.

It is also feasible that students may not want to show a private video to others. For example, they may wish to hide or blur a funny video about sports or a joke video about a lifestyle. Similarly, some students may share offensive videos which need to be disabled by the meeting host. For this purpose, video classification has become important. Video classification takes the video as input and classifies it into different categories. For privacy preservation in remote learning, the video needs to be classified into either privacy-sensitive or not-privacy-sensitive. Deep learning has achieved great progress for classifying videos. Among this work, convolutional neural network has been shown capable of classifying video with decent success (Karpathy et al., 2014). The convolutional model takes the frames of the videos and inputs them to a classification network, showing that the classification network can classify the video. It further shows that the video can be classified by using only one frame of the video. Pre-training the model on a large and more general video database is also shown to be helpful to improve the classification accuracy. Overall, the model is able to achieve 65% three-fold accuracy of classification, which is reasonable for practical video privacy classification (Karpathy et al., 2014).

To better classify video for privacy preservation, improved video classification accuracy is desirable. The root cause of the lack of video classification accuracy is the use of a single frame of video for video classification. Consequently, use of multiple frames for video classification is likely able to improve the classification accuracy. Research has shown success in capturing such temporal relationships of the video. It is known that the use of Stand-Alone Inter-Frame Attention can capture the intrinsic relation between frames and meanwhile attend to the correct features of the video (Long et al., 2022). Most importantly, the Inter-Frame Attention mechanisms can track video objects across video frames such that it is possible to more accurately classify videos. Results show that such inter-frame video attention can increase video classification accuracy from 65% to 75% for top-1 accuracy. Such improvement is meaningful for privacy preservation in remote learning to better classify videos.

Following the first step of recognizing text, image, and video and correctly classifying its privacy, the second step is to either blur or hide the privacy-sensitive content. For the purpose of blurring the shared content, an image filter is desired. The application of filters to blur privacy-sensitive content is relatively straightforward. Mostly, a filter is applied on the desired content to blur the specified region, achieving the goal. However, hiding *privacy-sensitive* content involves some work to crop the specified region. If the region is at the top or bottom, the crop will be easier to implement, but if the content is at the center of screen, cropping the privacy-sensitive content will yield a blank part representing missing content. Therefore, it will impact the quality of the shared screen. Under this condition, replacement of the privacy-sensitive content with other content is desired. The other content can be an icon or an image the students select so that it will be pleasant to view while also preserving privacy in remote learning.

Challenges of Dynamic Blurring or Hiding Shared Screen

It is worth noting that the blurring of specified *privacy-sensitive* regions may lead to discomfort during the remote learning session. Other students may regard the blurred content as a strange phenomenon. To alleviate this, hiding *privacy-sensitive* content may be more desirable. Meanwhile hiding the specified content is also problematic. Hiding content may introduce flickering of the screen if not implemented correctly, thereby also introducing visual discomfort. As such, care is needed to implement these techniques to ensure that remote learning can be conducted smoothly without any strange feelings associated with it.

Conclusion

This chapter has discussed three methods to preserve privacy in remote learning. The first method involves the use of a virtual background to replace the actual background. The virtual background technique has been widely adopted as the industry gold standard for video sessions in remote learning. While it has been widely used, most software including Zoom and Google Meet are not yet mature in terms of technology. Glitches still exist while using the virtual background, particularly when the actual background is dynamic rather than static, making it challenging to use this technique in remote learning. As such, a review of more state-of-the-art virtual background techniques is meaningful. It is expected that with these new techniques, it will become more feasible to conduct remote learning with a virtual background on/in an outdoor environment or highly dynamic indoor environment.

The second method proposes active noise suppression techniques for removal of background noises to preserve privacy. This chapter has systematically reviewed the conventional methods such as waveform spectrum and Gaussian distribution based statistical methods. Deep learning methods include spectrogram based convolutional neural network methods and Fourier domain based short-time Fourier transform methods. These state-of-the-art methods are proven rather effective in the suppression of background noises. The performance metrics including objective and subjective metrics are also given to evaluate the results of suppression. The advancement of audio and text analysis has shown it is practical to develop these background suppression models to effectively suppress *privacy-sensitive* sound to ensure the remote learning environment is free of privacy concerns.

The third method involves blurring or hiding sensitive shared content which is another critical task for preserving privacy. This chapter has reviewed methods for recognizing text, images, and videos. Through recognition of these types of content, it is feasible to either blur or hide the *privacy-sensitive* content. With state-of-the-art research, understanding the full text of the shared screen is feasible, therefore it is possible to recognize and understand the image and video to preserve privacy. Of course, there is other content such as animation GIF files and PDF attachments that are also *privacy-sensitive*. This content can be converted to either text, images, or videos first. Subsequently, content recognition can be performed to preserve privacy. The major concern of blurring or hiding content is the discomfort in viewing the blurred shared content, which can trigger suspicion from other viewers. As such, alternative operations such as hiding the content with another image may be preferable.

References

Bakkay, M. C., Rashwan, H. A., Salmane, H., Khoudour, L., Puig, D., & Ruicheck, Y. (2018). BSCGAN: Deep background subtraction with conditional generative adversarial networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 4018–4022). IEEE. https://doi.org/10.1109/ICIP.2018.8451603

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., & Keyhl, M. (2013). Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—temporal alignment. *Journal of the Audio Engineering Society*, 61(6), 366–384.

Black, P. E. (2019 February 11). Manhattan distance. In *Dictionary of Algorithms and Data Structures* [online]. https://www.nist.gov/dads/HTML/manhattanDistance.html

Bu, Y., Zou, S., Liang, Y., & Veeravalli, V. (2018). Estimation of KL divergence: Optimal minimax rate. In *IEEE Transactions on Information Theory*, 64(4), 2648–2674. https://doi.org/10.1109/TIT.2018.2805844

Chen, X., Hsieh, C.-J., & Gong, B. (2021). When vision transformers outperform ResNets without pre-training or strong data augmentations. ArXiv Preprint. ArXiv:2106.01548v3. https://doi.org/10.48550/arXiv.2106.01548

Chuang, Y. Y., Curless, B., Salesin, D. H., & Szeliski, R. (2001). A Bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/CVPR.2001.990970

Cioppa, A., Van Droogenbroeck, M., & Braham, M. (2020). Real-time semantic background subtraction. In 2020 *IEEE International Conference on Image Processing* (ICIP), (pp. 3214-3218). IEEE. https://doi.org/10.1109/ICIP40778.2020.9190838

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020 September 28). An image is worth 16×16 words: Transformers for image recognition at scale. In *ICRL* 2021, 1–21.

Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S. E., Thakker, M., Yoshioka, T., Gamper, H., & Aichner, R. (2022). Icassp 2022 deep noise suppression challenge. In *ICASSP 2022 – 2022 IEEE International Conference of Acoustics, Speech and Signal Processing* (ICASSP), 9271–9275.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log- spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2), 443–445. https://doi.org/10.1109/TASSP.1985.1164550

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., & Rubenstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4), 1–11. https://doi.org/10.1145/3197517.3201357

Feng, X., Liang, X., & Zhang, Z. (2016, September 17). A cluster sampling method for image matting via sparse coding. In: B. Leibe, J. Matas, N. Sebe, & M. Welling (eds) *Computer Vision—ECCV 2016*. https://doi.org/10.1007/978-3-319-46475-6_13

Gastal, E. S., & Oliveira, M. M. (2010, June 07). Shared sampling for real-time alpha matting. *Computer Graphics Forum*, 29(2), 575–584. https://doi.org/10.1111/j.1467-8659.2009.01627.x

Girshick, R. (2015). Fast R-CNN. In 2015 *IEEE International Conference on Computer Vision* (ICCV), 1440–1448. http://doi.org/10.1109/ICCV.2015.169.

Hani, M. (2020). Best practices for implementing remote learning during a pandemic. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 93(3), 135–141. https://doi.org/10.1080/00098655.2020.1751480

Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N. (2015 May 17). ViSQOL: An objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 13. https://doi.org/10.1186/s13636-015-0054-9

Hu, R., Singh, A., Darrell, T., & Rohrbach, M. (2020). Iterative answer prediction with pointer-augmented multimodal

transformers for TextVQA. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9989–9999. https://doi.org/10.1109/CVPR42600.2020.01001

Johnson, D. H. (2006). Signal to noise ratio. In *Scholarpedia*, 1(12). http://dx.doi.org/10.4249/scholarpedia.2088

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

Koyama, Y., Vuong, T., Uhlich, S., & Raj, B. (2020). Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks. ArXIV Preprint. ArXiv:2005.11611v3. https://doi.org/10.48550/arXiv.2005.11611

Levin, A., Lischinski, D., & Weiss, Y. (2007, December 18). A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2). https://doi.org/10.1109/TPAMI.2007.1177

Li, C., & Lalani, F. (2020, April 29). *The COVID-19 pandemic has changed education forever. This is how.* World Economic Forum. https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Liao, M., Pang, G., Huang, J., Hassner, T., & Bai, X. (2020 August). Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, 706–722. https://doi.org/10.1007/978-3-030-58621-8_41

Lieberman, J. (2020 May 27). Following pornographic bookmark incident, instructor says UM pushed him to resign. *The Miami Hurricane.* https://www.themiamihurricane.com/2020/05/27/following-pornographic-bookmark-incident-instructor-says-um-pushed-him-to-resign/

Long, F., Qiu, Z., Pan, Y., Yao, T., Luo, J., & Mei, T. (2022). Stand-alone inter-frame attention in video models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3182–3191. https://doi.org/10.1109/CVPR52688.2022.00319

Lotter, T., & Vary, P. (2005). Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Advances in Signal Processing.* 354850. https://doi.org/10.1155/ASP.2005.1110

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International jJournal of Remote sSensing*, 28(5), 823-870. https://doi.org/10.1080/01431160600746456

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Naderi, B., & Cutler, R. (2020). A crowdsourcing extension of the ITU-T recommendation p.835 with validation. ArXiv Prepublication. ArXiv:2010.13200v1. https://github.com/microsoft/P.808

Nocerino, N., Soong, F., Rabiner, L., & Klatt, D. (1985). Comparative study of several distortion measures for speech recognition. In *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 25–28. https://doi.org/10.1109/ICASSP.1985.1168478

Pandey, A., & Wang, D. (2019). TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6875–6879. https://doi.org/10.1109/ICASSP.2019.8683634

Patil, P. W., Dudhane, A., & Murala, S. (2021). Multi-frame recurrent adversarial network for moving object segmentation.

In 2021 IEEE *Winter Conference on Applications of Computer Vision* (WACV) (pp. 2301-2310). IEEE. https://doi.org/10.1109/WACV48630.2021.00235

Reddy, C. K. A., Shankar, N., Bhat, G. S., Charan, R., & Panahi, I. (2017). An individualized super-Gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device. *IEEE Signal Processing Letters*, 24(11), 1601–1605. https://doi.org/10.1109/LSP.2017.2750979

Reddy, C. K., Beyrami, E., Pool, J., Cutler, R., Srinivasan, S., & Gehrke, J. (2019). A scalable noisy speech dataset and online subjective test framework. In *Proc. Interspeech* 2019, 1816–1820. http://dx.doi.org/10.21437/Interspeech.2019-3087

Rix, A., Beerends, J., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In 2001 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 749–752. https://doi.org/10.1109/ICASSP.2001.941023

Sengupta, S., Jayaram, V., Curless, B., Seitz, S., & Kemelmacher-Shlizerman, I. (2020). Background matting: The world is your green screen. In 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 2288-2297). IEEE. https://doi.org/10.1109/CVPR42600.2020.00236

Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., & Hassner, T. (2021). TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 8798–8808. https://doi.org/10.1109/CVPR46437.2021.00869

Smith, R. W. (2013 February 04). History of the Tesseract OCR engine: What worked and what didn't. In *Proc. SPIE 8658, Document Recognition and Retrieval XX*, 865802. https://doi.org/10.1117/12.2010051

Srinidhi, S. (2019 November 27). *Understanding word n-grams and n-gram probability in natural language processing*. Towards Data Science. https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058

Srinivasan, S., Samuelsson, J., & Kleijn, W. B. (2007). Codebook-based Bayesian speech enhancement for nonstationary environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), 441–452. https://doi.org/10.1109/TASL.2006.881696

Sultana, M., Mahmood, A., & Jung, S. K. (2022). Unsupervised moving object segmentation using background subtraction and optimal adversarial noise sample search. *Pattern Recognition*, 129, 108719. https://doi.org/10.1016/j.patcog.2022.108719

Sun, J., Jia, J., Tang, C.-K., & Shum, H.-Y. (2004, August 01). Poisson matting. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers* (pp. 315-321). https://doi.org/10.1145/1186562.1015721

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:6105–6114. https://proceedings.mlr.press/v97/tan19a.html

Tezcan, M. O., Ishwar, P., & Konrad, J. (2021). BSUV-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction. *IEEE Access*, 9, 53849–53860. https://doi.org/10.1109/ACCESS.2021.3071163

Wang, Q., Li, S., Wang, C., Dai, M. (2020). Effective background removal method based on generative adversary networks. *Journal of Electronic Imaging*, 29. http://doi.org/10.1117/1.JEI.29.5.053014

Watcharasupat, K. N., Nguyen, T. N. T., Woon-Seng, G., Shengkui, Z., & Ma, B. (2022). End-to-end complex-valued multidilated convolutional neural network for joint acoustic echo cancellation and noise suppression. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 656–660. https://doi.org/10.1109/ICASSP43922.2022.9747034

Westhausen, N. L., & Meyer, B. T. (2020). Dual-signal transformation LSTM network for real-time noise suppression. In *Proc. Interspeech* 2020, 2477–2481. http://dx.doi.org/10.21437/Interspeech.2020-2631

Wolfe, P. J., & Godsill, S. J. (2001). Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing (Cat. No. 01TH8563)* (pp. 496–499). https://doi.org/10.1109/SSP.2001.955331

Xia, Y., Bruan, S., Reddy, C. K. A., Dubey, H., Cutler, R., & Tashev, I. (2020). Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech, and Language Processing (ICASSP)*, 871–875. https://doi.org/10.1109/ICASSP40776.2020.9054254

Xu, N., Price, B., Cohen, S., & Huang, T. (2017, November 09). Deep image matting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2017.41

Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), 7–19. https://doi.org/10.1109/TASLP.2014.2364452

Yang, Y., Cordeil, M., Beyer, J., Dwyer, T., Marriott, K., & Phister, H. (2020, October 13). Embodied navigation in immersive abstract data visualization: Is overview+detail or zooming better for 3D scatterplots? *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1214–1224. IEEE. https://doi.org/10.1109/TVCG.2020.3030427

Yin, D., Luo, C., Xiong, Z., & Zeng, W. (2020, April 03). PHASEN: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 9458–9465. https://doi.org/10.1609/aaai.v34i05.6489

Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1204–1213. https://doi.org/10.1109/CVPR52688.2022.01179

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., & Beyer, L. (2022). LiT: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18102–18112. https://doi.org/10.1109/CVPR52688.2022.01759

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N. C., Tatikonda, S., Duncan, J. S., & Liu, T. (2021). Surrogate gap minimization improves sharpness-aware training. In *ICLR 2022 Conference*, 1–24.