From RGB to Depth and Thermal: Mapping between Modalities to Alleviate Data Scarcity

Christian Stippel, Thomas Heitzinger, Rafael Sterzinger and Martin Kampel
TU Wien, Computer Vision Lab
Favoritenstr. 9/193-1, 1040 Vienna, Austria
{christian.stippel,thomas.heitzinger,rafael.sterzinger,martin.kampel}@tuwien.ac.at

Abstract

In human behavior analysis (HBA), RGB data is the predominant modality. However, its inherent limitations, such as sensitivity to lighting conditions and potential privacy infringements, are of particular concern in the context of HBA. Although thermal and depth data are less susceptible to these problems, datasets tailored to HBA within these modalities are scarce. In order to alleviate this issue, we present a generative approach for producing trimodal, in particular, human-centric datasets by mapping from RGB to thermal and depth. Using human segmentation masks from RGB images and thermal/depth backgrounds, our method employs conditional inpainting to produce high-quality synthetic depth/thermal data. Initial results showcase the potential of our proposed solution.

1. Introduction

Datasets in the domain of computer vision typically consist of RGB data due to its accessibility and affordability [9, 13]. Despite its popularity, this modality has drawbacks: RGB sensors are vulnerable to lighting variations, compromising performance in fluctuating conditions, and privacy concerns in sensitive areas, especially in HBA, as individuals are easily identifiable. In such cases, thermal or depth sensors are preferable, as they detect human presence without revealing identifiable characteristics and are less susceptible to lightning conditions.

While there are thermal and depth datasets containing labels for HBA such as OSU Thermal Pedestrian Dataset [2], ThermalWorld [8], and those by Heitzinger et al. [3, 4], they are not as comprehensive as their RGB counterparts. Remarkably, to our knowledge, only the dataset by Palermo et al. [11] provides a trimodal combination of RGB, thermal, and depth data for human-centric analysis.

While synthetic generation approaches exist to address this shortage, they all have their shortcomings: ThermalSynth [10] entails time-consuming processes and caters only to specific applications; MiDaS [12], an image translation model, excels in depth map generation but cannot provide absolute depth values; and ThermalGAN [8] relies on an older architecture which suggests potential room for improvements.

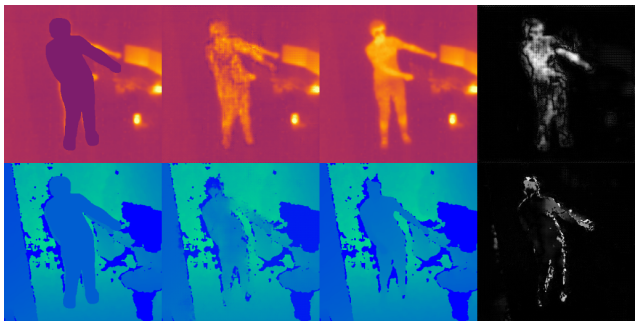

Figure 1. From RGB to depth and thermal: the top row features depth images, while the bottom row showcases thermal ones; images are arranged from left to right as follows: conditional input, inpainted output, ground truth, and absolute normalized error.

In this work, we propose a new approach to convert RGB into depth and thermal to level out the imbalance between these modalities while at the same time alleviating the aforementioned problems. In detail, once the mapping model has been trained, our approach for generating trimodal datasets needs only two elements:

1. RGB Video Datasets of People with a Static Camera
2. Background Depth and Thermal Frames

The advantage of our method lies in the ease of acquiring these two components: many RGB video datasets of people with static cameras are publicly available, and obtaining background depth and thermal frames is straightforward. Consequently, our approach simplifies generating trimodal datasets, offering a more efficient alternative to recording a conventional trimodal dataset from scratch.

Having acquired these components, we extract human segmentation masks from RGB data by utilizing Segment Anything Model (SAM) [7], a general-purpose segmentation network. Then, we combine the masks with the background thermal and depth images and use image-to-image techniques to insert humans into these modalities.

This strategy builds on the stability of image-to-image translation networks within the same modality, sidestepping the error-prone task of direct RGB to thermal or depth conversion.

2. Method: Mapping between Modalities

In our preprocessing pipeline, we utilize TRISTAR [14], a trimodal dataset, to create input-output pairs of images that are later used to train our mapping model:

1. **Bounding Box Extraction**: We extract bounding boxes from the segmentation masks that encapsulate the human figure in thermal and depth modalities.

2. **Mask Dilation**: We dilate the segmentation masks to capture the human figure and its vicinity. This process expands the boundaries of the masks with an 8×8 kernel, ensuring that the surrounding context is considered as well.

3. **Frame Modification**: We create a copy for each depth and thermal frame to modify it based on the dilated mask. Specifically, pixels within the mask are set to the respective modality's mean value (thermal or depth). Essentially, this step "erases" the human figures from frames, replacing them with a modality-neutral value.

For our architecture, we draw inspiration from the pix2pix framework [6]. In detail, our backbone model consists of two separate UNets, one dedicated to thermal and the other to depth inpainting. Each UNet consists of encoders, that comprise four convolutional blocks to progressively downsample the input and extract meaningful features, as well as decoders, that comprise the same number of deconvolutional blocks to again upsample these features.

In order to train our network, we start with an initial training phase using the Mean Squared Error (MSE) and only later introduce a Discriminator, based on the PatchGAN architecture, to further refine and enhance our predictions. The Discriminator is optimized every other epoch, while the Generator (UNets) is optimized every epoch. After the initial training phase, we modify the loss of the Generator: we combine the previously employed MSE loss, which ensures adherence to the images in our dataset, and the Binary Cross Entropy (BCE) loss, based on the Discriminator, to improve quality on a more granular level. Precisely, the total Generator loss is composed of the following weighted sum: $0.7 \cdot \text{MSE} + 0.3 \cdot \text{BCE}$.

After training, we can employ this model to create thermal and depth images using masks derived from RGB data combined with matching depth and thermal frames.

A performance comparison between the model trained only with the MSE loss and the one trained with the combination of MSE and BCE loss is depicted in Table 1. For a visual illustration of the mapping quality using MSE and BCE loss, see Figure 1. In regard to evaluation metrics, we compute the Root Mean Squared Error (RMSE) with a sliding window of eight, as well as the Fréchet Inception Distance (FID) [5] and the Kernel Inception Distance (KID) [1] using a pretrained Inception (v3) model. The metrics are calculated for the whole image to allow the pretrained Inception model to analyze the contrast of the person to the background.

Table 1. Comparison between UNet with MSE and MSE+BCE on various metrics for thermal/depth image generation; lower is better.

Modality	Metric	MSE	MSE+BCE
Thermal	**RMSE**	**0.095**	**0.095**
	KID	0.089 ± 0.003	$\mathbf{0.068 \pm 0.002}$
	FID	79.379	**63.882**
Depth	**RMSE**	0.139	**0.133**
	KID	$\mathbf{0.056 \pm 0.002}$	0.060 ± 0.003
	FID	**82.096**	84.905

Notably, we utilized the latest feature layer (2048) of Inception, pretrained on RGB data. This, combined with the difference in modalities (RGB versus depth and thermal), negatively impacts the expressiveness of the FID and KID metrics in our context. Hence, although our results generally showcase the usefulness of introducing an additional BCE loss, especially for the thermal modality, they need to be taken with a grain of salt.

3. Discussion and Conclusion

As the demand for thermal and depth datasets increases, especially in scenarios requiring enhanced privacy and performance in various lighting conditions, the search for viable solutions becomes increasingly crucial. In this paper, we introduced a new approach to fulfill this demand by transforming existing RGB datasets into corresponding thermal and depth counterparts. The initial results of our proposed approach, coupled with its minimal requirements – namely, an RGB video dataset of people featuring individuals in a static camera setting, along with background depth and thermal frames – underscore its viability and applicability in real-world scenarios. While our findings are promising, there remains room for improvement, for instance, further investigations with diffusion models.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[2] J Davis and M Keck. A Two-Stage Template Approach to Person Detection in Thermal Imagery. In *Proceeding of Workshop on Applications of Computer Vision (WACV)*, 2005.

[3] Thomas Heitzinger and Martin Kampel. A Foundation for 3D Human Behavior Detection in Privacy-Sensitive Domains. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 305. BMVA Press, 2021.

[4] Thomas Heitzinger and Martin Kampel. IPT: A Dataset for Identity Preserved Tracking in Closed Domains. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8228–8234. IEEE, 2021.

[5] 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.

[6] 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.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[8] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 606–624, 2018.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[10] Neelu Madan, Mia Sandra Nicole Siemon, Magnus Kaufmann Gjerde, Bastian Starup Petersson, Arijus Grotuzas, Malthe Aaholm Esbensen, Ivan Adriyanov Nikolov, Mark Philip Philipsen, Kamal Nasrollahi, and Thomas B Moeslund. ThermalSynth: A Novel Approach for Generating Synthetic Thermal Human Scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 130–139, 2023.

[11] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B Moeslund, and Sergio Escalera. Multi-modal RGB–Depth–Thermal Human Body Segmentation. *International Journal of Computer Vision*, 118:217–239, 2016.

[12] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International journal of computer vision*, 115:211–252, 2015.

[14] Christian Stippel, Thomas Heitzinger, and Martin Kampel. A Trimodal Dataset: RGB, Thermal, and Depth for Human Segmentation and Temporal Action Detection. In *DAGM German Conference on Pattern Recognition*. Springer, 2023.