VT-Intrinsic: Physics-Based Decomposition of Reflectance and Shading using a Single Visible-Thermal Image Pair

Zeqing Yuan Mani Ramanagopal Aswin C. Sankaranarayanan Srinivasa G. Narasimhan
Carnegie Mellon University

https://vt-intrinsic.github.io

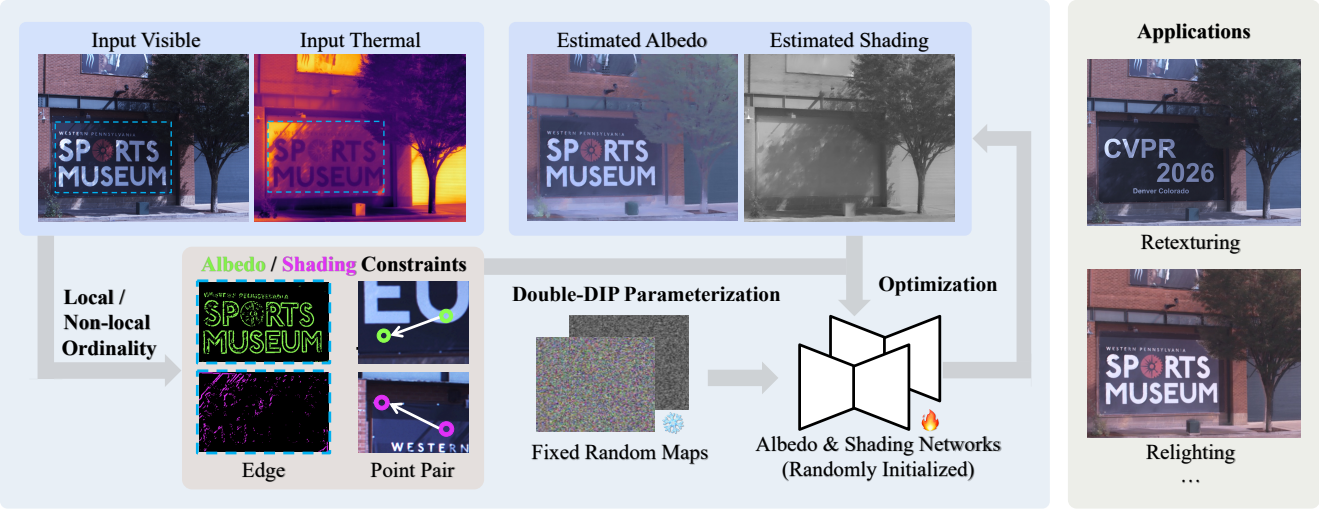

Figure 1. Through the tree's veil, sunlight weaves intricate shadows across a building façade. Visible and thermal images capture complementary cues of the reflected and absorbed light. Local and non-local visible–thermal ordinalities (Sec. 3) reveal albedo-/shading-dominant edges and point-pair ordinalities respectively, guiding an optimization using Double-DIP parameterization (Sec. 4). Our physics-based method reconstructs the complex shading and albedo without learned priors, whereas state-of-the-art models fail (see supplementary). The decomposition enables faithful retexturing and relighting via albedo and shading editing.

Abstract

Decomposing a scene into its reflectance and shading is a challenge due to the lack of extensive ground-truth data for real-world scenes. We introduce a novel physics-based approach for intrinsic image decomposition using a pair of visible and thermal images. We leverage the principle that light not reflected from an opaque surface is absorbed and detected as heat by a thermal camera. This allows us to relate the ordinalities (or relative magnitudes) between visible and thermal image intensities to the ordinalities of shading and reflectance. The ordinalities enable dense self-supervision of an optimizing neural network to recover shading and reflectance. We perform quantitative evaluations with known reflectance and shading under natural and artificial lighting, and qualitative experiments across diverse scenes. The results demonstrate superior performance over both physics-based and recent learning-based methods, providing a path toward scalable real-world data curation with supervision.

1. Introduction

Understanding how a scene appears from the interaction between *surface reflectance* (the intrinsic material color, often approximated as diffuse *albedo*) and *incident illumination* (*shading* determined by lighting and geometry) has long been a central pursuit in vision and imaging sciences [3]. Disentangling these physical factors is useful for various applications in graphics (recoloring, relighting, and compositing) and vision (object recognition and tracking). Recent learning-based methods have made progress by formulating this task in an end-to-end framework and inferring statistical priors from auxiliary datasets to constrain the otherwise ill-posed inverse problem [17]. However, collecting ground-truth data for real-world scenes remains infeasible, as measuring surface reflectance and shading requires specialized equipment and controlled procedures [19, 42].

In this paper, we introduce a novel physics-based framework that leverages a single auxiliary thermal image to decompose a visible image of a scene into its albedo and shad-

ing components. To see why a thermal image is useful here, we consider the underlying physical principles that govern albedo and shading. Shading corresponds to the total incident energy (or irradiance) at a scene point, while albedo represents the proportion of that energy reflected by the surface. For opaque objects, the unreflected portion of the incident energy is absorbed, which contributes to the thermal radiation. This radiation can be detected by a thermal camera in the long-wave infrared range (8–14 μm). A recent technique called JoLHT-Video [34] addressed this issue by modeling heat transport equations and estimating it from the transient heat flow observed by a thermal *video*. However, directly estimating the absorbed light is challenging as it requires controlled and visible-only lighting and transient thermal measurements from video, Inspired by this work, we pose the following question: *What can be achieved using only a single thermal image?*

Since absorption of light increases the temperature of an object, low-albedo regions—dark in the visible image—appear bright in the thermal image, whereas shading variations appear bright in both. Based on this observation, we relate visible–thermal intensity ordinalities between any two scene points to their albedo and shading ordinalities, *without* having to estimate the absorbed light. Specifically, the ordinality of neighboring scene points classifies edges as shading- or reflectance-dominant and defines an edge loss, while non-local ordinalities yield a point-pair loss. These new losses are used, together with the standard visible-image reconstruction loss, to optimize a neural network (e.g., Double Deep Image Prior [16], randomly initialized to parameterize albedo and shading), effectively providing dense self-supervision for intrinsic decomposition.

Our ordinality theory is derived using the Lambertian assumption and when illumination is confined to the visible spectrum (e.g., LED lighting). We further extend it to broadband sources containing infrared energy (e.g., sunlight, incandescent bulbs) by empirically observing—and statistically validating with [22]—that infrared albedo exhibits lower variation across common materials than visible albedo [9], thereby preserving ordinalities. Expert validation on diverse materials and natural scenes—including those moderately violating the Lambertian assumption—shows near-perfect agreement between our automatically estimated point-pair ordinalities and confident expert labels, confirming robustness across material types and generalization beyond idealized conditions.

We quantitatively evaluate our method on scenes with known reflectance (e.g. color charts) and known shading (e.g. object imaged under identical lighting but painted differently). We further test on visible-thermal pairs simulated from the MIT Intrinsic dataset [19]. Finally, we demonstrate qualitative results on complex indoor and outdoor scenes with notable improvements over both physics-based and learning-based methods trained on auxiliary datasets. Our dataset of visible-thermal image pairs can also provide supervision for learning methods in real-world scenes.

Limitations: Our method assumes dominant diffuse reflection, heat arising primarily from light absorption, and the absence of multiple colored illuminations. Performance can also be affected by the limited SNR and resolution of inexpensive microbolometer thermal cameras, particularly under weak illumination or in dynamic scenes. Despite these limitations, as thermal cameras improve and become more ubiquitous, understanding the interplay between light and heat holds strong potential for vision and graphics.

2. Related Work

2.1. Thermal Imaging for Physics-Based Vision

Thermal cameras have recently emerged as a powerful complement to visible sensing across geometry [32, 33, 39], materials [12], and appearance [2].

JoLHT-Video [34] showed that transient thermal video provides analytical cues for estimating absorbed and reflected light. However, it relies on (i) active lighting control, infeasible in-the-wild, (ii) visible-only illumination that excludes common sources (sun and incandescent bulbs), and (iii) radiometric calibration between visible and thermal cameras. In contrast, our method uses a *single* steady-state thermal image to extract reliable ordinal constraints without requiring video, calibration or lighting control.

2.2. Intrinsic Image Decomposition (IID)

Early optimization-based approaches: Retinex-style [25] methods rely on stringent assumptions that hinder generalization—such as smooth shading or reflectance [4], chromaticity-preserving shading variations [11, 14, 18], or local intensity similarity implying shared reflectance [37].
Learning-based approaches: Unsupervised learning based methods that decorrelate albedo and shading [29] or that enforce albedo consistency across changing illumination [27] improve upon hand-crafted priors. Supervised learning-based models are primarily trained on synthetic datasets [24, 26, 28, 35], which provide ground-truth albedo and shading but face a significant sim-to-real gap. Existing real-world datasets [5, 23, 42] offer sparse annotations used by models (e.g., to predict albedo ordinalities [44]), but are limited to small-scale indoor scenes. Intrinsic-v1 [6] expands to more diverse data by using model predictions as pseudo-ground truth, albeit imperfectly. Recent works [21, 30, 43] leverage diffusion priors for IID, yet as noted in [7], they suffer from hallucination.
Using auxiliary sensors: Cheng et al. [8] used near-infrared (NIR) images as shading proxies, but NIR albedo often varies across materials (albeit less than visible) and modern efficient lighting such as LEDs hardly emits NIR, limiting

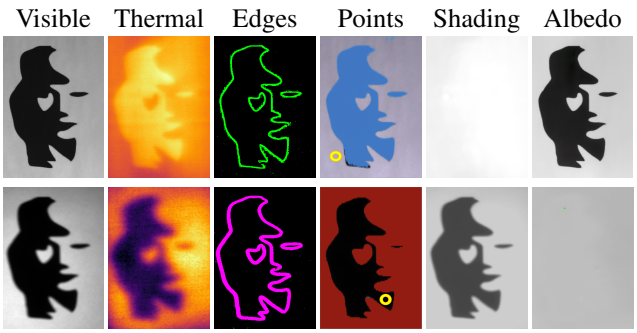

Figure 2. Printed (top) vs. projected (bottom) Roger Shepard's illusion [38]. **Top**: a *printed* paper lit by an incandescent bulb, where reflectance variations reveal a saxophone player. **Bottom**: the same pattern *projected* onto a uniform cardboard, where modulated shading reveals a lady's face. This comparison highlights the albedo-shading ambiguity and motivates modeling light–heat transport: reflectance induces inverse visible–thermal ordinalities, while shading yields consistent ones. Columns 3-4 show classified albedo- / shading-dominant edges (Sec. 3.1.1) and points of lower albedo / higher shading than ○ (Sec. 3.1.2). Our method decomposes correctly (right), whereas baselines fail (see supplementary).

generality and applicability. Sato et al. [36] used intensity of sparse LiDAR returns and enforce consistency with estimated albedos, yet LIDAR operates in NIR where albedo differs from visible [8]. While such NIR cues help in specific cases, our approach exploits the complementary relation between visible (reflected light) and thermal (proxy for absorbed light), enabling broader applicability.

3. Theory of Visible-Thermal Ordinality

We present the theoretical relationship between a visible and thermal image pair and show that the ordinality of their pixel intensities directly convey the ordinality of the underlying albedo or shading, as illustrated in Fig. 2. We first consider visible illumination (e.g., an LED), then extend our analysis to include invisible illumination component (e.g., infrared light in incandescent bulbs and sunlight).

3.1. Visible-only Illumination

Consider an opaque Lambertian scene imaged by visible and thermal cameras. The visible intensity at a pixel x is:

$$I_v(x) = g\rho(x)\eta(x), \tag{1}$$

where ρ is the albedo (reflectance), η the shading (irradiance), and $g = k/\pi$ a global scale determined by the camera gain k. For brevity, we omit x denoting a single pixel.

Light not reflected is absorbed by the surface and converted into heat, producing a heat source of intensity:

$$\mathcal{H} = (1-\rho)\eta. \tag{2}$$

This heat propagates through conduction, convection and radiation according to the heat transport equation. Though $\mathcal{H}$ is not directly measurable, it can be inferred from surface temperature, which is indirectly observed by a thermal camera. JoLHT-Video [34] modeled the light-heat transport using thermal video of the heating process to estimate $\mathcal{H}$ and solve for ρ and η. In contrast, we use a single thermal image I_t near thermal equilibrium, easily attainable within seconds under stable lighting.

Together, I_v and the absorbed heat image $\mathcal{H}$ impose local and non-local constraints on albedo and shading.

3.1.1. Local (Edge) Constraints

The spatial gradient of the visible image can be written as:

$$\nabla I_v = g(\nabla\rho)\eta + g\rho(\nabla\eta). \tag{3}$$

For most edges in natural images, one of the two terms on the right dominates—edges arise primarily from either albedo or shading variations [18, 25]. This creates a fundamental ambiguity, but the spatial gradients of the heat source provide complementary information:

$$\nabla\mathcal{H} = (-\nabla\rho)\eta + (1-\rho)\nabla\eta. \tag{4}$$

From (3) and (4), we have

$$\textbf{Albedo Edge}(\nabla\eta \to 0)\textbf{:}\ \text{sign}(\nabla I_v) = -\text{sign}(\nabla\mathcal{H}), \tag{5a}$$
$$\textbf{Shading Edge}(\nabla\rho \to 0)\textbf{:}\ \text{sign}(\nabla I_v) = \text{sign}(\nabla\mathcal{H}). \tag{5b}$$

This yields a simple criterion to distinguish albedo- and shading-dominant edges using the heat intensity image.

3.1.2. Non-Local (Point-Pair) Constraints

We generalize the above gradient analysis to compare point pairs, i.e., any two distinct pixels x_i and x_j in the scene.

$$I_v(x_i) = g\rho(x_i)\eta(x_i),\ \mathcal{H}(x_i) = (1-\rho(x_i))\eta(x_i), \tag{6a}$$
$$I_v(x_j) = g\rho(x_j)\eta(x_j),\ \mathcal{H}(x_j) = (1-\rho(x_j))\eta(x_j). \tag{6b}$$

When a pixel's visible intensity is lower (or higher) than another's while its thermal intensity is higher (or lower), the pixel's albedo is correspondingly lower (or higher).

Proposition 1 (Albedo Ordinality). *For pixels x_i, x_j with visible and heat intensities as in (6), if $I_v(x_i) < I_v(x_j)$ and $\mathcal{H}(x_i) > \mathcal{H}(x_j)$, then $\rho(x_i) < \rho(x_j)$, and vice versa.*

Conversely, when both visible and heat intensities are lower (or higher), its shading is also lower (or higher).

Proposition 2 (Shading Ordinality). *For pixels x_i, x_j with visible and thermal intensities as in (6), if $I_v(x_i) < I_v(x_j)$ and $\mathcal{H}(x_i) < \mathcal{H}(x_j)$, then $\eta(x_i) < \eta(x_j)$, and vice versa.*

Proofs for Prop. 1 and Prop. 2 are in the supplementary. The ordinalities here rely on $\mathcal{H}$, the heat from absorbed visible light. Next, we incorporate invisible light and relate $\mathcal{H}$ to thermal image intensity, I_t.

3.2. Visible and Invisible Illumination

Common light sources such as sunlight and incandescent lamps emit significant invisible radiation (e.g., infrared). While the visible camera captures only reflected light within its spectral response, heat generation arises from absorbed energy across all wavelengths. Thus, the equation for the heat source intensity has an additional term as follows:

$$\mathcal{H} = (1-\rho_v)\eta + (1-\rho_i)\frac{l_i}{l_v}\eta, \tag{7}$$

where ρ_i is the average albedo in the invisible band, l_i/l_v is the ratio of light intensity in the invisible and visible spectra.

While albedo variations are prominent in the visible spectrum, their counterparts in the infrared are much smaller [9]. Thus, we assume that ρ_i is locally constant within a region, allowing (7) to be approximated as:

$$\mathcal{H} = (\beta-\rho_v)\eta, \quad \text{s.t.} \quad \beta = 1+(1-\rho_i)l_i/l_v. \tag{8}$$

As β is locally constant, (5b) still holds as $\nabla\mathcal{H}$ is invariant to a constant offset in $\mathcal{H}$. Also, as $\beta > 1$, Prop. 1 and Prop. 2 holds whenever β is same for the two points.

3.3. Relating heat intensity to a single thermal image

While heat intensity is not directly observable, we show that a thermal image is a reliable proxy for ordinality constraints.

The heat transport equation at a surface point is:

$$\mathcal{C}_h\frac{\partial T}{\partial t} = \mathcal{H} + h_c(T_a - T) + 4\epsilon\sigma T_s^3(T_s - T) + \kappa\Delta T, \tag{9}$$

where $\mathcal{C}_h$ is the heat capacity, T the surface temperature, t time, h_c the convection coefficient, T_a the air temperature, ϵ the surface emissivity, σ the Stefan-Boltzmann constant, T_s the surrounding temperature, κ the thermal conductivity, and Δ denotes the Laplacian operator along the surface. A static scene under constant lighting reaches thermal equilibrium when the left side of (9) is zero, giving

$$\mathcal{H} = (h_c + 4\epsilon\sigma T_s^3)T - \kappa\Delta T - (h_c T_a + 4\epsilon\sigma T_s^4). \tag{10}$$

The image intensity measurement T_t made by a thermal camera is related to the temperature T as follows:

$$I_t = \epsilon U(T) + (1-\epsilon)U(T_s), \tag{11}$$

where U denotes the thermal camera's response function. Linearizing U as $U(T) = p_1T + p_2$ in (11), we get

$$T = a_1 I_t - a_2 \quad \text{s.t. } a_1 = \frac{1}{\epsilon p_1},\ a_2 = \frac{p_2 + p_1 T_s(1-\epsilon)}{\epsilon p_1}. \tag{12}$$

Substituting (12) in (10), we get

$$\mathcal{H} = c_1 I_t - c_2\Delta I_t - c_3, \tag{13}$$

where $c_1 = \frac{h_c + 4\epsilon\sigma T_s^3}{\epsilon p_1}$, $c_2 = \frac{\kappa}{\epsilon p_1}$, and $c_3 = (h_c + 4\epsilon\sigma T_s^3)(\frac{p_2 + p_1 T_s(1-\epsilon)}{\epsilon p_1}) + (h_c T_a + 4\epsilon\sigma T_s^4)$.

The thermal properties such as ϵ, and κ have small variations irrespective of the variation in albedo [41]. The environmental variables such as h_c, T_a, and T_s are also similar. Therefore, c_1, c_2 and c_3 are similar within a region. Also, thermal conductivity of many common materials, excluding metals, is low. Likewise, the Laplacian of a temperature field at steady state has a much smaller magnitude than absolute temperatures [41]. Therefore, we ignore the conduction term. Then, as $c_1 > 0$, the ordinal relationships between $\mathcal{H}$ at two points is the same as that of I_t.

Proposition 3. *In local regions, c_1, c_2 and c_3 are constant so that for any two pixels x_i, x_j, if $\mathcal{H}(x_i)$ is less (or more) than $\mathcal{H}(x_j)$, then $I_t(x_i)$ is also less (or more) than $I_t(x_j)$.*

3.4. Ordinality of Albedo and Shading

Using Prop. 3, we can extend the results from Eq. 5b to use thermal image intensities, as summarized below:

$$\textbf{Albedo Edge}(\nabla\eta = 0)\textbf{:}\ \text{sign}(\nabla I_v) = -\text{sign}(\nabla I_t), \tag{14a}$$
$$\textbf{Shading Edge}(\nabla\rho = 0)\textbf{:}\ \text{sign}(\nabla I_v) = \text{sign}(\nabla I_t). \tag{14b}$$

Similarly, we extend Prop. 1 and Prop. 2 to thermal image intensities, yielding the following ordinal relationships:

$$I_v(x_i) > I_v(x_j), I_t(x_i) > I_t(x_j) \Rightarrow \eta(x_i) > \eta(x_j), \tag{15a}$$
$$I_v(x_i) < I_v(x_j), I_t(x_i) < I_t(x_j) \Rightarrow \eta(x_i) < \eta(x_j), \tag{15b}$$
$$I_v(x_i) > I_v(x_j), I_t(x_i) < I_t(x_j) \Rightarrow \rho(x_i) > \rho(x_j), \tag{15c}$$
$$I_v(x_i) < I_v(x_j), I_t(x_i) > I_t(x_j) \Rightarrow \rho(x_i) < \rho(x_j). \tag{15d}$$

4. Method

Using the ordinalities as loss functions, we optimize the albedo and shading from a visible-thermal image pair. Let I_v be a $k-$channel visible image and I_t be the corresponding aligned thermal image. Let $\hat{\rho}$ and $\hat{\eta}$ be an estimate of the $k-$channel albedo and grayscale shading. Let $\bar{I}_v$ and $\bar{\rho}$ be the grayscale image and albedo estimate, respectively.

4.1. Local (Edge) Loss

Using Eq. 14, we label edges (A for albedo, S for shading) based on their local visible–thermal gradients (Fig. 2):

$$\mathcal{C}(x) = \begin{cases} \text{A} & |\nabla\bar{I}_v| > \epsilon_m, |\frac{\nabla\bar{I}_v \nabla I_t}{\|\nabla\bar{I}_v\|\|\nabla I_t\|}| > \epsilon_p, \\ \text{S} & |\nabla\bar{I}_v| > \epsilon_m, |\frac{\nabla\bar{I}_v \nabla I_t}{\|\nabla\bar{I}_v\|\|\nabla I_t\|}| < \epsilon_p, \end{cases} \tag{16}$$

where ϵ_m suppresses textureless regions and ϵ_p thresholds the cosine similarity between visible and thermal gradients.

Before computing ∇I_t, we apply Gaussian smoothing to reduce noise while maintaining gradient consistency.

With the class labels above, we formulate an edge loss that penalizes albedo gradients at shading-dominant pixels and vice versa, where Ω denotes all image pixels:

$$\mathcal{L}_{\text{edge}}(\bar{\rho}, \hat{\eta}, \mathcal{C}) = \frac{1}{|\Omega|}\Big[\sum_{\mathcal{C}(x)=S}\|\nabla\bar{\rho}(x)\|^2 + \sum_{\mathcal{C}(x)=A}\|\nabla\hat{\eta}(x)\|^2\Big]. \tag{17}$$

4.2. Non-Local (Point-Pair) Loss

During optimization, we use Poisson disk sampling [5] to generate random point pairs across the image. Using Eq. 15, each pair (x_i, x_j) is assigned a class label based on their normalized intensity differences δI_v and δI_t:

$$\mathcal{P}(x_i, x_j) = \begin{cases} S_+ & \delta I_v > \epsilon_d,\ \delta I_t > \epsilon_d, \\ S_- & \delta I_v < -\epsilon_d,\ \delta I_t < -\epsilon_d, \\ A_+ & \delta I_v > \epsilon_d,\ \delta I_t < -\epsilon_d, \\ A_- & \delta I_v < -\epsilon_d,\ \delta I_t > \epsilon_d, \end{cases} \tag{18}$$

where $\delta I_x = \frac{I_x(x_i) - I_x(x_j)}{Z_x}$ with normalization Z_x so that threshold ϵ_d is relative. The ordinal loss is a hinge-based formulation that enforces separation beyond a margin ε_m:

$$\mathcal{L}_{\text{ord}} = \frac{1}{|\mathcal{P}|} \sum_{(x_i, x_j)} \begin{cases} \max(\hat{\eta}_j - \hat{\eta}_i + \varepsilon_m, 0), & \mathcal{P}(x_i, x_j) = S_+, \\ \max(\hat{\eta}_i - \hat{\eta}_j + \varepsilon_m, 0), & \mathcal{P}(x_i, x_j) = S_-, \\ \max(\bar{\rho}_j - \bar{\rho}_i + \varepsilon_m, 0), & \mathcal{P}(x_i, x_j) = A_+, \\ \max(\bar{\rho}_i - \bar{\rho}_j + \varepsilon_m, 0), & \mathcal{P}(x_i, x_j) = A_-. \end{cases} \tag{19}$$

4.3. Regularization using Deep Image Prior

In complex real scenes, thermal noise can corrupt subtle gradients, and ordinal constraints alone cannot fully determine absolute albedo or shading values—they only restrict the solution space. Therefore, we adopt a variant of the Deep Image Prior [40] to parameterize albedo and shading, leveraging the inherent architectural prior in a randomly initialized network for regularization.

We employ a Double-DIP (DDIP) architecture [16] with two networks $\mathcal{N}(z_A, \Theta_A), \mathcal{N}(z_S, \Theta_S)$ to parameterize albedo and shading, respectively. Each uses a convolutional encoder–decoder with skip connections [40]. Θ_A, Θ_S are randomly initialized model weights and z_A, z_S are randomly sampled input noise vectors. The albedo network outputs a 3-channel image bounded to $[0, 1]^3$ via a sigmoid activation, while the shading network predicts a single channel constrained by a non-negativity penalty. We freeze z_A and z_S while only optimizing for Θ_A and Θ_S.

4.4. Optimization

Our complete objective function is as follows.

$$\mathcal{L}(\hat{\rho}, \hat{\eta}, I_v, I_t) = \|\hat{\rho} \cdot \hat{\eta} - I_v\|_2 + \lambda_1 \mathcal{L}_{\text{edge}}(\bar{\rho}, \hat{\eta}, \mathcal{C}(\bar{I}_v, I_t)) + \lambda_2 \mathcal{L}_{\text{ord}}(\bar{\rho}, \hat{\eta}, \mathcal{P}(\bar{I}_v, I_t)), \tag{20}$$

where $\lambda_1, \lambda_2 > 0$ are the respective loss weights. The thermal image is used only for edge or point pair losses, which operate on the mean albedo. The reconstruction loss is defined on the $3-$channel image.

5. VT-Intrinsic Dataset

Existing IID datasets lack thermal modalities, while current visible–thermal datasets (e.g., captured from vehicles or drones) focus on dynamic objects (people, cars) or non-light-absorption based heat sources (engines, people) that are out of scope for this work. So, we collected 600 visible–thermal image pairs (Fig. 3) across diverse stationary scenes under varying illumination to validate our method.

Figure 3. Visible–thermal image pair examples in the VT-Intrinsic dataset, covering diverse scenes including parks, schools, cathedrals, plazas, museums, and various urban streets.

Imaging System. We co-locate a FLIR Boson thermal camera (512×640 resolution, $24°$ HFOV, $\leq$ 50mK NEDT) with an IDS UI-3130 color camera (600×800 resolution, $27°$ HFOV) using a gold dichroic mirror (BSP-DI-25-2). For distant outdoor scenes, the cameras are placed side by side and aligned via homography.

Data Acquisition and Preprocessing. We captured 20 exposure-bracketed color images with geometrically spaced exposure times and merged them into a linear HDR image [13] after edge-aware demosaicing in OpenCV. Five frames were averaged to suppress sensor noise. The visible HDR and thermal images were aligned via homography.

Our dataset contributes in two key aspects: (i) a large collection of high-quality real-world outdoor images with diverse albedo–shading combinations (vs. predominantly indoor/synthetic prior datasets); (ii) abundant pseudo-ground-truth albedo and shading ordinalities for training and evaluation. As shown in Sec. 6.2.1, the thermal image produces reliable albedo/shading ordinalities across *arbitrary pixel pairs*—previously only available in limited form due to costly human labeling (IIW [5]).

6. Experiments

Datasets: As typical IID datasets lack associated thermal images, we construct the VT-Intrinsic dataset for qualitative evaluation. Obtaining ground truth albedo and shading for real-world scenes is impractical. Therefore, for quantitative evaluation, we collected images of a color chart under different illuminations: white LED light, incandescent bulb and sunlight. We also evaluate on the JoLHT-Video dataset [34], which contains four scenes of a color chart under varied illuminations and a *Painted-Mask* scene. In the supplementary, we evaluate on the MIT-Intrinsics [19] dataset by simulating an ideal thermal image using their pseudo-ground truth.

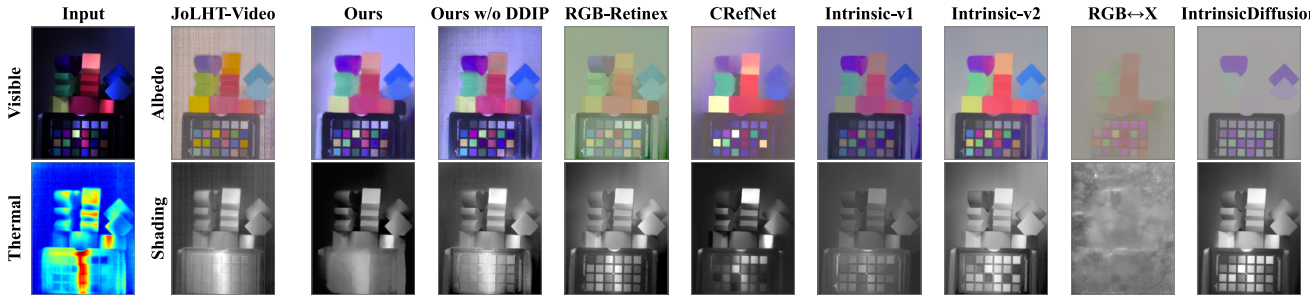

Figure 4. Results on a color-chart scene in JoLHT-Video dataset. Our method recovers the smooth line-light shading across the color chart.

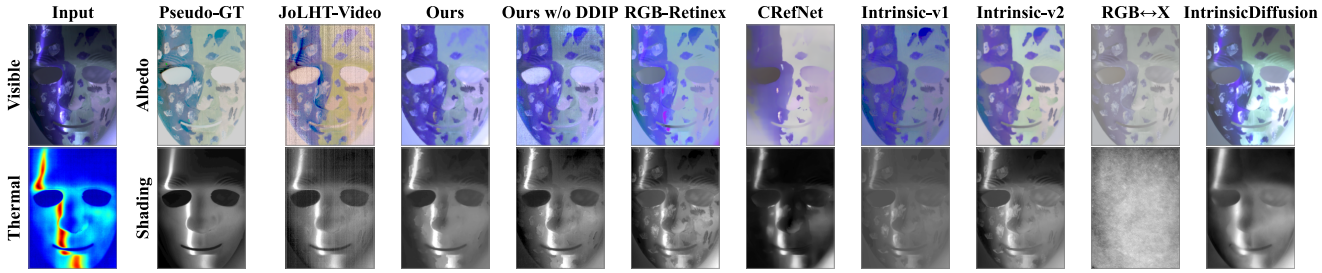

Figure 5. Results on *Painted-Mask* scene in JoLHT-Video dataset. Baselines show albedo texture in shading or highlight artifacts in albedo.

Metrics: We use the scale-invariant Mean Square Error (si-MSE) to evaluate albedo and shading quantitatively.

Baselines: We compare with state-of-the-art methods in three categories. *Learning-based:* Diffusion-based IntrinsicDiffusion [30] and RGB↔X [43], CNN-based Intrinsic-v1 [6] and Intrinsic-v2 [7], and Transformer-based CRefNet [31]. *Physics-based:* NIR-Priors [8], requiring a paired NIR image, and JoLHT-Video [34], demanding transient thermal video under controlled illumination. *Optimization-based:* RGB-Retinex [20] and Opt-LocalSmooth [37]. IntrinsicDiffusion, RGB↔X, and Intrinsic-v2 output colorful shading, while others grayscale.

6.1. Qualitative Evaluation

Fig. 6 presents comparisons with state-of-the-art baselines across various scenes. The first two cases demonstrate our ability to remove cast shadows from albedo (e.g., handrail and lantern shadows), while the next two highlight disentangling albedo texture from shading (e.g., rhino statue texture and checkerboard pattern). The final example is an homage to the classic Adelson's Checker-Shadow Illusion [1], where our method successfully separates the shadowed checker region from the cylinder shading.

Learning-based baselines often over-smooth albedo and shading due to strong statistical priors, producing flat grass shading or overly uniform wall colors. In contrast, our physics-based approach, guided solely by a single thermal image, better preserves details such as block-wise albedo variation, concrete texture, and natural shading gradients.

Fig. 4 and Fig. 5 show results on the JoLHT-Video dataset. Our method recovers smooth line-light shading comparable to JoLHT-Video using only a single thermal image, while other baselines exhibit clear albedo or shading leakage.

6.2. Quantitative Evaluation

6.2.1. Validating Ordinalities

We validated our ordinality theory on a wide range of real-world materials and scenes via expert annotations, and confirm statistically with a spectral reflectance dataset [22] that invisible component rarely overturn these ordinalities.

Patch Ordinality on Various Materials. The evaluation included 20 patches from CUReT dataset [10] and common objects (painted aluminum, plastic, wood, silk, leather, cloth, plaster, etc.). These patches were placed in different orientations under artificial and natural lighting. Experts confidently labeled 865 ordinalities across patches. Our prediction matched the expert labels with 98.59% accuracy in sunlight (albedo: 99.37%, shading: 97.01%) and 96.82% under white-LED (albedo: 94.62%, shading: 100%).

Point-Pair Ordinality on Diverse Scenes. We further evaluated on 100 real-world scenes in VT-Intrinsic dataset (Sec. 5), spanning materials such as stone, concrete, grass, vegetation, painted metal, plastic, and wood. Experts labeled the ordinalities in albedo or shading of 20 randomly sampled point pairs per image using the visible image as reference. Pairs with small intensity differences were excluded to avoid ambiguity. Experts confidently labeled 1,063 pairs and found 937 unclear. Ignoring the latter, our theory achieved 98.95% overall accuracy (albedo: 96.96%, shading: 99.62%), confirming the reliability of thermal-guided ordinal cues. More details are in the supplementary.

Statistical Robustness to Invisible Spectral Components. Violations of ordinalities due to invisible reflectance *are possible but statistically unlikely*. Above we validated ordinalities with expert labels for many materials and lighting conditions with non-visible components. As a further

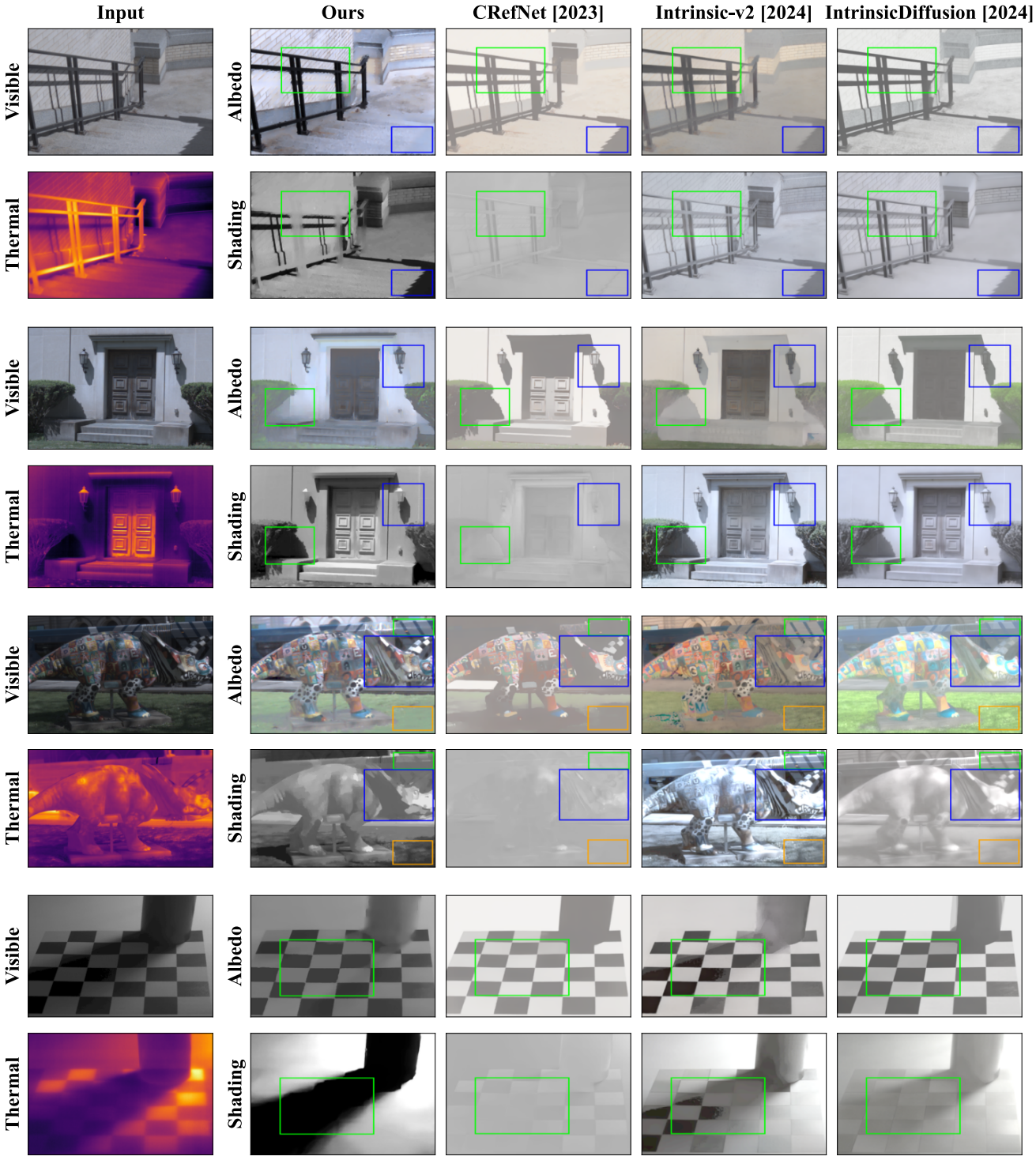

Figure 6. **Qualitative comparisons to state-of-the-art baselines.** The first two scenes show how our method removes cast shadows from albedo (e.g., shadow of handrail in case 1, lanterns in case 2). The next three demonstrate our ability to eliminate albedo texture from shading (e.g., rhino statue texture in case 3, checkerboard pattern in case 4). But baselines struggle with these challenges, despite their advantage of pre-training on large datasets, whereas our approach relies solely on physics-based information in a single thermal image. Baselines often over-smooth albedo and shading (e.g., smooth albedo on detailed ground and walls, flat shading on grass) due to reliance on priors. Diffusion-based baselines can offer appealing visual quality but sacrifice faithfulness (e.g. hallucinated albedo texture on the rhino statue in case 3). Images are tonemapped for visualization. Key differences are highlighted in bounding boxes. *More examples and baselines appear in the supplementary.*

Table 1. Results of si-MSE (↓) reported at 10^{-2} across datasets (Sec. 6.2). **Best** and second highlighted. Our method surpasses all learning-based approaches despite using no learned priors and achieves performance comparable to JoLHT-Video, which demands transient thermal video under controlled illumination. N/A indicates unavailable data, and ✗ denotes non-applicability.

Method ● Optimization-based ◆ Learning-based ★ Physics-based (w/ auxiliary sensor)	**JoLHT-Video Dataset**			**Color Chart w/ Different Illumination**		
	Painted Mask Albedo	Painted Mask Shading	Color Charts Albedo	White LED Albedo	Incandescent Albedo	Sunlight Albedo
● RGB-Retinex [20] (TPAMI'06)	25	0.30	3.4	2.42	2.33	2.73
● Opt-LocalSmooth [37] (CVPR'11)	45	0.35	7.1	2.41	4.21	2.04
◆ IntrinsicDiffusion [30] (SIGGRAPH'24)	37	0.25	2.9	4.12	3.33	4.85
◆ RGB↔X [43] (SIGGRAPH'24)	30	0.37	2.8	4.07	5.31	4.59
◆ Intrinsic-v2 [7] (ToG'24)	27	0.17	2.8	1.25	4.36	4.17
◆ Intrinsic-v1 [6] (ToG'23)	30	0.21	3.8	1.55	2.72	4.97
◆ CRefNet [31] (TVCG'23)	38	0.23	8.8	1.79	2.29	1.98
★ NIR-Priors [8] (ICCV'19)	N/A	N/A	N/A	✗	2.46	2.08
★ JoLHT-Video [34] (CVPR'24)	**8.4**	**0.05**	**2.0**	N/A	✗	✗
★ Ours	11	0.10	2.7	**0.37**	**1.06**	**1.19**

evaluation, we used the Solar Spectral Irradiance (ASTM-G173) and 427 materials from USGS Spectral Reflectance dataset [22], and confirmed that the ordinality of absorbed visible light matches the ordinality of total absorbed light (including UV and IR) in 94.2% cases out of all 90,951 or $\binom{427}{2}$ material pairs. The remaining 5.8% of violations occur mostly when absorptances are nearly identical.

6.2.2. Validating IID Method

We present evaluations on the color charts under different illuminations and JoLHT-Video dataset [34], and include in the supplementary (i) a simulation experiment to validate ordinality informativeness and (ii) an ablation study.

Color Chart under Different Illuminations. We imaged a color chart under white LED, incandescent and sunlight. Tab. 1 shows our method outperforming baselines under all illuminations. Incandescent and sunlight experiments demonstrate our robustness to albedo variations even in the invisible band that influence absorbed light. In contrast, physics-based baselines have limited applicability: JoLHT-Video assumes no invisible lighting component, and NIR-Priors requires NIR emission absent in white LEDs.

Using JoLHT-Video Dataset. The dataset [34] includes four color-chart scenes and a *Painted-Mask* scene with pseudo ground-truth obtained following [19], which are considerably challenging due to the strong lighting variations from line light (Fig. 4, Fig. 5). As shown in Tab. 1, our method outperforms all learning-based baselines without pre-trained priors, and achieves performance comparable to JoLHT-Video [34], which demands stricter conditions of calibrated transient thermal video and controlled lighting.

7. Limitations and Conclusion

This work explores photometric cues encoded in a single auxiliary thermal image, and presents physics-based optimization for albedo-shading separation. We showed its effectiveness on real scenes with a wide range of materials and lighting conditions. However, diffuse reflection dominates in these materials — metals, transparent objects and mirrors violate the visible image formation model. Our model also assumes that the heat arises primarily from light absorption — heat generated otherwise internally (engines, humans) or externally (hot air blower or fire) is not modeled. It also does not handle multiple colored illuminations. Finally, we rely on inexpensive microbolometer thermal cameras whose quality is lower compared to visible cameras — low SNR due to insufficient heat generation (overcast skies, dynamic objects) can degrade performance. Failure cases are shown in Fig. 7, with additional analysis on low-light and non-equilibrium thermal conditions in the supplementary. Despite these limitations, the improvements demonstrate the potential to scale supervision for learning algorithms. We hope our work inspires further exploration of light–heat interaction in computer vision and graphics.

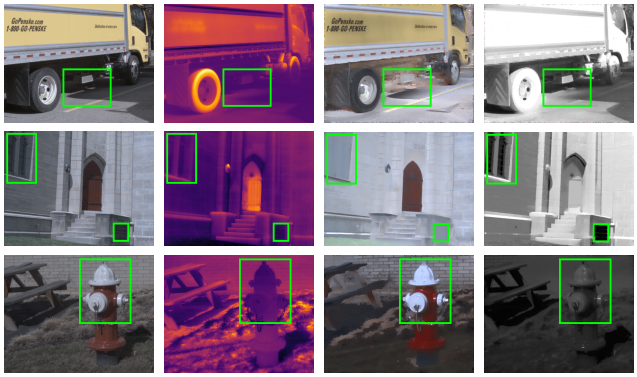

Figure 7. Corner cases: 1) Heat from a truck engine, unrelated to light absorption, elevates the thermal intensity of the road beneath it. 2) Non-opaque cathedral windows violate the visible image formation model.. 3) The metallic fire hydrant has low emissivity, resulting in poor thermal SNR, and exhibits specular highlights that challenge the common Lambertian assumption in IID.

Acknowledgements

This work was partly supported by NSF grants IIS210723, and NSF-NIFA AI Institute for Resilient Agriculture. We are sincerely grateful to Akihiko Oharazawa for his help with expert annotation, and to Sriram Narayanan and Gaurav Parmar for their insightful discussions.

References

[1] Edward Adelson. The checker shadow illusion. In *persci.mit.edu/gallery/checkershadow*, 1995. 6

[2] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N. Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023. 2

[3] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 1

[4] Harry G Barrow and Jay M Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978. 2

[5] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014. 2, 5

[6] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Trans. Graph.*, 43(1), 2023. 2, 6, 8

[7] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Trans. Graph.*, 43(6), 2024. 2, 6, 8

[8] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2521–2530, 2019. 2, 3, 6, 8

[9] Gyeongmin Choe, Srinivasa G. Narasimhan, and In So Kweon. Simultaneous estimation of near ir brdf and fine-scale surface geometry. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4

[10] Kristin Dana, Bram Ginneken, Shree Nayar, and Jan Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, 18:1–34, 1999. 6, 2

[11] Partha Das, Sezer Karaoglu, and Theo Gevers. Intrinsic image decomposition using physics-based cues and cnns. *Computer Vision and Image Understanding*, 223:103538, 2022. 2

[12] Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Aggelos Katsaggelos, Ashok Veeraraghavan, and Oliver Cossairt. Thermal spread functions (tsf): Physics-guided material classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1641–1650, 2023. 2

[13] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. *SIGGRAPH 97*, 1997. 5

[14] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. Color constancy at a pixel. *Journal of the Optical Society of America A*, 21(8):1453–1458, 2004. 2

[15] Sigmund Fritz. Illuminance and luminance under overcast skies. *J. Opt. Soc. Am.*, 45(10):820–825, 1955. 2

[16] Yossi Gandelsman, Assaf Shocher, and M. Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. *Computer Vision and Pattern Recognition*, 2018. 2, 5

[17] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. A survey on intrinsic images: Delving deep into lambert and beyond. *International Journal of Computer Vision*, 130(3):836–868, 2022. 1

[18] Gevers. Reflectance-based classification of color edges. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 856–861. IEEE, 2003. 2, 3

[19] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342, 2009. 1, 2, 5, 8

[20] Steven D. Hordley, Mark S. Drew, Graham D. Finlayson, and Cheng Lu. On the Removal of Shadows from Images . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28 (01):59–68, 2006. 6, 8

[21] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. *arXiv preprint arXiv: 2312.12274*, 2023. 2

[22] Raymond F. Kokaly, Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Todd M. Hoefen, Neil C. Pearson, Richard A. Wise, William M. Benzel, Heather A. Lowers, Ryan L. Driscoll, and Andrew J. Klein. USGS Spectral Library Version 7 Data. https://doi.org/10.5066/F7RR1WDJ, 2017. U.S. Geological Survey Data Release. 2, 6, 8

[23] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[24] Philipp Krahenbuhl. Free supervision from video games. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2964, 2018. 2

[25] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977. 2, 3

[26] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[27] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018. 2

[28] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7199, 2021. 2

[29] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single im-

age. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3248–3257, 2020. 2

[30] Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Wang. Intrinsicdiffusion: Joint intrinsic layers from latent diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6, 8, 1

[31] Jundan Luo, Nanxuan Zhao, Wenbin Li, and Christian Richardt. Crefnet: Learning consistent reflectance estimation with a decoder-sharing transformer. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6407–6420, 2024. 6, 8

[32] Yasuto Nagase, Takahiro Kushida, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Shape from thermal radiation: Passive ranging using multi-spectral lwir measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12661–12671, 2022. 2

[33] Sriram Narayanan, Mani Ramanagopal, Mark Sheinin, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. Shape from heat conduction. In *European Conference on Computer Vision*, pages 426–444. Springer, 2024. 2

[34] Mani Ramanagopal, Sriram Narayanan, Aswin C. Sankaranarayanan, and Srinivasa G. Narasimhan. A theory of joint light and heat transport for lambertian scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11924–11933, 2024. 2, 3, 5, 6, 8

[35] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10912–10922, 2021. 2

[36] Shogo Sato, Yasuhiro Yao, Taiga Yoshida, Takuhiro Kaneko, Shingo Ando, and Jun Shimamura. Unsupervised intrinsic image decomposition with lidar intensity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13466–13475, 2023. 3

[37] Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. Intrinsic images using optimization. In *CVPR 2011*, pages 3481–3487, 2011. 2, 6, 8

[38] Roger Shepard. Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art. 1990. 3

[39] 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 Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2075–2085, 2021. 2

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128:1867–1888, 2020. 5

[41] Michael Vollmer and Klaus-Peter Mollmann. *Fundamentals of Infrared Thermal Imaging*, chapter 1, pages 1–106. John Wiley & Sons, Ltd, 2017. 4

[42] Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. *International Conference on Computational Photography*, 2023. 1, 2

[43] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6, 8

[44] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 388–396, 2015. 2