Privacy-Preserving Pose Estimation for Human-Robot Interaction

Youya Xia\textsuperscript{1*}, Yifan Tang\textsuperscript{1*}, Yuhan Hu\textsuperscript{1} and Guy Hoffman\textsuperscript{1}

Abstract—Pose estimation is an important technique for nonverbal human-robot interaction. That said, the presence of a camera in a person’s space raises privacy concerns and could lead to distrust of the robot. In this paper, we propose a privacy-preserving camera-based pose estimation method. The proposed system consists of a user-controlled translucent filter that covers the camera and an image enhancement module designed to facilitate pose estimation from the filtered (shadow) images, while never capturing clear images of the user. We evaluate the system’s performance on a new filtered image dataset, considering the effects of distance from the camera, background clutter, and film thickness. Based on our findings, we conclude that our system can protect humans’ privacy while detecting humans’ pose information effectively.

I. INTRODUCTION

In this paper, we describe a method for privacy-preserving visual human-robot interaction (HRI). Our approach rests on effective human pose inference from a user-degraded camera. This would allow a user to cover the robot’s camera with a translucent film but still interact with the robot using nonverbal behaviors.

Visual observation of humans plays a vital role in understanding humans’ states and intentions for HRI \cite{12}. However, being monitored by cameras, especially at home, leads to privacy concerns \cite{3}. Prior work on protecting humans’ privacy in captured images focuses on using depth cameras or specially designed low-resolution cameras \cite{23}, \cite{32}. However, these methods require specialized equipment. In contrast, we propose to improve pose inference in images captured from a standard RGB camera which has been covered by a translucent film. The main contribution of this work is an image enhancement method that maintains privacy but enables robust pose estimation.

Pose estimation on clear images has been studied extensively \cite{14}, \cite{29}, but conducting pose estimation directly on degraded images is challenging. Researchers have proposed a number of image enhancement methods to facilitate pose estimation for degraded images \cite{17}, \cite{19}. However, these methods are generally based on a specific image formation model, such as low light or haze. We found that the privacy-preserving images in our case do not comply with existing models. Therefore, the existing models cannot solve the pose estimation problems for our purpose (see Section \ref{sec:related} for a detailed evaluation).

Our method is building on the recommendation put forth in Hu et al. \cite{15}, which proposed privacy-preserving HRI with cameras covered up by filtering materials. To detect humans’ poses from the degraded images, we propose a neural network based architecture for image enhancement. The proposed network is trained to produce an enhanced version of the input image. Afterwards, the enhanced image is passed into OpenPose \cite{4} to obtain the final estimated poses (Figure \ref{fig:1}). Experimental results indicate that our system can obtain human pose information effectively.

The paper thus makes the following contributions:

- A design for camera-based human-robot interaction, which can protect humans’ privacy while obtaining their pose information effectively;
- A neural network architecture designed for image enhancement specifically aimed at pose estimation from translucent-film-filtered images\cite{4};
- An extensive quantitative evaluation of the proposed system providing insights on the effects of distance, background clutter, and film thickness.

II. RELATED WORK

In this section, we review prior works on privacy-enhanced HRI and on image enhancement for pose estimation.

A. Privacy Protection in HRI

The camera used in many robots as an input to monitor a user’s state or intentions carries with it privacy concerns, especially in personal spaces such as the home. Researchers have proposed privacy-protecting strategies, such as blur filtration of all or part of the image \cite{5}, \cite{24}. As these methods aim to protect humans’ privacy through processing already-taken video data, they still allow robots to capture high-fidelity visual data.

Users often take control of their camera privacy by physically obstructing the lens of a camera placed in a private space. Building on this practice, Hu et al. propose a privacy-maintaining possibility for camera-based interaction with social robots \cite{15}. The authors suggest that through physically covering the robot’s cameras with a translucent

---

\*These authors contribute equally to the work.

\textsuperscript{1}Youya Xia, Yifan Tang, Yuhan Hu and Guy Hoffman are with Cornell University

\textsuperscript{1}The training and testing code is released at https://github.com/xiayax244/shadow_pose_estimation

---

Fig. 1: Left to right: the pose estimation results on the ground-truth clear image, the corresponding shadow image and enhanced image output by our enhancement module.
material, a robot can still exploit some interaction data in the
form of users’ shadows instead of high-fidelity images. The
paper, however, only described a concept sketch of this idea,
without proposing algorithms for inferring users’ poses from
full-body shadow images. Establishing an effective design for
such an algorithm is the aim of the current work, and at its
core lies an appropriate image enhancement module.

B. Image Enhancement for Pose Estimation

Image enhancement is an active field of computer vision,
with existing methods broadly falling into two categories:
single-image dehazing and low-light image enhancement.

a) Single-image Dehazing: Haze is a common type of
image degradation, leading to two kinds of work related to
single-image dehazing. The first line of work focuses on
estimating transmission maps and atmospheric light. It then
uses this atmospheric scattering model to remove haze. For
example, Li et al. (2017) provides an end-to-end network that
directly estimates the combination of atmospheric light and
transmission map [17]. The other line of work focuses on
learning a mapping from a hazy image to a haze-free image.
For instance, Li et al. (2020) proposes a convolutional neural
network (CNN) to learn a residual between hazy underwater
images and haze-free images [19]. But, both lines of work
do not have robust performances when images contain non-
uniform light or heavy haze [27], [28].

b) Low-Light Image Enhancement: Image generated
from low-light circumstances is another type of image degra-
dation. The early literature about low-light image enhance-
ment focuses on contrast enhancement [1], [5], [6], [25].
However, the early literature’s enhanced images suffer from
problems such as image distortion and artificial illumination.
State-of-the-art methodologies [10], [11], [22], [33] have
improved the naturalness of illumination and the realism
of images significantly. Nevertheless, they still encounter
failures of enhancement in regions with complex textures.

Both lines of state-of-the-art image enhancement method-
ologies focus on degraded images with explicit formation
models. Since there is no such formation model for im-
gages generated with a film-covered camera, implementing
the state-of-the-art methodologies in our method is not a
promising avenue. In Section V we explicitly evaluate the
use of existing image enhancement methods in comparison
with the method propose in this paper.

III. System

This section describes the full privacy-preserving pose
estimation pipeline, which includes the privacy-preserving
camera setup, the proposed network for image enhancement
module, and the pose estimation module.

A. Privacy-preserving camera setups

Based on the prior work of ShadowSense [15], in the pro-
posed system, the robots’ eyes are covered with a translucent
material which allows it to capture users’ full-body visual
data in the form of their shadows. This is a solution that can
easily be implemented by the end-user and leaves them in
full control of the image captured by the robot.

While adding privacy, conducting pose estimation directly
on the defined shadow images is challenging (see Section V
for detailed performance evaluation of pose estimation on
these images). To counter this issue, we present an image
enhancement module that can produce an enhanced version
of a shadow image tailored to human pose estimation.

![Image 2: The general structure of the MiniRes-based archi-
tecture. It is composed of an input layer with size 256 ×
256 × 3 and three Enhancement Modules (EM), each of which
has a respective optimization objective (i.e., structural loss,
perceptual loss and edge loss). The second and the third
modules are connected with the input layer via a short-cut
respectively.

Fig. 2: The general structure of the MiniRes-based archi-
tecture. It consists of 3 MiniRes blocks with a short-cut from the first
MiniRes block to the third MiniRes block.

![Image 3: The architecture of each Enhancement Module (EM).
It consists of 3 MiniRes blocks with a short-cut from the first
MiniRes block to the third MiniRes block.

B. MiniRes-based architecture for Image Enhancement

We propose a design for deep neural network (Figure 2),
composed of an input layer with size 256 × 256 × 3 and three
separate “Enhancement Modules” (EM). Each EM is
tuned to a specific optimization objective which is important
to pose estimation.

The first enhancement module aims to restore the shadow
image’s structural information so that the enhanced image
has higher structural similarity to the ground-truth clear
image. In the second enhancement module, the optimization
objective is to restore the shadow image’s perceptual features
so that the enhanced image has a smaller distance to the
ground-truth image in visual space. The final enhancement
module is supposed to restore the edge information of the
shadow image so that the enhanced image preserves the edge
information of the ground-truth image. The three EMs are
connected in sequential order, and the output of our network
is the output from the third enhancement module. Moreover,
to avoid classical degradation and vanishing gradient prob-
lems encountered by deep neural networks, we add a shortcut
from the input layer to the second and the third enhancement
modules, respectively.

Inside each of the enhancement modules (Figure 3), the
input is connected to a 3D Convolution layer whose dimen-
sion is 3 × 3 × 32 and a MaxPooling layer whose filter
size is 3 × 3. We then use three sequential copies of a
custom-designed component which is based on the ResBlock
module, and the pose estimation module.
structure in ResNet \cite{13}, an efficient network structure for various challenging vision tasks. We call this component MiniRes.

MiniRes blocks are different from original ResBlocks in two ways. First, to allow for the variance of shadow images created by different filters, we remove the two batch normalization layers from the original ResBlock architecture. Second, to speed up the inference process, we reduce the dimension of two 3D convolution layers from $3 \times 3 \times 64$ to $3 \times 3 \times 32$.

In addition, in each EM, a shortcut is added from the first to the third MiniRes Block to preserve information about the image’s content during optimization process. Finally, we connect the third MiniRes block to a 3D convolution layer with dimension $3 \times 3 \times 32$ to produce the output of size $256 \times 256 \times 3$ for each enhancement module.

### C. Loss Function Formulation

Each EM uses a module-specific loss function, with the total loss function defined as

$$L = L_{SL} + L_{PL} + L_{EL}$$

Here, $L_{SL}$ represents the structural loss, $L_{PL}$ represents the perceptual loss and $L_{EL}$ represents the edge loss, mapping to the three optimization objectives listed above.

- The **structural loss** $L_{SL}$ aims to restore the structural information of the shadow image’s content. To do so, we use the Structural Similarity Index Measure (SSIM) \cite{34}, which has shown success for repairing an image’s structural degradation \cite{20}, \cite{22}. For each pixel $x$, SSIM is calculated within a window size $11 \times 11$ around $x$ as follows:

$$SSIM(x) := \frac{(2\mu_x \mu_c(x) + d_1)(2\sigma_{xc}(x) + d_2)}{\mu_x^2(x) + \mu_c^2(x) + d_1(\sigma_x^2(x) + \sigma_c^2(x) + d_2)}$$

Here, $c$ and $e$ correspond to the network’s latent image and the paired ground-truth clear image. $\mu$, $\sigma^2$ correspond to expectation and variance of pixel values lying in each image window, and $\sigma_{xc}^2$ corresponds to the the pixels’ covariance between paired windows in $e$ and $c$. $d_1$ and $d_2$ are regulating constants which are set to 0.0001 and 0.0009 respectively. Given this definition, $L_{SL}$ is defined as:

$$L_{SL} := 1 - \frac{1}{3M} \sum_{i=1}^{M} (SSIM_r(e_i) + SSIM_g(e_i) + SSIM_b(e_i))$$

(3)

where $M$ is the number of pixels in the network’s latent image $e$ and $SSIM_h(e_i)$ means for each image channel $h \in \{r,g,b\}$, the SSIM value at pixel $i$ inside $e$.

- The **perceptual Loss** $L_{PL}$ ensures that the network’s output image is similar to the clear image in visual space.

$$L_{PL} := L_{MSE} + 2L_{MAE} + L_{ResNet}$$

We use a mixture of the Mean Absolute Error (MAE) (the $L_1$ norm) and the Mean Squared Error (MSE) (the $L_2$ norm), as the mixture has been proved to be a more robust metric for comparing global similarity in pixel space than using either norm alone \cite{9}. In addition, because ResNet has already been shown to be a robust feature extractor for high-level vision tasks, we use a feature extractor $\phi$ based on the top 6 layers of ResNet-50 \cite{13} pretrained on ImageNet \cite{7}. Then we compute the $L_2$ distance between the network’s latent image $e$ and the ground-truth image $c$ in the ResNet feature space:

$$L_{ResNet} := ||\phi(e) - \phi(c)||_2$$

(5)

- Finally, the **edge loss** $L_{EL}$ aims to restore the edge information of the ground-truth image. Specifically, we use an edge map extractor $\Omega$ which extracts the Sobel Edge maps of the network’s latent image $e$ and the corresponding ground-truth image $c$. Then, we computed the $L_2$ distance of those extracted maps as follows:

$$L_{EL} := ||\Omega(e) - \Omega(c)||_2$$

(6)

### D. Network Training

There exists no large-scale image dataset with film-filtered camera images. Based on prior literature of image enhancement \cite{17}, \cite{19}, training only on the limited number of shadow images that we would be able to collect does not produce robust performances. We therefore chose to train the proposed image enhancement module on the Outdoot Training Set (OTS) \cite{13}. The OTS is a large-scale synthetic dataset consisting of degraded images with different degrees of degradation. We train our proposed MiniRes-based network on the OTS using two NVIDIA™ RTX 2080 Ti cards.

### E. Pose Estimation Module

After the image enhancement module, the proposed system uses a pose estimation module to produce the final estimated poses in the enhanced image. Although various solutions for human body-pose detection have been proposed such as FastPose \cite{31} and other skeleton trackers \cite{16}, OpenPose is established as the state-of-the-art \cite{4}. In the last part of the proposed pipeline, we use OpenPose, as is, on the enhanced images as its inputs, to extract points lying inside humans’ bodies.

![Fig. 4: The figure shows the system when we put 1 layer of the plastic mask (i.e., one of our chosen filter layers) on top of the left camera.](image)

### IV. DATA COLLECTION

To test our image enhancement method, we collected a “shadow image” dataset using a binocular camera system (Figure 4), covering up one of the two cameras when
capturing an image. Using a binocular camera ensured that the ground-truth image and the paired shadow image include the same content. To reduce the impact of binocular disparity [26], we ensure that the distance between humans and the camera during the data collection process is at least 2m.

Fig. 5: The shadow images created by our chosen filter layers. Left to right: shadow images created by 1, 2 and 3 filter layers respectively.

Then, to create the shadow effect in one of the two cameras in the system, we used a freezer bag as the “base” filter. To evaluate the proposed system with different levels of privacy protection, we use 1, 2, & 3 layers of the freezer bag. Figure 5 shows the shadow effects created by selected layers. To test our proposed system’s robustness, we conducted the experiments under different background complexities (i.e., plain backgrounds, and complex backgrounds). Environments containing less than 4 objects are defined as plain backgrounds, denoted “private” spaces below. Similarly, we define environments containing more than 3 objects as complex backgrounds and we call them “public” spaces. In addition, since OpenPose uses a bottom-up approach for pose estimation, the farther distance between humans and cameras causes more challenges for the body part detection branch in its architecture. Aiming to understand the effectiveness of the MiniRes-based network for enhancing details (i.e., smaller body parts under farther distance) in the shadow image, we also conduct experiments under different proximity (i.e., close & far). Finally, to make sure there are no gender biases in our dataset, we collected images with both a female and a male actor.

We asked the two actors to conduct 3 activities (reading, exercising, and drinking) in both public and private spaces. The actors were also requested to repeat the activities with the same body part in the paired clear image and compare its location with the keypoint’s location associated with the same part in the paired clear image and calculate their L2 distance as L_{rec}. To count for the precision errors arising from binocular disparity which, in our dataset, is around 10 pixels, we regard k ∈ N_{te} if L_{rec} ≤ 10 pixels.

We will compare both metrics across filter layers, distance from camera, and background clutter.

A. Filter Layers: Effectiveness Evaluation of Privacy Protection

To provide a quantitative measure of the filter layers’ effectiveness for privacy protection, we use an established measure of image quality and estimate the subsequent degradation. Spatial–Spectral Entropy-based Quality (SSEQ) [21] is a comprehensive metric for measuring image quality based on images’ visual and structural information without using ground-truth images as references.

![Table I: The SSEQ scores for our chosen filter layers in both private and public spaces. We also include the Shadow Ratio scores in the table to demonstrate the each filter layers’ ability to protect humans’ identities.](image)

|       | SSEQ_{Clear} | SSEQ_{Shadow} | SR      |
|-------|--------------|---------------|---------|
| 1 Layer | 42.710       | 53.5125       | 0.2301  |
| 2 Layers| 42.7596      | 58.4768       | 0.3676  |
| 3 Layers| 39.0658      | 57.7562       | 0.4784  |

B. Image Enhancement Metrics

The proposed image enhancement module aims to facilitate the precision of pose estimation. So to evaluate the robustness of the trained MiniRes-based network, we propose two evaluation metrics based on the pose estimation results we obtain from the OpenPose module.

1. Detection: The first metric we propose is Detection Rate, DR := \frac{N_c}{N_e}. Here, N_{e} stands for the detected skeleton key points in the shadow images or enhanced images, while N_{c} stands for the detected key points in the paired ground-truth clear images. This metric only measures the number of detected features, without taking into account their accuracy.

2. Precision: Inspired by the mean Average Precision metric on pose estimation [2], we also use a second metric, the Shadow mean Average Precision, SmAP := \frac{N_{te}}{N_{e}}. Here, N_{te} stands for the number of the precisely detected key points in the shadow or enhanced images. Specifically, for each detected keypoint k in a shadow or enhanced image, we compare its location with the keypoint’s location associated with the same body part in the paired clear image and calculate their L2 distance as L_{rec}. To count for the precision errors arising from binocular disparity which, in our dataset, is around 10 pixels, we regard k ∈ N_{te} if L_{rec} ≤ 10 pixels.

V. EXPERIMENTS AND METRICS

This section conducts a quantitative evaluation of the proposed system, using the collected test shadow image dataset. We first evaluate the effectiveness of chosen filter layers, the robustness of the proposed image enhancement module quantitatively. Based on the proposed system’s performances, we then offer future researchers potential directions about how to choose suitable filters and proximity for setting up a privacy-preserving camera system for a social robot.
VI. Results

We evaluate the robustness of the trained network on the test dataset collected in public and private spaces using the proposed metrics (See our attached video for more experiments’ examples).

A. Private Space

We evaluate the effectiveness of our method for two different proximity settings and three layer settings. Figure 6 shows the mean DR and SmAP scores for each chosen filter layers on the private space dataset (low background clutter). Comparing scores for the original filtered (Shadow) and the enhanced images, we can see that the detection rate for shadow images degrades with increasing filter layers. SmAP (precision) is very low even with a single filter layer. In comparison, the enhanced images maintain high detection and accuracy scores even with increasing layers. Figure 7 shows an example image, comparing the pose estimation on the clear, filtered, and enhanced versions of the same image.

Besides, from the histograms in Figure 6, we find that proximity does not have a consistent effect on the shadow images’ pose estimation results. Both detection rates and precision are similar for people standing closer and farther from the camera.

B. Public Space

Figure 8 shows an example of an image taken in the public (cluttered) environment, demonstrating the pose estimation performance on the clear, filtered, and enhanced versions of the same image.

The results, seen in Figure 9, indicate a similar trend as in the private spaces. That said, un-enhanced shadow images degrade faster in public spaces, with loss of detection rates happening with 2 layers of filtering. In addition, compared with private spaces, for the same filter layers and proximity, the percentage of precisely detected key points on enhanced images is lower.
images in public spaces is smaller than the percentage in private spaces. Therefore, the public spaces are more challenging for image enhancement than private spaces.

C. Ablation Study

To evaluate each component’s importance in our defined loss function, we conduct an ablation study for our proposed network’s loss function. During the ablation study, to make more obvious distinctions of our network’s performances between original and modified loss functions, we choose datasets created from 2 filter layers and 3 filter layers. Specifically, for each type of spaces, we choose 3610 paired images generated from 3 filter layers and 3956 paired images generated from 2 filter layers. Then by changing the optimization objective in the enhancement module previously associated with \( L_{SL} \) to \( L_{PL} + L_{EL} \), we remove the structural loss from our optimization objective and retrain the network from scratch. Similarly, we train our network by removing the perceptual loss and edge loss respectively. In the ablation study table we list below, we use \( i_o \) to represent the original enhanced image output by the proposed network for \( i \) filter layers with \( i \in \{1, 2, 3\} \), \( i_g \), \( i_p \) and \( i_t \) to represent the enhanced image we obtain by removing one of the edge loss, perceptual loss, structural loss respectively.

| Layer/Method/Metric | 2\(_o\) | 2\(_m\) | 2\(_n\) | 3\(_o\) | 3\(_m\) | 3\(_n\) | 3\(_p\) |
|---------------------|------|------|------|------|------|------|------|
| DR                  | 0.959 | 0.605 | 0.730 | 0.954 | 0.531 | 0.582 |
| SmAP                | 0.901 | 0.069 | 0.049 | 0.887 | 0.044 | 0.062 |

TABLE II: The ablation study table we obtain by removing one of the perceptual loss, structural loss and edge loss.

We can infer two implications from the ablation study table (Table [II]). 1) The perceptual loss is more important than edge loss and structural loss for the performance of our proposed network since \( i_p \) suffers from the greatest decline in DR and SmAP scores compared with \( i_o \), with edge loss seemingly being the least contributing factor; 2) Since scores for the original enhanced images are higher than the enhanced images we obtain by removing structural loss, perceptual loss, and edge loss, respectively, the three components in our loss functions contribute to the network’s optimization process.

D. Comparison to Existing Image Enhancement Methods

There is no existing method proposed for boosting pose estimation’s performances on the type of shadow images we are interested in this work. That said, given the similarity of shadow images to hazy and low-light images in feature space, we evaluate the performances of two state-of-the-art methods focused on enhancing hazy and low-light images. Specifically, we choose AODNet [17], which focuses on hazy image enhancement, and MBLLEN [22], which focuses on low-light image enhancement for comparison study. Then we use the same paired images as the ablation study. Finally, we evaluate the pose estimation results (i.e., DR, and SmAP scores) of the enhanced images generated from the two state-of-the-art methods using OpenPose.

Table [III] shows the performance of our method (\( i_g \)) for \( i \) layers of filters, \( i \in \{1, 2, 3\} \). \( i_m \) indicated the output image that MBLLEN produces and \( i_o \) the output image that AODNet obtains. We can conclude that our proposed network boosts pose estimation on collected shadow images more effectively than AODNet and MBLLEN.

| Layer/Method/Metric | 2\(_o\) | 2\(_m\) | 2\(_n\) | 3\(_o\) | 3\(_m\) | 3\(_n\) | 3\(_p\) |
|---------------------|------|------|------|------|------|------|------|
| DR                  | 0.959 | 0.605 | 0.730 | 0.954 | 0.531 | 0.582 |
| SmAP                | 0.901 | 0.069 | 0.049 | 0.887 | 0.044 | 0.062 |

TABLE III: The evaluation table which shows the performances of AODNet and MBLLEN on the selected shadow image dataset.

VII. DISCUSSION

Based on our evaluation results, we offer future researchers potential directions about configuring the privacy-protected camera system.

- **Proximity**: From the four evaluation histograms in Figure [C] and [2] we conclude that the system can protect humans’ privacy and have robust performances on pose estimation when the distance between humans and camera system is between 2 and 4 meters. Therefore, we suggest that future designers can choose the proximity within range [2, 4]. Also, they are encouraged to explore a more suitable distance between humans and the system to balance the performances on privacy protection and pose estimation for their purposes.

- **Filters**: According to Table [I] and the four histograms, we can conclude that 1 filter layer is more optimal for pose estimation, while 3 filter layers are more optimal for privacy protection. 2 filter layers, on the other hand, has the medium score of shadow ratio and can be potentially chosen for balancing between privacy protection and pose estimation. Therefore, based on the SR value of Table [I] we propose that designers should choose filters with \( SR \in [0.36, 0.45] \) which is the SR value between the 2 and 3 filter layers.

VIII. CONCLUSION

This paper proposes a camera system that can protect humans’ privacy while obtaining their pose information effectively. It is based on an image enhancement neural network that uses structural, perceptual, and edge information and is trained on an off-context dataset. We evaluate the filter layers’ privacy protection ability and our proposed network’s robustness and find that our method regains the possibility for pose estimation even with a privacy filter. This can provide a useful solution for users wanting to cover a robot’s camera with a physical filter, but still interact with it. Finally, we offer future researchers potential directions about suitable filter choices and proximity between the camera system and humans. In future, we hope to incorporate social factors into our system design consideration by conducting human studies.
