Revealing True Identity: Detecting Makeup Attacks in Face-based Biometric Systems

Mohammad Amin Arab
Simon Fraser University
Burnaby, BC, Canada

Puria Azadi Moghadam
Simon Fraser University
Burnaby, BC, Canada

Mohamed Hussein
Information Sciences Institute
Marina del Rey, CA, USA

Wael Abd-Almageed
Information Sciences Institute
Marina del Rey, CA, USA

Mohamed Hefeeda
Simon Fraser University
Burnaby, BC, Canada

ABSTRACT

Face-based authentication systems are among the most commonly used biometric systems, because of the ease of capturing face images at a distance and in non-intrusive way. These systems are, however, susceptible to various presentation attacks, including printed faces, artificial masks, and makeup attacks. In this paper, we propose a novel solution to address makeup attacks, which are the hardest to detect in such systems because makeup can substantially alter the facial features of a person, including making them appear older/younger by adding/hiding wrinkles, modifying the shape of eyebrows, beard, and moustache, and changing the color of lips and cheeks. In our solution, we design a generative adversarial network for removing the makeup from face images while retaining their essential facial features and then compare the face images before and after removing makeup. We collect a large dataset of various types of makeup, especially *malicious* makeup that can be used to break into remote unattended security systems. This dataset is quite different from existing makeup datasets that mostly focus on cosmetic aspects. We conduct an extensive experimental study to evaluate our method and compare it against the state-of-the art using standard objective metrics commonly used in biometric systems as well as subjective metrics collected through a user study. Our results show that the proposed solution produces high accuracy and substantially outperforms the closest works in the literature.

CCS CONCEPTS

• **Security and privacy → Biometrics**; • **Computing methodologies → Biometrics**.

KEYWORDS

Presentation Attacks, Biometric Systems, Makeup Removal

ACM Reference Format:
Mohammad Amin Arab, Puria Azadi Moghadam, Mohamed Hussein, Wael Abd-Almageed, and Mohamed Hefeeda. 2020. Revealing True Identity: Detecting Makeup Attacks in Face-based Biometric Systems. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3394171.3413606

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413606

1 INTRODUCTION

Face-based biometric authentication systems are getting popular. They are being deployed, for example, at automated border control systems, airports, remote facilities, and more recently to unlock laptops and smartphones. Unlike other human biometrics such as fingerprint and iris, the face can easily be captured in a non-intrusive way from a distance, which contributes to the popularity of face-based biometric systems. Despite their popularity, these systems are susceptible to various attacks, such as an attacker presenting fake/printed photos, playing recorded videos on another device (e.g., smartphone), wearing facial masks, and changing their facial features by applying makeup.

In this paper, we focus on the makeup presentation attack, which is one of the most effective attacks on face-based biometric systems. Makeup can easily and considerably change the facial features of a person, making him/her appear as a totally different person. Makeup can alter the color and shape of eyebrows, lips, cheeks, beard, and moustache. It can also make a person look older/younger by adding/hiding wrinkles in different face areas. Thus, makeup can result in considerably lower matching scores of face images before and after applying makeup [8]. Therefore, makeup attacks pose a major challenge for face-based biometric authentication systems. It is worth mentioning that makeup can even cause difficulties for a human agent trying to verify the identity of a subject.

Makeup attacks can be broadly divided into two categories [13]: Impersonation and Concealment. In an impersonation attack, a person changes his/her facial features to appear like another person [22]. Impersonation can be used to attack biometric systems that employ white-lists. For example, an attacker can access a restricted facility by changing his/her facial features by applying makeup to look like one of the people who have legitimate access to that facility. Concealment, on the other hand, is used to attack systems that employ black-lists to deny access for people on these lists. In this case, a person tries to change his/her facial features such that the system becomes unable to match him/her against the black-list [9]. For instance, a banned spectator can apply makeup to fool the biometric system to gain access to a football stadium.

In this paper, we propose a novel solution to detect makeup attacks in face-based biometric systems. The high-level idea of our solution is simple: first remove the makeup from the presented face

image and then compare the face images with and without makeup. If the two images differ significantly, then the person may be trying to hide his/her true identity and therefore should be flagged. Realizing this simple idea is, however, quite challenging. First, makeup can come in numerous ways and forms, especially if it is purposely designed to break a security system that is not monitored by human agents. For example, the makeup may not look aesthetically appealing or even plausible; it is worn to fool the security system while there are no humans around to question the potentially unnatural looks. Examples of such makeup are shown in Figure 1b. We refer to this type of makeup as *malicious makeup*, which as shown in the figure, can be very subtle and may not be as colorful as the commonly used makeup, which we refer to as *cosmetic makeup*. We show samples of cosmetic makeup in Figure 1a for comparison. Given the subtlety, variety, and flexibility of applying the makeup anywhere in the face, detecting and removing malicious makeup is much more difficult than detecting and removing cosmetic makeup, which have been addressed in some recent works including [4–6].

The second difficulty in comparing face images with and without makeup is that removing the makeup should not introduce significant distortions. Otherwise, matching the face images after removing the makeup against databases may fail because of the distortions, not because of changing identities. This makes our problem more difficult than current works that remove a cosmetic makeup and apply another style of makeup, e.g., [14, 17, 26], which is known as makeup transfer. In makeup transfer, even if there were some distortions introduced during the makeup removal phase, most of these distortions would be covered by the new makeup. Whereas in our case, we need the faces without makeup to validate their identities. Third, the target domain of our problem is security applications, which have much more stringent requirements than cosmetic applications. Errors in cosmetic applications could annoy some users, whereas errors in security applications can lead to security breaches and/or high false positive/negative rates. Finally, despite all of these complexities, the system should be flexible and robust enough to allow legitimate users to wear normal makeup and still being easily validated.

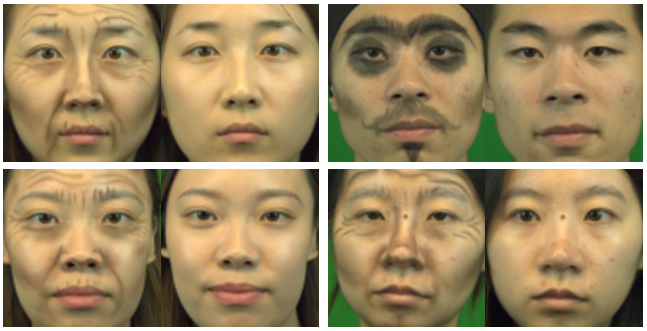
The contributions of this paper can be summarized as follows:

- New method for detecting malicious makeup attacks on face-based biometric authentication systems.
- Subjective study to collect realistic makeup datasets representing different types of potential attacks on face-based biometric authentication systems.
- Rigorous experimental evaluation that shows the robustness and accuracy of the proposed method, compared to the closest works in the literature. Our evaluation assesses standard objective metrics used in biometric systems as well as subjective metrics collected through a user study.

2 BACKGROUND

Face images are among the most commonly-used biometric in authentication systems. Such systems, however, are susceptible to various presentation attacks, including printed faces, artificial masks, prosthetics, and makeup attacks. To defend against various attacks and improve accuracy, authentication systems deployed in high security facilities usually employ additional modalities beside regular

(a) Cosmetic makeup samples from [7]



(b) Malicious makeup samples from our dataset

Figure 1: Cosmetic versus malicious makeup. The right image in each pair shows the same person without makeup.

RGB cameras. For example, stereo cameras can be used to estimate the depth in order to defend against printed faces. The results from different modalities are fused together to improve the accuracy of detecting presentation attacks.

As mentioned in Section 1, the focus of this paper is on detecting malicious makeup that is purposely applied to change the attacker's identity. In many cases, this makeup is very subtle and may not even be visible in regular RGB images. Thus, in this paper, we propose using a simple multispectral camera working in the Short Wave Infrared (SWIR) range of the electromagnetic spectrum. SWIR is a subset of the infrared band in the electromagnetic spectrum, approximately in the $0.9 - 1.7$ μm wavelength range, which is is not visible to human eyes. In the SWIR range, the makeup materials reflect/absorb the electromagnetic waves differently than human skins, regardless of the color of the human skin. Thus, this range provides the potential of detecting different makeup types and styles across various human races and skin colors.

We have performed multiple experiments to identify the most useful SWIR bands for the makeup removal process. Specifically, we used a multispectral camera to capture face images of multiple subjects wearing makeup. This camera (model: Xenics Bobcat 320) captures 6 SWIR bands: 940, 1050, 1200, 1450, 1550, and 1650 nm. More details about our capturing station and experiments are described in Section 5. Our experiments indicate that the 940 nm, 1050 nm, and 1200 nm bands provide the most useful information

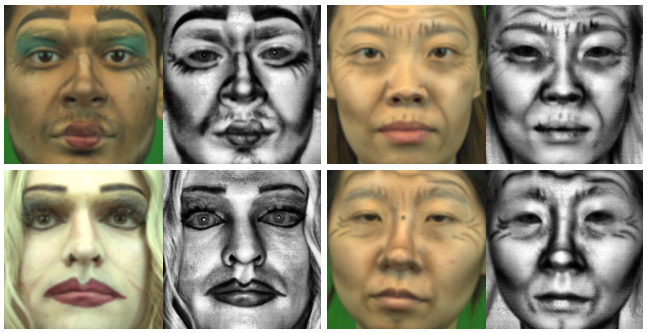


Figure 2: Samples from our experiments demonstrating the potential of using SWIR bands in detecting malicious makeup. Images are shown in pairs: left is the regular RGB image; right is the 1050nm SWIR band.

about makeup, which is complementary to the information present in RGB images. Other bands are either too dark or did not contain enough information. We show some examples from our experiments in Figure 2, which contains pairs of RGB and SWIR images. The SWIR image is the 1050 nm band. As the figure shows, the makeup appears clearly in the SWIR band, regardless of the skin color of the subject.

We note that one of the novel aspects of our work is using SWIR images along with the regular RGB images in detecting and removing subtle makeup to reveal the true identities of persons using face-based biometric systems.

3 RELATED WORK

Multiple works have recently addressed various aspects of handling makeup, including makeup detection, makeup removal, and makeup style transfer. We summarize the closest works in the following.

Makeup Detection: The work in this area heavily relies on colors, as face images in most of the available makeup datasets are colorful around the mouth and eye areas. For example, Rast et al. [19] use biologically-inspired features (BIFs), average skin tone, and histogram of oriented gradient to create makeup descriptive features. Bertacchi et al. [2] use the CMYK color model and neural networks to detect makeups. Chen et al. [6] extract facial features and feed them to a classifier to determine whether the face has makeup. Liu et al. [16] extract features from entropy maps using gradient orientation pyramid (GOP) [15], which are then used for classification. Kotwal et al. [13] propose a deep CNN model to detect age-induced makeup attacks.

Makeup Removal: With the improvement of deep learning in recent years, there has been an increasing attention to the problem of makeup removal. For example, Chang et al. [5] present an unsupervised learning approach based on CycleGAN that transfers images between makeup and no-makeup domains. Cao et al. [4] propose a deep learning bidirectional de-makeup network to remove makeup. And Wang et al. [24] present a locality-constrained coupled dictionary learning algorithm to remove the makeup.

Makeup Style Transfer: Other recent works focus on transferring the style of makeup from one person to another. For example, Li et al. [14] present BeautyGAN, which learns translation on instance-level through unsupervised adversarial learning. Liu et al. [17] utilize a deep localized makeup transfer network to transform makeup from one face image to another.

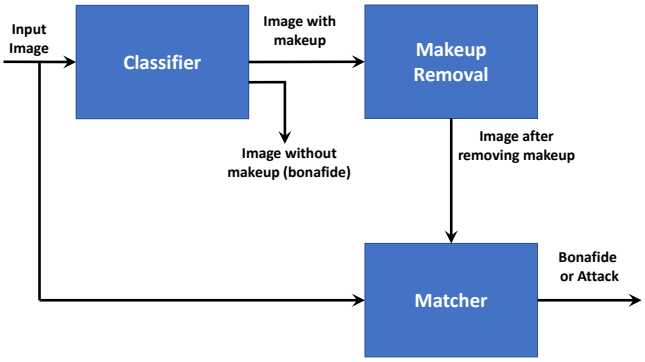


Figure 3: Overview of the proposed solution.

Unlike our work, however, all of the above works focus on *cosmetic* makeup, which is typically colorful and applied to smaller areas of the face, which makes detecting and removing it less challenging. Our work addresses *malicious* makeup, which is not necessarily colorful and can be arbitrarily applied to different areas of the face, since the objective of applying makeup in our case is not to look better, but rather to deceive the biometric authentication system. Examples of faces with cosmetic makeup from prior works are shown in Figure 1a, which are clearly quite different in nature from the malicious makeup samples in Figure 1b.

4 PROPOSED SOLUTION

A high-level overview of the proposed solution for the malicious makeup detection problem is illustrated in Figure 3. Our solution is composed of three components: Classifier, Makeup Removal, and Matcher. The Classifier is designed to separate images with makeup that could potentially be attacks from images that are not likely to be attacks (i.e., images with no makeup or light makeup that does not significantly change the identity of the person). An image with makeup will go through the Makeup Removal component, which tries to remove as much makeup as possible from this image while retaining the essential facial features and without introducing significant visual distortions. The resulting image without makeup is then compared against the input image using the Matcher to decide whether it is a makeup attack.

The details of the Makeup Removal and Matcher components are presented in the following two subsections. The Classifier component was intentionally designed to be simple and conservative in terms of not missing potential makeup attacks even if it comes at the expense of sending some images with light makeup to the Makeup Removal component. The Makeup Removal component is robust enough to handle images with light makeup. The Classifier is implemented as a CNN with 3 layers of convolution with max-pooling after each layer. We used dropout with probability of 20% as well as batch normalization for regularization. We note that our initial design did not include a classifier, which caused confusion for the Makeup Removal component, because it had to process a significant number of bonafides. Trying to remove little or no

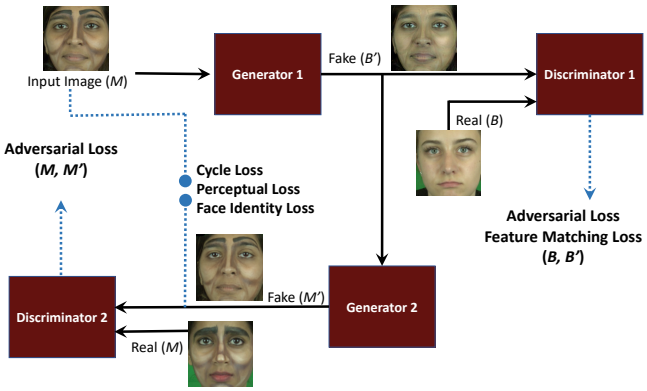


Figure 4: Design of the Makeup Removal component.

makeup from these images resulted in distortions that caused the Matcher to misclassify some of them.

4.1 Makeup Removal

The design of the Makeup Removal component is shown in Figure 4. It is based on CycleGAN [27], but with important changes in the structure of the generator and discriminator as well as multiple new losses added to the network design. CycleGAN provides a general domain-to-domain transformation. It cannot keep the fine details and structures of human faces, which are crucial in our case, especially that we are reconstructing the face of the *same* person, but without makeup. We started with CycleGAN but the results were not good (distorted faces, faces of other people, etc). We did not include these results in the paper, because they represent a weak baseline.

As describe below, we have substantially changed the internal design of CycleGAN: different structures for the discriminator and generator as well as added multiple losses. These changes made CycleGAN more suitable for transforming human faces while retaining their fine details and structures, which is by itself useful for many applications, aside from the makeup removal addressed in this paper.

Discriminator with Dilated Convolution: We changed the structure of the discriminator from the original CycleGAN to use *dilated convolution*, as shown in Figure 5b. Gokaslan et al. [10] show that using dilated convolution in the discriminator helps the generator to learn more about the context and decreases deformation.

Generator using Upsampling: The design of the generator in the original CycleGAN may produce checkerboard artifacts in some training stages, which could be tolerated in other applications by, for example, applying smoothing methods. These smoothing methods, however, hide/blur subtle facial features that are crucial in face-based authentication systems. We experimented with multiple, state-of-the-art, smoothing methods and they produced poor results for our makeup removal problem. As discussed in [18], deconvolution is the main reason for deformation artifacts in the generated images due to the uneven overlapping of kernels on the input image. To mitigate this problem, we present a new design for the generator in the Makeup Removal component, which does not use deconvolution. Rather, it uses upsampling and convolution. The design of our generator is shown in Figure 5a.

We compare the effect of using upsampling and convolution instead of deconvolution on the generated image quality in Figure 6, which shows the improved image quality using upsampling.

Losses: Next, we discuss the various added losses to improve the quality and accuracy of the Makeup Removal component. Recall that we are trying to map an image from domain M (Makeup attacks) to domain B (Bonafide images). CycleGAN provides two mappings $G_1 : M \rightarrow B$ and $G_2 : B \rightarrow M$. Furthermore, the discriminators try to distinguish between images (M or B) and their respected translated images ($G_1(M)$ or $G_2(B)$).

The basic CycleGAN [27] uses adversarial and cycle consistency losses [11]. These losses are not sufficient to capture the complexities of human faces and the fine details of different makeup styles, especially the malicious makeup that is designed to attack an authentication system. To address this problem, we added the following three new losses.

- *Perceptual loss:* which uses the structural similarity metric (SSIM) [21] to preserve the shape and structure of the reconstructed faces.
- *Feature matching loss:* ensures that the distribution of the generated images matches that of real images. It is the L1 norm between the feature maps (FMap) extracted by the last layer of discriminator for real and fake images.
- *Face identity loss:* guarantees a person's face is not changed after going back to the original domain. It is the L1 norm between the original and generated face features extracted from the last layer of the LightCNN [25].

Figure 4 shows how these losses are integrated with the Makeup Removal component.

In the Supplementary Materials submitted with this paper, we present the mathematical formula for each of these losses. We also present the results of an ablation study we conducted to analyze the impact of each loss.

Finally, we note that the goal of this paper was not to design new losses, as there are already many of them in the literature. We have experimented with several other losses, but they did not produce good results. We are not claiming contributions in this direction. Rather, we are proposing a new way of integrating three losses into a CycleGAN that we have carefully modified its discriminator and generator to produce accurate transformations of human faces.

4.2 Matcher

The existence of makeup, whether minimal/cosmetic or extensive/malicious, significantly affects the performance of face matching systems. Our initial design used a commercial matcher (VeriLook SDK - Neurotechnology), which produced poor accuracy. We propose a matcher design using local binary pattern (LBP) descriptors [1]. Matchers built using LBP descriptors were shown to outperform others for matching faces when there is makeup present [8].

LBP uses the difference between the intensity of the center pixel and its neighboring pixels in each $n \times n$ blocks. LBP for each pixel

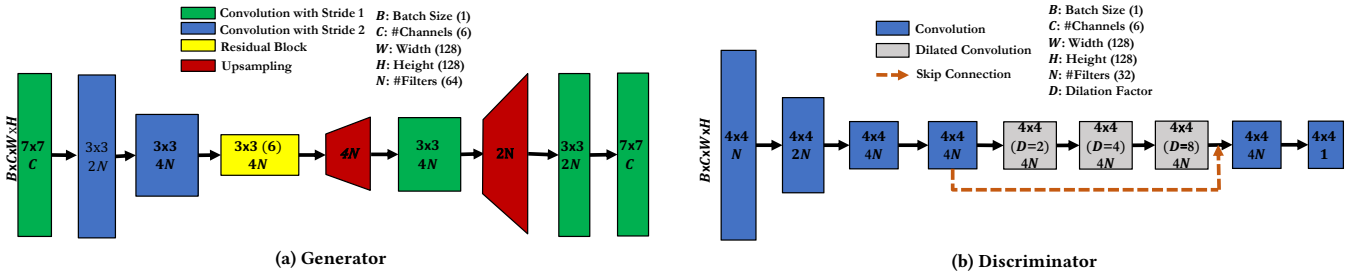


Figure 5: The structures of the generator and discriminator of the Makeup Removal component.

Figure 6: Effect of using upsampling (middle) instead of deconvolution (right) on the quality of the generated images from the input image (left).

x_c and its neighboring pixels x_i's is defined as:

$$LBP(x_c) = \sum_{i=0}^{n^2-1} \mathbb{I}(x_i - x_c > threshold) \times 2^i, \quad (1)$$

where $\mathbb{I}$ is the indicator function which returns one if the condition $x_i - x_c > threshold$ is satisfied and zero otherwise; the $threshold$ is defined by the user. Each LBP coded image is divided into multiple sub-regions and the histogram is calculated. Then, the histogram intersection similarity measure (φ) for comparing two LBP features H^{after} and H^{before} is computed as:

$$\varphi(H^{after}, H^{before}) = \sum_{i=1}^{L} \min(h_i^{after}, h_i^{before}), \quad (2)$$

where L is the number of histogram bins. Finally, the similarity measure is converted to a score to determine whether an image after removing the makeup is too different from the input image.

5 EVALUATION

5.1 Experimental Setup

Capturing Station: We setup a capturing station with an RGB camera and a multispectral camera. The model of the RGB camera is Intel RealSense D435 and has a resolution of $1,920 \times 1,080$. The model of the multispectral camera is Xenics Bobcat 320 and it operates in the $900 - 1700$ nm range, with spatial resolution of 320×256 pixels and frame rate of up to 100 Hz. We configured our station to capture 6 SWIR bands at: 940, 1050, 1200, 1450, 1550, and 1650 nm. As we discussed in Section 2, the most useful SWIR bands are 940, 1050, 1200 nm, and they all produce very similar images. We use these three bands as a form of data augmentation in the training of our model, although one SWIR band is sufficient for practical systems. The data is captured as videos and the middle frame is picked to eliminate closed eyes and squinting of subjects.

The capturing station has illumination boards with LEDs emitting light at wavelengths 940, 1050, 1200, 1450, 1550, and 1650 nm. The illumination boards are synchronized with the two cameras through a microcontroller. A chair was placed in front of the station for a subject to sit on during capturing at a distance of about 62cm, which was in focus for both cameras.

Data Collection: We collected realistic makeup data to evaluate the proposed method. The data collection process was approved by the Research Ethics Boards of our institutions. We recruited a professional makeup artist to apply different types of makeup on subjects. We divide the makeup types into four groups: old age, contour, fake mustache, and extensive. Old age makeup adds artificial wrinkles to the face. Contour makeup adds different shades to the skin of the face to change its visual shape and it may add fake eyebrows. Fake mustache makeup adds dark marks around the lips and it may add shade/marks around the eyes. Extensive makeup covers most of the face with layers of makeup materials and it may add artificial contours and eyebrows to the face. Examples of these makeups are shown in Figure 7.

The RGB photo of each subject was taken before applying any makeup, and it is referred to as the bonafide image. Then, the makeup artist applies one of the makeup types on the subject. Then, the RGB and SWIR images of the subject are taken by the capturing station. If the same subject participates multiple times, we apply a different type of makeup in each case. Then, the whole experiment is repeated for another subject.

In total, our dataset has 466 samples from 73 subjects (38 males and 35 females) who have different ethnicities and skin colors and cover different age groups. We apply multiple types of makeups, and we capture several shots with different poses for each makeup application. The data collection process lasted for several days. The final dataset contains the following:

- 233 samples (half of the dataset) with no makeup (labeled as bonafides),
- 193 samples labeled as makeup attacks, of which 95 samples have contour makeup, 43 samples have extensive makeup, 46 samples have old-age makeup, and 9 samples have fake mustache makeup, and
- 40 samples with cosmetic makeup, which are labeled as bonafides.

Figure 7: Samples of makeup attacks: (i) top-left: old-age makeup, (ii) top-right: contour makeup, (iii) bottom-left: fake mustache, and (iv) bottom-right: extensive makeup.

We note that previous makeup datasets mostly contained white female subjects, which may result in inaccurate results because of the introduced bias during training the models. On the other hand, our dataset contains females and males in different poses and facial states (e.g., faces with and without smiles, different head positions, eyes looking at various directions, etc.). Furthermore, the images were captured with different illuminations and background colors. As a result, we believe that our dataset is more representative of realistic scenarios.

Image Preprocessing and Model Training: We use the CNN-based face aligning method introduced by Bulat et al. in [3] to align SWIR and RGB images of each subject due to a slight difference in the angle of the SWIR and RGB cameras. After alignment, the images are cropped and resized to 256 x 256 pixels.

We divided our makeup dataset into three sets as follows:

- Training Dataset: 214 samples, of which 107 samples labeled as makeup attacks, and 107 have no makeup (labeled as bonafides).
- Validation Dataset: 92 samples, of which 39 samples labeled as makeup attacks, 7 with cosmetic makeup (labeled as bonafides), and 46 have no makeup (labeled as bonafides).
- Testing Dataset: 174 samples, of which 47 samples labeled as makeup attacks, 40 with cosmetic makeup (labeled as bonafides), and 87 have no makeup (labeled as bonafides).

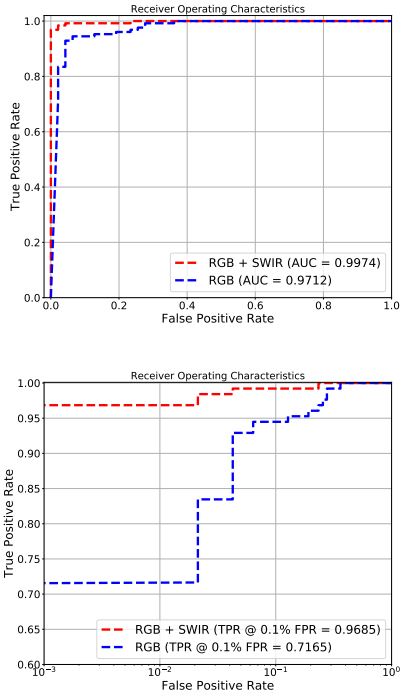
We note that the Testing Dataset contained images of different subjects than those in the Training and Validation Datasets.

We trained the network for 600 epochs and used the model that resulted in the lowest loss on the Validation Dataset. The details of our model parameters are given in the Supplementary Materials submitted with this paper.

5.2 Performance of the Whole System

We analyze the accuracy of the proposed makeup attack detection system using objective metrics commonly used in biometric authentication systems.

First, we plot the receiver operating characteristic (ROC) curve produced by our system (denoted by RBG+SWIR) on the Testing Dataset in Figure 8a. In the same figure, we also plot the ROC curve produced by our system using only the RGB images without any



(b) Zoomed–Log Scale

Figure 8: Performance of the proposed solution.

SWIR data (denoted by RGB). The figure shows that the area under the curve (AUC) for our system is 0.9974, indicating high accuracy. The AUC drops to 0.9714 when our system uses only RGB images. While it is still good accuracy, using only RGB images will lead to higher false positive (attack) rates, which is not desirable for authentication systems. To illustrate this, we focus on the range of false positive rates of less than 1% and plot it on a log-scale in Figure 8b. The figure shows that our system using RGB+SWIR channels produces a true positive rate (TPR) of 0.9685 when the threshold for the false positive rate (FPR) is set to 0.1%, which is high enough for most practical authentication systems. On the other hand, the TPR at the same FPR of 0.1% drops to 0.7165 when we use only RGB images.

We note that the evaluation of the whole system includes the Classifier and Matcher components. For the Classifier, we set the threshold such that the false negative rate is zero for the Validation Dataset, which is conservative to minimize the chances of missing attacks. For the Matcher, we used the Equal Error Rate (EER) computed on the Validation Dataset to find the best threshold. EER is the point on the ROC curve at which TPR equals FPR.

Next, we measure the metrics recommended by the ISO/IEC 30107-3 standard [12] for evaluating biometric systems:

- Attack Presentation Classification Error Rate (APCER): the proportion of attack presentations incorrectly classified as bonafide presentations.
- Bonafide Presentation Classification Error Rate (BPCER): the proportion of bonafide presentations incorrectly classified as presentation attacks.

- The Average Classification Error Rate (ACER): the average of APCER and BPCER.

Table 1 shows that our system achieves low error rates, using the standard performance metrics.

BPCER	APCER	ACER
2.37%	2.13%	2.25%

Table 1: Performance of the makeup attack detection system using standard metrics.

Finally, we show in Figure 9 the few cases in which our system failed: one false negative (missed makeup attack) and three false positives (bonafides classified as attacks). For the false negative case, although the makeup was partially removed, the matching score was slightly higher than the threshold. This is because the face image has a fake mustache and our dataset contained only 9 samples for fake mustaches, which were not enough to train the model to detect various shapes of fake/real mustaches. The false positive cases are mostly due to some distortions introduced in the produced images that made the matching scores slightly lower than the threshold. Although our dataset contained subjects from different ethnicities, some ethnicities did not have enough samples, like the ones shown in the figure. We believe that training with a dataset that contains more samples from different ethnic groups will further improve the performance.

Figure 9: The four failed cases resulted from our model: 1 false negative (top-left corner) and 3 false positives. Images are shown in pairs, where the left image is the input and the right image is the output of our model.

5.3 Comparisons against State-of-the-Art

We compare the key component of the proposed system (the Makeup Removal component) against the closest state-of-the-art makeup removal and style transfer methods in the literature [4, 14, 20]. The first method is the Bidirectional Tunable De-Makeup (referred to as BTD) network introduced in [4], which removes the makeup from faces. The second method is BeautyGAN [14], which transfers the style of the makeup from one face to another using unsupervised adversarial learning. We also compare against [20], which presents a framework for face manipulation in general. This method learns the equivalent residual images which are the difference between images before and after domain transformation using GANs. We implemented this method and used it for makeup removal, and we refer to it as RES (short for residual). We modified all methods to use RGB and SWIR bands (same as our method) for the sake of fairness and we trained all methods on the same datasets.

Comparison using Objective Metrics: We first compare all considered methods using two common objective metrics: structural similarity index (SSIM) and local binary patterns descriptor (LBP). SSIM measures the similarity between two images (the bonafide image and the image generated by the makeup removal component). LBP is a visual descriptor used to compare images [23], which we use to assess the similarity between the bonafide and generated images. Table 2 shows that our system produces better results than the others in both metrics.

	SSIM	LBP(normalized)
BeautyGAN [14]	0.6392	0.4633
BTD[4]	0.6939	0.4631
RES [20]	0.5805	0.1094
Ours	**0.7173**	**0.4665**

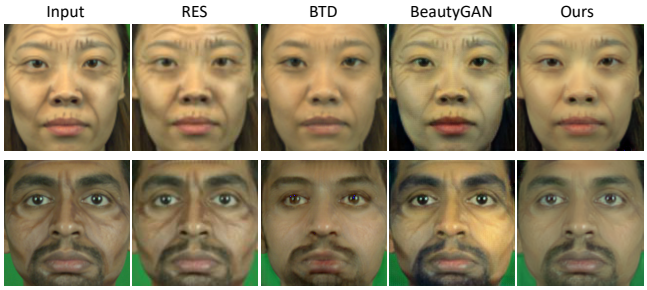
Table 2: Performance of the Makeup Removal component against the closest works in literature.

Comparison using Visual Samples: Next, we visually compare the images produced by our Makeup Removal component versus those produced by the other three works. In Figure 10, we present sample representative results, grouped based on the makeup type. As the figure shows, our method consistently outperforms other methods across all makeup types. Specifically, our method removes most of the makeup and introduces much less distortions in the generated images than other methods.

Comparison using Subjective Study: Finally, we conducted a subjective study to compare the quality of the produced images by our Makeup Removal component versus the images produced by other methods. We designed a simple web form displaying one row of images, where the input image with makeup is on the left and the four images produced by the different methods are on the right (the form is shown in the Supplementary Materials). Under the row of images, there are four rows of radio buttons for the participants to rank each produced image from 1 (Highest Quality) to 4 (Lowest Quality), based on whether the makeup was removed properly and the amount of distortion introduced. The order of the output images was randomized for each sample and for each participant, and the names of the removal methods were not shown.

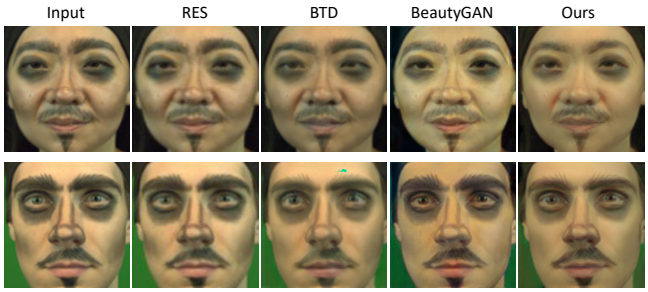
A total of 57 subjects participated in this study, 65% classified themselves as male, 33% as female, and 2% did not specify gender. The participants have various technical backgrounds and are from different age groups: 30% are between 18−24 years old, 56% between 25 − 35, and 14% are older than 35.

The summary of the results is given in Table 3, which shows that more than 77% of the participants ranked the images produced by our method as the highest. Adding the first two columns in the table indicates that more than 94% of the participants ranked our method either the highest or the second highest in terms of quality.

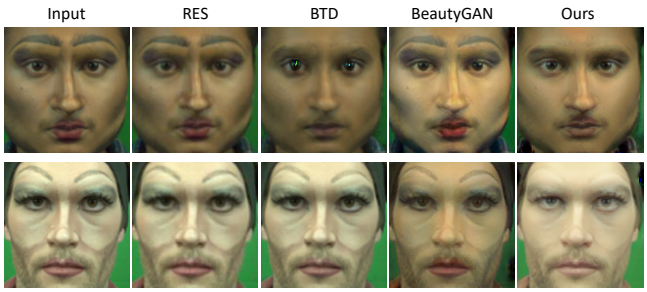


(a) Old Age Makeup Attacks

(b) Contour Makeup Attacks



(c) Fake Mustache Makeup Attacks



(d) Extreme Makeup Attacks

Figure 10: Comparison of the proposed Makeup Removal method against the closest works in the literature: RES [20], BTD [4], and BeautyGAN [14].

6 CONCLUSIONS

We presented a new solution to detect one of the hardest attacks on face-based biometric authentication systems: makeup attacks. We designed a makeup removal model based on generative adversarial networks, but with several critical changes to address the difficulty of removing the makeup while not changing the identity of the input face images or introducing extensive distortions in the output. Our complete solution included a classifier to separate images with makeup from others with light or no makeup that does not change the identity of the person. It also included a customized matcher to compare images with and without makeup. The idea of our system is novel in the face-based biometrics domain: detecting makeup attacks by first reconstructing the faces without makeup and robustly measuring the differences between faces with and without makeup. This is important because it allows the biometric system to detect various combinations of makeups, which can be numerous and changing with time. This is in contrast to prior systems, which train models on 'labeled data'. For such systems to be of practical use, huge amounts of labeled data are needed to cover possible makeup combinations.

	Rank 1	Rank 2	Rank 3	Rank 4
RES[20]	3.72%	10.01%	28.79%	**57.48%**
BeautyGAN[14]	2.58%	24.87%	**48.19%**	24.36%
BTD[4]	15.89%	**48.40%**	20.33%	15.38%
Ours	**77.81%**	16.72%	2.68%	2.79%

Table 3: Results of the subjective study comparing our Makeup Removal component versus others in literature.

We collected a unique dataset of what we call malicious makeup, which is a makeup purposely applied to deceive security systems, especially unattended ones where there are no humans to question the potentially weird looks of the makeup. We evaluated the proposed solution using standard metrics such as the Attack Presentation Classification Error Rate (APCER) and Receiver Operating Characteristic (ROC) curve. Our results show that the proposed solution is fairly accurate. For example, it achieves an APCER value of 2.13% and a true positive rate (TPR) of 0.9685 when the threshold for the false positive rate (FPR) is set to a very low value of 0.1%. In addition, we compared the proposed makeup removal model against the closest three works in the literature using a subjective study. Fifty seven subjects participated in the study, each evaluated 17 makeup removal cases covering different types of makeup. The results show that about 78% of the participants ranked the images produced by our method as the highest in terms of removing most of the makeup while not distorting or changing the identity of face images.

7 ACKNOWLEDGMENTS

This research is support in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. 2006. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 12 (Dec 2006), 2037–2041.

[2] Marcello Guariento Bertacchi and Ismar Frango Silveira. 2019. Facial Makeup Detection using the CMYK Color Model and Convolutional Neural Networks. In *Proceedings of the XV Workshop de Visão Computacional (WVC)*. 54–60.

[3] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). *International Conference on Computer Vision* (2017).

[4] C. Cao, F. Lu, C. Li, S. Lin, and X. Shen. 2019. Makeup Removal via Bidirectional Tunable De-Makeup Network. *IEEE Transactions on Multimedia* 21, 11 (Nov 2019), 2750–2761.

[5] H. Chang, J. Lu, F. Yu, and A. Finkelstein. 2018. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 40–48.

[6] C. Chen, A. Dantcheva, and A. Ross. 2013. Automatic facial makeup detection with application in face recognition. In *Proceedings of the International Conference on Biometrics (ICB)*. 1–8.

[7] Cunjian Chen, Antitza Dantcheva, Thomas Swearingen, and Arun Ross. 2017. Spoofing faces using makeup: An investigative study. In *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, 1–8.

[8] A. Dantcheva, C. Chen, and A. Ross. 2012. Can facial cosmetics affect the matching accuracy of face recognition systems?. In *Proceedings of the IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 391–398.

[9] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. 2020. Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network. *IEEE Transactions on Information Forensics and Security* 15 (2020), 42–55.

[10] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. 2018. Improving shape deformation in unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 649–665.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*. 2672–2680.

[12] ISO 30107-3:2017 2017. *Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*. Standard. International Organization for Standardization, Geneva, CH.

[13] K. Kotwal, Z. Mostaani, and S. Marcel. 2019. Detection of Age-Induced Makeup Attacks on Face Recognition Systems Using Multi-Layer Deep Features. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2019).

[14] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. BeautyGAN: Instance-Level Facial Makeup Transfer with Deep Generative Adversarial Network. In *Proceedings of the 26th ACM International Conference on Multimedia*. 645–653.

[15] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. 2010. Face Verification Across Age Progression Using Discriminative Methods. *IEEE Transactions on Information Forensics and Security* 5, 1 (March 2010), 82–91.

[16] K. Liu, T. Liu, H. Liu, and S. Pei. 2015. Facial makeup detection via selected gradient orientation of entropy information. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 4067–4071.

[17] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. 2016. Makeup like a superstar: Deep localized makeup transfer network. *arXiv preprint arXiv:1604.07102* (2016).

[18] Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and Checkerboard Artifacts. *Distill* (2016). https://doi.org/10.23915/distill.00003

[19] S. Rasti, M. Yazdi, and M. A. Masnadi-Shirazi. 2018. Biologically inspired makeup detection system with application in face recognition. *IET Biometrics* 7, 6 (2018), 530–535.

[20] Wei Shen and Rujie Liu. 2017. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4030–4038.

[21] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel. 2017. Learning to generate images with perceptual similarity metrics. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 4277–4281.

[22] H. Steiner, A. Kolb, and N. Jung. 2016. Reliable face anti-spoofing using multispectral SWIR imaging. In *Proceedings of the International Conference on Biometrics (ICB)*. 1–8.

[23] X. Tan and B. Triggs. 2010. Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. *IEEE Transactions on Image Processing* 19, 6 (2010), 1635–1650.

[24] Shuyang Wang and Yun Fu. 2016. Face behind makeup. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

[25] X. Wu, R. He, Z. Sun, and T. Tan. 2018. A Light CNN for Deep Face Representation With Noisy Labels. *IEEE Transactions on Information Forensics and Security* 13, 11 (Nov 2018), 2884–2896.

[26] L. Xu, Y. Du, and Y. Zhang. 2013. An automatic framework for example-based virtual makeup. In *Proceedings of the IEEE International Conference on Image Processing*. 3206–3210.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

8 SUPPLEMENTARY MATERIALS

This section provides additional information that we could not include in the paper itself because of space limitations.

8.1 Losses and their Impact on the Model

This subsection provides the details of the losses used in our Makeup Removal component.

If we denote the data distribution as $m \sim p_{makeup}(m)$ and $b \sim p_{bonafide}(b)$, we can formulate the loss function as follows:

$$L = L_{adv} + L_{cycle} + L_{perc} + L_{feature} + L_{face}, \tag{3}$$

where

$$\begin{aligned}
L_{adv} &= E_{b\sim p_{bonafide}(b)}[\log D_2(b) + \log(1 - D_1(G_1(b)))] \\
&\quad + E_{m\sim p_{makeup}(m)}[\log D_1(m) + \log(1 - D_2(G_2(m)))], \\
L_{cycle} &= E_{m\sim p_{makeup}(m)}[\|G_2(G_1(m)) - m\|_1 \\
&\quad + E_{b\sim p_{bonafide}(b)}[\|G_1(G_2(b)) - b\|_1], \\
L_{perc} &= MS - SSIM(m, G_2(G_1(m))) \\
&\quad + MS - SSIM(b, G_1(G_2(b))), \\
L_{feature} &= E_{m\sim p_{makeup}(m)}[\|FMap_2(b) \\
&\quad - FMap((D_2(G_2(m))))\|_1] \\
&\quad + E_{b\sim p_{bonafide}(b)}[\|FMap_1(m) \\
&\quad - FMap((D_1(G_1(b))))\|_1], \\
L_{face} &= E_{m\sim p_{makeup}(m)}[\|LightCNNFMap(G_2(G_1(m))) \\
&\quad - LightCNNFMap(m)\|_1] \\
&\quad + E_{b\sim p_{bonafide}(b)}[\|LightCNNFMap(G_1(G_2(b))) \\
&\quad - LightCNNFMap(b)\|_1].
\end{aligned} \tag{4}$$

Ablation Study. We have conducted an ablation study to analyze the impact of each of these losses. Sample results are shown in Figure 11.

Figure 11: Effect of the added losses on the generated face images. The left image is the input with makeup. The second image from the left is produced by the model before adding any of our losses. The third, fourth, and fifth images show the impact of each of the three losses.

8.2 Training Details

This subsection lists the values of all parameters used in the training of our deep learning model.

8.3 Subjective Study Form

This subsection provides more details on our subjective study.

Figure 12 shows a screenshot of the introductory page of the subjective study, and Figure 13 shows the form that the subjects completed.

Image size	256×256
Number of training samples	214
Number of validation samples	92
Number of test samples	174
Number of residual blocks in the generator	9
Number of filters in the first layer of the generator	64
Number of filters in the first layer of the discriminator	64
Probability of the dropout	20%
Batch normalization	after each layer

Table 4: The details of training our deep learning model and its hyperparameters.

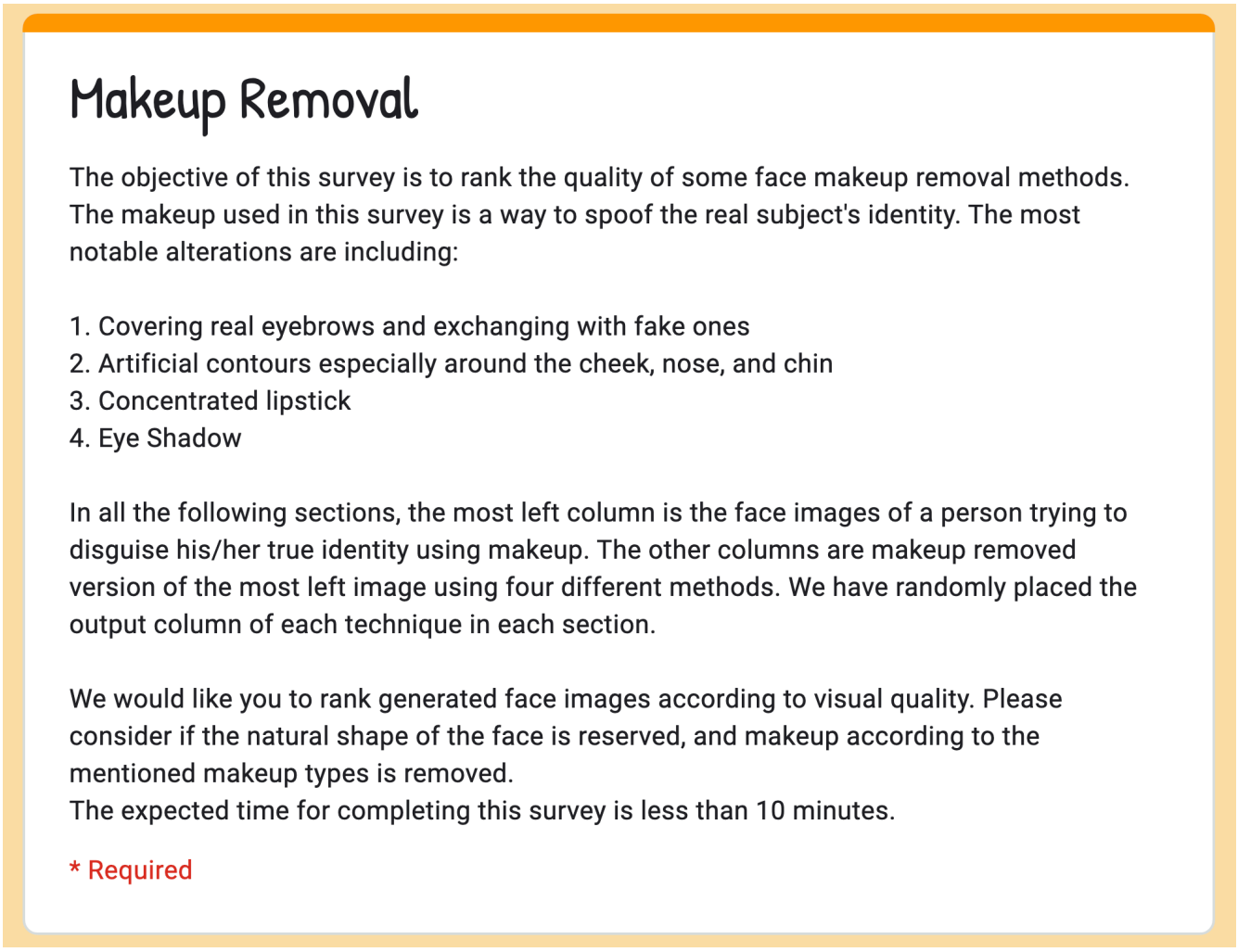


Figure 12: The introduction page of the subjective study.

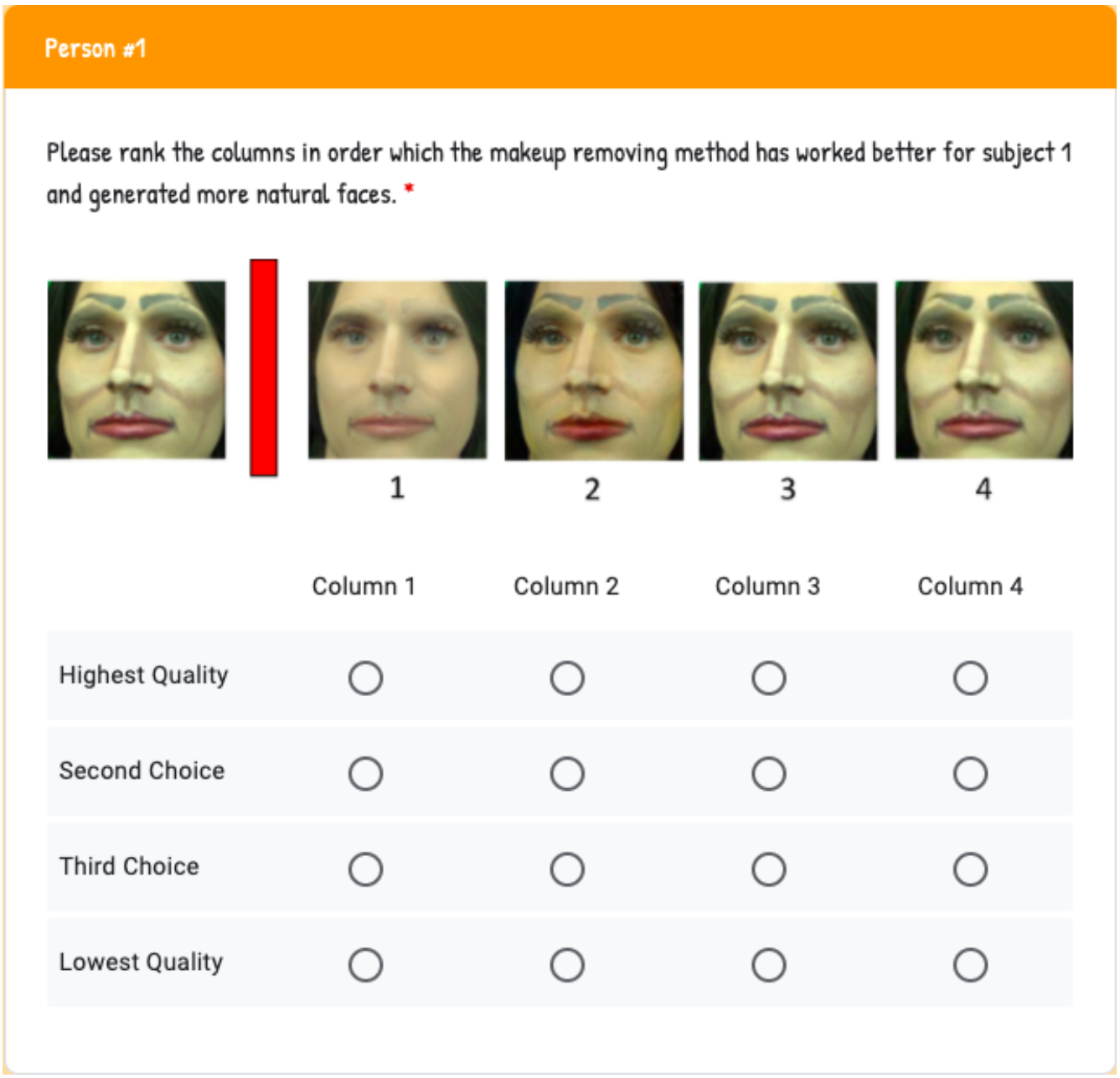


Figure 13: The web form for subjective study.