Projecting Trackable Thermal Patterns for Dynamic Computer Vision

Mark Sheinin, Aswin C. Sankaranarayanan, and Srinivasa G. Narasimhan
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

*Adding artificial patterns to objects, like QR codes, can ease tasks such as object tracking, robot navigation, and conveying information (*e.g.*, a label or a website link). However, these patterns require a physical application and they alter the object's appearance. Conversely, projected patterns can temporarily change the object's appearance, aiding tasks like 3D scanning and retrieving object textures and shading. However, projected patterns impede dynamic tasks like object tracking because they do not 'stick' to the object's surface. Or do they? This paper introduces a novel approach combining the advantages of projected and persistent physical patterns. Our system projects heat patterns using a laser beam (similar in spirit to a LIDAR), which a thermal camera observes and tracks. Such thermal patterns enable tracking poorly-textured objects whose tracking is highly challenging with standard cameras while not affecting the object's appearance or physical properties. To avail these thermal patterns in existing vision frameworks, we train a network to reverse heat diffusion's effects and remove inconsistent pattern points between different thermal frames. We prototyped and tested this approach on dynamic vision tasks like structure from motion, optical flow, and object tracking of everyday textureless objects.*

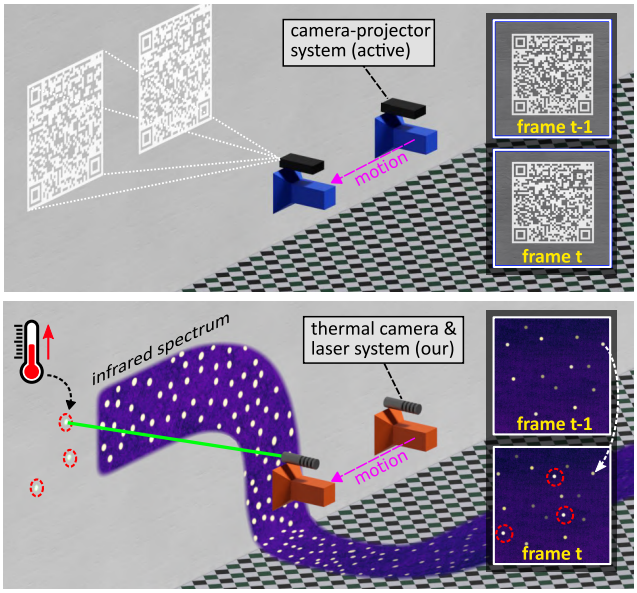

Figure 1. Projecting trackable thermal patterns. Visual navigation is challenging in environments lacking distinct features, like texture-less walls or highly repetitive tiling. **(Top)** A camera-projector system can introduce distinct features, but these features can not be used as a reference between frames because the projected pattern 'moves' along with the system. **(Bottom)** Our system comprises a thermal camera and a colocated laser projector. The laser 'paints' a temporary heat pattern that 'sticks' to objects' surfaces and can, therefore, be tracked in the thermal domain.

1. Introduction

Pattern matching is fundamental for navigating, understanding, and interacting with the world. In particular, environment modeling and navigation require recognizing the same patterns from different views and at different times. But many parts of our environment, whether naturally formed or man-made, have an appearance that makes these tasks challenging. For instance, as seen in Fig. 1(Top), visual navigation in a long hallway using a textureless wall or highly repetitive floor tiling will be difficult. This problem worsens at night when distant landscape features are not visible. Similarly, creating a 3D model of a smooth, textureless object is hard using its natural appearance.

In computer vision, the absence of a 'good' texture can be mitigated by adding artificial patterns to the object. Such patterns can be broadly classified into two categories: physical patterns directly imprinted on objects and projected patterns applied remotely. Physical patterns are helpful for dynamic tasks like object tracking because they can be designed with specific visual characteristics that make them easy to detect and track (*e.g.*, a QR code). However, applying such patterns requires direct contact with the object, and they alter the object's appearance. Conversely, projected patterns can be applied remotely and changed at will, making them ideal for 3D reconstruction and material property acquisition [23, 27]. However, projected patterns are ineffective for dynamic tasks because they do not 'stick' to the object's surface like physical patterns. This is illustrated by the blue camera-projector system in Fig. 1(Top). The sys-

tem moves along the hallway while projecting a pattern on the wall. However, the captured frames reveal no information about the camera motion because the pattern 'moves along' with the camera.

In this paper, we propose a novel approach that combines the benefits of both pattern types. Our method uses a laser to 'draw' a *heat pattern* on an object. The increase in surface temperature, which forms the pattern, is minimal, akin to a temperature rise from a hand imprint. While invisible to standard RGB cameras or human observers, these heat patterns are visible to thermal cameras, which image in the infrared spectrum. Thermal patterns can be applied remotely, just like standard projected structured light patterns, but have the advantage of remaining 'stuck' to the object's surface for a certain time duration. Therefore, thermal patterns allow a navigating agent to 'paint their own texture' on completely textureless object surfaces, thus aiding localization and mapping (see Fig. 1(Bottom)). Finally, these patterns vanish once the target surface returns to room temperature, leaving the object's appearance unaltered.

However, using thermal heat patterns presents unique new challenges. Most tracking algorithms assume that the appearance of tracked features remains consistent across frames (*e.g.*, optical flow). However, thermal patterns significantly change over time due to heat diffusion and entirely fade away after a while. This requires the continual projection of new pattern points not existing in the previous frames (red circles in Fig. 1), further complicating the tracking task between frames. To address these challenges, we develop a learning-based approach that mitigates appearance variation between frames, allowing thermal patterns to integrate into existing vision frameworks seamlessly.

We introduce a new class of structured light methods in the thermal regime, where the projected patterns 'stick' onto object surfaces. We systematically explore the space of possible projected patterns and determine which patterns suit dynamic vision tasks. We build a prototype and demonstrate its effectiveness in tasks like structure from motion, object tracking, and optical flow. This work opens a new avenue in fusing physical and projected patterns, promising novel possibilities in computer vision applications.

2. Related works

Thermal imaging has been adopted as a sensing modality for various 'classic' computer vision tasks, including depth reconstruction [5, 20, 21, 33, 34], object segmentation [7, 18], person detection, recognition and re-identification [13, 14, 32, 41, 42, 44], people tracking and pose estimation [6, 39], and more. Moreover, many works have focused on transferring thermal images into the visible-light domain [1, 12, 38] and improving the thermal images' quality via algorithmic means [25] or clever combinations of algorithms and hardware [15, 28].

Some researchers have also leveraged the unique properties of thermal radiation to accomplish novel vision tasks beyond standard RGB cameras. Tomohiro *et al*. leveraged thermal reflection for non-line-of-sight imaging [19], while Liu *et al*. used reflections to reconstruct a human shape. Tang *et al*. [36] used thermal images to infer past human positions while Brahmbhatt *et al*. inferred human grasping [4].

Finally, some past methods combined thermal imaging with active illumination. Dashpute *et al*. imaged a laser heating profile to classify materials [10], Tanaka *et al*. exploited the temperature rise from a far IR light source for light transport decomposition [35], and Erdozain *et al*. combined an infrared source with an infrared sensor for 3D scanning [11]. Our method goes beyond these prior works by exploiting the inherent persistence of projected thermal patterns, which we use to accomplish dynamic vision tasks.

3. Background

3.1. Thermal Imaging

Thermal cameras sense the light *emitted* from objects in the infrared spectrum. The amount of radiation emitted by an object, per wavelength, depends on the object temperature $\boldsymbol{T}$ [22]. The exact relationship between $\boldsymbol{T}$ and the signal measured by a Long Wave Infrared (LWIR) camera $S(\boldsymbol{T})$ is a function of the wavelength-dependant emission and camera response, the object emissivity, atmospheric transmission, temperature of surrounding objects and more [40]. Camera manufacturers usually model the camera readings using the Sakuma–Hattori equation:

$$S(\boldsymbol{T}) = \frac{c_1}{\exp \frac{c_2}{c_3 \boldsymbol{T} + c_4} - 1}, \tag{1}$$

where c_1, c_2, c_3 and c_4 are curve-fitting parameters [26]. In this paper, we avoid the cumbersome procedure of calibrating the constants in Eq. (1). Instead, our method relies on the spatial structure and the temporal difference between individual thermal frames, much like human vision, which is mostly invariable to the absolute scene light intensity.

3.2. Laser heating

The heating of a surface by a laser can be described by:

$$\frac{1}{\alpha}\frac{\delta \boldsymbol{T}}{\delta t} = \Delta \boldsymbol{T} + \frac{(1-\rho)\boldsymbol{P}^{\text{laser}}\delta}{\boldsymbol{k}} e^{-\delta z}, \tag{2}$$

where $\boldsymbol{P}^{\text{laser}}$ is the laser's power [43], Δ is the Laplace operator and z is the depth into the material. The constants α, $\boldsymbol{k}$, and δ are the material's thermal diffusivity, thermal conductivity, and absorption coefficient, respectively. These constants determine how 'well' the laser can heat a spot and how fast the generated heat will diffuse. The exponent term suggests that the laser exponentially decays into the material ($z > 0$). Observe that the material's albedo ρ at the

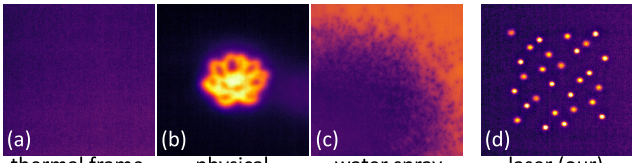

Figure 2. Trackable thermal patterns. **(a)** Thermal image before application. **(b)** A physically applied pattern by touching the object surface with a warm metal branding iron. **(c)** Spraying water yields an uncontrolled texture in the thermal domain, resulting in high- or low-frequency textures. **(d)** We propose a practical, remote, and completely controlled way for applying a heat texture on scene surfaces by steering a laser at desired scene points.

laser's wavelength determines the fraction of laser power absorbed and converted to heat. Therefore, dark objects require less laser power to generate the same temperature rise.

4. Projecting Trackable Thermal Patterns

A *trackable thermal pattern* is a thermal pattern that satisfies these two conditions: (a) it has a spatial structure that yields trackable feature points, and (b) these feature points remain trackable for at least two consecutive video frames for a static object and camera. There are many ways to generate trackable thermal patterns. For example, a heat pattern could be applied to an object's surface via direct physical contact (see Fig. 2(b)). Such patterns could be applied by a robot's manipulating arm or tracks, but their physical range is limited to the robot's reaching range. Moreover, physical application limits the pattern's programmability.

Another way to apply heat patterns is by spraying the surface with a liquid (see Fig. 2(c)). Spraying a liquid creates a random arrangement of water droplets on the object's surface. However, the trackability of the resulting thermal texture is inconsistent since it has a large variability in quality. Specifically, as seen in Fig. 2(c), the spraying may result in 'good' high-frequency texture regions and 'bad' regions lacking spatial contrast.

We propose a practical way for remotely creating controllable, trackable thermal patterns using a camera-projector system that can be fitted on autonomous vehicles and robots. Specifically, the system consists of a thermal camera and a steerable visible or near-infrared laser in a coaxial configuration that can create and observe arbitrary sequential patterns on environment surfaces energy-efficiently (Fig. 2(d)). The laser 'paints' the patterns by converting light energy to heat. Next, we describe the design of our camera-projector system and the projected patterns.

4.1. Designing the Camera-Projector System

Fig. 3 shows a schematic of our system. A thermal camera views the scene through an optical filter designed to reflect infrared light but transmit visible light. On the filter's second side, a projector consisting of a laser steered with a dual-axis galvanometer mirror system illuminates the scene. The camera and projector are temporally synchronized to control the projected pattern during each camera frame.

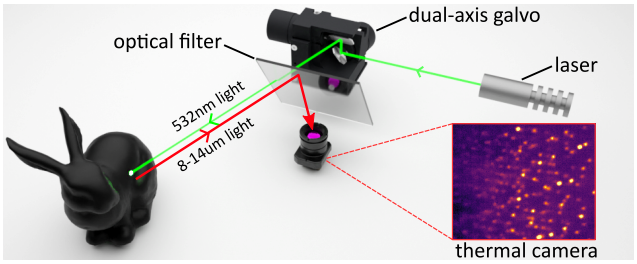

Figure 3. Our system comprises a thermal camera co-located with a dual-axis Galvo scanning mirror, which steers a laser beam at scene surfaces. The laser locally heats up object surface points, creating a trackable heat pattern in the infrared domain.

We set the camera and projector in a coaxial configuration (no baseline exists between the two). Thus, each camera pixel can be mapped to a corresponding outgoing projector ray. Then, for a static scene, a pattern can be designed in the thermal camera's image domain and predictably illuminate the desired scene points regardless of scene geometry. For a dynamic scene, the imaged pattern will deviate from the projected pattern due to object or system motion.

Let $I(\mathbf{x}, t)$ denote the scaled image intensity at pixel coordinates $\mathbf{x} \in \mathbb{R}^2$ and frame sample time t, where $t = 0$ denotes the sample time of the video stream's first frame. We denote t as the *sample time* because, as elaborated in the supplementary, unlike standard visible-light cameras, our thermal camera does not continuously collect light during the frame exposure time but samples the incident power of infrared light during continuous exposure [24]. During the camera's operation, the projector continuously projects a series of 'dots' toward the scene having index n, where $n = 0, 1, .., N-1$. Since the camera and laser are co-located, we denote each dot's projection direction using the camera's pixel coordinates $\mathbf{x}_n$. Each dot is steered to $\mathbf{x}_n$ at time t_n and remains at $\mathbf{x}_n$ until time t_{n+1}. Without loss of generality, suppose that all dots have the same duration $T^{\text{dot}} = t_{n+1} - t_n \forall n$. Then, define the projector illumination pattern P as the sequence of all projected points

$$P \equiv \Big((\mathbf{x}_n, t_n) \Big) \Big|_{n=0}^{N-1}. \tag{3}$$

4.2. Which Patterns Fit Dynamic Vision Tasks?

The pattern definition in Eq. (3) can describe an arbitrary sequential pattern. For example, when T^{dot} is a low-order fraction of the camera sample period T^{samp}, the resulting pattern will appear as a series of disconnected dots (Fig. 2(d)). Conversely, using $T^{\text{samp}} \gg T^{\text{dot}}$, one

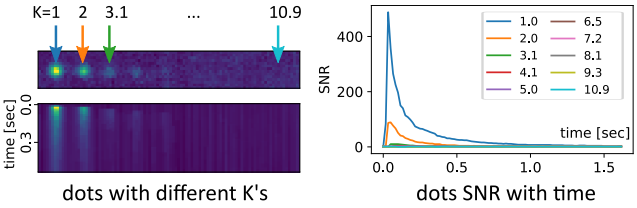

Figure 4. Laser dot projection times. **(Left)** A thermal image showing the relative intensity of laser dots projected at different durations T^{samp}/K, where K is the number of points per frame. Below is a plot of a single row over time (vertically). **(Right)** Dot SNR over time for various K.

can 'draw' seemingly continuous patterns like a circle or a square. Given this vast space of possible patterns, which pattern should one use for dynamic vision tasks?

The pattern dot duration T^{dot} is a key parameter that guides the pattern selection process. Let K denote the number of pattern points projected per frame,

$$K \equiv T^{\text{samp}}/T^{\text{dot}}. \tag{4}$$

For simplicity, we constrain K to take positive integer values. In other words, we assume the system projects at least one new point for each new video frame. The point duration T^{dot} determines the temperature rise at the object's surface, which in turn determines the signal-to-noise ratio (SNR) of each pattern dot over time.

Computing the exact temperature rise analytically using Eq. (2) requires meticulous calibrations. Instead, we empirically evaluate the SNR by projecting dots with different Ks on a black object, which serves as the 'best case scenario' for most environments. Fig. 4 shows the SNR of individual pattern dots over time for $K = 1, 2, .., 10$.[1] The plots in Fig. 4(Right) show that the SNR rapidly deteriorated for $K > 2$. This result suggests that to maintain the patterns' trackability for a reasonable duration, we must use 'discrete' patterns having only a few dots per frame as opposed to 'continuous' patterns where K is relatively large.

Different pattern types adhere to the mentioned constraint (*e.g.*, discrete-point patterns). Since scene motion can occur anywhere in the image domain, the frame must be evenly covered by P. Moreover, given the unknown and potentially changing direction of motion, a quasi-random coverage of the image domain is preferred over a sequential one. We tested various random dot pattern types that fit the demands above and concluded that Halton sequences perform best for our applications [17]. Now that we have determined which patterns to project, we must address the challenge of heat diffusion inherent to thermal patterns.

[1]The effective K in the experiment of Fig. 4 differs slightly from the intended integer values due to camera-projector synchronization constraints detailed in the supplementary.

5. Learning to Reverse Heat Diffusion

In the previous section, we showed how projected thermal patterns diffuse and evaporate, changing their appearance and thus requiring a constant influx of *new points* to be added to the object. This continuous variation in the imaged pattern can degrade the performance of vision algorithms that assume a consistent visual appearance between frame pairs (*i.e.*, brightness constancy assumption). We now describe a learning-based approach to 'undiffuse' the thermal frames and preserve their relative visual appearance.

Let $f = 0, 1, \ldots$ be the frame index, and $I(\mathbf{x}, t_f)$ is the frame corresponding to index f. For brevity, henceforth, we drop the time symbol and denote $I(\mathbf{x}, f) \equiv I(\mathbf{x}, t_f)$. Let $P_{f,m}$ denote the sub-sequence of pattern points projected between the sample of frame f and $f+m$:

$$P_{f,m} \equiv \Big((\mathbf{x}_n, t_n)_k \mid t_n \in [t_f, t_{f+m}) \Big) \Big|_{k=1}^{K}. \tag{5}$$

Namely, $P_{f,m}$ tells us which new points will appear in frame $f+m$ that did not exist in frame f. For example, when projecting $K=3$ points per frame, $P_{f,1}$ will contain exactly three points whose projection start time are t_f, $t_f + T^{\text{dot}}$ and $t_f + 2T^{\text{dot}}$. Similarly, $P_{f,m}$ will contain mK points that are projected between frames f and $f+m$. Then, the difference in scene appearance between frame f and frame $f+m$, for a static scene and static noiseless camera, is given by the *projection-diffusion* operator D:

$$I(\mathbf{x}, f+m) = D(I(\mathbf{x}, f) \mid P_{f,m},\ m). \tag{6}$$

Here, D does two operations: (a) diffusing the appearance of points existing in $I(\mathbf{x}, f)$ by the equivalent of m frames, and (b) adding the newly projected points defined by $P_{f,m}$.[2]

5.1. Reversing the Projection-Diffusion Operator

We aim to match the appearance of $I(\mathbf{x}, f)$ and $I(\mathbf{x}, f+m)$ before they serve as input to some dynamic vision task. But due to scene motion, we can not predict the new point locations in frame $f+m$, even given $P_{f,m}$. Thus, we instead seek the inverse operator

$$I(\mathbf{x}, f) \approx D^{-1}(I(\mathbf{x}, f+m) \mid P_{f,m},\ m). \tag{7}$$

The projection-diffusion reversal (PDR) operator D^{-1} reverses the diffusion of points that exist in frame f and removes the new points $P_{f,m}$. Eq. (7) is an approximation since the physical heat transfer is generally irreversible [9].

Modeling D analytically requires calibrating not only the material-specific parameters in Eq. (2), but also the camera's exact radiometric response in Eq. (1) and the signal's atmospheric attenuation. Moreover, the laser point's exact heat diffusion depends on the existing three-dimensional

[2]In Eq. (6), we neglect the newly projected points' diffusion since, in Sec. 5.1, we only seek D^{-1} where the new points are removed.

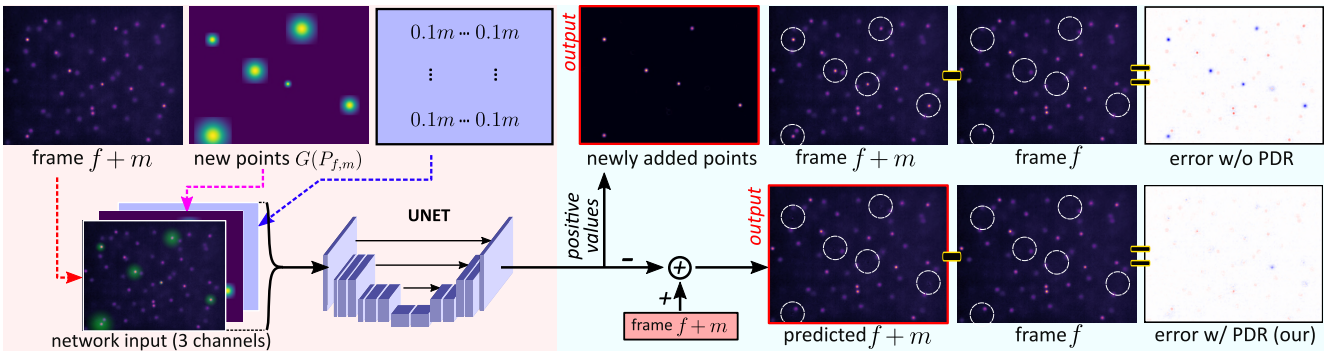

Figure 5. Reversing thermal projection-diffusion. Besides the effect of scene of camera motion, each successive thermal frame $f+m$ differs from the previous frame f in two additional ways: (a) it contains newly projected points not existing in frame f, (b) previously projected points (existing in frame f) have undergone heat diffusion. To correct these non-motion-related deviations, we train a neural network to reverse the projection-diffusion's effect, yielding a corrected $f+m$ frame with an appearance consistent with frame f.

heat profile, which is unobservable by the camera. Nevertheless, our applications only require reversing D for a few frames (*e.g.*, $m = 1, 2$), and each captured frame contains many 'data examples' of point diffusion over time (from recently projected points to points projected many frames ago). This plurality of points in each frame holds relevant cues about the diffusion-process parameters of the scene materials. This insight motivates us to use a data-driven solution and approximate D^{-1} using a neural network.

To approximate D^{-1}, we use a six-block Resnet [16, 45, 46], whose input consists of three channels (see Fig. 5). The first channel is $I(\mathbf{x}, f+m)$, the second channel is an image that represents $P_{f,m}$, denoted by $G(P_{f,m})$, and the third channel is a 'constant image' with value $0.1m$. The second and third channels encode information about the expected position of new points (to be removed by the network) and the amount of expected point diffusion, respectively. The image $G(P_{f,m})$ is a heat map with values in $[0, 1]$ which 'tells' the network the likely spatial positions of the new points $P_{f,m}$. But, as illustrated in Fig. 6, because the object might be moving, the observed dots' location will deviate from their predicted positions in $P_{f,m}$. Moreover, the deviation amount depends on the time difference between the dot projection time and the frame sample time $|t_n - t_{f+m}|$. The greater the time gap, the more likely it is to observe a more significant deviation. To account for the deviation from $P_{f,m}$, we construct $G(P_{f,m})$ by assigning projection-order-depended spatial uncertainty to each point as described in the supplementary.

Frame scaling and object room temperature. Dynamic tasks rely on the spatial distribution of the thermal frames (*i.e.*, the resulting imaged pattern), while the absolute temperature readings are irrelevant. Therefore, we scale the thermal frame's pixel values to fit $[0, 1]$ using:

$$I(\mathbf{x}, f) = (I^{\text{raw}}(\mathbf{x}, f) - a)/b, \tag{8}$$

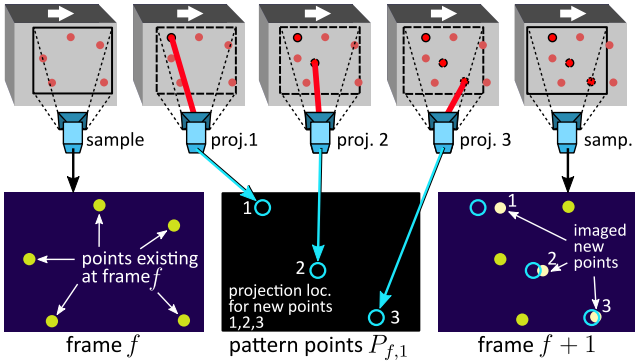

Figure 6. Motion during point projection. Three laser points indexed $1, 2$ and 3 are projected between frames f and $f+1$. The points' image plane projection coordinates are shown in cyan. However, the object moves to the right during the sequential point projection, causing the imaged point locations in frame $f+1$ to deviate from the projection coordinates. Point 1's deviation is larger than point 3 because point 3 is closer to the sampling time of $f+1$.

where $I^{\text{raw}}(\mathbf{x}, f)$ is raw 16-bit frame, and a, b are constants. But can our network learn to reverse dot diffusion at various room temperatures from just a few examples? In the supplementary, we show that for small variations around room temperature (*i.e.*, tens of degrees Kelvin), the camera response is approximately affine with $\boldsymbol{T}$:

$$S(\boldsymbol{T}) = \frac{c_1}{\exp \frac{c_2}{c_3 \boldsymbol{T} + c_4} - 1} \approx c_5 \boldsymbol{T} + c_6, \tag{9}$$

where c_5 and c_6 are constants. At any pixel $\mathbf{x}_0$, combining Eqs.(8-9) and derivating with respect to $\boldsymbol{T}$ yields:

$$\frac{\delta I(\mathbf{x}_0, .)}{\delta \boldsymbol{T}} = \frac{\delta}{\delta \boldsymbol{T}} \left(\frac{1}{b} \left(c_5 \boldsymbol{T} + c_6 - a \right) \right) = \frac{c_5}{b} \tag{10}$$

Eq. (10) shows that our scaling maintains the camera's affine response to $\boldsymbol{T}$, suggesting that the network can be roughly invariant to the object's ambient temperature under the conditions above.

5.2. Training procedure

As training data, we gather sequences of static scenes while projecting P. Each video provided many frame pairs of $I(\mathbf{x}, f)$ and $I(\mathbf{x}, f+m)$. Because the scene is static, we can define the loss function for a single pair as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}} || I(\mathbf{x}, f) - \hat{D}^{-1}\left(I(\mathbf{x}, f+m) \mid P_{f,m}, m\right)) ||_2^2, \quad (11)$$

$$\hat{D}^{-1}(\cdot) \equiv I(\mathbf{x}, f+m) - F_\theta\big(I(\mathbf{x}, f+m) \mid P_{f,m}, m\big), \quad (12)$$

where $F_\theta()$ denotes the PDR network. Unlike in dynamic scenes, the imaged pattern dot locations in static scenes will never deviate from the predicted points given by $P_{f,m}$. Therefore, training using the loss in Eq. (11) does not expose the network to examples representing scenes with motion. To correct for that, during training, we add a random spatial shift to each dot location in $P_{f,m}$ before generating image $G(P_{f,m})$. The random spatial shift magnitude is proportional to $|t_n - t_{f+m}|$ to account for the difference in timing between point projection and frame sample—more details in supplementary.

6. Application to Dynamic Tasks

In this Section, we detail how to employ the framework presented in Sections 4-5 for two dynamic tasks: Structure-from-Motion and computing optical flow. For both tasks, the system continuously captures the dynamic scene while projecting a dot pattern as described in Sec. 4.2. The optical flow is computed between pairs of temporally adjacent frames $I(\mathbf{x}, f)$ and $I^{\text{pdr}}(\mathbf{x}, f+m)$, where I^{pdr} is the output of the projection-diffusion reversal operator [3]

$$I^{\text{pdr}}(\mathbf{x}, f+m) \equiv \hat{D}^{-1}\left(I(\mathbf{x}, f+m)\right). \quad (13)$$

In SfM, the captured video frames are processed sequentially to detect and track each projected point from its first appearance until its visual quality drops below a predefined threshold, as detailed in the supplementary. The point tracking yields a plurality of point correspondences between temporally adjacent frames, which are manually fed into an off-the-shelf multi-view stereo pipeline to generate the camera trajectories and scene geometry (see Fig. 7).

Because the network's raw output $F_\theta\big(I(\mathbf{x}, f+m)\big)$ is subtracted from $I(\mathbf{x}, f+m)$ to reverse projection-diffusion, the output image's positive values belong to the newly added points $P_{f,m}$ (to be removed), while its negative values correspond to point intensity lost due to heat diffusion (to be reinforced). Therefore, we use the image

$$I^{\text{new}}(\mathbf{x}, f+m) \equiv \max(F_\theta\big(I(\mathbf{x}, f+m)\big), 0) \quad (14)$$

as input to a keypoint detector to add the newly projected points for tracking. Similarly to the optical flow case, we use the 'fully-reversed' frame $I^{\text{pdr}}(\mathbf{x}, f+m)$ as input to track points existing in $I(\mathbf{x}, f)$ (see Fig. 7).

[3] We drop the conditional terms in $\hat{D}^{-1}\big(\ldots \mid P_{f,m}, m\big)$ for brevity.

7. Prototype and Training Details

Hardware prototype. Our camera-projector system consists of a green laser and a FLIR thermal camera with a resolution of 640×512. The laser is steered toward the scene using a dual-axis galvo system with analog inputs generated by a data acquisition device (DAQ). We used a 150 mW for all experiments except the cart experiment of Fig 9, where a 1 W was used. Despite the high wattage of the latter laser, the power and temperature increase per surface point remains small due to the rapid point cycling. We capture video at 30 Hz and synchronize the galvo and camera by triggering both using individual but synchronized clocks generated by an Arduino Due board. See supplementary for a detailed part list, system image, and calibration details.

Projection-diffusion reversal network training. To train the network, we gathered a dataset of 13 static scenes with different materials and distances from the system. Each scene was a video lasting between 20 to 45 s, in which pattern projection is interleaved with periods having no pattern. The pattern-free periods are designed to cool off the accumulated surface heat caused by the continuous projection on the same object surface (as opposed to dynamic scenes where the projected surfaces continuously shift). We trained the network to reverse diffusion for one and two frames ($m = 1, 2$). The scaling in Eq. (8) was computed by finding the lowest and highest raw camera readings across all dataset scenes. See more details in the supplementary.

8. Experimental Evaluation

We tested our method for various dynamic vision tasks, projecting $K = 2, 3$ or 4 new points per frame in all experiments, depending on the material, the laser, and the distance to the camera. In generating frame pairs, we used $m = 1$ for large scene motions or $m = 2$ for small motions. In SfM, new points were detected using a Shi-Tomasi corner detector [31] and tracked using an ad hoc implementation based on the Lucas-Kanade method [3].

In Fig 7, we compute SfM on low-albedo, mostly textureless objects for which applying SfM directly on RGB frames fails. Nevertheless, our system generated accurate point correspondences, which were manually fed to COLMAP to yield precise camera motions and sparse 3D object shapes [29, 30]. Since generating ground truth camera motion for our textureless objects is hard (COLMAP failed on the RGB sequences), we assess the recovered motions by computing the motion's deviation from the expected 'perfect' circle. Thus, Fig. 7's yellow, blue, and green results with PDR fit a circle with R^2 of $0.99996, 0.9991, 0.999$, respectively. The blue and green results without PDR yield a fitting of 0.94 and 0.99, each. To assess shape recovery, we added texture to the plastic

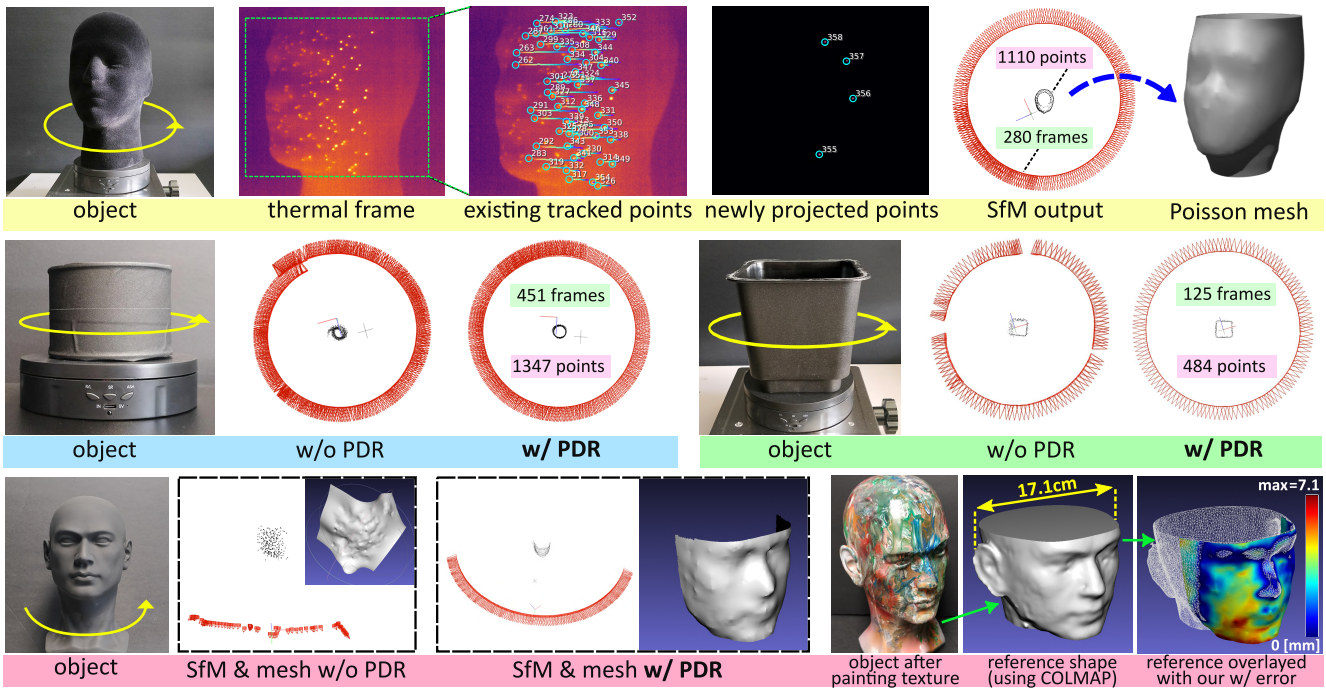

Figure 7. Structure from Motion. We scan dark, mostly textureless objects on a rotating stage. Our system continuously projects new 'heat points' on the object's surface during motion. The camera and projector lack any baseline, and the SfM relies solely on the tracked heat points. **(Top row)** Velvet-like styrofoam head. The third and fourth subfigures show tracked points and newly projected points outputted by the projection-diffusion reversal (PDR) operator, respectively. The stage rotates about $440°$. The recovered camera positions are highly accurate, fitting a circle with a coefficient of determination of R^2=0.99996. **(Middle row)** Scanning a cylindrical can and a rectangular planter. SfM without the PDR operator yields degraded results. **(Bottom row)** Black plastic head scan of about $160°$ rotation. Here, the reconstruction without the PDR operator failed to generate a face and register all the frames. We texturized the head with paint to recover a good reference mesh using COLMAP. After alignment, the maximum error between the reference and our recovery was $7.1\,\mathrm{mm}$.

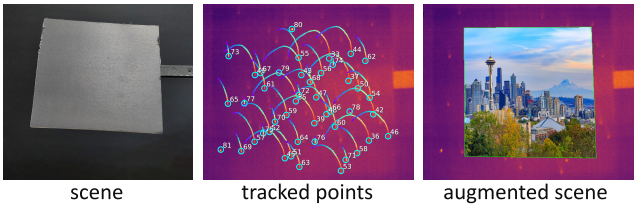

Figure 8. Augmented reality. Tracking the featureless plane's 3D pose enables superimposing a picture on the plane's surface.

head of Fig. 7(Bottom) using paint and recovered its shape with COLMAP. Post alignment and true-size scaling, the mean and max MeshLabs's Hausdorff distances between the painted head and our recovered mesh were $1.5\,\mathrm{mm}$ (0.8% w.r.t. shape scale) & $7.1\,\mathrm{mm}$ (3.7%), each [8].

In Fig. 8, we show an augmented reality application by projecting on a featureless black sheet and using the tracked points to superimpose a picture on the plane's surface. In Fig. 9, we put the system on a cart and use it for indoor localization. We direct the system towards the floor using a mirror and traverse in a loop around the office desks spanning around $20\,\mathrm{m}$. Fig. 9 shows that our system provides good camera motion tracking. The recovered motion shows good loop closure despite the lack of any matches connecting the first and last frames. For reference, we computed the cart's motion using a GoPro camera attached to the cart and directed at the room. The GoPro was affixed on the cart's right, about $25\,\mathrm{cm}$ away from the thermal camera, which explains the broader loop in Fig. 9. The mean absolute error compared to the reference was on the order of $14\,\mathrm{cm}$. Fig. 9's recovery used our PDR network; The supplement discusses the PDR performance gap in the cart experiment.

In Fig. 10, we apply our method to compute the optical flow for both rigid (Top row) and non-rigid object motions (Bottom row). We use RAFT to compute the pair-wise flow [37], where the current frame is corrected with the PDR network. Observe that the projected points yield sufficient texture to compute the flow of otherwise textureless objects. The recovered flow of Fig. 10(Top row) fits the expected circular flow model with R^2=0.98.

We tested the projection-diffusion network's effect on the downstream vision tasks. As expected, the network had a significant effect in cases where the image pairs exhibited a large appearance deviation. Such deviations occur for ma-

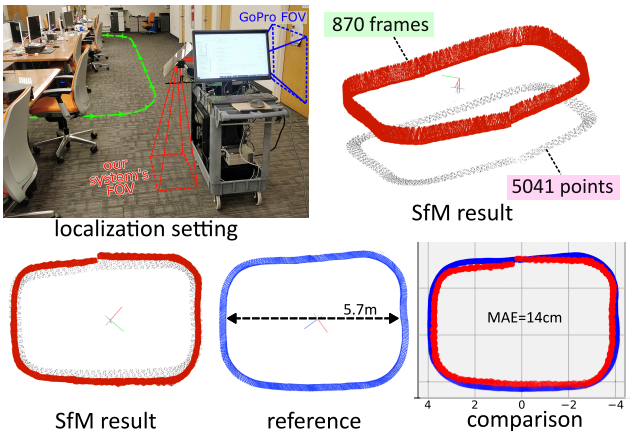

Figure 9. Indoor localization. We loop around office desks while pointing the system towards the floor. Our method yields accurate tracks and loop closure, with no shared feature points between the first and last frames. The blue reference camera poses stem from a COLMAP reconstruction using frames from a GoPro camera affixed to the cart's outer end, pointed at the office.

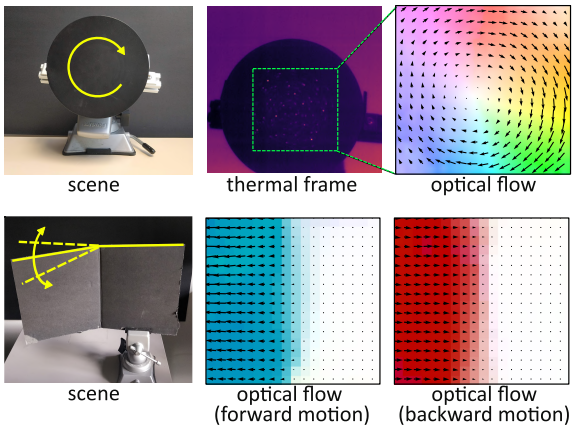

Figure 10. Optical Flow. **(Top row)** A rotating stage is positioned to face our system. We use RAFT to compute the optical flow between every pair of consecutive frames. **(Bottom row)** Non-rigid object motion (left plane folds while the right remains fixed).

terials that rapidly diffuse the heat or when the input frames are further apart (*i.e.*, $m > 1$). See Fig. 7 for visual comparisons. Conversely, the network had a negligible effect on materials where the point diffusion was relatively slow and the individual points had little overlay (*e.g.*, Fig. 9).

9. Discussion and Limitations

Loop closure. The thermal patterns are transient and disappear after a short time. Thus, at each frame, the existing pattern is only consistent with temporally adjacent frames, making feature-based loop closure impossible when relying *solely* on the projected patterns. Nevertheless, most environments should have some regions with sufficient natural infrared texture to allow loop closure.

Laser illumination. When using a laser to generate heat, unintended direct and specular reflection may pose safety issues for applications involving humans and animals. However, an 'eye-safe' laser with emission wavelengths greater than $1.4\,\mu\text{m}$ may be used instead of a green laser. See the supplementary for a laser safety discussion.

Material properties. We assume object materials that are responsive to heat projection. This responsivity, which manifests with pattern SNR and dot frame duration, depends on the material's absorption of the laser's wavelength (*i.e.* albedo), thermal conductivity and diffusivity, and emissivity in the camera's infrared range. Thus, our method will exhibit degraded performance on some materials like glass, metals, and more (see supplementary for examples).

Point density for 3D recovery. Our approach yields a sparse point cloud of the scene. However, expanding the model into a dense one using existing multi-view-stereo (MVS) approaches is not straightforward using thermal images. Nevertheless, future works may extend MVS approaches to handle the thermal domain and its particular characteristics (*e.g.*, heat diffusion and more).

Rapid scene motion. In this work, we assumed and verified experimentally that each dot projection time is short enough such that the scene is approximately static per dot. However, very rapid scene motions may smear the projected dots into curves. Expanding our method to accommodate these short curves can be achieved by including them in the training set (*e.g.*, by projecting short curves during training) and making slight adjustments to the input G channel.

10. Conclusion

This paper introduces a new class of structured light methods, where the projected patterns remain on the object in the thermal domain, combining the advantages of physical and projected patterns. By using a laser to project controllable heat patterns onto objects, we can remotely imprint patterns without altering an object's appearance, facilitating dynamic vision tasks for textureless surfaces. Like lidar, our system could be integrated into autonomous vehicles and robots to ease navigation in challenging environments, both urban and industrial (*e.g.*, inside pipelines or tunnels). Our work also provides a pathway for encoding digital codes into the thermal patterns, enabling communication between different agents [2]. Our approach could facilitate other studies like recovering material properties, structural analysis, medical imaging, and even artistic expression in fields such as computer graphics.

Acknowledgements: This work was partly supported by NSF grants IIS-2107236, CCF-1730147, and NSF-NIFA AI Institute for Resilient Agriculture. We thank J. Luiten for help with experiments and M. Ramanagopal for insightful discussions.

References

[1] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. Generating visible spectrum images from thermal infrared. In *Proc. CVPRW*, pages 1143–1152, 2018. 2

[2] Filippo Bergamasco, Andrea Albarelli, Emanuele Rodola, and Andrea Torsello. Rune-tag: A high accuracy fiducial marker with strong occlusion resilience. In *Proc. IEEE CVPR*, pages 113–120, 2011. 8

[3] Jean-Yves Bouguet et al. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation*, 5(1-10):4, 2001. 6

[4] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proc. IEEE CVPR*, pages 8709–8719, 2019. 2

[5] Chia-Yen Chen, Chia-Hung Yeh, Bao Rong Chang, Jun-Ming Pan, et al. 3d reconstruction from ir thermal images and reprojective evaluations. *Mathematical Problems in Engineering*, 2015, 2015. 2

[6] I-Chien Chen, Chang-Jen Wang, Chao-Kai Wen, and Shiow-Jyu Tzou. Multi-person pose estimation using thermal images. *IEEE Access*, 8:174964–174971, 2020. 2

[7] Junzhang Chen and Xiangzhi Bai. Atmospheric transmission and thermal inertia induced blind road segmentation with a large-scale dataset tbrsd. In *Proc. IEEE ICCV*, pages 1053–1063, 2023. 2

[8] Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, volume 17:2, pages 167–174. Blackwell Publishers, 1998. 7

[9] Wikipedia contributors. Irreversible process. https://en.wikipedia.org/wiki/Irreversible_process. 4

[10] Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Aggelos Katsaggelos, Ashok Veeraraghavan, and Oliver Cossairt. Thermal spread functions (tsf): Physics-guided material classification. In *Proc. CVPR*, pages 1641–1650, 2023. 2

[11] Jack Erdozain, Kazuto Ichimaru, Tomohiro Maeda, Hiroshi Kawasaki, Ramesh Raskar, and Achuta Kadambi. 3d imaging for thermal cameras using structured light. In *Proc. IEEE ICIP*, pages 2795–2799. IEEE, 2020. 2

[12] Lu Gan, Connor Lee, and Soon-Jo Chung. Unsupervised rgb-to-thermal domain adaptation via multi-domain attention network. In *Proc. IEEE ICRA*, pages 6014–6020. IEEE, 2023. 2

[13] Junfeng Ge, Yupin Luo, and Gyomei Tei. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298, 2009. 2

[14] Jingu Heo, Seong G Kong, Besma R Abidi, and Mongi A Abidi. Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. In *Proc. IEEE CVPRW*, pages 122–122. IEEE, 2004. 2

[15] Luocheng Huang, Zheyi Han, Anna Wirth-Singh, Vishwanath Saragadam, Saswata Mukherjee, Johannes E Fröch, Joshua Rollag, Ricky Gibson, Joshua R Hendrickson, Phillip WC Hon, et al. Broadband thermal imaging using meta-optics. *arXiv preprint arXiv:2307.11385*, 2023. 2

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE CVPR*, 2017. 5

[17] Ladislav Kocis and William J Whiten. Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software (TOMS)*, 23(2):266–294, 1997. 4

[18] Zülfiye Kütük and Görkem Algan. Semantic segmentation for thermal images: A comparative survey. In *Proc. IEEE CVPR*, pages 286–295, 2022. 2

[19] Tomohiro Maeda, Yiqin Wang, Ramesh Raskar, and Achuta Kadambi. Thermal non-line-of-sight imaging. In *Proc. IEEE ICCP*, pages 1–11. IEEE, 2019. 2

[20] ELEONORA Maset, Andrea Fusiello, Fabio Crosilla, R Toldo, and D Zorzetto. Photogrammetric 3d building reconstruction from thermal images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:25–32, 2017. 2

[21] Yasuto Nagase, Takahiro Kushida, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Shape from thermal radiation: Passive ranging using multi-spectral lwir measurements. In *Proc. CVPR*, pages 12661–12671, 2022. 2

[22] Max Planck. On the law of distribution of energy in the normal spectrum. *Annalen der physik*, 4(553):1, 1901. 2

[23] Marc Proesmans and Luc Van Gool. Reading between the lines—a method for extracting dynamic 3d with texture. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 95–102, 1997. 1

[24] Manikandasriram Srinivasan Ramanagopal, Zixu Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. Pixel-wise motion deblurring of thermal videos. *arXiv preprint arXiv:2006.04973*, 2020. 3

[25] Rafael E Rivadeneira, Angel D Sappa, Boris X Vintimilla, Dai Bin, Li Ruodi, Li Shengye, Zhiwei Zhong, Xianming Liu, Junjun Jiang, and Chenyang Wang. Thermal image super-resolution challenge results-pbvs 2023. In *Proc. IEEE CVPR*, pages 470–478, 2023. 2

[26] Fumihiro Sakuma and Susumu Hattori. Establishing a practical temperature standard by using a narrow-band radiation thermometer with a silicon detector. *Metrology Institute Report*, 32(2):p91–97, 1983. 2

[27] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanic, and Xavier Llado. A state of the art in structured light patterns for surface profilometry. *Pattern recognition*, 43(8):2666–2680, 2010. 1

[28] Vishwanath Saragadam, Zheyi Han, Vivek Boominathan, Luocheng Huang, Shiyu Tan, Johannes E Fröch, Karl F Böhringer, Richard G Baraniuk, Arka Majumdar, and Ashok Veeraraghavan. Foveated thermal computational imaging in the wild using all-silicon meta-optics. *arXiv preprint arXiv:2212.06345*, 2022. 2

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. IEEE CVPR*, 2016. 6

[30] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV*, 2016. 6

[31] Jianbo Shi et al. Good features to track. In *Proc. IEEE CVPR*, pages 593–600. IEEE, 1994. 6

[32] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proc. IEEE ICCV*, pages 11218–11228, 2023. 2

[33] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proc. IEEE CVPR*, pages 1043–1053, 2023. 2

[34] 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 *Proc. IEEE WACV*, pages 5798–5807, January 2023. 2

[35] Kenichiro Tanaka, Nobuhiro Ikeya, Tsuyoshi Takatani, Hiroyuki Kubo, Takuya Funatomi, Vijay Ravi, Achuta Kadambi, and Yasuhiro Mukaigawa. Time-resolved far infrared light transport decomposition for thermal photometric stereo. *IEEE TPAMI*, 43(6):2075–2085, 2019. 2

[36] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. What happened 3 seconds ago? inferring the past with thermal imaging. In *Proc. IEEE CVPR*, pages 17111–17120, June 2023. 2

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*, pages 402–419. Springer, 2020. 7

[38] 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 *Proc. IEEE CVPR*, pages 2961–2969, 2017. 2

[39] Andre Treptow, Grzegorz Cielniak, and Tom Duckett. Real-time people tracking for mobile robots using thermal vision. *Robotics and Autonomous Systems*, 54(9):729–739, 2006. 2

[40] Michael Vollmer. Infrared thermal imaging. In *Computer Vision: A Reference Guide*, pages 666–670. Springer, 2021. 2

[41] Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, and Hao Tang. Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In *Proc. IEEE ICCV*, pages 11122–11131, 2023. 2

[42] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proc. IEEE ICCV*, pages 11069–11079, 2023. 2

[43] Bekir Yilbas. *Laser heating applications: analytical modeling*. Elsevier, 2012. 2

[44] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proc. IEEE ICCV*, pages 11185–11195, 2023. 2

[45] Jun-Yan Zhu. pytorch-cyclegan-and-pix2pix. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix. 5

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE ICCV*, 2017. 5