Low-Resolution Thermal Sensor-Guided Image Synthesis

Sheng-Yang Chiu Yu-Chee Tseng Jen-Jee Chen
Pervasive AI Research Labs, College of Artificial Intelligence,
National Yang Ming Chiao Tung University, Taiwan
{syc.c, jenjee}@nycu.edu.tw, yctseng@cs.nycu.edu.tw

Abstract

Thermopile array sensors are cost-effective thermal imaging alternatives and are less vulnerable to privacy intrusion, light conditions, and obtrusiveness. While numerous occupant surveillance systems have been developed based on such sensors, low spatial resolutions prohibit them from deriving more sophisticated applications. To help relieve the limitation, we propose to enrich thermopile array sensors with additional non-thermal features and develop, to the best of our knowledge, the first low-resolution thermal-guided image synthesis model capable of producing realistic and attribute-aligned color images. These thermal heatmaps are regarded as semantic maps, but have very low resolutions. We propose an extension of SPADE (Spatially-Adaptive Denormalization), namely SPADE-SR, to incorporate the spatial property of a thermal heatmap into a conditional GAN through iterative Self-Resampling. Compared to SPADE, SPADE-SR yields better results in terms of image quality and reconstruction error while using significantly fewer model parameters. A new LRT-Human (Low-Resolution Thermal Human) dataset comprised of 22k (thermal heatmap, RGB image) pairs with various thermal and non-thermal coupling is derived to support our claims. Our work explores the cross-thermal-RGB modality paradigm and poses a great opportunity for thermopile array sensors in surveillance usages.

1. Introduction

Thermal imaging cameras capture emitted infrared radiation from heated objects. They have been used in various fields, ranging from cutting-edge space telescopes to everyday applications such as plumbing inspection, surveillance, or thermal screening. Due to the recent COVID-19 pandemic, many thermal-related applications have also been developed [6, 5]. The sheer ability to see temperatures makes thermal imaging an excellent way to detect human presence under various illumination conditions. Infrared solutions for common computer vision tasks such as image segmentation and depth estimation can be found [33, 37]. Some works jointly utilize RGB and thermal information for various tasks [35, 51].

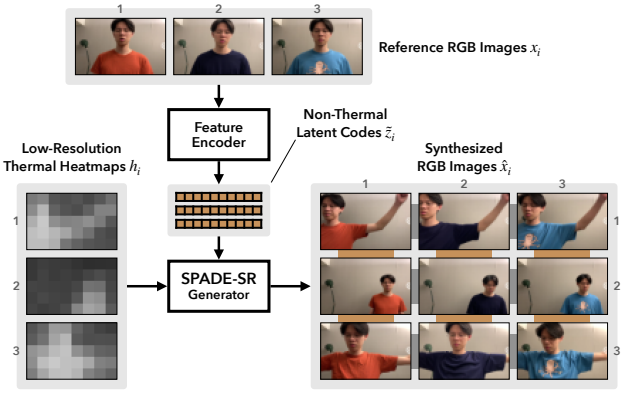

Figure 1. The proposed thermal-guided image synthesis model learns to generate realistic and attribute-aligned images based on low-resolution heatmaps and non-thermal codes. Our SPADE-SR (SPADE with *Self-Resampling*) generator is more effective at incorporating the unconventional semantic map into the network.

Despite thermal cameras' usefulness in various fields, their popularity is limited by cost, ranging from hundreds to hundreds of thousands of dollars. The active-cooled ones are photon detectors equipped with cryocoolers to detect infrared radiation, which can provide high image quality. The uncooled ones are thermal detectors that detect heat, which stand as cost-effective alternatives for everyday applications [12]. Recently, the infrared thermopile array sensor, an uncooled type of imager that has been actively integrated with MEMS, has gained more popularity [30, 31]. Numerous applications have been proposed, including fall/bed-exit alarm, convulsive movement detection, activity recognition, and posture classification. A thermopile array provides a cost-effective option compared to full-scale thermal cameras. While its lower spatial resolution naturally preserves the privacy of monitored occupants, the lack of sufficient details is a barrier to deriving more sophisticated applications, and the uncertainties from thermal data may lead to a model with over-fitting or ill-posed problems.

In this paper, we develop a deep-learning framework that expands the thermopile array sensor's capability by supplying extra RGB image features of the monitored occupant. The goal is to uplift a thermal heatmap to a natural, realistic RGB image with sufficient details fitting the heatmap. Instead of tackling an application problem such as human detection, we focus on modeling thermal and non-thermal uncertainties in a disentangled manner through a generative process to provide an image synthesis tool for further end usages. We propose a novel thermal-guided color image synthesis model by utilizing a low-resolution heatmap generated by a thermopile array sensor as a semantic map. By bridging thermal and color imaging information, our work explores this cross-modality paradigm and extends thermopile array sensors to more general surveillance usages. We depict our framework in Fig. 1.

We design a LRT-Human (Low-Resolution Thermal Human) dataset containing human-occupied scenarios with different lighting and clothing properties representing thermal and non-thermal variables. We capture the scene with a 64-pixel, 8×8 resolution Panasonic Grid-EYE infrared array sensor along with an RGB camera. These data are used to train a conditional generative adversarial network that synthesizes diverse and realistic color images based on given thermal information. In order to feed spatial conditions into the generator network, many works follow the content-style paradigm using Adaptive Instance Normalization (AdaIN) [21] or Spatial Adaptive Denormalization (SPADE) [44] to perform Semantic Image Synthesis. Although our task can be categorized as a spatial conditioning problem, it exists several differences. As opposed to the conventional content-style relation, our condition (low-resolution thermal heatmap) is pixelated and its structural relation to the output RGB image is quite coarse. Furthermore, the heatmap structure is limited to heated objects only. Our task is more similar to the Semantic Image Synthesis paradigm, where the output image should be conditioned spatially according to a semantic map. However, the thermal heatmap does not necessarily resemble the segmentation map because each thermal value is a continuous temperature, rather than discrete labels, and its resolution gap to the generated image is large.

Motivated by the above observations, we propose SPADE-SR, a *Self-Resampling* SPADE that is more effective in imposing spatial normalization at different resolutions throughout the network and in fixing the structural-misalignment problem caused by the resolution gaps between heatmaps and RGB images. We evaluate SPADE-SR on LRT-Human and our results demonstrate that SPADE-SR can synthesize high-quality images based on disentangled thermal and non-thermal attributes while outperforming SPADE using significantly fewer model parameters. Our work makes the following contributions:

- We propose a novel thermal-guided image synthesis model based on a low-resolution thermopile sensor that generates high-quality and attribute-aligned color images (FID 8.3).
- SPADE-SR improves over SPADE by incorporating an unconventional low-resolution thermal heatmap as a semantic map using a self-learned feature up-sampling branch.
- A new LRT-Human dataset containing 22k (thermal heatmap, RGB image) pairs with various thermal and non-thermal coupling is developed.

2. Related Works

2.1. Thermal Imaging

From space telescopes that search extragalactic stars to fever detection in the pandemic, thermal imaging plays an essential role in modern technological developments. A thermal image sensor is made up of an array of detectors that is capable of capturing infrared radiation. The Planck's law suggests that any object with a temperature above absolute zero would emit radiation in the thermal infrared spectrum (3 to 15 μm wavelengths). Actively-cooled thermal detectors pick up photons like regular cameras; such sensors provide high image quality, but need to work under very low temperatures so that they do not flood by their own radiation. Uncooled thermal detectors, typically found in everyday applications such as thermal screening or building inspection, work by measuring the change of resistance or voltage when the detector material is heated by infrared radiation and are more cost-effective [12].

Through thermal cameras, human bodies become easily visible against the environment. This makes an excellent application to Internet-of-Things. However, uncooled micro-bolometer detectors cost at least several hundred dollars. Recently, thermopile array sensors [30, 31] have gained more popularity due to their lower price. Integrating IoT with thermopile array sensors is an alternative solution to RGB cameras when cost, privacy, light condition, and obtrusiveness are concerns [46].

Thermopile array sensor-based monitoring tasks can be categorized into activity recognition, posture recognition, and localization. For activity recognition, [29] utilizes spatial-temporal information and proposes a CNN and LSTM based model to recognize five human activities. Also from a spatial-temporal viewpoint, [42] includes optical-flow features and proposes a CNN and GRU based model to recognize seven human activities. For posture recognition, a system to recognize 26 yoga postures through thermopile array sensors is proposed in [13]. LFIR2Pose [25] is a CNN-based key-point estimation model. For localization, sensors are usually placed on the ceiling. A CNN-based

classifier [3] is designed for human counting and localization via a Canny edge detector. [45] proposes a multi-human localization and tracking system.

Other applications include convulsive movement detection [17], contactless respiratory monitoring [49], hand gesture recognition [4], and fall detection [36]. A fuzzy logic-based dynamic background removal algorithm [15] is able to take out non-human heat sources from a heatmap.

2.2. Conditional GAN

Generative Adversarial Networks (GANs) [14] have demonstrated their ability to generate complex high-dimensional data such as high-fidelity images. Training a GAN involves a process of updating a generator G and discriminator D alternatively and needs careful hyper-parameter tuning. To improve training stability and generation quality, some works [1, 34] focus on loss function, and some on gradient regularization [16, 40, 38], training strategy [28, 20], or network architecture [50, 27].

Conditional GANs (cGANs) can generate samples concerning some premise information. The vanilla cGAN [39] provides an additional label to G through concatenation, and D is to recognize whether the data-label pair is valid or not. AC-GAN [43] also provides a class vector to G, but D does not see the data-label pair and has to predict the label from the data, while G is encouraged to generate data that maximizes the corresponding class probability. Other works such as Conditional Batch Normalization [11] for feeding class vector, and projection discriminator [41] for increasing generation diversity.

A single categorical label describes an image globally. Some information such as segmentation map, heatmap, and keypoint are in spatial forms, and they describe an image locally. Reshaping such information into a 1D vector would loss critical spatial properties. Pix2pix [24] and CycleGAN [53] tackle the problem from an image-to-image translation viewpoint. However, a deterministic mapping is assumed between input and output and it does not suit multi-modal problems. Another formulation is to factorize an image into a content-style form, where content influences spatial structures (e.g., edge and shape) and style controls spatial-irrelevant attributes (e.g., color, texture, and semantic) [9, 22, 54, 7], where Adaptive Instance Normalization (AdaIN) [21] is utilized to modulate style information. On the other hand, Semantic Image Synthesis focuses on generating semantically aligned images from a given segmentation map [44, 48, 55, 47]. The state-of-the-art models are variants of Spatial Adaptive Denormalization (SPADE) [44], which re-scales and shifts feature maps spatially according to the segmentation map.

GANs and cGANs have been successfully used for image synthesis in different applications. Compared to existing works, our work is unique in feeding in low-resolution thermal images as an input source in synthesizing natural RGB images, thus making cross-modality image synthesis possible. A recent study on cross-modality prediction of future videos from previous videos through wearable sensor data is proposed in [26].

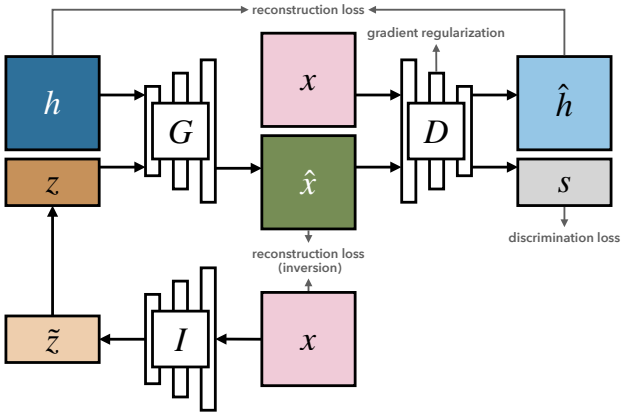

Figure 2. The overall image synthesis model.

3. Methodology

3.1. Overview

Consider a sensor rig accommodating a thermopile sensor and an RGB camera that synchronously collect data. Our goal is to generate a realistic RGB image x from a low-resolution heatmap h obtained from the thermal sensor. One might attempt to train a deep neural network that performs direct image translation $x = f(h)$ with a reconstruction loss to achieve this. While being straightforward, such a mapping is considered ill-posed because (1) most RGB-related information simply does not present in the thermal domain, and (2) the super-resolution process itself is ill-posed. It may simply output a pixel-wise average RGB image of all possible solutions from a given thermal heatmap, leading to blurry and unrealistic image quality.

This work approaches the problem from a generative viewpoint. We introduce a noise variable $z \sim P_{noise}(z)$ to the function $x = f(z, h)$ to model non-thermal uncertainties. We derive a conditional GAN (cGAN), specifically in the form of AC-GAN, where h plays as a condition. We select AC-GAN for our problem formulation because the thermopile sensor data is often noisy, and its low-resolution and progressive scanning nature might produce hard-to-notice misalignment between h and x. AC-GAN not only recognizes the ground-truth heatmap-RGB relation but also encourages high mutual information between the generated image $f(z, h)$ and h [8]. To incorporate h into the generator network, we propose SPADE-SR, a Spatial Adaptive Denormalization method with *Self-Resampling*. In contrast, the original SPADE resizes semantic map beforehand.

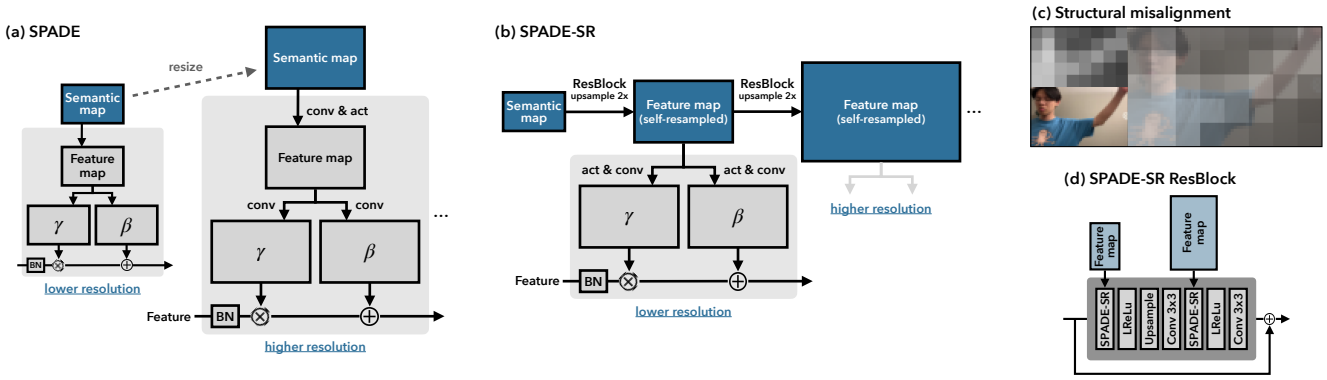

Figure 3. (a) SPADE. (b) SPADE-SR. (c) Illustration of structural misalignment due to resolution gap. (d) SPADE-SR ResBlock

The dataset is denoted as $S = \{d_1, d_2, \ldots, d_K\}$, where each $d_i = (x_i, h_i), i = 1 \ldots K$, consists of an RGB image $x_i \in \mathbb{R}^{H_x \times W_x \times 3}$ and a thermal heatmap $h_i \in \mathbb{R}^{H_h \times W_h \times 1}$. Our goal is to generate an RGB image $\hat{x} = G(z, h) \in \mathbb{R}^{H_x \times W_x \times 3}$, where h is a heatmap and $z \sim \mathcal{N}(0, 1), z \in \mathbb{R}^{128}$, is a noise variable representing non-thermal attributes independent of h. The discriminator D takes an image and outputs a reconstructed heatmap $\hat{h} \in \mathbb{R}^{H_h \times W_h \times 1}$ and a realness score $s \in \mathbb{R}^1$. Once the training of G and D is completed, we then train an inversion encoder I that converts a source image x to its corresponding non-thermal code $\tilde{z}$. We then use the code $\tilde{z} = I(x)$ in combination with different heatmaps to generate new RGB images based on that specific $\tilde{z}$. The overall diagram of our model is depicted in Fig. 2. More details are given below.

3.2. SPADE-SR

In SPADE, features go through multiple up-sampling stages. In each stage, these features go through Batch Normalization [23] first and then denormalized spatially using scaling (γ) and shifting (β) parameters according to a semantic map. The parameters are learned by passing the semantic map through two convolutional layers, including a shared one followed by two separated ones for γ and β, respectively. In order to impose spatial normalization at different resolutions throughout the network, the semantic map is resized (interpolated) to match required resolutions, as shown in Fig. 3a. However, for our low-resolution thermal heatmaps, up-sampling them drastically to higher resolutions might fail because images with large resolution gaps do not align well, as illustrated in Fig. 3c. With SPADE, the two convolutional layers are too shallow to handle such gaps effectively. While simply adding layers may solve the problem, the model size would significantly increase.

In contrast, our SPADE-SR employs a dedicated semantic network branch that performs a learned resizing (self-resampling) process that effectively transforms the semantic map into rich, multi-resolution features based on its own needs before inferring the parameters γ and β, as shown in Fig. 3b. It is interesting to note that removing Batch Normalization in the layers of semantic branch yields a huge difference and is crucial to the results. This observation is in line with the viewpoint of SPADE, which states that normalization would "wash out" the conditional information.

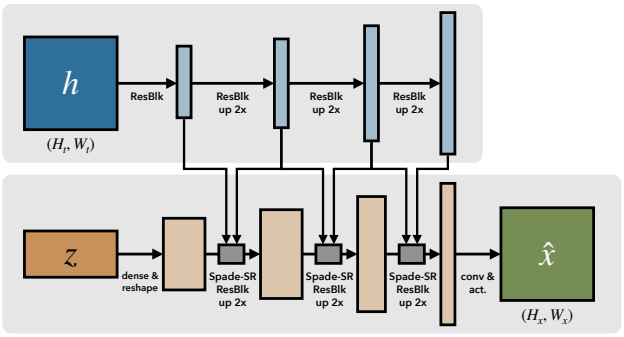

Figure 4. Structure of G.

3.3. Models

Generator In Fig. 3d, we show the proposed SPADE-SR ResBlock, which is the building block of our generator network G, as depicted in Fig. 4. SPADE-SR Resblock is modified from the Residual Block (ResBlock) [19] by replacing the Batch Normalization with SPADE-SR. The G consists of two network branches. The main branch is built upon layers of SPADE-SR ResBlock, iteratively transforming the noise vector z into an RGB image $\hat{x}$. The semantic branch is built with standard ResBlocks without Batch Normalization to process the thermal heatmap h. The generated image (H_x, W_x) is eight times the heatmap (H_t, W_t). So there are three up-sampling stages (three SPADE-SR ResBlocks). The layer details of G are shown in Table 1.

Discriminator The discriminator D is to differentiate between a real image x and a fake image $\hat{x}$ by returning a realness score s. Aside from this task, there is an auxiliary task for D to infer the heatmap h, either from x or

Semantic Branch	Main Branch
$h \in \mathbb{R}^{H_h \times W_h \times 1}$	$z \in \mathbb{R}^{128}$
ResBlk $1 \rightarrow ch$	Linear $128 \rightarrow H_h \times W_h \times 512$
ResBlk up $ch \rightarrow ch$	SPADE-SR Blk up $512 \rightarrow 512$
ResBlk up $ch \rightarrow ch$	SPADE-SR Blk up $512 \rightarrow 256$
ResBlk up $ch \rightarrow ch$	SPADE-SR Blk up $256 \rightarrow 128$
-	BN, LReLU, 3×3 Conv $128 \rightarrow 3$
-	Tanh

Table 1. Layer specification of G.

$\hat{x}$. This auxiliary task represents the conditional distribution $Q(h|x)$. D has two output heads, with one dense layer for s and another ResBlock and convolution layer for $\hat{h}$.

3.4. Loss Functions

GAN Phase G and D are trained in an adversarial manner, where G tries to fool D by producing realistic images and D tries to tell real from fake images apart. In this work, we adopt the hinge loss [34] as the realness rating for D. For G and D, their loss to minimize are:

$$\begin{aligned} L_{adv}^{G} &= -\frac{1}{n}\sum_{i=1}^{n} s_i^{fake} \\ L_{adv}^{D} &= \frac{1}{n}\sum_{i=1}^{n} relu(1 + s_i^{fake}) + \frac{1}{n}\sum_{i=1}^{n} relu(1 - s_i^{real}), \end{aligned} \tag{1}$$

where s^{fake} and s^{real} are D's realness rating for $\hat{x}$ and x, respectively. To impart training stability, we apply R1-regularization [38]:

$$R_1 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla D(x_i)\|^2. \tag{2}$$

We train D and G to minimize a heatmap reconstruction loss, encouraging $G(z, h_i)$ not to lose the information of h_i. We adopt a loss threshold, similar to the hinge loss, to cope with noisy thermal data:

$$\begin{aligned} L_{mi}^{DG} = \frac{1}{2n}\sum_{i=1}^{n} &relu(|h_i - \hat{h}_i^{fake}| - 0.05) + \\ &relu(|h_i - \hat{h}_i^{real}| - 0.05), \end{aligned} \tag{3}$$

where $\hat{h}^{fake}$ and $\hat{h}^{real}$ are reconstructed thermal heatmaps of $\hat{x}$ and x inferred by D, respectively. The total loss of the D and G are defined as:

$$\begin{aligned} L_{total}^{D} &= L_{adv}^{D} + \gamma L_{mi}^{DG} + \beta R_1 \\ L_{total}^{G} &= L_{adv}^{G} + \gamma L_{mi}^{DG}. \end{aligned} \tag{4}$$

We set $\beta = 5.0$ and $\gamma = 2.0$ in our implementation.

Inversion Phase After the GAN training phase, we then train the inversion encoder I. The first objective of I is the L1 pixel-wise error. To further ensure that $\tilde{z}$ produces not only similar but plausible images, an adversarial loss L_{adv}^{I} is added by utilizing the pretrained D into the inversion phase [52]:

$$\begin{aligned} L_{rec}^{I} &= \frac{1}{n}\sum_{i=1}^{n} |x_i - G(\tilde{z}_i, h_i)| \\ L_{adv}^{I} &= \frac{1}{n}\sum_{i=1}^{n} s_i^{rec}, \end{aligned} \tag{5}$$

where s_i^{rec} is D's realness rating for the inverted image $G(\tilde{z}_i, h_i)$. We set $\lambda = 0.1$ in our implementation. The total loss of I is defined as:

$$L_{total}^{I} = L_{rec}^{I} - \lambda L_{adv}^{I}. \tag{6}$$

4. Experiments

4.1. Dataset

We developed the LRT-Human (Low-Resolution Thermal-Human) dataset, which was collected on a simple sensor rig that accommodates a thermopile sensor, an RGB camera, and an MCU. Note that there exist other thermal-human datasets [10, 2], which, when downsampled, have a similar form to ours. However, LRT-Human focuses on modeling thermal and non-thermal attributes with less challenging image details. The dataset consists of 22K (low-resolution thermal heatmap, RGB image) pairs with a human occupant under various scene configurations in a way that thermal and non-thermal attributes can be verified explicitly. To match the thermopile sensor and RGB camera spatially, we calibrate them by finding a translation and scaling parameter that maximizes the overlapping area between heatmaps and human masks of RGB images acquired by the publicly available pretrained Mask R-CNN [18]. Fig. 5 displays some random samples with different human poses and positions under several clothing and lighting conditions. Table 2 contains more dataset specifications.

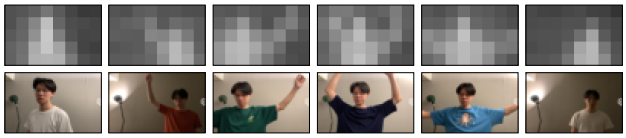

Figure 5. Random sample pairs from the LRT-Human dataset.

4.2. Evaluation

We set dimensions H_x, W_x, H_h and W_h to match the dataset specification. Thermal heatmaps between 19 to 30

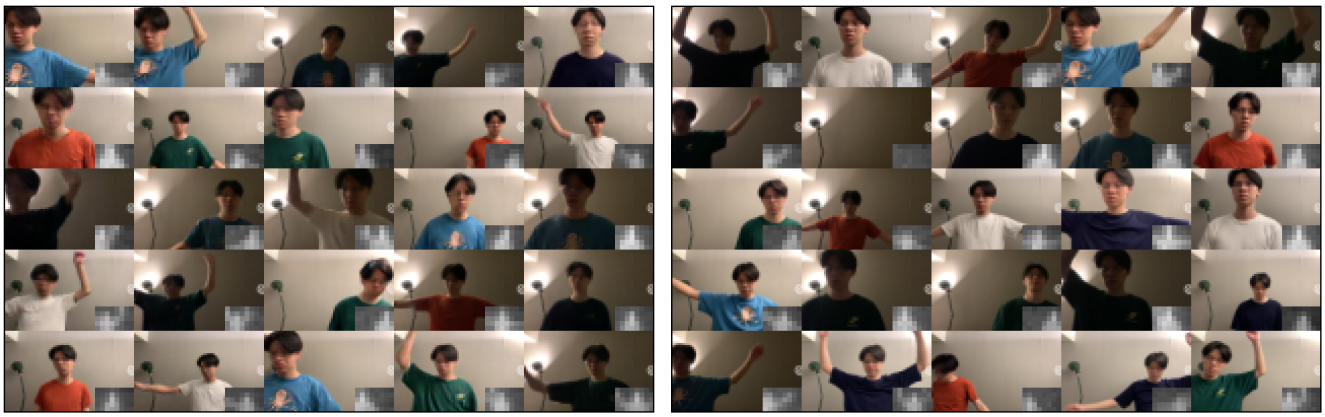

Figure 6. Left: generated image samples. Right: real image samples.

Name	Value
Dataset Parameters	
Thermopile array sensor	Panasonic AMG8833
Thermal heatmap resolution	8×5 $(w \times h)$
RGB camera	Apple iPhone XS
RGB image resolution	64×40 $(w \times h)$
Field of view	$60^\circ \times 37.5^\circ$ $(w \times h)$
# of data pairs	$21,959$
Thermal Attributes	
# of human occupant	$\{0, 1\}$
Occupant position	Random positions
Occupant pose	Random poses
Non-Thermal Attributes	
# of clothing colors	5
# of lighting conditions	2

Table 2. Specifications of the LRT-Human dataset.

degree Celsius are re-scaled linearly to $[-1 : 1]$. We use Adam [32] optimizer with $\beta_1 = 0$, $\beta_2 = 0.999$, and learning rates $= 2 \times 10^{-4}$ and 5×10^{-5} for D and G, respectively. For the inversion phase, we use Adam with $\beta_1 = 0$, $\beta_2 = 0.999$, and learning rates $= 2 \times 10^{-4}$ and 5×10^{-5} for the D and I, respectively. The batchsize is set to 256, and the update ratio for D to G (GAN) and for D to I (inversion) are both 1:1. We train the GAN phase with 400 epochs and inversion phase with 10 epochs. We apply exponential moving average 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 13 hours on a single Nvidia V100.

Generation Quality The Fréchet inception distance (FID) [20] is evaluated for image generation quality by comparing the feature distributions between real and generated images, with 0 as the lowest score (best image generation).

$$\text{FID} = ||\mu - \mu_w||_2^2 + \text{tr}(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2}), \quad (7)$$

where μ and μ_w are the means and Σ and Σ_w are the covariance matrices of the features for real and generated images. The features are a 2048-dimension vectors taken out from the intermediate layer of the Inception network. A comparison of generated and real image samples is shown in Fig. 6. We reach an FID score of 8.32, and we note that lower scores might be achieved by further training.

Disentanglement This work aims to generate RGB images from independent thermal and non-thermal sources. These two sources should not be entangled, meaning that each source is only responsible for its designated purpose. To evaluate the disentanglement performance, we randomly sample k RGB images from the dataset and use I to retrieve their $\tilde{z}$. We then randomly sample n thermal heatmaps and use them to generate $k \times n$ images. We observe the visual changes between these two factors in Fig. 7. These two attributes pose a clear and independent relation, where heatmaps control occupants' poses and non-thermal attributes $\tilde{z}$ control clothing and lighting styles, proving the disentanglement property.

Reconstruction We further evaluate the robustness of G by conducting a reconstruction error measurement. In the LRT-Human dataset, samples in the same clip have the same non-thermal attributes (i.e., same clothing and lighting). Consider any clip. First, we randomly sample an RGB image from the clip and get its non-thermal code $\tilde{z}$ from I. Second, we use this $\tilde{z}$ combined with all heatmaps in the clip and generate an entire fake clip. Last, we calculate the mean absolute error between the two clips. This helps evaluate the pose/position correctness of the generated images and tests the consistency of non-thermal attributes across the whole fake clip since only a single $\tilde{z}$ from a random image in the clip is used. SPADE-SR scores an average of

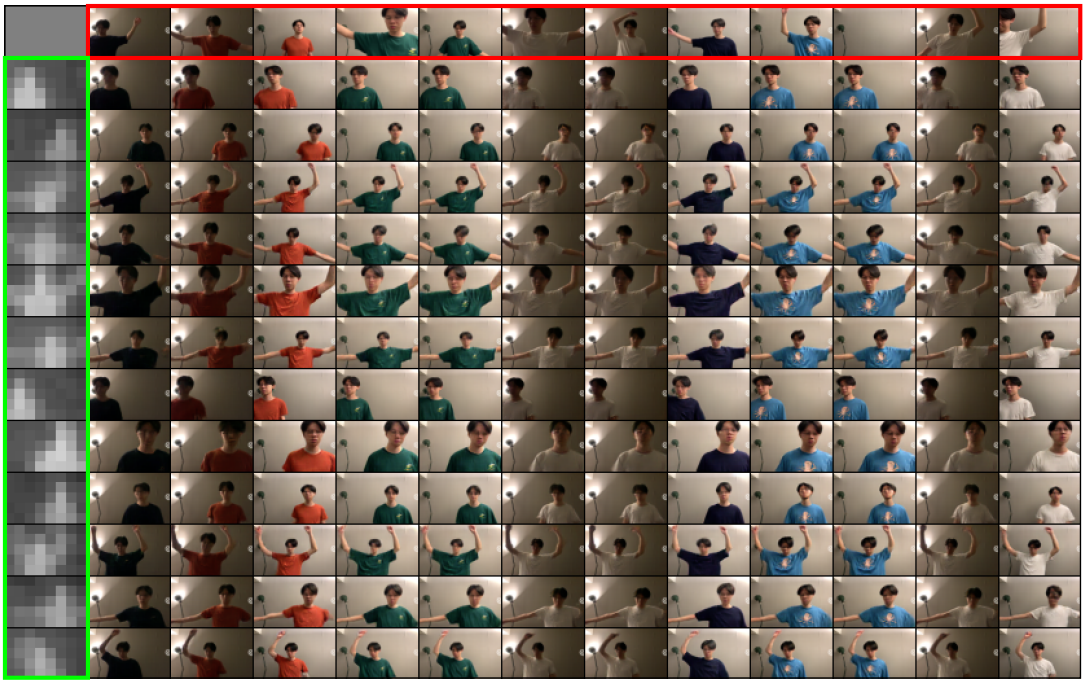

Figure 7. Disentanglement evaluation. The top row (red) shows the source images from which non-thermal codes are retrieved from I, and the left-most column (green) shows the heatmaps. Each cell represents a generated image given the corresponding non-thermal code and heatmap. The model performs well-each column has similar clothing and lighting style while each row has similar human pose and position. Notice the source image with no occupant (3rd from right) still results in a valid inverted latent code that generates plausible images.

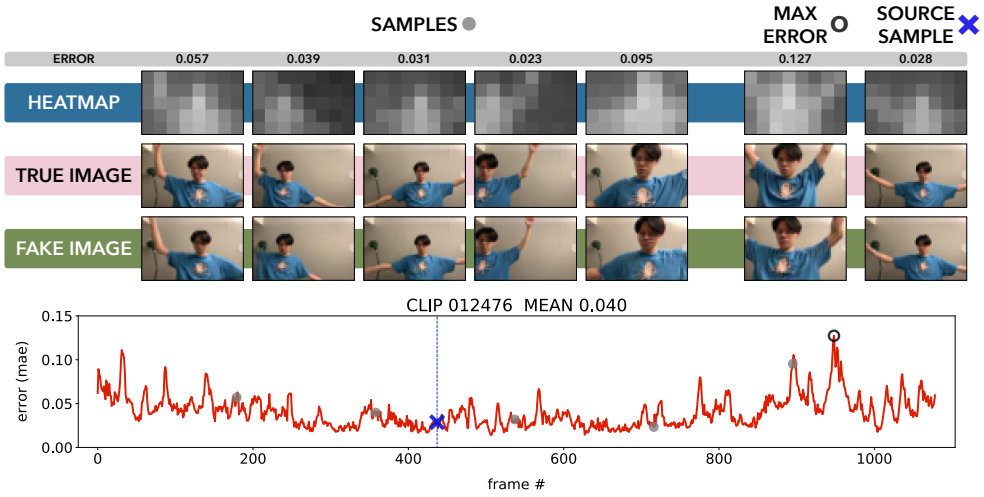

Figure 8. Visualization of image samples and frame-wise reconstruction error across a clip. Notice that the error at the source frame is not necessarily the lowest, suggesting that the non-thermal latent code $\tilde{z}$ is generalized, not tailor-learned.

0.0272 in mean absolute error across all clips in ten runs. We select some clips and plot the frame-wise reconstruction errors along with image samples for visualization in Fig. 8. We have three discoveries. First, SPADE-SR performs well on replicating an entire clip, generating similar poses/positions and non-thermal attributes at every frame. Second, most large errors in a clip are not caused by an incorrect $\tilde{z}$, but by image imperfections or slight postural differences due to the low-resolution nature of heatmaps or minor RGB-heatmap miss-alignment in the dataset. Third, there is no trend that the errors of the source frames are the lowest, suggesting that SPADE-SR learns a generalized la-

model	FID ↓	recon. ↓	param. ↓
SPADE-64	11.68	0.0334	12.6M
SPADE-128	9.62	0.0278	15.3M
SPADE-SR-32	9.51	0.0275	**11.3M**
SPADE-SR-128	**8.03**	**0.0266**	16.3M
SPADE-SR-64	8.32	0.0272	12.9M

Table 3. Comparison on models.

tent code $\tilde{z}$, instead of memorizing a tailored code at that specific frame.

4.3. Comparison

Table 3 compares SPADE and SPADE-SR in various model sizes on the LRT-Human dataset. SPADE-SR outperforms SPADE in both FID score and reconstruction error while using fewer parameters, even at one-fourth of channel size (32 v.s. 128). After multiple trials, we observe that SPADE with 64 channels would always collapse, generating samples with missing attributes (e.g., clothing color). We did not further test smaller channel sizes such as 32 because we suspect that it has reached SPADE's limit. We select SPADE-SR with 64 channels as our standard model because it strikes a balance between performance and model size.

5. Conclusions

This work presents, to the best of our knowledge, the first thermal-guided image synthesis model based on low-resolution heatmaps that is able to generate high-quality and attribute-aligned images. With the Self-Resampling tweak, we propose SPADE-SR that improves over SPADE by performing a learned resizing process instead of interpolation. We evaluate our model on the proposed LRT-Human dataset and SPADE-SR demonstrates outstanding generation quality, disentangling property, and reconstruction error. SPADE-SR also outperforms SPADE in model size. On the application side, we will develop privacy-preserving surveillance based on SPADE-SR as our future work. An obvious limitation of our current work is that we only consider one occupant in the scene. In order to improve its generality, it deserves to consider multiple occupants. Future directions also include assessing the influence of heatmap resolutions under different task difficulties, utilizing spatial-temporal features, learning intermediate information such as pose and keypoint, learning to handle unseen image samples, and further experimenting SPADE-SR on large-scale datasets.

Acknowledgements

This research is co-sponsored by ITRI, Pervasive Artificial Intelligence Research (PAIR) Labs, and National Science and Technology Council (NSTC). This work is also financially supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University (NYCU) and Ministry of Education (MOE), Taiwan.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Guillaume-Alexandre Bilodeau, Atousa Torabi, Pierre-Luc St-Charles, and Dorra Riahi. Thermal–visible registration of human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64:79–86, 2014.

[3] Mondher Bouazizi, Chen Ye, and Tomoaki Ohtsuki. Low-resolution infrared array sensor for counting and localizing people indoors: When low end technology meets cutting edge deep learning techniques. *Information*, 13(3):132, 2022.

[4] Daniel S Breland, Simen B Skriubakken, Aveen Dayal, Ajit Jha, Phaneendra K Yalavarthy, and Linga Reddy Cenkeramaddi. Deep learning-based sign language digits recognition from thermal images with edge computing system. *IEEE Sensors Journal*, 21(9):10445–10453, 2021.

[5] Rafael Y Brzezinski, Neta Rabin, Nir Lewis, Racheli Peled, Ariel Kerpel, Avishai M Tsur, Omer Gendelman, Nili Naftali-Shani, Irina Gringauz, Howard Amital, et al. Automated processing of thermal imaging to detect covid-19. *Scientific Reports*, 11(1):1–10, 2021.

[6] Karen Cardwell, Karen Jordan, Paula Byrne, Susan M Smith, Patricia Harrington, Mairin Ryan, and Michelle O'Neill. The effectiveness of non-contact thermal screening as a means of identifying cases of covid-19: a rapid review of the evidence. *Reviews in Medical Virology*, 31(4):e2192, 2021.

[7] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2021.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[10] S. Cosar and N. Bellotto. Human re-identification with a robot thermal camera using entropy-based sampling. *Journal of Intelligent & Robotic Systems*, 2019.

[11] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Mod-

ulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25(1):245–262, 2014.

[13] Munkhjargal Gochoo, Tan-Hsu Tan, Shih-Chia Huang, Tsedevdorj Batjargal, Jun-Wei Hsieh, Fady S Alnajjar, and Yung-Fu Chen. Novel iot-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning. *IEEE Internet of Things Journal*, 6(4):7192–7200, 2019.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[15] Nanhao Gu, Bo Yang, and Tong Zhang. Dynamic fuzzy background removal for indoor human target perception based on thermopile array sensor. *IEEE Sensors Journal*, 20(1):67–76, 2019.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[17] Ouday Hanosh, Rashid Ansari, Naoum P Issa, and A Enis Cetin. Convulsive movement detection using low-resolution thermopile sensor array. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 300–301, 2020.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[25] Saki Iwata, Yasutomo Kawanishi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, and Tomoyoshi Aizawa. Lfir2pose: Pose estimation from an extremely low-resolution fir image sequence. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2597–2603. IEEE, 2021.

[26] Kun-Ru Wu Jia-Yan Li, Chao-Ho Lin and Yu-Chee Tseng. SensePred: Guiding video prediction by wearable sensors. *IEEE Internet of Things Journal (to appear)*.

[27] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[29] Takayuki Kawashima, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase, Daisuke Deguchi, Tomoyoshi Aizawa, and Masato Kawade. Action recognition from extremely low-resolution thermal image sequence. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[30] Masafumi Kimata. Trends in small-format infrared array sensors. *SENSORS, 2013 IEEE*, pages 1–4, 2013.

[31] Masafumi Kimata. Uncooled infrared focal plane arrays. *IEEJ Transactions on Electrical and Electronic Engineering*, 13(1):4–12, 2018.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[33] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[34] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[35] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[36] Zhixin Liu, Ming Yang, Yazhou Yuan, and Kit Yan Chan. Fall detection and personnel tracking system using infrared array sensors. *IEEE Sensors Journal*, 20(16):9558–9566, 2020.

[37] Yawen Lu and Guoyu Lu. An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3833–3843, 2021.

[38] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[39] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[40] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[41] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

[42] Igor Morawski and Wen-Nung Lie. Two-stream deep learning architecture for action recognition by using extremely low-resolution infrared thermopile arrays. In *International Workshop on Advanced Imaging Technology (IWAIT) 2020*, volume 11515, page 115150Y. International Society for Optics and Photonics, 2020.

[43] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[44] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[45] Dongning Qu, Bo Yang, and Nanhao Gu. Indoor multiple human targets localization and tracking using thermopile sensor. *Infrared Physics & Technology*, 97:349–359, 2019.

[46] Mikko Rinta-Homi, Naser Hossein Motlagh, Agustin Zuniga, Huber Flores, and Petteri Nurmi. How low can you go? performance trade-offs in low-resolution thermal sensors for occupancy detection: A systematic evaluation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–22, 2021.

[47] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2020.

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[49] Qi Zhan, Wenjin Wang, and Xiaorong Ding. Examination of potential of thermopile-based contactless respiratory gating. *Sensors*, 21(16):5525, 2021.

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[51] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021.

[52] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[54] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[55] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.