# Hand Hygiene Quality Assessment Using Image-to-Image Translation

Chaofan Wang(✉), Kangning Yang, Weiwei Jiang, Jing Wei, Zhanna Sarsenbayeva, Jorge Goncalves, and Vassilis Kostakos

School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
chaofanw@student.unimelb.edu.au

**Abstract.** Hand hygiene can reduce the transmission of pathogens and prevent healthcare-associated infections. Ultraviolet (UV) test is an effective tool for evaluating and visualizing hand hygiene quality during medical training. However, due to various hand shapes, sizes, and positions, systematic documentation of the UV test results to summarize frequently untreated areas and validate hand hygiene technique effectiveness is challenging. Previous studies often summarize errors within predefined hand regions, but this only provides low-resolution estimations of hand hygiene quality. Alternatively, previous studies manually translate errors to hand templates, but this lacks standardized observational practices. In this paper, we propose a novel automatic image-to-image translation framework to evaluate hand hygiene quality and document the results in a standardized manner. The framework consists of two models, including an Attention U-Net model to segment hands from the background and simultaneously classify skin surfaces covered with hand disinfectants, and a U-Net-based generator to translate the segmented hands to hand templates. Moreover, due to the lack of publicly available datasets, we conducted a lab study to collect 1218 valid UV test images containing different skin coverage with hand disinfectants. The proposed framework was then evaluated on the collected dataset through five-fold cross-validation. Experimental results show that the proposed framework can accurately assess hand hygiene quality and document UV test results in a standardized manner. The benefit of our work is that it enables systematic documentation of hand hygiene practices, which in turn enables clearer communication and comparisons.

**Keywords:** Hand hygiene · Handrub · Six-step hand hygiene technique · Healthcare-associated infections · Nosocomial infections

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16449-1_7.


L. Wang (Eds.): MICCAI 2022, LNCS 13437, pp. 64–73, 2022.
https://doi.org/10.1007/978-3-031-16449-1_7

## 1 Introduction

Healthcare-Associated Infections (HAIs) or nosocomial infections are a major patient-safety challenge in healthcare settings [22]. Appropriate hand hygiene is a simple and cost-efficient measure to avoid the transmission of pathogens and prevent HAIs [22]. However, research has found that hand hygiene quality in healthcare settings is generally unsatisfactory [17,18].

Typically, hand hygiene quality can be assessed by two methods: microbiological validation and Ultraviolet (UV) tests. Microbiological validation mainly uses samples from the fingertips (EN 1500) [14] or through the glove juice method (ASTM E-1174) before and after the World Health Organization (WHO) six-step hand hygiene technique. This approach evaluates hand hygiene quality in terms of bacteria count reduction [22]. Conversely, UV tests require subjects to use hand disinfectants mixed with fluorescent concentrates to perform the handrub technique, and then measure the skin coverage of the fluorescent hand disinfectants [4]. A strong correlation between the visual evaluation of UV tests and the degree of bacterial count reduction has been reported [6]. Compared to microbiological validation, UV tests can deliver an immediate and clearly visible result of skin coverage with hand disinfectants [19].

By assessing hand hygiene quality from UV tests, electronic hand hygiene monitoring systems could provide on-time intervention and periodic personalized hygiene education to Healthcare Workers (HCWs) to improve their hand hygiene practices [21]. The documented quality results can also be utilized to quantify hand hygiene technique effectiveness and provide corresponding improvement recommendations. However, subjects' hands come in diverse sizes and shapes, and their gestures and finger positions may differ across observations, resulting in difficulties in assessing hand hygiene quality and documenting its result through a standardized method that cannot be achieved by registration. Previous studies rely on manual annotation or traditional machine learning algorithms to analyze UV test results, which typically consider the presence, count, size, and/or location of the uncovered areas from the observations during the UV tests or the collected UV test images. Traditional machine learning algorithms can also summarize error distribution in terms of predefined hand regions. However, they lack the ability to further locate errors inside the hand regions or provide detailed morphology information [10,12,15]. Such information is crucial to enable consistent feedback and comparisons of hand hygiene quality. While manual annotation has been used to documented the size and location of uncovered areas on a normalized hand template, it is restricted to small sample sizes and lacks of standardized observational practices [4,5]. Thus, manual annotation can only be used for coarse-grained estimations of hand hygiene quality.

In this paper, we propose a novel deep learning-based framework to overcome these issues and evaluate hand hygiene quality on a large scale and in a standardized manner. Our contributions are twofold: 1. a method for segmenting hands from the background and classifying the hand areas covered with fluorescent hand disinfectants; 2. an approach for translating segmented hands into normalized hand templates to provide standardized high-resolution visualizations of hand hygiene quality.

## 2 Methods

Our framework includes two sub-models: 1. an Attention U-Net model localizes and segments hands from UV test images and identifies areas covered with fluorescent hand disinfectants (Fig. 1); and 2. a U-Net-based generator subsequently convert the segmented hands into normalized hand templates (Fig. 2).

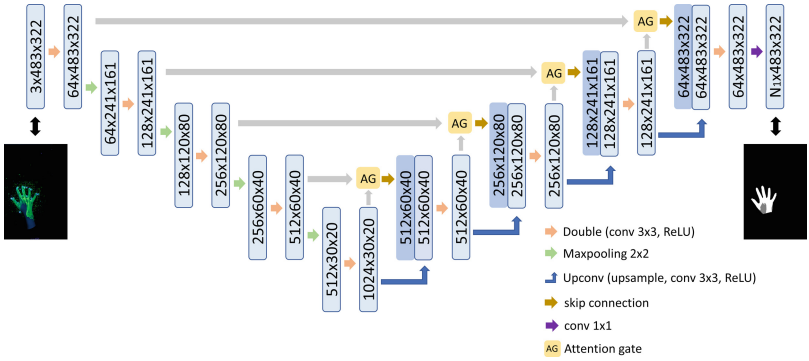


**Fig. 1.** Attention U-Net architecture. Input UV test images are progressively downsampled and upsampled by successive contracting path (left side) and expansive path (right side) to output images after hand segmentation and area classification. $N_1$ represents two classes, namely hand areas covered with fluorescent hand disinfectants (white) and uncovered hand areas (gray), and these two class images are then combined for better visualization. Attention gates highlight salient image regions and provide complimentary details to the upsampling network.

### 2.1 Hand Segmentation and Area Classification

Taking UV test images as input, we first apply the cascaded fully convolutional neural networks (i.e., U-Net [13]) with an attention gate (AG) mechanism [9] to capture spatial features and select informative feature responses. As shown in Fig. 1, this model contains three parts. Firstly, it has a contraction module consisting of alternating layers of convolution and pooling operators, which is used to capture local contextual information (like shapes or edges) progressively via multi-layers receptive fields and extract fine-grained feature maps. Secondly, it has a symmetrical expansion module where pooling operators are replaced by up-convolution operators, which is used to reconstruct and refine the corresponding hand segmentation and area classification images through successively propagating the learned correlations and dependencies to higher resolution layers. Moreover, it has multiple AGs connected to the correspondingly contracting and expansive paths to provide attention weights over the extracted different scales of feature maps in order to amplify feature responses in focus regions and simultaneously suppress feature responses in irrelevant background regions.

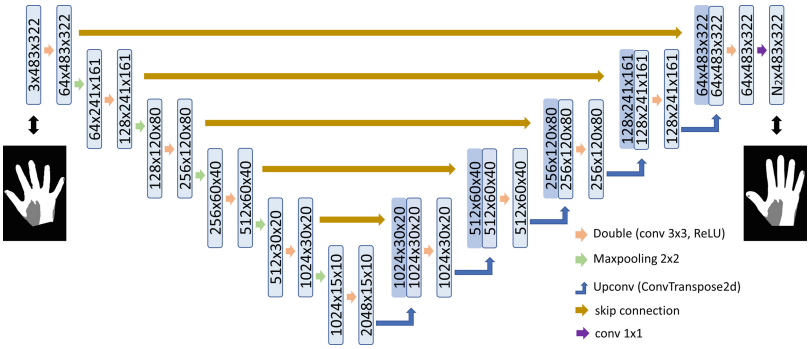


**Fig. 2.** U-Net-based generator architecture. Enlarged segmented hand images are translated into normalized hand templates via a fully convolutional network. $N_2$ represents two classes, namely hand areas covered with fluorescent hand disinfectants (white) and uncovered hand areas (gray), and these two class images are then combined for better visualization.

In this way, the finer details provided by each AG can supplement the corresponding upsampled coarse output through the skip layer fusion [7]. As defined by [9], for feature map $x^l$ from a contracting convolutional layer $l$, we first computed the attention coefficients $\alpha^l$ by:

$$\alpha_i^l = \sigma_2(\psi^{\mathrm{T}}(\sigma_1(W_x^{\mathrm{T}} x_i^l + W_g^{\mathrm{T}} g_i + b_g)) + b_\psi) \tag{1}$$

where $x_i^l$ is the vector of each pixel $i$ in $x^l$, $g_i$ is a gating vector that is collected from the upsampled coarser scale and used for each pixel $i$ to determine focus regions, $\sigma_1$ refers to the rectified linear unit (ReLU) activation function, $\sigma_2$ refers to the sigmoid activation function that is used to normalize the attention distribution, and $W_x$, $W_g$, $\psi$, $b_g$, and $b_\psi$ are the trainable parameters.

### 2.2 Translation Between Segmented Hands and Hand Templates

To further convert the enlarged segmented hands into normalized hand templates that would be easy to compare, we built a U-Net-based generator. The main idea is to take advantage of the translation equivariance of convolution operation. Given a segmented hand image $X_s \in \mathbb{R}^s$, we seek to construct a mapping function $\phi : \mathbb{R}^s \rightarrow \mathbb{R}^t$ via a fully convolutional network to translate it into a normalized hand template. Unlike the Attention-based U-Net used for localization and segmentation, we do not introduce AGs into the generator architecture, since the self-attention gating is at the global scale, which lacks some of the inductive biases inherent to convolutional networks such as translation equivariance and locality [3].

### 2.3 Loss Function

Cross-entropy loss function is the most commonly used for the task of image segmentation. However, experiments have shown that it does not perform well in the presence of the imbalance problem (*i.e.,* segmenting a small foreground from a large context/background) [16]. As suggested by Chen *et al.* [2], in this study, we combine the cross-entropy loss and dice loss for leveraging the flexibility of dice loss of class imbalance and at the same time using cross-entropy for curve smoothing. We trained both sub-models using this joint loss function:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\beta(t_i \log p_i) + (1-\beta)[(1-t_i)\log(1-p_i)] - \frac{2\sum_{i=1}^{N} p_i t_i + \epsilon}{\sum_{i=1}^{N}(p_i + t_i) + \epsilon} \quad (2)$$

where $p_i$ and $t_i$ stand for pairs of corresponding pixel values of prediction and ground truth [8], $N$ is *the number of samples* $\times$ *the number of classes*, $\epsilon$ is the added constant to avoid the undefined scenarios such as when the denominator is zero, $\beta$ controls the penalization of FPs and FNs.

## 3 Experiments

### 3.1 Dataset Construction

To the best of our knowledge, there is no publicly available UV test-related dataset. Thus, we conducted a lab study with four tasks categories (*e.g.,* Shapes, Equally Split, Individual WHO Handrub Steps, Entire WHO Handrub Technique) to collect images with different skin coverage with fluorescent hand disinfectants (details in Appendix Fig. 2). The study protocol was reviewed and approved by the University of Melbourne's Human Ethics Advisory Group. From the lab study, we collected 609 valid UV test images for both hands and both sides (1218 images when separating left and right hands) from twenty-nine participants to evaluate the effectiveness of the proposed models.

For each of the 1218 images, we first labeled the ground truth for the hand segmentation and area classification image. Labeling the ground truth for hand segmentation consists of two steps: hand contour detection and wrist points recognition from the image taken under white light (Appendix Fig. 1b). Hand contour was recognized by the red-difference chroma component from the YUV system and Otsu's method for automatic image thresholding [11], while two authors manually marked wrist points. We acquired the ground truth for hand segmentation by cropping hand contour with wrist points. Then, we labeled hand areas covered with fluorescent hand disinfectants from the image taken under UV light (Appendix Fig. 1c). Since the fluorescent concentrate used in the experiment glows green under a UV lamp, we transferred these images to the Hue Saturation Value (HSV) color system and used the H channel to detect areas within the green color range with a threshold.

Regrading the ground truth for hand translation, manual translation between segmented hands and hand templates is impractical due to a lack of a standardized translation process. Instead, we decided to generate synthetic translation data by sampling triangles and trapezoids on the same relative positions in segmented hand and hand template pair to train the hand translation model, thereby obtaining the mapping information. We first split segmented hands and hand templates into 41 triangles based on the landmarks generated by MediaPipe (and manual labels for the standard hand templates) and finger-web points (convexity defects of hand contours) calculated by OpenCV [1,20,23]. Then for each of the 41 triangle pairs, we randomly sampled triangles or trapezoids within the triangle on the segmented hands and translated them into the corresponding positions within the triangle on the hand template through homography (shown in Fig. 4, and more examples can be found in our dataset repository) [1]. Furthermore, we resized the segmented hands to cover the image to remove irrelevant background and facilitate the training process.

### 3.2 Implementation Details

We implemented both the hand segmentation and area classification model and the hand translation model in Pytorch with a single Nvidia GeForce RTX 3090 (24GB RAM). The hand segmentation and area classification model was trained for 30 epochs, while 40 epochs were used for the translation model, and both models were trained with a batch size of 16. We resized the input images for both models to $3 \times 483 \times 322$ pixels (16% of the original image). We used RMSprop optimization with an initial learning rate of $10^{-5}$, a weight decay of $10^{-8}$, and a momentum of 0.9. On this basis, we applied the learning rate schedule: if the Dice score on the validation set not increased for 2 epochs, the learning rate would be decayed by a factor of 0.1. For data augmentation, we flipped images horizontally to increase the size of the dataset for the hand segmentation and area classification model, and we also performed rotation ($\pm 20°$) and resize ($\pm 5\%$) towards the dataset for hand translation model to increase its generalizability. Code and example dataset repository is available at https://github.com/chaofanqw/HandTranslation.

### 3.3 Evaluation Metrics

To evaluate the performance of both models, we conducted five-fold cross-validation. For each round, we retained six participants' data for testing, and the remaining participants' data were then shuffled and partitioned into the training and validation sets with a 90% and 10% breakdown respectively. The trained model with the highest Dice coefficient on the validation set was then evaluated on the test set, and the Dice coefficient and Intersection over Union (IOU) score across all five-folds were then averaged and reported.

Furthermore, to evaluate the performance for the hand segmentation and area classification model, we further compared its results with two other state-of-the-art segmentation models, namely U-Net and U-Net++, through the same

five-fold cross-validation. Also, to visualize the real-life performance of the hand translation model, we employed the model to translate the segmented hands with different skin coverage (collected from the lab study) to hand templates.

## 4 Results

For the hand segmentation and area classification model, the Attention U-Net achieved the highest mean Dice coefficient (96.90%) and IOU score (94.02%). Meanwhile, U-Net++ achieved a comparable performance to Attention U-Net (Dice coefficient: 96.87%, IOU score: 93.95%), and both models outperformed U-Net (Dice coefficient: 96.64%, IOU score: 93.54%). Figure 3 provides the qualitative results of the hand segmentation and area classification over different hands, sides, and task categories.

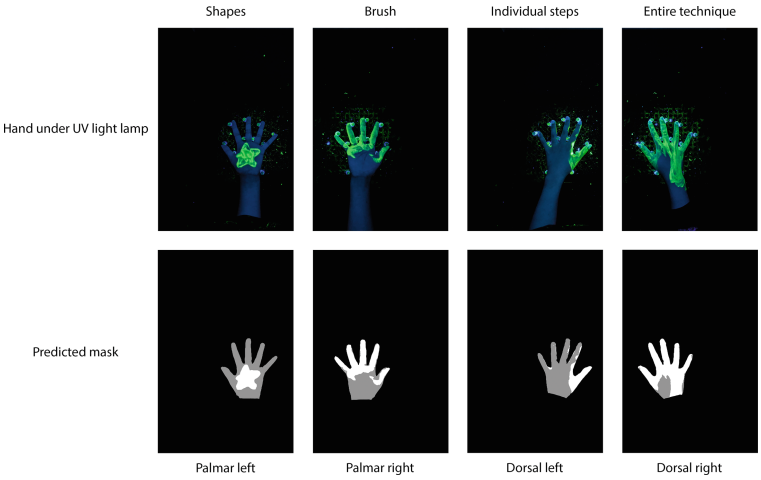


**Fig. 3.** Qualitative evaluation results over different hands, sides, and task categories. The upper row shows the original color images taken under UV light. The bottom row exhibits the predicted hand segmentation and area classification results, where black indicates background, gray indicates uncovered hand areas, and white indicates hand areas covered with hand disinfectants.

We further evaluated the Attention U-Net performance for each participant and task category. For the participant-wise model performance, the highest was seen by P30 (Dice coefficient: 97.96%, IOU score: 96.02%), while the lowest was seen by P2 (Dice coefficient: 95.68%, IOU score 91.75%). For the task-wise model performance, the highest was of the "Shapes" task (Dice coefficient: 97.07%, IOU score: 94.33%), while the lowest was of the "Individual WHO Handrub Steps" task (Dice coefficient: 96.89%, IOU score: 93.98%).

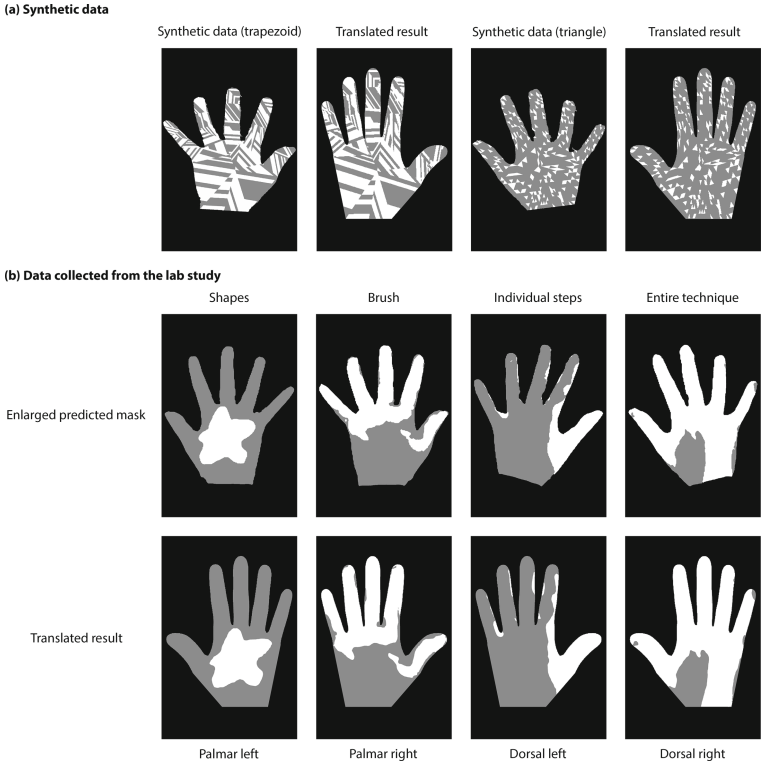


**Fig. 4.** Qualitative evaluation results over synthetic data and the lab study data. (a) The synthetic images with trapezoids or triangles within the segmented hands and the corresponding hand translation results. (b) The upper row shows the enlarged hand segmentation and area classification results from the lab study, and the bottom row exhibits the corresponding hand translation results.

For the hand translation model, the proposed system achieved the Dice coefficient of 93.01% and the IOU score of 87.34% on the synthetic dataset. Figure 4a provides qualitative results of the hand translation for the synthetic data. We then evaluated the trained model on the lab study data of segmented hands with different skin coverage. However, the model with the best performance on the synthetic dataset tends to overfit the shape of trapezoid and triangle and generates rough contours for the areas covered with hand disinfectants. Thus, to avoid overfitting, we chose to use the model trained with eight epochs for translating the lab study data to hand templates based on visual inspections (Fig. 4b).

## 5 Conclusion

We present an image-to-image translation framework to evaluate hand hygiene quality through UV test images. The proposed framework adopts an attention U-Net to segment hands from the background and classifies the areas covered with hand disinfectants and a U-Net-based generator to translate segmented hands into normalized hand templates. Trained on the presented dataset, experimental results show that the proposed framework can accurately evaluate hand hygiene quality and document UV tests results in a standardized manner. The documented quality results can be then used to summarize the frequently untreated areas caused by standardized hand hygiene techniques or HCWs' respective techniques and evaluate their effectiveness [4,20]. Due to the nature of aforementioned application scenarios, translation model errors can be mitigated after summarizing the data over a large study population. Moreover, since the translation model tends to overfit sampling patterns of triangles and trapezoids of the generated synthetic dataset, future studies can aim to investigate other synthetic data generation methods to further improve the hand translation model performance.

**Acknowledgement.** This work is partially funded by NHMRC grants 1170937 and 2004316. Chaofan Wang is supported by a PhD scholarship provided by the Australian Commonwealth Government Research Training Program.

## References

1. Bradski, G.: The OpenCV library. Dr Dobbs J. Softw. Tools (2000). https://doi.org/10.1111/0023-8333.50.s1.10
2. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_50
3. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Kampf, G., Reichel, M., Feil, Y., Eggerstedt, S., Kaulfers, P.M.: Influence of rubin technique on required application time and hand coverage in hygienic hand disinfection. BMC Infect. Dis. (2008). https://doi.org/10.1186/1471-2334-8-149
5. Kampf, G., Ruselack, S., Eggerstedt, S., Nowak, N., Bashir, M.: Less and less-influence of volume on hand coverage and bactericidal efficacy in hand disinfection. BMC Infect. Dis. (2013). https://doi.org/10.1186/1471-2334-13-472
6. Lehotsky, A., Szilagyi, L., Bansaghi, S., Szeremy, P., Weber, G., Haidegger, T.: Towards objective hand hygiene technique assessment: validation of the ultraviolet-dye-based hand-rubbing quality assessment procedure. J. Hospital Infection **97**(1), 26–29 (2017). https://doi.org/10.1016/j.jhin.2017.05.022, https://linkinghub.elsevier.com/retrieve/pii/S0195670117302943
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

8. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
9. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
10. Öncü, E., Vayısoğlu, S.K.: Duration or technique to improve the effectiveness of children' hand hygiene: a randomized controlled trial. Am. J. Infect. Control (2021). https://doi.org/10.1016/j.ajic.2021.03.012
11. Otsu, N.: Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Syst. Man Cybern. (1979). https://doi.org/10.1109/tsmc.1979.4310076
12. Rittenschober-Böhm, J., et al.: The association between shift patterns and the quality of hand antisepsis in a neonatal intensive care unit: an observational study. Int. J. Nurs. Stud. (2020). https://doi.org/10.1016/j.ijnurstu.2020.103686
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Standardization, E.C.: European Standard, EN1500:2013, CHEMICAL DISINFECTANTS AND ANTISEPTICS. HYGIENIC HANDRUB. TEST METHOD AND REQUIREMENTS (PHASE 2/STEP 2). European Committee for Standardization (2013)
15. Szilágyi, L., Lehotsky, A., Nagy, M., Haidegger, T., Benyó, B., Benyó, Z.: Steryhand: A new device to support hand disinfection. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2010 (2010). https://doi.org/10.1109/IEMBS.2010.5626377
16. Taghanaki, S.A., et al.: Combo loss: handling input and output imbalance in multi-organ segmentation. Comput. Med. Imaging Graph. **75**, 24–33 (2019)
17. Taylor, L.J.: An evaluation of handwashing techniques-1. Nursing times (1978)
18. Taylor, L.J.: An evaluation of handwashing techniques-2. Nursing times (1978)
19. Vanyolos, E., et al.: Usage of ultraviolet test method for monitoring the efficacy of surgical hand rub technique among medical students. J. Surg. Educ. (2015). https://doi.org/10.1016/j.jsurg.2014.12.002
20. Wang, C., et al.: A system for computational assessment of hand hygiene techniques. J. Med. Syst. **46**(6), 36 (2022). https://doi.org/10.1007/s10916-022-01817-z
21. Wang, C., et al.: Electronic monitoring systems for hand hygiene: systematic review of technology (2021). https://doi.org/10.2196/27880
22. World Health Organization (WHO): WHO guidelines on hand hygiene in health care (2009)
23. Zhang, F., et al.: MediaPipe hands: on-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)