TU WIEN Informatics

Multimodal Machine Learning to alleviate Data Scarcity

Trimodal Datensatz Erzeugung

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Visual Computing

eingereicht von

Christian Stippel, Bsc.
Matrikelnummer 11778254

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: PD. Dr. techn. Dipl. Ing. Martin Kampel
Mitwirkung: Dipl.Ing. Thomas Heitzinger, Bsc.

Wien, 30. Jänner 2024

Christian Stippel | Martin Kampel

TU WIEN Informatics

Multimodal Machine Learning to alleviate Data Scarcity

Trimodal Dataset Generation

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Visual Computing

by

Christian Stippel, Bsc.
Registration Number 11778254

to the Faculty of Informatics

at the TU Wien

Advisor: PD. Dr. techn. Dipl. Ing. Martin Kampel
Assistance: Dipl.Ing. Thomas Heitzinger, Bsc.

Vienna, January 30, 2024

Christian Stippel	Martin Kampel

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Christian Stippel, Bsc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 30. Jänner 2024

Christian Stippel

Danksagung

Zuallererst möchte ich meiner Familie meinen Dank aussprechen. Ihre Unterstützung und ihr Glaube an mich während meiner gesamten akademischen Reise haben mir viel ermöglicht. Ohne ihre Ermutigung und ihr Verständnis wäre dieser Weg ungleich schwieriger gewesen.

Ein besonderer Dank gilt meinen Freunden, die mich unterstützt haben. Ihr wart immer da, um mir bei schwierigen Entscheidungen zu helfen, mich in stressigen Zeiten aufzumuntern und Erfolge mit mir zu feiern. Eure Freundschaft, euer Humor und eure Ratschläge haben mir geholfen, die Herausforderungen und Belastungen des Studiums zu bewältigen.

Insbesondere möchte ich mich bei jenen Freunden bedanken, die mich beim Schreiben der Publikationen und meiner Masterarbeit unterstützt haben. Ein besonderes Dankeschön daher an Wolfgang Koch, Simon Reiser, Katharina Scheucher, Christian Sallinger, Jakob Kolhas, Anton Mihalkevich, Saeed Helali, Sebastian Steiner, Benjamin Schwendinger, Mathias Wess und Christian Pratallesi für die Hilfe beim Labeln des TRISTAR Datensatzes und Korrekturlesen der Paper und der Masterarbeit.

Ein ganz besonderer Dank gilt Rafael Sterzinger. Seine Unterstützung war während meiner gesamten akademischen Reise von unschätzbarem Wert. Rafael stand mir nicht nur bei der Masterarbeit, sondern auch bei zahlreichen Herausforderungen und Projekten während des Studiums zur Seite. Sein unermüdlicher Einsatz, sei es durch fachliche Beratung oder einfach durch ermutigende Worte in schwierigen Zeiten, hat mir geholfen, Hindernisse zu überwinden und meine Ziele zu erreichen. Für seine Hingabe, Geduld und die zahlreichen Stunden, die er mir gewidmet hat, bin ich zutiefst dankbar.

Schließlich möchte ich meinen Betreuern Thomas Heitzinger und Martin Kampel meinen herzlichen Dank aussprechen. Eure Expertise, eure Leidenschaft für die Forschung und euer Engagement für meine akademische Entwicklung haben tiefen Eindruck bei mir hinterlassen. Eure Anleitung war entscheidend für meinen Erfolg und hat mir geholfen, mein Potenzial voll auszuschöpfen. Vor allem möchte ich mich für eure Spontanität und Flexibilität bedanken.

Acknowledgements

First and foremost, I would like to express my gratitude to my family. Their support and belief in me throughout my academic journey has made a lot possible. Without their encouragement and understanding, this journey would have been much more difficult.

A special thanks to my friends who have supported me. You have always been there to help me make difficult decisions, cheer me up in stressful times and celebrate successes with me. Your friendship, humor and advice have helped me to overcome the challenges and stresses of studying.

In particular, I would like to thank those friends who supported me in writing the publications and my Master's thesis. A special thank you to Wolfgang Koch, Simon Reiser, Katharina Scheucher, Christian Sallinger, Jakob Kolhas, Anton Mihalkevich, Saeed Helali, Sebastian Steiner, Benjamin Schwendinger, Mathias Wess and Christian Pratallesi for their help in labeling the TRISTAR dataset and proofreading the paper and the Master's thesis.

A very special thank you to Rafael Sterzinger. His support has been invaluable throughout my academic journey. Rafael stood by my side not only during my Master's thesis, but also through numerous challenges and projects during my studies. His tireless efforts, be it through expert advice or simply words of encouragement during difficult times, have helped me overcome obstacles and achieve my goals. I am deeply grateful for his dedication, patience and the countless hours he has devoted to me.

Finally, I would like to express my sincere gratitude to my supervisors Thomas Heitzinger and Martin Kampel. Your expertise, passion for research and commitment to my academic development have left a deep impression on me. Your guidance was crucial to my success and helped me to realize my full potential. Above all, I would like to thank you for your spontaneity and flexibility.

Kurzfassung

Die Forschung im Bereich der Computer Vision, insbesondere in der Analyse menschlichen Verhaltens, hat sich überwiegend auf RGB-Datensätze gestützt, die trotz ihres Informationsreichtums Einschränkungen in Bezug auf Lichtverhältnisse und Datenschutzbedenken aufweisen. Um diese Herausforderungen zu adressieren, präsentiert diese Arbeit einen umfassenden Ansatz, der RGB-Daten durch thermische und Tiefendaten ergänzt, um robustere und datenschutzfreundlichere Alternativen zu bieten.

Wir führen TRISTAR ein, ein öffentliches Trimodales Segmentierungs- und Aktionsarchiv, das registrierte Sequenzen von RGB-, Tiefen- und Thermaldaten in verschiedenen Umgebungen umfasst. Dieser Datensatz beinhaltet Annotationen für die semantische Segmentierung von Menschen, per Bild Annotationen für die zeitliche Aktionsdetektion und das Verständnis von Szenen. Benchmark-Modelle, die sich auf die Segmentierung von Menschen und die Aktionsdetektion konzentrieren, zeigen signifikante Verbesserungen bei der Verwendung von Thermal- und Tiefenmodi.

Darüber hinaus entwickeln wir eine generative Technik zur Erstellung trimodaler Datensätze, indem wir RGB-Daten mittels Unsupervised Learning in Thermal- und Tiefenbilder übersetzen. Diese Methode hat das Potential Lösung in Szenarien mit begrenzter Datenverfügbarkeit oder herausfordernden Bedingungen zu sein.

Abstract

Research in computer vision, particularly in human behavior analysis, has predominantly relied on RGB datasets, which despite their information richness have limitations in terms of lighting conditions and privacy concerns. To address these challenges, this work presents a comprehensive approach that augments RGB data with thermal and depth data to provide more robust and privacy-friendly alternatives.

We introduce TRISTAR, a public trimodal segmentation and action archive comprising registered sequences of RGB, depth and thermal data in different environments. This dataset includes annotations for semantic segmentation of humans, per image annotations for temporal action detection and scene understanding. Benchmark models focusing on human segmentation and action detection show significant improvements when using thermal and depth modes.

In addition, we are developing a generative technique to create trimodal datasets by translating RGB data into thermal and depth images using unsupervised learning. This method has the potential to be a solution in scenarios with limited data availability or challenging conditions.

Contents

CHAPTER 1

Introduction

Computer vision is indispensable in a variety of research fields. Its utility spans object recognition [LMB+14, RDGF16], scene reconstruction [KZ02, MAMT15], and advanced image processing, contributing to academic and commercial advancements. The quality and type of data used are critical factors in its effectiveness. Traditional datasets predominantly employ Red-Green-Blue (RGB) imaging due to its accessibility and richness of detail under ideal conditions.

The prominence of RGB data was solidified with landmark datasets that have become integral to computer vision research. One of the pioneering datasets, "ImageNet", introduced an extensive collection of labeled RGB images suitable for image classification [DDS+09]. This dataset has millions of images that helped significantly progress Machine Learning (ML), especially in the training of Convolutional Neural Networks (CNNs) [KSH12]. A CNN is a deep learning model primarily used for processing data, such as images, with a grid-like topology. It employs convolutional layers to learn spatial hierarchies of features from input images automatically and adaptively. Another influential dataset called "Microsoft COCO" (Common Objects in Context) emerged after ImageNet [LMB+14]. COCO focuses on object recognition and segmentation in complex everyday scenes, expanding on the traditional image classification task.

RGB data is not limited to object recognition or image classification. It is also essential to Human Behavior Analysis (HBA) and used in various domains such as security, health monitoring, and human-computer interaction [Pop10]. Researchers can analyze complex patterns of human behavior by interpreting movements, postures, and gestures captured in RGB images and videos [JLD12]. For example, the healthcare sector leverages RGB data for patient monitoring, analyzing physical responses or changes to prescribed treatments [KTK15]. RGB data is widely used in HBA mainly due to its availability and ease of acquisition. This is because employing standard RGB cameras eliminates the need for specialized equipment, making it accessible to a broader audience. It also enables continuous monitoring without requiring direct human oversight.

While RGB cameras are helpful for detailed monitoring, their visual nature can be a disadvantage in certain situations: In settings where lighting conditions or privacy concerns are critical, the explicit nature of RGB may not be suitable. Including behavioral and physical details in HBA can lead to situations where privacy is at risk, with the potential of unauthorized access or misuse. This can compromise an individual's privacy and make them vulnerable to potential harm [BEG00].

The lack of privacy preservation can be observed when looking at human faces. Figure 1.1 compares faces in RGB, showcasing that a person can be easily identified in this modality.

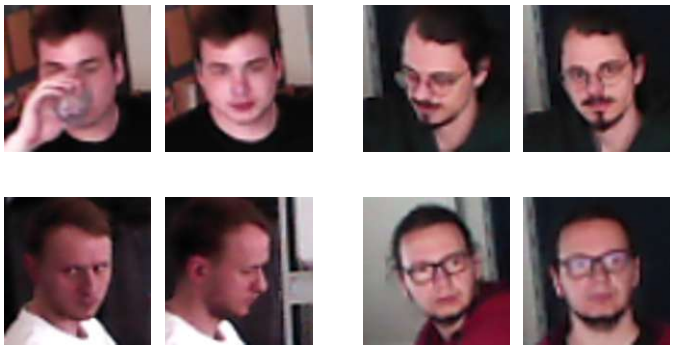


Figure 1.1: RGB faces demonstrating the potential for re-identification. The left is the original face, and the right is another face of the same sequence. Faces are re-identifiable.

On top of privacy concerns, RGB cameras face significant challenges when used in poorly lit conditions. In scenes with bad lighting, such as in natural or nocturnal settings, the camera's ability to capture reliable data can be severely impaired. This results in compromised data quality that can cause inaccuracies in HBA. Figure 1.2 illustrates the potential limitations encountered in the RGB modality when operating in different lighting conditions, underscoring the importance of using more robust imaging alternatives that perform reliably under a broader range of lighting environments.

Figure 1.2: Illustration of the limitations of RGB cameras in varying lighting conditions, highlighting the challenges in low-light scenarios for HBA.

The limitations of RGB cameras in terms of their dependence on optimal lighting conditions restrict their applicability in various HBA scenarios. Based on this and the

previously raised privacy concerns, there is a need for more robust imaging alternatives that can perform reliably across a broader range of lighting environments.

Alternative forms of imagery, such as thermal and depth data, can address these challenges. Thermal cameras use heat signatures, and depth sensors employ structured light or time-of-flight techniques, enabling operation in complete darkness or uneven lighting [FAT11]. Depth sensors generate images based on calculated depth instead of light reflection. Structured light depth sensors may not perform well with reflective surfaces or when the measured surface is hit at a shallow angle. Additionally, the lack of texture information may make it difficult to differentiate between humans and "human-shaped objects". Nevertheless, these sensors can enhance results if the background is far away and is not affected by lighting conditions.

However, these technologies are not without drawbacks. Depth sensors, which generate images based on distance measurements rather than light reflection, may struggle with reflective surfaces or when encountering surfaces at shallow angles. The absence of texture information in depth sensing can also challenge distinguishing between actual humans and objects with silhouettes similar to humans.

Despite these issues, depth sensors can be unaffected by lighting conditions. They can still improve the accuracy of HBA systems, mainly when the background is significantly from the subject [SFC+11]. This feature ensures uninterrupted HBA across various environmental conditions, facilitating applications like 24/7 patient monitoring, nighttime surveillance, or studies of nocturnal human activities. Figure 1.3 demonstrates the limitations of using RGB for privacy preservation instead of depth and thermal imaging.

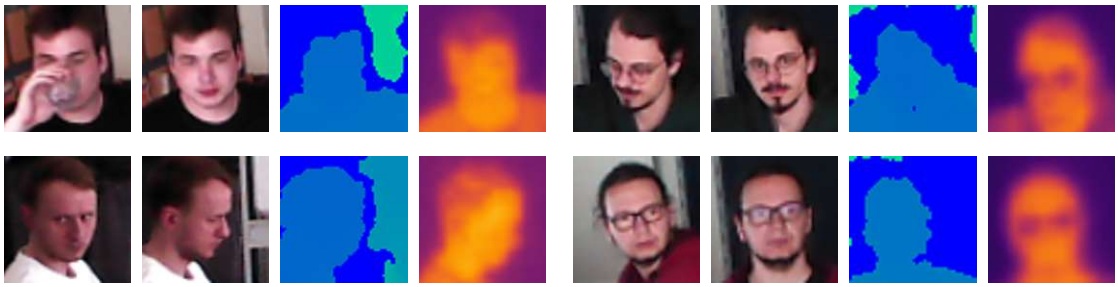


Figure 1.3: Comparison of RGB, thermal, and depth data regarding privacy concerns. The first image shows the original face, while the rest display the face from a different angle using RGB, depth, and thermal modality.

Visual data obtained under sub-optimal lighting conditions often contains "noise", necessitating sophisticated post-processing methods or making downstream tasks like action recognition infeasible. This can be computationally intensive, leading to delays in real-time monitoring applications and increased operational costs, thus undermining the cost-effectiveness of RGB. Additionally, RGB does not work in settings with no light.

The comparison between depth and thermal modalities and RGB under different lighting conditions is illustrated in Figure 1.4. When the background is spatially separated, depth

provides a clear outline of the human shape. Thermal images are well-suited for HBA because humans have a different heat signature when compared to the background.

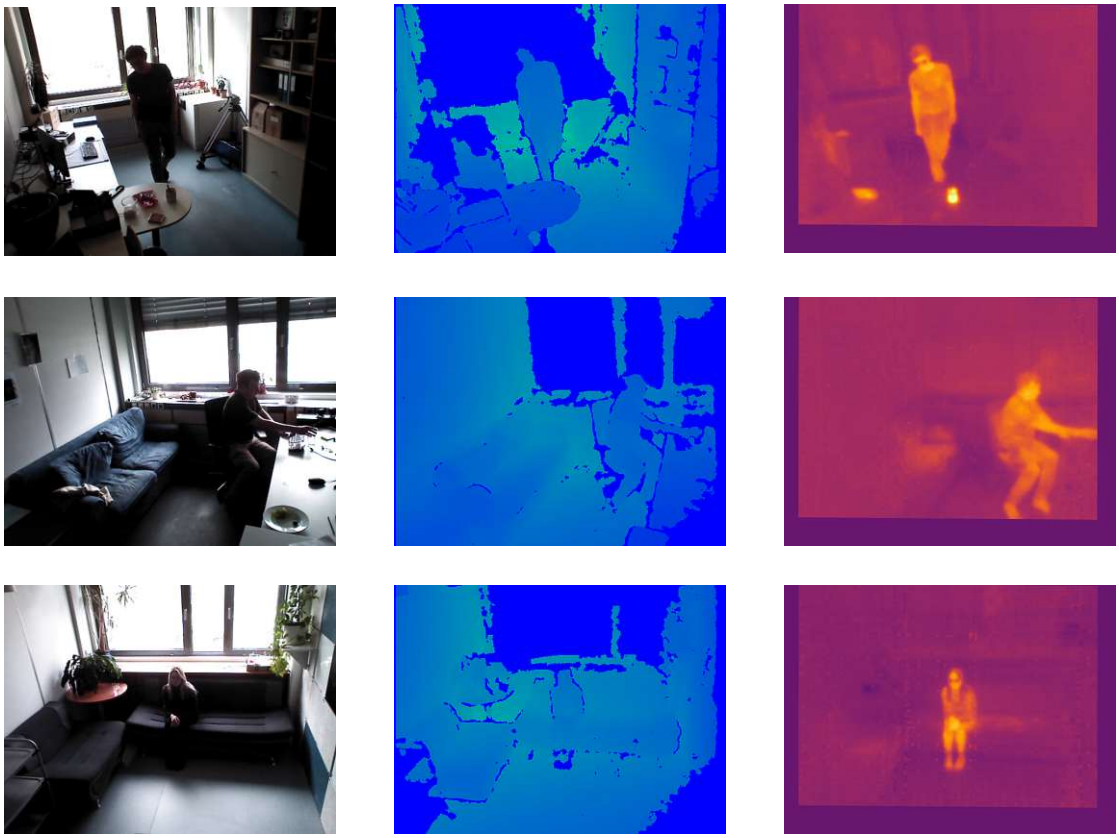


Figure 1.4: Samples of RGB, Depth, and Thermal modalities under different lighting conditions are shown in the figure below. The left frame shows RGB, while the middle and right frames display depth and thermal modalities. As can be seen, the human subject is more visible in the depth and thermal frames, even under poor lighting conditions.

The figure is divided into three columns that represent different modalities. The left column illustrates the RGB modality and highlights its dependence on ambient lighting. The middle column, dedicated to the depth modality, displays its ability to capture objects' spatial layout and contours regardless of lighting, offering a more consistent performance. Finally, the right column focuses on the thermal modality, demonstrating its proficiency in detecting heat signatures, which makes it also highly effective in diverse lighting conditions, including complete darkness.

Our qualitative samples hint towards depth and thermal being a more robust alternative than RGB. Based on our qualitative samples, it appears that depth and thermal imaging may be a more reliable option compared to RGB. One of the focuses of this thesis is to investigate whether this claim is valid and determine if there is a significant variance in performance for downstream tasks, such as action recognition.

However, even if depth and thermal can improve on RGB, the data available varies significantly among the data modalities. Depth and thermal availability are considerably limited, which hinders the development of robust, multimodal vision systems and restricts their potential in various contexts. The reasons behind this disparity include the historical focus of research, technological advancements, and practical challenges in data collection and processing [BRSB23]. No commercial devices allow you to record the registered depth and thermal [SK22]. Thus, the second focus of this thesis is to investigate and propose potential solutions to overcome the challenge of data scarcity in thermal and depth modalities.

The previously explained issues with RGB and the possible solution with supplementing or replacing RGB with depth or thermal imaging prompt the following research questions:

1. Can the use of depth and thermal modalities be an alternative to RGB for HBA?

Given the limitations of RGB datasets, exploring the potential of depth and thermal modalities to enhance or substitute RGB in HBA is beneficial. Through empirical evidence, we demonstrate the situations where depth and thermal modalities can be a viable alternatives to RGB. We publish our dataset TRImodal Segmentation and acTion ARchive (TRISTAR) [SHK23].

2. Can we translate RGB to Depth and Thermal Data with unsupervised learning?

Acquiring registered RGB, depth, and thermal data can be challenging because no publicly available sensors can record trimodal datasets [SK22]. This question examines the feasibility of using unsupervised learning to translate RGB data into these modalities, addressing the data acquisition gap and exploring the effectiveness of data generation methodologies in producing accurate depth and thermal representations from RGB inputs. Our results for this question are published at OAGM [SHSK23].

3. Can unsupervised learning serve as an effective data augmentation strategy for HBA?

Here, we test if combining a small subset of our training dataset with a more significant synthetic part performs similarly or better than the original dataset. We show the effect of augmentation in our third publication at PeRConAI [SHSK24].

This thesis presents a series of contributions, evaluations, and novel methodological advancements that collectively contribute to the domain of HBA. My key contributions are summarized as follows:

- An extensive state-of-the-art evaluation of current datasets in the field, beginning with a focus on RGB datasets in HBA then progressing to depth and thermal datasets. This evaluation demonstrates that, while depth datasets for HBA exist and thermal datasets are even less common, most labeled datasets are RGB. The

work also delves into sensor fusion to provide a view of dataset availability and applicability.

- A detailed evaluation of state-of-the-art methods for image translation, with a particular emphasis on paired image translation techniques. The *pix2pix* architecture is explored in greater detail to provide a deeper understanding of its underlying mechanisms and role in image translation tasks.

- An investigation into related work surrounding segmentation and action recognition, providing a comprehensive review of current methodologies and their performances.

- The creation and public release of a novel dataset, TRISTAR, encompasses over 15,000 frames with action recognition labels and human segmentation masks. The peer-reviewed paper titled "A Trimodal Dataset: RGB, Thermal, and Depth for Human Segmentation and Temporal Action Detection" was presented at the German Conference on Pattern Recognition (GCPR).

- The development and showcase of innovative RGB to depth and thermal translation approaches. Initial tests on direct image inpainting were published as preliminary results at the Austrian Association of Pattern Recognition (AARP) Workshop. Moreover, a novel image-to-image translation pipeline for static camera scenes was developed. This fully autonomous pipeline facilitates the translation from RGB to depth and thermal, given static RGB datasets and suitable depth and thermal background datasets. The findings were published at the PerConAI workshop, associated with a conference on pervasive computing.

- An in-depth evaluation that synthesizes all the findings published in the papers. It is demonstrated that the proposed methods enable translation from RGB to depth and thermal and support the training of new models on these translated datasets.

The collective efforts and findings presented in this thesis significantly advance the field of HBA, providing valuable resources for future research and applications.

The remainder of this thesis is structured as follows. Chapter 2 provides an in-depth overview of the existing literature. It critically examines datasets, synthetic data generation, and image translation techniques, focusing on their application in the field of HBA. Chapter 3 is dedicated to our own recorded dataset. It delves into the technical setup and the data acquisition process, providing detailed insights into the creation and specifics of our trimodal dataset TRISTAR [SHK23]. We use TRISTAR in the remaining work as a basis for all experiments. In Chapter 4, the thesis explores innovative methodologies for translating RGB data into thermal and depth modalities. Here, we explain concepts like UNet [RFB15] for image translation or ImageBind [GENL+23] for querying a database of images. The downstream tasks of our synthetic datasets, like human segmentation and action recognition, are discussed in Chapter 5. This chapter details the evaluation tasks established for our translation models, emphasizing the real-world applicability

and relevance of the research. Chapter 6 presents the outcomes of this thesis. It encompasses a thorough presentation of the evaluation metrics and methodologies employed, an analysis of our inpainting techniques, and a discussion of the results obtained from various modalities and action recognition experiments. Finally, Chapter 7 summarizes the key insights, discusses the broader impacts of the work, discusses the implications of our findings, and suggests directions for future research. This thesis presents a series of contributions, evaluations, and novel methodological advancements that collectively contribute to the domain of HBA.

CHAPTER 2

Related Work

Exploring datasets and methodologies in HBA is valuable because it helps in understanding underlying patterns and trends, developing accurate models, and gaining a more comprehensive understanding of data, especially in the context of RGB, depth, and thermal data. With their historical significance and wide application, the landscape of RGB datasets offers a foundational perspective on HBA research. An in-depth analysis reveals the strengths and limitations inherent to this modality, underscoring its critical role in the field.

We examine the advantages of depth and thermal data over RGB and discuss potential strategies for overcoming the limitations of RGB. These modalities can address the shortcomings of RGB, particularly in low-light environments or scenarios that require increased privacy. By incorporating depth and thermal data alongside RGB, we can improve our analysis of human behavior.

As the natural depth and thermal datasets are limited, we delve into the realm of data synthesis. We provide an overview of approaches for synthesizing depth and thermal images and highlight the potential of image-to-image translation methods.

Finally, we examine progress in image segmentation and human action recognition. These domains are essential for comprehending the relative efficacy of genuine and synthesized data in HBA research, which serves as a foundation for future comprehensive ablation studies.

2.1 Datasets

The increase in datasets, specifically in computer vision and machine learning, is impressive [BRSB23]. While ample RGB datasets are popular due to their simplicity of collection and wide range of applications, other modalities are less explored. The number of depth and thermal datasets is smaller than RGB datasets in HBA. Color images are

preferred because they are easier to collect. Only recent publications show devices for capturing RGB, depth, and thermal at the same time [SK22]. However, thermal and depth data can be advantageous and even necessary in certain situations. The following sections compare the most prominent datasets of each modality and highlight the gap in availability between RGB, depth, and thermal datasets.

2.1.1 Color Datasets

A few significant RGB datasets have greatly influenced the development of computer vision. One of the pioneering datasets, ImageNet, offers a vast collection of annotated images that have facilitated advancements in image classification and object detection [DDS+09]. The COCO dataset, Common Objects in Context, complements ImageNet and sets a new dataset scale and diversity standard. It has enabled the development of advanced deep-learning models that focus on objects within their everyday environments, providing a diverse array of complex scenes for model training. This has added depth to object detection and segmentation, making it stand out from other datasets [LMB+14].

Additionally, PASCAL VOC is another significant dataset that has contributed to the progress in semantic segmentation, a fundamental computer vision task [EVGW+10]. It includes images across 20 object categories and is notable for its detailed annotations spanning object detection, segmentation, and classification. The ADE20K dataset, "MIT Scene Parsing Benchmark", further expanded the scope of available RGB data [ZZP+17]. It stands out for its extensive range of object and stuff categories, combined with detailed annotations. This dataset includes a variety of scenes, from urban landscapes to interior settings, making it a valuable resource for training segmentation algorithms. The Cityscapes dataset [COR+16] offers a specialized focus on urban environments. With high-definition images from various cities and detailed segmentation masks, it has become an essential resource for applications such as autonomous driving. Lastly, the Mapillary Vistas dataset [NORBK17] demonstrates the potential of crowdsourcing in dataset creation. It provides diverse, street-level imagery with detailed annotations, suitable for training models in dynamic urban scenarios.

In the realm of HBA, RGB datasets have enabled many applications. The richness and variety of these datasets have allowed for extensive research and development in areas such as facial recognition, emotion detection, and action recognition, which are critical in HBA. In video content, the Charades dataset [SDFG17] introduced a temporal component, focusing on action recognition. This dataset comprises video clips that capture a range of human activities in domestic settings, annotated to provide insights into human-object interactions. Datasets like the FER-2013 [GEC+13] and the AffectNet [MHM17] have provided substantial resources for facial expression recognition and emotion detection. These datasets, with their vast array of annotated facial images, have been pivotal in training models to understand subtle human emotions, an essential aspect of HBA. The Kinetics dataset [KCS+17] and UCF101 [SZS12] have been central to the progress in action recognition. They offer extensive collections of video data that capture a wide range of human activities, enabling algorithms to recognize and interpret complex

human actions in various contexts. Samples from the Kinetics and UCF101 datasets are illustrated in Figures 2.1a and 2.1b, respectively, showcasing the diversity of activities and scenarios covered.

(a) Sample frames from the Kinetics dataset [KCS+17].

(b) Sample frames from the UCF101 dataset [SZS12].

Figure 2.1: Sample frames from the Kinetics and UCF101 datasets showcasing a variety of human activities and actions.

The availability of diverse RGB datasets has significantly contributed to the depth and breadth of research in HBA. These datasets cover various scenarios and environments, from controlled laboratory settings to unstructured real-world scenes. They provide a rich resource for training and evaluating models that need to operate in diverse and often challenging conditions.

In conclusion, the progression of RGB datasets, from foundational ones like ImageNet and COCO to those tailored explicitly for HBA tasks, is a cornerstone in advancing the field of computer vision and, particularly, HBA. Their extensive coverage enables researchers to explore and innovate in understanding and analyzing human behavior without collecting new datasets.

While RGB datasets have been instrumental, they are less effective under poor lighting conditions, where depth and thermal modalities excel due to their ability to capture spatial and temperature variations. Our TRISTAR dataset adds depth and thermal datasets, and the benchmark models implemented show improvements when supplementing RGB with depth and thermal.

2.1.2 Depth Datasets

Depth datasets have emerged as pivotal resources in computer vision, playing a foundational role in tasks that necessitate understanding scene geometry, such as 3D reconstruction, scene understanding, and object detection in cluttered environments. The information in these datasets, encapsulating the distance between the camera and the scene's objects, brings dimensionality to scene understanding that RGB data alone cannot offer.

Among outdoor depth datasets, KITTI [GLU12] stands out. Originating from the Karlsruhe Institute of Technology and Toyota Technological Institute in Chicago, KITTI provides a diverse array of data, including depth maps, captured in urban and rural settings. Comprising over 93,000 depth maps, its annotations encompass a variety of tasks ranging from optical flow to 3D object detection. Another significant outdoor dataset is Cityscapes [COR+16]. While it contains RGB data and segmentation masks, Cityscapes offers depth information for its high-definition images sourced from 50 cities. When juxtaposed with the rich segmentation masks, the depth annotations provide a comprehensive understanding of urban environments, with applications extending from pedestrian detection to scene parsing.

NYU Depth, which New York University curated, is a valuable resource for analyzing indoor environments. This dataset offers various paired RGB and depth images covering indoor scenes, such as kitchens and bedrooms. It also provides detailed segmentation masks for over 1,000 object and stuff categories and depth information, which can be used to predict 3D room layouts and improve indoor scene understanding [SHKF12].

Additionally, the Middlebury Stereo dataset [SS02] has played a crucial role in stereo vision research. It offers a range of stereo images with ground truth depth data, essential for depth estimation and 3D reconstruction studies.

The SUN RGB-D dataset [SLX15] further enriches the collection of indoor depth data. It includes indoor scenes captured with depth sensors, providing a comprehensive resource for object recognition and scene understanding tasks.

For tracking within enclosed spaces, the IPT dataset [HK21b] is applicable. It is designed explicitly for tracking tasks and offers depth data captured in constrained environments. It is particularly suitable for applications like surveillance or robotics, where tracking in cluttered, constrained spaces is paramount.

Regarding HBA, the depth modality has received limited attention in research, with only a few publications available on the topic [HK21a, Esc12, AMY18]. The limited exploration of depth data in HBA can be attributed to several factors. Firstly, collecting depth data that accurately captures human behavior in a wide range of scenarios is challenging. It requires sophisticated depth-sensing technology and scenarios where human subjects are involved in diverse activities. Secondly, interpreting depth data for understanding complex human behaviors demands advanced algorithms that can process and analyze 3D spatial information in the context of human actions and interactions.

Figure 2.2 shows samples from the IPT and MIPT datasets of Heitzinger et al. They solve the privacy issue, but no action labels are included. Additionally, MIPT even contains thermal data.

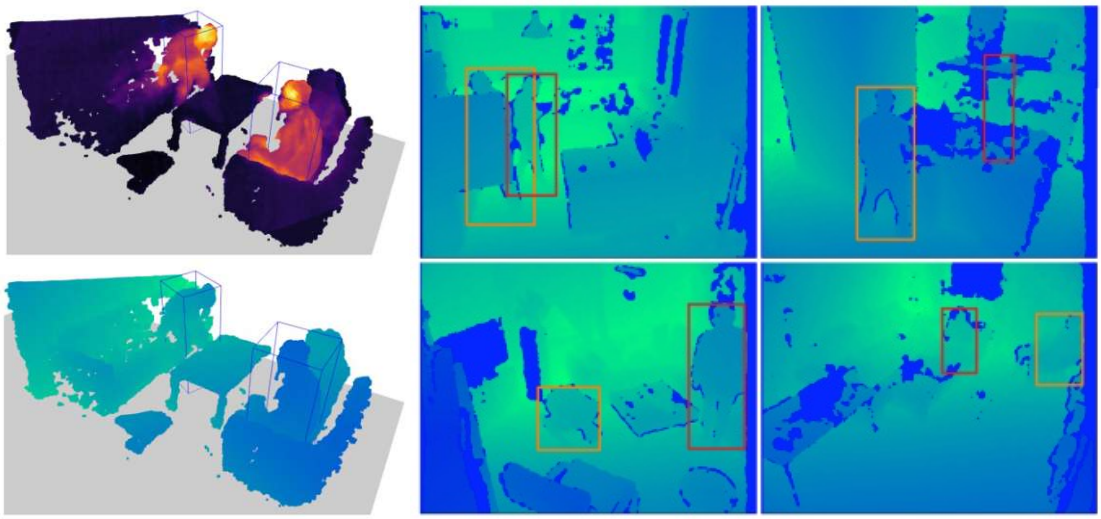


Figure 2.2: Sample frames from the IPT and MIPT datasets, showcasing depth imaging while addressing privacy concerns [HK21b, HK21a].

There is an abundance of depth datasets available. However, most of them focus on tracking and positioning. Fewer datasets specifically deal with human activity when compared to RGB datasets. RGB activity datasets are frequently available. If you require a new activity recognition dataset, finding one or using foundation models is possible.

2.1.3 Thermal Datasets

Thermal imaging, with its ability to capture temperature distributions, has gained significant traction within the computer vision community for its manifold applications and distinctive advantages [HK21a, HK21b, KKH+18]. Unlike its RGB counterpart, thermal imaging remains impervious primarily to illumination variations. This characteristic often proves invaluable in low-light scenarios or when discerning living entities based on their heat signatures. Furthermore, thermal images offer insights into the intrinsic physiological state of subjects, opening up novel research avenues that RGB imaging might not cater to.

Another thermal dataset is the OSU Thermal Pedestrian Dataset [DK05]. Curated by Ohio State University, this dataset is a comprehensive compilation of thermal images spotlighting pedestrians, encompassing a myriad of environmental settings. From varying ambient temperatures to different times of the day, this dataset ensures that models trained on it can detect pedestrians in diverse conditions. However, it's worth noting that its primary emphasis lies in pedestrian detection, with limited scope for HBA. The Terravic Facial Infrared Database can be used for facial recognition. It offers a color

and thermal facial image mix [Mie05]. This allows for a richer understanding of facial features and characteristics, proving particularly beneficial in biometric authentication scenarios where the subtle heat variations of facial landmarks can aid identification.

To leverage thermal data for person re-identification, Kniaz et al. introduced Thermalgan [KKH+18]. Within their work, they also published the ThermalWorld dataset. ThermalWorld explicitly focuses on personal re-identification. MIPT also includes some thermal samples [HK21a].

Thermal datasets are not as widely used as other datasets, even though they clearly distinguish humans in HBA as they have a different heat signature.

2.1.4 Combinations of Modalities

The PST900 dataset [SRZ+20] is one resource that proposes long-wave infrared (LWIR) imagery as a supporting modality for semantic segmentation using learning-based techniques. This dataset provides 894 synchronized and calibrated RGB and thermal image pairs with per-pixel human annotations across four distinct classes. In addition to presenting a unique dataset, the authors introduce a novel passive calibration target.

Another notable resource is the InfAR action dataset [GDL+16], which focuses on action recognition using infrared data. To our knowledge, only a single dataset exists that combines RGB, thermal, and depth data [PCB+16] for human segmentation. This dataset comprises 5,274 frames recorded in three shots in three distinct office scenes. Figure 2.3 shows samples from this unique dataset.

Figure 2.3: Sample frames from the Multi-modal RGB-Depth-Thermal Human Body Segmentation dataset, illustrating the integration of different data modalities for human segmentation in office environments [PCB+16].

While there is one dataset, the background variation will likely lead to overfitting these scenes. Our dataset comprises 18 distinct views, which mitigates the risk of overfitting. Finally, Brenner et al.'s survey [BRSB23] provides a systematic literature review of the

fusion of RGB-D and thermal sensor data, highlighting the progress made in this area over the past decade.

2.2 Synthetic Data

The field of synthetic dataset generation is well-explored, with significant research focused on remodeling scenes. For example, ThermalSynth [MSG+23] employs specialized shaders for this purpose. However, this approach is time-intensive and primarily utilized for rare action modeling, limiting its broader applicability.

In contrast, the problem of translating images to depth maps is addressed using different methods. For instance, the MiDaS model [RLH+20] can do depth estimation up to a scale and shift. However, this approach falls short in its inability to predict absolute depth values, an aspect that newer developments in the field aim to tackle. Recent models [BBW+23], [SKNF23] leverage diffusion models for depth prediction, improving the relative depth approach.

As for the translation of RGB to thermal images, ThermalGAN [KKH+18] stands out. However, its effectiveness might be compromised due to its reliance on an older architecture, highlighting a potential area for further improvement.

2.3 Image Translation

In digital image processing, image inpainting offers innovative methods for embedding objects into background scenes. This approach is beneficial when transforming images from standard RGB formats to other modalities like depth or thermal imaging. This section highlights critical advancements in image inpainting, emphasizing its applicability in conditional inpainting using masks, specifically for human figures.

A groundbreaking innovation in this field is the development of conditional Generative Adversarial Networks (cGANs). In cGANs, the generator and discriminator networks are conditioned on supplementary data, such as class labels or information from different image modalities. This strategy is exceedingly effective for applications including photo refinement, artistic style transfer, and even in the analysis of medical imagery, where contextual interpretation is crucial.

Deep learning models have facilitated a significant leap forward in image translation technology, especially those incorporating a U-Net architecture within a Generative Adversarial Network (GAN) framework. The pioneering model in this domain is Pix2Pix, which uniquely combines a U-Net architecture with a PatchGAN discriminator for adept image-to-image translation [IZZE17].

The U-Net architecture was initially developed for biomedical image segmentation. It features a dual-pathway design: a contraction path to assimilate context and a symmetrically expanding path for precise localization. This configuration is particularly effective for tasks requiring spatial awareness in image processing.

Within the GAN structure, the U-Net acts as the generator, creating images that closely resemble real images from the desired domain. Conversely, the discriminator is trained to differentiate between these synthetic images and authentic images from the target domain. Both the generator and discriminator undergo simultaneous training in a competitive setting. The generator strives to produce increasingly convincing images, while the discriminator improves its ability to identify synthetic creations.

The discriminator architecture used is called the PatchGAN discriminator. The PatchGAN discriminator divides the input image into overlapping patches. Each of these patches is independently classified as real or fake. Essentially, it assesses whether each patch is drawn from the distribution of patches in the real images. The PatchGAN can understand and critique finer details and textures by focusing on smaller regions of the image. This is important in tasks like style transfer or photo-realistic image generation, where details matter. However, because patches are part of the whole image, they also capture some global context, allowing the model to consider the overall coherence of the picture. One advantage of this approach is that it is computationally more efficient than processing the entire image simultaneously. Also, because it focuses on local features, it tends to be more effective in capturing high-frequency information, which is crucial for generating sharp and realistic images.

While Pix2Pix works well, further work is needed to improve RGB image-to-image translation results. Zhao et al. [ZCS+21] proposed a new technique that combines image-conditional and modulated unconditional generative architectures to overcome existing inpainting algorithms' limitations. These algorithms tend to fail when dealing with large missing regions. The co-modulation technique used by Zhao et al. can help achieve better results in inpainting large human-shaped masks within images. The study by Suvorov et al. [SLM+22] introduces a technique called Large Mask Inpainting (LaMa). This method can deal with large missing areas and complex geometric structures in high-resolution images. It achieves this by utilizing Fast Fourier Convolutions (FFCs) and a high receptive field perceptual loss, which expands the effective receptive field of the inpainting network and the loss function. As a result, this approach leads to significantly improved inpainting results. In our context, this technique could enable effective human inpainting into complex background scenes.

2.4 Segmentation and Action Recognition

This thesis uses segmentation as part of its mapping pipeline and action recognition for evaluation. For segmentation, a combination of a variation of You Only Look Once (YOLOv7) [WBL23] and Segment Anything (SA) [KMR+23] is used. Action Recognition is done with various 3D Convolutional Neural Networks [TBF+15, HKS17].

YOLOv7 is a fast, real-time object detection system that views images in a single glance, predicting object locations and classifications simultaneously [WBL23]. Unlike traditional systems that perform separate steps for object localization and classification, YOLO

unifies these processes, enabling it to quickly and accurately identify multiple objects in complex scenes.

SA introduces a novel approach to image segmentation, encompassing a unique task, an efficient model, and an extensive dataset [KMR+23]. This project has created the largest segmentation dataset, featuring over 1 billion masks across 11 million licensed images that respect privacy. The model at the heart of SA is designed to be promptable, enabling it to adapt zero-shot to new image distributions and tasks without requiring specific training for each new scenario. Evaluations of the model's capabilities on various tasks have demonstrated its impressive zero-shot performance, often matching or surpassing results from fully supervised models.

3D Convolutional Neural Networks performed well on action recognition tasks [TBF+15]. Unlike traditional 2D convolutions that process single images, 3D convolutions extend this by considering the temporal dimension, making them well-suited for analyzing video data. 3D convolutions extract features from spatial and temporal dimensions by applying filters to consecutive frames, thus capturing the inherent motion in actions. A prominent example is the 3D ResNet, an extension of the Residual Network architecture to 3D convolutions [HKS17]. It leverages residual learning, using shortcut connections to skip one or more layers in a deep 3D CNN framework. 3D ResNet models are particularly effective in action recognition tasks, as they can learn complex and nuanced patterns of motion and appearance that define various actions in video sequences.

CHAPTER 3

Dataset

Constructing a specialized dataset is essential to address our research questions empirically, particularly in translating RGB to depth and thermal imagery and evaluating HBA through action recognition in video sequences. This dataset plays a critical role in several key areas:

- Comparing RGB, depth, and thermal modality for HBA.
- Paired translation from RGB to depth and thermal with inpainting.
- Comparing synthetic vs real parts of the dataset.

Additionally, the dataset aims to close the gap between depth, thermal, and RGB modality. While RGB data is a cornerstone in various computer vision applications, as evidenced in widely recognized datasets like ImageNet [DDS+09] and Microsoft COCO [LMB+14], its dependency on lighting conditions and inherent privacy concerns pose significant challenges. In scenarios where accuracy and confidentiality are of utmost importance, such as in sensitive environments or applications requiring reliable operation under diverse lighting conditions, relying solely on RGB data can be insufficient. We call our dataset TRImodal Segmentation and acTion ARchive (TRISTAR).

3.1 Motivation

Our dataset comprises sequences of registered RGB, depth, and thermal images. Additionally, each frame in the dataset is annotated with human segmentation masks and action labels. These annotations are crucial for HBA, enabling the detailed study of human actions within the captured scenes.

Key tasks that arise from temporal action labels of our dataset include:

- **Action Identification**: Identifying the type of action being performed, such as running, jumping, or dancing.

- **Temporal Localization**: Pinpointing the start and end times of the action within the video sequence. This involves segmenting the video into sections and labeling each with the corresponding action.

- **Contextual Understanding**: Analyzing the context of the action, which may include background scenes, objects, and interactions with other entities in the video.

Figure 3.1 illustrates a representative sample from our trimodal dataset. This figure showcases the unique aspects of each modality in the following order: RGB, depth, and thermal imaging alongside a human segmentation mask.

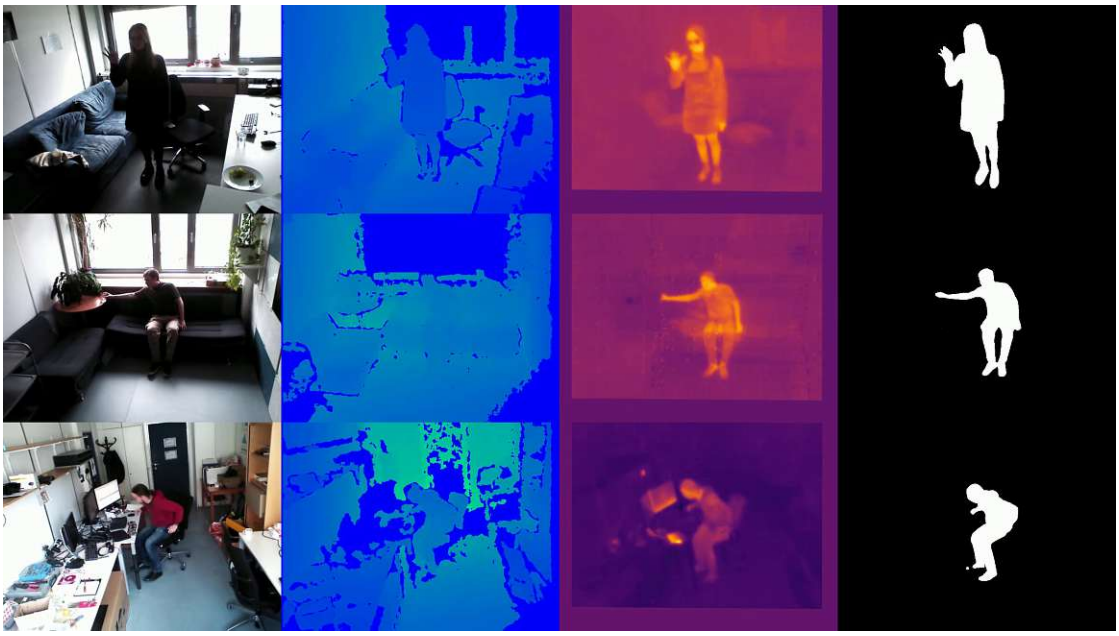
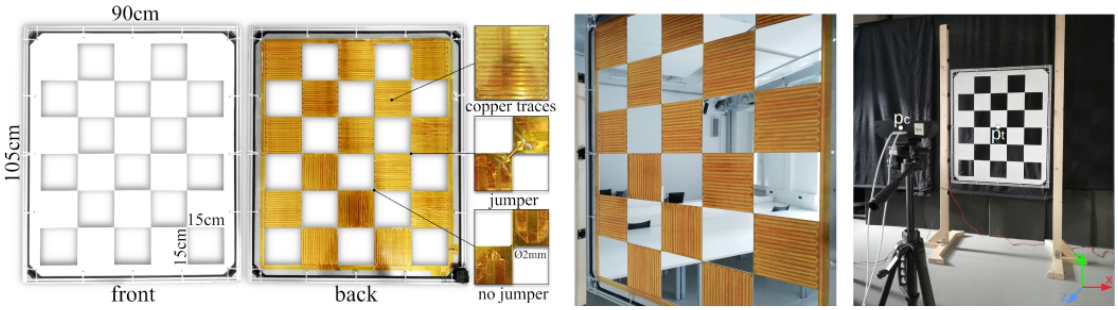


Figure 3.1: Examples from our trimodal dataset, encompassing RGB, depth, thermal imaging, and human segmentation mask from [SHK23].

The RGB image provides a detailed color representation of the scene, which is essential for understanding texture and appearance. The depth image offers spatial information, encompassing the distance of objects from the camera. In contrast, the thermal image captures the temperature distribution in the scene, which can be particularly useful for identifying living beings and understanding environmental conditions. The human segmentation mask distinguishes human figures from their surroundings.

3.2 Sensor Setup

The CTCAT, as outlined by Strohmayer et al. [SK22], combines RGB, structured light depth, and uncooled radiometric thermal cameras. The resolution of images is standardized to 640x480 pixels despite each camera's unique resolution. The alignment of the different modalities is achieved using a custom-made, heated checkerboard calibration pattern with holes. Figure 3.2 shows images of the heated checkboard calibration pattern.



(a) multi-modal calibration target (b) heating element (c) calibration setup

Figure 3.2: Figure from [SK22] to show the calibration process. (a) Front and back view of the multi-modal geometric calibration target, showing the copper traces of the custom heating element. (b) Close-up view of the custom heating element. (b) Geometric calibration setup.

Our dataset is recorded using a stand, a Bluetooth keyboard, a portable monitor, and a camera setup based on the by Strohmayer et al. obtained calibration parameters [SK22]. Figure 3.3 illustrates our camera setup.

Figure 3.3: The camera setup with the CTCAT unit and the captured scene from [SHK23].

Central to this setup is the CTCAT unit, depicted in detail on the top right and in action

on the left. On the left, it combines a portable external monitor and a stand to capture the entire scene. Adjacent to the CTCAT unit, the figure showcases a trimodal sample. This setup forms the backbone of our image capture strategy, as described in [SHK23].

3.3 Dataset Design

Building upon the work in the sensor setup, we present our resulting trimodal dataset. Our dataset encompasses an array of office scenes recorded using a trimodal sensor arrangement integrating RGB, thermal, and depth data.

To create our multimodal dataset, we draw inspiration from the Charades Dataset [SVW+16] and the work of Palmero et al. [PCB+16]. Our dataset focuses on office environments, selected for their various scenarios and activities. Actions are chosen based on their occurrence in real office settings, as detailed in Table 3.1. Different office spaces are included, such as open doors of adjacent offices and meeting rooms, each with varying interactions and lighting conditions.

Table 3.1: List of Actions, States, Transitions, and Locations used for labeling.

Label	Items
Action Classification	`put_down`, `pick_up`, `drink`, `type`, `wave`
State	`sit`, `walk`, `stand`, `lie`
Transitions	`get_down`, `get_up`
Location	`out_of_view`, `out_of_room`, `in_room`

TRISTAR contains diverse settings visually represented in Figure 3.4.

Figure 3.4: Variety of office locations and lighting conditions in the dataset.

The figure includes three images that show different light conditions. On the left, there is a hallway that connects various offices. In the middle is a kitchen corner with sofas with poor lighting conditions. Finally, on the right, a meeting room is shown. In addition to these three settings, our dataset includes various recordings of diverse offices. This

variety of settings is essential for ensuring our findings are robust and applicable across various office layouts and lighting conditions.

3.4 Ground Truth Generation

The ground truth generation process involves using pretrained YOLOv7 and YOLOv8 models to detect human bounding boxes in the RGB images [WBL23, JCQ23]. These detections are then used as conditions for SAM for preliminary human mask generation [KMR+23]. Figure 3.5 shows an example of the resulting mask when pretrained YOLOv6 [LLJ+22] is combined with SAM. The green rectangle shows the result of YOLOv6, and the slightly transparent green mask shows the result of the SAM.

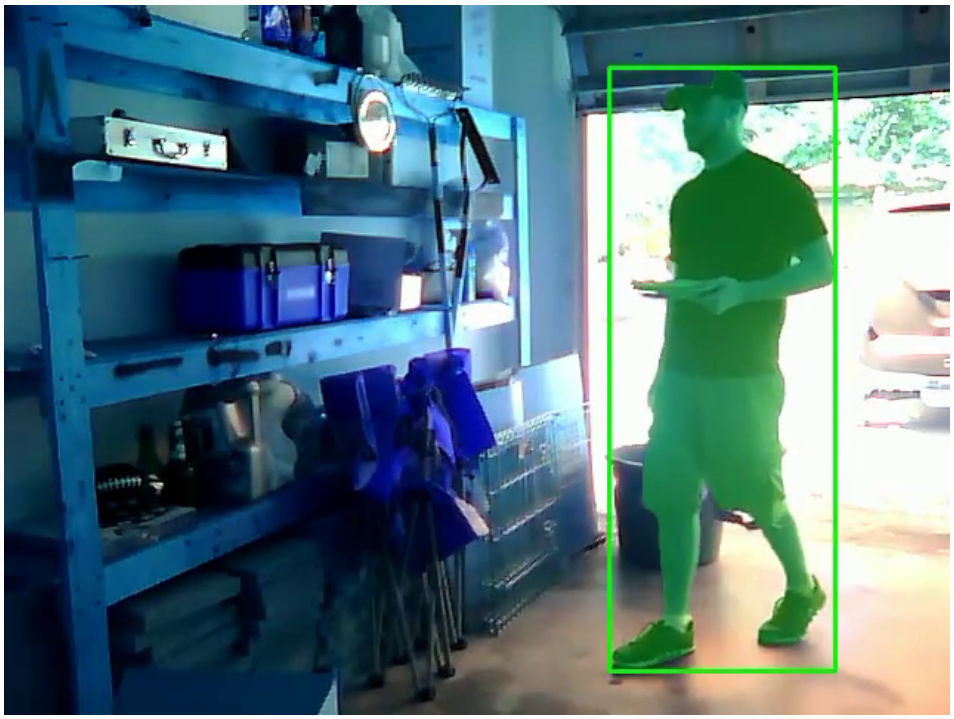


Figure 3.5: Result of applying YOLOv6 and SAM to an image of the charades dataset.

Using the human masks from the pre-labeling stage as a basis, a team of annotators labeled the RGB images, with 15,618 frames labeled in total. For increased accuracy, annotators can access corresponding thermal and depth data, color-mapped to RGB. The Label Studio[1] platform facilitates this large-scale annotation task. Figure 3.6 depicts the human segmentation annotation process.

Dense per-frame labeling of 14 classes, including actions, states, and transitions, is performed for action labeling. A spreadsheet system aids in this process, categorizing each

[1] `https://labelstud.io/` (Last accessed on 22.01.2024).

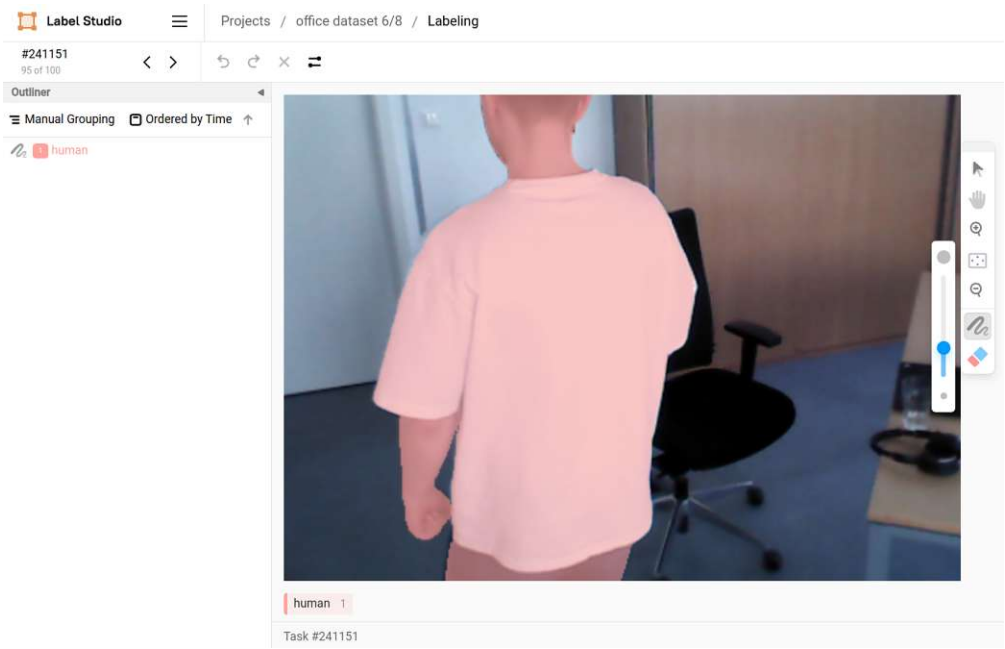


Figure 3.6: Illustration of the manual human segmentation annotation process using Label Studio.

frame's parameters' actions, states, and locations. The goal is to achieve temporal action segmentation or detection by identifying specific actions within the frames. Figure 3.7 demonstrates the action label annotation process using a spreadsheet.

	A	B	C	D	E	F	G	H
1	source	person	shot	frame	actions	transitions	state	location
2	20.csv	Thomas Leopold	20	000185_0000030580.png			sit	in_room
3	20.csv	Thomas Leopold	20	000186_0000030877.png			sit	in_room
4	20.csv	Thomas Leopold	20	000187_0000031060.png			sit	in_room
5	20.csv	Thomas Leopold	20	000188_0000031149.png			sit	in_room
6	20.csv	Thomas Leopold	20	000189_0000031239.png			sit	in_room
7	20.csv	Thomas Leopold	20	000190_0000031329.png			sit	in_room
8	20.csv	Thomas Leopold	20	000191_0000031419.png			sit	in_room
9	20.csv	Thomas Leopold	20	000192_0000031508.png			sit	in_room
10	20.csv	Thomas Leopold	20	000193_0000031598.png			sit	in_room

Figure 3.7: The action label annotation process using a spreadsheet for temporal action segmentation.

3.5 Dataset Analysis

Our dataset encompasses 10 unique office environments with 18 camera angles, 101 shots, and 15,618 frames. Each frame is annotated to provide human masks for semantic segmentation and dense labels for temporal action detection and scene understanding. Table 3.2 summarizes the contents of the dataset.

Table 3.2: Details of the Trimodal Dataset.

Content	Indoor Human Behavior
Modalities	Registered RGB, Depth, Thermal
Type of Data	Sequences
Resolution	640x480
Frame Rate	8.7 fps
#Offices	10
#Camera Angles	18
#Shots	101
#Frames	15,618
#Individuals	8
#Actions	14

The dataset split into training, validation, and test sets is structured to ensure a robust training. Figure 3.8 shows the distribution of individuals of our dataset. Each bar represents a number of frames the person occurs in.

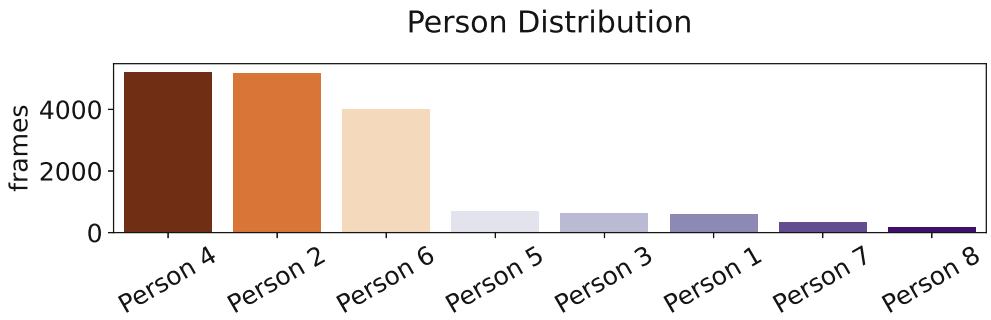


Figure 3.8: Distribution of individuals in our dataset.

Persons 1, 2, and 3 are the most frequently captured individuals. They constitute most of the training dataset and a significant portion of the validation set. This choice aims to provide a rich and diverse range of behaviors and interactions for the model to learn from during the initial training phase.

In contrast, the remaining individuals appear less frequently in the dataset and are predominantly used in the validation and test sets. Furthermore, the dataset takes into account the Variety of office environments. Different views and office settings are

exclusively reserved for the validation and test sets. Including unique views and settings in the validation and test sets is essential for assessing the model's adaptability and performance in previously unseen or novel office environments and individuals.

The diversity of actions is illustrated in Figure 3.9, where the number of frames for each action label is presented. As shown, the action 'type' has the highest occurrence, consistent with office settings where typing is a common and often prolonged activity. The actions 'wave,' 'drink,' 'pick_up,' and 'put_down' are recorded with similar frequencies, yet 'pick_up' and 'put_down' are less frequent as these represent shorter duration tasks.

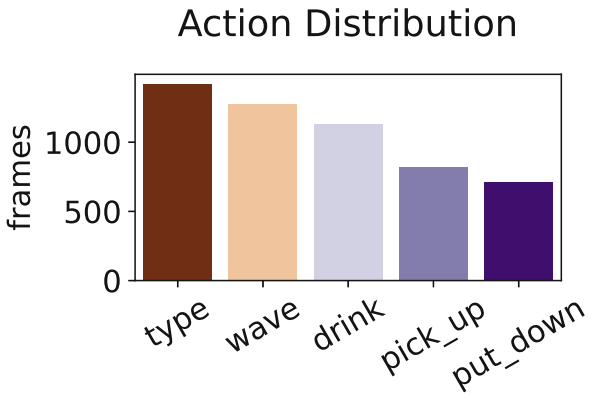


Figure 3.9: Bar chart representing the distribution of action labels in the dataset.

Figure 3.10 shows the distribution of states within the dataset. Walking is the predominant state, reflecting the high mobility within the office environment. Sitting and standing are observed almost as frequently as each other, typical for an office setting where people switch between these two states. Lying down occurs rarely, which is expected due to the lack of appropriate furniture in such settings.

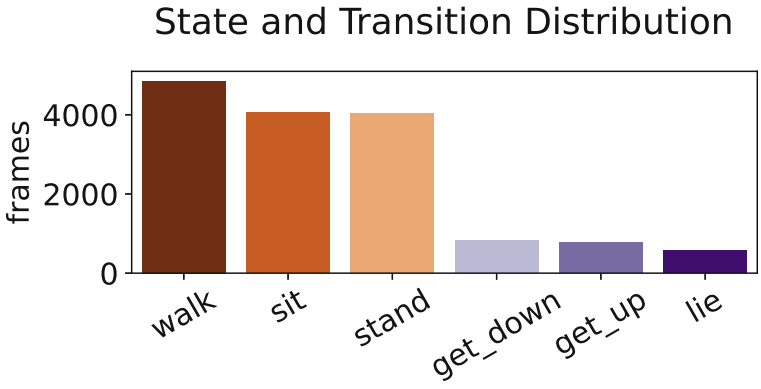


Figure 3.10: Bar chart representing the distribution of action labels in the dataset.

The transitions 'get_down' and 'get_up' are the rarest states, indicating that once a

person is sitting, walking, or lying, they tend to maintain that state for extended periods rather than transitioning frequently. Location labels like `in_room`, `out_of_view`, and `out_of_room` are also marked within the dataset.

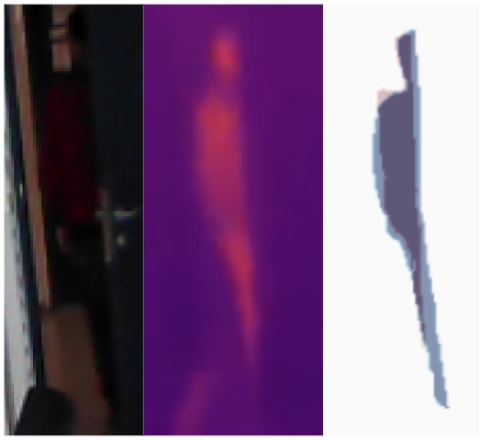
3.6 Dataset Quality Evaluation

We perform a quality assessment to ensure the consistency of our human segmentation labels. First, we test the quality of human masks. The Jaccardian Index, also known as Intersection over Union (IoU), is a commonly used metric in image segmentation tasks to quantify the accuracy of the predicted segmentation. It is defined as the size of the intersection divided by the size of the union of the sample sets. Mathematically, the IoU can be represented as:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

where A and B are the ground truth and predicted segmentation areas, respectively.

In our assessment, the Jaccardian Index averaged **0.948** across 1106 double-labeled frames, indicating a high degree of accuracy in our labeling process. Figures 3.11a and 3.11b show typical errors identified during this evaluation. Label agreement for actions, transitions, states, and locations was also high, demonstrating the robustness of our labeling process.



(a) First case: small labeled area due to a person leaving the room.

(b) Second case: discrepancy due to sloppy labeling.

Figure 3.11: Visualization of errors in RGB, thermal, and segmentation masks.

Figure 3.12 presents a confusion matrix for state transitions, explaining the model's performance in identifying various states. The matrix particularly highlights the difficulty in distinguishing between the `get_down` state and states like `walk` or `stand`.

This confusion primarily stems from the inherent challenge of precisely defining the transition points in human motion. For instance, when an individual starts sitting down from a standing position (get_down), it is easily confused with standing, leading to misclassifications between these states.

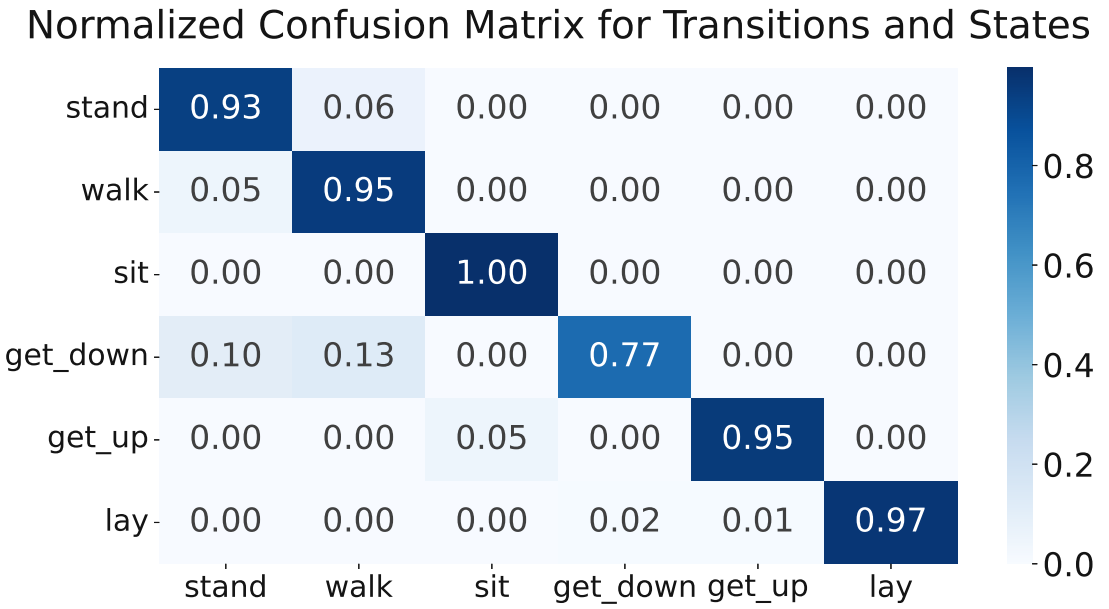


Figure 3.12: Confusion matrix for state transitions in our double-labeled dataset.

The matrix also reveals that the transition to `get_down` is frequently confused with the `stand` and `walk` states. This is attributed to the subtlety and brevity of the get_down action, which can be easily overlooked.

3.7 Challenges & Limitations

Despite the contributions of our trimodal dataset to the field of HBA, it is necessary to recognize its limitations, particularly in comparison to the extensive RGB datasets available in the domain. These limitations primarily revolve around the dataset's size and diversity.

Our dataset, while the largest among existing trimodal datasets, still falls short in size when compared to the vast collections of RGB-only datasets. The extensive size of RGB datasets is a key factor in training more robust and sophisticated models, as the depth and variety in these large datasets contribute significantly to the generalizability and performance of machine learning models. Our dataset's size limits the complexity of models that can be trained and the extent of generalization achievable. Although it is adequate to address specific tasks within the dataset's domain, there no statement can be

made about whether the models can be used in different indoor settings or with various camera configurations.

Another limitation lies in the diversity of the dataset. While our dataset offers a variety of scenarios and captures a range of human behaviors, it does not encompass the same level of diversity observed in larger RGB datasets. These RGB datasets often include a broad spectrum of environments, activities, and subjects, which are necessary for models to operate in diverse real-world settings. The limited diversity in our dataset affects its effectiveness in training models adept at recognizing and analyzing human behavior across a wide array of situations. These limitations underline the need for continued development of trimodal datasets in HBA to create more comprehensive and diverse datasets.

CHAPTER 4

Mapping RGB to Thermal and Depth

We introduce a novel approach to increase the quantity of trimodal HBA datasets. Specifically, we concentrate on converting RGB data, the most commonly used modality in HBA, into thermal and depth data to overcome limitations associated with RGB datasets. This conversion allows us to produce new trimodal datasets, which are rare in the field of HBA. Our main contributions here include:

- Utilization of Accessible Resources: Our pipeline takes advantage of the abundance of RGB datasets featuring individuals performing various actions and easily obtainable background depth and thermal frames. Obtaining these datasets is a relatively simple task due to their wide availability.
- Development of a Translation Pipeline: We map RGB to depth or thermal images by conditioning the translation process on suitable depth and thermal backgrounds, enhancing the accuracy of the transformation.
- Evaluation with Action Recognition: We assess our methodology by training action recognition models on real and synthetic datasets, demonstrating its utility in scenarios with limited depth and thermal data.

The remainder of this chapter is structured as follows. First, we explain using image inpainting to directly draw humans into thermal and depth. Second, we introduce a more complex pipeline that conditions the UNets on a background frame of the respective modality, the cropped RGB, and an SDF of the mask.

4.1 Modality Translation with Image Inpainting

Our first translation process focuses on simply inpainting existing registered thermal and depth images. Figure 4.1 provides a high-level visualization of our input and output process, encapsulating the data flow through the translation process.

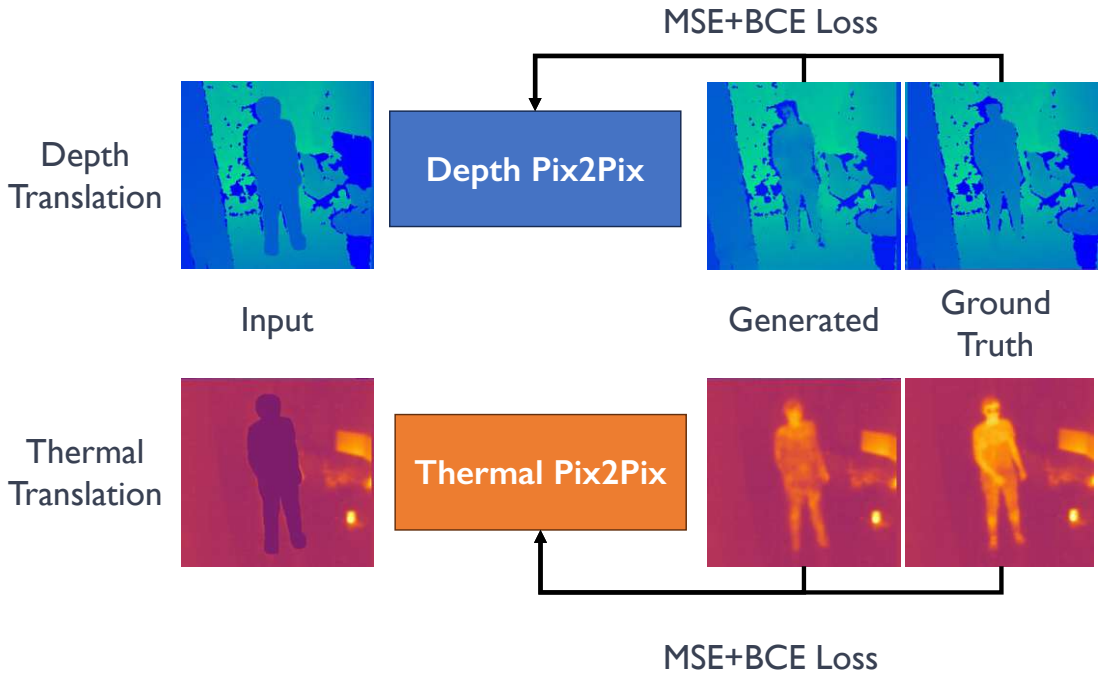


Figure 4.1: Visualization of our methodology's input and output process, depicting the transformation from RGB to depth and thermal modalities.

To translate depth and thermal images, we perform the following steps:

1. We preprocess our TRISTAR dataset to contain pairs of normal frames and frames with removed humans, as shown on the left in Figure 4.1.

2. We train UNet models to translate the frames with removed humans to the original thermal and depth images.

3. During inference, one can manually select appropriate background thermal and depth images and use human masks obtained from static labeled RGB videos to translate labeled RGB to labeled depth and thermal datasets.

The following section explains the preprocessing, inpainting, and architecture of the UNet.

4.1.1 Dataset Preprocessing

The process begins with the TRISTAR dataset [SHK23], which serves as the basis for creating input-output pairs of images. These pairs are essential for training our mapping model.

Initially, we extract squared bounding boxes from the segmentation masks. Following the bounding box extraction, we perform a mask dilation using an 8×8 kernel. This dilation process captures the immediate context surrounding the human figures. Once we have these dilated masks, we modify the corresponding depth and thermal frames. The modification involves setting the pixels within these dilated masks to a mean value representative of each modality. Effectively, this step 'erases' the human figures from the frames, leaving us with a neutral background primed for the inpainting process. Finally, we crop and resize both the edited and the original frame to the previously obtained bounding box. Details can be observed in Algorithm 1.

Algorithm 4.1: Transform Frame Image

Result: Transform frame image based on mask

Function TransformFrame(F, M);
$Y, X \leftarrow \text{Where}(M == 1)$;
if $Y.size == 0$ ***or*** $X.size == 0$ **then**
  return CenteredCrop(F), CenteredCrop(F);
end
$x_{\min}, x_{\max} \leftarrow \min(X), \max(X)$;
$y_{\min}, y_{\max} \leftarrow \min(Y), \max(Y)$;
$\Delta x \leftarrow \text{Int}((x_{\max} - x_{\min}) \times \rho)$;
$\Delta y \leftarrow \text{Int}((y_{\max} - y_{\min}) \times \rho)$;
$x_{\min} \mathrel{-}= \Delta x$;
$x_{\max} \mathrel{+}= \Delta x$;
$y_{\min} \mathrel{-}= \Delta y$;
$y_{\max} \mathrel{+}= \Delta y$;
$L \leftarrow \max(x_{\max} - x_{\min}, y_{\max} - y_{\min})$;
$C_x, C_y \leftarrow (x_{\min} + x_{\max})//2, (y_{\min} + y_{\max})//2$;
$x_{\min}, x_{\max} \leftarrow C_x - L//2, C_x + L//2$;
$y_{\min}, y_{\max} \leftarrow C_y - L//2, C_y + L//2$;
$F_{\text{mod}} \leftarrow F.copy()$;
$F_{\text{mod}}[M == 1] \leftarrow \text{Int}(\mu_{\text{frame}})$;
$F_{\text{out}} \leftarrow \text{CropAndResize}(F_{\text{mod}}, x_{\min}, x_{\max}, y_{\min}, y_{\max})$;
$F_{\text{gt}} \leftarrow \text{CropAndResize}(F, x_{\min}, x_{\max}, y_{\min}, y_{\max})$;
return $F_{\text{out}}, F_{\text{gt}}$;

- F: The original frame image to be transformed.
- M: The mask indicates regions of interest in the frame.

- Y, X: Arrays of y and x indices where the mask is 1 (region of interest).
- $x_{\min}, x_{\max}, y_{\min}, y_{\max}$: The minimum and maximum x and y coordinates of the masked region.
- $\Delta x, \Delta y$: Padding added to the x and y dimensions of the cropped region, calculated as a percentage (ρ) of the region's width and height.
- ρ: Padding ratio.
- L: The side length of the square cropping region is determined by the larger dimension of the masked area after padding.
- C_x, C_y: The center coordinates of the cropping region.
- F_{mod}: The modified frame where the masked region is set to a constant value μ_{frame}.
- μ_{frame}: Mean frame value or a predetermined constant to fill in the masked region.
- F_{out}: The cropped and resized output frame based on the modified frame.
- F_{gt}: The ground truth cropped and resized frame from the original frame.

Figure 4.2 visualizes the stages of the algorithm.

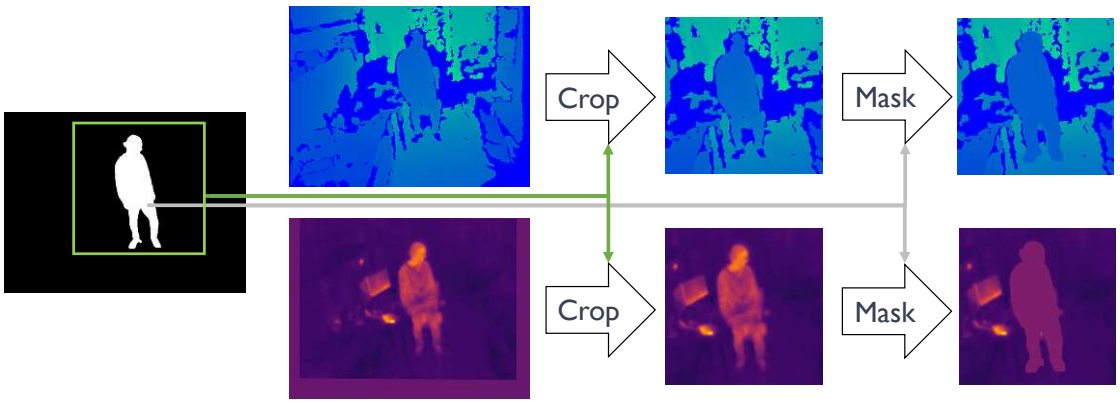


Figure 4.2: Illustration of the preprocessing algorithm applied to frame images. The process involves identifying and transforming regions of interest within the frame based on the mask.

4.1.2 Translation Architecture

The process architecture draws inspiration from the Pix2Pix framework [IZZE17], a framework for image-to-image translation tasks. Pix2Pix's adversarial training approach, which employs a conditional generative adversarial network (cGAN), is particularly suited

for tasks that aim to generate images indistinguishable from authentic images in a target domain. This framework is chosen for its ability to learn a mapping from input to output images and to model the loss function necessary to train this mapping, making it highly applicable for thermal and depth inpainting tasks.

In our implementation, we utilize two separate UNets for thermal or depth inpainting. The UNets are components of the Pix2Pix architecture, consisting of convolutional and deconvolutional blocks designed for feature extraction and image reconstruction. The strength of UNets lies in their architecture, which enables precise localization and the use of context information, which is essential for tasks like inpainting.

The Pix2Pix framework, particularly with its use of UNets, differs from other image-to-image translation methods, such as those based on variational autoencoders (VAEs) or standalone GANs. While VAEs are excellent for generating new images, they may lack the precision of detail that UNets provide. Standalone GANs, on the other hand, may not always ensure spatial consistency, which is crucial in tasks like inpainting or modality translation. The Pix2Pix framework, with its conditional GAN setup, ensures that the generated images are realistic and aligned spatially and contextually with the input images.

4.1.3 Discriminator Loss

Training our network involves a two-phase approach, each phase targeting a specific aspect of the model's performance. Initially, we optimize the network using Mean Squared Error (MSE) loss. This stage is necessary for ensuring that our model's output aligns closely with the target images pixel-wise.

In the subsequent phase, we introduce a discriminator based on the PatchGAN architecture to further refine the model's outputs. The choice of PatchGAN is deliberate: it assesses the authenticity of local image patches, making it adept at detecting finer details and textures that contribute to the overall realism of the image. This discriminator learns to distinguish between real and generated images, providing an adversarial component that encourages the model to produce more realistic and convincing outputs.

Therefore, the final model output is a blend of fidelity to the original dataset, ensured by MSE loss and enhanced image quality, by BCE loss from the discriminator.

4.1.4 Preliminary Result

Figure 4.3 visualizes the result of the Pix2Pix translation.

This approach is presented as a peer-reviewed paper at AAPR 2023 [SHSK23]. While our inpainting approach generates similar heating signature and depth values where it should, the figure clearly outlines room for improvement. Additionally, for inference, the background has to be selected manually.

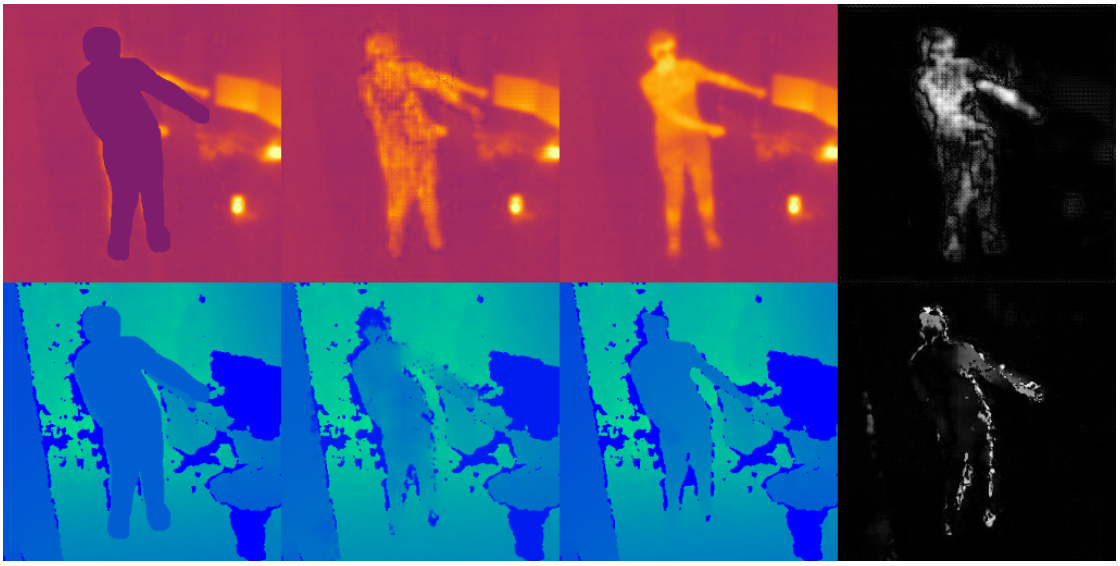


Figure 4.3: Illustration of the mapping process from RGB to depth and thermal modalities, showcasing conditional input, inpainted output, ground truth, and error analysis.

4.2 Multimodal Input

The single modality approach encounters limitations that can impact the overall performance and applicability of the model:

- **Limited Contextual Information:** Single modality inputs, such as depth or thermal data alone, provide a restricted scene view. This limitation can lead to inaccuracies, as the model might miss out on vital contextual cues.
- **Inference Challenges:** Models trained on single modality data can struggle with complex inference tasks, particularly in diverse and dynamic real-world settings. This is due to the lack of diverse information that other modalities could provide.
- **Manual Intervention:** The reliance on manual processes, such as selecting the background, poses a challenge for scalability and automation, limiting the model's practicality in various applications.

We propose a second image-to-image translation methodology that transforms RGB into corresponding depth and thermal data to address these issues, effectively bridging the gap between these modalities. The core idea is to condition the UNet on multiple inputs instead of one:

- Depth or Thermal Background
- Cropped and masked RGB frame

- Signed Distance Function (SDF)

The process involves two primary stages: First, matching backgrounds are located using [GENL+23], and human masks are segmented by integrating YOLOv6 [LLJ+22], a state-of-the-art object detection model, with the Segment Anything Model (SAM) [KMR+23]. ImageBind is a technique that aligns various modalities into a unified embedding space. This embedding is used for identifying matching backgrounds across thermal and depth frames.

Second, we use a Pix2Pix model to create accurate thermal and depth translations conditioned on the background, signed distance function, and cropped RGB and masked RGB. Then, we use a custom algorithm to merge the cropped inpainted frame with the original background frame.

Figure 4.4 provides a high-level overview of our pipeline.

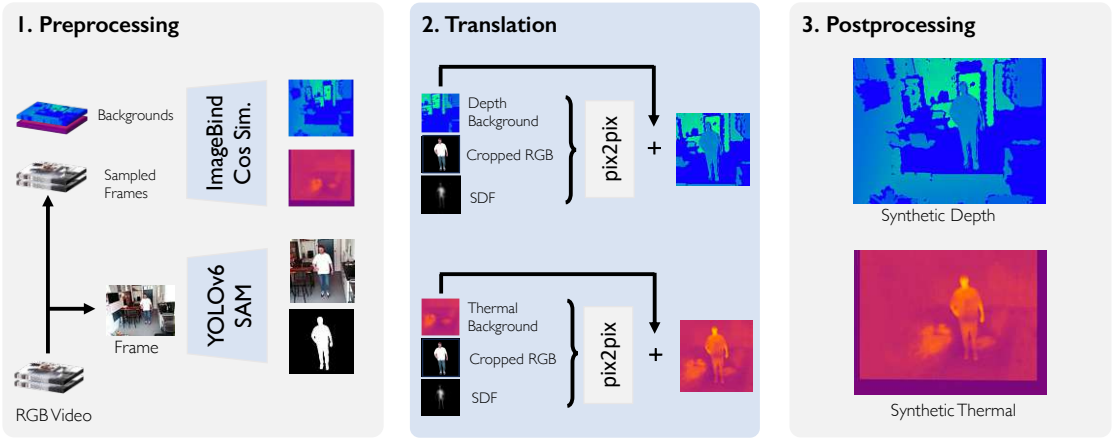


Figure 4.4: Overview of our proposed methodology, illustrating the integration of ImageBind [GENL+23] for obtaining matching backgrounds, YOLOv6 and Segment Anything Model for segmenting human masks from RGB, and Pix2Pix for modality translation from [SHSK24].

This method improves traditional image inpainting techniques because additional information is given with different inputs. The following sections explain each step in detail.

4.2.1 Locate Background

We utilize ImageBind [GENL+23], a model capable of learning a joint embedding across multiple modalities, to locate background frames in thermal and depth that closely match the given RGB data. Formally, let $f_{\text{RGB}}(I)$ and $f_{\text{thermal}}(T)$ denote the functions that compute embeddings for an RGB image I and a thermal image T respectively. For a set of RGB images from the same sequence $\{I^1, I^2, \ldots, I^n\}$, embeddings are given by $E^i_{\text{RGB}} = f_{\text{RGB}}(I^i)$. The process for thermal embeddings E^j_{thermal} follows similarly.

To calculate similarity measures for an RGB image I^i to a set of thermal images $\{T^1, T^2, \ldots, T^m\}$, we employ the cosine similarity $S_C(A, B)$ and compute a score vector as:

$$\mathbf{S}^i_{\text{thermal}} = \begin{bmatrix} S_C(E^i_{\text{thermal}}, E^1_{\text{RGB}}) \\ S_C(E^i_{\text{thermal}}, E^2_{\text{RGB}}) \\ \vdots \\ S_C(E^i_{\text{thermal}}, E^m_{\text{RGB}}) \end{bmatrix} \tag{4.1}$$

Where the cosine similarity is defined as

$$S_C(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{4.2}$$

Since all computed similarities originate from the same RGB sequence, we compute the average over the cosine similarity between each background embedding and RGB image. An average score vector is obtained by aggregating scores for multiple RGB images:

$$\bar{\mathbf{S}}^i_{\text{thermal}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}^i_{\text{thermal}} \tag{4.3}$$

The index of the background closest to our RGB images is determined by

$$\text{index}_{\text{thermal}} = \arg\min_j (\bar{\mathbf{S}}^j_{\text{thermal}}) \tag{4.4}$$

The same approach is used for finding matching depth or other domain backgrounds.

4.2.2 Obtain Human Masks

For obtaining human segmentation masks from RGB images, our methodology utilizes a combination of YOLOv6 [LLJ+22], a powerful object detection model, and the SAM [KMR+23], a versatile segmentation tool. Specifically, YOLOv6 is first employed to accurately detect human figures within the RGB images, which are then precisely segmented using the SAM in the cropped regions identified by YOLOv6. This combination is specifically chosen for its effectiveness in accurately detecting and segmenting human figures from diverse backgrounds in RGB images. Moreover, the flexibility of this approach allows for easy adaptation to various contexts beyond HBA, such as animal behavior analysis, by simply replacing the object detector with one that is more suited to the new subject matter.

4.2.3 Pre-Process Data

Figure 4.5 displays extracting and cropping the RGB image and preparing the normalized Signed Distance Field (SDF).

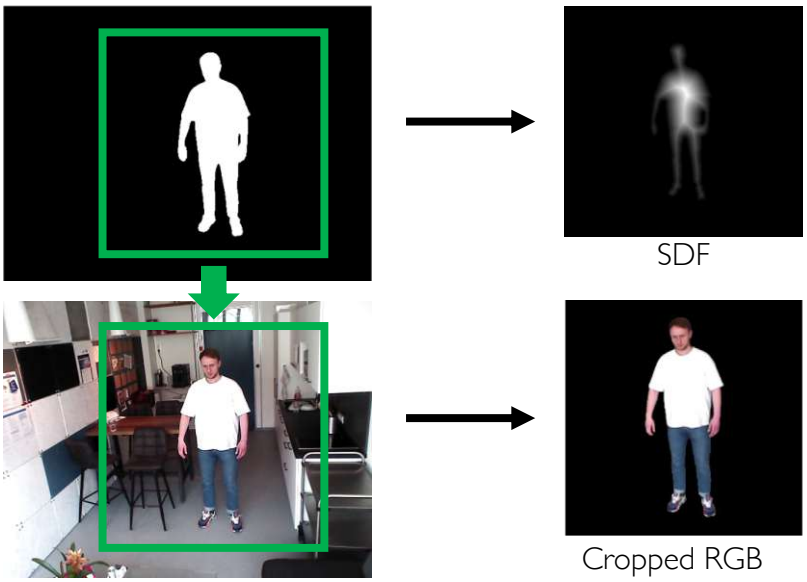


Figure 4.5: Visualization of extracting RGB and the normalized Signed Distance Field.

The preprocessing begins with creating an SDF based on the previously extracted human masks. The SDF indicates the distance of each pixel to the nearest boundary of the human subject, with negative values inside and positive values outside the subject. In our approach, we invert the SDF, setting negative values to zero, and apply min-max normalization. This inversion and normalization are crucial as they enhance the model's ability to discern spatial relationships within the image, particularly between the person and their surrounding environment.

Subsequently, we extract the subject from the RGB image using a bounding box slightly larger than the binary mask. This extraction process is replicated for both the background and the SDF. Additionally, we apply a masking technique to remove the background from the RGB image, focusing the model's attention solely on the human figure. This step, referred to as "Cropped RGB" in our translation process, significantly simplifies the translation task and contributes to the stability of the model's training.

In the final preprocessing step, all inputs are resized to a uniform dimension of 256×256 pixels. This standardization is essential for maintaining consistency across the dataset and ensuring compatibility with our translation models, which are optimized for inputs of this specific size.

4.2.4 Translate Data

The core of our data translation process is a network architecture inspired by the Pix2Pix framework [IZZE17]. Our approach, however, diverges from the conventional usage of Pix2Pix, as we focus on translating between modalities rather than within a single modality. To achieve this, our network utilizes a unique five-channel input comprising the RGB data, depth or thermal background, and the normalized Signed Distance Field (SDF). This combination is specifically tailored to facilitate the translation of RGB data into depth or thermal modalities.

Our backbone architecture is a standard UNet [RFB15], enhanced with an EfficientNet-B4 [TL19] as its encoder. The choice of EfficientNet-B4 is driven by its proven efficiency and effectiveness in various image processing tasks, making it well-suited for our complex translation objectives. Figure 4.6 shows the EfficientNet-B4 architecture.

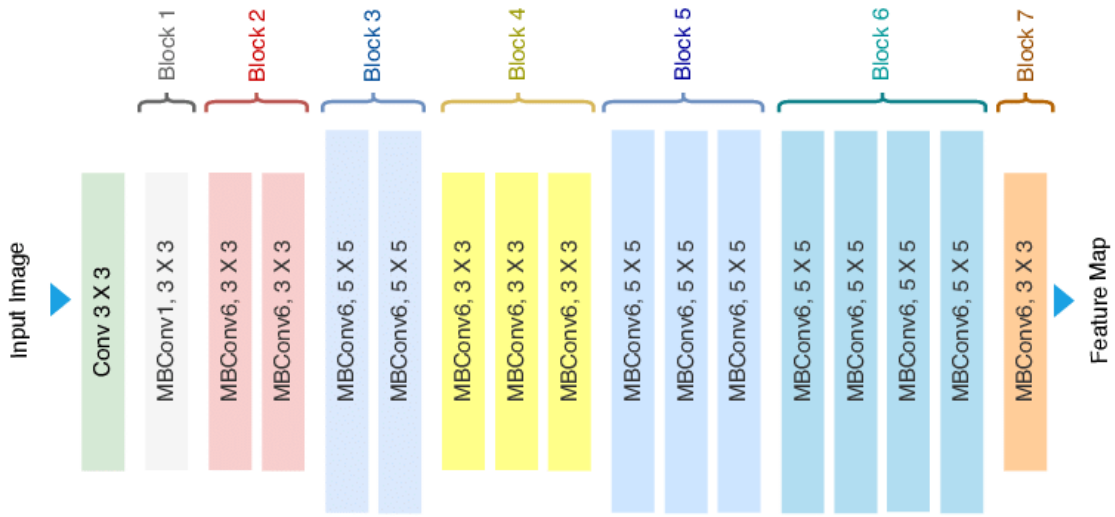


Figure 4.6: The architecture of EfficientNet-B4, highlighting its key components and efficiency in image processing [TL19].

Furthermore, we simplify the prediction task for our model by manually adding the background to the output post-prediction. This step ensures that the model primarily focuses on accurately inserting the human figure into the scene rather than reconstructing the entire background from scratch.

The training of our network is guided by a joint objective that involves minimizing both the L1 loss and the BCE loss. We decide to choose the L1 loss over L2 loss in our network training due to its effectiveness in preserving image details and avoiding blurred outputs. L1 loss is more robust to outliers and emphasizes absolute differences, ensuring small discrepancies significantly impact the loss. This leads to sharper and more detailed image generation. The L1-loss is calculated about the ground truth thermal or depth frame, while the BCE-loss is derived from a PatchGAN discriminator as described

in [IZZE17]. The PatchGAN discriminator evaluates local image patches, providing an adversarial challenge to our model. This encourages the model to focus on minimizing the L1-distance error and significantly enhancing the perceptual quality of the translated images. By addressing both these aspects, we aim to ensure that the translated images are accurate in terms of content and visually convincing and realistic.

4.2.5 Post-Process Data

The concluding stage of our pipeline is the post-processing of data, focusing on integrating translated cropped subject images into their original backgrounds. This step ensures the seamless blend of the translated images within the original context, thus preserving the scene's naturalness and coherence. The post-processing procedure is done in the following five stages:

1. **Dilation of the Original Mask:** Initially, the original mask is dilated using an 8×8 kernel. This expansion aids in preparing the mask for the subsequent blending phase, ensuring a smooth transition between the translated image and the original background.

2. **SDF Computation:** A dilated mask SDF is computed. Within the original mask, SDF values are set to zero to focus exclusively on the border areas for blending.

3. **Inversion and Normalization of the SDF:** The computed SDF is inverted and normalized by dividing by its maximum value, transforming it into a format suitable for the blending process.

4. **Alignment and Extraction of Masks and Translated Images:** The masks and translated images are adjusted to align with the original image dimensions. The translated and interpolated masks are extracted accordingly.

5. **Merging of Translated Image into the Original Image:** Finally, the translated image is merged into the original image. Pixels within the original mask are directly replaced, and weighted blending is performed at the border using the normalized SDF values to ensure a smooth transition.

The use of a normalized SDF and careful mask alignment ensures seamless transitions between frames. The final merging step, combining pixel replacement and weighted blending, maintains the natural aesthetics of the scene.

4.3 Results

The effectiveness and accuracy of our translation method are shown in Figure 4.7. These results demonstrate the model's capability to accurately transform RGB images into

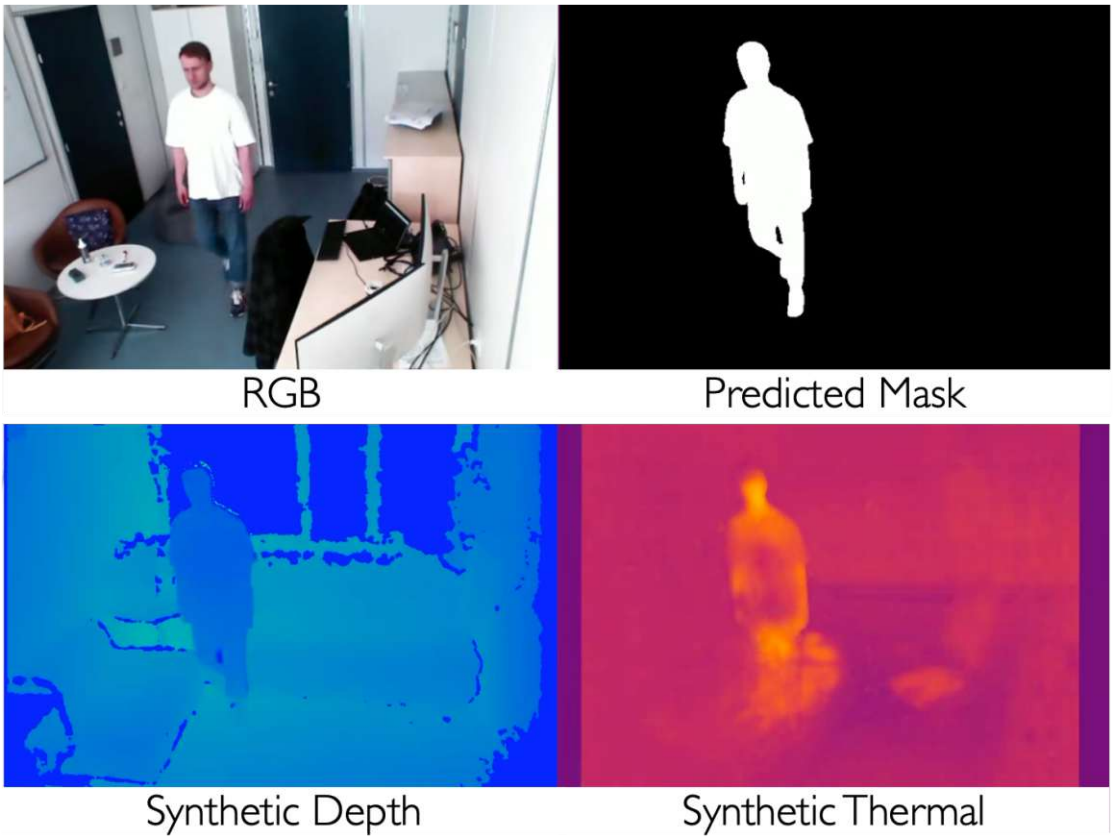


Figure 4.7: Illustration of the RGB to depth and thermal data transformation quality.

depth and thermal data, particularly highlighting the detailed rendering of human features in these modalities.

For example, the face is visible within the thermal frame instead of the single modality. The synthetic thermal data clearly shows that the head is significantly hotter than the upper body, which is typical for thermal images of humans.

To summarize, our final approach contributes in several significant ways:

1. Simplification of the prediction task using depth and thermal backgrounds, which aids in isolating the subject of interest from its surroundings.

2. Implement an adjusted Pix2Pix framework to ensure precise translation of subject details from RGB to the target modality.

3. Interpolation between generated details and prepared backgrounds, facilitating seamless integration of the subject into the new modality.

4. Conditioning on the cropped and masked RGB image, providing pixel-wise information of object surface characteristics for more accurate temperature value prediction.

5. Use a normalized signed distance function (SDF) to add spatial context to the translations.

Details are described in Algorithm 2.

Algorithm 4.2: Adjust and Integrate Translated Frame

Result: Adjust and integrate translated frame to create final frame

Function AdjustAndIntegrateFrame($D, T, M, y_{\min}, y_{\max}, x_{\min}, x_{\max}$);
$H, W \leftarrow \text{Shape}(M)$;
$O_D, O_T \leftarrow \text{Copy}(D), \text{Copy}(T)$;
\# Preparation steps;
$K \leftarrow \text{Ones}((8, 8), \text{bool})$;
$D_M \leftarrow \text{BinaryDilation}(M, \text{structure} = K)$;
$I_M \leftarrow D_M$ **and not** M;
\# Compute SDF for dilated mask and normalize it;
$S \leftarrow \text{ComputeSDF}(D_M)$;
$S[M] \leftarrow 0$;
if $Max(S) \neq 0$ **then**
  $S \mathrel{/}= \text{Max}(S)$;
end
\# Adjust masks and SDF to align with the generated images;
$y_{\min} \leftarrow \max(0, y_{\min})$;
$x_{\min} \leftarrow \max(0, x_{\min})$;
$y_{\max} \leftarrow \min(H, y_{\max})$;
$x_{\max} \leftarrow \min(W, x_{\max})$;
$T_M \leftarrow M[y_{\min} : y_{\max}, x_{\min} : x_{\max}]$;
$T_I \leftarrow I_M[y_{\min} : y_{\max}, x_{\min} : x_{\max}]$;
\# Apply translations and integrations;
$O_D[M] \leftarrow D[T_M]$;
$O_T[M] \leftarrow T[T_M]$;
$O_D[I_M] \leftarrow (\text{Blend}(O_D, D, S))[I_M]$;
$O_T[I_M] \leftarrow (\text{Blend}(O_T, T, S))[I_M]$;
return O_D, O_T;

- D: Depth frame to be adjusted and integrated.
- T: Thermal frame to be adjusted and integrated.
- M: Binary mask indicating regions of interest.

- H, W: Height and width of the mask.
- O_D, O_T: Original depth and thermal frames, respectively.
- K: Kernel used for binary dilation.
- D_M: Dilated mask obtained from binary dilation of M.
- I_M: Interpolated mask, derived from the difference between the dilated mask and the original mask.
- S: Signed Distance Field (SDF) computed from the dilated mask.
- T_M: Translated mask, cropped based on the bounding box coordinates.
- T_I: Translated interpolated mask, cropped similarly to T_M.

CHAPTER 5

Action Recognition and Segmentation

Our evaluation of the ablation studies and downstream tasks of our dataset and synthetic dataset pipeline is organized into a series of tasks that demand different deep learning architectures. The assessment of modalities and synthetic datasets is concentrated on the following tasks:

- **Human Mask Segmentation:** This process involves identifying and isolating the human figures from the background in each video or image frame. Using varying models, the algorithm detects the outline of human subjects and segments them, effectively separating them from other elements in the scene. The accuracy of human mask segmentation directly impacts the performance of subsequent analyses.

- **Temporal Action Recognition:** Temporal Action Recognition refers to identifying and classifying specific actions or behaviors exhibited by subjects over time within a video sequence. Unlike static image recognition, this involves analyzing the temporal dynamics and changes in consecutive frames to recognize actions. This task is essential for applications, including video surveillance, sports analysis, and healthcare monitoring, where understanding and interpreting human actions and activities over time is necessary.

To achieve human mask segmentation, we employ U-Net [RFB15] and a pre-trained DeepLabV3[CZP+18]. For action and activity recognition, we use a 3D ConvNet [TBF+15] and a 3D ResNet .

5.1 U-Net

The U-Net architecture, initially developed for biomedical image segmentation, is a CNN known for its efficiency for segmentation tasks [RFB15]. U-Net's architecture is characterized by its symmetric U-shape, designed to efficiently capture context and localization information. The network consists of two primary paths: the contraction path (encoder) and the expansion path (decoder). Figure 5.1 shows the U-Net architecture and explains its name.

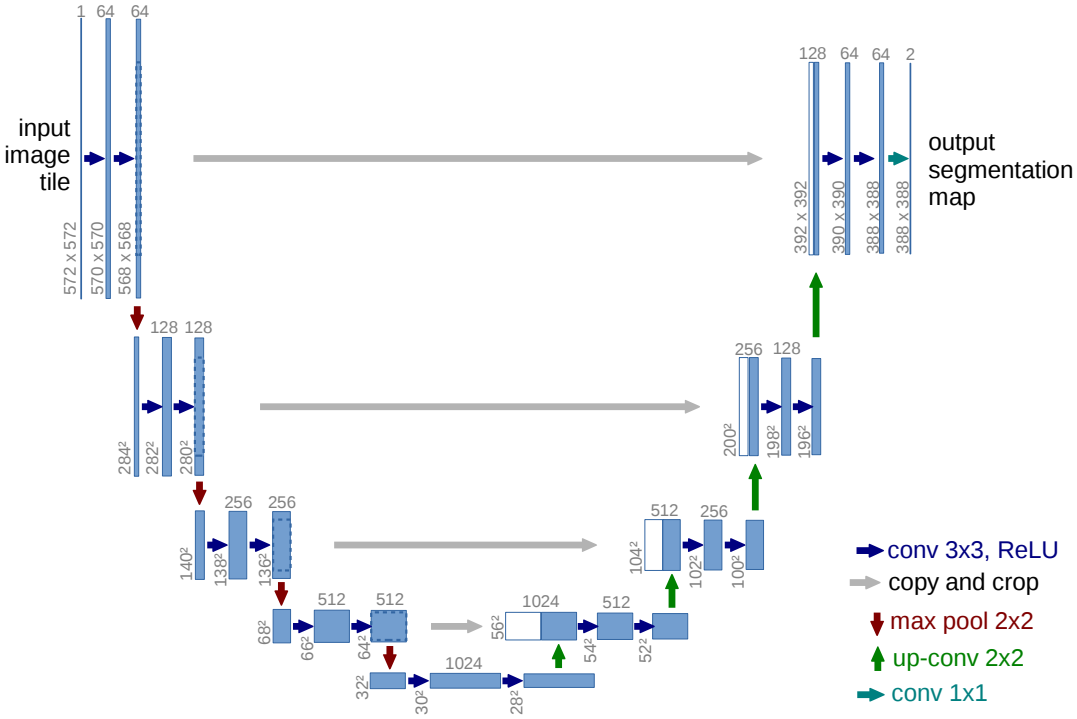


Figure 5.1: The U-Net architecture demonstrates its symmetric U-shape with an encoder (contraction path) and a decoder (expansion path) from [RFB15].

The contraction path follows the typical architecture of a CNN. It comprises repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. With each downsampling step, the network doubles the number of feature channels.

The expansion path involves upsampling the feature map and a 2×2 up-convolution that halves the number of feature channels. This is followed by a concatenation with the correspondingly cropped feature map from the contraction path and two 3×3 convolutions, each followed by a ReLU. This path increases the resolution of the output.

A key feature of the U-Net are the skip connections that connect the contraction path to the expansion path. These connections help the network localize and learn representations for segmentation by combining general and local features.

The final layer of the U-Net is a 1×1 convolution that maps each feature vector to the desired number of classes. In the case of binary segmentation, it commonly maps to one channel with sigmoid, where 0 marks the background and 1 the object, or two channels that represent the object and the background.

For our evaluation, we utilize a U-Net to segment human masks from images. The network's ability to effectively capture context and detailed localization information makes it well-suited for this task.

5.2 DeepLabV3

DeepLabV3 is an advanced semantic segmentation model. It is designed to segment complex images into specific classes efficiently and was introduced by Chen et al. [CZP+18]. DeepLabV3 performs superior to U-Nets, especially when pretrained on extensive datasets like COCO [LMB+14]. Figure 5.2 visualizes the DeepLabV3 architecture.

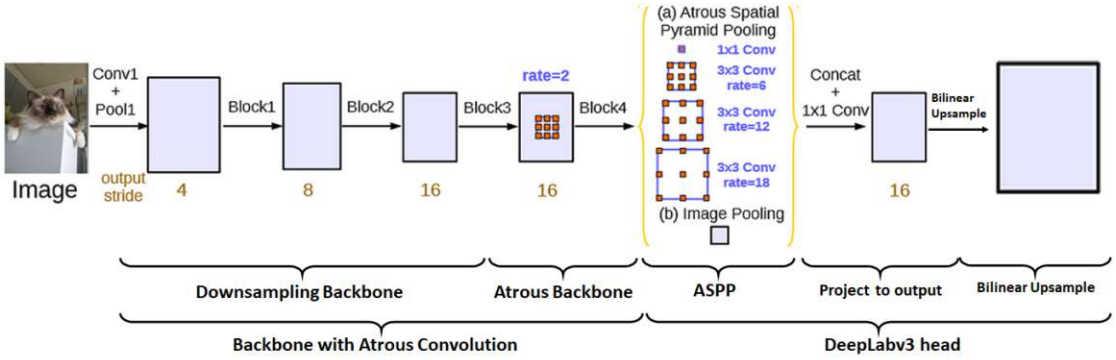


Figure 5.2: The DeepLabV3 architecture, demonstrating its key components such as Atrous Convolution and Atrous Spatial Pyramid Pooling (ASPP) from [CZP+18].

DeepLabV3 extends upon U-Nets by integrating the concept of dilated or Atrous convolutions, which enables the model to capture multi-scale contextual information without losing resolution. This is particularly beneficial for segmenting objects of different sizes in an image.

One of the distinguishing features of DeepLabV3 is the Atrous Spatial Pyramid Pooling (ASPP) module. ASPP probes an incoming feature map with filters at multiple scales, allowing the network to capture both local details and broader context.

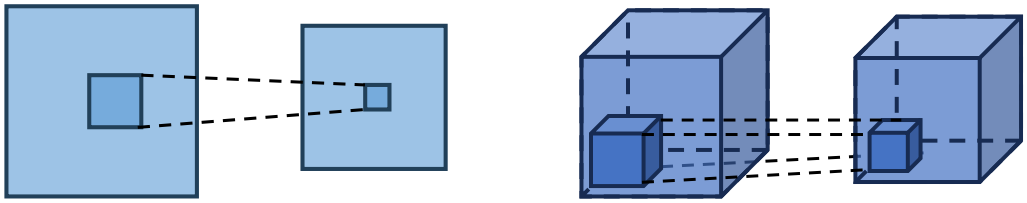
DeepLabV3 typically uses a modified Xception or ResNet as a feature extractor, pretrained on a large-scale dataset like ImageNet. This backbone is used for initial feature extraction before the ASPP module. It adopts an encoder-decoder structure, where the encoder utilizes atrous convolutions for feature extraction, and the decoder refines the segmentation results, focusing on object boundaries for more precise segmentation. DeepLabV3 is particularly effective for human segmentation tasks due to its ability to handle varied object scales and complex scenes.

We employ a DeepLabV3 model for human segmentation that is pretrained on the COCO dataset. The first layer channels of the model are modified to align with the modalities we want to use.

5.3 3D ConvNet

3D Convolutional Networks (3D CNN) have emerged as a powerful tool in computer vision, particularly for tasks that involve understanding spatial and temporal dynamics in video data [TBF+15]. In this section, we explore the architecture and application of 3D CNN in our research, mainly focusing on their role in analyzing video sequences for action recognition and other dynamic tasks.

Figure 5.3 presents a visual comparison between 2D and 3D Convolutional Networks. Specifically, Figure 5.3a showcases a 2D Convolution, where the convolutional operation is applied over two dimensions, typically height and width. This is commonly used in image processing tasks. In contrast, Figure 5.3b displays a 3D Convolution, illustrating how convolutional layers extend to three dimensions, including the temporal dimension alongside height and width. This approach is beneficial in analyzing data with a time component, such as videos.



(a) 2D Convolution (b) 3D Convolution

Figure 5.3: Comparison between 2D and 3D Convolutional Networks. Left: The structure and processing of a 2D Convolutional Network, focusing on spatial features. Right: The architecture of a 3D Convolutional Network, highlighting its ability to process both spatial and temporal dimensions in video data.

In a 3D CNN, convolutional layers apply 3D kernels to the input data. To extract features, these kernels move along the three axes: height, width, and time to pull features. This process helps capture spatial features like shapes and textures and temporal features like movements and actions. The following equation can represent the operation of a 3D convolution:

$$V'_{xyz} = \sum_{i,j,k} K_{ijk} \cdot V_{(x+i)(y+j)(z+k)} \tag{5.1}$$

where V' is the output volume, V is the input volume, K is the 3D kernel, and x, y, z are the spatial and temporal coordinates of the output volume.

Similar to 2D ConvNets, 3D ConvNets also utilize pooling layers to reduce the spatial dimensions of the feature maps, thereby reducing the computational load and the risk of overfitting. The pooling operation is also performed in three dimensions. After several convolutional and pooling layers, the network uses fully connected layers to integrate the learned features into a format suitable for classification or other high-level tasks.

For downstream action recognition tasks, we utilize 3D ConvNets due to their capability to capture and interpret temporal dynamics in videos. These networks can recognize and classify complex actions by processing sequences of frames. The ability to analyze the progression of movements over time makes 3D ConvNets particularly effective in understanding and identifying various actions and behaviors within video sequences.

The 3D ConvNet deployed in this thesis has been structured to process and analyze multimodal data. This model comprises four primary convolutional blocks, where each block consists of a 3D convolutional layer followed by a MaxPooling layer. These layers work to progressively reduce spatial dimensions while capturing higher-level features. Subsequent to these convolutional blocks, our model employs an Adaptive Average Pooling layer, which condenses the feature map into a more manageable size. Finally, the network is equipped with three distinct classifier sequences, each for specific classification tasks with varying output dimensions, employing ReLU activation, Dropout for regularization, and appropriate output layers for multilabel and mutually exclusive label classification.

5.4 3D ResNet

3D Residual Networks (3D ResNets) adapt the ResNet architecture for 3D data, such as videos or volumetric images. These networks have gained prominence for their effectiveness in handling the spatial-temporal aspects of video data, making them particularly suitable for tasks like action recognition and scene understanding in videos.

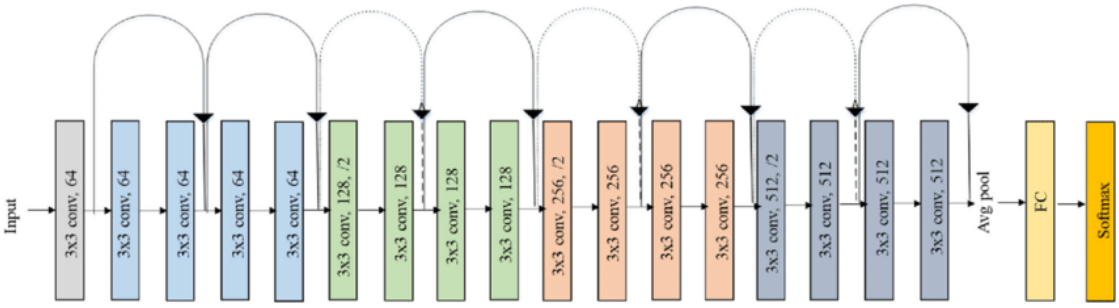


Figure 5.4: The architecture of a 3D Residual Network demonstrates its ability to process spatial and temporal information in video data [HKS17].

3D ResNets extend the concept of residual learning to three dimensions. The architecture introduces residual connections (or skip connections) in 3D convolutional neural networks. These connections allow the network to bypass one or more layers, which helps alleviate the vanishing gradient problem and enables the training of deeper networks.

In a 3D ResNet, the convolutional layers are three-dimensional, processing data across spatial dimensions (height and width) and the temporal dimension (depth or time). This approach allows the network to extract features that encapsulate spatial and temporal information, making it efficient for video analysis.

The key feature of 3D ResNets is the residual connections that skip one or more layers by performing identity mapping. These connections help preserve the information from earlier layers and reduce the training difficulties in deep networks. They also enable the network to learn more refined and complex features from the data.

Like other CNN architectures, 3D ResNets also include pooling layers for dimensionality reduction and fully connected layers for classification. The pooling layers in 3D ResNets operate in three dimensions, consolidating both spatial and temporal features.

3D ResNets, are a specialized form of 3D Convolutional Networks. They are essentially 3D ConvNets that incorporate specific layer combinations and skip connections, which enhance their ability to understand temporal sequences. We employ 3D ResNets alongside our custom 3D ConvNet to create a more comprehensive and comparable with other work.

5.5 Other concepts

Modality Fusion is a concept in multi-modal data processing, where information from different modalities (like RGB, depth, and thermal data) is combined to enhance the performance of machine learning models. Fusion can occur at different stages: early (beginning), middle, or late (end).

Early fusion combines raw data from all modalities at the beginning of the process. While it allows the model to learn from the raw combined data, it may also increase the complexity and computational cost. Our research focuses on early fusion, which directly integrates raw data.

Middle fusion involves combining features after they have been processed individually to a certain extent. This method allows each modality to be processed independently to extract relevant features before merging which can lead to a more efficient and effective learning process as the model leverages independent and combined features.

Late fusion combines the outputs of separate models for each modality at the final stage. This approach is beneficial when different modalities contribute independently to the final decision but may lack the synergy of combined feature learning.

CHAPTER 6

Evaluation & Results

In order to accurately assess the effectiveness of the modalities and datasets discussed in this thesis, a comprehensive evaluation is necessary. By focusing on action recognition and segmentation tasks, we obtain a better understanding of how these modalities and datasets perform in real-world scenarios.

The first part of the analysis is an ablation study, aimed at identifying the most effective combinations of the modalities RGB, depth, and thermal. This study aims to identify the individual and combined effects of the approaches on model performance.

Furthermore, we examine the role of a discriminator within a U-Net architecture, particularly its impact on the quality of generated outputs. This investigation includes empirical tests to determine whether the inclusion of a discriminator leads to enhanced accuracy and segmentation quality.

Another aspect of the study involves experimenting with different input conditions of the U-Net. By varying the input configurations, the goal is to find the optimal setup that enhances the network's efficiency in processing multi-modal data for segmentation tasks.

Finally, we focus on action recognition models. This evaluation is conducted in two stages: initially assessing the models' performance on real data to establish a baseline, and subsequently evaluating their effectiveness with synthetic data. This part of the evaluation tests the viability of synthetic data in training robust models and explores its potential to augment data to improve the models' generalizability and performance in diverse conditions.

6.1 Evaluation Metrics and Methodologies

In the domain of image segmentation, action recognition tasks, and image translation, several metrics are employed to assess the performance of models quantitatively.

6.1.1 Image Segmentation

Intersection over Union (IoU), also known as the Jaccard Index, is a commonly used metric for image segmentation. It measures the overlap between the predicted segmentation and the ground truth, quantitatively assessing the model's accuracy. IoU is defined as the size of the intersection divided by the size of the union of the two sets (predicted segmentation and ground truth).

Given a predicted segmentation mask P and a ground truth mask G, the IoU is calculated as:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{6.1}$$

$|P \cap G|$ represents the intersection, common area, of the predicted and ground truth masks, and $|P \cup G|$ denotes their union, the total area covered by both masks.

The IoU score ranges from 0 to 1, where 1 indicates perfect overlap, and 0 indicates no overlap. Higher IoU scores correspond to more accurate segmentation. This metric is useful in scenarios where the balance between precision, i.e., how much of the predicted segmentation is relevant, and recall i.e., how much of the appropriate segmentation was predicted, are crucial.

6.1.2 Image Generation

The Frechet Inception Distance (FID) is a widely used metric for evaluating the quality of images generated by machine learning models, particularly generative models like Generative Adversarial Networks (GANs). FID measures the distance between feature vectors calculated for authentic and generated images, typically extracted using the InceptionV3 model [HRU+17, SVI+16].

The FID score is calculated as follows:

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{6.2}$$

Where:

- μ_r and Σ_r are the mean and covariance of the feature vectors extracted from the real images.
- μ_g and Σ_g are the mean and covariance of the feature vectors extracted from the generated images.
- Tr denotes the trace of the matrix, which is the sum of the elements on the main diagonal.

Lower FID scores indicate that the generated images are more similar to the real images, implying better quality. This metric is particularly useful for quantitatively assessing

the performance of generative models as it captures the similarity in terms of both the content and the style of the images.

The InceptionV3 model, used in computing FID, is a deep CNN with high accuracy in image classification tasks [SVI+16]. By using this model to extract feature vectors, FID leverages its ability to distinguish detailed and semantic features in images.

Similar to FID, the Kernel Inception Distance (KID) measures the similarity between real and generated images. However, KID employs a kernel-based approach to calculate the distance between the distributions of features extracted from real and generated images using the Inception model [BSAG18].

The KID score is computed using the following approach:

$$\text{KID} = \mathbb{E}_{x,x' \sim P_r}[k(x, x')] + \mathbb{E}_{y,y' \sim P_g}[k(y, y')] - 2\mathbb{E}_{x \sim P_r, y \sim P_g}[k(x, y)] \tag{6.3}$$

where:

- P_r and P_g are the distributions of features, based on the Inception model, for real and generated images.
- $k(\cdot, \cdot)$ is a characteristic kernel function, such as a polynomial kernel.
- $\mathbb{E}$ denotes the expectation.

Unlike FID, which uses a linear layer of the Inception model, KID leverages a kernel function to compute the mean embedding of the feature vectors in a reproducing kernel Hilbert space. This makes KID more robust to outliers and variations in sample size. Lower KID scores indicate more significant similarity between the distributions of real and generated images, suggesting higher quality of the generated images.

The kernel-based approach in KID allows for a comparison between the distributions of features, providing a reliable measure of the quality of generated images, especially in scenarios where robustness to outliers and sample size variations is an issue.

6.1.3 Action Recognition

In action recognition, especially in the context of per-frame multi-label classification, evaluating models' performance requires using specific metrics. These metrics are Accuracy, Recall, Precision, and F1 Score.

Accuracy in this context refers to the proportion of correctly identified labels for each frame. It measures the model's overall correctness across all classes and labels within each frame. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correctly Predicted Labels}}{\text{Total Number of Labels}} \tag{6.4}$$

Recall, or Sensitivity, assesses the model's ability to identify all relevant instances of a label per frame correctly. It is useful for understanding how well the model captures the occurrence of specific actions within each frame:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6.5}$$

Precision evaluates the proportion of predicted labels for a particular action that is correct, reflecting the model's ability to accurately recognize actions in each frame without overgeneralizing:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6.6}$$

The F1 Score is beneficial in the context of multi-label classification as it provides a balance between Precision and Recall. It is the harmonic mean of these two metrics, offering a single measure to assess the model's accuracy in identifying multiple labels per frame:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.7}$$

These metrics collectively provide a comprehensive evaluation of a model's performance in action recognition tasks, where correctly identifying and classifying each action and activity in every frame is necessary.

6.2 Benchmarking TRISTAR

Our comprehensive evaluation process focused on training a U-Net and DeepLabV3 for human segmentation and an action detection model using RGB, thermal, and depth modalities. To create a fair comparison across these different modalities, we standardized each modality's input using z-normalization based on the respective training set's mean and variance. We then update the size of each model's input channels to fit the combination of modalities. For instance, when combining depth and thermal data, the input channel size was is set to two, whereas utilizing all modalities expands the input channel size to five.

6.2.1 Split

Our model's benchmarking involves dividing our dataset into training, validation, and test splits based on the shot level rather than the frame level. This ensures that all frames within a single shot are exclusively assigned to one of these sets, eliminating the risk of information leakage. We also minimize overlaps in offices and subjects within the different shots. The data is distributed as 63.77%for training, 18.23% for validation, and 18.00% for the test set, corresponding to 9,959 frames in training, 2,848 in validation, and 2,811 in testing.

6.2.2 Human Segmentation

For the task of human segmentation, we employ an U-Net [RFB15] and DeepLabv3 [CZP+18] architectures. We use a DeepLabv3 model pretrained on the COCO dataset, then fine-tune it on TRISTAR. An early fusion technique is implemented, normalizing the input frames from each modality and concatenating them to form the multi-modal input. The DeepLabv3 model, designed initially for RGB inputs, is adapted to include input channels for thermal and depth modalities. In scenarios including the RGB modality, we replicate the original RGB weights in the new input layer. For additional modalities we duplicate the R channel. After ten epochs of training at a learning rate of 0.0001, the model with the lowest validation loss is selected for testing. Surprisingly, as shown in Tables 6.1a and 6.1b, the best performance is achieved by excluding the RGB modality, highlighting the importance of the thermal modality in environments with RGB clutter.

Table 6.1: Results for Segmentation using U-Net and DeepLabv3 on the test set. The input layers channel are updated to accommodate the concatenation of the models.

(a) Results for U-Net.

RGB	Depth	Thermal	Loss	IoU
–	–	✓	0.040	0.659
–	✓	–	0.055	0.580
–	✓	✓	**0.020**	**0.775**
✓	–	–	0.147	0.356
✓	–	✓	0.062	0.673
✓	✓	–	0.071	0.553
✓	✓	✓	0.025	0.726

(b) Results for DeepLabv3.

RGB	Depth	Thermal	Loss	IoU
–	–	✓	0.041	0.660
–	✓	–	0.045	0.622
–	✓	✓	**0.023**	**0.806**
✓	–	–	0.050	0.586
✓	–	✓	0.041	0.670
✓	✓	–	0.086	0.494
✓	✓	✓	0.048	0.619

6.2.3 Action Recognition

Our approach to temporal action detection is based on the method presented in prior research. We use the same early fusion technique as in the segmentation task for the multi-modal inputs. Initialized with random weights, the model processes sets of eight frames at a time, with the first seven providing context and the eighth frame as the prediction target. The architecture includes four 3D convolution pooling blocks with ReLU activation for feature extraction, global average pooling, and two MLPs for classification. The results, detailed in Table 6.2, show that the combination of depth and thermal modalities lead to the highest performance in action classification, reinforcing the advantages of non-RGB modalities in complex scenes.

6.3 Inpainting Results

A key aspect of our evaluation involves comparing the performance of the model trained solely with MSE loss against the model trained with a combination of MSE and BCE loss.

Table 6.2: Results for Temporal Action Recognition using custom 3D Convolution Architecture on the test set.

RGB	Depth	Thermal	Loss	Accuracy	Precision	Recall
–	–	✓	2.367	0.903	0.796	0.620
–	✓	–	2.504	0.889	0.749	0.577
–	✓	✓	**2.347**	**0.907**	**0.813**	0.626
✓	–	–	2.659	0.876	0.704	0.537
✓	–	✓	2.346	0.904	0.799	0.623
✓	✓	–	2.465	0.897	0.758	**0.629**
✓	✓	✓	2.349	0.901	0.783	0.618

Where BCE is the loss created by using the discriminator. To quantitatively assess the performance, we compute the RMSE, FID, and KID. These evaluations are conducted using a pretrained Inceptionv3 model, which allows us to analyze the pixel-level accuracy and the generated images' semantic and perceptual quality. The metrics are calculated for the entire image, enabling a comprehensive analysis that includes the person's contrast against the background.

Table 6.3 presents the results of this performance comparison, highlighting the differences between the models trained with MSE loss and the combined MSE and BCE loss in terms of RMSE, KID, and FID. Lower values in these metrics indicate better performance, with particular attention given to improvements in the perceptual quality of the images as reflected in the FID and KID scores.

Table 6.3: Comparison between U-Net 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	**0.068 ± 0.002**
	FID	79.379	**63.882**
Depth	**RMSE**	0.139	**0.133**
	KID	**0.056 ± 0.002**	0.060 ± 0.003
	FID	**82.096**	84.905

It is crucial to acknowledge that in our evaluation, we utilized the latest feature layer (2048) of the Inception model, originally pretrained on RGB data. Using a model trained on a different modality than our target modalities (depth and thermal) presents a limitation in our assessment. Specifically, this mismatch can negatively impact the expressiveness and reliability of the FID and KID metrics in our context.

While powerful in evaluating image quality, the FID and KID metrics are fundamentally designed for and trained on RGB data. When applied to depth and thermal data,

their ability to capture and assess the quality of these modalities may be limited. This limitation arises because the features extracted by the Inception model are inherently attuned to the characteristics of RGB images, such as color and texture, which may not be directly applicable or fully representative of depth and thermal data.

Therefore, while our results indicate the usefulness of incorporating an additional BCE loss-particularly for the thermal modality-it is important to interpret these findings carefully. Further research and development of evaluation metrics specifically tailored for depth and thermal data would be beneficial in obtaining a more accurate understanding of model performance in these modalities.

6.4 Ablation Study Input Modalities

In our evaluation, we conduct an ablation study to determine the impact of various input combinations on the performance of our translation model. These combinations include different configurations of backgrounds, masks, and RGB images. Additionally, we extend our evaluation to include a comparative analysis, wherein we train an action recognition model on both synthetic and real data from our TRISTAR dataset [SHK23]. This dataset is integral to our evaluation, serving multiple purposes:

- Providing background thermal and depth images for conditioning our translation models.
- Supplying data for training our translation models.
- Acting as a benchmark for comparing action recognition models trained on synthetic versus real data.

6.4.1 Ablation Study Translation

Our evaluation focuses on the effectiveness of conditioning the translation model on different combinations of the following inputs:

- **Background**: If a suitable depth or thermal background is included.
- **SDF**: If the SDF is included.
- **Crop**: If a masked and cropped RGB is included.
- **Add**: If the background is added to the person at the end.

We assess the model's performance using three key metrics: FID [HRU+17], KID [BSAG18] for the semantic similarity between source and target dataset as well as MSE for the pixel similarity. For our study, we input normalized depth and thermal data into the Inception model by duplicating the frame three times to create a three-channel grayscale-like "RGB"

image. To adapt the pre-trained model to our non-RGB data, we extract lower-level features from an early layer of the model. Table 6.4 showcases the results of our ablation study for the depth modality.

Table 6.4: Results of Depth Analysis in the Ablation Study

Background	SDF	Crop	Add	FID	KID	MSE
✓	✓	✓	✓	**16.20021**	**17.91997**	0.55078
✓	✓	✓		34.29326	33.63969	**0.53175**
	✓	✓		61.38142	57.60419	1.76927
✓		✓	✓	19.01336	20.07544	0.57610
✓			✓	22.27072	23.33244	0.57251

Combining all input conditions (adding background, SDF, cropped RGB, and adding the background post-translation) yields the best performance, as indicated by the lowest FID and KID scores. This finding suggests that the translation under these conditions is more semantically similar to real images. Lower FID and KID values signify a closer alignment with the distribution of real image features. However, a slightly higher MSE score of approximately 0.55 as opposed to 0.53, though moderate, points to potential discrepancies at a per-pixel-level comparison, which may not always correlate with perceptual image quality.

Conversely, not including the background in the final translation results in a marginal increase in numerical accuracy but significantly elevates both FID and KID scores. This implies that while the model might be more precise numerically, the generated images' semantic integrity and perceptual quality are compromised.

Omitting the background altogether while retaining SDF and cropped RGB leads to the highest increases in FID, KID, and MSE scores. This drastic increase across all metrics highlights the critical role of the background in preserving both the semantic and pixel-level quality of the images.

The exclusion of certain conditions, such as SDF or cropped RGB, generally results in a moderate score increase. This indicates that each condition has a balanced impact on both the numerical accuracy and perceptual quality of the images, underscoring the importance of each element in maintaining the overall integrity and realism of the synthesized images. Table 6.5 shows the results of our ablation study for the thermal modality.

As with the depth modality, having a background for the model to condition is crucial for performance and outweighs the impact of not using the cropped RGB and SDF. The relatively stable FID and KID scores indicate this. However, the numerical performance, i.e., MSE, worsens when not using the cropped RGB and SDF. This suggests that while the overall semantic integrity of the images may be maintained, precise pixel-level accuracy is negatively affected.

Table 6.5: Results of Thermal Analysis in the Ablation Study

Background	SDF	Crop	Add	FID	KID	MSE
✓	✓	✓	✓	1.27067	0.42648	**0.35773**
✓	✓	✓		3.56303	1.78465	0.53721
	✓	✓		15.5289	9.11526	1.56718
✓		✓	✓	**1.0324**	**0.31805**	0.53392
✓			✓	1.12273	0.34114	0.55877

6.5 Action Recognition Results

Our evaluation begins with a performance comparison of models trained exclusively on either synthetic or real data. We design a 3D ConvNet for action recognition within the TRISTAR dataset, based on our proposal for benchmarking TRISTAR [SHK23]. The quantitative assessment of these models focuses on key metrics, including accuracy, precision, recall, and F1 score.

The outcomes of this comparison are summarized in Table 6.6. Our findings reveal a decrease in performance, notably a $\sim 12\%$ reduction in F1 score when models are trained solely on synthetic data compared to real data. While this reduced performance aligns with expectations, considering the inherent differences between synthetic and real data, the performance of the model trained on synthetic data is still acceptable within the context of our study's objectives.

Table 6.6: Action Recognition Performance using 3D ConvNet

Test Metric	Synthetic Data	Real Data
Accuracy	0.8669	**0.907**
F1 Score	0.5799	**0.707**
Precision	0.6449	**0.813**
Recall	0.5268	**0.626**

These results show the viability of using synthetic data for training action recognition models. While there is an observable decrease in specific metrics, the overall performance of models trained on synthetic data offers promising insights, especially considering the challenges and limitations of acquiring extensive real datasets in specific domains.

Next, we introduced a second model, designed closely in line with the ResNext architecture [HKS17]. Our initial model served as a baseline, whereas the 3D ResNet model is taken from an established, research-validated framework, offering advanced features tailored explicitly for handling complex, high-dimensional data. This model is trained using three distinct data setups: solely synthetic data, a mixture comprising 10% real and 90% synthetic data, and exclusively real data. This approach allows us to evaluate

the efficiency of our methodology in varying data environments, particularly emphasizing its potential as a data augmentation technique.

The performance of the ResNet model under these different training conditions is summarized in Table 6.7. As anticipated, the model trained exclusively on synthetic data has slightly lower performance metrics than the other setups. However, an interesting observation is that the model trained with a mix of synthetic and real data showcases performance metrics almost identical to those trained purely on real data. This finding underscores the effectiveness of our synthetic data generation methodology, particularly in scenarios where real trimodal data is limited or difficult to obtain.

Table 6.7: Action Recognition Performance using ResNet Model

Test Metric	Synthetic Data	Synthetic Augmentation	Real Data
Accuracy	0.8712	**0.90462**	0.90409
F1 Score	0.5898	0.69637	**0.69684**
Precision	0.6636	**0.78246**	0.77610
Recall	0.5307	0.62734	**0.63227**

These results illustrate the potential of synthetic data, when used in conjunction with real data, to achieve comparable outcomes to training with real data alone. It suggests that our synthetic data generation approach can be a reliable augmentation step, enhancing the training process and improving model performance in real-world applications.

Figure 6.1 shows qualitative samples of two shots from our synthetic dataset. As can be seen, our method's depth and thermal heat signatures appear very plausible.

The background of our synthetic data generation pipeline is not perfectly aligned, indicating that it may not be an exact replica of real-world scenarios. However, this approach can still be a useful augmentation step in the training process. It can help models for downstream task to focus on the person and prevent overfitting on the background.

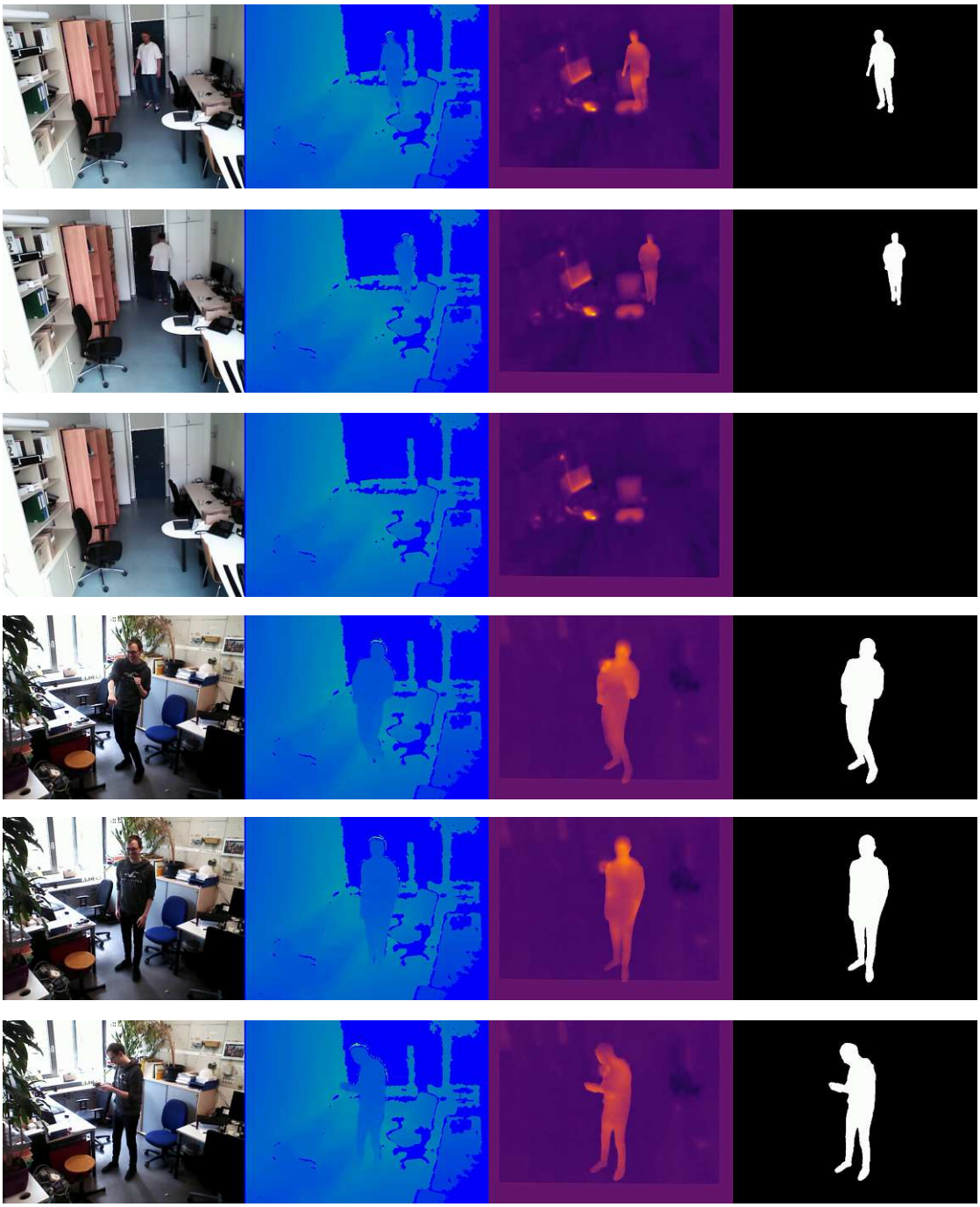


Figure 6.1: Qualitative synthetic data samples showing depth and thermal heat signatures. Despite some background discrepancies, the depth and thermal representations are convincingly realistic.

CHAPTER 7

Conclusion

This thesis explored multimodal machine learning, focusing on integrating depth and thermal for human behavior analysis.

Our work consists of dataset collection and an ablation study for various modalities, implementing novel translation methods, and developing advanced models for action recognition and segmentation. Our implementation of RGB to depth and thermal translation methods shows the potential of image translation in extending the utility of existing data. This finding is particularly noteworthy as it underscores the value of depth and thermal modalities in scenarios where RGB data may be limited or ineffective.

Moreover, our research has shown that translating RGB to depth and thermal images is feasible and practical for training models from scratch. This approach allows for the extension of training datasets using only RGB frames, which facilitates the development of models that can use depth and thermal modalities.

Additionally, we have explored using this translation methodology for data augmentation purposes. Our experiments have revealed that models trained with natural and synthetic data can perform comparably to those trained exclusively on accurate data. We have demonstrated that models perform similarly even when trained with a dataset comprising 10% accurate and 90% synthetic data.

The following list summarizes our contributions:

- Creating a new trimodal dataset with action recognition and human mask labels.
- Evaluating the modalities and showing empirically that the combination of depth and thermal modality can be superior to RGB for specific scenarios, proofing their usefulness.
- Creating a novel pipeline for RGB to depth and thermal translation.

- A detailed evaluation of the synthetically generated datasets.

In addition to our empirical findings and methodological advancements, a notable highlight of our work is the publication of three peer-reviewed papers. The first one to present TRISTAR our trimodal dataset at GCPR[SHK23], the second to show single modality translation with Pix2Pix [SHSK23] and finally the entire translation pipeline at PeRConAI [SHSK24]. Each of these papers contributes disseminates our results and contributions.

Our research has also practical implications. The methodologies developed can be leveraged in surveillance, human-computer interaction, and healthcare, understanding human behavior and actions in poorly lit or privacy-sensitive settings. In conclusion, this thesis contributes substantially to synthetic data generation and its application in understanding human behavior.

CHAPTER 8

Further Work

Our work is not without limitations. In particular, the size and variety of our dataset TRISTAR and diversity and the constraints of the translation methods employed is limiting.

One of this study's primary limitations is the dataset's limited size and scope. With around 15,618 labeled frames it is the largest trimodal dataset. However, the significant size does not encompass the vast array of human actions and interactions in more dynamic or diverse settings which are present in RGB datasets. This limitation is significant in behavior analysis and synthesis, where a broader spectrum of data is required for generalizable and accurate model training. The dataset's constraint regarding the variety of actions also poses a challenge, as it limits the model's ability to learn and replicate a broader range of human behavior. Further work should consider recording additional datasets with a greater variety of images.

Another limitation is related to the translation methods introduced. The need for suitable backgrounds in the database constrains these methods. The dependence on the availability of appropriate backgrounds restricts the versatility of the translation process, as it cannot be universally applied across all possible scenarios.

Furthermore, the assumption of a static camera in our methodology introduces limitations. The static camera restricts the model's applicability in more dynamic environments where the camera is moving. In real-world applications, especially in interactive environments, the ability to process and translate data from moving cameras can be necessary. The static camera assumption thus limits the model's practicality and adaptability in such scenarios.

These limitations highlight several areas for future research and development. Expanding the dataset to include more diverse human actions and interactions would improve accuracy and applicability.

Additionally, exploring translation methods that are less dependent on the background database and can adapt to dynamic camera movements would open up new avenues for practical applications. In terms of future work, integrating Diffusion Models could be another improvement, given their capability to generate high-quality, diverse images. This approach can significantly enhance the synthesis of realistic human actions and interactions in diverse environments. The diffusion process could replace the direct UNet approach used within our pipeline.

Moreover, a comprehensive evaluation and benchmarking against established architectures like MiDaS [RLH+20] for depth estimation or ThermalGAN [KKH+18] for thermal image synthesis would provide valuable insights.

List of Figures

List of Tables

Bibliography

[AMY18] Heba Hamdy Ali, Hossam M Moftah, and Aliaa AA Youssif. Depth-based human activity recognition: A comparative perspective study on feature extraction. *Future Computing and Informatics Journal*, 3(1):51–67, 2018.

[BBW+23] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[BEG00] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10, 2000.

[BRSB23] Martin Brenner, Napoleon H Reyes, Teo Susnjak, and Andre LC Barczak. Rgb-d and thermal sensor fusion: A systematic literature review. *arXiv preprint arXiv:2305.11427*, 2023.

[BSAG18] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DK05] J Davis and M Keck. A two-stage approach to person detectionin thermal imagery. In *Proceeding of Workshop on Applications of Computer Vision (WACV)*, 2005.

[Esc12] Sergio Escalera. Human behavior analysis from depth maps. In *International Conference on Articulated Motion and Deformable Objects*, pages 282–292. Springer, 2012.

[EVGW+10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[FAT11] Sergi Foix, Guillem Alenya, and Carme Torras. Lock-in time-of-flight (tof) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, 2011.

[GDL+16] Chenqiang Gao, Yinhe Du, Jiang Liu, Jing Lv, Luyu Yang, Deyu Meng, and Alexander G Hauptmann. Infar dataset: Infrared action recognition at different times. *Neurocomputing*, 212:36–47, 2016.

[GEC+13] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[GENL+23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[HK21a] 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.

[HK21b] 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.

[HKS17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[HRU+17] 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.

[IZZE17] 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.

[JCQ23] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023.

[JLD12] Zhuolin Jiang, Zhe Lin, and Larry Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):533–547, 2012.

[KCS+17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[KKH+18] 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.

[KMR+23] 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.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[KTK15] Priyanka Kakria, NK Tripathi, and Peerapong Kitipawang. A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors. *International journal of telemedicine and applications*, 2015:8–8, 2015.

[KZ02] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part III 7*, pages 82–96. Springer, 2002.

[LLJ+22] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[LMB+14] 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.

[MAMT15] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[MHM17] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[Mie05] Roland Miezianko. Terravic research infrared database. *IEEE OTCBVS WS Series Bench*, 2005.

[MSG+23] 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.

[NORBK17] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[PCB+16] 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.

[Pop10] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[RDGF16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[RLH+20] 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.

[SDFG17] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 585–594, 2017.

[SFC+11] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[SHK23] 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.

[SHKF12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[SHSK23] Christian Stippel, Thomas Heitzinger, Rafael Sterzinger, and Martin Kampel. From rgb to depth and thermal: Mapping between modalities to alleviate data scarcity. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2023.

[SHSK24] Christian Stippel, Thomas Heitzinger, Rafael Sterzinger, and Martin Kampel. Closing the gap in human behavior analysis: A pipeline for synthesizing trimodal data. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2024.

[SK22] Julian Strohmayer and Martin Kampel. A compact tri-modal camera unit for rgbdt vision. In *2022 the 5th International Conference on Machine Vision and Applications (ICMVA)*, pages 34–42, 2022.

[SKNF23] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.

[SLM+22] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[SLX15] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[SRZ+20] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002.

[SVI+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[SVW+16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[TL19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[WBL23] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023.

[ZCS+21] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

[ZZP+17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.