

---

# Face Detection in the Operating Room: Comparison of State-of-the-art Methods and a Self-supervised Approach

Thibaut Issenhuth · Vinkle Srivastav ·  
Afshin Gangi · Nicolas Padoy

Received: date / Accepted: date

**Abstract** *Purpose:* Face detection is a needed component for the automatic analysis and assistance of human activities during surgical procedures. Efficient face detection algorithms can indeed help to detect and identify the persons present in the room, and also be used to automatically anonymize the data. However, current algorithms trained on natural images do not generalize well to the operating room (OR) images. In this work, we provide a comparison of state-of-the-art face detectors on OR data and also present an approach to train a face detector for the OR by exploiting non-annotated OR images.

*Methods:* We propose a comparison of 6 state-of-the-art face detectors on clinical data using Multi-View Operating Room Faces (MVOR-Faces), a dataset of operating room images capturing real surgical activities. We then propose to use self-supervision, a domain adaptation method, for the task of face detection in the OR. The approach makes use of non-annotated images to fine-tune a state-of-the-art detector for the OR without using any human supervision.

*Results:* The results show that the best model, namely the tiny face detector, yields an average precision of 0.556 at Intersection over Union (IoU) of 0.5. Our self-supervised model using non-annotated clinical data outperforms this result by 9.2%.

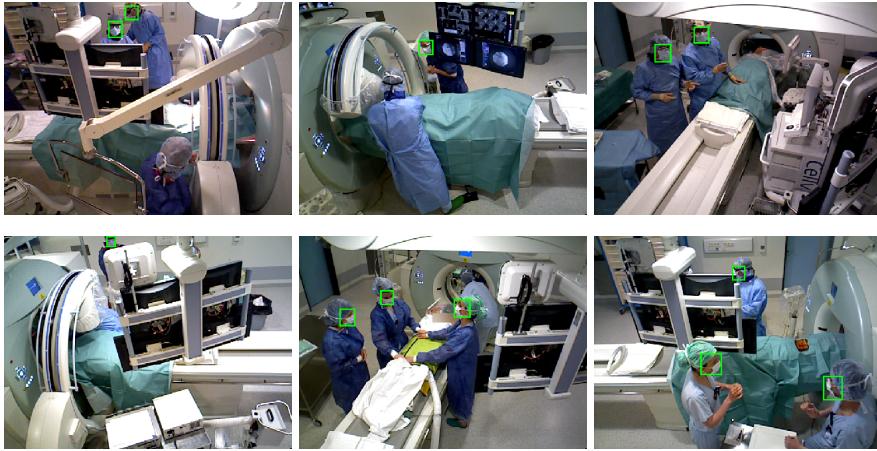
*Conclusion:* We present the first comparison of state-of-the-art face detectors on operating room images and show that results can be significantly improved by using self-supervision on non-annotated data.

**Keywords** Face Detection · Semi-supervised Learning · MVOR-Faces Dataset · Visual Domain Adaptation · Operating Room

---

Thibaut Issenhuth · Vinkle Srivastav · Nicolas Padoy  
ICube, University of Strasbourg, CNRS, IHU Strasbourg, France  
E-mail: issenhuth@unistra.fr · srivastav@unistra.fr · npadoy@unistra.fr

Afshin Gangi  
Radiology Department, University Hospital of Strasbourg, France



**Fig. 1** Examples of images collected in the OR, illustrating the challenges for face detection systems: occlusion with medical equipment or other persons, masked faces, absence of visible skin. Ground-truth is shown with green bounding boxes.

## 1 Introduction

Modern ORs are high-tech environments, where surgical activities are increasingly captured with cameras. The automatic understanding of this visual data by extracting rich and meaningful information is a promising way of developing machine intelligence and context-aware systems in the clinical environment. It will help to improve the workflow in hospitals by developing better decision support systems for the clinicians [1]. Face detection in the operating room is one of the key steps needed to develop intelligent context-aware systems for the automatic analysis of human activities. It can indeed serve for person detection and identification as well as for the anonymization of sensitive OR data.

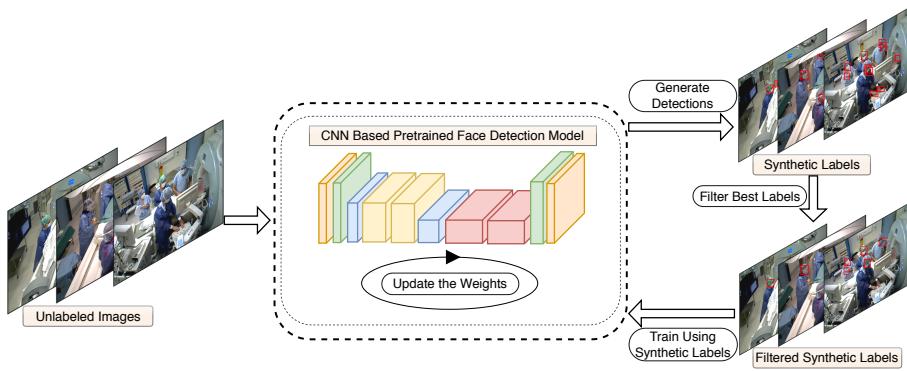
Face detection is a very active research topic in computer vision. Before the rise of deep learning, traditional methods for face detection used machine learning algorithms on top of hand-crafted features [2]. With the advent of deep learning architectures based on convolutional neural networks (CNNs), the performance of face detectors has drastically improved. CNNs are trained end-to-end and are able to learn semantically rich and robust data representations that yield great accuracy. The face detection architectures are often inspired by deep object detectors, whether they are one-stage [3, 4] or two-stage detectors [5]. The one-stage detectors generally divide the image into a grid of boxes and directly classify and regress the localization of objects in each box. The two-stage networks first use a Region Proposal Network (RPN) [6] to extract Region of Interests (ROIs), then a second network to classify and localize each ROI more accurately. These detectors manage to handle the variety of scales, by setting up strategies to specifically detect small faces. They perform contextual reasoning and use image or features pyramids to achieve robustness. The success of these methods can also be attributed to the availability of large-

scale annotated dataset. Indeed, WIDER Faces [7], the standard dataset for training and testing face detection methods in the wild, contains 32,203 images with 393,703 labeled faces. Apart from the bounding box based face detection methods, faces can also be extracted from the face keypoints of human pose estimators, which is performed by mainly two types of approaches: bottom-up and top-down. Bottom-up approaches [8,9] first detect all the keypoints, then assemble them into skeletons, whereas top-down approaches [10,11,12] first detect persons, often with standard object detectors, and then detect keypoints for each detected person using a single person pose estimator. The top-down approaches resolve better the keypoint to person assignment and therefore largely outperform bottom-up approaches on the standard public datasets [13, 14].

The automatic recognition of activities during real surgeries to develop intelligent context-aware assistance systems is a recent field that has started to gain traction in the medical as well as computer vision community [15, 16,17]. Work on analyzing humans in OR videos have generally focused on person bounding box detection and on human pose estimation, using either RGB or RGB-D data [18,19,20]. So far, face detection in the OR has however received very little attention, besides the recent work [21], described below. Furthermore, current state-of-the-art face detectors, even those close to the human-level performance, do not generalize well to OR images. Their inferior generalization can be explained by the fact that they have been trained on natural images, whereas OR images are very specific and challenging: persons' faces are often occluded due to equipment clutter, masks, and glasses. Figure 1 shows some examples of the challenging situations occurring inside the OR.

One standard approach to overcome this visual domain difference is to use transfer learning, which adapts a method by fine-tuning its parameters on an annotated target dataset. In [21], authors recently proposed such a method to detect the faces in the OR by finetuning the Faster-RCNN detector [5] on OR videos. The video dataset consists of youtube OR videos, which have been manually annotated with face bounding boxes. In [21], the results are further improved by using temporal smoothing. Manual annotation can, however, be expensive and time-consuming, whereas non-annotated data is in abundance and often inexpensive. Therefore, this work aims at distilling knowledge from non-annotated OR data to improve the baseline performance of a face detector.

This paper investigates face detection and visual domain adaptation for the OR environment. We first present a comparison of 6 state-of-the-art face detectors. We consider methods where faces can be obtained either directly from bounding box based methods or from face keypoints generated by human pose estimators. We evaluated these methods on the MVOR-Faces, an extension of the MVOR dataset [22] augmented with face bounding box annotations. To the best of our knowledge, this paper presents the first comparison of state-of-the-art face detectors in an OR environment. We also select one detector, SSH [3], and propose to improve it by using an iterative self-supervised method. Several variants of self-supervised methods have been recently used to improve the quality of the synthetic annotations, for instance by using temporal ensembles



**Fig. 2** An iterative approach to adapt a face detector to a target operating room. First, we obtain a trained face detector (SSH [3]) and unlabeled images of the same operating room. Then, we repeat the following steps: (a) use the detector to generate the labels (b) filter the detections with a heuristic to create good quality synthetic annotations (c) retrain the detector using synthetically generated annotations.

[23] or by combining the results of different geometric transformations [24]. In this work, we found it effective to iteratively generate synthetic annotations and fine-tune the model. This approach significantly improves the original model, and largely outperforms the best face detectors on all metrics.

## 2 Methodology

### 2.1 Comparison of State-of-the-art Face Detector

We present below the state-of-the-art methods for face detection used in our comparison. In this study, the faces are represented by bounding boxes. We consider 4 methods where the faces are directly obtained as the output of the detector and 2 methods where the face bounding boxes are generated from face keypoints detected by human pose estimators. These methods are selected based on their ranking on standard public datasets, namely the WIDER dataset [7] for bounding box based face detectors and the COCO dataset [13] for human pose estimators. For reproducibility, we only choose open-source methods.

#### 2.1.1 Bounding Box Based Face Detectors

**Faster-RCNN face detector [5].** The Faster-RCNN, originally designed as a generic two-stage object detector, was trained for the face detection task on the WIDER Faces dataset. First, the RPN generates ROIs with a sliding window approach on deep feature maps. At each sliding window location, anchors, bounding boxes of different scales and aspect ratios, are predicted as either background or ROI. Then, ROIs are pooled and used as input for a second network, which classifies the face and regresses for the exact coordinates of its bounding box.

**Finding tiny faces [25].** This method is specifically conceived to detect faces of different scales. The input of the algorithm is an image pyramid, with three versions of the image: one downsampled, the original and one upsampled image. Each rescaled image is processed by a shared pyramidal CNN, which predicts binary heatmaps for bounding box templates of different sizes.

**SSH: Single Stage Headless Face Detector [3].** This is a one-stage face detector that includes a context module, namely a set of convolutional layers to increase the effective receptive field and different branches to achieve scale-invariance. It uses three different detector networks to predict small, medium and large face anchors. SSH achieves a similar accuracy than the tiny face detector [25], while maintaining real-time performance.

**S<sup>3</sup>FD: Single Shot Scale-invariant Face Detector [4]:** This is also a one-stage face detector inspired by the RPN [6] and SSD [26] architectures. They design strategies to increase the number of positive anchors matching tiny faces during training. Their CNN architecture includes feature maps and detectors that are specific to a range of scales and uses a max-out background label on small anchors to reduce the number of false positives.

### 2.1.2 Human Pose Estimation Based Face Detectors

Human pose estimation aims to localize the anatomical keypoints of all the persons present in an image. These anatomical keypoints are spread across the whole body including the face keypoints. Current state-of-the-art methods are trained on the COCO dataset, which includes five face keypoints (nose, left eye, right eye, left ear, right ear). We fit a bounding box of fixed size, 30x30 pixels, at the average location of these five keypoints to extract the face bounding box. We now describe the pose estimation methods that we used for the evaluation.

**OpenPose [8].** This is one of the best bottom-up approaches for human pose estimation. First, a CNN predicts confidence heat-maps for all keypoints and part affinity fields for each joint. These maps are then used to construct a graph of keypoints and body joints. Finally, this graph is parsed using a bi-partite graph matching algorithm to produce a set of human poses.

**AlphaPose [10].** This is one of the state-of-the-art top-down methods. It uses Faster-RCNN [6] to detect persons. Then, cropped human bounding boxes are processed by a single person pose estimator, which is composed of several modules, including spatial transformers, to refine the keypoint detections in the bounding box. This method successfully handles the problems caused by inaccurate and duplicate bounding boxes.

## 2.2 Iterative Self-supervised Approach for Face Detection in the OR

We use the following two steps to improve the selected state-of-the-art face detector on OR images: 1. Generation of the unlabeled dataset 2. Iterative refinement using a self-supervised approach.

### 2.2.1 Generation of the Unlabeled Dataset

We use an unlabeled dataset of 20k images generated from videos captured in the OR. The videos were collected on days different from the ones of the test dataset to ensure the absence of overlap between the unlabeled dataset and the test dataset. We then use OpenPose [8], a multi-person pose estimator, on the OR videos to get the approximate number of persons in each frame. The computational efficiency of OpenPose allows us to make the inference on the entire dataset in a reasonable time. We divide the images into four categories: images with one, two, three, and four or more detected persons. Since OpenPose also gives a confidence score for each detected skeleton, we average the scores of the detected skeletons and take the 5k highest-scored images from each category (i.e., 20k images overall). This selection method ensures that the images contain persons in different numbers.

### 2.2.2 Iterative Refinement using Self-supervised Approach

We utilize an iterative self-supervised approach to adapt the state-of-the-art model to the target OR dataset. This approach consists of fine-tuning the model on a subset of its own detections. We use SSH [3], pre-trained on WIDER Faces, as the CNN-based model for the iterative refinement. We choose this model because it has high computational efficiency and also yields state-of-the-art results on WIDER Faces. This detector is then used to generate synthetic labels on the unlabeled dataset. To select quality face bounding boxes, we use a simple yet effective heuristic criteria: with a dataset of  $N$  images, we select the best  $2*N$  detections. Since we have approximately 2.5 persons/image in the unlabeled dataset,  $2*N$  best detections contain the face bounding boxes with a high recall. These synthetically annotated images are then used to finetune the original model. We perform these steps iteratively to improve the detections and the detector at each iteration as shown in Fig. 2. It is to be noted that no validation set is available as we did not use any supervised annotation. Therefore, our experiments differ from the traditional deep learning experiments, which fine-tune the hyper-parameters based on the performance on a validation set. We mainly conduct the fine-tuning experiments with a different number of training batches before relabelling and different iteration numbers. We present the result of each experiment on the test-set.

## 3 Experimental Setup

### 3.1 Test Dataset (MVOR-Faces)

We compare the state-of-the-art face detectors on MVOR-Faces, a dataset of operating room images captured during real surgical procedures. MVOR-Faces is an extension of the public MVOR dataset [22], which consists of 732 multi-view frames (2196 images) recorded in an interventional room. In the MVOR

dataset, faces of persons without a mask and nude parts of patients are fully blurred, and the persons with masks are blurred only on the eyes. MVOR-Faces contains the same images as MVOR, except that the eyes of the persons wearing a mask are not blurred. Also, it contains the manually annotated face bounding box of all visible faces wearing a mask. All fully-visible faces and nudity zones are still blurred in the MVOR-Faces as needed for anonymity. Overall, the dataset contains 2262 face bounding boxes for 2196 images.

### 3.2 Evaluation Metrics

We use the standard metrics for object detection from COCO [13], i.e. Average Precision (AP) and Average Recall (AR).  $AP^{IoU}$  is the average precision at a fixed intersection over union (IoU), and AP is the average of  $AP^{IoU}$  at different IoU thresholds. While the public implementation averages between IoU of 0.5 and 0.95 with a step of 0.05, we average between an IoU of 0.3 and 0.95 to support a slightly looser metric. The consideration of a slightly looser metric is motivated by the fact that face detection in a medical context is quite challenging: clinicians wear mask, glasses, and hats, and are often occluded. Therefore, a looser metric reduces the bias in favor of the face detectors.

## 4 Results and Discussion

### 4.1 Comparison of State-of-the-art Face Detectors

Table 1 shows the comparison of state-of-the-art face detectors. The first four methods in Table 1 directly output face bounding boxes, as described in section 2.1.1. The next two methods detect the human skeletons, including face keypoints (i.e. ears, eyes and nose). We extract the face bounding boxes from face keypoints as specified in Section 2.1.2. Unless otherwise stated, we use the exact same models provided by the authors, without modifying any hyper-parameter.

On AP(0.3:0.95), Tiny Face detector [25] is the best model, with 0.340; SSH [3] and S3FD [4] are close, with respectively 0.314 and 0.302. On AP(0.3), AlphaPose is the best model with 0.785. With this looser metric, human pose estimators perform better. Indeed, in the OR environment, when clinicians wear mask and hats, face detectors cannot rely on the same features as in the outside environment, such as the mouth shape and the nose. Human pose estimators, which also detect other body keypoints, are more robust than face detectors. However, they do not localize the bounding boxes accurately enough to perform well on stricter metrics. For comparison, we also provide the results on the original MVOR dataset, where the eyes are blurred, in Table 2. The results show a significant drop in the performance highlighting the importance of the eyes for face detection.

Overall, results of state-of-the-art detectors show a large margin for improvement on the MVOR-Faces dataset. With an IoU of 0.5, which is a less strict



(a) Faster-RCNN [6]



(b) S3FD [4]



(c) Tiny Face [25]



(d) SSH [3]

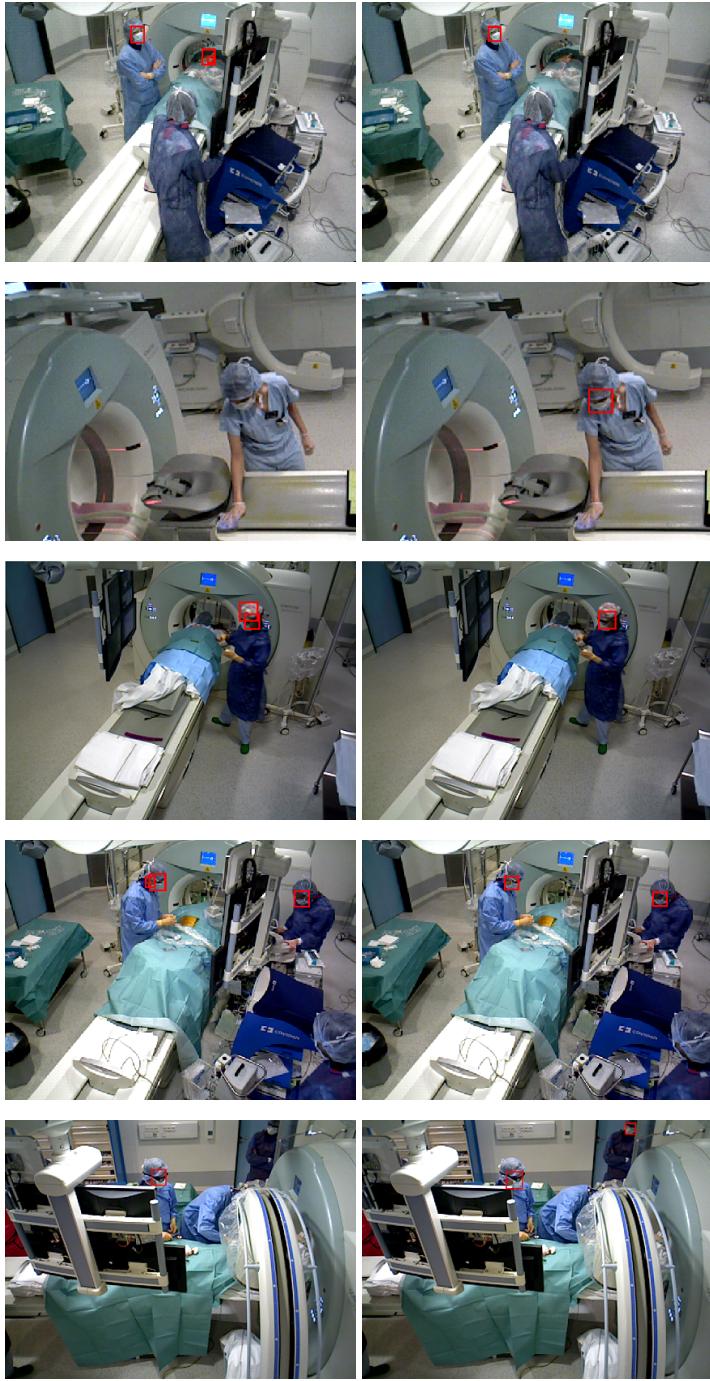


(e) AlphaPose [10]



(f) OpenPose [8]

**Fig. 3** Qualitative results from the face detectors evaluated on MVOR-Faces. The displayed detections were selected based on a score threshold of the detector corresponding to a recall threshold of 70% at an IoU of 0.3.



**Fig. 4** Comparison of the original SSH model (left column) with the best self-supervised model trained with our iterative approach (right column). To filter the displayed detections, we use the score threshold corresponding to a recall threshold of 70% at an IoU of 0.3, as in Fig. 3. The self-supervised model detects much harder examples, with occlusion or uncommon poses.

metric, the AP of the best model is only 0.556. On the WIDER Faces dataset, tiny face detector [25] achieves 0.819 using the same metric, while SSH [3] reaches 0.944 and S3FD [4] 0.958. Qualitative results shown in Fig. 3 illustrate some of the mistakes made by the state-of-the-art detectors on the MVOR-Faces dataset, e.g. multiple detections, false positives, false negatives.

#### 4.2 Iterative Self-supervision

As mentioned in section 2.2, we use SSH [3] for the self-supervised process. We conduct several experiments on self-supervision to demonstrate the interest of this iterative approach. During training, we use the following hyper-parameters: stochastic gradient descent with a learning rate of 0.04, momentum of 0.9 and weight decay of  $5e^{-4}$ . The batch size is 2. Anchors, which correspond to a location ( $x, y$ ) in the image and a predefined bounding box size (width, height), are considered as positives if their IoU with a ground-truth bounding box is greater than 0.5, as negatives otherwise. During inference, we use an image pyramid of four levels, as the authors. The aspect ratio of each rescaled image is preserved. The weights are initialized with the ones provided by the authors, after training on the WIDER Faces dataset.

In Table 4, we provide the test-results of our proposed iterative process, with different hyper-parameters (number of iterations and number of training batches used before relabelling). When relabelling, we filter the detections with the same criteria as explained in section 2.2, i.e.  $2^*N$  best detections where  $N$  is the number of images. The training is done with the same parameters as mentioned above. The model which performs best on the test dataset is achieved at iteration 4 when training with 2000 batches before regenerating the labels. The model outperforms the state-of-the-art with a large margin on all metrics. On AP(0.5), it outperforms tiny face detector [25] by more than 9%, and the original SSH model by 13.1%.

In Fig. 4, we compare a few detections from the original SSH and the best self-supervised model on the test-set. The latter detects much harder examples, with occlusion or uncommon poses, and has fewer false positives.

In Table 3, we show an ablation study of self-supervision for domain adaptation, with no relabelling of target images by the self-supervised model. The 20k images of the unlabeled dataset are annotated with the detections of the original SSH model. We filter the predictions with the same criteria: since we have 20k images, we take the best 40k detections. Then, the model is fine-tuned by training on synthetically annotated images on 15k training batches. We observe a quick saturation on the test-set, MVOR-Faces: the AP(0.3:0.95) reaches 0.372 after 1k batches and 0.378 after 10k batches (i.e., with one epoch on the entire unlabeled dataset). At 15k batches, the AP(0.3:0.95) is back at 0.372. The quick saturation of this process highlights the interest of our iterative approach.

Detector	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Faster-RCNN [6,5]	0.254	0.651	0.407	0.345
S3FD [4]	0.302	0.627	0.486	0.395
Tiny Face [25]	0.340	0.734	0.556	0.428
SSH [3]	0.314	0.704	0.517	0.421
AlphaPose [10]	0.279	0.785	0.463	0.358
OpenPose [8]	0.240	0.776	0.365	0.316
Self-supervised SSH	<b>0.402</b>	<b>0.800</b>	<b>0.648</b>	<b>0.474</b>

**Table 1** Results of state-of-the-art face detectors on MVOR-Faces. First four methods are bounding box based face detectors. AlphaPose and OpenPose are human pose estimators, from which face bounding boxes are generated from the face keypoints. Results show the margin for improvement on the MVOR-Faces dataset.

Detector	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Tiny Face [25]	0.237	0.627	0.369	0.331
SSH [3]	0.229	0.600	0.368	0.368
AlphaPose [10]	0.239	<b>0.742</b>	0.370	0.323
Self-supervised SSH	<b>0.306</b>	0.711	<b>0.492</b>	<b>0.403</b>

**Table 2** Comparative study: results of state-of-the-art face detectors on MVOR [22], the public version of MVOR-Faces. Here, clinicians wearing a mask are blurred around the eyes. Results show the significant decrease in the performance as compared to MVOR-Faces shown in Table 1.

Number of training batches	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Original SSH model [3]	0.314	0.704	0.517	0.421
1000	0.372	<b>0.781</b>	0.597	<b>0.462</b>
2000	0.373	0.769	0.593	0.458
3000	0.374	0.770	0.595	0.458
5000	0.372	0.770	0.598	0.454
10000	<b>0.378</b>	0.778	<b>0.608</b>	0.462
15000	0.372	0.781	0.600	0.457

**Table 3** Comparative study: training SSH model without re-generating the synthetic labels. One training batch is composed of two images. Here, the self-supervised model is trained on the images annotated by the original SSH model and initialized with SSH weights. The AP saturates quite fast. After 1k batches, the AP does not significantly increase. The best results are much lower than best results with the iterative approach in Table 2.

## 5 Conclusion

We propose the first broad evaluation of state-of-the-art face detectors on OR images. Since the results show a large margin for improvement, we also propose to use an iterative self-supervised approach to adapt a face detector to a given OR. It consists of gathering images of the target environment, generating synthetic annotations with a model trained on a manually annotated dataset, and retraining it iteratively using the synthetic labels. This method is generic and applicable to any OR configuration. Our self-supervised detector outperforms the state-of-the-art on MVOR-Faces by a large margin, namely by more than 6% on AP(0.3:0.95). By significantly improving the accuracy of face detection, we show that self-supervision is a promising direction to

Iteration	Number of training batches before relabelling	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
	Original SSH model [3]	0.314	0.704	0.517	0.421
1	1000	0.373	0.782	0.598	0.462
2		0.365	0.783	0.590	0.461
3		0.367	0.785	0.592	0.450
4		0.365	0.788	0.594	0.456
1	2000	0.373	0.773	0.596	0.459
2		0.385	0.796	0.622	0.466
3		0.393	<b>0.808</b>	0.630	0.465
4		<b>0.402</b>	0.800	<b>0.648</b>	<b>0.474</b>
1	3000	0.373	0.772	0.595	0.457
2		0.377	0.793	0.608	0.452
3		0.383	0.806	0.614	0.459
4		0.371	0.797	0.604	0.448

**Table 4** The iterative process of self-supervision with different hyper-parameters. One iteration consists of three steps: (1) Generate predictions on the unlabeled dataset with the last model. (2) Filter detections: select the best  $2N$  detections on  $N$  images. (3) Retrain the model on 1k, 2k or 3k training batches. When training with 2k or 3k batches before relabelling, it improves the results from the baseline approach (see in Table 3). One batch is composed of two images.

transfer state-of-the-art computer vision approaches to the medical context, where annotations are challenging to generate.

**Acknowledgements** This work was supported by French state funds managed by the ANR within the Investissements d’Avenir program under references ANR-16-CE33-0009 (DeepSurg), ANR-11-LABX-0004 (Labex CAMI) and ANR-10-IDEX-0002-02 (IdEx Unistra). The authors would also like to thank the members of the Interventional Radiology Department at University Hospital of Strasbourg for their help in generating the dataset.

## References

- Kenny Chen, Paolo Gabriel, Abdulwahab Alasfour, Chenghao Gong, Werner K Doyle, Orrin Devinsky, Daniel Friedman, Patricia Dugan, Lucia Melloni, Thomas Thesen, et al. Patient-specific pose estimation in clinical environments. *IEEE Journal of Translational Engineering in Health and Medicine*, 2018.
- Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages I–I, 2001.
- Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *ICCV*, pages 4885–4894, 2017.
- Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S<sup>3</sup>fd: Single shot scale-invariant face detector, 2017.
- Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn, 2017.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.
- Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016.
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. 2017.

9. Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, pages 34–50, 2016.
10. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.
11. Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.
12. Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. 2018.
13. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.
14. Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014.
15. Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Multi-stream deep architecture for surgical phase recognition on multi-view rgbd videos. In *MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2016.
16. Lena Maier-Hein, Swaroop Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science: enabling next-generation surgery. *Nature Biomedical Engineering*, 2017.
17. Serena Yeung, N Lance Downing, Li Fei-Fei, and Arnold Milstein. Bedside computer vision-moving artificial intelligence from driver assistance to patient safety. *NEJM*, 378(14):1271, 2018.
18. Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Articulated clinician detection using 3d pictorial structures on rgbd data. *Medical image analysis*, 35:215–224, 2017.
19. Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. A multi-view rgbd approach for human pose estimation in operating rooms. In *WACV*, pages 363–372, 2017.
20. Vasileios Belagiannis, Xinchao Wang, Horesh Ben Shitrit, Kiyoshi Hashimoto, Ralf Stauder, Yoshimitsu Aoki, Michael Krantzfelder, Armin Schneider, Pascal Fua, Slobodan Ilic, et al. Parsing human skeletons in an operating room. *Machine Vision and Applications*, 27(7):1035–1046, 2016.
21. Evangelos Flouty, Odysseas Zisisopoulos, and Danail Stoyanov. Faceoff: Anonymizing videos in the operating rooms, 2018.
22. Vinkle Srivastav, Thibaut Issenhuth, Kadkhodamohammadi Abdolrahim, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgbd operating room dataset for 2d and 3d human pose estimation. In *MICCAI-LABELS-2018*, 2018.
23. Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.
24. Ilija Radosavovic, Piotr Dollr, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning, 2018.
25. Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, 2017.
26. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.