

Privacy-Preserving Video Conferencing via Thermal-Generative Images

Sheng-Yang Chiu*, Yu-Ting Huang*, Chieh-Ting Lin*, Yu-Chee Tseng,
Jen-Jee Chen, Meng-Hsuan Tu, Bo-Chen Tung, and YuJou Nieh .

Abstract—Due to the COVID-19 epidemic, video conferencing has evolved as a new paradigm of communication and teamwork. However, private and personal information can be easily leaked through cameras during video conferencing. This includes leakage of a person’s appearance as well as the contents in the background. This paper proposes a novel way of using online low-resolution thermal images as conditions to guide the synthesis of RGB images, bringing a promising solution for real-time video conferencing when privacy leakage is a concern. SPADE-SR [1] (Spatially-Adaptive De-normalization with Self Resampling), a variant of SPADE, is adopted to incorporate the spatial property of a thermal heatmap and the non-thermal property of a normal, privacy-free pre-recorded RGB image provided in a form of latent code. We create a PAIR-LRT-Human (LRT = Low-Resolution Thermal) dataset to validate our claims. The result enables a convenient way of video conferencing where users no longer need to groom themselves and tidy up backgrounds for a short meeting. Additionally, it allows a user to switch to a different appearance and background during a conference.

Keywords: conditional GAN, confidentiality and privacy, image synthesis, sensor system, video conference

I. INTRODUCTION

Due to the COVID-19 pandemic, there are growing demands for people to work from home. Online or virtual conferencing has become a major way of communication for teamwork. This requires people to activate webcams and make themselves and their backgrounds visible to everyone else in a video conference room. For this purpose, users must be well-groomed for an extended period of time with proper backgrounds. To facilitate online conferencing, solutions such as Mesh [2], Horizon Workrooms [3], and Animaze by Facerig have been developed. While attractive, these applications require special equipment such as Microsoft HoloLens 2 [4] and Meta Quest 2 [5], which may take extensive time to set up. Professional video-based tools dedicated to online conferencing include Zoom, Webex, Meet, and Teams [6]. On the other hand, social tools such as Line, Messenger, and Discord also support conferencing capability.

Most of the aforementioned tools involve the use of built-in cameras, which may easily run into the risk of revealing users’ personal information and privacy. Such information may leak out from the users themselves, the backgrounds, or the contents in the background. A significant amount

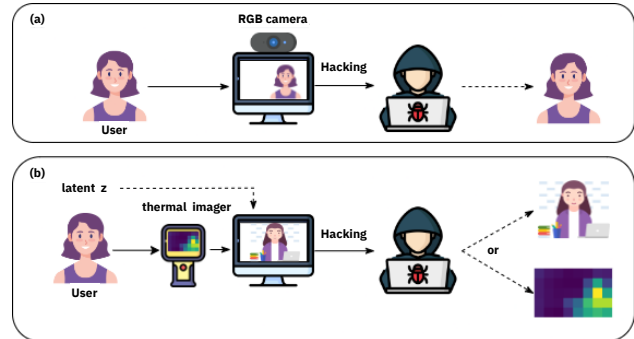


Fig. 1. Video conferencing scenarios and potential privacy leakage risks: (a) traditional and (b) thermal-generative.

of private information may be compromised either directly from the video contents or indirectly from a side channel by eavesdropping. For example, a device infected with Trojan horses may allow a hacker to capture images from a victim’s camera [7].

This work proposes an image generation system for video conferencing. Instead of transmitting real images, we transmit fake real-time images synthesized from (i) a pre-taken, privacy-free image scene provided by the user, which is converted to a form of latent code, and (ii) a sequence of ultra-low-resolution real-time thermal images taken from a thermal sensor. These low-resolution thermal images contain little private information and identifiable objects even if they are compromised. In addition, it is more convenient because users no longer need to dress up in a tidy home environment for short routine meetings. We compare the traditional and the proposed thermal-generative conferencing scenarios in Fig. 1.

In the literature, sensor-based applications have been intensively studied [8], [9], [10]. Reference [11] studied the coverage issue of sensors in a 3D environment. Networking issues of wireless sensors have been discussed in [12]. Recently, several works have applied deep learning techniques for multi-modality sensor data fusion. The work [13] studied person tracking by fusing video data from a drone and wearable sensor data from people on the ground. The work [14] addressed cooperative object identification from a drone camera and wireless sensors on the ground. A computer vision-assisted instant alert mechanism was proposed in [15]. A new boundary localization scheme through fusing multi-modality sensor data was presented in [16]. A novel SensePred framework for wearable sensor-assisted video prediction was proposed in [17]. These references motivate us

*Equal contribution.

All the authors are with College of Artificial Intelligence, National Yang Ming Chiao Tung University (NYCU), Taiwan; Yu-Chee Tseng also holds a joint appointment with College of Computer Science, NYCU, and with Miin Wu School of Computing, National Cheng Kung University, Taiwan.

PAIR-LRT-Human dataset is available on GitHub (<https://github.com/acqxi/PAIR-LRT-Human>)

to study fusing sensor and video data for privacy-preserving conferencing.

The rest of this paper is organized as follows. Section II introduces some related works. Section III presents the proposed privacy-preserving video conferencing system. The experiment results are in Section IV. We conclude our work in Section V.

II. RELATED WORKS

Generative adversarial networks [18] are considered one of the most creative frameworks to produce high-fidelity, high-quality pictures. One of them, DCGANs [19], which consists of convolutional layers, can generate images from a dataset. Conditional GANs [20] extend the field to text-to-image and image-to-image syntheses by controlling some specific labels. AC-GAN [21] requires only one label and it can generate high-resolution images from noise.

The image-to-image generation has made significant progress due to the works Pix2pix [22] and cycleGAN [23]. Pix2pix uses supervised learning with given paired images to generate images and cycleGAN can use a special training cycle to generate an image from different domains by unsupervised learning. Examples include the conversions between different domains, such as painting images to scenery photos [24], [25], [26], [27], facial expressions transfer by styleGAN [28], and full-body thermal images to color images [29].

With the advance of technologies, using thermal images to produce visible images has become a reality recently [30], [31]. There is considerable progress in thermal infrared image colorization, such as cGAN. Some works try to convert a street view thermal image to a realistic photo [29] and some transform facial thermal images into realistic human faces [32], [33], [34]. On the other hand, in order to generate more realistic images, [24], [35], [36], [37] present solutions to generate a synthesized image by a segmentation map. The state-of-the-art model in this issue is Spatial Adaptive Denormalization (SPADE) [24], which is based on GANs to shift segmentation maps to feature maps. Without relying on segmentation maps, SPADE-SR [1] is a variant of SPADE designed for transferring low-resolution upper-body thermal images to color images.

Recently, video conferencing has become a popular way of communication. Video conferencing also faces more and more security and privacy risks. The work [38] indicates that names, faces, and ages may be leaked through video conferencing. With the progress of face recognition [39], when facing a face recognition attack [40], one direct thought is to use GANs to generate appearances of the user with different clothes and expressions. However, the inputs to GANs are still actual images of the user. This motivates us to consider using a latent code of the user combined with a low-resolution thermal image to synthesize images for video conferencing. In this work, we shall adopt SPADE-SR [1] to achieve our goal.

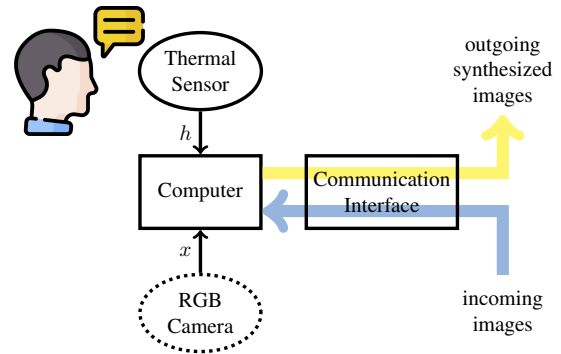


Fig. 2. Hardware Architecture.

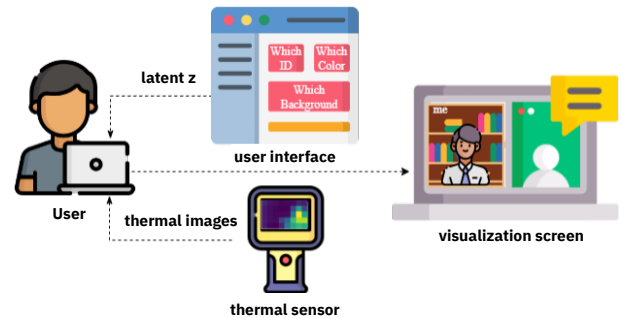


Fig. 3. Setting up a privacy-preserving video conference.

III. PRIVACY-PRESERVING VIDEO CONFERENCING

A. Hardware Architecture

Our hardware architecture is shown in Fig. 2. It only requires an edge computing client, such as a laptop, that is equipped with a low-resolution thermal imaging sensor. The RGB imaging sensor is optional as it is only required during the training stage to take a photo set representing the user(s). In the online conferencing stage, the user only needs to pick a latent code z from a latent code set Z that represents a pre-taken photo x in the training stage. The edge computing client simply takes z and the current thermal image as inputs and continuously produces synthesized images, which are transmitted to the conferencing partner(s).

B. Configuring a Video Conference

The process to configure a video conference is shown in Fig. 3. From the pre-taken photo set, the user can pick the appearance and background to his preference. The selection is mapped to a latent code $z \in Z$. Then the thermal sensor continuously streams thermal images to the laptop, which will convert them to fake images on the conferencing screen.

The deep learning model includes a training phase and an online phase, as shown in Fig. 4. In the training phase, the first step is to collect a dataset S that consists of a number of pairs (x, h) , where x is an RGB image and h is a thermal image that are taken simultaneously. Then we train our thermal-GAN model (refer to the next subsection),

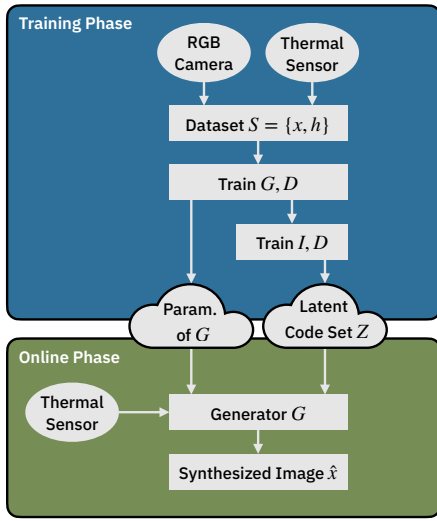


Fig. 4. Training and online phases.

which includes training G and D of GANs and training D and I of Inversion. The outputs include (i) the parameters of G and (ii) a latent code set $Z = \{\tilde{z} | \tilde{z} = I(x), (x, \cdot) \in S\}$. That is, Z is obtained by applying I on each x in the dataset S .

C. Thermal-GAN

Our aim is to produce a high-quality RGB image \hat{x} using a low-resolution heatmap h acquired from a thermal sensor. The synthesized image \hat{x} should be as close to the original image x as possible such that $\{x, h\} \in S$. To accomplish this, we will use GANs, specifically the SPADE-SR (SPADE with Self-Resampling) model [1], to build our video conferencing system. This model takes a thermal heatmap h and a latent code $z \sim P_{noise}(z)$ as inputs and produces an image $\hat{x} = f(z, h)$ by following the conditional GAN paradigm, or more precisely, the AC-GAN, where h serves as a condition. SPADE-SR follows AC-GAN since the thermal sensor data has a very low resolution, which can safeguard users' privacy. Moreover, sensor data is often quite noisy. The design not only recognizes the relationship between h and x , but also encourages high mutual information between the generated image \hat{x} and h [41].

Fig. 5 shows the architecture of SPADE-SR [1]. The dataset is denoted as $S = \{d_1, d_2, \dots, d_K\}$, where each $d_i = (x_i, h_i), i = 1 \dots K$, consists of an RGB image $x_i \in R^{H_x \times W_x \times 3}$ and a thermal heatmap $h_i \in R^{H_h \times W_h \times 1}$. G generates an RGB image $\hat{x} = G(z, h) \in R^{H_x \times W_x \times 3}$, where h is a heatmap and $z \sim \mathcal{N}(0, 1), z \in R^{256}$, is a noise variable representing non-thermal attributes independent of h . The discriminator D takes an image and outputs a reconstructed heatmap $\hat{h} \in R^{H_h \times W_h \times 1}$ and a realism score $s \in R^1$. I is an inversion encoder that converts a source image x to its matching non-thermal code \tilde{z} . The code $\tilde{z} = I(x)$ is then used in conjunction with other heatmaps to synthesize more RGB images.

Fig. 6 shows the architecture of G . It is made up of two

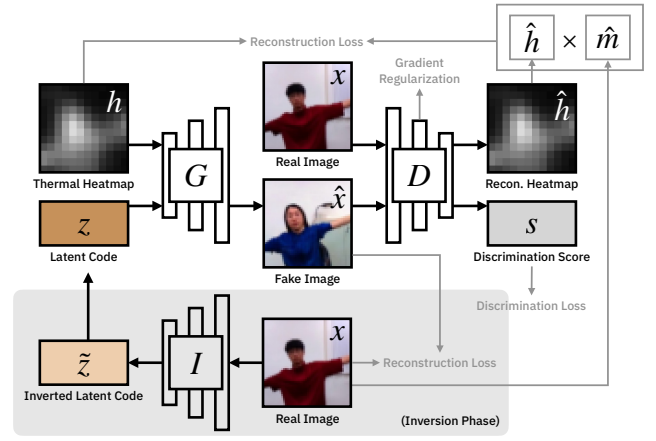


Fig. 5. The SPADE-SR model for privacy-preserving image synthesis.

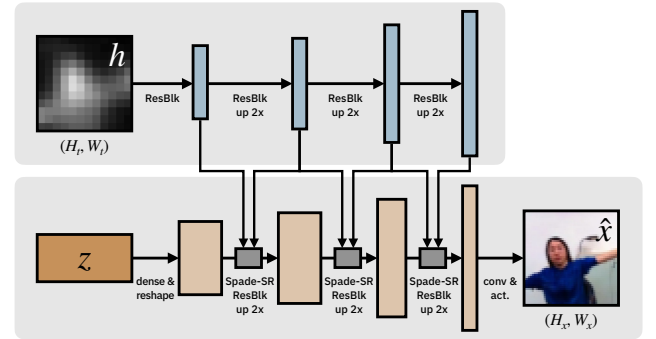


Fig. 6. The generator G and the resampling process to produce \hat{x} .

network branches. The main branch is composed of SPADE-SR ResBlock layers [1] that successively convert the noise vector z into an RGB image \hat{x} . To process the thermal heatmap h , the semantic branch continuously resamples itself to a larger feature map through ResBlocks. There are three phases of upsampling (three SPADE-SR ResBlocks), so the generated image (H_x, W_x) is eight times larger than (H_t, W_t) .

D. Loss Functions

We propose three loss functions to facilitate the training of our model. The first one pertains to the reconstruction loss between the original input feature h and its predicted counterpart \hat{h} . The second loss function relates to the discrimination loss of the discriminator D , while the third one corresponds to the reconstruction loss between the original RGB image x and its predicted version \hat{x} . The latter two loss functions are based on the definitions presented in SPADE-SR [1].

For the first reconstruction loss, a mask m is obtained from the RGB ground truth image x after performing person segmentation. The mask m has a size of (H_x, W_x) . We resize it to the size of the thermal image (H_t, W_t) , denoted as $\hat{m} = [\hat{m}_{i,j}]$. We then apply \hat{m} on h and \hat{h} and compare the results using the hinge L1 loss, which ensures that \hat{h} is similar to h while limiting the deviation no more than a small ϵ . This

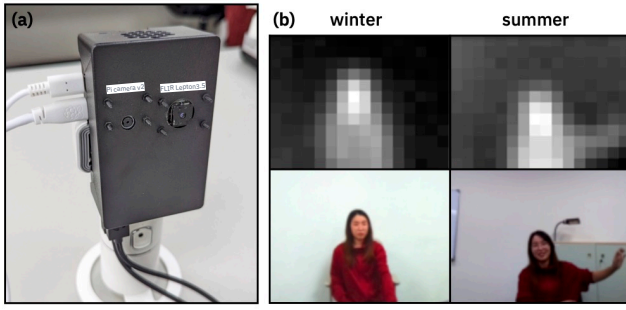


Fig. 7. (a) Hand-made data collection device based on Raspberry Pi 4B. (b) Recording environment and two data samples. There was a bigger temperature difference between the subject and the background during winter.

can be expressed as:

$$L_{\text{rec}} = \sum_{i,j} \mathbb{I}(\hat{m}_{i,j} = 1) \cdot \left[\max(0, |h_{i,j} - \hat{h}_{i,j}| - \epsilon) \right]. \quad (1)$$

Here, $h_{i,j}$ and $\hat{h}_{i,j}$ represent the pixel values at position (i, j) in h and \hat{h} , respectively. The function $\mathbb{I}(\cdot)$ is an indicator function that is equal to 1 when the condition in parentheses is true, and 0 otherwise. The mask enforces the model to focus only on the part with human thermal.

IV. EXPERIMENT RESULTS

A. Dataset

1) *Data Collection Process*: We have created a PAIR-LRT-Human (PAIR Lab Low-Resolution Thermal-Human images) dataset. The collection device is a hand-crafted Raspberry Pi with a FLIR Lepton3.5 thermal camera and a Pi camera v2, as shown in Fig. 7(a). Before collecting data, fisheye calibration on the camera is required, followed by pairing the camera with the thermal sensor to ensure that there is as much overlapping between their views as possible. Because the thermal sensor has a lower rate of 8.7 FPS, we regulate the sampling rate of the pi camera. That is, whenever a thermal image is recorded, we read out an RGB image.

The experiment was kept with as few irrelevant heat sources as possible. There was only one person moving into the area. The individual must face the device while wearing light clothes and doing a wide range of actions. The room temperature was under control; however, as can be seen from the figure, the background wall temperatures were still different in summer and winter. This dataset includes two persons, two scenarios, and three clothing colors, for a total of 12 different settings. Each environment has up to 11K thermal-RGB image pairs, for a total of approximately 33K pairs.

2) *Data Preprocessing*: We aim to align the two types of images. For each RGB image, we down-sample it to $(H=96, W=128, C=3)$ by pixel averaging and then normalize each value to $[-1, 1]$. For each thermal image, we blur it using a Gaussian filter with kernel size α [42], add random Gaussian noise [43], down-sample it via pixel averaging [44] to $(H=12, W=16, C=1)$, and normalize each value

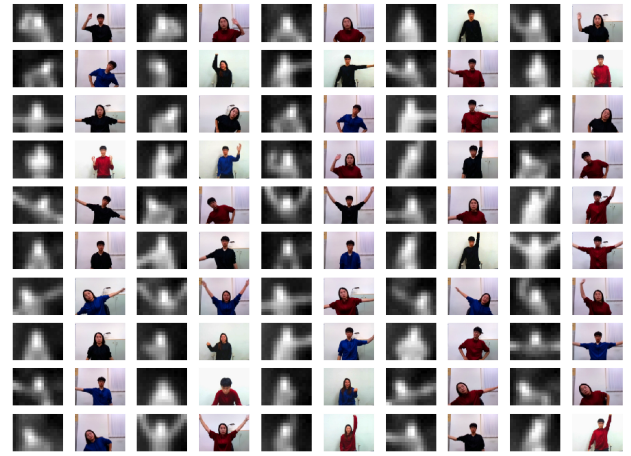


Fig. 8. Random samples of thermal-RGB pairs from the PAIR-LRT-Human dataset after data pre-processing.

TABLE I

COMPARISON OF PAIR-LRT-HUMAN AND LRT-HUMAN DATASETS.

| | PAIR-LRT-Human | LRT-Human |
|--------------------|----------------------------|------------------------------|
| thermal sensor | FLIR Lepton3.5 | Pana. AMG8833 |
| heatmap resolution | 16×12 | 8×5 |
| RGB camera | Pi camera v2 | iPhone XS |
| RGB resolution | 128×96 | 64×40 |
| field of view | $71^\circ \times 57^\circ$ | $60^\circ \times 37.5^\circ$ |
| No of data pairs | (33, 228) | (21, 959) |
| human occupant | 1 | none or 1 |
| standing position | random | random |
| pose | upper-body | upper-body |
| clothing colors | 3 | 5 |
| env conditions | 3 | 2 |
| No of persons | 2 | 1 |

to $[-1, 1]$. The down-sampling step is to ensure that the thermal images have a very low resolution that contains almost no meaningful information for human eyes. Our goal is to synthesize RGB images 8 times larger than thermal images. Fig. 8 contains some random samples in PAIR-LRT-Human with various human poses and positions under various clothing and lighting situations.

A comparison between PAIR-LRT-Human and LRT-Human created by [1] is shown in Table I.

B. Validations

1) *Model Training and Testing*: We adjust dimensions (H_x, W_x) of x and (H_h, W_h) of h to match the dataset specification. Our goal is to achieve $H_x = 8 \times H_h$ and $W_x = 8 \times W_h$. To train the GAN, the Adam [45] optimizer is used with $\beta_1 = 0$, $\beta_2 = 0.999$, and learning rates = 2×10^{-4} and 5×10^{-5} for D and G , respectively. During the inversion phase, we utilize Adam with $\beta_1 = 0$, $\beta_2 = 0.999$, and learning rates = 2×10^{-4} and 5×10^{-5} for the D and I , respectively. The batch size is set to 128, and the update ratios for D to G (GAN) and for D to I (inversion) are both

1. We set 400 epochs to train the GAN phase, and 50 epochs to train the inversion phase. We apply exponential moving averages on model parameters for evaluation using decay rates of 0.999 and 0.99 for the GAN and inversion phases, respectively. The total training time for GAN and inversion phases takes about 26.5 hours on four Nvidia V100 GPUs. The image generation rate reaches 11.7 fps with one single input and up to 389 fps with multiple inputs.

2) *Generation Quality*: We measure the RGB image synthesis quality by FID (Fréchet Inception Distance) [46]. FID calculates the distance between two feature maps extracted by the Inception network from two input images. It helps measure the similarity of two images when they are describing the same person. It is important to note that the feature maps used are retrieved from a specific layer of the Inception network, which are in the form of a 1D vector of length 2048. The FID score is defined as follows:

$$FID = \|\mu_1 - \mu_2\|_2^2 + Tr\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}\right), \quad (2)$$

where μ_1 and μ_2 are the mean vectors, and Σ_1 and Σ_2 are the covariance matrices of the feature maps of the original and the generated reference sets of images, respectively. Tr denotes the trace operator and $\|\cdot\|_2$ denotes the Euclidean norm. Since FID calculates the distance between two one-dimensional vectors, a smaller distance or a lower FID score indicates that the two vectors are closer, which means that the corresponding images are more similar. As the result, our method achieves an FID score of 9.6.

To visualize the synthesis quality, we show some examples in Fig. 9. We randomly select a portion of the dataset as the test set and use the heatmap of a test data pair and the corresponding person’s latent code as input. We compare the generated RGB image and heatmap with the ground truth. The results demonstrate that our method is capable of producing results that are visually similar to the ground truth.

3) *Avoiding Entanglement*: We synthesize RGB images from thermal and RGB image features. It is expected that these features are not entangled, which implies that each feature is only relevant to the real features that have been ascribed to it. To evaluate the disentanglement performance, we randomly pick k RGB images from the dataset and use the inversion encoder to get the features of RGB images; then we randomly sample n thermal images and use them to produce $k \times n$ images. We place these produced images in a $k \times n$ grid to see how one attribute changes graphically under the influence of the other attribute. As shown in Fig. 10, these two attributes have a clear and independent relationship - the heatmap controls the pose and position of the person and the latent code affects the clothing, background, and gender information.

4) *Privacy Protection*: With respect to the trade-offs between thermal image resolution and privacy protection, in Table II we evaluate privacy protection by using a fine-tuned YOLO v7 model [47] to detect people in a thermal image. Intuitively, if a person in a thermal image can be



Fig. 9. Some samples of ground truth and generated images. Green rows are random samples from PAIR-LRT-Human, and blue rows are synthesized RGB and thermal images.

TABLE II
PRIVACY PROTECTION DEGREE COMPARISON.

| Thermal Image Resolution | YOLO v7 Accuracy | Privacy Protection Degree |
|--------------------------|------------------|---------------------------|
| 160x120 | 80.7% | Low |
| 16x12 | 0% | High |
| 8x5 | 0% | High |

detected by the model, there is a higher risk that some details in the thermal image may reveal the person’s private information. Therefore, the higher accuracy in detecting people in thermal images means a lower degree of privacy protection. Conversely, a lower accuracy in detecting people indicates a higher degree of privacy protection. Our experiment results show that when using the fine-tuned YOLO v7 model on thermal images of a resolution of 160×120 , we achieve 80.7% accuracy. However, for thermal images with resolutions of 16×12 and 8×5 , the accuracy is 0%. This supports our claim that reducing the thermal image resolution can increase privacy protection.

5) *Demonstration*: A live video demonstration is provided on the Internet¹. Fig. 11 reports some failure cases.

6) *Discussion and Limitations*: Below, we discuss some limitations of our current work.

- Since we only use very low-resolution thermal images for image generation, it is still quite difficult to synthesize images with sufficient facial details. However, if higher-resolution thermal sensors are used, more personal information may be disclosed. This is a dilemma. Future work may consider using higher resolutions for facial expressions synthesis and lower resolutions for upper body synthesis.

¹A demo video is available in YouTube (<https://youtu.be/angJdjjt3Fs>)



Fig. 10. Disentanglement evaluation. The top row shows the source images and the leftmost column shows the heatmaps. Each cell represents a generated image given the corresponding non-thermal code and heatmap.

- While our system performs well, there are still some failure cases in our current results, which are shown in Fig.11. We shall undergo more data collection and data augmentation to improve our dataset. For example, data collection in different environments with various backgrounds and external thermal sources can be conducted.
- Currently, to run our video conferencing system, one has to pick a latent code z in the existing latent code set Z . These latent codes are pre-recorded. That is, one is not able to use an image not pre-trained by our system. To do so, the model needs to be retrained.
- Our system has been validated with a small set Z . That is, it is more suitable for use by a single person or a small working group. How to scale up its size deserves further study.

V. CONCLUSIONS

In this paper, we have addressed the privacy protection problem of video conferencing. To tackle this challenge, we propose an image generation system using thermal sensors based on the state-of-the-art method SPADE-SR. We also propose to improve the loss function of SPADE-SR by using a detected human mask on the heatmap to separate external thermal sources from human thermal sources. What this system actually uploads to the Internet are synthetic images, so even if the system is hacked, our system is more secure in terms of privacy protection. The privacy protection capability is validated via a human detection model that tries to identify people from a thermal image. We evaluate our model on the newly proposed PAIR-LRT-Human dataset and it demonstrates excellent image generation quality with an

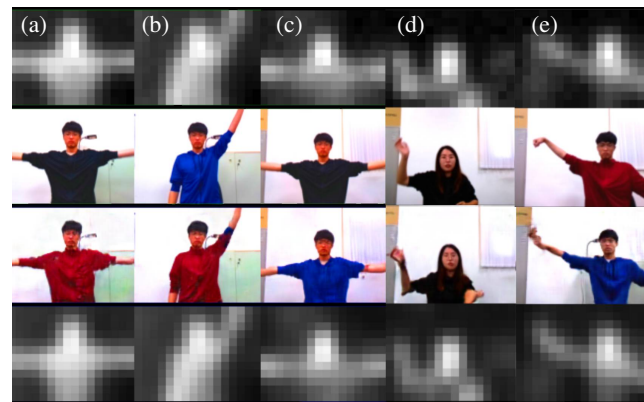


Fig. 11. Failure cases. (a)(b)(c) are speculated to be caused by a small number of images that fail to generate correct colors due to the inability of reconstruction loss in the inverse stage to accurately identify colors. (d) may be due to the fact that the model can only verify its own results by regenerating low-resolution thermal images, so it is not too important whether the arm shape is correct. (e) is a complex pose that may cause motion errors due to insufficient learning.

FID score of 9.6 and disentanglement properties. In terms of synthesis speed, our model reaches 11.7 fps with one single input, and up to 389 fps with multiple inputs.

REFERENCES

- [1] S.-Y. Chiu, Y.-C. Tseng, and J.-J. Chen, "Low-resolution thermal sensor-guided image synthesis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2023, pp. 60–69.
- [2] Microsoft mesh. [Online]. Available: www.microsoft.com/en-us/mesh
- [3] Horizon workrooms. [Online]. Available: <https://www.oculus.com/workrooms/>

- [4] Microsoft hololens 2. [Online]. Available: www.microsoft.com/en-us/hololens
- [5] Meta quest 2. [Online]. Available: <https://store.facebook.com/tw/en/quest/products/quest-2>
- [6] R. Singh and S. Awasthi, "Updated comparative analysis on video conferencing platforms-zoom, google meet, microsoft teams, webex teams and gotomeetings," *EasyChair Preprint*, vol. 4026, pp. 1–9, 2020.
- [7] T. Pultarova, "Webcam hack shows vulnerability of connected devices," *Engineering & Technology*, vol. 11, no. 11, pp. 10–10, 2016.
- [8] Y.-C. Wang, Y.-Y. Hsieh, and Y.-C. Tseng, "Multiresolution spatial and temporal coding in a wireless sensor network for long-term monitoring applications," *IEEE Transactions on Computers*, vol. 58, no. 6, pp. 827–838, 2009.
- [9] C.-H. Wu and Y.-C. Tseng, "Data compression by temporal and spatial correlations in a body-area sensor network: A case study in pilates motion recognition," *IEEE Transactions on Mobile Computing*, vol. 10, no. 10, pp. 1459–1472, 2011.
- [10] M.-S. Pan, Y.-A. Chen, T.-C. Chien, Y.-F. Lee, and Y.-C. Tseng, "Automatic lighting control system and method," Nov. 30 2010, uS Patent 7,843,353.
- [11] C.-F. Huang, Y.-C. Tseng, and L.-C. Lo, "The coverage problem in three-dimensional wireless sensor networks," *Journal of Interconnection Networks*, vol. 08, no. 03, pp. 209–227, 2007.
- [12] M.-S. Pan, H.-W. Fang, Y.-C. Liu, and Y.-C. Tseng, "Address assignment and routing schemes for zigbee-based long-thin wireless sensor networks," in *VTC Spring 2008 - IEEE Vehicular Technology Conference*, 2008, pp. 173–177.
- [13] A. Mukashev, L.-D. Van, S. Sharma, M. F. Tandia, and Y.-C. Tseng, "Person tracking by fusing posture data from uav video and wearable sensors," *IEEE Sensors Journal*, vol. 22, no. 24, pp. 24 150–24 160, 2022.
- [14] L. Van, L. Zhang, C. Chang, K. Tong, K. Wu, and Y. Tseng, "Things in the air: tagging wearable iot information on drone videos," *Discov. Internet Things*, vol. 1, no. 1, 2021. [Online]. Available: <https://doi.org/10.1007/s43926-021-00005-8>
- [15] Y.-Y. Tseng, P.-M. Hsu, J.-J. Chen, and Y.-C. Tseng, "Computer vision-assisted instant alerts in 5g," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020, pp. 1–9.
- [16] Y.-T. Liu, J.-J. Chen, Y.-C. Tseng, and F. Y. Li, "An auto-encoder multitask lstm model for boundary localization," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10 940–10 953, 2022.
- [17] J.-Y. Li, J. C.-H. Lin, K.-R. Wu, and Y.-C. Tseng, "Sensepred: Guiding video prediction by wearable sensors," *IEEE Internet of Things Journal*, pp. 1–1, 2022.
- [18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
- [21] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.
- [22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [24] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.
- [25] X. Huang, A. Mallya, T.-C. Wang, and M.-Y. Liu, "Multimodal conditional image synthesis with product-of-experts gans," *arXiv preprint arXiv:2112.05130*, 2021.
- [26] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
- [27] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Gp-gan: Towards realistic high-resolution image blending," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2487–2495.
- [28] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.
- [29] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 174–181.
- [30] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1143–1152.
- [31] N. Damer, F. Boutros, K. Mallat, F. Kirchbuchner, J.-L. Dugelay, and A. Kuijper, "Cascaded generation of high-quality color visible face images from thermal captures," *arXiv preprint arXiv:1910.09524*, 2019.
- [32] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1161–1165, 2018.
- [33] Y. Luo, D. Pi, Y. Pan, L. Xie, W. Yu, and Y. Liu, "Clawgan: Claw connection-based generative adversarial networks for facial image translation in thermal to rgb visible light," *Expert Systems with Applications*, vol. 191, p. 116269, 2022.
- [34] K. Mei, Y. Mei, and V. M. Patel, "Thermal to visible image synthesis under atmospheric turbulence," *arXiv preprint arXiv:2204.03057*, 2022.
- [35] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.
- [36] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," *arXiv preprint arXiv:2012.04781*, 2020.
- [37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.
- [38] D. Kagan, G. F. Alpert, and M. Fire, "Zooming into video conferencing privacy and security threats," *arXiv preprint arXiv:2007.01059*, 2020.
- [39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [40] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.
- [41] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
- [42] R. Deriche, "Recursively implementing the gaussian and its derivatives," Ph.D. dissertation, INRIA, 1993.
- [43] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *The Annals of mathematical statistics*, vol. 34, no. 1, pp. 152–177, 1963.
- [44] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *2010*

IEEE International Conference on Image Processing. IEEE, 2010, pp. 2805–2808.

- [45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [47] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” *arXiv preprint arXiv:2207.02696*, 2022.