2D-Pose Based Human Body Segmentation for Weakly-Supervised Concealed Object Detection in Backscatter Millimeter-Wave Images*

Lawrence Amadi[1][0000−0003−2913−4056] and Gady Agam[1]

Visual Computing Lab, Illinois Institute of Technology, Chicago IL 60616, USA
lamadi@hawk.iit.edu
agam@iit.edu

Abstract. The detection and localization of anomalies in backscatter images of a person is a standard procedure in airport security screening. Detecting a concealed item on a person and localizing the item to a specific body part requires the ability to recognize and segment distinct body parts. This can be challenging for backscatter images compared with RGB images due to lacking chromaticity cues and the limited availability of annotated backscatter images. To address this problem, we propose a weakly-supervised method for anomaly detection on human body parts which is based on an unsupervised body segmentation procedure that uses keypoints from a pretrained pose estimator to segment backscatter images without significant performance degradation. The paper presents a method for adapting a pretrained RGB pose estimator to segment human body parts in millimeter-wave images. We then train a body part-aware anomaly detection classifier to detect foreign objects on the body part. Our work is applied to TSA's passenger screening dataset containing backscatter millimeter-wave scan images of airport travelers with binary labels that indicate whether a concealed item is attached to a body part. Our proposed approach significantly improves detection accuracy on 2D images from the baseline approach with a state-of-the-art performance of 97% F1-score and 0.0559 log-loss on TSA-PSD test set.

Keywords: Pose refinement · Body segmentation · Object detection.

1 Introduction

Backscatter images such as X-ray, MRI, and millimeter-wave scanner images are predominantly used to examine internal organs or beneath the clothing of persons. These images are generated by specialized scanners and are essential for computer-aided medical diagnosis and security screening inspections. Backscatter images are also characterized by very low chromaticity and illumination. The low distinctive visibility makes human inspection difficult and also pose a challenge for computer diagnostic software. In the US, the HIPAA Privacy Rule considers these intrusive backscatter images of persons their personal data and

* Obtain source code at github.com/lawrenceamadi/RaadNet

protects it against unauthorized inspection. Hence, in the case of airport security screening where full-body backscatter millimeter-wave scan (MWS) images of persons must be inspected to make sure that they do not conceal prohibited or harmful items under their garment, a computer vision algorithm is used to analyze and detect anomalies in the images. When an anomaly is detected, the algorithm must also localize it to a specific body part as it recommends a follow-up pat-down search of the indicated body part. Therefore, localizing anomalies to specific body parts is as important as high precision anomaly detection of concealed items to streamline pat-down searches shorten airport security screening queues. This requirement demands an algorithm capable of recognizing different body parts in backscatter MWS images. Similarly, for medical imaging, there is value in developing an anomaly detection algorithm that is body-organ aware to assist physicians in diagnosis.

Fig. 1: Example of our refined 2D pose (2^{nd}) and unsupervised body segmentation (4^{th}) on MWS images compared to SOTA HRNet 2D pose (1^{st}) [28] and DensePose body segmentation (3^{rd}) [14].

Designing a body-aware anomaly detection algorithm for backscatter images is challenging for two reasons. First, unlike RGB images, there are hardly sufficiently large datasets of backscatter MWS images with body part bounding-box or pixel-level annotations to supervise the training of a backscatter body part segmentation deep neural network. Second, as demonstrated in Fig. 1, directly applying pretrained RGB body segmentation models (e.g. DensePose [14]) to MWS images fails to produce meaningful segmentation because of lacking chromatic and illumination cues.

This work makes the following contributions:

1. We introduce an unsupervised procedure for segmenting body parts in MWS images by estimating bounding-polygons for each body part.
2. We then propose a weakly-supervised, RoI-attentive, dual-chain CNN classifier that detects anomalies given multi-view images of a cropped body part.

Our approach leverages multi-view information to refine sub-optimal poses generated by RGB-pretrained human pose estimators. The refined keypoints are then used to estimate bounding-polygons that enclose each body part. Subsequently, the bounding-polygons are used to crop regions of the images that represent each body part and the images are fed to our body-aware anomaly detection neural network.

1.1 TSA Passenger Screening Dataset

Our unsupervised body segmentation method is evaluated on the Transportation Security Administration Passenger Screening Dataset (TSA-PSD) [32] which contains backscatter full-body scans of persons acquired by a High Definition Advanced Imaging Technology (HD-AIT) millimeter wave scanner (MWS). The dataset contains 2,635 scans of airport travelers. Each scan is encoded as a **Projected Image Angle Sequence File (*.aps*)** and contains a sequence of 16 2D images captured from different viewpoints such that the person appears to be spinning from left to right when the images are played back frame by frame. 1,247 of the 2,635 scans are the annotated train-set with binary labels that indicate whether an object is concealed in a body part of the scan subject. TSA outlines 17 body parts of interests. They include right and left forearms, biceps, abdomens, upper and lower thighs, calves, ankles, chest, upper back, and groin. Hence, a scan with 16 images has 17 binary labels. There are no pixel-level or image-level ground-truth annotation of concealed items or body parts. There are also no binary labels per body part, per frame.

2 Related Work

2D pose estimation on RGB images is widely studied in [3, 4, 16, 24, 28, 30]. However, these models require keypoint annotations to learn proper pose encoding. Similarly, state-of-the-art human body segmentation neural networks [8, 10, 12, 14, 15, 17, 19, 20, 22, 27, 33, 34, 37] rely on bounding-box or pixel-level annotations of body parts.

Anomaly object detection in persons, luggage, cargo containers, and scenes are studied in [1, 2, 5, 13, 23, 25, 29]. A majority of the leading methods are based on deep neural networks. Riffo and Mery [25] propose a shape implicit algorithm for detecting specific threat objects (razor blades, shurikens, and guns) in x-ray images of luggage. Although their method can be modified to detect threat items on the human body, their object-specialized approach is not expected to generalize to unknown objects that are not encountered during training because their algorithm is designed to detect specific items, not general anomalies. A popular approach for detecting threat objects uses AlexNet classifier [2] and predefined fixed region-of-interest (RoI) bounding-boxes to segment body parts. The fixed RoI bounding-boxes do not account for variations of body part size, positioning, and orientation on a per person basis. This limitations makes this approach more suitable for the less mobile torso body parts (e.g. chest, back) and limited viewpoints. Another approach that uses AlexNet for anomaly detection [13] combines 2D and 3D data to segment the body parts and generate threat or benign labels for each cropped image. This enables supervised training of the model using a set of cropped body parts with assigned threat labels. This approach allows for a simpler neural network architecture because of the 1-1 mapping of cropped images and labels. However, generating false labels for cropped images can degrade to the accuracy of the classifier. Note, anomalies in TSA-PSD body parts are typically visible in 6 frames or less.

Other concealed item detection algorithms applied to TSA-PSD are either designed to use 3D volumetric data, or a combination of 3D and 2D images. We give a high-level description of the proprietary TSA-PSD classifiers as reported by the Department of Homeland Security as details of the state-of-the-art classifiers are not released to the public. Jeremy Walthers (1^{st}) approach used an array of deep learning models customized to process images from multiple views. Sergei Fotin (2^{nd}) and Oleg Trott (5^{th}) adopted an approach that fuses 2D (10-41 MB per file) and 3D (330 MB per file) data sources to make object and location predictions. Despite their high accuracy, the 1^{st} and 2^{nd} approach may be less suitable for real-time use because the inference time for an array of neural networks or very large files can be substantial. David Odaibo and Thomas Anthony (3^{rd}) developed an algorithm that uses specialized 3D image-level annotations to train a 2-stage identification model. It is unclear whether the annotations were automated or manually labeled. Location based models (4^{th}), automatic image segmentation with a collection of specialized models trained on cropped body part images (6^{th}), separately trained models with image augmentation (8^{th}), and the use of synthetic data and cross-image analysis (7^{th}) are other techniques used to improve model detection accuracy.

3 Proposed Method

We approach concealed item detection and association to body parts as a two-stage problem. First, we segment the human body parts in the frames of each scan to generate 17 sequences of $n \leq 16$ cropped images (Sec. 3.1). Each sequence corresponds to a body part and contains cropped images of that body part from different viewpoints. Since the presence of a concealed item is often visible in 6 frames or less, never from all viewpoints, we must detect anomalies in body parts on a per image-sequence basis. In the second stage (Sec. 3.2), we train a deep CNN anomaly detector that processes a sequence of cropped images of a body part and classifies it as benign or abnormal when a foreign object is detected in any of the cropped images. Note, a single CNN detector is trained for all body parts. This makes our detector simpler and lightweight compared to other state-of-the-art classifiers that use an array of body part or gender specialized networks. In contrast, our network uses a novel region-attentive architecture that makes it body part aware.

3.1 Body Part Segmentation from 2D Poses

Our approach for adapting an RGB *Human Pose Estimation (HPE)* network to perform unsupervised body segmentation begins with correcting local-optima keypoint locations in confidence map output of a RGB-pretrained 2D pose estimator.The corrected keypoint positions in each frame are further optimized using RANSAC bundle adjustment [35] to consolidate a global-optimum 3D pose. A new set of coherent 2D poses are derived by projecting the global-optimum 3D pose back to the 2D frames. The refined keypoints in each frame are then used to estimate bounding-polygons that segments the body parts in each frame.

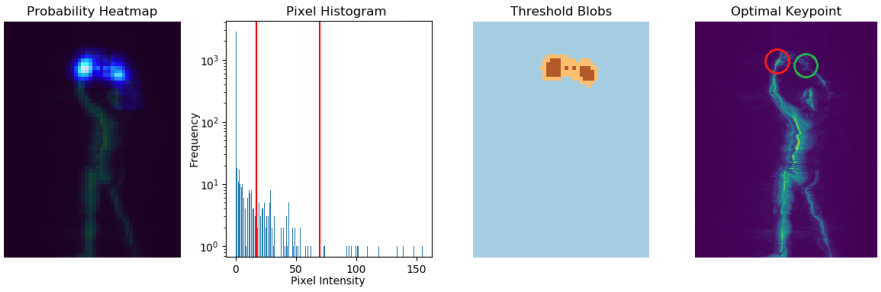

Fig. 2: Outcome of our keypoint selection algorithm for Right Wrist. *A*-*D* from left-right. *A* depicts the right wrist 2D confidence map. *B* is the histogram of the confidence map. The *red* lines indicate the multi-Otsu thresholds used to segment *A* the confidence map to 3 layers; *blue*, *orange*, and *brown* (detected blobs) in *C*. *D* shows our algorithm selects the correct position of the right wrist (*green* circle) instead of the location of highest confidence (*red* circle).

Keypoint Selection from HPE Confidence Maps. Without keypoint annotations to train a pose estimator on the TSA dataset, we use a Deep-HRNet 2D pose estimator [28] pretrained on the COCO dataset [21] to estimate 15 keypoints of persons in MWS images. They include right and left wrists, elbows, shoulders, hips, knees, ankles, head, neck, and pelvis keypoints. Deep-HRNet is preferred to other 2D HPE networks [3,4,24] because of its high-resolution architecture. Compared to the others, it estimates more realistic poses of the subjects in backscatter MWS images. Typically, the position of a keypoint is derived from the corresponding 2D confidence map output of the pose estimator as the pixel location with the highest confidence. We observed that this naive method often produced incorrect keypoint estimates because the Deep-HRNet estimator often generated confidence maps with more than one concentrations of high confidence scores (i.e. blobs) for backscatter MWS images. In such cases, the naive selection will default to the leftmost blob even when the correct keypoint position is in one of the other blobs.

We implement a keypoint selection post-processing procedure that selects the better-positioned blob and keypoint location given the occurrence of multiple blobs in the confidence map. The premise of our keypoint selection algorithm is that the relative positioning (left, center, or right) of joints in each frame is consistent across all scans because subjects assume a standard posture (standing erect with hands raised) when being scanned. And their pose is sustained while each frame is captured from rotating viewpoints. We begin by segmenting the confidence map into three layers using multi-Otsu binarization [36] to determine the confidence threshold for each layer. Using a modified *island-finder* algorithm, we traverse the segmented confidence map (now a 2D matrix with three unique values; 0, 1, 2) to identify all blobs. The blobs are grouped into three clusters by spatial proximity. The cluster nearest to the expected keypoint position (left, center, or right) is selected. We then compute the $argmax$ of confidence scores in the chosen blob cluster to retrieve the pixel position of the keypoint. Fig. 2 illustrates the outcome of this procedure.

Multi-view Coherent Pose Optimization. We observed that even after refining the 2D pose in each MWS frame, some keypoints may still be sub-optimally estimated in some frames. Subsequently, causing inaccurate segmentation of body parts associated with the keypoint. Therefore, we ought to correct incoherent poses across all frames for each scan. In Algorithm 1 we describe a pose optimization procedure that takes advantage of the multiple viewpoints of TSA-PSD MWS images to reconstruct consistent 2D poses across all frames. The 2D position of each keypoint, across all frames of a scan, are optimized independently using RANSAC bundle adjustment [6, 9, 11, 31, 35].

Algorithm 1 Per Keypoint RANSAC Bundle Adjustment

Input: $P \leftarrow \{(x, y) : \forall\ frames\ f_1 \dots f_{16}\}$ ▷ 2D pixel positions of keypoint in each frame
Output: $P^{"}$ ▷ 2D pixel positions of keypoint in each frame after bundle adjustment
$n \leftarrow 0$, $I \leftarrow \emptyset$
while $n \leq 100$ **do** ▷ iterate over subset of keypoints for 3D bundle adjustment
  $R \leftarrow \{(x, y) :\subset P\}$ ▷ randomly selected subset, 1 every 4 consecutive frames
  $p^{3D} \leftarrow (x, y, z)_R$ ▷ 3D point is regressed from R via least squares optimization
  $P^{'} \leftarrow \{(x^{'}, y^{'}) : \forall\ frames\}$ ▷ 2D positions after projecting p^{3D} to each frame
  $I^{'} \leftarrow \{(x, y) :\subset P\}$ ▷ note inlier points based on Euclidean dist. between P & $P^{'}$
  if $|I^{'}| > |I|$ **then**
    $I \leftarrow I^{'}$ ▷ retain the largest inlier set
  end if
end while
$p^{3D} \leftarrow (x, y, z)_I$ ▷ least squares bundle adjusted 3D point regressed from I 2D points
$P^{"} \leftarrow \{(x^{"}, y^{"}) : \forall\ frames\}$ ▷ final 2D positions after projecting p^{3D} to each frame

Estimating Bounding-Polygons for Body Segmentation. After refining the keypoints, we segment the body parts in each frame by defining an oriented bounding-box around each body part. Vertices of the bounding-polygon (a quadrilateral with 4 vertices) are estimated from a subset of keypoints associated with a given body part. We define two types of body parts. ***Limb Body Parts*** are segmented using a pair of keypoints. They include forearms, biceps, upper and lower thighs, calves, and ankles. We refer to the pair of keypoints used to segment limb body parts as *anchor keypoints*. ***Torso Body Parts*** are segmented using a set of four keypoints. They include chest, back, abs, and groin. We refer to the keypoints used to segment torso body parts as *pillar keypoints*.

Limb Body Part Segmentation. We begin by computing the angle between the *anchor keypoints* and the y-axis. The image is then rotated by the computed angle so that the limb is vertically aligned. We extract the luminance channel of the rotated image and remove noise with a Gaussian filter. We then sum the pixel intensity along the horizontal axis of a rectangular region enclosing the limb. This produces a *pixel intensity curve*. The rectangular region is vertically bounded by the y-coordinates of the rotated keypoints, and horizontally bounded by a predefined width for each limb. Next, we fit a degree 6 polynomial line to the computed pixel intensity curve and extract the x-coordinates of the

rightmost and leftmost local minima of the polynomial line. The x-coordinates of the local minima define the width of an axis-aligned bounding-box around the limb. Similarly, the y-coordinates of the rotated keypoints define the height. The bounding-box is transformed to an oriented bounding-polygon when its vertices are inversely rotated.

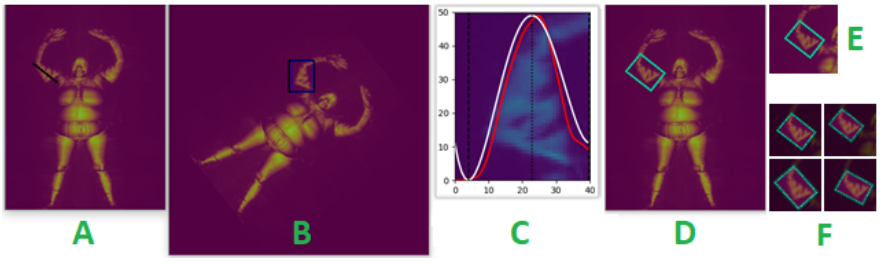

Fig. 3: Visualization of intermediate stages of limb segmentation for the Right Elbow. Black line in *A* links the Right Elbow and Shoulder keypoints. The frame is rotated to vertically align the pillar keypoints in *B*. *C* shows the computed *pixel intensity curve* (red line) and fitted polynomial (white) line. *D* shows the estimated bounding-polygon. *F* are examples of (shift, zoom, rotate) cropped image augmentation (with RoI mask) generated from *E*.

Torso Body Part Segmentation. Pillar keypoints of torso body parts typically outline a quadrilateral region containing two or more body parts. For example, the right shoulder, neck, pelvis, and right hip keypoints segment the right-Abdomen and half of the upper chest (see image A in Fig. 4). To precisely capture the intended body part, we shift one or more edges of the quadrilateral. The edges that are moved, the direction (horizontally or vertical), and the extent they are shifted is guided by a predefined configuration for each torso body part. The new vertices of the bounding-polygon are computed as the points of intersection of the adjusted quadrilateral edges. This procedure is illustrated in Fig. 4 for the right abdomen.

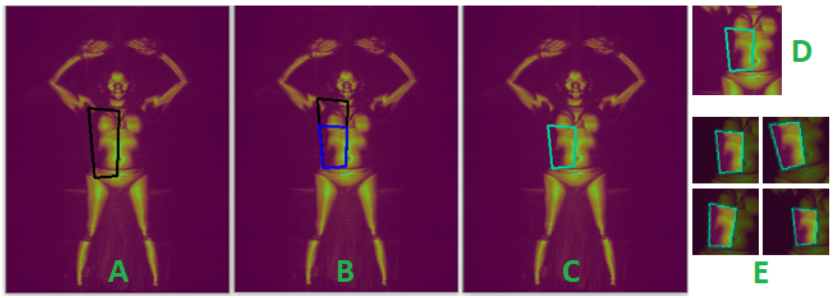

Fig. 4: Visualization of torso segmentation for the Right Abdomen. Vertices of black quadrilateral in *A* are the Pelvis, Neck, Right Shoulder and Hip keypoints. The top edge of the black quadrilateral in *B* is shifted downwards, resulting in the blue polygon. The green polygon in *C* encloses the RoI for the Right Abdomen. *E* shows examples of (shift, zoom, rotate) randomly generated cropped image augmentation from *D*, overlaid with RoI mask.

Each bounding-polygon defines the region-of-interest (RoI) of a body part. The segmented body parts are cropped, in excess, by a standard 160×160 pixel window such that the RoI is contained in the cropped image (see E and F in Fig. 3 and Fig. 4) before down-sampling to 80×80 pixels. This approach preserves the aspect ratio of the body parts in contrast to directly resizing the RoI to the standard size. We found our network performed better at detecting concealed items when the aspect ratio of images are not altered. We have chosen a generous standard size of 160 sq. pixels to accommodate the sizes of all body parts and all subjects, big and small. Another reason for cropping in excess of the RoI is because the demarcation between neighboring body parts is not absolute and when a concealed object spans the boundary of two body parts must only be attributed to one. By over-cropping the RoI and incorporating the RoI mask and our proposed Region Composite Vectors into the network, our model learns to associate concealed objects to the dominant body part. We have designed our network architecture to use the RoI mask to refocus attention on the RoI when detecting anomalies, but only after extracting features from the entire cropped images.

To summarize, given each scan, we compile 17 sequences of cropped images for each body part (from multiple viewpoints). Each sequence contains 12 images because we observed that the maximum number of frames where a body part is visible is 12. Cropped images are re-sampled and augmented to compensate for body parts that are visible in less than 12 frames (as low as 3 for chest and back). Finally, images are downsampled by a factor of 2.

3.2 RaadNet: RoI Attentive Anomaly Detection Network

We design an anomaly object detection network that classifies a sequence of cropped images of a segmented body part as benign or abnormal. Indicating the presence of a concealed item in one or more of the cropped images. The network, illustrated in Fig. 5, takes an input sequence of n=12 cropped images for each body part, their corresponding RoI binary masks, and *Region Composite Vectors* (RCVs). The $\boldsymbol{RCV}$ of a cropped image is a vector of size 17 defined in Eq. (1) as the *intersect-over-union* (IoU) between the body part's RoI I_i and the RoI of all body parts in the given frame I.

$$RCV_i(I) = \langle IoU(I_0, I_i),\ \ldots,\ IoU(I_{16}, I_i)\rangle, \quad i \leq 16 \tag{1}$$

RCVs numerically summarize the proportion contribution of the body parts captured in a cropped image. We expect most of the IoU components of the vector will be 0, with only a few having values greater than 0 (as is the case for adjacent body parts). The component corresponding to dominant body part will have the highest value. We found RCVs provide cues to the network that helps it better resolve overlap conflict when a concealed item is partially contained in the RoI. The cropped images pass through feature extraction blocks and the masks are downsampled to match the dimensions of the extracted features. We use the first 5 blocks of MobileNetV2 [26] (pretrained on ImageNet [7]) for

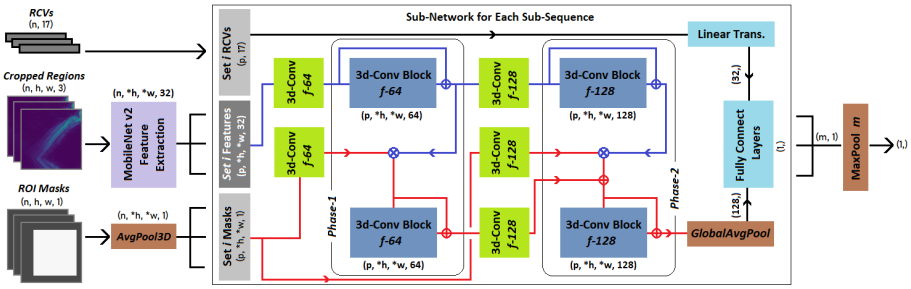

Fig. 5: An instance of RaadNet, our 2-phase, dual-pipeline *(indicated by the blue and red lines)* anomaly detection network that takes as input cropped images of a body part, their RoI masks and RCVs, and outputs the probability that a concealed item is in either of the cropped images. We use $n{=}12$ cropped images per body part. $h,w{=}80$ are the height and width of the cropped images. $^*h,^*w{=}10$, $p{=}3$, $m{=}4$. Each sub-sequence of images p is passed through the same sub-network (enclosed in large rectangle). Residual convolution blocks (in dark-blue) contains two 3D convolution layers with $kernel{=}3$ and f filters. Convolutions are accompanied by batch normalization and Re-LU activation. The fully connected block (in light-blue) contains 5 dense layers of sizes 128,64,64,16,1 followed by a sigmoid activation. $\otimes$ and $\oplus$ are element-wise multiplication and addition operations. Notice that a deeper network can be created by increasing the number of phases.

feature extraction. The extracted features, downsampled masks, and RCVs are separated into $m{=}4$ sub-sequences. Each containing $p{=}n/m$ contiguous temporal components that are fed to a dual-pipeline, multi-phase sub-network. During each phase, the *Image-pipeline* (*blue path* in Fig. 5) encodes textures of the entire cropped images, while the *RoI-pipeline* (*red path* in Fig. 5) is designed to extract textures precisely from the RoIs in the cropped images. This is achieved by computing the element-wise multiplication of the residual convolution block output of the *Image-pipeline* and the convoluted RoI masks. The resulting tensor is passed to the residual block of the *RoI-pipeline*. The dual pipelines ensures the network can detect anomalies that partially appear on the boundary of the body part's RoI and aids the network to decide whether to attribute the detected anomaly to the RoI. This is especially useful when an object is not fully contained in the RoI but well captured in the cropped image. The output of the residual block in the *RoI-pipeline* of the final phase and the RCVs are fed to a fully connected block which outputs the probability that a concealed item is present in one or more of the cropped images in the sub-sequence. The final classification of the body part is the max probability aggregate of the sub-sequences.

Training and Inference with Ensemble Classifiers. Our 2-phase, dual-pipeline network has about $2.47M$ parameters ($5.19M$ flops). We train 3 classifiers on overlapping, equal-sized, subsets of the training set. This is done with a *3-folds* stratified learning scheme where each classifier is trained on 2 subsets and validated on the other subset. Each classifier is trained for 80 epochs with a batch size of 64 using Adam optimizer [18] and a dynamic learning rate starting at $1e{-}3$ and decreased to $5e{-}5$ between the 9^{th} and 72^{nd} epoch by a non-linear cosine function. During training, we re-sample sequences of cropped images with concealed items and augment all images by moving the cropped window about the RoI, zooming, horizontal flipping, and adjusting image contrast. At inference, the verdict is aggregated as the mean probability of the ensemble classifiers.

4 Experiments and Results

We evaluate the correctness of our 2D-pose refinement procedure in Tab. 1, the accuracy of our anomaly detection network in Tab. 2, and compare the performance of our state-of-the-art ensemble classifier to proprietary algorithms and other top detectors applied to TSA-PSD in Fig. 6.

4.1 Evaluation of 2D Pose Refinement for MWS Images

We evaluate our proposed keypoint correction process to show the relevance of our approach that adapts a RGB pretrained pose encoder to estimate more accurate poses on backscatter MWS images without supervision. Tab. 1 shows the Mean Per Joint Position Error (MPJPE) computed between predicted keypoint positions and manually labeled ground-truth positions. The final stage of our pose refinement (*Coherent-Pose*) decreases the error of estimated keypoints by 68%. We go on to show that this boost in accuracy is carried over to the anomaly detection network when trained with better segmented images. Note, however, that the consolidation of globally optimum coherent poses can sometimes come at the expense of local optima keypoint positions. We observe this consequence in the right and left hip and right knee keypoints where the coherent poses degrade the accuracy of the refined poses. This is because the pixel location of these keypoints are particularly volatile from frame to frame as the viewpoint of the person changes.

mm	R.Sh	R.Eb	R.Wr	L.Sh	L.Eb	L.Wr	R.Hp	R.Ke	R.Ak	L.Hp	L.Ke	L.Ak	*Avg.*
Generic Pose	115.7	191.5	119.1	100.4	173.4	117.6	88.49	115.3	137.3	79.91	105.8	134.5	123.2
Refined Pose	55.2	36.9	35.2	52.3	44.4	42.6	**65.1**	**38.5**	42.9	**61.6**	36.9	43.6	46.3
Coherent Pose	**40.2**	**32.5**	**19.9**	**38.1**	**31.4**	**23.6**	82.8	38.9	**22.4**	78.1	**35.8**	**21.0**	**38.7**

Table 1: Accuracy of 2D-poses derived by the naive keypoint selection (*Generic Pose*), our proposed method for guided keypoint selection (*Refined Pose* of Sec. 3.1) and correcting incoherent pose estimation (*Coherent Pose* of Sec. 3.1). Evaluated on the mean L2-norm between manually labelled keypoint locations and estimated keypoint locations of 50 scans (800 images). *R.Sh* refers to Right Shoulder, *L.Ke* Left Knee. *Avg.* is the mean over all keypoints.

4.2 Concealed Item Detection with RaadNet.

We conduct ablation experiments on our anomaly detection network trained on different types of segmented body part images and varying inputs. We present a comprehensive evaluation of our methods in comparison to published works on concealed item detection on the TSA dataset in Tab. 2. Our proposed method of using refined 2D keypoints to segment the human body parts consistently outperforms other published work on 2D concealed item detection in TSA-PSD in all metrics. Our RaadNet detector, trained on body part images segmented by coherent keypoints, RoI masks, and RCVs, performs at an average F1-Score of 98.6% on a disjoint validation-set and 0.0751 log-loss on the test-set. The

Body Part Anomaly Detection Methods	Validation						Test
	Avg.F1 ↑	*F1-Sc.* ↑	*Preci.* ↑	*Recall* ↑	*Acc.* ↑	*L.loss* ↓	*L.loss* ↓
FastNet (*) [23]	.8890	-	-	-	-	-	-
AlexNet-1 (*) [2]	-	-	-	-	-	**.0088**	-
AlexNet-2 (*) [13]	.9828	-	-	-	-	-	.0913
RaadNet +Fixed RoI Seg. in [2]+Mask+RCV (*Bsl-1*)	.9761	.9555	.9628	.9487	.9723	.0201	.1384
RaadNet +Unrefined-Pose Seg.+Mask+RCV (*Bsl-2*)	.9184	.7526	.8652	.6659	.9108	.1282	.1608
Ours RaadNet +Coherent-Pose Seg. (*Abl-1*)	.9775	.9581	.9655	.9505	.9766	.0143	.0934
Ours RaadNet +Coherent-Pose Seg.+Mask (*Abl-2*)	.9637	.9540	.9550	.9531	.9687	.0131	.0886
Ours RaadNet +Coherent-Pose Seg.+Mask+RCV (*Opt.*)	**.9859**	**.9738**	**.9941**	**.9544**	**.9946**	.0097	**.0751**

Table 2: Comparison of RaadNet and our proposed body segmentation to relevant published work on TSA-PSD. *Bsl-1* is our baseline network trained with fixed RoI segmented body parts with RoI masks and RCVs. *Bsl-2* is our baseline RaadNet trained with images segmented using original DeepHRNet keypoints (without refinement), masks, and RCVs. *Abl-1*, *Abl-2* and *Opt.* are our networks trained with body parts segmented using refined keypoints (Sec. 3.1). *Abl-1* is without RoI masks and RCVs, *Abl-2* is without RCVs, and *Opt* is with all three inputs; cropped images of body parts, RoI masks, and RCVs.

log-loss ϵ, on the test-set of TSA-PSD is defined in Eq. (2) between predicted threat probabilities $\hat{y}$ and the ground-truth binary label y of all $N{=}17 \times 1338$ body parts and scan subjects in the test-set. "ln" is natural logarithm.

$$\epsilon = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)] \tag{2}$$

By performing more precise body part segmentation on MWS images using refined 2D keypoints, we improved our network's ability to accurately detect concealed items by 53% (0.0751 test-set log-loss). Outperforming the published state-of-the-art method [13] (at 0.0913 test log-loss) by 17%. This is further extended to a 38% decrease in log-loss by our 3-ensemble classifiers. Note that the methods in the top 3 rows of Tab. 2 do not directly compare to our results because those works detect anomalies in a small subset of body parts (e.g. chest, thigh, arm, back). Whereas, our method detects concealed items on all body parts and the reported values reflect the cumulative performance on all body parts. To aid comparison with previous works, we show their best results reported for a single body part. Our ablation study highlights the importance of RCV and RoI mask. As expected, the use of masks improves the confidence of classification in *Abl-2* (decrease in log-loss). Although, at the expense of classification accuracy which is recovered by supplying RCVs. This is because RoI masks may exclude parts of objects on the boundary of body parts, whereas, RCVs inform the network how much of the objects are contained in the RoI. Hence, allowing the network to make a better decision of attributing detected concealed objects to body parts.

4.3 Comparison to TSA-PSD Proprietary Classifiers

The anomaly detection accuracy of RaadNet is further improved with 3-ensemble classifiers. Achieving up to 0.05592 mean log-loss on the test-set (7^{th} overall in Fig. 6). Note, mean log-loss is the only evaluation metric reported for the test-set

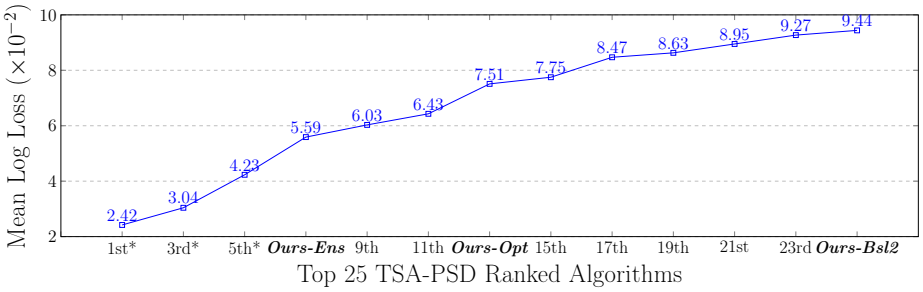

Fig. 6: Top-ranked TSA-PSD algorithms on the Kaggle Leaderboard. (*) indicates algorithms reported to have used 3D image files. Our ensemble RaadNet ranks 7^{th}, placing at topmost proprietary category. Making our method the only published, comprehensive work that places in the top-8.

because the ground-truth labels are private to TSA. This makes our proposed method the only comprehensive, fully-disclosed work that places at the top-8 proprietary category on the TSA Leaderboard. Details of the top-11 algorithms are proprietary and undisclosed to the public. Most of the top-8 methods are reported to use a combination 3D volumetric data (*330 MB* per file) and 2D (*10-41 MB*) image data (1^{st}). Whereas, we use only the smallest 2D image data available (*10 MB,* ***.aps*** *files*). RaadNet may be directly compared to the 6^{th} which use 2D data and multiple classifiers specialized for each body part. In contrast, we use only 3-ensemble classifiers, each component classifier trained on a disjoint subset of all body parts. We observed that all our baseline classifiers have a higher rate of false-negatives than false-positives. In other words, RaadNet is more likely to miss a concealed item than to generate false alarms. The difference narrows and false-negative rates decreases as more precise body part segmentation is used in *Ours-Ens* and *Ours-Opt*. Highlighting the importance of accurate body segmentation in backscatter MWS images.

5 Conclusion

We have shown how improved 2D human pose estimation, and the consequent improvement of body part segmentation can lead to a significant performance boost on body part anomaly detection task. Adapting 2D pose encoders trained on RGB images to estimate the keypoints of persons in backscatter MWS images is non-trivial without ground-truth annotations but, as we show, can be very rewarding when done well. Our keypoint refinement procedure and unsupervised body part segmentation algorithm described in Sec. 3.1 enables us to accurately segment the body parts of persons in MWS images. Subsequently, this allows us to train our anomaly detection network on cropped images of segmented body parts. With precise segmentation of body parts in backscatter MWS, we design a simple and effective body-part-aware neural network architecture that can be trained with weak supervision on all body parts in tandem.

Acknowledgements. Funding for this project was provided by the National Science Foundation, NSF grant BCS 2040422.

References

1. Ajami, M., Lang, B.: Using rgb-d sensors for the detection of unattended luggage. In: . (11 2016)
2. Bhattacharyya, A., Lind, C.H.: Threat detection in tsa scans using alexnet. In: . (2018)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. Computer Vision and Pattern Recognition(CVPR) (2017)
4. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7103–7112 (2017)
5. Cheng, G., Han, J., Guo, L., Liu, T.: Learning coarse-to-fine sparselets for efficient object detection and scene classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1173–1181 (2015)
6. CURTIS, A.R., POWELL, M.J.D., REID, J.K.: On the Estimation of Sparse Jacobian Matrices. IMA Journal of Applied Mathematics **13**(1), 117–119 (02 1974). https://doi.org/10.1093/imamat/13.1.117
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR09 (2009), dataset available at http://image-net.org/index
8. Fang, H., Lu, G., Fang, X., Xie, J., Tai, Y.W., Lu, C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 70–78 (2018)
9. Fioraio, N., di Stefano, L.: Joint detection, tracking and mapping by semantic bundle adjustment. 2013 IEEE Conference on Computer Vision and Pattern Recognition pp. 1538–1545 (2013)
10. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. ArXiv **abs/1808.00157** (2018)
11. Grisetti, G., Guadagnino, T., Aloise, I., Colosi, M., Corte, B.D., Schlegel, D.: Least squares optimization: from theory to practice. Robotics **9**, 51 (2020)
12. Gruosso, M., Capece, N., Erra, U.: Human segmentation in surveillance video with deep learning. Multimedia Tools and Applications **80** (01 2021). https://doi.org/10.1007/s11042-020-09425-0
13. Guimaraes, A.A.R., Tofighi, G.: Detecting zones and threat on 3d body for security in airports using deep machine learning. ArXiv: Computer Vision and Pattern Recognition **abs/1802.00565** (2018)
14. Güler, R.A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: Densereg: Fully convolutional dense shape regression in-the-wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2614–2623 (2017)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. Computer Vision and Pattern Recognition (CVPR) (2018)
16. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6981–6990 (2019)
17. Hynes, A., Czarnuch, S.: Human part segmentation in depth images with annotated part positions. Sensors **18**, 1900 (06 2018). https://doi.org/10.3390/s18061900
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)

19. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 3260–3271 (2022)
20. Lin, K., Wang, L., Luo, K., Chen, Y., Liu, Z., Sun, M.T.: Cross-domain complementary learning using pose for multi-person part segmentation. IEEE Transactions on Circuits and Systems for Video Technology **PP**, 1–1 (05 2020). https://doi.org/10.1109/TCSVT.2020.2995122
21. Lin, T.Y., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014), dataset available at http://cocodataset.org/#home
22. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: ECCV (2018)
23. Maqueda, I.G., de la Blanca, N.P., Molina, R., Katsaggelos, A.K.: Fast millimeter wave threat detection algorithm. 2015 23rd European Signal Processing Conference (EUSIPCO) pp. 599–603 (2015)
24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
25. Riffo, V., Mery, D.: Automated detection of threat objects using adapted implicit shape model. IEEE Transactions on Systems, Man, and Cybernetics: Systems **46**, 472–482 (2016)
26. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4510–4520 (2018)
27. Saviolo, A., Bonotto, M., Evangelista, D., Imperoli, M., Menegatti, E., Pretto, A.: Learning to segment human body parts with synthetically trained deep convolutional networks. In: IAS (2021)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. ArXiv **abs/1902.09212** (2019)
29. Thangavel, S.: Hidden object detection for classification of threat. In: . pp. 1–7 (01 2017). https://doi.org/10.1109/ICACCS.2017.8014719
30. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition pp. 1653–1660 (2014)
31. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment - a modern synthesis. In: Workshop on Vision Algorithms (1999)
32. TSA: Passenger screening challenge dataset (2017), https://kaggle.com/c/passenger-screening-algorithm-challenge/data
33. Xia, F., Wang, P., Chen, X., Yuille, A.L.: Joint multi-person pose estimation and semantic part segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6080–6089 (2017)
34. Yang, L., Song, Q., Wang, Z., Jiang, M.: Parsing r-cnn for instance-level human analysis. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 364–373 (2019)
35. Yaniv, Z.: Random sample consensus (ransac) algorithm , a generic implementation release. In: proceedings (2010)
36. Zhang, J., Hu, J.: Image segmentation based on 2d otsu method with histogram analysis. 2008 International Conference on Computer Science and Software Engineering **6**, 105–108 (2008)
37. Zhang, S.H., Li, R., Dong, X., Rosin, P.L., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.: Pose2seg: Detection free human instance segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 889–898 (2019)