

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Deep Thermal Optical Flow

Permalink

<https://escholarship.org/uc/item/8fb1r3k8>

Author

Mohan Kumar, Harshith

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Deep Thermal Optical Flow

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Computer Science

by

Harshith Mohan Kumar

March 2025

Thesis Committee:

Professor Vishwanath Saragadam, Co-Chairperson

Professor Amit Roy-Chowdhury, Co-Chairperson

Professor Hang Qiu

Professor Greg Ver Steeg

The Thesis of Harshith Mohan Kumar is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

Acknowledgments

I would like to express my deepest gratitude to my thesis advisors, Dr. Vishwanath Saragadam and Dr. Amit Roy-Chowdhury, for their unwavering support, invaluable guidance, and insightful feedback throughout my research journey. Their mentorship has been the cornerstone of my academic and professional growth.

I am also deeply thankful to Dr. Greg and Dr. Hang for graciously serving on my thesis committee. Their exceptional teaching and mentorship have profoundly shaped my academic development and left a lasting impact on my work.

My heartfelt thanks go to my fellow students at UCR—Saketh, Arnab, Pratyay, and Alper—for their camaraderie and support. Their friendship and collaboration have made my Master’s experience truly memorable and enriching.

A special note of appreciation goes to Piyush, Pranav, and Rohit, who helped me navigate the challenges of settling in Riverside. Their kindness and support made this new environment feel familiar and welcoming.

Above all, I owe my deepest gratitude to my parents. Their unwavering belief in me, unconditional love, and countless sacrifices have been the foundation of my resilience and success. This journey is as much a testament to their support as it is to my efforts.

To my parents for all the support.

ABSTRACT OF THE THESIS

Deep Thermal Optical Flow

by

Harshith Mohan Kumar

Master of Science, Graduate Program in Computer Science
University of California, Riverside, March 2025
Professor Vishwanath Saragadam, Co-Chairperson
Professor Amit Roy-Chowdhury, Co-Chairperson

Optical flow problem focuses on predicting motion between consecutive frames, typically within the RGB domain. While RGB-based models have reached a high level of maturity, optical flow prediction in alternative modalities, such as thermal imagery, remains underexplored. Given that thermal imaging excels in challenging conditions like fog and darkness, developing a thermal optical flow model is of critical importance. We propose a first-of-its-kind data-driven approach to optical flow, with a novel cross-modal supervision. To achieve this, we collected more than a thousand, spatially aligned and synchronized RGB and thermal video sequences across diverse environments and types of motion, including humans, vehicles, animals, and nature. We then leveraged this dataset to generate pseudo-ground-truth optical flow estimates by using state-of-the-art optical flow techniques on the RGB images, and then used them to fine-tune the optical flow network with thermal images as input. Our contribution culminates in a high-quality, synchronized RGB-Thermal video dataset, with state-of-the-art baselines for optical flow in the thermal domain.

Contents

1 Introduction	1
2 Prior Work	4
2.0.1 Optical Flow	4
2.0.2 Thermal Imaging	5
3 Proposed Work	7
4 Experiments	14
5 Conclusions	18
Bibliography	19

LIST OF FIGURES

1.1 Overview. Our dataset is pivotal in distilling knowledge from RGB optical flow models into a thermal optical flow models. In contrast to RGB-based approaches that fail under night time, our thermal model leverages daytime knowledge from RGB to operate efficiently at night.....	2
3.1 Framerate influence. Predicted occlusion masks illustrate the gradual increase in motion ambiguities as the frame rate decreases from 60 to 15 frames per second.....	9
3.2 Resolution influence. Optical flow visualizations show the loss of detail in downsampled, lower-resolution input compared to the original higher- resolution input	9
3.3 Dataset Snapshot. The visualization above shows a snapshot of the diverse set of images captured by our synchronized optical setup. We captured indoor scenes, outdoor scenes, scenes with moving humans, with moving vehicles, day time, and night time. Our dataset offers a significant contribution to computer vision in the thermal wavelengths.....	10
3.4 Hardware Setup. Our optical setup consists of a FLIR Boson+ camera, and a FLIR Blackfly S camera, both colocated with a dichroic mirror that transmits visible wavelengths and reflects LWIR wavelengths. Both cameras are synchronized by an external trigger generated by an Arduino at 60 frames per second	11
3.5 Calibration Setup. (a) A halogen lamp is positioned near a calibration poster board. (b) When activated, the black squares absorb and emit more heat than the white squares, creating a distinct checkerboard pattern in the thermal image	12
4.1 Qualitative Results. RGB and LWIR frames are shown alongside the outputs of three methods: pretrained (no fine-tuning), photometric fine-tuning, and pseudo fine-tuning, with RGB ground truth provided for reference. The Charbonnier metric (lower is better) is included for quantitative evaluation. Across the examples, the pseudo fine-tuning approach demonstrates superior performance, supported by the lowest Charbonnier scores	16

LIST OF TABLES

3.1 Comparing Datasets. Key attributes include synchronization (Sync), calibration (Calib), frames per second (FPS), thermal resolution (Thermal Res.), RGB resolution (RGB Res.) in megapixels (Mp), and the number of videos (Vids). Our dataset, UCRT, stands out with the highest FPS, thermal resolution, and number of videos, making it the most comprehensive for cross-spectral optical flow tasks.....	8
4.1 Pretrained Models. Comparing optical flow at low resolution (768x512) with high resolution (1536x1024) rgb inputs for different supervised synthetic pre-trained weights across the sintel (S), KITTI (K), FlyingChairs (C) and FlyingThings3D (T) datasets	15
4.2 Quantitative Results. End-point error (EPE) and accuracy thresholds (1px, 3px, 5px) are reported for three methods: Pseudo fine-tuning, Photometric fine-tuning, and Pre Trained (no fine-tuning). Lower EPE values indicate better performance, while higher accuracy values are preferable. The Pseudo fine-tuning method achieves the lowest EPE, demonstrating its effectiveness.....	16

Chapter 1

Introduction

Thermal images enable clear imagery even in poor weather conditions, such as fog, scatters of rain, night time driving, and even visual clutter. A key component of many vision tasks is the ability to estimate dense pixel correspondences between video sequences, which can be leveraged to solve applications such as super-resolution [5, 4, 74, 65], video synthesis [21, 45, 25], and autonomous driving [18, 72, 42]. Optical flow with visible (RGB) images can be reliably and robustly estimated with state-of-the-art (SoTA) deep network based techniques. However, robustly estimating optical flow with thermal images is an unexplored and more importantly, a highly challenging task. The challenges with thermal optical flow include lack of strong spatial texture for dense pixel correspondences, and lack of ground truth optical flow, which is readily available in the visible domain, thanks to a large pool of simulated optical flow data.

In this paper, we propose a novel approach to thermal optical flow computation via a cross-modal supervision. We first identify optical flow tend to be modality invariant,

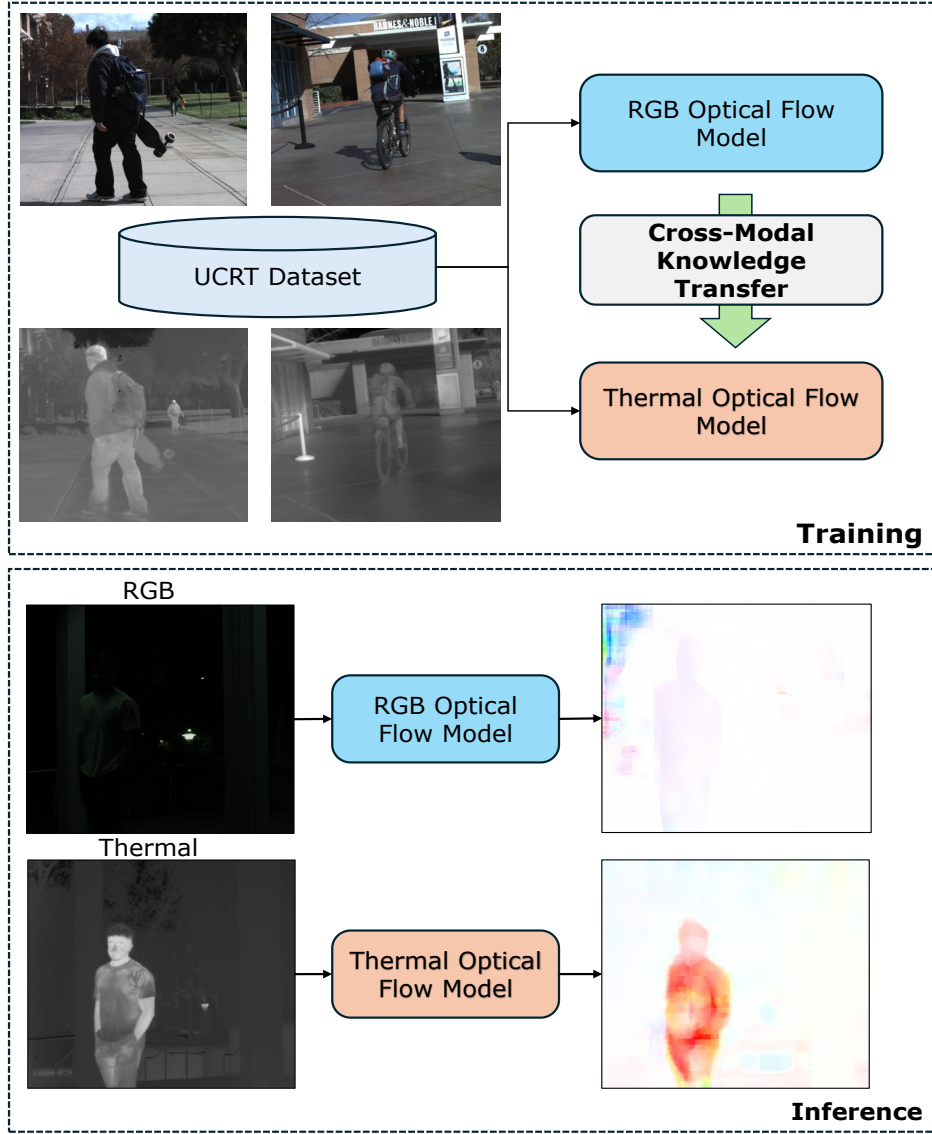


Figure 1.1: **Overview.** Our dataset is pivotal in distilling knowledge from RGB optical flow models into a thermal optical flow models. In contrast to RGB-based approaches that fail under night time, our thermal model leverages daytime knowledge from RGB to operate efficiently at night.

implying that the optical flow estimated in RGB match the optical flow estimated with thermal images. Further, we leverage the observation that optical flow can be reliably estimated on real world RGB images with SoTA deep network-based models. We can thereby leverage this optical as a ground truth, for thermal images captured at the same time as the

RGB images. We built a first-of-its kind large and diverse dataset consisting of spatially aligned and synchronized RGB and thermal videos sequences. For each data point consisting of two subsequent RGB and thermal video frames, we first generate an optical flow on the RGB image pair. We then use this optical flow as ground truth to supervise the training of a deep network that takes thermal images as input and outputs optical flow. We then fine-tuned the parameters of the deep network with a mean squared error loss between the ground truth optical flow (computed from the RGB images), and self-supervised warping loss, that enables robustness to minor differences in thermal and RGB optical flows. We provide the first thermal optical flow baseline. We demonstrate the brightness-invariance property of thermal imaging, makes them well suited for optical flow estimation particularly in scenarios where traditional visible imaging struggles. Additionally, we benchmark our method using two popular optical flow architectures: RAFT [61], a CNN-based approach, and FlowFormer++ [48], a transformer-based approach.

Our key contributions can be summarized as follows:

- Novel high-quality dataset: We introduce a new dataset of high-resolution, high-frame-rate RGB-Thermal pixel-aligned video sequences featuring over 1,000 diverse scenes and with over 22,500 aligned image pairs for robust training and evaluation.
- First thermal optical flow baseline: We propose the first network for estimating optical flow from thermal imagery.
- Cross-spectral supervision: We develop a cross-spectral supervision strategy that leverages pre-trained RGB models to generate occlusion masks and pseudo-labels, effectively bridging the domain gap between RGB and thermal modalities.

Chapter 2

Prior Work

We cover the relevant works on optical flow, primarily with RGB images, and the advantages and challenges associated with thermal images.

2.0.1 Optical Flow

Early optical flow algorithms relied on combining photometric and smoothness losses [68, 20, 14], improved later by approaches that included occlusion masks [66, 26] to enable sharper optical flow around object boundaries. Other approaches have leveraged simulated data through data augmentation [33, 31, 73, 69]. While unsupervised approaches are potentially a promising approach for dealing with data like thermal images, they have limited accuracy. This can be improved with precise ground truth from simulated data.

The paradigm of data-driven optical flow estimation shifted with FlowNet [13, 24], the first end-to-end convolution neural network that directly estimated optical flow. Subsequent innovations have introduced various inductive biases to enable robust optical flow [59, 61, 22, 48, 67, 12], all leveraging ground truth from simulated data. A large

contribution to this was the availability of optical flow datasets.

Early datasets like Middlebury and FlyingChairs focused on small-scale, synthetic scenes with simple motion patterns [2, 13]. More complex datasets such as MPI Sintel and FlyingThings3D emerged to introduce realistic rendering, diverse motion and challenging occlusions [8, 39]. To generalize applicability to real-world scenarios, datasets such as KITTI and HD1K [18, 28] were introduced. These datasets leveraged high-precision LiDAR scanners to estimate sparse optical flow ground truth. While obtaining ground truth optical flow is possible in RGB domain, owing to precise renderers, the same is nearly impossible in thermal domain due to lack of reliable simulators.

2.0.2 Thermal Imaging

Illumination invariance of thermal imaging has provided significant advantages over visible counterparts in core computer vision tasks that include material classification [47, 9], object detection [64, 56, 11, 10], reconstruction [44], depth estimation [43, 53, 52, 34, 51, 54] and tracking [7, 75]. Unlike visible imaging, which relies on reflected light, thermal imaging directly captures emitted radiation, making it robust to lighting variations and capable of operating in complete darkness.

Thermal images suffer from fixed pattern noise [1], lack of texture and hence a smooth monochromatic appearance, and ghosting artifacts with detectors with long time constants, or large motion in the scene [3]. This particularly affects optical flow that relies on nearly noise-free images with high amounts of texture in the scene. Many of these drawbacks can be solved with a large pool of relevant thermal images. However, existing thermal image datasets are woefully insufficient to solve the problem of thermal optical flow.

Unlike visible datasets, thermal datasets are sparse in both quantity and image resolution. Some [6, 16, 62] address this limitation by synthetically converting visible images to the thermal domain using generative models. Others [29, 30, 63, 55, 19, 23] focus on collecting paired lwir-rgb images through stereo calibration or specialized optical hardware. However, these datasets often lack the diversity, resolution, and temporal consistency required for robust optical flow estimation, highlighting the need for a dedicated thermal optical flow dataset.

Chapter 3

Proposed Work

Our goal is to estimate the dense forward optical flow field $\mathbf{F}_{1 \rightarrow 2}$ from thermal image pairs (I_t^{ir}, I_{t+1}^{ir}) by learning an optimal function f_{θ^*} such that $\mathbf{F}_{1 \rightarrow 2} = f_{\theta^*}(I_1^{ir}, I_2^{ir})$. We learn the optimal parameters θ^* by finding $\arg \min \mathcal{L}(\theta)$. The global loss function $\mathcal{L}$ combines occlusion-aware unsupervised losses from [57][27] with our cross-modal supervision and is defined as:

$$\mathcal{L}(\theta) = \underbrace{\gamma_{ph}\mathcal{L}_{ph}(\theta) + \gamma_{sm}\mathcal{L}_{sm}(\theta)}_{\text{Unsupervised}} + \underbrace{\gamma_{cross}\mathcal{L}_{cross}(\theta)}_{\text{Cross-modal supervised}} \quad (3.1)$$

We employ a combination of unsupervised photometric loss and supervised RGB pseudo-labels. We achieve this by computing the endpoint error (EPE) between the predicted thermal flow and the pseudo-labels generated from a pre-trained RGB optical flow model. The EPE is defined as the charbonnier penalty of the difference between the two flow outputs:

$$\mathcal{L}_{cross}(\theta) = \frac{1}{HW} \sum \sqrt{|\mathbf{F}_{1 \rightarrow 2}^{ir} - \mathbf{F}_{1 \rightarrow 2}^{rgb}|^2 + \epsilon^2} \quad (3.2)$$

Dataset	Sync	Calib	FPS	Thermal Res.	RGB Res. (Mp)	Videos
UCRT (ours)	✓	✓	60	640x512	1.57	1088
Caltech Aerial [29]	✓	✓	30	640x512	0.58	37
M3FD [30]	✓	✓	-	640x512	0.76	-
FLIR ADAS [15]	✓	✗	24	640x512	0.76	8
Freiburg Thermal [63]	✓	✓	-	640x512	1.25	41
PST900T [55]	✓	✓	-	320x256	0.92	-
MFNet [19]	-	-	-	320x256	0.31	-
KAIST [23]	✓	✓	20	320x256	0.31	41

Table 3.1: **Comparing Datasets.** Key attributes include synchronization (Sync), calibration (Calib), frames per second (FPS), thermal resolution (Thermal Res.), RGB resolution (RGB Res.) in megapixels (Mp), and the number of videos (Vids). Our dataset, UCRT, stands out with the highest FPS, thermal resolution, and number of videos, making it the most comprehensive for cross-spectral optical flow tasks.

where ϵ is a small constant used for numerical stability. This cross-spectral supervision bridges the initial domain gap between RGB and thermal modalities.

Following [57] we apply the loss $L(\theta)$ to the intermediate predictions of the flow network f_θ . Architectures like RAFT and FlowFormer++ iteratively refine intermediate predictions at full resolution; therefore, for convergence, it is essential to apply the unsupervised losses at the entire sequence. We exponentially decay the weight of these losses in earlier iterations [61]. $L_{seq} = \sum_{i=1}^N \gamma^{N-i} \mathcal{L}_\gamma$, where N is the total number of flow iterations, and γ is the decay factor. Our experiments use $N = 12$ and $\gamma = 0.77$.

For reliable optical flow prediction, two key factors are essential: high frame rates and high resolution. High frame rates ensure temporal smoothness and reduce motion ambiguities as shown in Fig. 3.1 enabling a more accurate tracking of fast-moving objects. Similarly, high-resolution inputs preserve fine-grained details in the scene, as shown in Fig. 3.2, which are crucial for capturing subtle motion patterns. Beyond these, robust

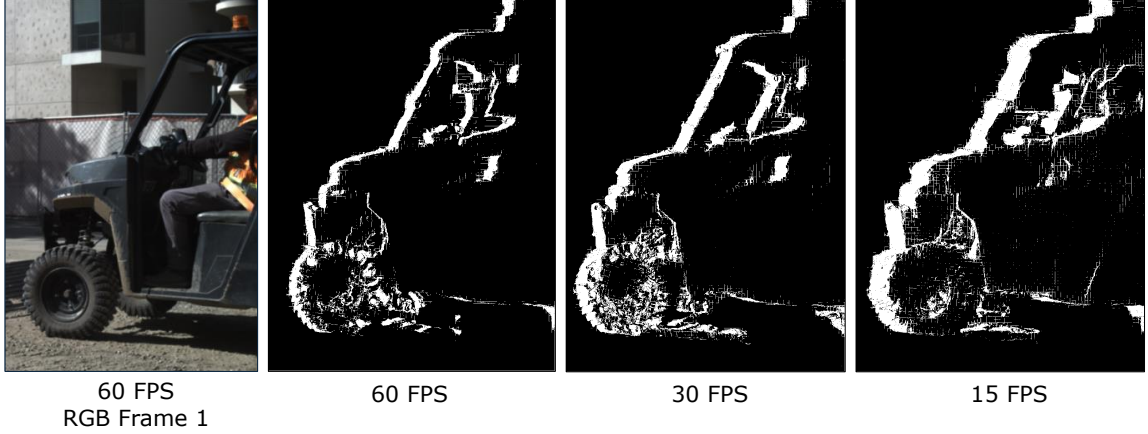


Figure 3.1: **Framerate influence.** Predicted occlusion masks illustrate the gradual increase in motion ambiguities as the frame rate decreases from 60 to 15 frames per second.



Figure 3.2: **Resolution influence.** Optical flow visualizations show the loss of detail in downsampled, lower-resolution input compared to the original higher-resolution input.

training of a cross-spectral optical flow model requires additional considerations. Spatial and temporal synchronization across video frames is necessary to maintain consistency in motion estimation, while diverse motion patterns in the training data help the model generalize to various real-world scenarios. To accomplish this task, we collect the first RGB-Thermal high resolution, high frame rate dataset with over 1000 diverse burst video sequences. As shown in Table [3.1](#), our dataset is uniquely suited for training optical flow models.

Our dataset was collected across a diverse 600-acre campus space, which includes



Figure 3.3: **Dataset Snapshot.** The visualization above shows a snapshot of the diverse set of images captured by our synchronized optical setup. We captured indoor scenes, outdoor scenes, scenes with moving humans, with moving vehicles, day time, and night time. Our dataset offers a significant contribution to computer vision in the thermal wavelengths.

varied terrains, visualized in Fig. 3.3. The data collection was performed using handheld equipment, ensuring flexibility in capturing a wide range of dynamic motion and scenes. The thermal images were captured in a 16-bit pre-automatic gain control (AGC) mode to ensure that the raw data is preserved. The RGB data was downsampled from 3.1MP to 1.57MP for efficient storage.

Let the dataset D_{ucrt} be a collection of temporally and spatially aligned RGB-LWIR videos $\{(V_i^{rgb}, V_i^{lwir})\}_{i=0}^N$ consisting of τ consecutive frames where $V_i^M = \{I_t^M\}_{t=0}^\tau$ for $M \in \{rgb, lwir\}$ spectral modalities. For each video sequence, the thermal and color images are spatially related by a homography transformation $\{H_i\}$, which aligns their respective

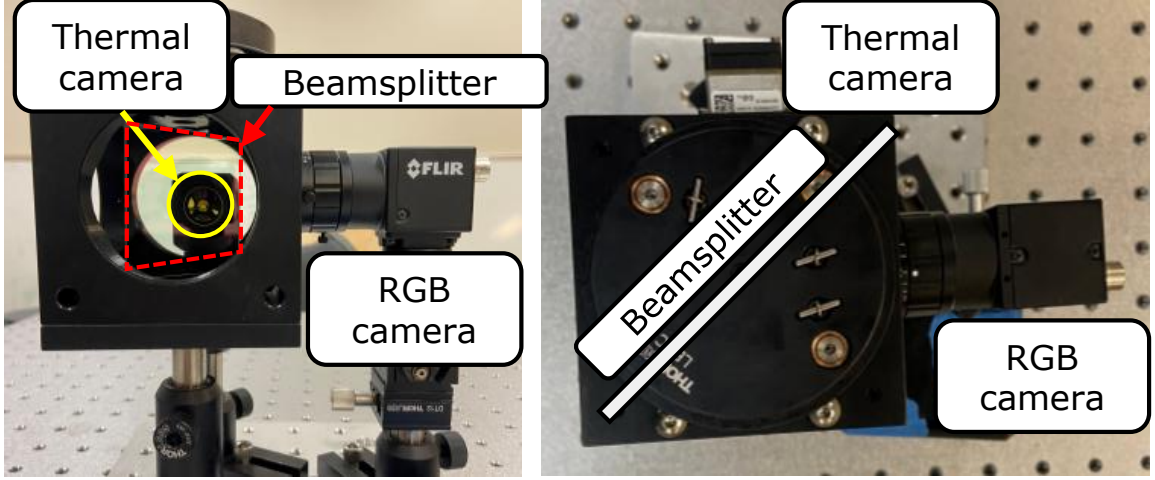


Figure 3.4: **Hardware Setup.** Our optical setup consists of a FLIR Boson+ camera, and a FLIR Blackfly S camera, both colocated with a dichroic mirror that transmits visible wavelengths and reflects LWIR wavelengths. Both cameras are synchronized by an external trigger generated by an Arduino at 60 frames per second.

coordinate systems. Therefore, each frame pair (I_t^{lwir}, I_t^{rgb}) includes:

- A thermal image $I_t^{lwir} \in \mathbb{R}^{H \times W}$
- A color image $I_t^{rgb} \in \mathbb{R}^{H \times W \times C}$

In the following sections, we will present our hardware setup, outline the synchronization and calibration procedures, describe the data collection workflow, and conclude with a comparative analysis against other RGB-thermal datasets. The optical imaging system, illustrated in Fig. 3.5a, comprises an optical beamsplitter, a thermal camera, and a visible camera. We use a BlackFly S 3.1 megapixel color machine vision camera with a 12mm focal length lens and an uncooled Boson 14mm 32Fov thermal camera. A 52mm square indium tin oxide (ITO) dichroic optical beamsplitter is mounted on a three-axis kinematic prism platform, positioned perpendicularly to both the thermal and visible cameras. The purpose of the beamsplitter is to transmit wavelengths from the visible band (400-800nm) of the

incident light and to reflect the thermal band (2000-14000nm). This coaxial configuration eliminates any disparity by ensuring that there is no baseline between the two cameras, therefore allowing their optical centers to coincide.

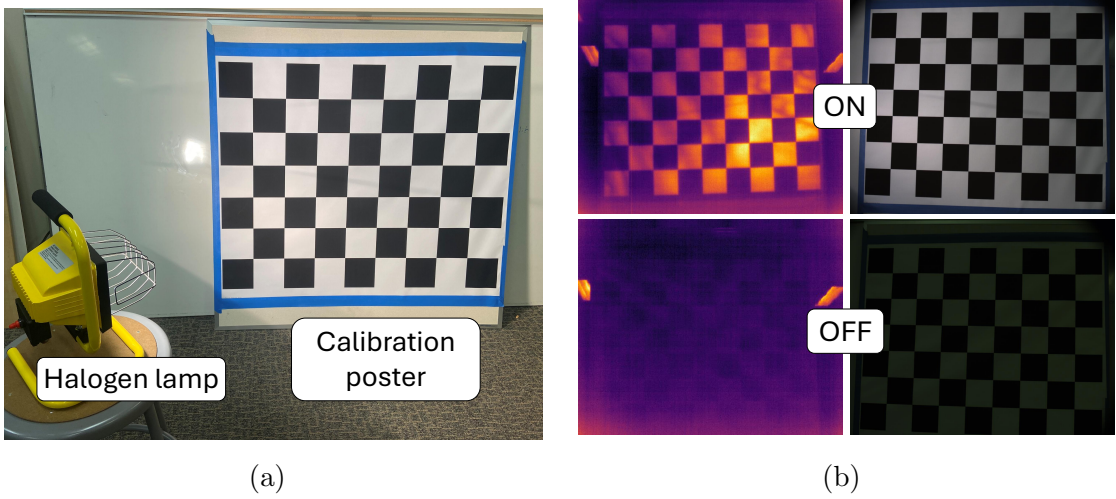


Figure 3.5: **Calibration Setup.** (a) A halogen lamp is positioned near a calibration poster board. (b) When activated, the black squares absorb and emit more heat than the white squares, creating a distinct checkerboard pattern in the thermal image.

With the goal of estimating optical flow in mind, we prioritize the acquisition of high frame rates to minimize motion ambiguity between consecutive frames through hardware-triggered temporal synchronization. We achieve this configuration by dedicating an Arduino Uno to output a steady 25% duty cycle PWM signal at 60 Hz. At the rising edge, both the RGB and LWIR cameras capture a frame, ensuring temporal alignment.

Following [55, 35], we perform spatial calibration using a 60x75 cm poster-printed portable checkerboard target affixed to a rigid pin board. We use a halogen bulb to heat the poster paper as shown in Fig. 3.5. We frequently collect calibration data for each day totaling over 600 images. We implement a robust calibration pipeline extending Zhang’s

classical checkerboard method [76]. Our process can be broken down into two stages, namely correspondence detection and RANSAC optimization.

For correspondence detection, synchronized image pairs undergo spectral-specific preprocessing. Then chessboard corners are detected in both modalities and any ambiguity in ordering is resolved through mean squared error verification. We further refine the homography estimation using RANSAC with adaptive iteration control. In each iteration, N correspondence pairs are randomly sampled to compute a candidate homography H_k via direct linear transform. The best homography H^* is selected based on maximal inlier consensus.

Chapter 4

Experiments

To evaluate the effectiveness of our proposed thermal optical flow framework, we conducted experiments using two state-of-the-art optical flow architectures: RAFT [61] and FlowFormer++ [48]. For both networks, we use the AdamW optimizer and one-cycle learning rate scheduler. The initial learning rate is set to $1e-4$ with a weight decay of $1e-4$ is used to regularize the model weights. We train the networks for 10k iterations using a batch size of 4. We experimented with pretrained weights from multiple datasets, including FlyingChairs, Sintel, KITTI and FlyingThings3D, and we report our findings in Table 4.1. To initialize the networks, we use pretrained weights from the Sintel dataset, which provide a strong starting point. Our results indicate that the weights pretrained on the Sintel dataset are the most effective.

Our dataset is split into training, validation and test sets with an 80/10/10 ratio, ensuring a balanced evaluation of the model’s performance. Before training, the RGB images and their corresponding pseudo-optical flow labels are downsampled and warped

	Method	Census	Charbonnier
FlyingThings	RAFT- <i>low</i>	0.540	0.385
	RAFT- <i>high</i>	0.427	0.360
FlyingChairs	RAFT- <i>low</i>	0.546	0.434
	RAFT- <i>high</i>	0.433	0.403
KITTI	RAFT- <i>low</i>	0.548	0.400
	RAFT- <i>high</i>	0.434	0.375
Sintel	RAFT- <i>low</i>	0.538	0.383
	RAFT- <i>high</i>	0.419	0.356

Table 4.1: **Pretrained Models.** Comparing optical flow at low resolution (768x512) with high resolution (1536x1024) rgb inputs for different supervised synthetic pre-trained weights across the sintel (S), KITTI (K), FlyingChairs (C) and FlyingThings3D (T) datasets.

using a homography matrix to align them with the thermal images. The images are then cropped to a resolution of 448×512 to standardize the input size and reduce computational overhead. To improve the robustness of the thermal optical flow model, we apply photometric augmentations to the thermal input images. Specifically, we use contrast limited adaptive histogram equalization (CLAHE) and sharpness adjustments, each applied randomly with a 20% probability. Additionally, we pre-compute occlusion masks from high-resolution RGB image pairs $(I_t^{rgb}, I_{t+1}^{rgb})$. These masks are used to identify regions where photometric consistency cannot be enforced, ensuring that the model focuses on reliable regions during training.

In the absence of ground truth optical flow, we rely on average endpoint error (EPE) using high-resolution RGB optical flow. Additionally, we evaluate photometric consistencies through the structural similarity index measure (SSIM), census transform, and generalized Charbonnier photometric loss.

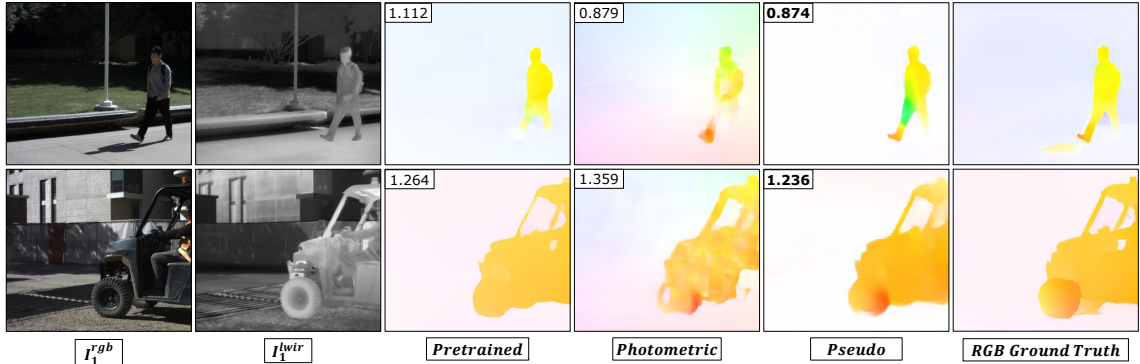


Figure 4.1: **Qualitative Results.** RGB and LWIR frames are shown alongside the outputs of three methods: pretrained (no fine-tuning), photometric fine-tuning, and pseudo fine-tuning, with RGB ground truth provided for reference. The Charbonnier metric (lower is better) is included for quantitative evaluation. Across the examples, the pseudo fine-tuning approach demonstrates superior performance, supported by the lowest Charbonnier scores.

Method	EPE	1px	3px	5px
Pseudo	1.374	0.620	0.904	0.930
Photometric	1.583	0.646	0.904	0.930
Pre-Trained	1.407	0.681	0.896	0.927

Table 4.2: **Quantitative Results.** End-point error (EPE) and accuracy thresholds (1px, 3px, 5px) are reported for three methods: Pseudo fine-tuning, Photometric fine-tuning, and Pre-Trained (no fine-tuning). Lower EPE values indicate better performance, while higher accuracy values are preferable. The Pseudo fine-tuning method achieves the lowest EPE, demonstrating its effectiveness.

We compared the performance of RAFT and FlowFormer++ under different settings, including the use of pretrained weights and our unsupervised cross-modal fine-tuning strategy.

We conducted an ablation study to evaluate the impact of cross-modal supervision on RAFT’s performance. As shown in Table 4.2 leveraging pseudo-labels generated from pre-trained RGB models significantly enhances the model’s performance. Our experiments reveal that pseudo fine-tuning converges to optimal performance when assessed using both photometric and supervised metrics.

Furthermore, Figure 4.1 provides a qualitative comparison between RGB and LWIR optical flow predictions. The results demonstrate that the pseudo fine-tuned framework produces accurate and consistent flow fields in the thermal domain, underscoring the effectiveness of our cross-modal supervision approach.

Chapter 5

Conclusions

In this work, we introduce the first thermal optical flow network, addressing a significant gap in the field. We present a novel high-resolution, high-frame-rate RGB-thermal dataset, consisting of over 1,000 diverse scenes to enable robust training and evaluation. Our cross-spectral supervision strategy leverages pre-trained RGB models to generate occlusion masks and pseudo-labels, effectively bridging the domain gap between RGB and thermal modalities. By combining unsupervised learning with RGB pseudo-supervision, we demonstrated improved performance in challenging conditions where traditional methods struggle. Our experiments with RAFT and FlowFormer++ architectures highlight the effectiveness of our approach, showcasing its potential for real-world applications. This work lays the foundation for further research in thermal optical flow, offering a new baseline.

Bibliography

- [1] *Fundamentals of Infrared Thermal Imaging*, chapter 1, pages 1–72. John Wiley Sons, Ltd, 2010.
- [2] Simon Baker, Stefan Roth, Daniel Scharstein, Michael J. Black, J.P. Lewis, and Richard Szeliski. A database and evaluation methodology for optical flow. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.
- [3] Fanglin Bao, Shubhankar Jape, Andrew Schramka, Junjie Wang, Tim E. McGraw, and Zubin Jacob. Why thermal images are blurry. *Opt. Express*, 32(3):3852–3865, Jan 2024.
- [4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep Burst Super-Resolution . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9205–9214, Los Alamitos, CA, USA, June 2021. IEEE Computer Society.
- [5] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep Reparametrization of Multi-Frame Super-Resolution and Denoising . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2440–2450, Los Alamitos, CA, USA, October 2021. IEEE Computer Society.
- [6] Neeraj Bhat, Navneet Saggi, and S. Kumar Pragati. Generating visible spectrum images from thermal infrared using conditional generative adversarial networks. pages 1390–1394, 06 2020.
- [7] Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, and Milind Tambe. BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos . In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1736–1745, Los Alamitos, CA, USA, March 2020. IEEE Computer Society.
- [8] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

- [9] Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Aggelos Katsaggelos, Ashok Veeraraghavan, and Oliver Cossairt. Thermal Spread Functions (TSF): Physics-Guided Material Classification . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1641–1650, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.
- [10] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow From Anywhere: Pseudo Multi-Modal Object Detection in Thermal Imagery . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1029–1038, Los Alamitos, CA, USA, June 2019. IEEE Computer Society.
- [11] Dinh Phat Do, Taehoon Kim, Jaemin Na, Jiwon Kim, Keonho Lee, Kyunghwan Cho, and Wonjun Hwang. D3T: Distinctive Dual-Domain Teacher Zigzagging Across RGB-Thermal Gap for Domain-Adaptive Object Detection . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23313–23322, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.
- [12] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [13] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.
- [14] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [15] Inc. FLIR Systems. Flir adas dataset. Thermal imaging dataset for Advanced Driver Assistance Systems (ADAS).
- [16] Lu Gan, Connor T. Lee, and Soon-Jo Chung. Unsupervised rgb-to-thermal domain adaptation via multi-domain attention network. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6014–6020, 2022.
- [17] Mathias Gehrig, Mario Millhauser, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras . In *2021 International Conference on 3D Vision (3DV)*, pages 197–206, Los Alamitos, CA, USA, December 2021. IEEE Computer Society.
- [18] 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, 2012.

- [19] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017.
- [20] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1–3):185–203, August 1981.
- [21] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [22] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. *ECCV*, 2022.
- [23] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2016.
- [25] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.
- [26] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–572. Springer, 2020.
- [27] Rico Jonschkowski, Austin Stone, Jonathan T. Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 557–572, Berlin, Heidelberg, 2020. Springer-Verlag.
- [28] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gusefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.
- [29] Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Caltech aerial rgb-thermal dataset in the wild. In *European Conference on Computer Vision*, pages 236–256. Springer, 2025.

- [30] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.
- [31] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6489–6498, 2020.
- [32] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [33] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8770–8777, 2019.
- [34] Yawen Lu and Guoyu Lu. An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image. pages 3832–3842, 01 2021.
- [35] Thomas Luhmann. *Geometric Calibration of Thermographic Cameras*. 05 2013.
- [36] Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, and Shuaicheng Liu. Efficient Meshflow and Optical Flow Estimation from Event Cameras . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19198–19207, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.
- [37] Xinglong Luo, Kunming Luo, Ao Luo, Zhengning Wang, Ping Tan, and Shuaicheng Liu. Learning Optical Flow from Event Camera with Rendered Dataset . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9813–9823, Los Alamitos, CA, USA, October 2023. IEEE Computer Society.
- [38] Rémi MARSAL, Florian CHABOT, Angélique Loesch, and Hichem SAHBI. Brightflow: Brightness-change-aware unsupervised learning of optical flow. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2060–2069, 2023.
- [39] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.
- [40] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the Thirty-Second AAAI*

- Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’18/IAAI’18/EAAI’18. AAAI Press, 2018.
- [41] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert E. Carroll. Burst denoising with kernel prediction networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2017.
 - [42] Zhixiang Min, Yiding Yang, and Enrique Dunn. Voldor: Visual odometry from log-logistic dense optical flow residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
 - [43] Yasuto Nagase, Takahiro Kushida, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Shape from Thermal Radiation: Passive Ranging Using Multi-spectral LWIR Measurements . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12651–12661, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.
 - [44] Sriram Narayanan, Mani Ramanagopal, Mark Sheinin, Aswin C. Sankaranarayanan, and Srinivasa G. Narasimhan. Shape from heat conduction. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVIII*, page 426–444, Berlin, Heidelberg, 2024. Springer-Verlag.
 - [45] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.
 - [46] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2232–2241, 2017.
 - [47] Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, and Chandra Kambhamettu. Material classification with thermal imagery . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4649–4656, Los Alamitos, CA, USA, June 2015. IEEE Computer Society.
 - [48] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023.
 - [49] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision (ECCV)*, pages 628–645, 2022.
 - [50] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth, and ego-motion by contrast maximization. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, pages 1–18, 2024.

- [51] Ukcheol Shin, Kyunghyun Lee, Seokju Lee, and In So Kweon. Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss. *IEEE Robotics and Automation Letters*, 7(2):1103–1110, 2021.
- [52] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep Depth Estimation from Thermal Image . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1053, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.
- [53] Ukcheol Shin, Kwanyong Park, Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Self-supervised Monocular Depth Estimation from Thermal Images via Adversarial Multi-spectral Adaptation . In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5787–5796, Los Alamitos, CA, USA, January 2023. IEEE Computer Society.
- [54] Ukcheol Shin, Kwanyong Park, Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Self-supervised monocular depth estimation from thermal images via adversarial multi-spectral adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5798–5807, 2023.
- [55] 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, 2020.
- [56] Pranjay Shyam and HyunJin Yoo. Lightweight Thermal Super-Resolution and Object Detection for Robust Perception in Adverse Weather Conditions . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7456–7467, Los Alamitos, CA, USA, January 2024. IEEE Computer Society.
- [57] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [58] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2021.
- [59] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.
- [60] Kenichiro Tanaka, Nobuhiro Ikeya, Tsuyoshi Takatani, Hiroyuki Kubo, Takuya Fumatomi, Vijay Ravi, Achuta Kadambi, and Yasuhiro Mukaigawa. Time-Resolved Far Infrared Light Transport Decomposition for Thermal Photometric Stereo . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(06):2075–2085, June 2021.

- [61] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 402–419, Berlin, Heidelberg, 2020. Springer-Verlag.
- [62] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O’Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. CATS: A Color and Thermal Stereo Benchmark . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 134–142, Los Alamitos, CA, USA, July 2017. IEEE Computer Society.
- [63] Johan Vertens, Jannik Zörn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020.
- [64] Vibashan Vs, Domenick Poster, Suya You, Shuowen Hu, and Vishal M. Patel. Meta-UDA: Unsupervised Domain Adaptive Thermal Object Detection using Meta-Learning . In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3697–3706, Los Alamitos, CA, USA, January 2022. IEEE Computer Society.
- [65] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using HR optical flow estimation. *IEEE Transactions on Image Processing*, 2020.
- [66] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4884–4893, 2018.
- [67] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Reza Tofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022.
- [68] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pages 3–10. Springer, 2016.
- [69] Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. Unsamflow: Unsupervised optical flow guided by segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19027–19037, 2024.
- [70] Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. Unsamflow: Unsupervised optical flow guided by segment anything model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19027–19037, 2024.

- [71] Shuai Yuan, Xian Sun, Hannah Kim, Shuzhi Yu, and Carlo Tomasi. Optical flow training under limited label budget via active learning. In *European Conference on Computer Vision*, pages 410–427, 2022.
- [72] Shuai Yuan, Shuzhi Yu, Hannah Kim, and Carlo Tomasi. Semarflow: Injecting semantics into unsupervised optical flow estimation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9566–9577, October 2023.
- [73] Shuai Yuan, Shuzhi Yu, Hannah Kim, and Carlo Tomasi. Semarflow: Injecting semantics into unsupervised optical flow estimation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9566–9577, 2023.
- [74] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [75] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8876–8885, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.
- [76] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.