ORIGINAL ARTICLE

DisguisOR: holistic face anonymization for the operating room

Lennart Bastian[1] · **Tony Danjun Wang**[1] · **Tobias Czempiel**[1] · **Benjamin Busam**[1] · **Nassir Navab**[1]

Abstract

Purpose Recent advances in Surgical Data Science (SDS) have contributed to an increase in video recordings from hospital environments. While methods such as surgical workflow recognition show potential in increasing the quality of patient care, the quantity of video data has surpassed the scale at which images can be manually anonymized. Existing automated 2D anonymization methods under-perform in Operating Rooms (OR), due to occlusions and obstructions. We propose to anonymize multi-view OR recordings using 3D data from multiple camera streams.

Methods RGB and depth images from multiple cameras are fused into a 3D point cloud representation of the scene. We then detect each individual's face in 3D by regressing a parametric human mesh model onto detected 3D human keypoints and aligning the face mesh with the fused 3D point cloud. The mesh model is rendered into every acquired camera view, replacing each individual's face.

Results Our method shows promise in locating faces at a higher rate than existing approaches. DisguisOR produces geometrically consistent anonymizations for each camera view, enabling more realistic anonymization that is less detrimental to downstream tasks.

Conclusion Frequent obstructions and crowding in operating rooms leaves significant room for improvement for off-the-shelf anonymization methods. DisguisOR addresses privacy on a scene level and has the potential to facilitate further research in SDS.

Keywords Anonymization · Face detection · Multi-view operating rooms · Surgical data science · Surgical workflow recognition

Introduction

The past years have seen an increase in video acquisitions in hospitals and surgical environments. In the field of surgical data science (SDS), the analysis of endoscopic and laparoscopic frames is already an established research direction [1]. It aims to build cognitive systems capable of understanding the procedural steps of an intervention, for example, recognizing and localizing surgical tools [2]. Closely related to the endoscopic frames are videos from externally mounted cameras, capturing the surgical scene from an outside perspective [3]. These rich information sources build the foundation for

analyzing and optimizing the workflow, essential for developing context-aware intelligent systems, improving patient quality of care, and advancing anomaly detection. However, video recordings of surgeries are still considered problematic due to strict privacy regulations established to protect both patients and medical staff. As manually anonymizing video frames no longer become feasible at scale, it is imperative to develop automatic de-identification methods to advance future research and facilitate SDS dataset curation.

Surgical operating rooms are frequently crowded and packed with medical equipment. Cameras can only be mounted at particular positions, leading to perspectives not usually found in conventional datasets [4]. This poses challenges even for advanced anonymization methods, as they tend to perform poorly under partial occlusions and obscure camera angles [5]. A few methods address the specific challenges of OR anonymization [5, 6] from individual cameras. Recent works propose addressing the OR's unique challenges by combining multi-view RGB-D data (*multi-view*)

Lennart Bastian and Tony Danjun Wang have been contributed equally to this study.

✉ Lennart Bastian
Lennart.Bastian@tum.de

1 Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany

to compensate for missed information in surgical workflow recognition [3, 7–9]. The existence of such *multi-view* OR recordings requires anonymizing all camera views, as a failed anonymization in a single view breaches the privacy of the entire scene.

We propose a novel anonymization approach for *multi-view* recordings, which leverages 3D information to detect faces where conventional methods fail. We utilize a 3D mesh to accurately replace each detected person's face, preserving privacy as well as data integrity in all camera views. In Fig. 1, we compare single and *multi-view* approaches, highlighting the advantages of scene-level anonymization. We additionally show that in comparison with existing methods, our face replacement yields images that harmonize well with the surgical environment as measured by image similarity. Our main contributions can be summarized as follows:

- We present a novel framework for accurate *multi-view* 2D face localization by leveraging 3D information. We further emphasize the necessity for consistent anonymization across all camera views using our proposed *holistic recall*.
- We present a training-free, mesh-based anonymization method yielding complete control during the 3D face replacement step while generating more realistic results than existing state-of-the-art approaches.
- The images anonymized by our framework can be effectively utilized by downstream methods, as shown through experiments on image quality assessment and downstream face localization.

Related works

Face detection

With the advent of public face detection benchmark datasets like WIDERFACE [4], numerous deep learning-based face detectors were introduced in recent years [11–13]. Such methods typically regress a bounding box onto the region where a face can be successfully identified in the image. As WIDERFACE consists of annotated images from everyday scenarios, face detectors trained on this dataset can suffer in complex and crowded OR environments [5]. Occlusions and obstructions from medical equipment or personnel in close quarters, masks, and skull caps can lead to missed predictions and, ultimately, incomplete anonymizations. While we also use 3D data for anonymization, our work diverges from 3D face recognition [14], where a scan of a 3D face is matched to a catalogue of face scans.

Image anonymization

Identity scrubbing can be achieved by removing the sensitive area, blurring, or pixelization [3]. In the OR, standardized scrubs and gloves already obscure many possible landmarks, leaving the face as the primary identifier that could be used for re-identification, as previously established [5, 6]. A recent line of work has proposed to replace faces with artificially generated faces using GANs [10, 15] or parametric face models [16]. Such replacement methods tend to yield a more realistic-looking output, and the resulting anonymized area resembles the input more closely, which can positively affect downstream applications [15]. However, these methods typically contain a separate branch to handle face detection [10] and thus suffer similarly in OR environments due to partial obstructions.

Human pose estimation

Using human pose estimation as an additional context to localize faces has been demonstrated as valuable [5]. The torso, shoulders, and arms provide useful cues for localizing faces occluded under a surgical mask and skull cap. Beyond mere 2D human keypoint detection, a significant emphasis has also been placed on regressing keypoints from multiple camera views in a shared 3D space [17, 18]. 3D human pose detection can be especially beneficial for multi-person scenarios such as surgical ORs, where ubiquitous occlusions can lead to poor performance in individual camera views [19, 20]. Regressing a 3D human shape from a single input image is also an active area of research [21]. However, such methods would suffer similarly to partial occlusions. To avoid this shortcoming, we leverage the 3D nature of *multi-view* OR acquisitions.

Methods

An overview of our proposed method DisguisOR is shown in Fig. 2. We use the *multi-view* OR dataset introduced in Bastian et al. [7] depicting veterinary laparoscopic procedures, expanded to all four cameras available in the acquisitions. Each camera's color and depth images are combined into a colored 3D point cloud using the Azure Kinect framework. Subsequently, the four partial point clouds are registered into one global coordinate space by minimizing the photometric reprojection error over keypoints on a large visual marker. Our pipeline thus uses RGB images and depth maps from each camera, along with a fused point cloud of the entire scene as input.

Fig. 1 Holistic Face Anonymization From Scene Level Representation. Four views of a *multi-view* OR acquisition are visualized on the left, highlighting person 1 in each view. Conventional methods detect and anonymize faces in each image individually (green boxes) through either blackening (BLN), pixelization (PXL), blurring (BLR), or face replacement (FRP) [3, 6, 10]. Our framework (purple boxes) leverages *multi-view* information to localize faces in 3D, enabling consistent anonymization in all four images
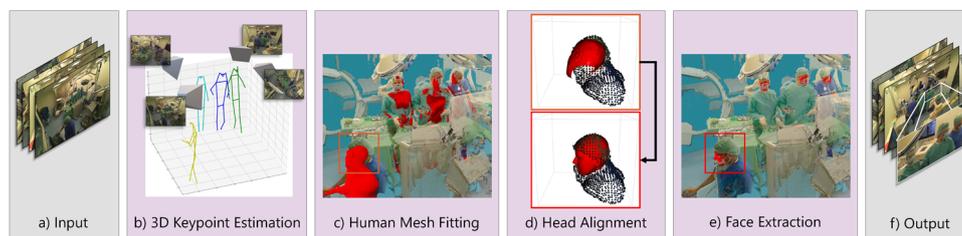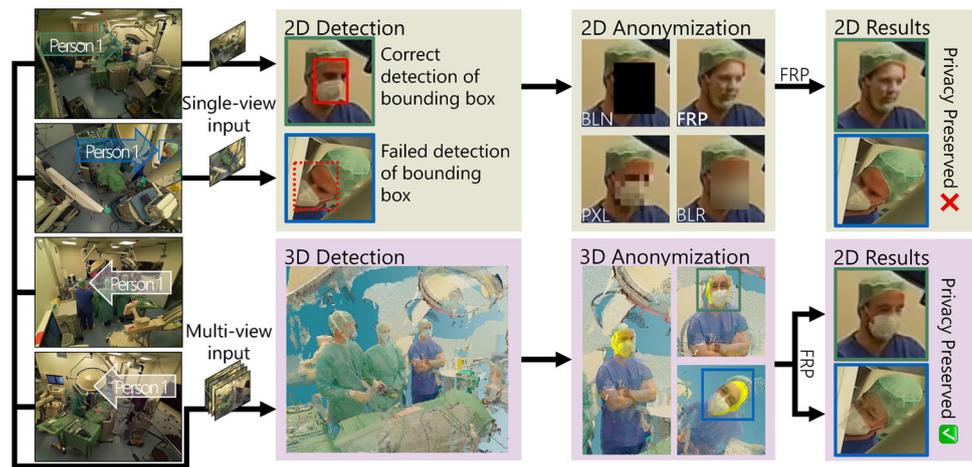
Fig. 2 Anonymization Pipeline of DisguisOR. Four RGB images **a** and a fused 3D point cloud as input. Human pose keypoints are first detected in individual views and fused in 3D for each person in the scene **b**. Fitting of parametric human mesh model onto keypoints of each person in the scene **c**. Further geometrical refinement of the mesh positioning **d** and face extraction **e**. 3D mesh texturization and back-projection to each camera view **f**

Multi-person 3D Mesh Regression

We adopt an unsupervised 3-stage approach to fit a 3D mesh [22] for each person in the scene. 2D human keypoints are first detected [23] in each camera view, and regressed in a global coordinate frame with VoxelPose [18]. As neither 2D nor 3D human poses are available as ground truth, we use an existing detector [23] trained on COCO to estimate 2D human keypoints from an image. In order to combine poses from each view in a robust manner, Voxelpose must first be trained to learn how multiple 2D poses from each camera can be optimally combined in 3D. To achieve this, we follow the procedure described in [18] and synthetically generate ground truth by sampling existing 3D human poses from the Panoptic dataset [24] and placing them at random locations in the 3D space. These poses are then projected back into each 2D image-plane and used as input to guide Voxel-Pose through the 2D-to-3D multi-person regression task. The trained model ultimately combines 2D human poses from multiple views into one joint 3D human pose for each person in the scene. We then perform an additional temporal smoothing on each 3D human pose sequence to interpolate missing poses and reduce noise (for details, see suppl.).

3D Human Representation

In order to adequately represent the face of each individual in the scene, we propose to use the statistical parametric human mesh model SMPL [22], which we regress onto each 3D human pose obtained as output from Voxelpose. While temporal smoothing yielded less noisy keypoint estimates, we noticed that the 3D mesh model did not always align with the 3D point cloud of an individual, resulting in inaccurate face localizations. To resolve this issue, we perform a rigid registration between the head segment of the SMPL model and the point cloud. More specifically, we crop the point cloud around the estimated head of the SMPL model and align the model with the point cloud using the probabilistic point-set registration method FilterReg [25], and subsequently fine-tune using iterative closest point (ICP) [26]. As a final step, we extract the face from the SMPL mesh, which should now be aligned with the 3D location of an individual's face.

Rendering the faces in 2D

Thus far, our pipeline estimates a global 3D face mesh that overlaps with each person in the scene. In order to yield a 2D anonymization, these meshes can now be projected back into

all camera views, replacing the face of each individual with a unique template (see Fig. 5). However, a 3D face might be occluded in a particular view, for example, due to an OR light and, therefore, not visible. To mitigate false-positive predictions, we check whether a 3D face is visible in 2D by looking for a disparity between the camera's depth map and the 3D face mesh (for details, see suppl.). We then utilize the Poisson Image Editing technique [27] to harmonize the face template and the background image for a more natural appearing face replacement. The template can also be changed for each individual to influence factors such as age, sex, or ethnicity.

Ground truth curation

We identified three distinct scenarios of varying difficulties from the complete dataset [7]. We then annotated each visible face manually in all camera views for a total of 4913 face bounding boxes. The annotation criteria were adopted to closely match the style of the WIDERFACE dataset [4]. The three scenarios are chosen to specifically represent the varying characteristics in the OR. They differ in the number of individuals present, their attire, and the degree of obstructions (Fig. 3).

- *Easy Evaluation Scenario* Up to four people in the scene, all wearing surgical masks and hospital scrubs with only a few face obstructions. A total of 1310 faces.
- *Medium Evaluation Scenario* Five or six people in the scene with regular face obstructions caused by the position of the surgical lights. A total of 2317 faces.
- *Hard Evaluation Scenario* Four people are present in the room. Clinicians additionally wear skull caps and gowns. The surgical lights frequently obstruct the faces in two of the views. A total of 1286 faces.

Experiments

Face localization

We compare the proposed method's face localization performance with that of DSFD [11], a state-of-the-art detector also used in DeepPrivacy [10]. We use the model pre-trained on WIDERFACE [4] provided by the authors. We additionally evaluate the self-supervised domain adaption (SSDA) strategy proposed by Issenhuth et al. [5]. Here we also use DSFD as the face detection backbone, fine-tuning it on 20k unlabeled images as proposed, with the suggested hyperparameters.

In addition to recall, we propose to evaluate *multi-view* OR anonymization with what we coin *holistic recall*. The *holistic recall* considers a face as detected only if it was identified in all camera views where it is at least partially visible. We argue that this is more suitable than image-wise evaluation, as a missed detection of a face in a single view results in a breach of anonymization for that individual.

We calculate the smallest rectangle outlining the rendered mesh to generate face predictions for evaluation. As the proposed method does not rank the output detections with a confidence score, the commonly used average precision (AP) score is not defined. Therefore, we additionally report precision and F1-score for all three methods in the supplementary materials. Furthermore, the four cameras are categorized as either a *surgical camera* (SC) or *workflow camera* (WFC), depending on the perspective of the camera (see Fig. 3). The images and angle of acquisition in WFCs are more similar to what might be found in public face detection datasets [4], while SCs may acquire the scene from above, and individuals are more frequently obscured by OR equipment.

Image quality

We compare the images anonymized by our approach to those altered by several conventional anonymization methods, such as blurring (61 × 61 kernel), pixelization (8 × 8 pixels), blackening, as well as the established GAN-based model DeepPrivacy [10] (see 2D anonymization Fig. 1). To disentangle image quality and face detection, we only evaluate image quality on faces detected by both our method and Deep Privacy, totaling 3786 faces. We evaluate the effectiveness of our face replacements on the cropped ground-truth bounding boxes with three established image quality metrics. The fréchet inception distance (FID) [28] measures the overall realism by calculating the distribution distance of the original and generated set of images. Learned perceptual image patch similarity (LPIPS) [29] reflects the human perception of an image's realism by computing the difference between activations of two image patches for a standard neural network. The structural similarity index measure (SSIM) [30] calculates the quality of an image pixel-wise based on luminance, contrast, and structure.

Finally, we conduct additional experiments on the downstream behavior of off-the-shelf methods on our anonymized faces (see suppl.)

Results

Face Detection

Figure 4 depicts the performance of our proposed method in comparison with two existing baselines. In the easy evaluation scenario, both DSFD and DisguisOR perform comparably, while the SSDA achieves a 9% higher *holistic recall*. In the medium and hard scenarios, DisguisOR outper-

Fig. 3 Overview of the Three Scenarios and All Four Camera Views. We categorize the scenes by their complexity. Surgical cameras (SC) are characterized by persistent obstructions and unusual viewing angles. Workflow cameras (WFC) exhibit ordinary viewing angles with fewer obstructions
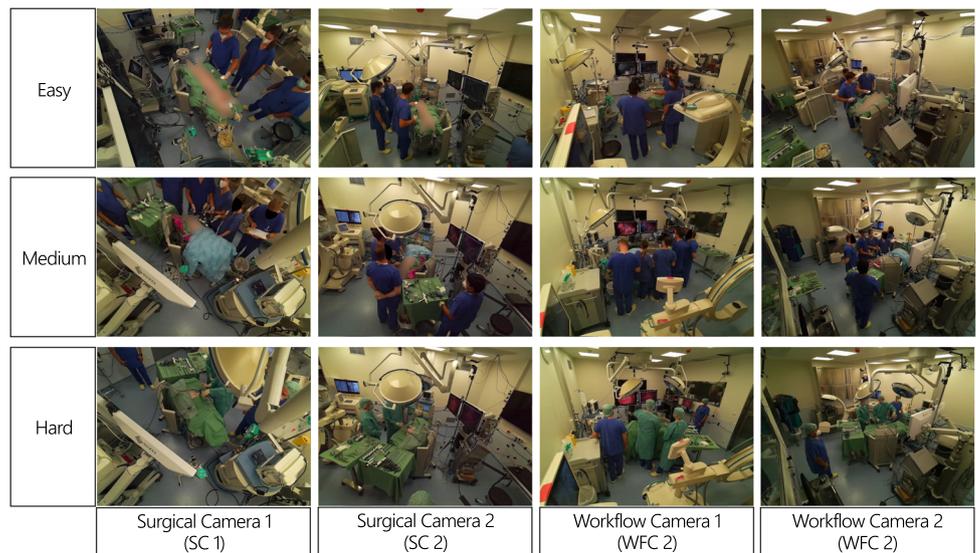
Fig. 4 Face Localization Performance of DSFD [11], the Self-Supervised Domain Adaption (SSDA) Method of [5], and DisguisOR Over All Scenarios. The *holistic recall* considers a face as detected only if the same face was successfully detected in all camera views where it is at least partially visible. Both recall and *holistic recall* are reported for IOU@0.4. Best values for each camera denoted in bold
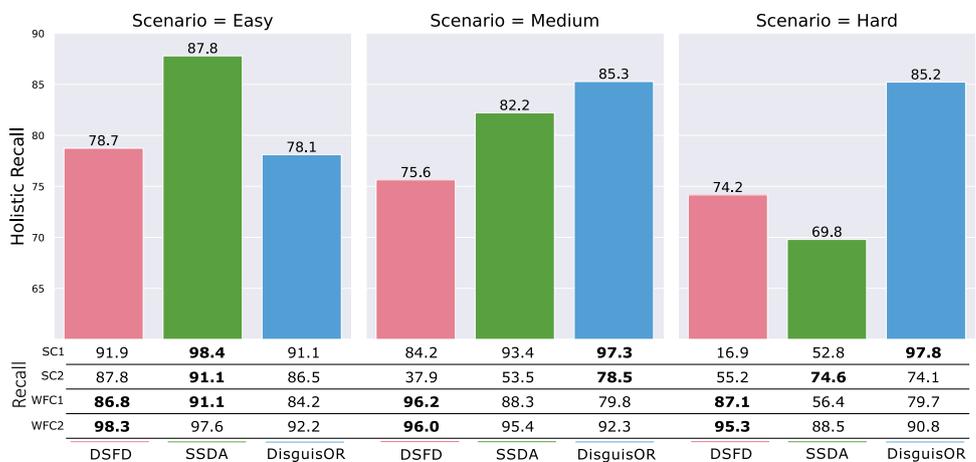
forms DSFD and SSDA in *holistic recall* by 10% and 3%, and 11% and 16%, respectively.

These disparities are largely due to a poor detection rate in the surgical cameras, which are acquired from unusual camera angles and contain frequent obstructions (Fig. 3). DisguisOR is able to better cope with the increased occlusions and the number of individuals present in these scenarios, highlighting the proposed method's robustness under partial visibility. By combining information from multiple cameras, DisguisOR yields a geometrically consistent detection—if an individual face has been accurately localized in the 3D scene, it can be more consistently identified in each individual image. While DSFD achieves a significantly higher accuracy in the easy scenario when refined via SSDA, the human pose backbone of DisguisOR underwent no such refinement and would likely also see some performance improvements.

SSDA underperforms the baseline DSFD model in the hard scenario, as well as DisguisOR in both the medium and hard scenarios. This could be because these more challenging

detection candidates are less represented in the training data distribution or not detected with a high confidence score, and thus not pseudo-labeled frequently enough.

The recall rates over individual cameras reflect the characterizations of surgical and workflow camera views. While DSFD generally achieves slightly higher recall rates on workflow camera views (WFC1, WFC2), DisguisOR achieves

Table 1 Comparison of different anonymization techniques based on image quality metrics. An arrow depicts whether a smaller (down) or larger (up) value is more favorable with respect to each metric

Method	FID ↓ [28]	LPIPS ↓ [29]	SSIM ↑ [30]
Blackening	194.03	0.5392	0.1864
Pixel	173.34	0.4037	0.6080
Blur	164.57	0.3688	0.6014
DeepPrivacy [10]	94.85	0.2276	0.6294
DisguisOR	**35.24**	**0.1341**	**0.8143**

Best values for each metric denoted in bold

Fig. 5 Faces From Our Dataset Anonymized With Different Methods. Two unaltered faces **a**, **d**, faces anonymized with DeepPrivacy [10] **b**, **e** and anonymized with DisguisOR **c**, **f**. Note that DeepPrivacy fails to incorporate face masks in its generated faces. Rendering faces with a texture that reflects the setting (in this case a texture with a mask) can yield more consistent face replacement results

much higher recall rates in the surgical camera views (SC1, SC2), see Fig. 3. SSDA improves recall for DSFD in surgical cameras, although it still falls short of DisguisOR in medium and hard scenarios. The surgical cameras in the hard scenario are especially challenging for face detectors, as severe occlusions, unusual camera angles, and surgical scrubs drastically impair the face detectors' detection rate. In the case of faces in SC1 of the hard scenario (see person 1 in Fig. 1), DSFD achieves a recall rate of 16.9%. Using SSDA increases this recall rate to 52.8%, which DisguisOR still outperforms with a recall rate of 97.8%.

Our method is somewhat limited by the field-of-view (FOV) of the depth sensors during an acquisition. This partially explains the comparable performance with DSFD in the easy scenario, as individuals frequently move along the edge of the scene where depth coverage is limited. The 3D reconstruction we use to triangulate faces could also be performed without the use of the slightly more costly depth sensors, albeit less accurately.

Image quality

In Table 1, we measure the quality of images altered by baseline approaches and our proposed method. As expected, conventional obfuscations like blackening, pixelization, and blurring achieve inferior results across all three metrics. DeepPrivacy [10] is designed to generate synthetic faces instead of applying conventional privacy filters, explaining the improved results on all image quality metrics compared to the conventional methods. Our method further improves upon these results as the replacement of the face information can be precisely controlled, even enabling the replacement of people wearing masks without creating corrupted or unnatural faces. In Fig. 5, we illustrate examples where DeepPrivacy replaces the face mask of a person with an unnatural mouth (e), while our method manages to blend the template and original image (f) more effectively.

Conclusion

Existing anonymization methods do not effectively leverage *multi-view* data as they consider individual views indepen-

dently. OR cameras are frequently mounted in unconventional positions and therefore suffer from heavy occlusions, making multiple views essential for accurately acquiring details of a procedure. Our 3D face detection framework DisguisOR enables consistency over each camera, preventing missed detections in a single view that would breach an individual's anonymity. Therefore, we emphasize the use of scene-level anonymization with our proposed *holistic recall* metric to consider the recall of faces detected jointly in all camera views. We validate our face detection approach based on recall on individual camera views as well as *holistic recall*, demonstrating that our method achieves state-of-the-art results under challenging scenarios and views.

Furthermore, anonymization methods must balance the discrepancy between anonymizing data and retaining its downstream utility. We show that our framework reduces this discrepancy by yielding more realistic face replacements compared to existing methods. The modularity of our anonymization approach provides us with fine-grained control of the face replacement, allowing us to vary parameters such as age, gender, or ethnicity. Existing datasets could even be augmented with faces representing a broad demographic, combating bias induced by unrepresentative training sets. We are convinced that our method will facilitate further research by reducing the burden of manually annotating existing and future *multi-view* data acquisitions.

Declarations

References

1. Czempiel T, Paschali M, Keicher M, Simson W, Feussner H, Kim ST, Navab N (2020) TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In: MICCAI 2020,

2. Garrow CR et al (2021) Machine learning for surgical phase recognition: a systematic review. Ann Surg 273(4):684–693

3. Srivastav V, Issenhuth T, Kadkhodamohammadi A, de Mathelin M, Gangi A, Padoy,N (2018) MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. arXiv preprint arXiv:1808.08180

4. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: A face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5525-5533)

5. Issenhuth T, Srivastav V, Gangi A, Padoy N (2019) Face detection in the operating room: comparison of state-of-the-art methods and a self-supervised approach. Int. J. Comput. Assist. Radiol. Surg. 14:1049–1058

6. Flouty E, Zisimopoulos O, Stoyanov D (2018) Faceoff: Anonymizing videos in the operating rooms. CoRR **abs/1808.04440**

7. Bastian L, Czempiel T, Heiliger C, Karcz K, Eck U, Busam B, Navab N (2022) Know your sensors-a modality study for surgical action classification. Comput Methods Biomech Biomed Eng Imaging Visual. https://doi.org/10.1080/21681163.2022.2152377

8. Schmidt A, Sharghi A, Haugerud H, Oh D, Mohareri O (2021) Multi-view surgical video action detection via mixed global view attention. In: MICCAI, Springer pp. 626–635

9. Sharghi A, Haugerud H, Oh D, Mohareri O (2020) Automatic operating room surgical activity recognition for robot-assisted surgery. In: MICCAI, Springer pp. 385–395

10. Hukkelas H, Mester R, Lindseth F (2019) Deepprivacy: A generative adversarial network for face anonymization. In: Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, Proceedings, Part I 14 (pp. 565-578)

11. Li J, Wang Y, Wang C, Tai Y, Qian J, Yang J, Huang F (2019) DSFD: dual shot face detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5060-5069)

12. Zhu Y, Cai H, Zhang S, Wang C, Xiong Y (2020) Tinaface: Strong but simple baseline for face detection. arXiv preprint arXiv:2011.13183

13. Minaee S, Luo P, Lin Z, Bowyer K (2021) Going deeper into face detection: a survey. arXiv preprint arXiv:2103.14983

14. Zhou S, Xiao S (2018) 3d face recognition: a survey. HCIS 8(1):1–27

15. Cai Z, Xiong Z, Xu H, Wang P, Li W, Pan Y (2021) Generative adversarial networks: a survey toward private and secure applications. ACM Comput Surv (CSUR) 54(6):1–38

16. Sun Q, Tewari A, Xu W, Fritz M, Theobalt C, Schiele B (2018) A hybrid model for identity obfuscation by face replacement. In: Proceedings of the European conference on computer vision (ECCV) (pp. 553-569)

17. Liu W, Bao Q, Sun Y, Mei T (2022) Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective. ACM Comput Surv 55(4):1–41

18. Tu H, Wang C, Zeng W (2020) End-to-end estimation of multi-person 3d poses from multiple cameras. CoRR **abs/2004.06239**

19. Hu H, Hachiuma R, Saito H, Takatsume Y, Kajita H (2022) Multi-camera multi-person tracking and re-identification in an operating room. J Imaging 8(8):219

20. Özsoy E, Örnek EP, Eck U, Czempiel T, Tombari F, Navab N (2022) 4d-or: semantic scene graphs for or domain modeling. Springer, Berlin

21. Kolotouros N, Pavlakos G, Daniilidis K (2019) Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4501–4510

22. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Graph (TOG) 34(6):1–16

23. Geng Z, Sun K, Xiao B, Zhang Z, Wang J (2021) Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14676–14686)

24. Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y (2015) Panoptic studio: A massively multi-view system for social motion capture. In: (ICCV)

25. Gao W, Tedrake R (2019) Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11095-11104)

26. Besl PJ, McKay ND (1992) A method for registration of 3-d shapes. IEEE TPAMI 14(2):239–256

27. Pérez P, Gangnet M, Blake A (2003) Poisson image editing. ACM Trans Graph 22(3):313–318

28. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, vol 30. Annual conference on neural information processing systems. Long Beach, CA, USA, pp 6626–6637

29. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 586-595)

30. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612