Heat Assisted Detection and Ranging

Fanglin Bao,[1] Xueji Wang,[1] Shree Hari Sureshbabu,[1] Gautam Sreekumar,[2]
Liping Yang,[1] Vaneet Aggarwal,[3] Vishnu N. Boddeti,[2] and Zubin Jacob[1,*]

[1]*Birck Nanotechnology Center, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA*
[2]*Michigan State University, East Lansing, MI 48824, USA*
[3]*School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA*

(Dated: March 14, 2023)

Machine perception uses advanced sensors to collect information of the surrounding scene for situational awareness [1–7]. State-of-the-art machine perception [8] utilizing active sonar, radar and LiDAR to enhance camera vision [9] faces difficulties when the number of intelligent agents scales up [10, 11]. Exploiting omnipresent heat signal could be a new frontier for scalable perception. However, objects and their environment constantly emit and scatter thermal radiation leading to textureless images famously known as the 'ghosting effect' [12]. Thermal vision thus has no specificity limited by information loss while thermal ranging, crucial for navigation, has been elusive even when combined with artificial intelligence (AI) [13]. Here we propose and experimentally demonstrate heat-assisted detection and ranging (HADAR) overcoming this open challenge of ghosting and benchmark it against AI-enhanced thermal sensing. HADAR not only sees texture and depth through the darkness as if it were day, but also perceives decluttered physical attributes beyond RGB or thermal vision, paving the way to fully-passive and physics-aware machine perception. We develop HADAR estimation theory and address its photonic shot-noise limits depicting information-theoretical bounds to HADAR-based AI performance. HADAR ranging at night beats thermal ranging and shows an accuracy comparable with RGB stereovision in daylight. Our automated HADAR thermography reaches the Cramér-Rao bound on temperature accuracy, beating existing thermography techniques. Our work leads to a disruptive technology that can accelerate the Fourth Industrial Revolution (Industry 4.0) [14] with HADAR-based autonomous navigation and human-robot social interactions.

The emerging Industry 4.0 of smart technologies [15] calls for a future with scalable human-robot social interactions since it is expected that one in ten vehicles will be automated by 2030 [16] and 20 million robot helpers will be serving people [17]. Each of these agents will collect information about its surrounding scene through advanced sensors to make decisions without human intervention. However, simultaneous perception of the scene by numerous agents (scalable perception) is fundamentally prohibitive for active modalities due to signal interference and eye safety [10, 11]. *Quasi*-passive approaches like cameras are an alternative but they rely on ambient illumination. Furthermore, cameras cannot compete with human perception even though important strides [18] have been made recently based on deep learning [19, 20]. It causes phenomena like phantom braking [9] in automated vehicles due to the visual ambiguity and lack of physical context in perception. A paradigm shift of *fully*-passive perception beyond traditional vision is urgently needed that can boost the AI industry (Fig. 1a-b).

An attractive approach to scalable perception is using the fully passive heat signal originating from infrared thermal radiation. Exploiting heat signals for imaging [21–24] has well-known advantages, *e.g.*, to see through the darkness or solar glare as well as bad weather [25], and not surprisingly, it has been the natural choice of predators (snake) when hunting prey (rat) at night [26]. Nevertheless, fundamental obstacles exist for heat-assisted perception. Physical attributes of the scene, namely, temperature (T, physical status), emissivity (e, material fingerprint) and texture (X, surface geometry) are mixed in photon streams, as objects and environment constantly emit and scatter thermal radiation. This is manifested as the ghosting effect [12] related to lack of texture in thermal imaging. Ghosting limits thermal imaging only to night vision enhancement without any specificity even when combined with AI algorithms (see Tab. S3 in the Supple. Info. for a review).

TeX decomposition and TeX vision We address the ghosting effect with an approach we call TeX decomposition, which vividly recovers the texture from the cluttered heat signal and also accurately disentangles temperature and emissivity at the Cramér-Rao bound. Representing these decluttered TeX attributes in HSV color space ($\text{Hue} = e, \text{Saturation} = T, \text{Brightness} = X$) leads to a paradigm shift of TeX vision with physical context for machine perception (Fig. 1b-c). TeX vision empowers AI algorithms to reach information-theoretic bounds, which has thus far been elusive for traditional RGB or thermal vision. Fig. 1c shows TeX vision for on- and off-road scenes at night overcoming the ghosting ef-

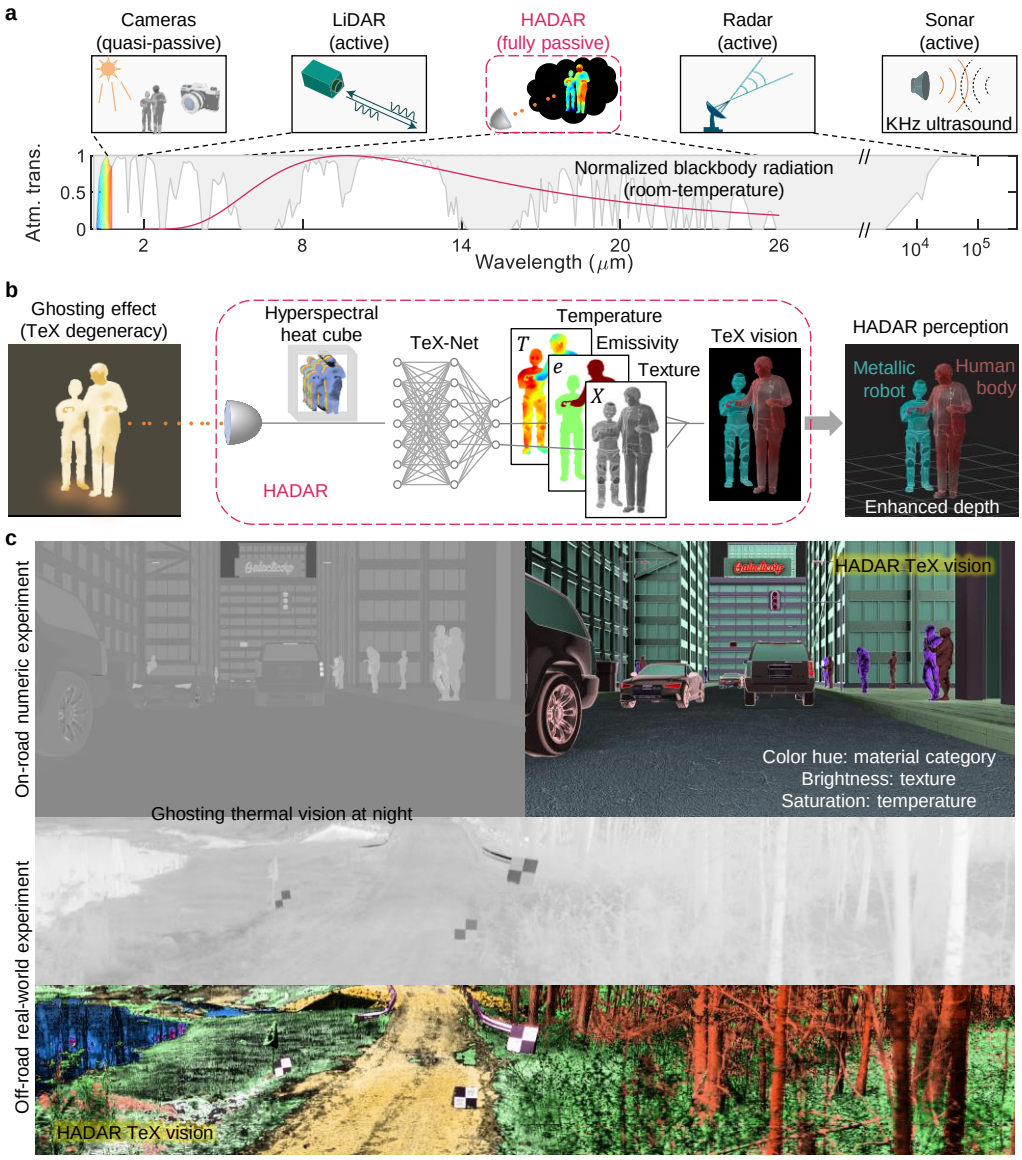


FIG. 1. HADAR as a paradigm shift in machine perception. a, Fully passive HADAR makes use of heat signals, as opposed to active sonar, radar, LiDAR, and quasi-passive cameras. Atmospherical transmittance window (white area) and temperature of the scene determine the working wavelength of HADAR. b, HADAR takes thermal photon streams as input, records hyperspectral-imaging heat cubes, addresses the ghosting effect through TeX decomposition (see Extended Data Fig. 1 for TeX-Net and see Methods for all decomposition methods) and generates TeX vision for improved detection and ranging. c, TeX vision demonstrated on our HADAR database and outdoor experiments (see Extended Data Figs. 2∼4) clearly shows that HADAR sees textures through the darkness with comprehensive understanding of the scene.

fect (also see Supplementary movies for video demonstrations). Our demonstrations of HADAR include detection and ranging based on TeX vision, for both real-world level HADAR database and outdoor experiments. We provide detailed comparisons with state-of-the-art AI-enhanced thermal sensing and prove that HADAR provides universal performance enhancement. This can lead to adoption of TeX vision as an industry standard.

For intuitive clarity, we first explain the origin of the ghosting effect using an example of thermal radiation (visible) from a light bulb. Fig. 2 shows Monte Carlo path tracing simulations of rays emanating from a bulb, with reflection of environmental emission taken into account. Geometric textures on the bulb surface can be seen only when the bulb is off. We emphasize that this texture revealed by reflection is completely lost in direct emission when the bulb is switched on, a familiar scenario from daily experience. Since every object in a complex scene emits and scatters thermal radiation, they are thermal light sources with no texture like a shining bulb. The total heat signal leaving an object α has two

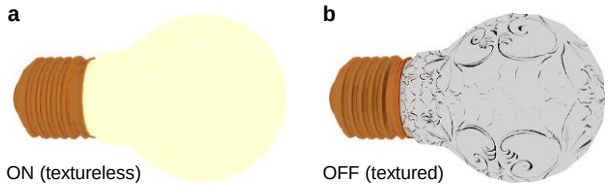


FIG. 2. Monte Carlo path tracing simulation of a light bulb to explain the 'ghosting effect'. Geometric texture on a light bulb can only be seen when the bulb is off whereas this texture is completely missing when it is glowing. The blackbody radiation can never be turned off leading to loss of texture for thermal images. This ghosting effect presents the long-standing roadblock for heat-assisted machine perception.

additive contributions,

$$S_{\alpha\nu} = e_{\alpha\nu} B_\nu(T_\alpha) + [1 - e_{\alpha\nu}] X_{\alpha\nu}, \tag{1}$$

where the first term is direct thermal emission (textureless), and the second term carrying texture is the environmental emission entering the detector after scattering from the object. Here ν in the subscript denotes wavenumber (spectrum) dependence. The key difference with a shining bulb is that blackbody radiation B_ν is fundamentally governed by Planck's law and cannot be switched off. Textureless thermal imaging is thus widely regarded as impossible to use for quantitative insight about a scene. The environmental thermal illumination on object α from all other objects β is given by $X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta} S_{\beta\nu}$, with $V_{\alpha\beta}$ being the thermal lighting factor. Ghosting effect is exacerbated for high emissivity materials in nature such as skin and plants ($e \approx 1$) as the total collected signal consists of dominant direct emission and only a weak scattered signal. We note that $S_{\alpha\nu}$ is invariant under joint transformations of temperature T, emissivity e and texture X (see Methods), which we address as TeX degeneracy. In addition to the ghosting effect, this TeX degeneracy renders the separation of temperature-emissivity as a major roadblock [27] to quantitative thermal sensing.

We recover the texture by breaking TeX degeneracy and discretizing spectral emissivity $e_{\alpha\nu}$ into $e_\nu(m_\alpha)$ in a material library, $\mathcal{M} = \{e_\nu(m) | m = 1, 2, \cdots, M\}$, that contains all possible spectral emissivity in the scene. This opportunity of dimensional reduction is available naturally in smart applications where materials usually have industrial standards [28]. The material library explains the physics but requires on-site collection/calibration. We have also provided a generalized HADAR theory that does not require an input of material library (see Sec. SVC of the Supple. Info.). Our approach of TeX-Net uses Eq. (1) to design physics-based loss and uses a 3D convolutional neural network to learn spatio-spectral features, in recovering texture X, temperature T and emissivity e. With general HADAR performance shown in Extended Data, here we demonstrate the fundamental limits as well as real-world performance of HADAR.

HADAR identifiability We develop HADAR estimation theory to address fundamental limits of object identification from its thermal infrared signature. We believe this will be crucial in guiding public policy for the industrial revolution where decision accuracy of machine perception can be bounded by physical laws as opposed to training data volume. HADAR is distinct from hyperspectral imaging where material difference is determined by the Euclidean distance between their reflectance spectra. In stark contrast, HADAR identifiability is determined by multi-parameter estimation of temperature, emissivity and texture (see Fig. S6 and relevant contexts in Supple. Info.). We exploit the multi-parameter Cramér-Rao bound and propose semantic distance to categorize objects based on their intrinsic material properties. Fig. 3 shows a pertinent example of human vs. robot identification. A human-shaped target (Fig. 3a) could be a human (organic skin or fabrics material) or robot (metallic) with distinct emissivity (top inset), but they will produce a visually indistinguishable incident spectrum on the detector (bottom inset; modelled by FLIR A325sc). We define HADAR identifiability as the maximum Shannon information of the target material that one can retrieve from N incident photons. It holds for all scenes (see Extended Data Fig. 6 for generalization to multi-material scenes) and is given by

$$I = \log_2 \left[1 + \mathrm{erf}[\sqrt{\frac{N d_0^2}{2(1+\gamma)}}] \right], \tag{2}$$

where $\gamma \equiv \gamma_1 N + \gamma_0$ is the detector's electronic-noise power normalized by the photonic shot-noise power. We introduce d_0 as the semantic distance between two materials with known spectral emissivity defined using single-photon Fisher information matrix (see Methods).

The insight from Eq. (2) is that the shot-noise limit arising from the discrete nature of photons sets the information-theoretical upper bound to the performance of all identification algorithms. Here we reach the bound with machine-learning-based approaches widely deployed for perception. We generate multiple spectra for human and robot with Monte Carlo simulation in the shot-noise limit, and use machine learning for material classification. Fig. 3b shows machine learning performance (red circles) is indeed bounded by the theoretical limit (red curve). Our theory also applies to realistic detectors with common noise sources (Flicker noise: cyan dashed; Johnson-Nyquist noise: cyan dash-dotted; mixed noise: cyan solid; modelled by FLIR A325sc) and corresponding algorithmic performance (see Fig. S7 of Supple. Info.). Fig. 3c shows the minimum photon number required to identify the human-shaped target, which is determined by unit statistical distance ($\sqrt{N} d_0 = 1$, $I \approx 0.75$ bits). The minimum photon number for given semantic distance or vice versa, the minimum semantic distance for

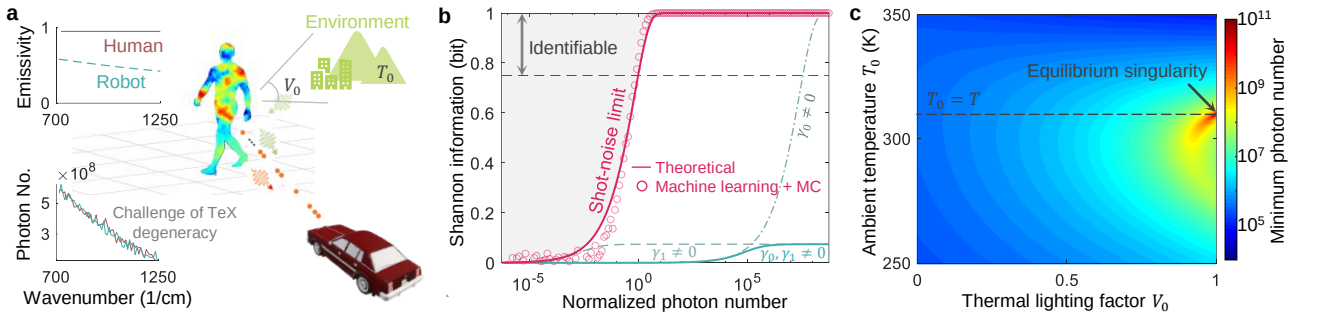


FIG. 3. Shot-noise limit of HADAR identifiability. TeX degeneracy limits HADAR identifiability, as in the illustrative human-robot identification problem (a). Top inset: distinct emissivity of human (graybody) and robot (aluminum). Bottom inset: Near-identical incident spectra for human (37 C°, red) and robot (72.5 C°, blue). b, HADAR identifiability (Shannon information) as a function of normalized photon number Nd_0^2. We compare theoretical shot-noise limit of HADAR (red solid) and Machine learning performance (red circles) on synthetic spectra generated by Monte Carlo simulations. We also consider realistic detectors with Johnson-Nyquist noise ($\gamma_0 = 3.34e5$), Flicker noise ($\gamma_1 N = 3.34e5$), or mixed noise ($\gamma_1 N = \gamma_0 = 3.34e5$). Identifiability criterion (gray dashed) is $\sqrt{N}d_0 = 1$. Minimum photon number $1/d_0^2$ required to identify a target is usually large due to the TeX degeneracy, dependent of the scene as well as the thermal lighting factor, as shown in c for scene a. Particularly, it diverges at singularity $V_0 = 1$ and $T = T_0$ when the target is in thermal equilibrium with environment.

given photon number, sets fundamental limits to object identification beyond training volume, providing a theoretical foundation for designing public policies.

HADAR depth resolution Depth of objects is a critical scene attribute for autonomous navigation. Daylight RGB stereovision already has widespread applications [8], but infrared thermal ranging is elusive. We demonstrate that HADAR ranging at night beats thermal ranging, with depth accuracy comparable to RGB stereovision in daylight. Our approach of HADAR ranging exploits stereovision based on the TeX vision, but to show the importance of texture in ranging and better capture the physics, here we focus on the scattered signal that can be reconstructed through TeX decomposition. Real-world HADAR ranging will be discussed later. For a concise car/pedestrian scene, thermal imaging loses textures due to TeX degeneracy (Fig. 4a) and leads to inaccurate ranging (Fig. 4d). HADAR (Fig. 4b) recovers texture comparable to grayscale optical imaging (Fig. 4c). We note that the HADAR ranging result (Fig. 4e) is comparable to RGB stereovision (Fig. 4f). Quantitatively, the absolute ranging error (cyan data points in insets) with respect to the ground truth along white dashed lines shows $\sim 100\times$ accuracy improvement in HADAR vs. thermal ranging (the improvement is scene dependent).

We derive the fundamental limit on HADAR ranging providing a rigorous theoretical foundation for future autonomous navigation applications. HADAR ranging error δz of a window (block or feature area) equals the disparity error between corresponding window positions in stereo matching [29], up to a dimensionless coefficient. Its fundamental limit is given by

$$\sqrt{N}\delta z \geq \sqrt{2(1+\gamma)(\sigma_{\rm d}^2 + \sigma_{\rm c}^2)}, \qquad (3)$$

where $\sigma_{\rm d}$ is the diffraction-induced uncertainty in estimating a point source position from the incident photon distribution. Here, $\sigma_{\rm c}$ is the photonic correspondence uncertainty in locating the same point source between stereo images in an extended scene with N observed photons. The physical significance of the photonic correspondence uncertainty is the indistinguishability of photons of the same frequency from different point sources. It is given by the Cramér-Rao bound of window-position estimation in the ideal image of the scene (see Sec. SIIC of the Supple. Info.). Theoretical bounds of δz with computed $\sigma_{\rm c}$ along white dashed lines (red curves in insets of Fig. 4d-e) are consistent with numeric experiments (cyan data points), showing a 2-orders-of-magnitude accuracy improvement in HADAR.

Real-world HADAR perception We now experimentally demonstrate HADAR in real-world scenes. Our HADAR prototype-1 for low-end applications is based on commercial FLIR thermal camera with custom designed spectral modules (see Extended Data Fig. 10). We propose a paradigm shift of physics-driven perception with HADAR TeX attributes, as opposed to existing vision-driven perception [18]. We use an outdoor scene at night with a car, human being and Einstein cut-out to mimic a human geometrically, and illustrate how HADAR addresses phantom braking. Fig. 5 shows that both RGB optical imaging (Fig. 5a) and sparse LiDAR point cloud (Fig. 5c; Velodyne Puck VLP-16) cannot distinguish the human body with the real-scale Einstein cardboard. Furthermore, LiDAR has difficulties in detecting the black car due to low reflection, whereas the optical camera cannot see objects in the dark. HADAR detects people in the corresponding material region (skin+fabrics) and clearly distinguishes it from the cardboard, overcoming the phantom braking problem. See Extended Data

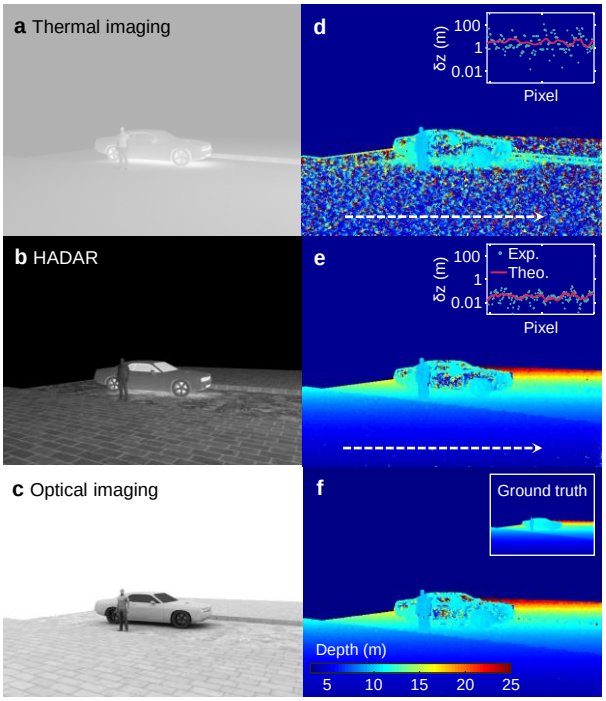


FIG. 4. Fundamental limit of HADAR ranging. a and d, Ranging based on raw thermal images shows poor accuracy due to ghosting. b and e show recovered textures and enhanced ranging accuracy ($\sim 100\times$) in HADAR, as compared with thermal ranging. We also show the optical imaging (c) and RGB stereovision (f) for comparison. Insets in d and e show the depth error δz in Monte Carlo experiments (cyan points) in comparison with our theoretical bound (red curve), along white dashed lines.

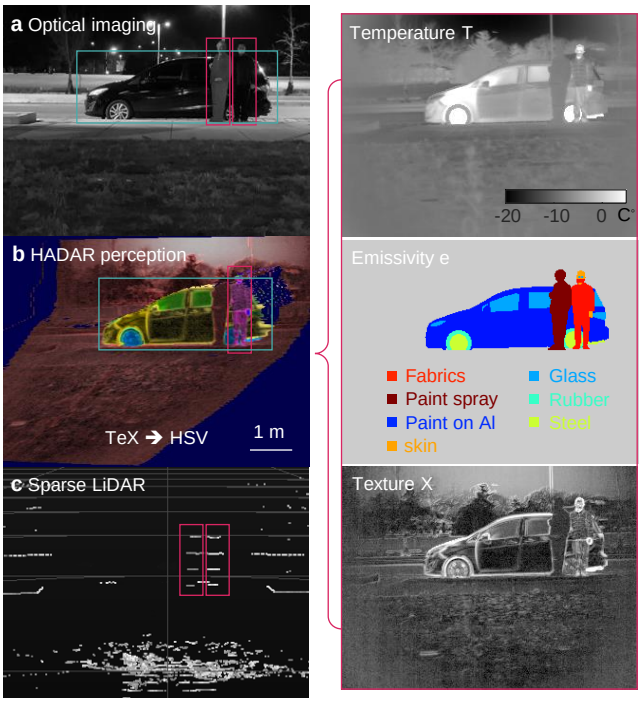


FIG. 5. Physics-driven HADAR perception in Indiana, USA. For an outdoor scene of a human body, an Einstein cardboard, and a black car at night, vision-driven object detection yields two human bodies (error) and one car from optical imaging (a), and two human bodies and no car (error) from LiDAR pointcloud (c). HADAR perception based on TeX physical attributes has comprehensive understanding of the scene and accurate semantics (b; one human body and one car) for unmanned decisions.

Figs. 7 and 8 for more details about HADAR detection and semantics. Major advantages of HADAR perception utilizing physical context will be found in autonomous navigation and wildlife monitoring, where multiple physical attributes beyond visual appearance are desired either for safety guarantees [30] or scientific purposes [31].

Our HADAR prototype-2 for high-end applications is based on a pushbroom hyperspectral imager (see Methods). We use an off-road scene to demonstrate that TeX vision sees textures through the darkness with physical context, and that HADAR ranging at night beats thermal ranging, with accuracy comparable to RGB stereovision in daylight. Real-world TeX vision with material identification and texture recovery has been shown in Fig. 1c and can be found in Extended Data Figs. 3 and 4 with more details. Fig. 6 shows the stereovision metric statistics based on TeX vision at night, thermal vision at night, and RGB vision in daylight. The comparison of metrics (normalized by RGB depth metrics) in Fig. 6b clearly demonstrates that HADAR ranging at night beats thermal ranging and matches RGB stereovision in daylight, abbreviated as 'TeX_night $\sim$ RGB_day $>$ IR_night'. See Fig. S19 of the Supple. Info. for general HADAR ranging performance over various scenes.

HADAR thermography The COVID-19 pandemic has brought about the urgent need of remote thermography for fever surveillance. Unmanned and high-speed infrared surveillance can significantly relieve the risk to healthcare workers and help limit spread of the virus. However, large scale temperature screening with existing noncontact infrared thermometer or infrared thermography is ineffective due to lack of adaptivity to emissivity (complexion/skin variability), age, gender, circadian variations and distance of the target [33, 34]. As illustrated above, HADAR with TeX vision can identify spectral emissivity, estimate distance, and recover textures, promising in advanced adaptivity for more accurate temperature estimation. Here, we have also experimentally demonstrated that HADAR thermography can automatically recognize emissivity and reach the Cramér-Rao bound on temperature accuracy (see Extended Data Fig. 9). This goal has been elusive due to TeX degeneracy which limits temperature accuracy in real-world environments. Unmanned HADAR thermography reaching the Cramér-Rao bound is therefore promising for the smart healthcare industry including early reliable skin cancer

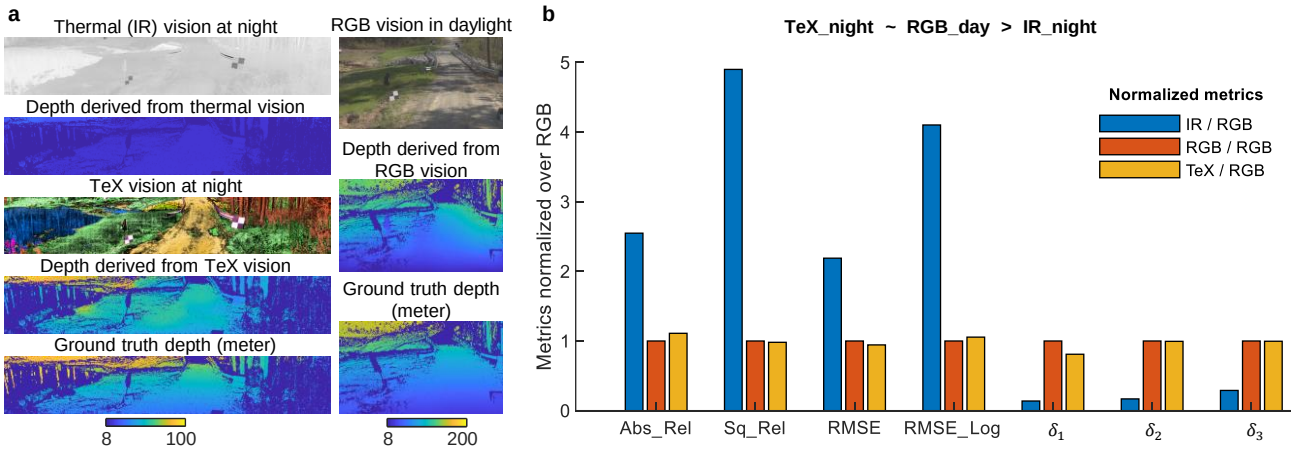


FIG. 6. HADAR ranging (TeX vision + AI) at night beats state-of-the-art thermal ranging (thermal vision + AI) at night and matches RGB stereovision in daylight, abbreviated as 'TeX_night ∼ RGB_day > IR_night'. a, It can be clearly seen that thermal imaging is impeded by the ghosting effect, while HADAR TeX vision overcomes the ghosting effect providing a fundamental route to extracting thermal textures. This texture is crucial for AI algorithms to function optimally. To prove the HADAR ranging advantage, we used GCNDepth (pre-trained on the KITTI dataset) [32] for monocular stereovision, as the state-of-the-art AI algorithm. Ground truth depth is obtained through a high-resolution LiDAR. Depth metrics are listed in Tab. I. We normalized the depth metrics over that of RGB stereovision. The comparison of normalized metrics (b) clearly demonstrates 'TeX_night ∼ RGB_day > IR_night', *i.e.*, HADAR sees texture and depth through the darkness as if it were day. See Methods for the definitions of used depth metrics. See Secs. SIIIA and SV of the Supple. Info. for more details.

Real-world performance (depth)	Error				Accuracy (%)		
	Abs_Rel	Sq_Rel	RMSE	RMSE_Log	δ_1	δ_2	δ_3
Thermal vision + AI	0.61	13.44	22.96	1.24	7.72	14.82	28.18
RGB vision + AI	0.24	2.74	10.49	0.30	55.88	87.96	97.12
TeX vision + AI	0.27	2.69	9.90	0.32	45.25	87.52	96.77

TABLE I. Depth metrics statistics associated with Fig. 6 revealing HADAR ranging advantage.

detection [35].

Outlook We proposed and demonstrated HADAR for fully-passive and physics-aware machine perception. Our shot-noise limits of detection and ranging set the benchmark and call for heat exploitation in the quantum regime where single photon detectors are being developed beyond visible spectral range into the thermal infrared [36]. Practical challenges exist, such as, real-time data acquisition, spatio-spectral motion blur, and functionality-cost optimization. Nevertheless, we believe HADAR will lead to a new chapter in the Fourth Industrial Revolution with applications in autonomous navigation, healthcare, agriculture, wildlife monitoring, geosciences and defense industry.

* zjacob@purdue.edu

[1] C. Rogers, A. Y. Piggott, D. J. Thomson, R. F. Wiser, I. E. Opris, S. A. Fortune, A. J. Compston, A. Gondarenko, F. Meng, X. Chen, G. T. Reed, and R. Nicolaescu, A universal 3d imaging sensor on a silicon photonics platform, Nature **590**, 256 (2021).

[2] D. Floreano and R. J. Wood, Science, technology and the future of small autonomous drones, Nature **521**, 460 (2015).

[3] Y. Jiang, S. Karpf, and B. Jalali, Time-stretch lidar as a spectrally scanned time-of-flight ranging camera, Nat. Photonics **14**, 14 (2020).

[4] L. Maccone and C. Ren, Quantum radar, Phys. Rev. Lett. **124**, 200503 (2020).

[5] J. Tachella, Y. Altmann, N. Mellado, A. McCarthy, R. Tobin, G. S. Buller, J.-Y. Tourneret, and S. McLaughlin, Real-time 3d reconstruction from single-photon lidar data using plug-and-play point cloud denoisers, Nat. Commun. **10**, 4984 (2019).

[6] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, Soli: Ubiquitous gesture sensing with millimeter wave radar, ACM Trans. Graph. **35**, 142 (2016).

[7] A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F. N. C. Wong, J. H. Shapiro, and V. K. Goyal, First-photon imaging, Science **343**, 58 (2014).

[8] A. Geiger, P. Lenz, and R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in *2012 IEEE conference on computer vision and pattern recognition* (IEEE, 2012) pp. 3354–3361. Also see, *e.g.*, https://www.tesla.com/autopilot; https://waymo.com/tech/

[9] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks, in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 293–308.

[10] G. B. Popko, T. K. Gaylord, and C. R. Valenta, Interference measurements between single-beam, mechanical scanning, time-of-flight lidars, Opt. Eng. **59**, 1 (2020).

[11] J. Hecht, Lidar for self-driving cars, Opt. Photon. News **29**, 26 (2018). Eye safety requires the emitting power of an agent to scale down as the inverse of the number of agents.

[12] K. P. Gurton, A. J. Yuffa, and G. W. Videen, Enhanced facial recognition for thermal imagery using polarimetric imaging, Opt. Lett. **39**, 3857 (2014).

[13] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu, Cats: A color and thermal stereo benchmark, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). Also see `http://thermalradar.com/`, where ranging has to be active.

[14] K. Schwab, The fourth industrial revolution: what it means, how to respond, Foreign Affairs **12**, 2015 (2015).

[15] B. L. Risteska Stojkoska and K. V. Trivodaliev, A review of internet of things for smart home: Challenges and solutions, J. Cleaner Prod. **140**, 1454 (2017).

[16] `https://mailchi.mp/statista/autonomous_cars_20200206?e=145345a469`.

[17] `https://resources.oxfordeconomics.com/how-robots-change-the-world`.

[18] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. **70**, 41 (2018).

[19] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436 (2015).

[20] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science **349**, 255 (2015).

[21] R. Gade and T. B. Moeslund, Thermal cameras and applications: a survey, Mach. Vis. Appl. **25**, 245 (2014).

[22] K. Tang, K. Dong, C. J. Nicolai, Y. Li, J. Li, S. Lou, C.-W. Qiu, D. H. Raulet, J. Yao, and J. Wu, Millikelvin-resolved ambient thermography, Sci. Adv. **6**, eabd8688 (2020).

[23] M. Henini and M. Razeghi, *Handbook of infrared detection technologies* (Elsevier, 2002).

[24] A. Haque, A. Milstein, and F.-F. Li, Illuminating the dark spaces of healthcare with ambient intelligence, Nature **585**, 193 (2020).

[25] K. Beier and H. Gemperlein, Simulation of infrared detection range at fog conditions for enhanced vision systems in civil aviation, Aerosp. Sci. Technol. **8**, 63 (2004).

[26] E. Newman and P. Hartline, Integration of visual and infrared information in bimodal neurons in the rattlesnake optic tectum, Science **213**, 789 (1981).

[27] A. Gillespie, S. Rokugawa, T. Matsunaga, J. S. Cothern, S. Hook, and A. B. Kahle, A temperature and emissivity separation algorithm for advanced spaceborne thermal emission and reflection radiometer (aster) images, IEEE Trans. Geosci. Remote Sens. **36**, 1113 (1998).

[28] A. Baldridge, S. Hook, C. Grove, and G. Rivera, The aster spectral library version 2.0, Remote Sens. Environ. **113**, 711 (2009).

[29] R. Szeliski, *Computer Vision – Algorithms and Applications*. (Springer, London, 2011).

[30] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Inf. Fusion **14**, 28 (2013).

[31] C. E. R. Lopes and L. B. Ruiz, On the development of a multi-tier, multimodal wireless sensor network for wild life monitoring, in *2008 1st IFIP Wireless Days* (IEEE, 2008) pp. 1–5.

[32] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network, Neurocomputing (2022).

[33] W. F. Wright and P. A. Mackowiak, Why temperature screening for coronavirus disease 2019 with noncontact infrared thermometers does not work, Open Forum Infect. Dis. **8**, 10.1093/ofid/ofaa603 (2020).

[34] P. Ghassemi, T. J. Pfefer, J. P. Casamento, R. Simpson, and Q. Wang, Best practices for standardized performance testing of infrared thermographs intended for fever screening, PLoS One **13**, 1 (2018).

[35] C. Magalhaes, J. M. R. Tavares, J. Mendes, and R. Vardasca, Comparison of machine learning strategies for infrared thermography of skin cancer, Biomed. Signal Process. Control **69**, 102872 (2021).

[36] D. V. Reddy, R. R. Nerem, S. W. Nam, R. P. Mirin, and V. B. Verma, Superconducting nanowire single-photon detectors with 98% system detection efficiency at 1550 nm, Optica **7**, 1649 (2020).

METHODS

TeX degeneracy For an object α, its spectral radiance $S_{\alpha\nu}$ given by Eq. 1 in the main text is invariant if we change its physical attributes $\{T_\alpha, e_{\alpha\nu}, X_{\alpha\nu}\}$ to $\{T'_\alpha, e'_{\alpha\nu}, X'_{\alpha\nu}\}$, where T'_α is an arbitrary temperature value, $X'_{\alpha\nu}$ is an arbitrary spectral texture curve, and the spectral emissivity curve $e'_{\alpha\nu}$ is given by

$$e'_{\alpha\nu} = \frac{e_{\alpha\nu}[B_\nu(T_\alpha) - X_{\alpha\nu}] + [X_{\alpha\nu} - X'_{\alpha\nu}]}{B_\nu(T'_\alpha) - X'_{\alpha\nu}}.$$

Here, ν is the wavenumber, and B is the blackbody radiation. Physical states having distinct triplets of TeX attributes but having the same observed heat signal S_ν are addressed as TeX degeneracy.

TeX decomposition We exploited a material library $\mathcal{M}$ and the mathematical structure of X to overcome TeX degeneracy. In this paper, we proposed several approaches to fully solve TeX decomposition, depending on the specific problem. Learning based TeX-Net (see Extended Data Fig. 1 for the architecture) utilizing both spatial and spectral information in decomposing TeX is our general solution, whereas we have also provided the analytical inverse function, least-squares estimator, and TeX-SGD (semi-global decomposition) as non-machine-learning baselines (see Sec.SIIIB of the Supple. Info.). For the HADAR database and HADAR prototype-2 experiments, we used TeX-Net and TeX-SGD. For human-robot identification in Fig. 3, we used a one-dimensional 3-layer convolutional neural network (with ReLU activation) followed by a softmax classifier to recognize material category m_α from collected radiation spectrum $S_{\alpha\nu}$. Categorical cross entropy was used as the loss function, and Adam optimizer was used. For Extended Data Fig. 7, we used the analytical inverse function of T and X following a material classifier as mentioned above. In HADAR prototype-1 experiments, we used the least-squares estimator for TeX decomposition. In Fig. 5, the material library was drawn from the NASA JPL ECOSTRESS spectral library [37], with emissivity shown in Sec.SIVD of the Supple. Info. In HADAR prototype-2 experiments, we used a semantic library, instead of the material library, estimated from the data itself, see Sec.SVB of the Supple. Info. With the least-squares estimator, we verified that TeX decomposition is crucial for vision applications and goes beyond the traditional TE (temperature-emissivity) separation approach, see Fig. S20 of the Supple. Info.

TeX vision and pseudo-TeX vision Motivated by coloring convention in existing literature where different colors represent different categories, we use HSV format to represent TeX with mapping $\mathrm{H} = e, \mathrm{S} = T, \mathrm{V} = X$. In this TeX vision, different hues of color represent different materials, saturation indicates temperature, and brightness gives textures. The texture recovered in TeX vision is from increased information in sensor data, in contrast to state-of-the-art approaches such as Automatic gain control (AGC). AGC is also applied to TeX to get better visualization. See Sec. SIIIC of the Supple. Info. for more details about TeX vision and how the TeX vision image is formed.

As TeX vision requires the input of hyperspectral heat cubes, we also propose pseudo-TeX vision to extend its applications to common thermal datasets without spectral resolution. See Sec. SIIID of the Supple. Info. for details.

Monte Carlo path tracing The HADAR database is an LWIR (long-wave infrared) stereo-hyperspectral database mostly synthesized by exploiting Planck's law and Kirchhoff's law in Blender Cycles renderer. The database has been made public and is available at https://github.com/FanglinBao/HADAR, where detailed descriptions can be found. In this paper, Lambertian (diffusive) reflectance was used for simplicity. Samples per pixel was 2000. We also implemented path tracer according to Eq. 1, per wavenumber, with OpenGL (version 4.6) and Compute Unified Device Architecture (CUDA, version 10.2) on GPU. For Fig. 4, ray depth was 8 with (thermal imaging) or without (HADAR perception) direct emission at the final step. Rendering wavenumber was $769\,\mathrm{cm}^{-1}$. The ground and the sky were at $20\,\mathrm{C}^\circ$. The emissivity pattern of the ground was generated by mapping a regular road image to emissivity between 0.8 and 1, to maximize texture loss in thermal vision. Synthetic textures for all other scenes were surface normal textures. The optical image was rendered without direct emission from objects but with sky illumination. Image size is 640×480. For Fig. 2, ray depth was 8 with normal texture on an opaque glass bulb. For Extended Data Fig. 7, ray depth was 1 for 11 discrete wavenumbers within $715 \sim 1250\,\mathrm{cm}^{-1}$.

HADAR estimation theory Here, we provide a short answer to the question of 'How many photons are needed to identify the target material'. Full derivations of fundamental bounds for both detection and ranging are given in Sec. SII of the Supple. Info. Identifying the target between two candidate materials $e_{1\nu}$ and $e_{2\nu}$ is mapped to estimating the fraction g of a mixture of these two materials, $e_\nu = (1-g)e_{1\nu} + ge_{2\nu}$, with $g = 0$ indicating one material and $g = 1$ the other. The normalized spectrum $p_{\alpha\nu} \equiv S_{\alpha\nu}/\int S_{\alpha\nu}\mathrm{d}\nu$ describes the spectral probability density for one incident photon. The Fisher information matrix (FIM) regarding unknown parameters $\Theta = \{g, T, V_0\}$ reads $J_{ij} = NJ^0_{ij}/(1+\gamma)$, where $J^0_{ij} = \int \frac{\partial_i p_{\alpha\nu} \cdot \partial_j p_{\alpha\nu}}{p_{\alpha\nu}} \mathrm{d}\nu$ is the single-photon FIM, N is the total number of photon, T is the temperature, V_0 is the thermal lighting factor of the environment, $i, j \in \Theta$, and $\gamma \equiv \gamma_1 N + \gamma_0$ is the electronic-noise power normalized by the shot-noise power. The Cramér-Rao bound $\sigma^2 \equiv [1/J]_{gg}$ puts a lower bound to the variance of any

unbiased estimator of g, and the statistical distance is $d \equiv 1/2\sigma$. Only depending on material, semantic distance $d_0 \equiv 1/2\sigma_0$ with $\sigma_0^2 \equiv [1/J^0]_{gg}$ describes how different two materials are with each other, under the TeX degeneracy. The related but distinct concept of statistical distance depicts the overall distinguishability of two spectra but depends on the detector and measurement time. Our semantic distance approach captures the intrinsic identifiability of objects from the scene alone. The detection probability (true positive rate) is given by $P = [1 + \mathrm{erf}(d/\sqrt{2})]/2$. The Shannon information of material is given by $I = \log_2 P - \log_2(1/2)$.

In evaluating the theoretical bound on HADAR identifiability in Fig. 3, we used $V_0 = 0.5$ (α suppressed) and $T_0 = 20\,\mathrm{C}^\circ$. Target distance was 30 m when input signal was evaluated. In the Monte Carlo simulations in Fig. 3b, we first found the nearest robot condition ($T = 83.46\,\mathrm{C}^\circ$, $V_0 = 0.17$) to human ($T = 37\,\mathrm{C}^\circ, V_0 = 0.5$). We used 150 sampling normalized photon numbers, and then for each normalized photon number, we generated 5000 spectra ($715 \sim 1250\,\mathrm{cm}^{-1}$, $\Delta\nu = 1\,\mathrm{cm}^{-1}$) for each of two candidates with Monte Carlo simulation in the shot-noise limit. At last, we used machine learning (25% spectra for training, 25% for validation, and 50% for test) for material classification, and the test accuracy was used to compute Shannon information for each normalized photon number. The dimensionality curse for high spectral resolution (536 bands used) leads machine learning to over-fitting, and slight deviation between Monte Carlo simulation and theoretical prediction can be observed in Fig. 3b. Once the dimensionality curse is relieved, perfect agreement can be reached, see Fig. S7 in Supple. Info. where all spectra are down-sampled into 3 spectral bands (dimension = 3) for both theory and machine learning.

In evaluating the theoretical bound on ranging error in Fig. 4 without photon number, we used $J_x = (\partial_x N_{iq})^2/(N_{iq} + \sigma^2)$ [Eq. S37 in Supple. Info.] instead. Variance was estimated by matching corresponding pixels according to the ground truth disparity and computing the signal fluctuation. Finite difference was used to approximate derivative. $b = 0.2\,\mathrm{m}$ and $f = 1.4\,\mathrm{cm}$. The block size in correlation-based sub-pixel block matching was 5×5 (see Fig. S10 of Supple. Info. for AI results).

Guiding public policy The HADAR identifiable criterion is $\frac{Nd_0^2}{1+\gamma} = 1$, which means one can identify the target material if $\frac{Nd_0^2}{1+\gamma} \geq 1$. The semantic distance between human body (skin) and robot (aluminum) in Fig. 3 is calculated to be $d_0 \approx 0.001$. This requires $\frac{N}{1+\gamma} \geq 10^6$ to identify the target if the environment is at $T_0 = 20\,\mathrm{C}^\circ$ and $V_0 = 0.5$ (see Fig. 3c). The observed photon number N is related to the human-robot scene, as well as the f-number (focal length f over the aperture size D), exposure time t, and pixel size A_p, see the heat signal model in Sec. SI of the Supple. Info. Eventually, the above identifiable criterion leads to the minimum requirement of the hardware configurations, $\frac{tA_\mathrm{p}}{(1+\gamma)(f/D)^2} \geq 5 \times 10^{-16}$. This minimum requirement of the hardware will guide the public policies in the AI industry. For example, the lowest detectivity (or highest NEP), the smallest aperture size, the highest frame rate and hence the maximum travelling speed, *etc.*, must meet the above inequality so as to be able to identify human vs. robot. If the detector doesn't meet the above requirement, its collected data will be insufficient in information. No matter how much data is collected and used to train a neural network (how large the training volume is), machine learning cannot perform well (see the machine learning performance in Fig. 3b of the main text when the photon number is insufficient, *i.e.*, the normalized photon number is below 1). If the detector is given, *e.g.*, the FLIR A325sc camera, we have $\frac{tA_\mathrm{p}}{(1+\gamma)(f/D)^2} = 8.16 \times 10^{-18}$ in one image frame. To meet the criterion, we must have $d_0 > 0.0078$, which means the FLIR A325sc camera can only distinguish sufficiently different material pairs in one image frame, such as, organic skin vs. glass mannequin ($d_0 = 0.049$). This minimum semantic distance identifiable by the given detector will also guide the public policies in the AI industry. For example, in which scenario the given camera can be used, and in which scenario the camera cannot. Likewise, our fundamental limit of HADAR ranging accuracy can also put requirements on hardware configurations or restrict travelling speed, etc., and guide the public policies.

Our fundamental limits bound the average machine learning performance due to the shot noise and detector noise. Lucky evaluation events could occur but they will fluctuate around the average bounds, as can be seen in Fig. 3b and insets of Fig. 4. Human error or software bugs are not considered in our bounds, but our bounds are useful because they depict the optimal performance of machine learning when human error and software bugs are completely corrected. Therefore, our bounds related to physical laws of thermal photonic information theory can be used as a guidance to public policies.

Thermal camera specifications Our FLIR A325sc thermal camera is a science-grade high-performance radiometric camera (price $\sim \$10,000$). It is equipped with an uncooled Vanadium Oxide (VOx) microbolometer detector that produces thermal images of 320×240 pixels. Detector pitch is $25\,\mu$m. Pixel size is approximated as $12\,\mu$m. Time constant is 12 ms. Focal length is 18 mm. And f-number is 1.3. Noise equivalent temperature difference (NETD) is typically $< 50\,$mK and characterized to be 47.8 mK. FLIR A325sc was available when the experiments in this paper were designed and conducted. We note that now it has been discontinued and replaced by a more advanced model, FLIR A655sc. The latter has a 640×480 pixel array with typical NETD $< 30\,$mK, but it is twice as expensive. A better camera will give better HADAR data. Since the advantage of HADAR over traditional thermal vision comes from the spectral resolution

and the theory we used to interpret the hyperspectral data, FLIR A325sc presents a better functionality-cost balance.

FTIR specifications Our Nicolet iS50 is equipped with a cooled (liquid Nitrogen) MCT-A detector, with sensor element size being 1 mm. Its special detectivity is $4.7\times10^{10}\,\mathrm{cm}\cdot\mathrm{Hz}^{1/2}/\mathrm{W}$, with preamplifier bandwidth being 175 kHz. Spectral resolution $\Delta\nu = 0.48\,\mathrm{cm}^{-1}$ within $769 \sim 1332\,\mathrm{cm}^{-1}$ is used in this paper. Aperture of external optics is 2 inches. Focal length of external optics is about 10 cm. Optical efficiency is approximated as 0.9 in deriving the Cramér-Rao bound on temperature accuracy in Extended Data Fig. 9.

Prototype HADAR calibration and data collection Our HADAR prototype-1 was built with FLIR A325sc as the detector, plus 10 thermal infrared filters (price $\sim$ \$10,000) from Spectrogon to retrieve spectral resolution. A gold mirror was also mounted on the filter wheel to monitor the status of the detector in real time. Once started, the detector was left to stabilize for at least 30 min to warm up. In experiments when the detector exchanged heat with the scene, mirror signal was checked so that data collected with very different detector status was ignored. The filter transmittance curves were characterized by Nicolet iS50. The spectral response curve of the camera was calibrated with standard blackbody source (EOI. Inc. DCN1000N7). In experiments, a uniform reference object was used to further calibrate camera's self radiation pattern, as well as the side effect of the filter wheel acting as an out-of-focus diaphragm. The experimental data collected were left and right heat cubes of dimension Height $\times$ Width $\times$ Channel $= 240 \times 320 \times 10$. Number of channels was the number of filters. See Extended Data Fig. 10 for HADAR prototype-1 calibration and data collection.

The HADAR prototype-2 is based on a pushbroom hyperspectral imager with a cooled HgCdTe sensor. To collect the real-world experimental data, we formed a partnership with DARPA (The Defense Advanced Research Projects Agency, through the Invisible Headlights project) and the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center, U.S. Army). The pushbroom hyperspectral imager gives 256 spectral bands, but its price is over a million dollars. The focal length is 50 mm, and the f-number is f/0.9. It uses a 256×256 focal plane array with 40 μm pitch pixels. The typical noise of the sensor is around 1 'microflick', which at 10 μm wavelength corresponds to about a 1000:1 signal-to-noise ratio. Explicitly, for a 300 K temperature scene at 10 μm, the noise equivalent temperature difference is around 63 mK. Denoising and extrinsic calibrations can be found in Sec. SV of the Supple. Info.

In our proof-of-concept experiments, we used the filter-wheel approach to demonstrate the HADAR prototype-1. The filter-wheel approach is time consuming but cost effective, suitable for low-end HADAR applications. In contrast, HADAR prototype-2 with a pushbroom sensor was demonstrated for high-end HADAR applications. HADAR can also be implemented by other approaches with mosaic sensors, gratings, prisms, interferometers, or Fabry-Perot cavities, depending on the desired spectral resolution, spatial resolution, data acquisition speed, or functionality-cost balance.

Computational efficiency and deploy-ability (1) Our TeX-Net has about 0.5M weights in total. The evaluation of our TeX-Net (GPU Nvidia RTX A6000 48GB) takes 42.4 ms. Data collection of the currently used pushbroom hyperspectral imager takes around 1s, but the filter-wheel approach can be optimized down to around 10ms with high-speed filter wheel (*e.g.*, Telops multispectral cameras). Overall, our results show that HADAR data collection and processing can support up to 20 Hz TeX vision frame rate. Pursuing higher frame rate motivates further research on new hyperspectral imaging sensors to collect thermal infrared data and photonic neural networks for TeX decomposition. (2) Our generalized HADAR theory does not require the input of a material library and hence is free of on-site library collection/calibration. This enables real-time HADAR applications. Our HADAR prototype-2 experiment is a field test with the HADAR sensor mounted on a car. Corresponding TeX vision results on the DARPA IH test data shows the deploy-ability of HADAR, see Extended Data Figs. 3 and 4.

Standard depth metrics Let $pred$ and gt denote predicted and ground truth depth, respectively. D represents the set of all predicted depth values. $|\cdot|$ returns the number of elements, and $||\cdot||$ returns the absolute value. The standard depth metrics used in Fig. 6 are defined as below.

Absolute and Relative Error,

$$\text{Abs\_Rel} = 1/|D| \cdot \Sigma_{pred\in D}||gt - pred||/gt.$$

Squared Relative Error,

$$\text{Sq\_Rel} = 1/|D| \cdot \Sigma_{pred\in D}||gt - pred||^2/gt.$$

Root Mean Squared Error,

$$\text{RMSE} = \sqrt{1/|D| \cdot \Sigma_{pred\in D}||gt - pred||^2}.$$

Root Mean Squared Log Error,

$$\text{RMSE\_Log} = \sqrt{\frac{1}{|D|}\sum_{pred\in D}||\log(gt) - \log(pred)||^2}.$$

δ_t Accuracy,

$$\delta_t = \frac{1}{|D|}|\{pred \in D|\max(\frac{gt}{pred}, \frac{pred}{gt}) < 1.25^t\}|.$$

HADAR thermography Note that existing thermal imaging measures the total radiance and approximates emissivity e_ν as a default parameter or manually input constant e. This causes the temperature readout to be biased and incorrect. Furthermore, when two

different materials at different temperatures happen to emit the same total radiance, thermal imaging predicts the same temperature leading to the thermal camouflage effect [38, 39] (the integral version of TeX degeneracy; Extended Data Fig. 9b). We use 2 stripes of tapes on plastics (not shown) and 3 patches of tapes on silicon (3 rectangles in Extended Data Fig. 9a-d) to read out the temperature without influence from non-trivial emissivity of sample materials. Non-uniform heating effect and self-radiation of the thermal camera (FLIR A325sc) are calibrated and removed from data in Extended Data Fig. 9b before TeX decomposition. In HADAR TeX decomposition, two materials are identified automatically and temperature is estimated accordingly, revealing an otherwise hidden HADAR alphabet pattern.

In statistical analysis (Extended Data Fig. 9e-f, based on Nicolet iS50), response curve and environmental radiation are calibrated, in addition to the dark noise caused by device's self radiation, before experiments. After calibration, 20 measurements are taken for each of 16 heating powers. Conventional thermal imaging with default emissivity ($e = 0.95$) severely deviates from the ground truth obtained by the thermocouple. Manually input emissivity is calibrated at the lowest heating power, but this approach also deviates as the heating power increases, since the calibration is inaccurate once again caused by thermal camouflage. These issues are overcome with HADAR thermography which estimates temperature unbiasedly. We note that the root-mean-square-error (RMSE) beats both infrared as well as contact thermography. This is not surprising as even contact thermography using thermocouples have inevitable errors arising from noisy physical contacts.

The temperature difference between tumor cells and regular cells in skin cancer could be as high as $0.25\,\mathrm{C}^\circ$. However, the signal captured by a thermal camera is the radiance S that includes scattering contributions from the environment (X) along with direct emission from the tumor cells. Having a hot object (other people, instruments) in the patient room (or, considering X or not) makes a striking difference in estimated temperatures. As an example, the emissivity of skin can be well approximated as a constant of 0.95, and we assume that the environment is a blackbody ($X = B$) to approximately see the errors arising from ignoring the environment. The presence/absence of X is equivalent to a 5% relative difference of $B(T)$, which corresponds to $3\,\mathrm{C}^\circ$ temperature variation around the standard $37\,\mathrm{C}^\circ$ temperature. This error arising from ignoring the environmental signal is much larger than the temperature difference caused by tumor cells. To minimize this effect, accurate thermography is limited to 'either an open-area, outdoor environment under clear sky (cloud free), or using a cold-plate setup', which restricts the indoor applications for fever surveillance. Since TeX vision decomposes S, HADAR can reach the Cramér-Rao bound of temperature by properly estimating e and X and hence is promising for reliable skin cancer detection.

Data availability The data supporting the findings of this study is available in the paper. The HADAR database is available at https://github.com/FanglinBao/HADAR.

Code availability The custom-designed codes are provided along with the HADAR database at https://github.com/FanglinBao/HADAR.

Acknowledgement This work was supported by the IH (Invisible Headlights) project from DARPA (Defense Advanced Research Projects Agency). We thank the Army night-vision team (Infrared Camera Technology Branch, DEVCOM C5ISR Center, U.S. Army) for the help in collecting HADAR prototype-2 experimental data. We thank Ziyi Yang for her help in experiments.

Author contributions F.B. and Z.J. conceived the idea. F.B. led and Z.J. supervised the project. F.B. developed and L.Y. contributed to the HADAR estimation theory. F.B. generated the HADAR database and designed experiments. X.W. built HADAR prototype-1 and conducted experiments. F.B. analyzed experimental data. S.H.S., G.S., F.B., V.A., and V.N.B. contributed to machine learning. F.B. prepared and all authors revised the manuscript.

Competing interests The authors declare no competing interests.

Correspondence and requests for materials should be addressed to Z. Jacobs.

[37] A. Baldridge, S. Hook, C. Grove, and G. Rivera, The aster spectral library version 2.0, Remote Sens. Environ. **113**, 711 (2009).

[38] Y. Qu, Q. Li, L. Cai, M. Pan, P. Ghosh, K. Du, and M. Qiu, Thermal camouflage based on the phase-changing material gst, Light: Sci. Appl. **7**, 26 (2018).

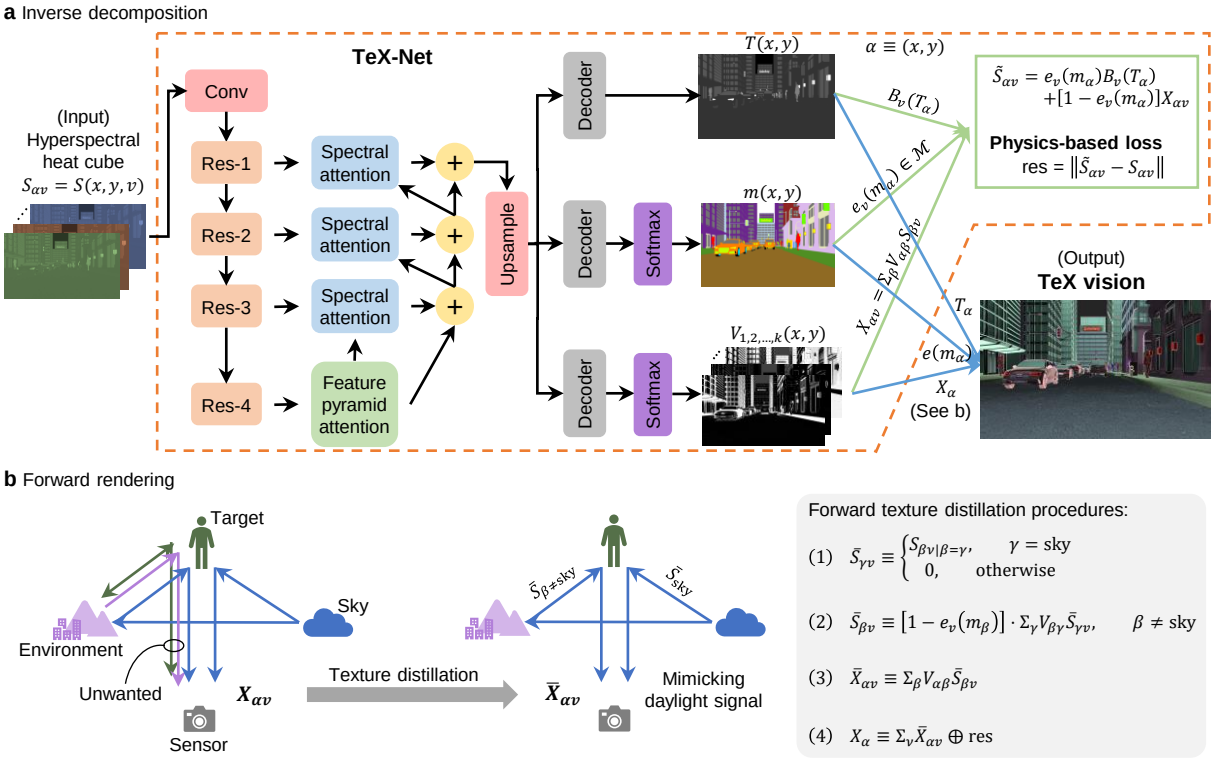
[39] M. Li, D. Liu, H. Cheng, L. Peng, and M. Zu, Manipulating metals for adaptive thermal camouflage, Sci. Adv. **6**, 10.1126/sciadv.aba3494 (2020).

[40] H. Li, P. Xiong, J. An, and L. Wang, Pyramid attention network for semantic segmentation, in *British Machine Vision Conference 2018, Newcastle, UK* (2018) p. 285.

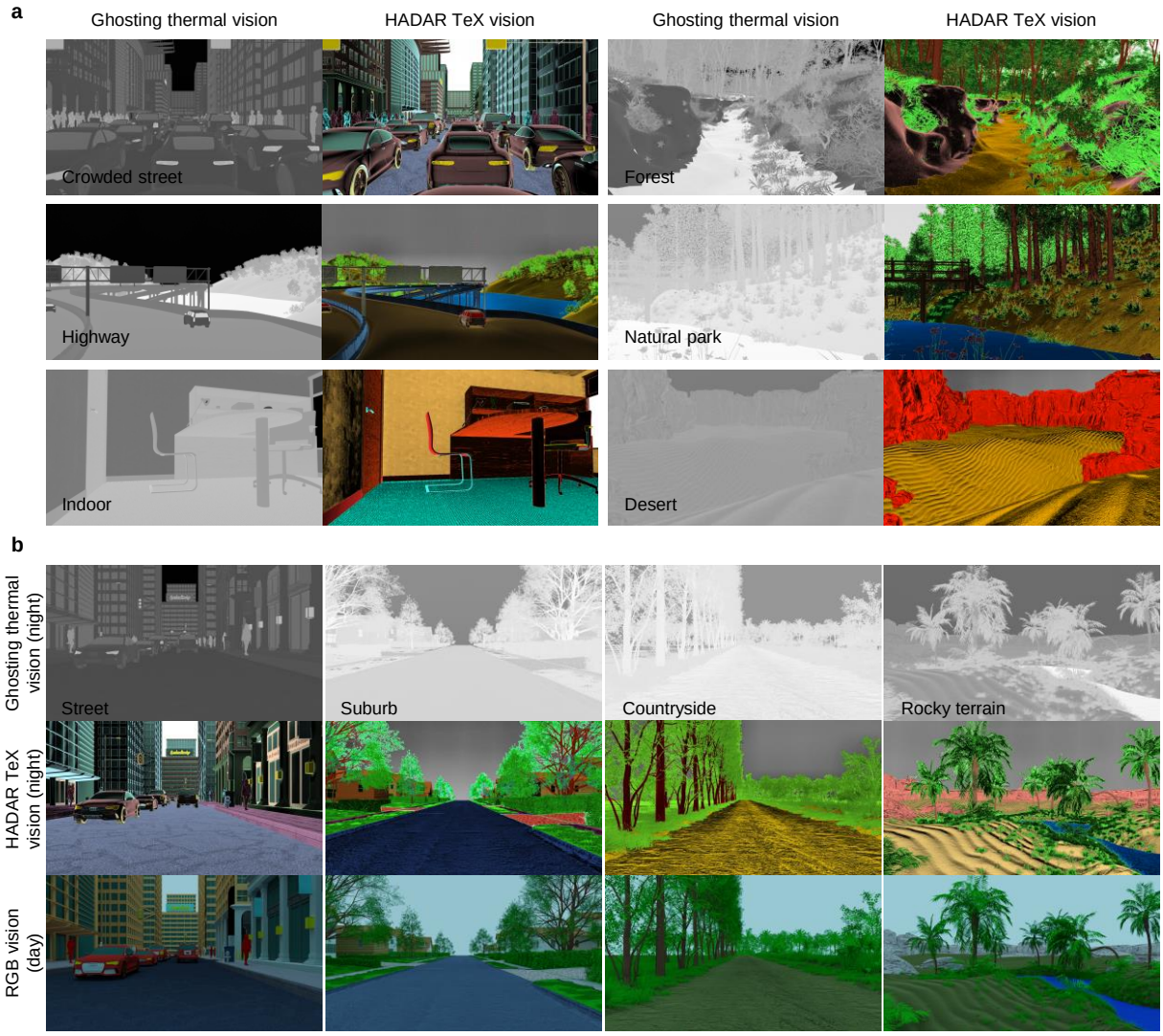
[41] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, Deeppruner: Learning efficient stereo matching via differentiable patchmatch, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

[42] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network, Neurocomputing (2022).

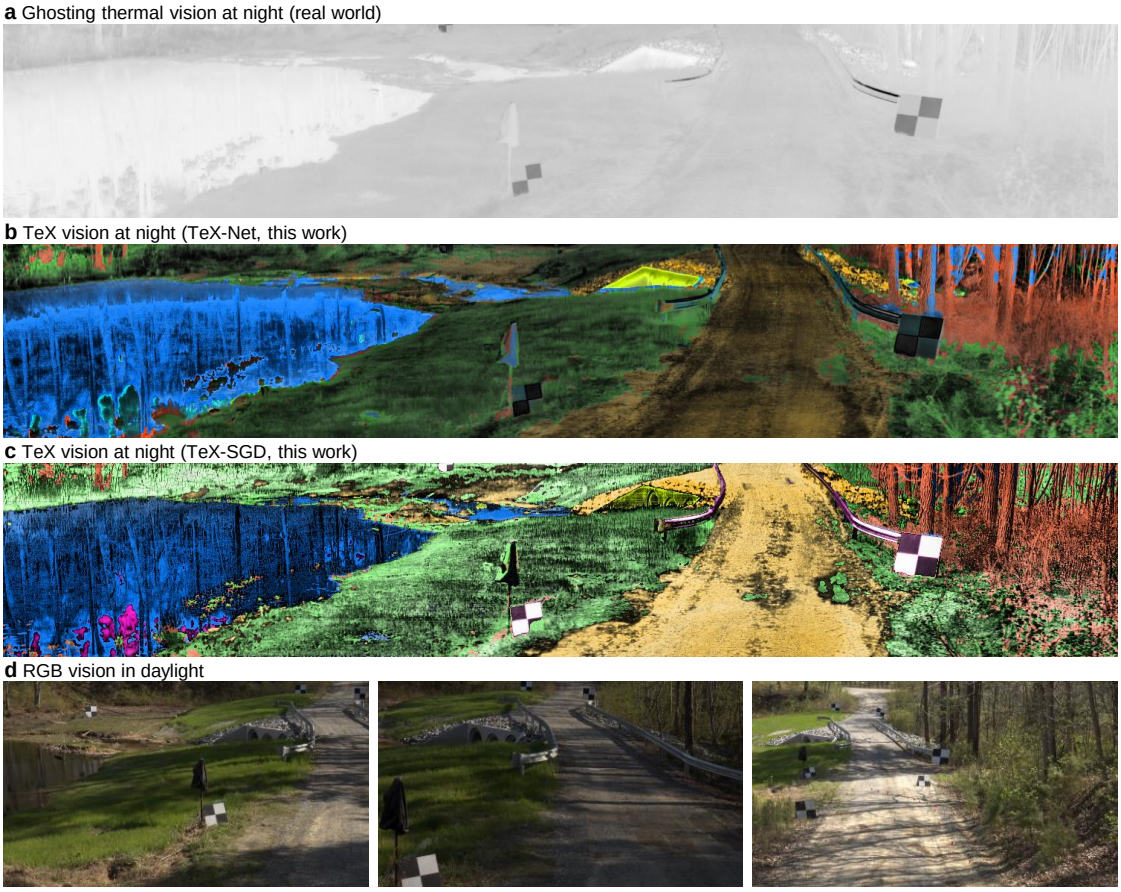
[43] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, Dual attention network for scene segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019) pp. 3146–3154.



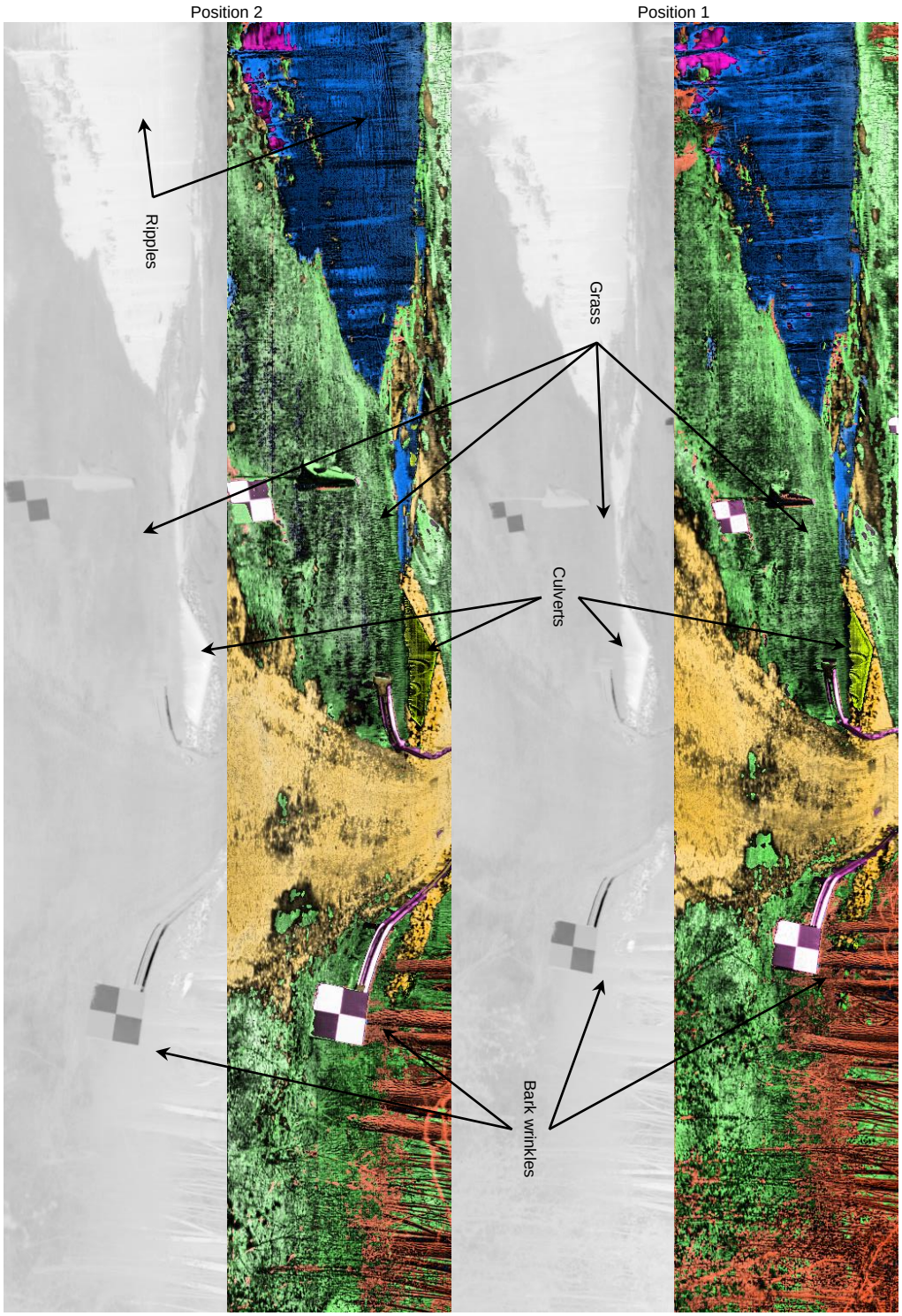
Extended Data Fig. 1. HADAR TeX vision algorithms. a, Architecture of our TeX-Net for inverse TeX decomposition. TeX-Net is physics-inspired for three aspects. Firstly, TeX decomposition of heat cubes relies on both spatial patterns and spectral thermal signatures. This inspires the adoption of spectral and pyramid (spatial) attention layers [40] in the UNet model. Secondly, due to TeX degeneracy, the mathematical structure, $X_{\alpha\nu} = \sum_\beta V_{\alpha\beta} S_{\beta\nu}$, has to be specified to ensure the uniqueness of inverse mapping, and hence it is essential to learn thermal lighting factors V instead of texture X. That is, TeX-Net cannot be trained end-to-end. Here, α, β, and γ are indices of objects, and ν is the wavenumber. X_α is constructed with V and $S_{\beta\nu}$ indirectly, where $S_{\beta\nu}$ is the down-sampled $S_{\alpha\nu}$ to approximate k most significant environmental objects. Thirdly, the material library $\mathcal{M}$ and its dimension are key to the network. TeX-Net can either be trained with ground truth T, m, and V in supervised learning, or alternatively, with material library $\mathcal{M}$, Planck's law $B_\nu(T_\alpha)$, and the mathematical structure of $X_{\alpha\nu}$ in unsupervised learning. In supervised learning, the loss function is a combination of individual losses with regularization hyper-parameters. In unsupervised learning, the loss function defined on the re-constructed heat cube is based on physics models of the heat signal. In practice, a hybrid loss function with T, e, V contributions (50%) in addition to the physics-based loss (50%) is used. In this work, we have also proposed a non-machine-learning approach, the TeX-SGD (Semi-Global Decomposition), to generate TeX vision. TeX-SGD decomposes TeX pixel per pixel, based on the physics loss and a smoothness constraint, see Sec. SIIIA-B of the Supple. Info. for more details. Res-1/2/3/4 are ResNet50 with downsampling. The plus symbol is addition operation followed by upsampling. b, Texture distillation reconstructs the part of scattered signal that originates only from sky illuminations. The texture distillation process is to mimic daylight signal as X to form TeX vision, and it is done by evaluating the HADAR constitutive equation in a forward way, with the physical attributes solved out in TeX-SGD or TeX-Net. It removes the unwanted effect of other environmental objects being the light source which is unfamiliar in daily experience. The process can be described in 4 steps. Here, step (1) is the initialization that keeps only the sky illumination on and turns other radiations off. Step (2) is the iterative HADAR constitutive equation without direct emission. Evaluating it multiple times gives the multiple scattering effect. Note that the ground truth texture partly remains in the physics-based loss, res, due to cutoffs on scattering and/or number of environmental objects. The final estimated texture in step (4) is a fusion of distilled scattered signal $\bar{X}_{\alpha\nu}$ and the physics-based loss res. Arrows in (b) indicate thermal radiation emitted/scattered along the arrow direction. The TeX-Net code, pre-trained weights, and a sample implementation of texture distillation is available at https://github.com/FanglinBao/HADAR.



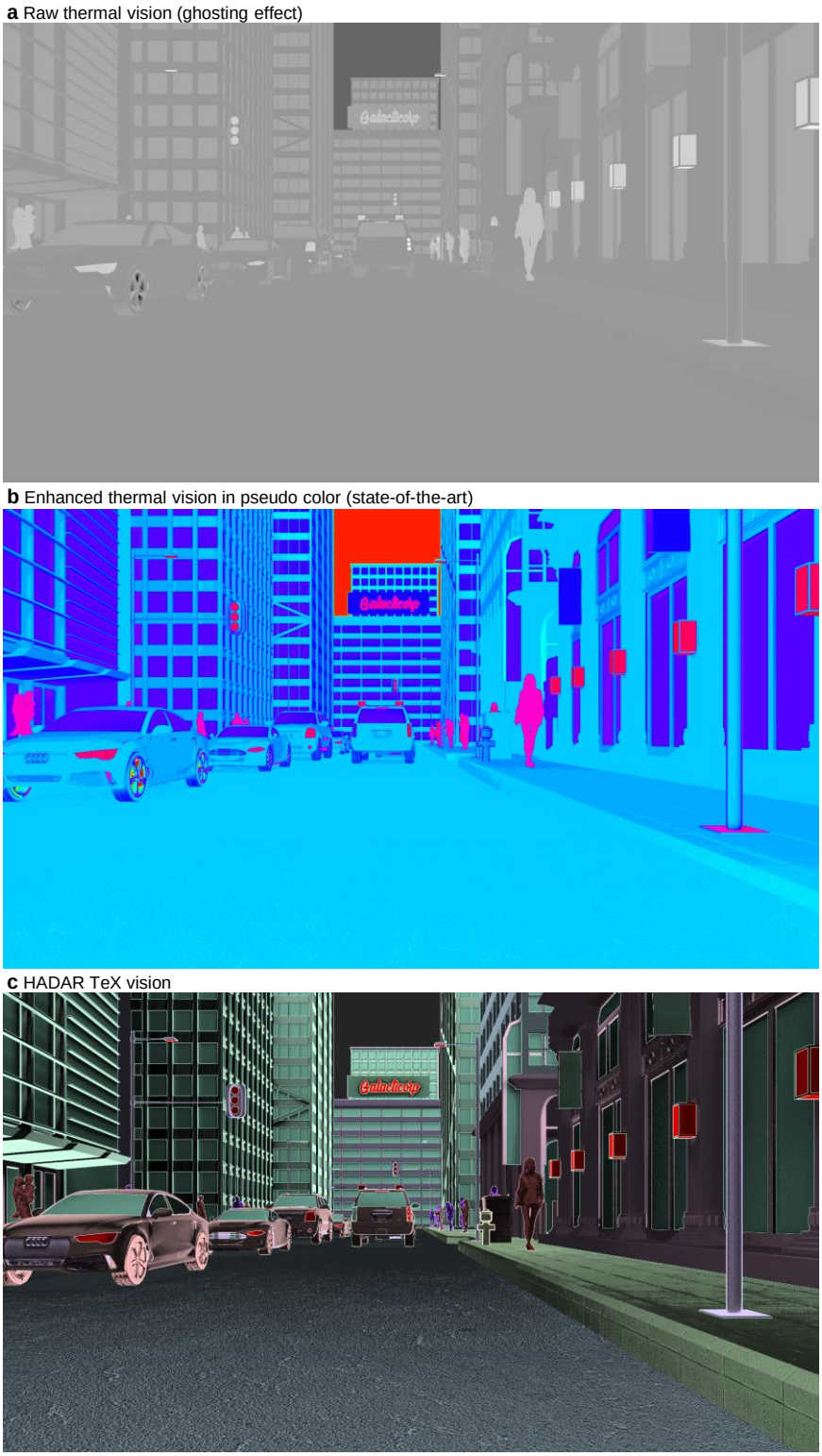
Extended Data Fig. 2. HADAR database and demonstrated TeX vision show that HADAR overcomes the ghosting effect in traditional thermal vision and sees through the darkness as if it were day. TeX vision (color hue H = material e; saturation S = temperature T; brightness V = texture X) provides intrinsic attributes and enhanced textures of the scene to enable comprehensive understanding. Our HADAR database consists of 11 dissimilar night scenes covering most common road conditions that HADAR may find applications in. Particularly, the indoor scene is designed for robot helpers in smart home applications, while others are for various self-driving applications. Scene-11 is a real-world off-road scene and shall be shown in Extended Data Fig. 3. The database is a long-wave infrared stereo-hyperspectral database with crowded (*e.g.*, Crowded street) and complicated (*e.g.*, Forest) scenes, having multiple frames per scene and 30 different kinds of materials. The database is available at https://github.com/FanglinBao/HADAR. See Fig. S18 in the Supple. Info. for the TeX-Net performance on the HADAR database.



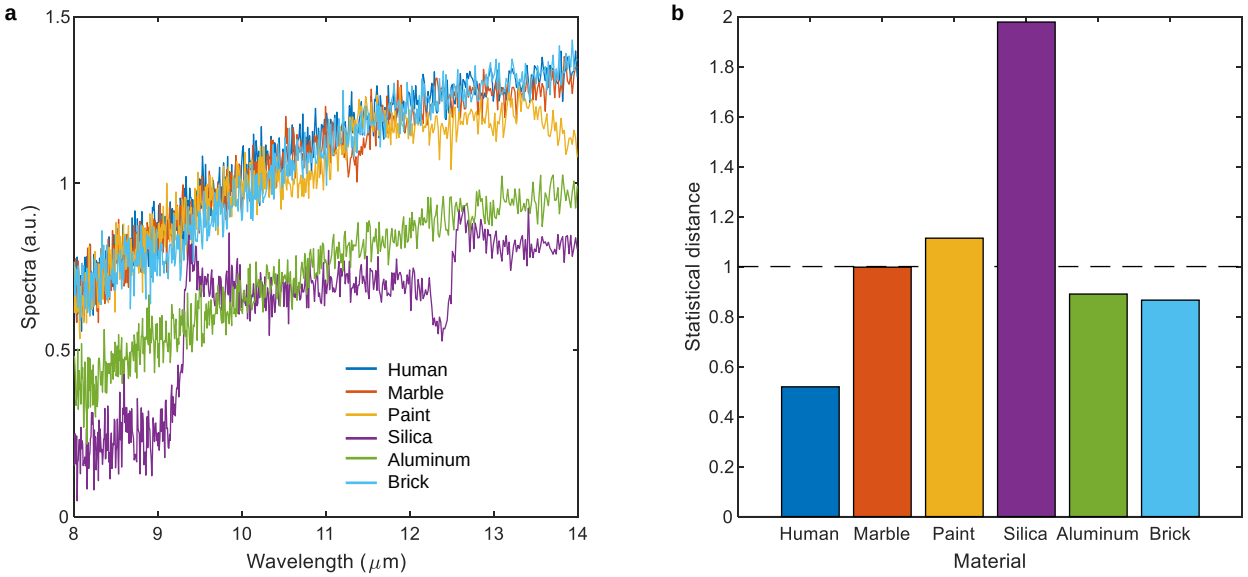
Extended Data Fig. 3. HADAR TeX vision demonstrated in real-world experiments (Scene-11 of the HADAR database) overcomes the ghosting effect in traditional thermal vision and sees through the darkness as if it were day. Here, TeX vision was generated by both TeX-Net and TeX-SGD (semi-global decomposition) for comparison. We used a semantic library instead of the exact material library for the TeX vision, see Sec. SVC of the Supple. Info. for more details. The semantic library consists of tree (brown), vegetation (green), soil (yellow), water (blue), metal (purple), and concrete (chartreuse). Water gives mirror images of trees and part of the sky beyond the view. Most of the water pixels can be correctly estimated as 'water', except for a small portion corresponding to sky image that has been estimated as 'metal', since metal also reflects the sky signal. TeX-Net utilizes both spatial information and spectral information for TeX decomposition, and hence its TeX vision is spatially smoother. In contrast, TeX-SGD mainly makes use of spectral information and decomposes TeX pixel per pixel. Compared with TeX-Net, we observed that TeX-SGD is better at material identification and texture recovery for fine structures, such as, bridge fence, bark wrinkles, and culverts. Note that the current TeX-Net was trained partially with TeX-SGD outputs. The above observations are not used to claim performance ranking between TeX-SGD and TeX-Net. Both TeX-Net and TeX-SGD confirm that HADAR TeX vision has achieved a semantic understanding of the night scene with enhanced textures comparable to RGB vision in daylight.



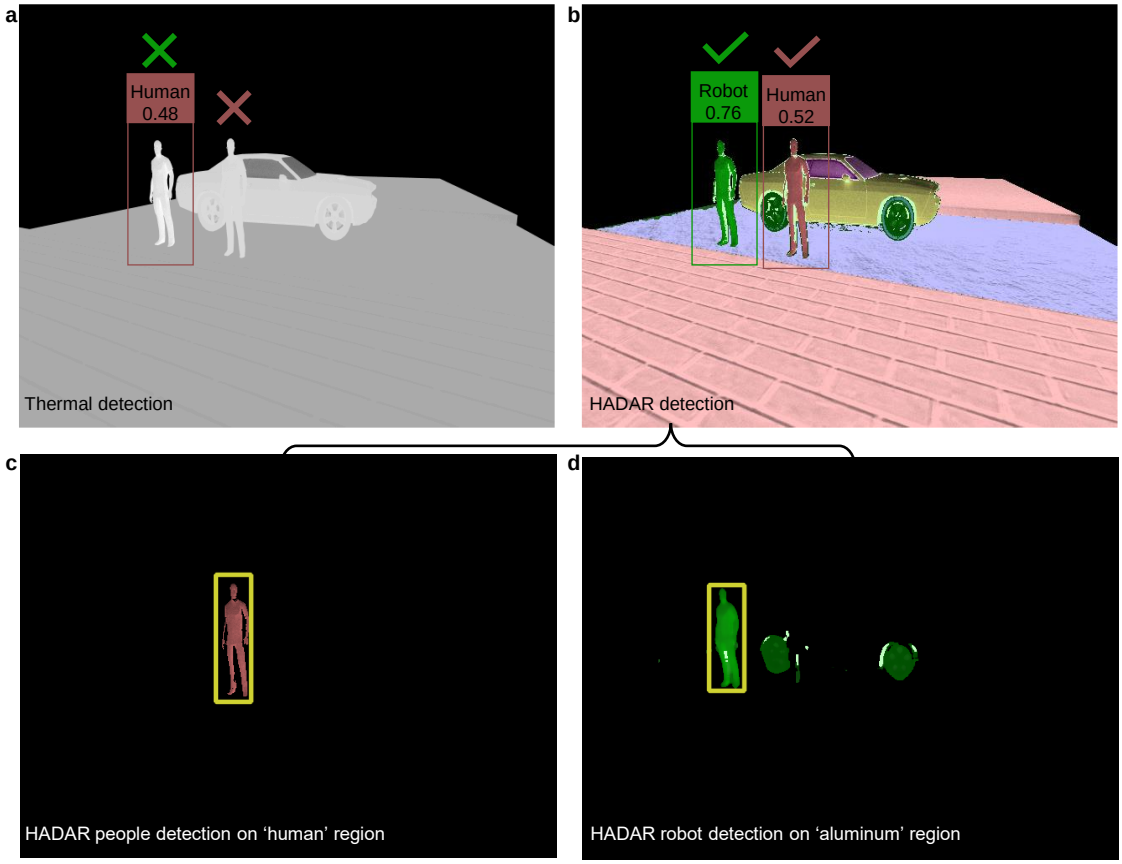
Extended Data Fig. 4. HADAR TeX vision recovers textures and overcomes the ghosting effect. Here, TeX vision is generated by TeX-SGD (semi-global decomposition). From right to left are TeX/thermal/TeX/thermal vision of an off-road night scene at two different positions. HADAR recovers fine textures such as water ripples, bark wrinkles, culverts, in addition to the great details of the grass lawn. The HADAR prototype-2 sensor is a focal plane array focusing at infinity. Close objects exhibit focus blur, while distant objects are beyond the spatial resolution to show fine details. Therefore, fine textures are mostly observed in a certain distance range.



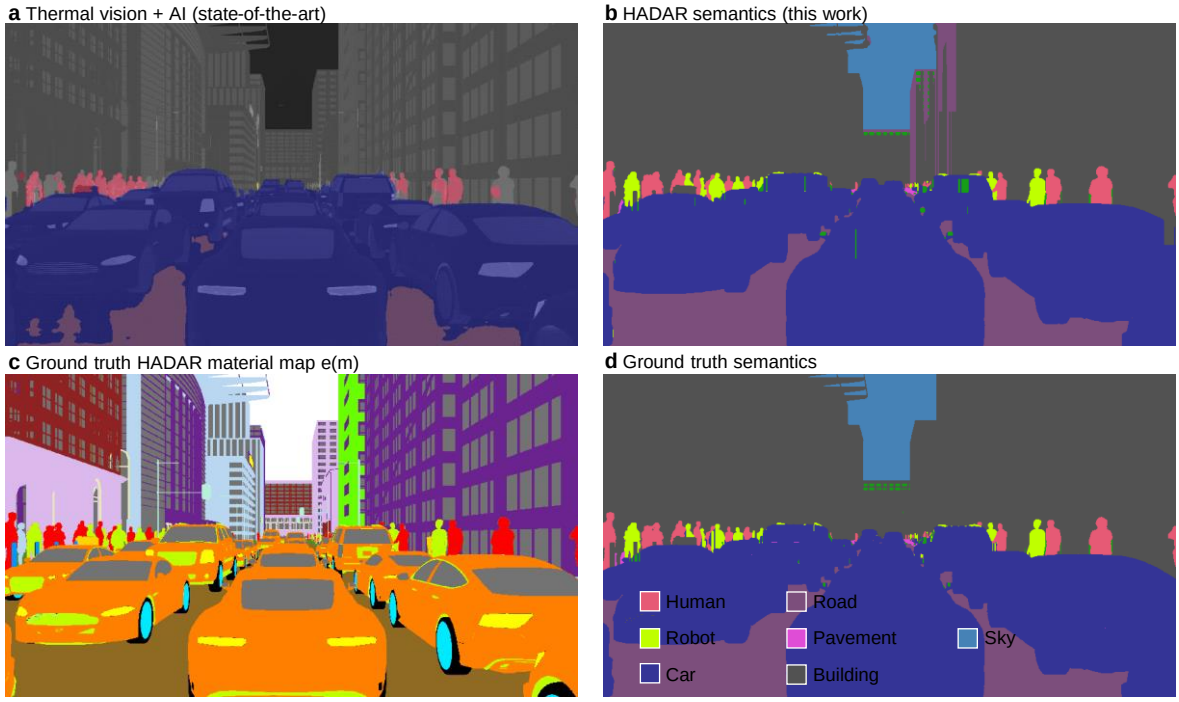
Extended Data Fig. 5. HADAR TeX vision overcomes the ghosting effect in traditional thermal vision and beats state-of-the-art approach to enhance visual contrast. This scene consists of multiple humans (dark red in TeX vision), robots (purple), cars and buildings at a summer night. Geometric textures of the road and pavements are vivid in TeX vision but invisible in raw thermal vision and poor in enhanced thermal vision. The mean texture density (standard deviation, see Sec. SIID of the Supple. Info. for more details) in TeX vision is 0.0788, about 4.6 folds more than the texture density of 0.0170 in the state-of-the-art enhanced thermal vision. This scene is the Street-Long-Animation in the HADAR database.



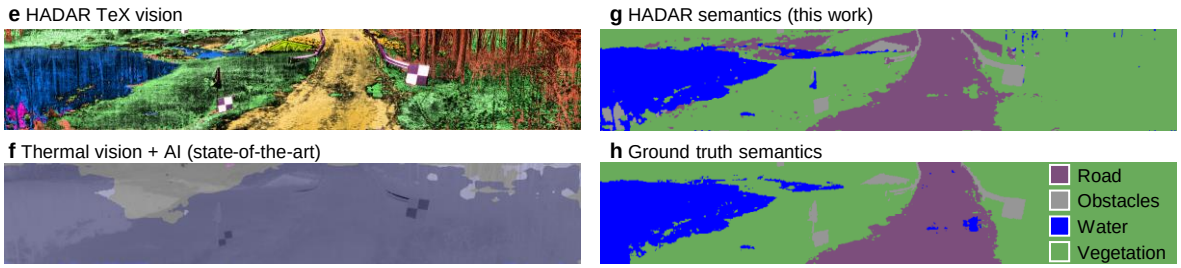
Extended Data Fig. 6. HADAR estimation theory for multi-material library. a, Sample incident spectra of 5 materials generated by Monte Carlo simulations. $T = 60\,\mathrm{C}^\circ, T_0 = 20\,\mathrm{C}^\circ$ and $V_0 = 0.5$. b, Minimum statistical distance of each material. Spectra of silica and paint have non-trivial features that are distinct with other materials in the library. Statistical distance larger than 1 (dashed line) consistently indicates that silica and paint are identifiable. Note that aluminum is similar to human skin under TeX degeneracy and non-identifiable, as discussed in Fig. 3, even though with the same temperature its spectrum is much weaker than human skin. Emissivity of human skin was approximated as a constant 0.95. Other emissivity profiles were drawn from NASA JPL ECOSTRESS spectral library. This figure intuitively shows that HADAR identifiability based on semantic/statistical distance is an effective figure of merit to describe identifiability. For more details in generalizing HADAR estimation theory to multiple materials, see Sec. SIIB of the Supple. Info.



Extended Data Fig. 7. HADAR detection (TeX vision + AI) beats widely used state-of-the-art thermal detection (conventional thermal vision + AI). a, Human body detection results based on thermal imaging. b, Human and robot identification results based on HADAR. Detection is performed by thermal-YOLO (YOLO-v5 fine tuned on the thermal automative dataset, https://github.com/MAli-Farooq/Thermal-YOLO-And-Model-Optimization-Using-TensorFlowLite), with detection score/confidence shown together with the bounding box. Due to TeX degeneracy and the ghosting effect, human body, robot (aluminum at 72.5 C°), and the car (paint at 37 C°) emit similar amount of thermal radiation, and hence the human body visually merges into the car in thermal imaging while the robot is mis-recognized as a human body. With our proposed TeX vision which captures intrinsic attributes, HADAR can distinguish them clearly and yield correct detection. Explicitly, we first extract the material regions corresponding to human (c) and robot (d), and then we perform people detection individually, and at last we combine detection results together to form the final HADAR detection (b). We observed that the above results showing the advantage of HADAR TeX vision vs. thermal vision is robust and independent of the AI algorithms. Standard computer vision toolbox (People detector in Matlab R2021b) also confirms the results. We also observed that HADAR detection is robust against wrong material predictions, even though few road pixels under the car and around the human leg are predicted as 'aluminum' in (b) and (d).

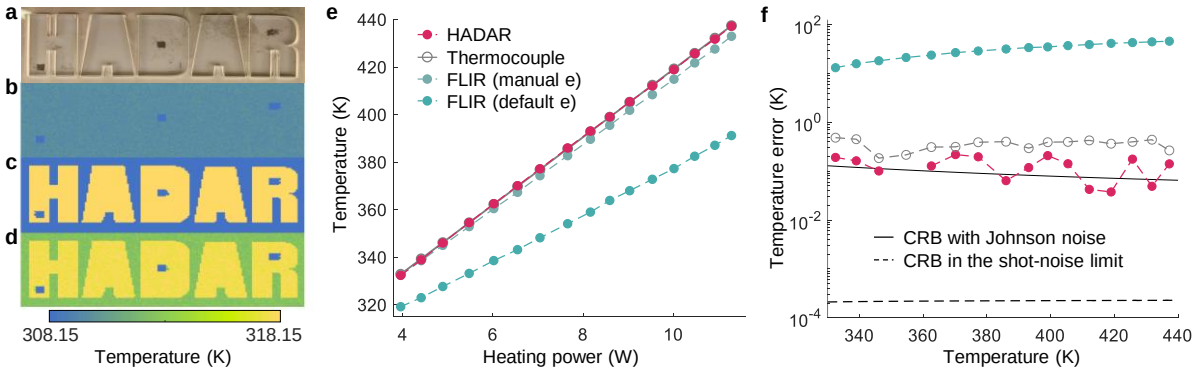


mIoU %	Human	Robot	Car	Road	Pavement	Building	Sky
Thermal vision + AI	33	0	90	67	25	69	16
TeX vision + non-AI	**94**	**84**	**98**	**92**	**87**	**84**	**90**

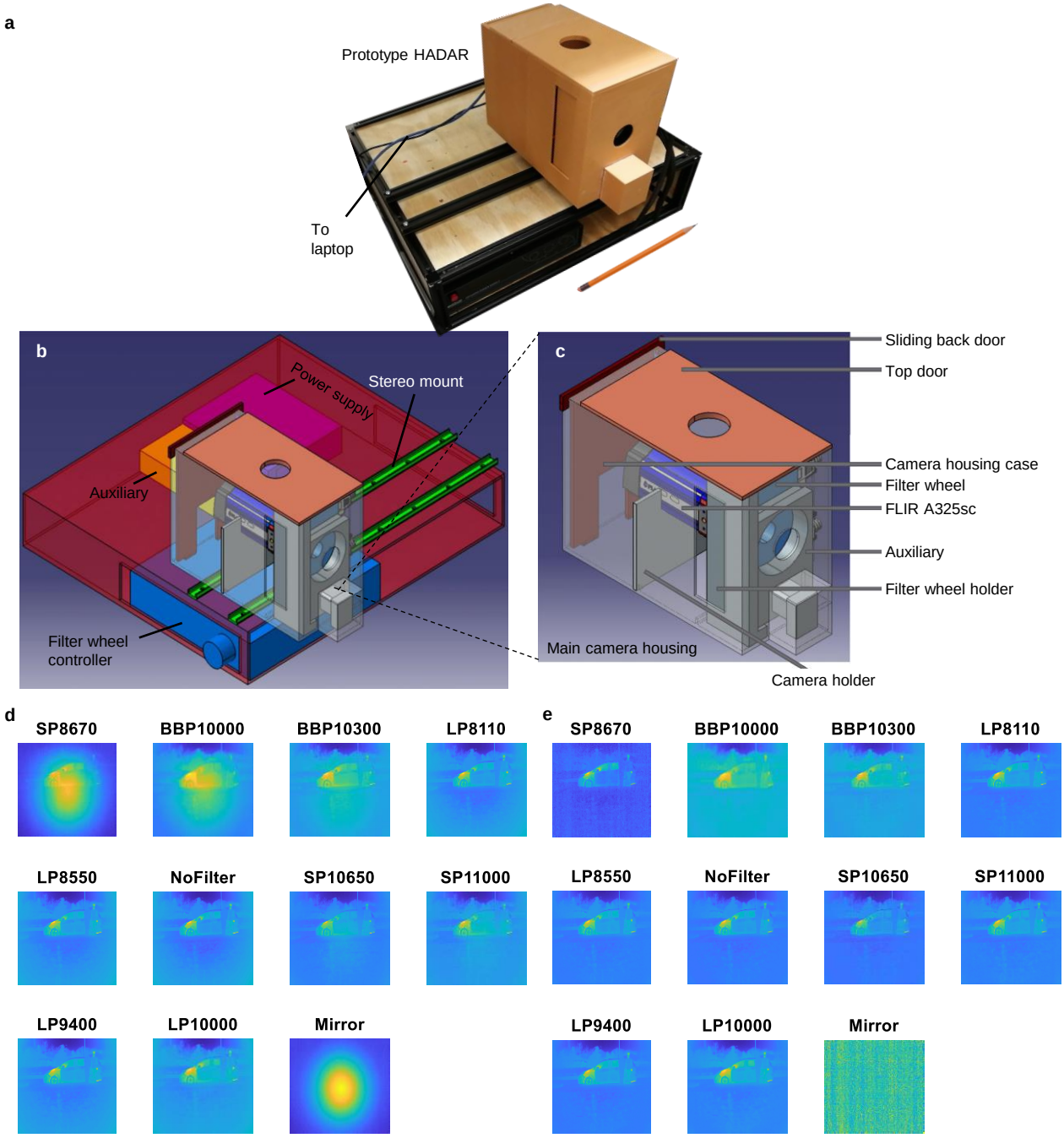


mIoU %	Road	Obstacles	Water	Vegetation
Thermal vision + AI	0.0	0.4	0	0
TeX vision + non-AI	**77.4**	**66.4**	**89.7**	**90.0**

Extended Data Fig. 8. HADAR physics-driven semantic segmentation beats state-of-the-art vision-driven semantic segmentation (thermal vision + AI). a, Thermal semantic segmentation with DANet (pre-trained on the Cityscapes dataset) [43]. b, HADAR semantic segmentation transformed from the material map in estimated TeX vision. c, Ground truth material map in the ground truth TeX vision. d, Semantic segmentation transformed from (c) to approximate the ground truth segmentation, see Sec.SIIIE of the Supple. Info. for more details of the non-machine-learning transformation. Statistics in the upper table were done on the first 4 on-road scenes in the HADAR database with 5-fold cross validation. (e-h) and the lower table show the typical performance comparison between HADAR vs. thermal semantics, where the off-road scene is beyond DANet's training set. We have also observed consistent results on other non-city scenes in the HADAR database (not shown). This real-world off-road scene is a general example to show the importance of material fingerprint in detection/segmentation. Since AI enhancement is only used in thermal semantics, the advantage of HADAR semantics is clearly from TeX vision with physical attributes. In the future, learning-based approaches to convert material map to semantic segmentation with the help of spatial information may further improve HADAR semantics. mIoU: Pixel-wise mean intersection over union. Ground truths of the real-world scene were manually annotated.



Extended Data Fig. 9. Unmanned HADAR thermography reaching the Cramér-Rao bound (CRB). By exploiting spectral information and automatically identifying the target, HADAR maximizes temperature accuracy beyond traditional methods. Demonstrated in a-d is a HADAR alphabet sample made of plastics at 312.15 K on an unpolished silicon wafer at 317.15 K. a, Optical image. b, Thermograph using FLIR A325sc shows camouflage and lack of information. c, HADAR material readout. d, HADAR temperature read-out. b and d share the same colorbar clearly demonstrating the HADAR advantage. Shown in e-f is the measurement of a uniform n-type SiC sample kept on a heating plate with varying heating power. e, Mean temperature readout shows HADAR is unbiased and beats commercial infrared thermograph. f, Root-Mean-Square-Error shows HADAR reaches the CRB for the given detector and imaging system. HADAR also beats commercial thermocouple in precision.



Extended Data Fig. 10. Prototype HADAR calibration and data collection. a, Experimental setup of our HADAR prototype-I. b-c, 3D schematics of our prototype HADAR. In our prototype HADAR, we used a sturdy stereo mount to take stereo heat-cube pairs. d, Raw HADAR data with detector's self radiation reflected by filters. e, HADAR signal calibrated with a uniform reference object, to remove detector's self radiation. For more details of calibration, see Sec.SIV of the Supple. Info.

Supplementary Information for 'Heat-Assisted Detection and Ranging'

Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N. Boddeti, and Zubin Jacob

CONTENTS

The proposed heat-assisted detection and ranging (HADAR) is a completely-passive remote sensing technique. HADAR retrieves targets' temperature (T), material emissivity (e), texture (X), and subsequently, semantics and distance, by collecting three-dimensional hyperspectral data cubes in the thermal infrared spectrum, $\mathcal{C}_i(x, y, q)$, which we call heat cubes. Here, $\mathcal{C}$ is the electronic output data (*e.g.*, current, voltage, or photon counts) of the detector, i indexes the heat cube from the i-th detector, q indexes the used filters (or the spectral bands), and x and y are pixel coordinates on the image plane of the detector. This supplementary information explains in detail what is recorded in heat cubes $\mathcal{C}$, the theory of thermal textures, the fundamental limits in estimating material (material classification, semantic segmentation) and distance (ranging), explicit inverse algorithms used in solving TeX from $\mathcal{C}$, hardware calibrations, and prototype HADAR experimental details.

SI. HEAT SIGNAL AND HEAT CUBES

Denoting S the thermal radiation (heat signal) of an object, this section shows the mathematical model of S and how it is recorded in heat cubes $\mathcal{C}$. Fig. S1 is a comprehensive summary of the explicit models of heat cubes with various hyper-spectral/multi-spectral (HS/MS) components and various detectors. Based on these diverse models, Fig. S2 shows the unified model of heat cubes with input heat signal S. The following subsections will explain heat signals, imaging systems and detectors separately, and then analyze thermal textures.

A. Thermal radiation of the target and the environment

The implemented HADAR in this paper is based on traditional imaging system consisting of focusing components like lens or parabolic mirrors, as shown in Fig. S3(a), but we emphasize that HADAR could also apply to lensless imaging and doesn't restrict itself in the ray-optics regime. In traditional imaging systems, the scene of interest is mapped to rasterized sensor pixel arrays. Each pixel focuses on a small object element, $\alpha = \alpha(x, y)$, with area A_α determined by the imaging resolution. The total heat flux leaving α along z direction, as shown in Fig. S3(b), has two additive contributions,

$$S_{\alpha\nu}(\tilde{z}) = e_{\alpha\nu}(\tilde{z})B_\nu(T_\alpha) + \int r_{\alpha\nu}(\tilde{z}, \tilde{\rho})\bar{V}_{\alpha\beta}S_{\beta\nu}(\tilde{\rho})\,\mathrm{d}A_\beta, \tag{S1}$$

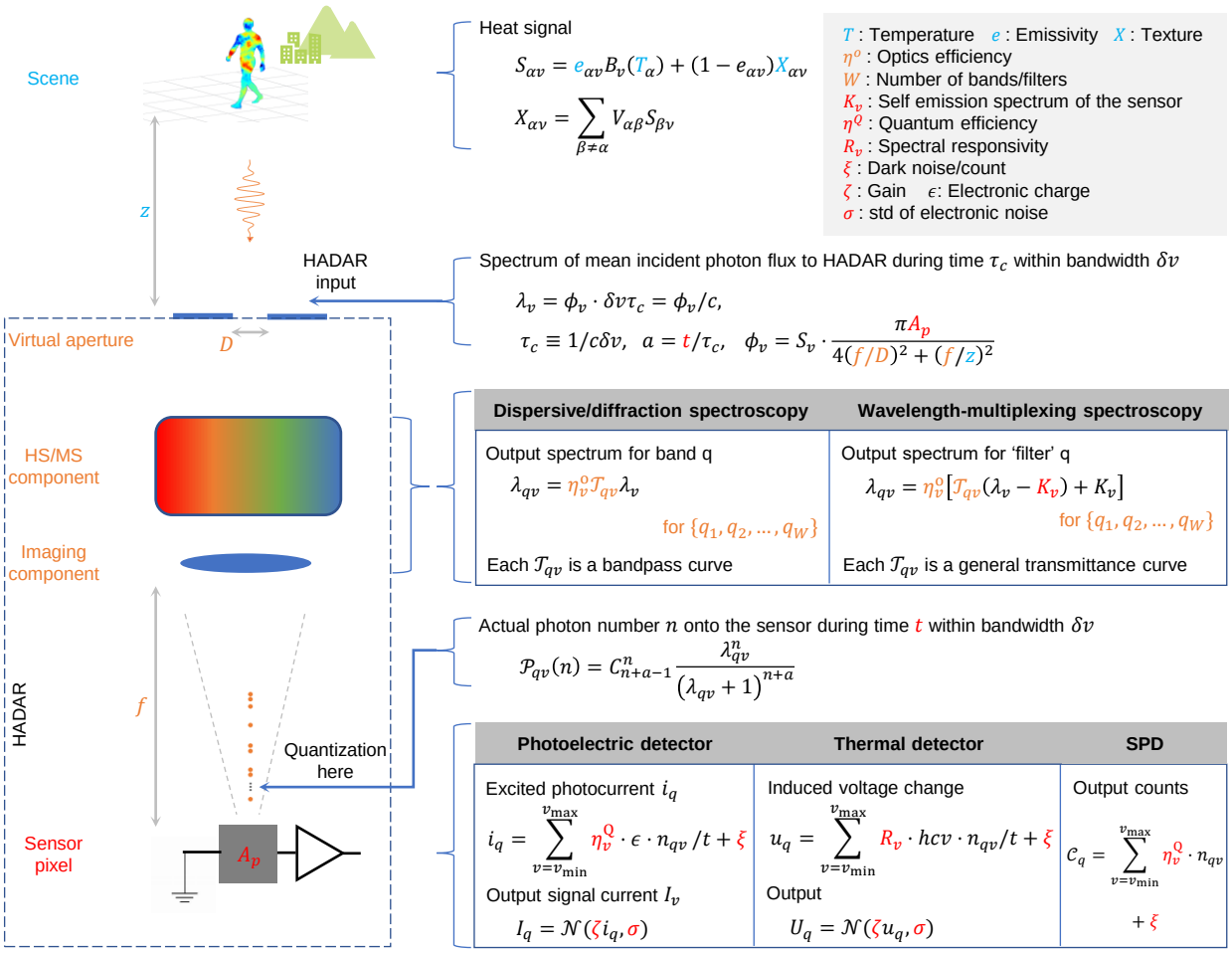


FIG. S1. Diverse models of heat cubes with various hyper-spectral/multi-spectral (HS/MS) components and detectors. Left side from top to bottom is the signal flowing from the scene through the aperture and optical components to the sensor. Right side gives the corresponding equations to model each process. Definition of parameters is given in the shaded box. A full list of parameters can be found in Fig. S2 and will be explained in the context. $\mathcal{N}$: normal distribution. C^n_{n+a-1} is the binomial coefficient. h: Planck's constant. SPD: single-photon detector.

where the first term is direct thermal emission from α, and the second term is the environmental emission from all other infinitesimal object elements β entering the detector after scattering from α. Here, $e_{\alpha\nu}$ is the spectral emissivity, a unique material signature. In principle, $e_{\alpha\nu}$ has angular dependence, determined by the local surface normal $\tilde{A}_\alpha$ and the observing direction $\tilde{z}$. $B_\nu(T_\alpha)$ is the blackbody radiation at temperature T_α, governed by Planck's law. $r_{\alpha\nu}(\tilde{z}, \tilde{\rho})$ is the reflectance distribution function with light passing from $-\tilde{\rho}$ to $\tilde{z}$ direction. $\bar{V}_{\alpha\beta} = \frac{F(-\tilde{\rho}\cdot\tilde{A}_\beta)F(\tilde{\rho}\cdot\tilde{A}_\alpha)}{\pi\rho^2}$ is the differential view factor from β to α satisfying

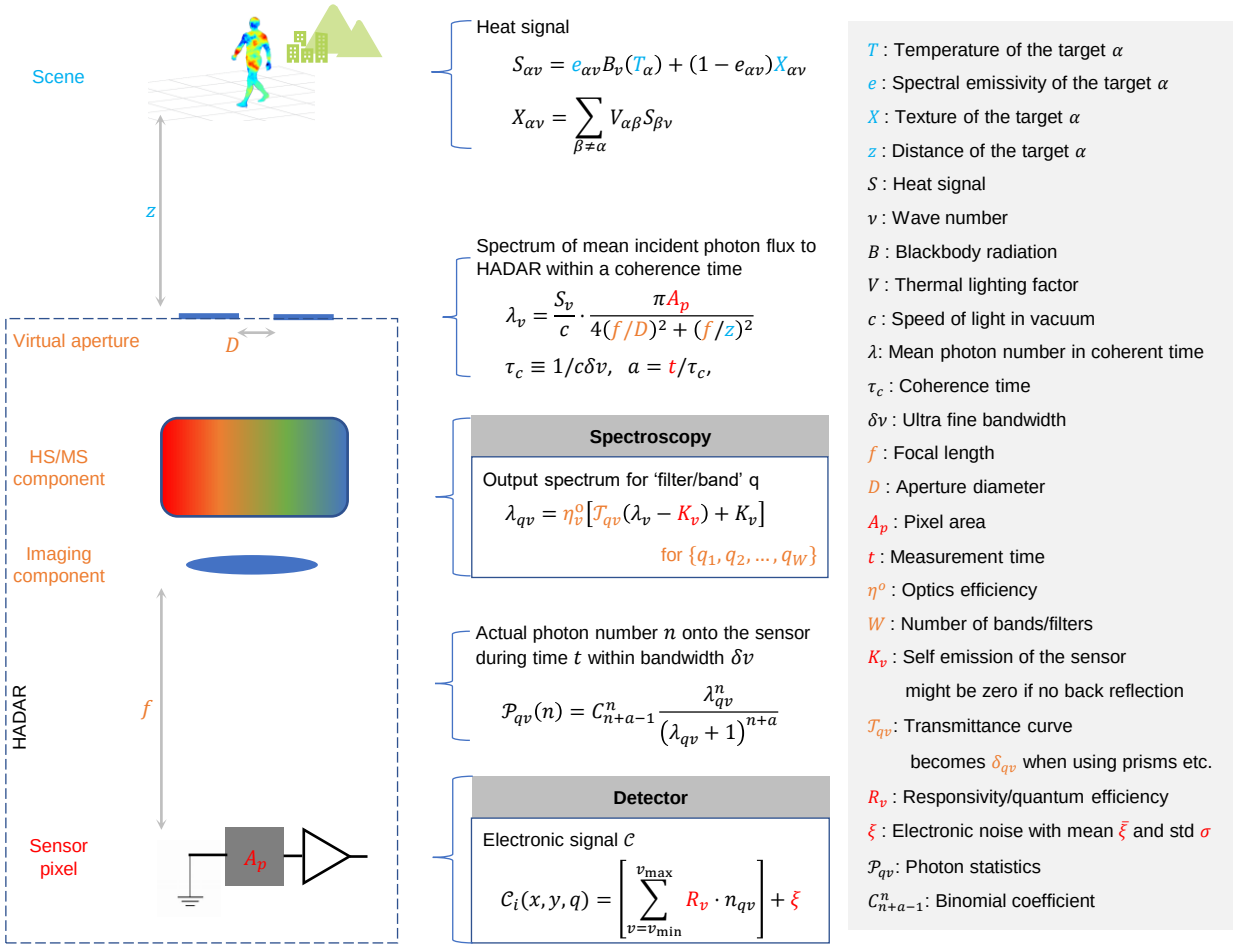


FIG. S2. The unified model of heat cubes $\mathcal{C}$ with input heat signal S. Left side from top to bottom is the signal flowing from the scene through the aperture and optical components to the sensor. Right side gives the corresponding equations to model each process.

$\int \bar{V}_{\alpha\beta}\,\mathrm{d}A_\beta = 1$. The integral $\int \cdot\,\mathrm{d}A_\beta$ is over all the emitting surfaces in the entire scene, and $F(x) \equiv \max(0, x)$. Eq. (S1) is consistent with the rendering equation in computer graphics to simulate real-world scenes.

For most natural objects with rough surfaces, the angular dependence of $e_{\alpha\nu}$ can be ignored, and the Lambertian (diffusive) reflectance applies, *i.e.*, $r_{\alpha\nu}$ is also angular independent. Furthermore, Kirchhoff's law implies that $r_{\alpha\nu} = 1 - e_{\alpha\nu}$, and hence the heat signal leaving α reduces to

$$S_{\alpha\nu} = e_{\alpha\nu}B_\nu(T_\alpha) + [1 - e_{\alpha\nu}]X_{\alpha\nu}, \tag{S2}$$

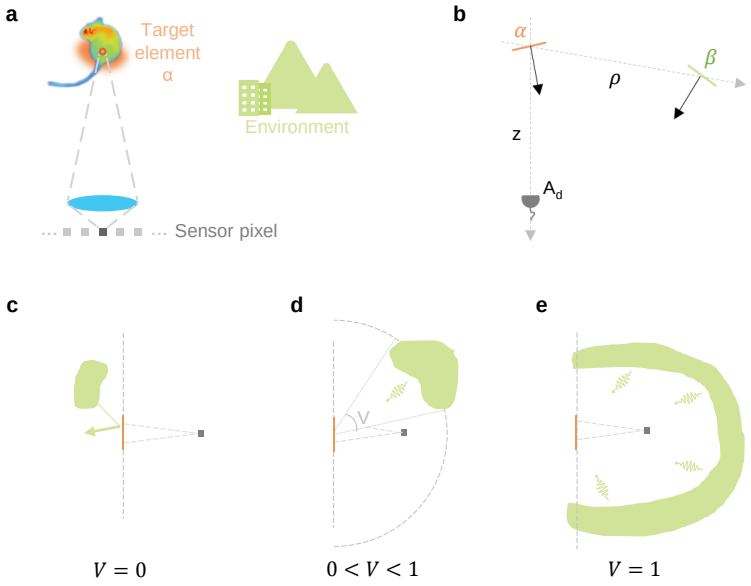


FIG. S3. a, Schematic of the HADAR imaging. One sensor pixel of the detector focuses on a small object element, and environment radiation can only be collected through scattering off the object element. b, Mathematical model of the scattering process. Target element of area A_α with emissivity e_α is at a distance of z away from the detector with aperture A_d. A small area of the environment A_β with emissivity e_β is at a distance of ρ away from the target. c-e, Illustrations of the thermal lighting factor V. c, $V = 0$, no emitting environment on the detector side, *i.e.*, no radiation can be scattered into the detector by the target. d, $0 < V < 1$, a fraction of environment emits on the detector side. e, $V = 1$, the target is fully surrounded by emitting environment on the detector side.

with

$$X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta} S_{\beta\nu}, \tag{S3}$$

and $V_{\alpha\beta} = \int_\beta \bar{V}_{\alpha\sigma}\, \mathrm{d}A_\sigma$. By integrating the differential view factor over uniform objects, the scattering term in Eq. (S1) simplifies to a summation over compact objects in Eq. (S3). Hereafter, we denote β a uniform and compact object with finite size in the scene. Now, the normalization condition becomes $\sum_\beta V_{\alpha\beta} = 1$. Conventionally, $V_{\alpha\beta}$ is called the view factor in radiative heat transfer. In our scenario, it depicts the fraction of thermal illumination from a finite object β that can be scattered by object α into the detector, and hence we call it the thermal lighting factor in the paper, to be more intuitive. Since V captures the local surface normal, X in Eq. (S3) carries the geometric surface texture of object α under

thermal illumination of its environment. X will be further discussed in Sec.SID.

Heat signal $S_{\alpha\nu}$ given by iterative equations (S2) and (S3) is an infinite series, consisting of multiple scattering contributions. To better illustrate the multiple scattering process and the concept of thermal lighting factor, we show a simpler example of only one uniform environmental object ($V < 1$, the rest of the environment is deep space). In this case,

$$S_{1\nu} = e_{1\nu}B_\nu(T_1) + [1 - e_{1\nu}]V_{10}e_{0\nu}B_\nu(T_0) + [1 - e_{1\nu}]V_{10}[1 - e_{0\nu}]V_{01}e_{1\nu}B_\nu(T_1) + \cdots , \quad \text{(S4)}$$

where

$$V_{10} = \frac{1}{\pi A_1}\frac{F(-\tilde{\rho}\cdot\vec{A_0})F(\tilde{\rho}\cdot\vec{A_1})}{\rho^2}, \quad \text{(S5)}$$

$$V_{01} = \frac{1}{\pi A_0}\frac{F(-\tilde{\rho}\cdot\vec{A_0})F(\tilde{\rho}\cdot\vec{A_1})}{\rho^2}. \quad \text{(S6)}$$

In Eq. (S4), the 0-th order term is the radiation directly from the target, the 1-st order term is the radiation emitted by the environment and scattered by the target, the 2-nd order term is the radiation emitted by the target, scattered by the environment back to the target again, and then scattered by the target, and so forth for higher order terms. Some examples of V_{10} are visualized in Fig. S3(c-e). Worth noting is that, in computer graphics, a truncation to multiple scattering at l-th order (*i.e.*, each light ray is bounced at most l times, $l = 4 \sim 8$) is sufficient to render high-quality and vivid movies. Besides, usually a few objects dominate in the scattering contribution, such as, sky, ground, and buildings. To give a quantitative intuition, we take an example of a pedestrian standing 5 m away from a car. The thermal lighting factor of the pedestrian on the car is $V \leq \frac{1}{25\pi} \approx 0.0127$, while the thermal lighting factor of the ground/sky on the car is about 0.5. With finite number of objects and bounces, it becomes possible to solve temperature T and thermal lighting factor V, and classify materials e_ν, from observed heat signal S_ν.

B. Hyperspectral imaging

In acquiring hyperspectral data cubes, imaging systems considered in this paper consist of wavefront modulation optical components for focusing (*e.g.*, lens or parabolic mirrors), and wavelength modulation optical components to retrieve spectral information (*e.g.*, diffraction gratings, dispersive prisms, filters or interferometers). In such systems as shown in Fig. S2,

the mean photon number of the heat signal input to HADAR, λ_ν, is connected to the heat flux per unit area per unit solid angle per wave number, S_ν, by the following relation,

$$\lambda_\nu = S_\nu \times \tau_c \delta\nu \times \frac{\pi A_\mathrm{p}}{4(f/D)^2 + (f/z)^2} = \frac{S_\nu}{c} \frac{\pi A_\mathrm{p}}{4(f/D)^2 + (f/z)^2}, \tag{S7}$$

where $\delta\nu$ is the bandwidth, $\tau_c \equiv 1/c\delta\nu$ is the coherence time, c is the speed of light in vacuum, f is the focal length, A_p is the pixel area, and the multiplication factor in the right-hand side of the above equation is the target area A_α times the integral of the solid angle of the aperture with respect to the target element. It gives the fraction of signal emitted by the target that is collected by the pixel. When the distance z is much larger than the aperture D, $z \gg D$, the f-number (f/D) dominates in the denominator, and the factor is proportional to $A_\alpha D^2/z^2$, according to the ray-optics geometry relation $A_\alpha/z^2 = A_\mathrm{p}/f^2$. In the opposite limit when $z \ll D$, the factor reduces to πA_α. For clarify, We have suppressed the subscript of α or (x, y) wherever there is no risk of confusion. When the measurement/integration time is t, $a \equiv t/\tau_c$, the mean photon number is simply given by $N_\nu \equiv a\lambda_\nu$.

When a set of filters is used for spectrum reconstruction, the mean photon number reaching the sensor is

$$\lambda_{q\nu} = \eta^\mathrm{o}_\nu \left[\mathcal{T}_{q\nu}\lambda_\nu + \mathcal{R}_{q\nu}K_\nu + \mathcal{E}_{q\nu}F_\nu\right], \tag{S8}$$

where η^o_ν is the optics efficiency accounting for the transmittance of optical components like lens. $\mathcal{T}_{q\nu}$ is the transmittance curve, $\mathcal{R}_{q\nu}$ is the reflectance curve, and $\mathcal{E}_{q\nu}$ is the absorptance curve of the q-th filter, satisfying $\mathcal{T}_{q\nu} + \mathcal{R}_{q\nu} + \mathcal{E}_{q\nu} = 1$. K_ν is the self-emission spectrum of the sensor, and F_ν is the self-emission spectrum of the filter. Under the assumption that either $\mathcal{E}_{q\nu} \to 0$ or $F_\nu \approx K_\nu$, the above equation can be simplified as

$$\lambda_{q\nu} = \eta^\mathrm{o}_\nu \left[\mathcal{T}_{q\nu}(\lambda_\nu - K_\nu) + K_\nu\right]. \tag{S9}$$

Eq. (S9) can be generalized to interferometer-based spectroscopy when the back reflection of the interferometer is significant. The fact that $\lambda_\nu - K_\nu$ is encoded by filters/interferometers in the experimental data makes the back reflection K play a role of 'negative' dark noise, as observed in Ref. [1] and also discussed here in Sec. SVI.

Parameters of imaging systems used in this paper are summarized in Tab. S1. Parameters of detectors are explained in the following subsection. Filter characterization is given in Sec. SIV

Imaging system	Optics efficiency η_ν^o	Focal length f (mm)	f-number (f/D)
iS50	90%	128.0	2.5
A325sc	99%	18.0	1.3

TABLE S1. Specifications for imaging systems used in this paper. iS50: Nicolet iS50 Fourier-transform infrared spectrometer. A325sc: FLIR A325sc thermal camera. D: diameter of the aperture.

C. Detectors and recorded signals

Actual recorded signal about the thermal radiation spectrum depends on specific detectors.

Existing infrared intensity detectors which respond to light field intensity are mainly based on thermal effect or photodiodes. The former absorbs thermal radiation and converts it into temperature change of the sensor, while the latter absorbs thermal radiation and generates electron-hole pairs. Their recorded signals, despite diverse readout mechanisms [*e.g.*, for thermal effect: temperature-dependent resistance (bolometers/microbolometers), thermal-electric effect (thermocouples or thermopiles), thermal expansion (Golay cells). For photodiodes: photovoltaic mode, photoconductive mode, avalanche mode] or sensor materials [*e.g.*, Vanadium oxide (VO_x), silicon, Mercury Cadmium Telluride (MCT), Indium Gallium Arsenide (InGaAs)], all suffer from inherent electronic noise (Johnson-Nyquist noise, Flicker noise, *etc.*). When considering signal-to-noise ratio (SNR), it is always convenient to map the electronic noise to the signal side and define noise-equivalent power (NEP). This is simply to rewrite $\sum_\nu R_\nu n_{q\nu} + \xi$ in Fig. S2 as $\bar{R}(n_q + \xi/\bar{R})$, where $n_q = \sum_\nu n_{q\nu}$ is the total photon number, $\bar{R} = (\sum_\nu R_\nu n_{q\nu})/n_q$ is the mean responsivity, and $\xi/\bar{R}$ is the noise-equivalent power in photon number. Hereafter, we denote $\xi/\bar{R}$ as ξ for simplicity, and the heat cube becomes

$$\mathcal{C}_q = \bar{R}(n_q + \xi). \tag{S10}$$

For comprehensive modelling, it's worth stressing that spectral quantum efficiency and readout/gain noise of the photodetector is usually packaged in NEP, or noise-equivalent temperature difference (NETD), or the special detectivity D^*. The response time Δt as well as the pixel size A_p shall be specified later. Within the scope of this paper, the actual value of

gain factor or responsitivity of the detector $\bar{R}$ is irrelevant, and hence we will focus on $n_q + \xi$ instead of $\mathcal{C}_q$ when talking about heat cubes hereafter. Nonlinearity is further assumed to be negligible.

Single photon detector is also being pushed to work in the thermal radiation spectrum of interest, typically $715 \sim 1250$ cm^{-1} (or equivalently $8 \sim 14$ μm). Superconducting nanowire single photon detector (SNSPD) is especially promising for long-wave infrared detection due to low Cooper-pair bond energy (∼meV). WSi-based SNSPD has been demonstrated to be able to operate at $1 \sim 7$ μm recently [2, 3], with high quantum efficiency (93%) and low dark count rate (< 1 cps). In the foreseeable future, single-photon level heat signal resolution for the desired heat spectrum is becoming possible.

As an ideal case, photon-number-resolving detectors (PNRDs), with perfect quantum efficiency and absolutely zero dark or electronic noise, record the exact photon number hitting the sensor, putting a fundamental shot-noise limit to the maximum amount of information one can retrieve from heat signal.

Here we discuss existing intensity detectors, in comparison with ideal detectors. For convenience, the following contents related to radiation power shall be discussed in terms of photon number instead of field intensity. All discussions would also apply to intensity detectors when photon number equivalent to the given intensity is considered. As a probabilistic problem, the actual photon number hitting the detector fluctuates around the mean photon number in Eq. (S9). Within a narrow band $\Delta\nu$ and the coherent time τ_c, thermal photon statistics is known to obey the Bose-Einstein distribution which suggests that the probability of registering n photons is given by [4]

$$\mathcal{P}_{q\nu}(n) = \frac{\lambda_{q\nu}^n}{(\lambda_{q\nu}+1)^{n+1}}. \tag{S11}$$

In common cases of HADAR applications, the spectral bandwidth is $\Delta\nu > 1$ cm^{-1} within the heat spectrum. The corresponding coherence time is hence below 33 ps. During an integration time t (whose minimum is determined by the response time Δt of the detector) longer than the coherence time, $t = a\tau_c$, the probability of registering n photons can be solved out as a typical *'n identical balls in a identical boxes'* problem, which yields a negative binomial distribution,

$$\mathcal{P}_{q\nu}(n) = C_{n+a-1}^n \frac{\lambda_{q\nu}^n}{(\lambda_{q\nu}+1)^{n+a}}, \tag{S12}$$

where $C^n_{n+a-1} \equiv \Gamma(n+a)/\Gamma(n+1)\Gamma(a)$ is the binomial coefficient. For intensity detectors, $a \gg 1$ and $N_{q\nu} \equiv a\lambda_{q\nu} \gg 1$ hold, and hence an asymptotic approximation follows if $\lambda_{q\nu} \ll 1$,

$$\mathcal{P}_{q\nu}(n) \approx \mathrm{Pois}(n; N_{q\nu}) \approx \mathcal{N}(n; N_{q\nu}, N_{q\nu}), \tag{S13}$$

where $\mathcal{N}$ is the Gaussian distribution. The probability of observing n_q with filter/band q is generally the convolution of $\mathcal{P}_{q\nu}$ for all wave numbers. Under the above asymptotic approximation, it can be given by

$$\mathcal{P}_q(n) = \mathcal{N}(n; N_q, N_q), \tag{S14}$$

where $N_q \equiv \sum_\nu N_{q\nu}$. Approximating the noise-equivalent photon number ξ by a Gaussian distribution of mean $\bar{\xi}$ and variance σ^2, $\mathcal{N}(\xi; \bar{\xi}, \sigma^2)$, one readily has the combined distribution for $n \leftarrow n_q + \xi$,

$$\bar{\mathcal{P}}_q(n) = \sum_{i=0}^{n} \mathcal{P}_q(i)\mathcal{N}(n-i; \bar{\xi}, \sigma^2) = \mathcal{N}(n; N_q + \bar{\xi}, N_q + \sigma^2). \tag{S15}$$

Actually, Johnson-Nyquist noise and Flicker noise are known to obey Gaussian distribution at one time instant. In general, the variance is their sum $\sigma^2 = \sigma_\mathrm{J}^2 + \sigma_\mathrm{F}^2$. White or pink noise spectrum indicates how noise evolves in time, and this is taken into account by the following scaling law. Define $N \equiv \sum_\nu N_\nu \equiv \sum_\nu a\lambda_\nu$ as the total input photon number through the aperture of HADAR. For the signal itself, $N \propto a \propto t$. For Johnson-Nyquist noise, $\sigma_\mathrm{J} \propto \sqrt{t}$, while for the Flicker noise, $\sigma_\mathrm{F} \propto t$. We define $\gamma \equiv \sigma^2/N$ as the ratio of the electronic noise power to the shot-noise power (normalized electronic noise power). It follows that $\gamma = \gamma_0 + \gamma_1 N$, where $\gamma_0 = \sigma_\mathrm{J}^2/N$ and $\gamma_1 = \sigma_\mathrm{F}^2/N^2$ are constants independent of time. In comparison, SNR is defined as N_q/σ. One can readily verify that it's possible to improve SNR by long-time integration for Johnson-Nyquist noise, while it's not the case for the Flicker noise. After all, the joint probability for a measured spectrum, $\mathbf{n} \equiv [n(q_1), n(q_2), \cdots]$, to be occurring is

$$\mathcal{P}(\mathbf{n}) = \prod_q \bar{\mathcal{P}}_q(n). \tag{S16}$$

Specific parameters for mentioned detectors in this paper are given in Tab. S2. Since quantum efficiency for state-of-the-art intensity detectors can easily reach over 50% [5] and is not the main restriction of HADAR, we assumed all quantum efficiency to be unity. The mean and std of the electronic noise are evaluated at the input of the detector, in terms

Detector	D^*(cm·$\sqrt{\text{Hz}}$/W)	NETD (mK)	$\sqrt{A_\text{p}}$ (μm)	Δt (μs)	$\bar{\xi}$ (count)	σ (count)	γ
MCT	4.70e10	/	1000	2.86	-1.46e8	4.53e4	/
μ-bolometer	/	47.80	12	12000	1.15e9	1.77e6	6.68e5
PNRD	/	/	/	/	0	0	0

TABLE S2. Detector specifications. MCT: liquid-Nitrogen cooled MCT detector for Nicolet iS50 Fourier-transform infrared spectrometer. See Ref. [1] for negative dark noise $\bar{\xi}$. μ-bolometer: uncooled microbolometer detector for FLIR A325sc thermal camera. PNRD: photon-number-resolving detector. D^*: special detectivity. NETD: noise-equivalent temperature difference. $\sqrt{A_\text{p}}$: pixel size. Δt: time constant. σ: std of noise-equivalent photon number. γ: the electronic noise power normalized by the shot-noise power. γ of the FLIR camera is modelled as if it is measuring the radiance per band for each of 536 bands in the $715 \sim 1250\,\text{cm}^{-1}$ spectral range.

of noise-equivalent photon number within the heat spectrum $715 \sim 1250$ cm^{-1}. The FLIR camera is evaluated at 30 C°, and we use $\gamma_0 = \gamma_1 N = \gamma/2$ at the given time constant, as approximated from experimental characterization. The FTIR spectroscopy is evaluated as a whole 'device' since the MCT detector is integrated inside. Accordingly, noise parameters of FTIR are evaluated as if FTIR was measuring spectrum band by band. In this sense, the back reflection of the sensor's emission K_ν in Eq. (S9) plays the role of dark noise, $\bar{\xi} = -\sum_\nu K_\nu$, which becomes negative. Experimental characterization reveals that the electronic noise of FTIR is purely Johnson-Nyquist noise. When compared with realistic detectors, corresponding ideal detectors take $\bar{\xi} = 0$, and $\sigma = 0$, with all other parameters matched.

D. Thermal textures and the TeX vision

For a well-calibrated detector with uniform responsivity/gain and dark noise across the pixel array, the spatial variation in its recorded image $\mathcal{C}(x, y)$ is, on average, from the

variation of S_α with respect to α. According to Eq. (S2), we have

$$\begin{aligned}\delta S_{\alpha\nu} = &\delta T_\alpha \cdot e_{\alpha\nu}\partial_T B_\nu \\ &+ \delta e_{\alpha\nu} \cdot [B_\nu(T_\alpha) - X_{\alpha\nu}] \\ &+ \delta X_{\alpha\nu} \cdot (1 - e_{\alpha\nu}).\end{aligned} \tag{S17}$$

The above equation explains the origin of thermal textures (the spatial variation in heat signals). Explicitly, there are 3 types of thermal textures. The first line in Eq. (S17) is from the temperature contrast, called the T-type thermal texture. The second line is from non-uniform material, called the e-type thermal texture. The third line carrying local surface normals of the target, as mentioned in Sec. SIA, is called the X-type thermal texture. We emphasize that T-type and e-type thermal textures might be significant near boundaries of objects where temperature jump and material change occur, but they are usually weak within uniform objects. Within uniform objects, it is the X-type thermal texture that captures the conventional geometric texture and visual details of the object. These origins of textures also apply to optical imaging under solar illumination, where the T-type texture is negligible and the e-type texture is caused by reflectivity contrast. In the visible-light spectrum, the surface texture on a light bulb is a good example to understand the X-type texture. Applying Eq. (S2) to imaging the surface of a bulb (uniform material, the bulb is switched off, as shown in Fig.2 in the main text) under solar illumination, we know that the direct emission (first term) is negligible and X is what is recorded in the image showing the geometric texture of the bulb surface. Another common example is imaging a lawn in daylight. Again, it is the X-type texture that gives the visual details of the lawn. Since it is important but usually ignored in thermal imaging, we call the 'X-type thermal texture' the 'texture' in the main text for short.

Generally, what is captured by detectors is a mixture of all three types of thermal textures. In traditional thermal imaging, T-type thermal texture dominates, and e-type and X-type textures are weak since $e_{\alpha\nu} \approx 1$. This is why thermal imaging is widely taken as the temperature contrast. T-type thermal texture could give the contour of objects that have different temperature with the background but is poor in details, exhibiting the 'ghosting effect'. In optical imaging under solar illumination, only the scattering term $(1 - e_{\alpha\nu})X_{\alpha\nu}$ in Eq. (S2) is recorded. Existing object detection, semantic segmentation and stereo depth estimation based on optical images, make use of the e-type and X-type textures. Therefore, to

overcome the 'ghosting effect' in thermal imaging and implement HADAR with comparable performance to optical detection and ranging, the key is recovering and enhancing e-type and X-type textures in the scattering term $(1-e_{\alpha\nu})X_{\alpha\nu}$, with the help of spectral resolution. However, we remind that materials' response to light (spectral features of emissivity) in the visible-light spectrum is different with that in the thermal infrared spectrum. Furthermore, the wavelength of the thermal infrared is around one order in magnitude larger than that of the visible light. They also result in different textures in thermal images and optical images.

We further discuss the structure of X in two different aspects. (1), According to Eq. (S3), we have, $\delta X = \Sigma\delta V \cdot S + \Sigma V \cdot \delta S$. The first term depending on local surface normals is the geometric texture. The second term depending on different environmental illuminations is commonly seen in mirror images formed on a flat surface. (2), X contains the scattered signal originated from all environmental objects. However, in daily experience, people are familiar with the part of scattered signal, denoted as $\bar{X}$, that is originated only from the sky illumination. The process to reconstruct $\bar{X}$ from X is shown in Extended Data Fig. 1b. In the simple example of imaging a light bulb mentioned above, where there is no other environmental objects than the sky, $\bar{X}$ is exactly the same as X. For the simple car-pedestrian scene in Fig. 4 in the main text, the synthesized ground has a very limited size and the sky illumination is dominant in most part of the image, and hence $\bar{X}$ is close to X. In this paper, we will interchangeably use $\bar{X}$ and X whenever possible, both implying the distilled texture $\bar{X}$ that people are familiar with.

Eq. (S2) and Eq. (S17) suggest a natural and physics-based TeX representation of the heat signal, which we call TeX vision. Instead of the grayscale thermal vision representing the integral $\int S_\nu \, \mathrm{d}\nu$, or the multi-band representation showing three most significant components (three spectral bands with least mutual information, or three principal components) of S_ν in RGB channels, TeX vision uses the HSV color format with mapping $\mathrm{H} = e, \mathrm{S} = T$ and $\mathrm{V} = X$, and captures the full physics information in heat signal. TeX vision presents different materials with different color hues, helpful for semantic segmentations. Furthermore, TeX vision encodes e-type and X-type textures in separate channels with temperature contrast, allowing for visually-enhanced textures. It combines the textures in the scattering term (used in optical imaging) and the temperature contrast (used in traditional thermal imaging), providing a better application potential. Sec. SIIIC explicitly shows how a TeX vision image is formed.

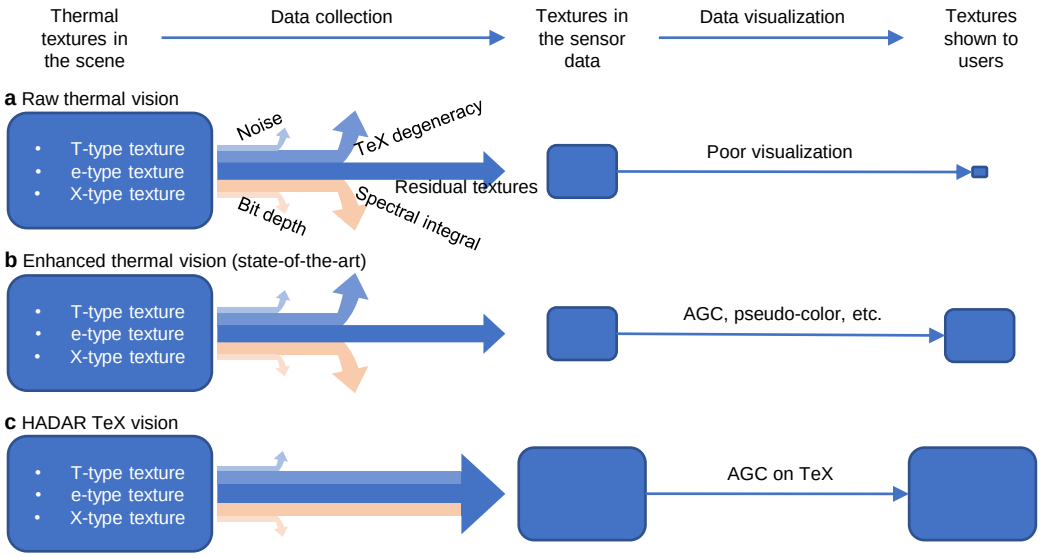


FIG. S4. Schematic diagram of texture flow from the scene to users. State-of-the-art thermal vision enhances visual contrast by improving data visualization. In comparison, HADAR recovers textures in the sensor data. As post-processing like FLIR AGC (automatic gain control, a variant of the histogram equalization algorithm) cannot increase information, HADAR improving data collection has better fundamental potentials than thermal imaging.

Now we explain four common channels where thermal textures can be lost. (1), TeX degeneracy leads to the same amount of radiance S even with different T/e/X value configurations, as shown in Methods – TeX degeneracy; (2), Spectral integral $\int S_\nu \, \mathrm{d}\nu$ loses spectral resolution and textures; (3), Weak textures will be buried by detector noise; and (4), Weak textures will be further lost in finite bit depths of the detector's readout circuit and numeric round off error. One extra aspect that affects visual contrast is related to the visual response and acuity of human eyes to different colors and intensities, or for artificial intelligence, the different sensitivity to different colors/intensities determined by particular training datasets. The texture flow in Fig. S4 summarizes the theory of thermal textures. Fig. S5 further illustrates the difference of our HADAR approach against traditional approaches in texture recovery. As an illustration, Fig. 4b in the main text shows the scattering signal at one single wavenumber full of textures, while Fig. 4a shows thermal imaging losing texture through channel (1).

With decluttered physical quantities, HADAR TeX vision provides a more promising platform and more insightful data to artificial intelligence (AI) algorithms for detection and

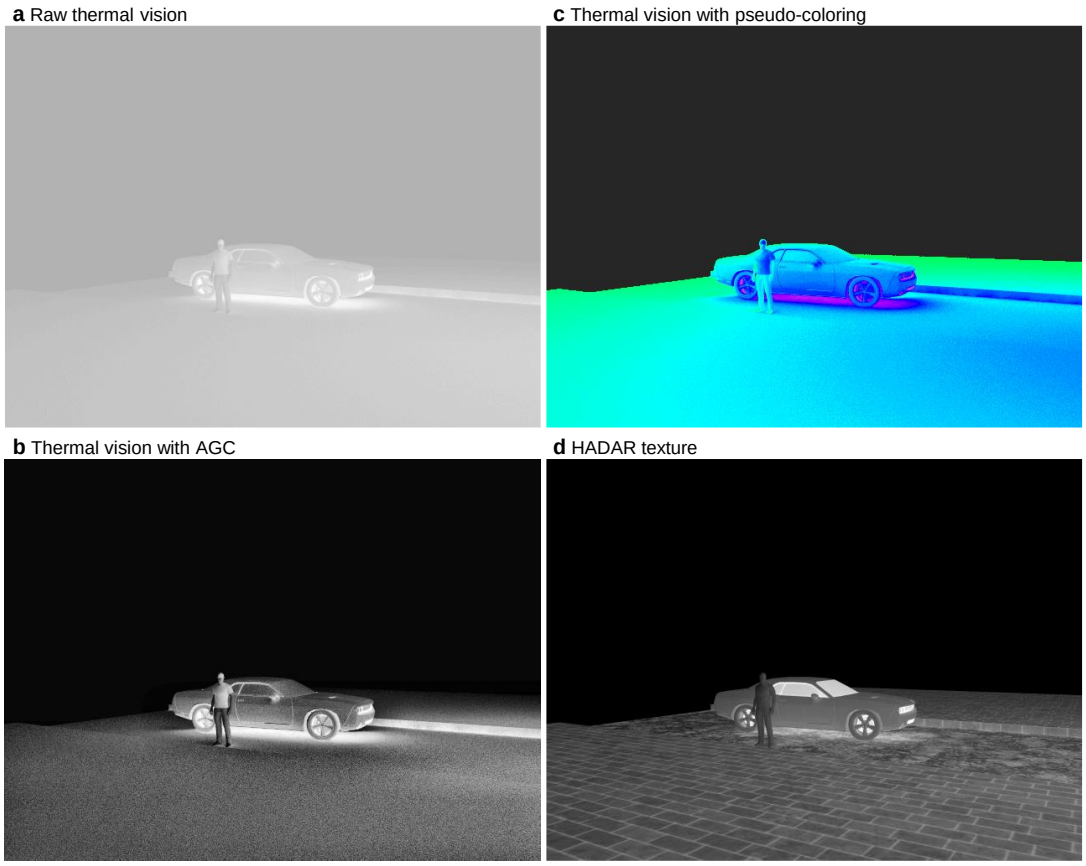


FIG. S5. HADAR texture beats state-of-the-art approaches to improve visual contrast. State-of-the-art approaches like FLIR AGC (automatic gain control, a variant of histogram equalization algorithm) or pseudo coloring are nonlinear mapping of the raw data to grayscale or RGB data, to better visualize the signal variation existed in data. They are post processing and cannot add textures/information to data, well known as the Data Processing Inequality [6] — 'post-processing cannot increase information'. a-c, When texture is already lost in sensor data, state-of-the-art approaches (post processing) cannot recover the texture. d, However, HADAR collects hyperspectral data and has more texture/information in the sensor data, enabling recovery of otherwise inaccessible textures.

ranging. The following overview summarizes the potential advantages of HADAR and TeX vision against state-of-the-art AI-enhanced thermal sensing.

Thermal sensing + AI (dataset + AI algorithm)	Object detection		Ranging	
	based on temperature†	based on material	based on temperature†	based on texture
ASL-TID + LatentSVM [25] / YOLO [26]	✔	✕	✕	✕
BU-TIV + SDD-Net [31]	✔	✕	✕	✕
CVC-FIR + YOLO [26] / LinSVM [27,28]	✔	✕	✕	✕
KAIST + YOLO [26] / AdaBoost [29]	✔	✕	✕	✕
LSI + CNN [30]	✔	✕	✕	✕
LITIV [32] + YOLO [26]	✔	✕	✕	✕
OSU-T + YOLO [26] / AdaBoost [33] / SVM [38]	✔	✕	✕	✕
Terravic + YOLO [26]	✔	✕	✕	✕
VOT-TIR [34] + Multiple AI therein	✔	✕	✕	✕
TIDPD + YOLO [35]	✔	✕	✕	✕
FLIR-ADAS + CNN [36]	✔	✕	✕	✕
Tetra + SVM [39]	✔	✕	✕	✕
Ref. [37] + SVM	✔	✕	✕	✕
Ref. [40] + SVM	✔	✕	✕	✕
CATS + CNN [41]	✔	✕	poor	✕
HADAR (TeX + AI)	✔	✔	✔	✔

TABLE S3. HADAR provides intrinsic physical attributes and enhanced textures enabling comprehensive understanding of the scene beyond AI-enhanced conventional thermal sensing. State-of-the-art thermal imaging, even equipped with AI, is limited solely to night vision enhancement without specificity or ranging accuracy. Identification based on material fingerprint is impossible for thermal imaging due to TeX degeneracy, and thermal ranging has poor performance due to the ghosting effect. HADAR with TeX vision can lead to improved AI performance, including identification based on material fingerprint (Figs. 5, Extended Data Figs. 7-9), and enhanced ranging based on recovered textures (Figs. 4 and 6, Fig. S19). †: Precisely, it's radiance contrast in thermal imaging. SVM: Support Vector Machine; CNN: Convolutional Neural Network; YOLO: You-Only-Look-Once.

SII. HADAR ESTIMATION THEORY I: FUNDAMENTAL LIMITS

With the theoretical model of heat signal given in the last section, here we focus on fundamental limits of HADAR, mainly regarding material and depth resolution in detection and ranging problems. For more details about the accuracy of temperature estimation, we refer the readers to Ref. [7].

Recall that the heat signal leaving object α is $S_{\alpha\nu} = e_{\alpha\nu}B_\nu(T_\alpha) + [1 - e_{\alpha\nu}]X_{\alpha\nu}$, with $X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta}S_{\beta\nu}$. Starting with T_α, $e_{\alpha\nu}$, and $V_{\alpha\beta}$ for all compact and finite objects, Monte Carlo path tracing can solve $S_{\alpha\nu}$ asymptotically with the l-th order scattering-cutoff solution $\tilde{S}^l_{\alpha\nu}$. The residual error $\delta_{\alpha l} \equiv |\tilde{S}^l_{\alpha\nu} - S_{\alpha\nu}| \to 0$ when l increases. In the inverse problem starting with $S_{\alpha\nu}$ for a few objects in a limited view of the entire scene, there are infinite solutions due to the TeX degeneracy given in Methods Section – TeX degeneracy. To break the TeX degeneracy, we assume that the spectral emissivities of objects in the scene are standard and can be characterized by a material library, $\mathcal{M} = \{e_\nu(m)|m = 1, 2, \cdots, M\}$. In this way, $e_{\alpha\nu}$ is discretized into $e_\nu(m_\alpha)$ and the dimension of solution space has been reduced significantly. This opportunity of dimensional reduction is available naturally in smart applications where materials usually have industrial standards [8]. Moreover, usually there are only a few giant and uniform objects (like sky, ground and buildings) that have significant scattering contributions in $X_{\alpha\nu}$, since the thermal lighting factor V for most other details (tiny objects or non-uniformity) is negligibly small, as explained in Sec. SIA. Let k denote the maximum number of significant environmental objects considered in the scene, whose spectral emissivity must be one out of M curves in the library. Now, the parameter set $\{klM\}$ determines the complexity of the inverse problem and also controls the accuracy of the solution of $T_\alpha, e_\nu(m_\alpha)$ and $X_{\alpha\nu}$. In the limit of $k, l, M \to \infty$, the potential TeX solution will converge to the ground truth, but the inverse problem will be extremely complicated to solve. Traditional thermal imaging for autonomous navigation is in the opposite limit, $l = k = 0$. As demonstrated in proof-of-concept experiments in this paper, we have moved forward the first step to make k and l finite. The raw solution of TeX with small $\{klM\}$ values already shows the advantage of TeX vision against traditional thermal vision and the advantage of HADAR against sonar, radar, LiDAR and cameras. When considering the fundamental limit of material estimation, we demonstrate $k = l = 1$ in this paper ($k = 1$ means one environmental object plus the deep space of zero radiation).

When processing HADAR prototype-1 experimental data, we have used $k = 2$ and $l = 1$ ($k = 2$ means two environmental objects without deep space). When processing HADAR prototype-2 experimental data and our synthetic HADAR database, we have used $k = 2$ and $l = \infty$, as shall be shown in Sec. SIII. The heat signal with $k = l = 1$ becomes

$$S_{\alpha\nu} = e_{\alpha\nu} B_\nu(T_\alpha) + [1 - e_{\alpha\nu}] V_\alpha B_\nu(T_0), \tag{S18}$$

where ambient temperature T_0 is assumed to be known (easy to measure).

A. Material estimation — two-material library

This subsection is devoted to addressing the question of how many photons are needed to identify the target material. As opposed to continuous parameters like temperature whose fundamental accuracy can be described by the variance of their estimator and given by the Cramér-Rao bound, material in a library is a categorical/ordinal parameter, upon which no variance can be defined. In fact, even if ordinal parameter is treated as a continuous parameter with discretization constraint, constrained Cramér-Rao bound [9] will yield a trivial bound (Var ≥ 0), and this implies one can never get an unbiased estimator for ordinal parameters. To address the accuracy of categorical/ordinal parameter estimation, we first start with the exact detection probability — the probability of predicting m when the ground truth material index is m (also known as the recall, or true positive rate; here we adopt the terminology as in Ref. [10]). The exact detection probability is analytically intractable, but we provide the numeric algorithm for it. We then develop an effective analytical theory for detection probability based on the Cramér-Rao bound of an auxiliary parameter. At last, we derive the Shannon information about the material that we can retrieve from experimental data. In the following contents, we assume the spectral resolution is obtained from dispersive prisms or diffraction gratings with fine bandwidth, for simplicity. Therefore, subscript q is identical to ν in Eq. (S16). However, we emphasize that the theory applies to filter-based experiments as well.

Start with hyperspectral imaging (HSI) where a library of reflectance spectra for potential materials, $\mathcal{R} = \{r_\nu^m | m = 1, 2\}$, is available. In the W-dimensional Hilbert hyperspace $\mathbb{R}^W$ where each axis represents the reflection signal at certain wave number, $r(\nu_w), w = 1, 2, \cdots, W$, the library $\mathcal{R}$ is a set of two isolated dots, as shown in Fig. S6a.

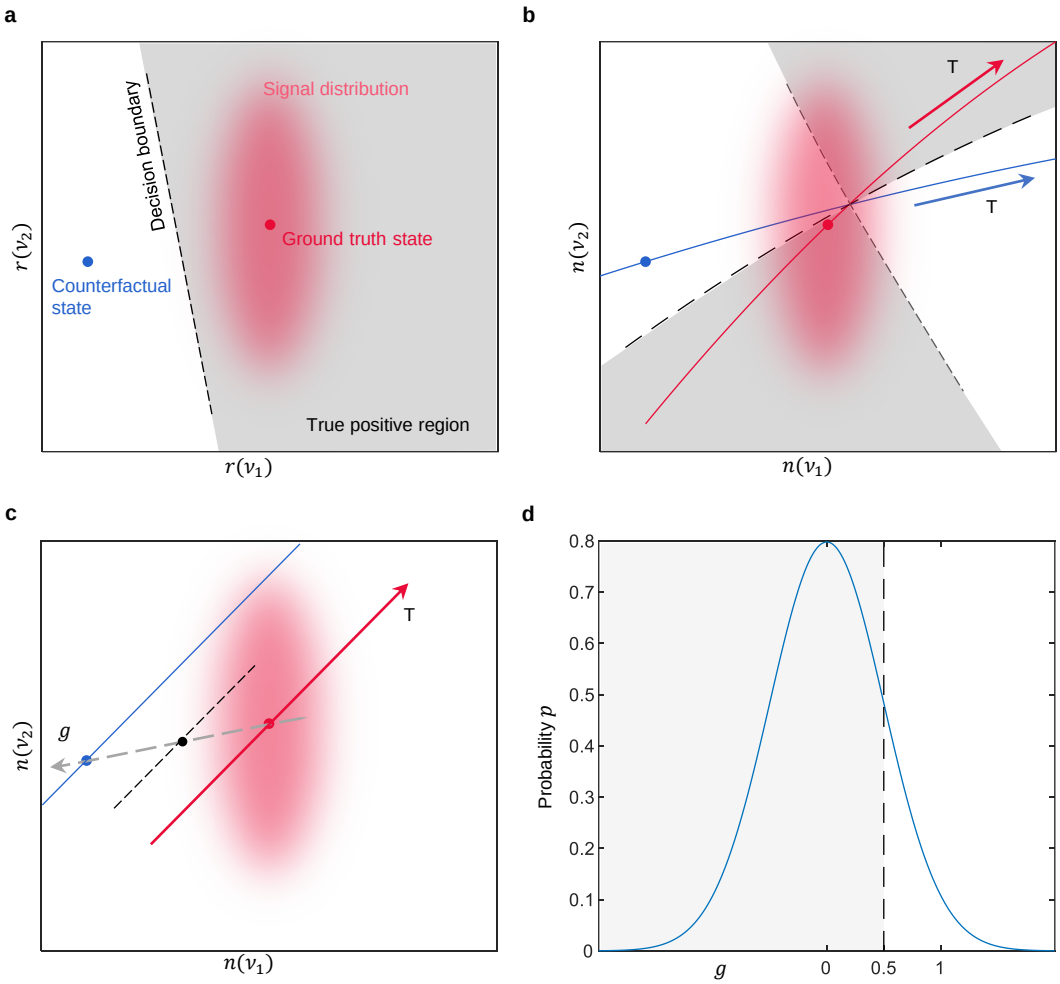


FIG. S6. Hyperspace representation of (a) HSI data and material library and (b-c) HADAR heat signal and material sheets. 2 out of W dimensions are shown for clarity. One material corresponds to one unique dot for HSI, but evolves to the V-T surface (V not shown for clarity) for HADAR. Red dot/curve: ground truth of the target material. Blue dot/curve: counterfactual target material. Dashed curve: the equiprobable hypersurface (decision boundary) in which each point has the same occurring probability for both materials. Red shaded spot represents probability distribution of experimental data due to noise. Gray dashed arrow represents the auxiliary g direction. (d) Probability distribution of signal along g direction.

The difference between two materials is characterized by the Euclidean distance of those two dots [11, 12]. For a given target state (red dot), the recorded experimental data point may distribute around with probability $\mathcal{P}$ given by Eq. (S16) [illustrated as the red shaded spot]. The probability of getting a correct material estimation (detection probability), based on

experimental data point, can be given by the integral of the probability distribution of the signal over the gray shaded area (true positive region). The boundary of the gray shaded area is the equiprobable hypersurface (decision boundary, dashed line) in which each point has the same occurring probability $\mathcal{P}$ for two possible states of the target. Put another way, either the target is the red dot or the blue dot, the probability to observe a data point in the hypersurface is the same. We emphasize that the equiprobable hypersurface might not be identical to the mid-perpendicular (equidistant) plane, since the probability distribution of signal is asymmetric along different axes.

Because temperature T and thermal lighting factor V are both unknown in HADAR, one cannot directly get spectral emissivity curve out of $S_{\alpha\nu}$ in Eq. (S2). Instead, we have to move to the W-dimensional Hilbert hyperspace $\mathbb{R}^W$ where each axis represents the heat signal $n(\nu_w), w = 1, 2, \cdots, W$, as shown in Fig. S6b. Here, materials at fixed V and T are isolated dots, but each material corresponds to one sheet of V-T surface given by Eq. (S18), and generally different material surfaces might intersect. Since each possible target state $\{mTV\}$ will have a different probability to produce a given signal point, now we have a maximum probability for each material to produce the given signal, where maximization is over T and V. We can define equal-maximum-probability hypersurfaces (dashed curves) accordingly, in which each point has the same maximum observation probability for both possible material surfaces. For a given target state (red dot), the detection probability P, now has to be integrated over irregular domains (gray shaded area). Assume, in a sufficiently wide domain covering the red shaded spot, $n(\nu)$ is meshed into G grids, $\{n_1(\nu), n_2(\nu), \cdots, n_G(\nu)\}$, and outside the domain, the data probability distribution is negligible. It follows that the total lattice sites of the hyperspace is G^W, and the V-T surface can also be discretized into a subset of lattice sites, whose size is proportional to G^2. The following algorithm 1 solves P, with computational complexity being MWG^{W+2}. While the size M of the library increases, W should increase accordingly to capture spectral features of all materials.

In light of the exponential computational complexity of the above algorithm, we seek an effective analytical theory for the detection probability. To that end, it's helpful to point out the following observation. Simplify the problem, by setting $W = 2$ and approximating two material surfaces as parallel lines, as shown in Fig. S6c. Note that material surfaces are only possible to be parallel when V and T are locked in a specific way. For simplicity, we admit the V-T locking without solving its explicit form, and re-define T as the remaining

Algorithm 1: Detection Probability Calculation Algorithm

Input: The target state $N_\nu(m_\alpha, T_\alpha, V_\alpha)$, in the discretized hyperspace, and mean heat signal $N_\nu(m, T, V)$ for all materials m in the library $\mathcal{M}$, for all temperature T and thermal lighting factor V.

Initialize probability vector, $\mathbf{P}_{M\times 1} = 0$;

/* $\mathbf{P}(m)$ is the probability to predict the target as material m, $m = 1, 2, \cdots, M$. */

for $n_i(\nu_1) \in \{n_1(\nu_1), n_2(\nu_1), \cdots, n_G(\nu_1)\}$

  ...

  for $n_j(\nu_W) \in \{n_1(\nu_W), n_2(\nu_W), \cdots, n_G(\nu_W)\}$

    Calculate probability $\mathcal{P}(\mathbf{n})$ for the ground truth target state to produce the signal point $\mathbf{n} = [n_i(\nu_1), \cdots, n_j(\nu_W)]$, according to Eq. (S16);

    for $m \in \{1, 2, \cdots, M\}$

      for **b** *in lattice sites of the* V-T *plane of* $N_\nu(m, T, V)$

        Calculate probability $\mathcal{P}_s(m, \mathbf{b})$ for possible target state **b** to produce signal **n** according to Eq. (S16);

      end

    end

    Find material of the maximum probability: $m = \arg\max_m\{\max_{\mathbf{b}}(\mathcal{P}_s)\}$;

    $\mathbf{P}(m) = \mathbf{P}(m) + \mathcal{P}(\mathbf{n})$;

  end

end

Output: Probability vector $\mathbf{P}$, with detection probability $P = \mathbf{P}(m_\alpha)$.

dimension of a material line. Worth noting is that, even the scenario is oversimplified, it provides insight of the detection probability and leads to analytical expression that effectively captures asymptotic behaviour of P. Connect points of the same T on both material lines, and define g-axis along that, from the target material (red line, $g = 0$) to the counterfactual material (blue line, $g = 1$). Again, we assume that g-T is a flat space, for simplicity, and that the distribution function of signal (red shaded spot) remains the same for different target states near the ground truth state. In fact, g is a continuous measure of distance between two materials. Now geometry laws assert that the equiprobable hypersurface (black dashed

line) overlaps with the $g = 1/2$ line, no matter what the orientation of g-axis is.

Theorem 1. *For a recorded spectrum* $\mathbf{n} = [n_1, n_2]$*, where* $n_1 \equiv n(\nu_1)$*,* $n_2 \equiv n(\nu_2)$*,* $n_1 \sim \mathcal{N}(N_1, \sigma_1^2)$*, and* $n_2 \sim \mathcal{N}(N_2, \sigma_2^2)$*, when there exists a linear coordinate transform,*

$$\begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} A_1 & B_1 \\ A_2 & B_2 \end{pmatrix} \begin{pmatrix} g \\ T \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \tag{S19}$$

where T *represents the material line,* g *is the material mixture fraction, and material lines are parallel, the detection probability* P *is exactly given by*

$$P = \Pr(g < 1/2) = \int_{-\infty}^{1/2} \mathcal{N}(g; 0, \sigma_{CRB}^2)\, \mathrm{d}g, \tag{S20}$$

as illustrated in Fig. S6d, where σ_{CRB}^2 *is the Cramér-Rao bound of* $E(g)$*, and* $E(g)$ *is the expectation value of* g*.*

Proof. Denote tensor Λ the linear transformation matrix,

$$\Lambda^{\mu}{}_{\alpha} = \begin{pmatrix} A_1 & B_1 \\ A_2 & B_2 \end{pmatrix}. \tag{S21}$$

The associated metric tensor of the g-T space is $\eta_{\alpha\beta} = \delta_{\mu\nu}\Lambda^{\mu}{}_{\alpha}\Lambda^{\nu}{}_{\beta}$, where $\delta_{\mu\nu}$ is the Kronecker delta tensor, as the metric tensor for the n-hyperspace. Since material lines are parallel and signal distribution is invariant, the equiprobable hypersurface coincides with the $g = 1/2$ line, and hence the detection probability is given by

$$P = \int_{-\infty}^{1/2}\int_{-\infty}^{\infty} \sqrt{\det(\eta)} \cdot \frac{1}{2\pi\sigma_1\sigma_2} \cdot \exp\left\{-\frac{1}{2}\left[(\frac{n_1 - N_1}{\sigma_1})^2 + (\frac{n_2 - N_2}{\sigma_2})^2\right]\right\} \mathrm{d}T\mathrm{d}g. \tag{S22}$$

Substituting n with the transform expressions and integrating over T, we have

$$P = \int_{-\infty}^{1/2} \frac{1}{\sqrt{2\pi}} \frac{\det(\Lambda)}{\sqrt{(B_1\sigma_2)^2 + (B_2\sigma_1)^2}} \cdot \exp\left\{-\frac{g^2}{2}\frac{\det(\Lambda)^2}{(B_1\sigma_2)^2 + (B_2\sigma_1)^2}\right\} \mathrm{d}g. \tag{S23}$$

On the other hand, the mean spectrum $[N_1, N_2]$ is also related by the coordinate transform to the expectation values of g and T, which are denoted as $\mathbf{\Theta} = \{\theta_1, \theta_2\} \equiv \{E(g), E(T)\}$. The Fisher information matrix (FIM) for the unknown parameter set $\mathbf{\Theta}$ reads,

$$J_{ij} = \langle \partial_i \log\mathcal{P}(\mathbf{n}) \cdot \partial_j \log\mathcal{P}(\mathbf{n}) \rangle = \sum_{\mu=1}^{2} \frac{\partial_i N_\mu \partial_j N_\mu}{\sigma_\mu^2}, \tag{S24}$$

where $\mathcal{P}(\mathbf{n}) = \mathcal{N}(n_1, N_1, \sigma_1^2) \times \mathcal{N}(n_2, N_2, \sigma_2^2)$, $i, j \in \mathbf{\Theta}$, and average $\langle\cdot\rangle$ is taken over the whole n-hyperspace. The Cramér-Rao bound for an unknown parameter i puts a lower bound to the variance of its unbiased estimator $\hat{i}$,

$$\mathrm{Var}(\hat{i}) \geq \mathrm{CRB}(i) \equiv \left[\frac{1}{J}\right]_{ii}. \tag{S25}$$

Explicitly, we have

$$\mathrm{CRB}[E(g)] = \frac{(B_1\sigma_2)^2 + (B_2\sigma_1)^2}{\det(\Lambda)^2}. \tag{S26}$$

Substituting Eq. (S26) into Eq. (S23) proves Eq. (S20). □

It is straightforward to derive the FIM with the full expression of heat signal probability in Eq. (S16),

$$J_{ij} = \langle \partial_i \log\mathcal{P}(\mathbf{n}) \cdot \partial_j \log\mathcal{P}(\mathbf{n}) \rangle = \sum_{q=1}^{W} \frac{\partial_i N_q \partial_j N_q}{N_q + \sigma^2}. \tag{S27}$$

Here, the FIM depends on specific hyper-spectral/multi-spectral (HS/MS) components like gratings/filters used to obtain the spectral resolution. In order to get an upper bound of FIM that is independent of specific optical components, we note

$$\begin{aligned} \frac{\partial_i N_q \partial_j N_q}{N_q + \sigma^2} &= \frac{(\sum_\nu \partial_i N_{q\nu})(\sum_\nu \partial_j N_{q\nu})}{\sum_\nu (N_{q\nu} + \sigma_\nu^2)} \\ &\leq \sum_\nu \frac{\partial_i N_{q\nu} \partial_j N_{q\nu}}{N_{q\nu} + \sigma_\nu^2} \\ &\leq \sum_\nu \frac{\partial_i N_\nu \partial_j N_\nu}{N_\nu + \sigma_\nu^2} \\ &\leq \frac{N}{1+\gamma} \sum_\nu \frac{\partial_i p_\nu \partial_j p_\nu}{p_\nu}. \end{aligned} \tag{S28}$$

The first equality is to use $N_q = \sum_\nu N_{q\nu}$ and to expand $\sigma^2 = \sum_\nu \sigma_\nu^2$. The expansion of σ^2 is arbitrary and will be determined below. The inequality in the second line can be proved mathematically (not shown) but has an intuitive interpretation — the Fisher information with spectral resolution is not lower than the Fisher information without spectral resolution. In the third line, N_ν is $N_{q\nu}$ with perfect optical transmittance and zero back reflection from the sensor. In the last line, we have used the total number of photons, $N = \sum_\nu N_\nu = \sum_\nu a\lambda_\nu$, in the heat spectrum incident into HADAR, and we have used $p_\nu \equiv N_\nu/N$ to denote the spectral probability distribution of one incident photon. The expansion of σ^2 in the first line is $\sigma_\nu^2 = \sigma^2 p_\nu$, and we have used $\gamma = \sigma^2/N$. Combining the above two equations, we will meet

a double summation $\sum_{q,\nu}$. With grating-type spectroscopy, this double summation is the summation over the full heat spectrum, while with filter-type spectroscopy, the summation over q is repeating the measurement W times. With either kind of HS/MS components, this double summation is equivalent to the summation over the full heat spectrum and the full measurement time. Consequently, the fundamental FIM that is determined by the input photons and irrelevant to optical components can be rewritten as

$$J_{ij} = \frac{N}{1+\gamma} J_{ij}^0 = \frac{N}{1+\gamma} \sum_{\nu} \frac{\partial_i p_\nu \partial_j p_\nu}{p_\nu}. \tag{S29}$$

In deriving Eq. (S29), we have followed the rigorous model of heat signal. However, since Eq. (S29) has been optimized over imaging systems, it has a simpler interpretation. While N is the total input photons and $1+\gamma$ is the degradation by realistic detector noise, J_{ij}^0 is the single-photon FIM that can be directly written down with the spectral probability distribution. For ideal PNRDs, $\gamma = 0$ and Eq. (S29) gives the shot-noise limit. In deriving Eq. (S29), we have also used an approximation in Eq. (S13) that $\lambda_{q\nu} \ll 1$. When this approximation doesn't hold, the variance of photon number for thermal sources of negative binomial distribution equals $N_{q\nu}(1+\lambda_{q\nu})$ instead of $N_{q\nu}$. The extra factor of $1+\lambda_{q\nu}$ accounts for photon bunching, when more than one photon hit HADAR simultaneously and cannot be resolved. This photon bunching effect has been ignored in considering the shot-noise limit.

Materials in a standard library are categorical parameters. Theorem 1 introduces the material mixture fraction g, which is a continuous measure of the distance ($g \in \mathbb{R}$) between two materials representing a 'virtual shift' of the target material from the ground truth $e_\nu(m_\alpha)$ [*i.e.*, $g = 0$] to the counterfactual $e_\nu(m)$ [*i.e.*, $g = 1$]. Theorem 1 suggests that distinguishing two materials is equivalent to estimating the continuous fraction g of a mixture of those two materials and then discretizing g. Therefore, we immediately have an effective theory in algorithm 2 for fundamental limit on HADAR material estimation, where generally $W > 2$ and material surfaces are not parallel. However, we emphasize that Eq. (S30) is not a signal mixture, as g axis is the equal-T contour. In principle, a mixture of signal would involve objects with different temperatures. Axes g and T are not necessary to be orthogonal. By computing FIM, one is quantifying sensitivities of the spectral distribution, in n-hyperspace, with respect to variations of unknown parameters. Put another way, FIM gives correlations of directional derivatives of the information carried by heat signal, with

Algorithm 2: Fundamental Limit on HADAR Material Estimation ($M = 2$)

Input: The material library $\mathcal{M} = \{e_\nu(m)|m = 1, 2, \cdots, M\}$, the target state $\{m_\alpha, T_\alpha, V_\alpha\}$, and the detection system configuration.

Initialize the probability vector, $\mathbf{P}_{M\times 1} = 0$;

/* $\mathbf{P}(m)$ is the probability to predict the target as material m, $m = 1, 2, \cdots, M$. */

Set $m = 1$ or 2, but $m \neq m_\alpha$;

Replace the emissivity in Eq. (S18) with

$$e_{\alpha\nu} = [1 - g] \cdot e_\nu(m_\alpha) + g \cdot e_\nu(m); \tag{S30}$$

Calculate λ_ν according to Eq. (S7) and obtain $p_\nu = \lambda_\nu / \sum_\nu \lambda_\nu$;

Calculate the single-photon Fisher information matrix, J^0_{ij}, for the unknown parameter set $\mathbf{\Theta} = \{g, T_\alpha, V_\alpha\}$ according to Eq. (S29);

Calculate the single-photon Cramér-Rao bound in material estimation, $\sigma_0 = \sqrt{[1/J^0]_{gg}}$;

Calculate the semantic distance between materials, $d_0 \equiv 1/2\sigma_0$, and obtain the statistical (Mahalanobis) distance $d = \sqrt{N/(1+\gamma)}d_0$;

/* Semantic distance d_0 is an intrinsic metric to quantify material difference, irrelevant to photon number nor detector noise. */

Calculate the detection probability P, according to Eq. (S20) which further simplifies to

$$P = \frac{1}{1+\epsilon}, \tag{S31}$$

where $\epsilon = (1 - \mathrm{erf}(d/\sqrt{2}))/(1 + \mathrm{erf}(d/\sqrt{2}))$;

Update probability vector $\mathbf{P}$, $\mathbf{P}(m_\alpha) = P$, $\mathbf{P}(m) = \epsilon/(1+\epsilon)$;

Calculate the Shannon information about the target material, $I = \log_2(P) - \log_2(1/2)$, which is the maximum amount of information that can be retrieved by HADAR;

Output: Shannon information I, semantic distance d_0, statistical distance d, and probability vector $\mathbf{P}$ with detection probability $P = \mathbf{P}(m_\alpha)$.

respect to parameters to be estimated. Off-diagonal terms of the FIM describe the coupling between different parameters, indicating how changing of parameter j would affect the directional derivative of the information with respect to parameter i. As shown in Fig. S6b, at $T = T_0$ where the material library is represented by red and blue dots, the derivative of

probability distribution (or information carried by heat signal) with respect to g is large. This means HADAR is sensitive to material change. At $T = T_1 > T_0$ where red and blue dots move closer to the intersection point or even coincide with each other, the derivative of probability distribution with respect to g is vanishing, which means HADAR becomes insensitive to material change. The matrix inversion in computing the Cramér-Rao bound takes the coupling into account and leads to an effective sensitivity. The statistical distance effectively describes the distinguishability of the target state (red dot) with counterfactual states (blue curve), while the semantic distance distills the distinguishability to a metric that is intrinsic to material difference. The above theory rigorously describes the following intuitive picture. Since temperature T and thermal lighting factor V are both unknown, HADAR has to estimate multiple parameters $\{mTV\}$ simultaneously from heat signal S to identify the material. The material difference at the given target state (red dot) is no longer characterized by the Euclidean distance between the red dot and the blue dot (with the same V and T as the red dot). Instead, it is characterized by the shortest 'distance' of the red dot to the blue curve, where the shortest distance is exactly captured by the multi-parameter statistical (Mahalanobis) distance. We emphasize that high-dimensional integral for the detection probability in algorithm 1 now reduces to a one-dimensional integral, as shown in Fig. S6d. In extreme scenes where $V_\alpha = 0$ and two materials are orthogonal $[e_\nu(1) \cdot e_\nu(2) \equiv 0$ for all $\nu]$, it can be shown that $d_0 \to \infty$, and hence one incident photon suffices to identify the target ($P \to 1$) through measuring the frequency. In the opposite limit where thermal lighting factor $V_\alpha = 1$ and $T_\alpha = T_0$, one can show semantic distance $d_0 \to 0$, and hence it is impossible to identify the target ($P \to 1/2$) no matter how many photons we have. We call the latter situation an equilibrium singularity. In this case, any target would form a cavity and be in thermal equilibrium with the environment, with photon number of the radiation field given by Boltzmann's distribution. It consistently leads to the blackbody radiation spectrum $S_{\alpha\nu} \equiv B_\nu(T_\alpha)$ as given in Eq. (S18), and hides every material feature of the target. One typical example of the equilibrium singularity is the standard cavity-based blackbody source commercially available. As long as the cavity is enclosed (with a tiny hole, $V_\alpha \approx 1$, $T_\alpha = T_0$), the output spectrum is a blackbody spectrum whatever the material is used inside the cavity. Another phenomenon of the equilibrium singularity can be commonly seen in a closed office room ($V_\alpha = 1$, $T_\alpha \approx T_0$). Thermal imaging of walls, desks and chairs inside the office (observed, not shown) would appear uniform of no texture

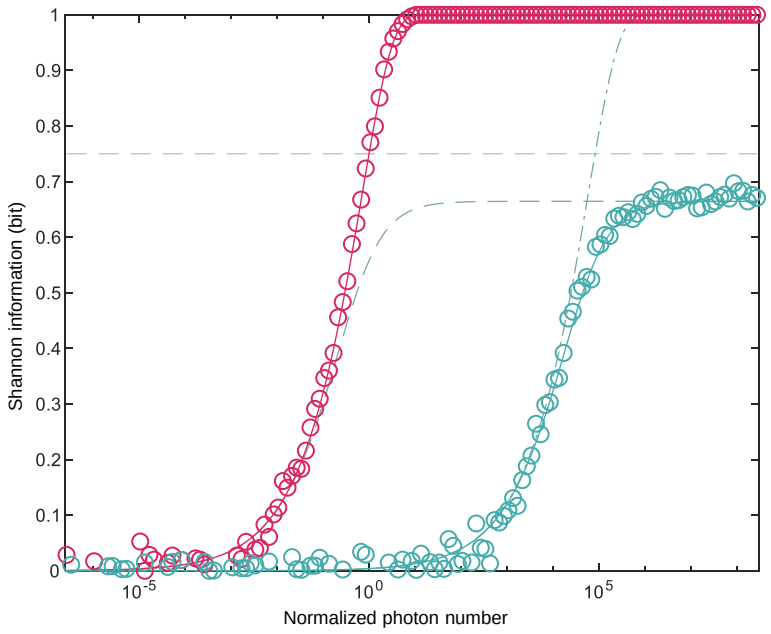


FIG. S7. Shot noise limit of HADAR identifiability for 3 spectral bands, corresponding to the scene of Fig. 3a in the main text. Normalized photon number is Nd_0^2. $\gamma_0 = \gamma_1 N = 6.85e4$ is used here for better visual clarity. Red curve: theoretical shot-noise limit. Red circles: machine learning with Monte Carlo simulations for the shot-noise limit. Cyan solid curve: mixed noise. Cyan circles: machine learning with Monte Carlo simulations for the mixed noise. Cyan dashed: Flicker noise. Cyan dash-dotted: Johnson-Nyquist noise. This figure shows perfect agreement between Monte Carlo simulations and theoretical predictions, as compared with Fig. 3b in the main text.

even they are made of very different materials. The objects are indeed emitting different amounts of thermal radiation. However, the scattered signal from the environment which is in thermal equilibrium with the object completely balances these differences. Numeric experiments of Monte Carlo simulation and machine learning classification further verifies our proposed algorithm 2, as shown in Fig. S7.

Remark 1. *Before HADAR detection, we've assumed that each material in the library is equiprobable to appear in the scene, and hence the initial information about material contained in the scene is* $-\log_2(1/2) = 1$ *bit. Even though different materials may have different volume occupation in the n-hyperspace, enclosed by the equiprobable hypersurface and axes, the assumption would still hold since materials have different populations in the physical scene and the scene is dynamic. Spatial and temporal averaging would render the over-*

all probability to be approximately equal. The asymmetric n-hyperspace volume occupation actually implies the biased estimator is unbalanced. In real applications, boundary effects (existence of axes) usually can be ignored and volume occupation would be effectively equal. After all, one can easily generalize the theory to material library with different appearing probability.

B. Material estimation — multi-material library

To identify an object α in a multi-material library $\mathcal{M} = \{e_\nu(m)|m = 1, 2, \cdots, M\}$ with $M > 2$, the detection probability P has the following asymptotic behaviors,

- if for $m \in \{1, 2, \cdots, M\}$ and $m \neq m_\alpha$, there are $\bar{M}$ materials satisfying $d(m, m_\alpha) \ll 1$, $\bar{M} < M$, then, $P \to 1/(\bar{M} + 1)$;
- if, $\forall m \in \{1, 2, \cdots, M\}$ and $m \neq m_\alpha$, $d(m, m_\alpha) \gg 1$, then, $P \to 1$.

In generalizing algorithm 2, we define material pairs $\{m, m_\alpha\}$ for all $m \neq m_\alpha$. For each material pair, we follow algorithm 2 to compute the semantic distance $d_0(m, m_\alpha) = 1/2\sigma_0(m, m_\alpha)$ and the statistical distance $d(m, m_\alpha) = \sqrt{N/(1+\gamma)}d_0(m, m_\alpha)$. We generalize the prediction probability vector as

$$\begin{aligned} \mathbf{P}(m) &= \frac{\epsilon(m, m_\alpha)}{\sum_{m'} \epsilon(m', m_\alpha)}, \\ \epsilon(m, m_\alpha) &\equiv \frac{1 - \mathrm{erf}[d(m, m_\alpha)/\sqrt{2}]}{1 + \mathrm{erf}[d(m, m_\alpha)/\sqrt{2}]}, \end{aligned} \tag{S32}$$

where summation is taken over $m' = 1, 2, \cdots, M$. When $m = m_\alpha$, $d(m, m_\alpha) = 0$ and $\epsilon(m, m_\alpha) = 1$. One can readily verify that the detection probability, $P = \mathbf{P}(m_\alpha)$ as given in Eq. (S32), meets all asymptotic behaviours listed above and recovers the detection probability in algorithm 2 when $M = 2$. The Shannon information about the target material is given by $\log_2(P) - \log_2(1/M)$. Prediction probability distribution might be useful in understanding machine-learning-based multi-material classification, but it is less important when considering whether a particular material m_α in the library is identifiable or not. The key figure of merit for this is the minimum semantic distance. The underlying physics is that, whether one material can be identified or not depends on its most similar material in the library, not others. Therefore, we define HADAR identifiability of the target material m_α as $I \equiv \log_2(P^{\min}) - \log_2(1/2)$, where the minimum detection probability $P^{\min}$ is

given by Eq. (S31) for the material pair that has the minimum semantic distance. When $M = 2$, HADAR identifiability reduces to the Shannon information of the target material. The overall HADAR material estimation theory is summarized in algorithm 3. Monte Carlo simulations to demonstrate material identification in the multi-material library is given in Extended Data Fig. 6. The minimum semantic and/or statistical distance of each material in the library will decrease when new materials are introduced into the library. Distinct materials are less affected, while similar materials are more affected. Semantic and statistical distance will increase with better spectral resolution (more spectral bands). The above analysis implies the following scaling laws. For fixed spectral resolution, the more materials in a library, the more difficult it is to distinguish each of them. For a fixed number of materials in the library, higher the spectral resolution, the easier it is to distinguish each of them. And to distinguish more materials in a library, better spectral resolution and better sensors are required.

C. Depth estimation

This subsection is denoted to addressing the question of how many photons are needed to reach a desired ranging accuracy. Unlike active ranging where the absolute phase or time-of-flight of signal is used and even one photon suffices to estimate distance, passive ranging suffers from the loss of absolute phase in imaging. Photon position in the image plane records the spatial phase gradient instead, and corresponding multi-photon windows Ω with the same photon distribution in different imaging systems are needed to statistically retrieve at least two gradients to reconstruct the phase and recover targets' position. In this paper, ranging is based on monocular/binocular stereo vision, as shown in Fig. S8. But we stress again that HADAR is not limited to monocular/binocular stereo vision. In binocular stereo vision, a point source target at position (x, y, z) is imaged with left and right focal-plane cameras. Two cameras are on the x axis, with position being $(-b/2, 0, 0)$ for the left and $(b/2, 0, 0)$ for the right. The target falls at $(x_{\mathrm{L}}, y_{\mathrm{L}})$ and $(x_{\mathrm{R}}, y_{\mathrm{R}})$ on two cameras, respectively. Here, subscripts indicate cameras' local coordinate systems whose origins are at the center of the cameras. Within ray optics, we can write down the geometric relations for two yellow-shaded similar triangles, and solve out the position of the target with image

Algorithm 3: Fundamental Limit on HADAR Material Estimation ($M \geq 2$)

Input: The material library $\mathcal{M} = \{e_\nu(m)|m = 1, 2, \cdots, M\}$, the target state $\{m_\alpha, T_\alpha, V_\alpha\}$, and the detection system configuration.

Initialize the probability vector, $\mathbf{P}_{M\times 1} = 0$;

/* $\mathbf{P}(m)$ is the probability to predict the target as material m, $m = 1, 2, \cdots, M$. */

for $m \in \{1, 2, \cdots, M\}$ *but* $m \neq m_\alpha$

  Repeat steps $3 \sim 7$ in algorithm 2 for material pair (m, m_α). Obtain semantic distance $d_0(m, m_\alpha) \equiv 1/2\sigma_0$ and statistical distance $d(m, m_\alpha) = \sqrt{N/(1+\gamma)}d_0(m, m_\alpha)$;

end

Search the minimum semantic distance among all $\{m, m_\alpha\}$ pairs,

$$d_0^{\min} = \min_m\{d_0(m, m_\alpha)\}, \tag{S33}$$

and get the corresponding minimum statistical distance $d^{\min}$;

Calculate prediction probabilities,

$$\mathbf{P}(m) = \frac{\epsilon(m, m_\alpha)}{\sum_{m'} \epsilon(m', m_\alpha)}, \quad \forall m \in \{1, 2, \cdots, M\} \tag{S34}$$
$$\epsilon(m, m_\alpha) \equiv \frac{1 - \mathrm{erf}[d(m, m_\alpha)/\sqrt{2}]}{1 + \mathrm{erf}[d(m, m_\alpha)/\sqrt{2}]},$$

where summation is taken over $m' = 1, 2, \cdots, M$;

Calculate the minimum detection probability $P^{\min}$ with the minimum statistical distance $d^{\min}$ substituted in Eq. (S31);

Calculate the HADAR identifiability of the target material, $I = \log_2(P^{\min}) - \log_2(1/2)$, which is between $0 \sim 1$;

/* The identifiable criterion is given by $d^{\min} \geq 1$, which gives $I \approx 0.75$. */

Output: HADAR identifiability I, minimum semantic distance $d_0^{\min}$ and minimum statistical distance $d^{\min}$, and prediction probability vector $\mathbf{P}$.

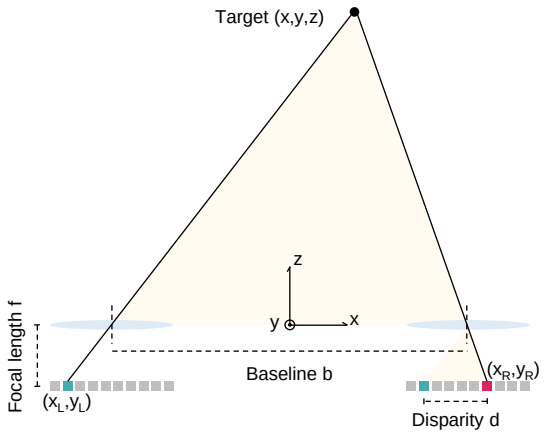


FIG. S8. Schematic of binocular stereo vision.

positions

$$x = -\frac{b}{d}\frac{x_{\mathrm{L}} + x_{\mathrm{R}}}{2},$$
$$y = -\frac{b}{d}\frac{y_{\mathrm{L}} + y_{\mathrm{R}}}{2},$$
$$z = \frac{bf}{d}, \tag{S35}$$

where disparity $d \equiv x_{\mathrm{R}} - x_{\mathrm{L}}$. Therefore, the ranging error δz is given by the disparity error δd through

$$\delta z = \frac{z^2}{bf}\delta d. \tag{S36}$$

When disparity error is fixed, *e.g.*, $\delta d = 1$ pixel, Eq. (S36) gives the quadratic scaling law of ranging error with respect to distance z. Note that in monocular stereo vision or multi-view stereo vision, δz is also proportional to δd but with different coefficients. This subsection will give the fundamental limit of the disparity error δd. For practical sources of the disparity error, such as, image distortion, camera out of calibration, and pixel locking, we refer the readers to Ref. [13]. To distinguish with practical disparity errors, here we call the physics origins of disparity error as the photonic disparity error.

We start with a continuous image plane without pixelization. Disparity error for a given target feature is from searching the areas (windows) on left and right image planes that correspond to the given feature. Fig. S9a shows the schematic of the correspondence problem commonly known in computer vision. The underlying scene S is general and could be any signals, *e.g.*, heat signal or optical signal. The models for imaging system and detector can be found in Sec. SI, especially in Fig. S2. One extra mechanism we have considered here is

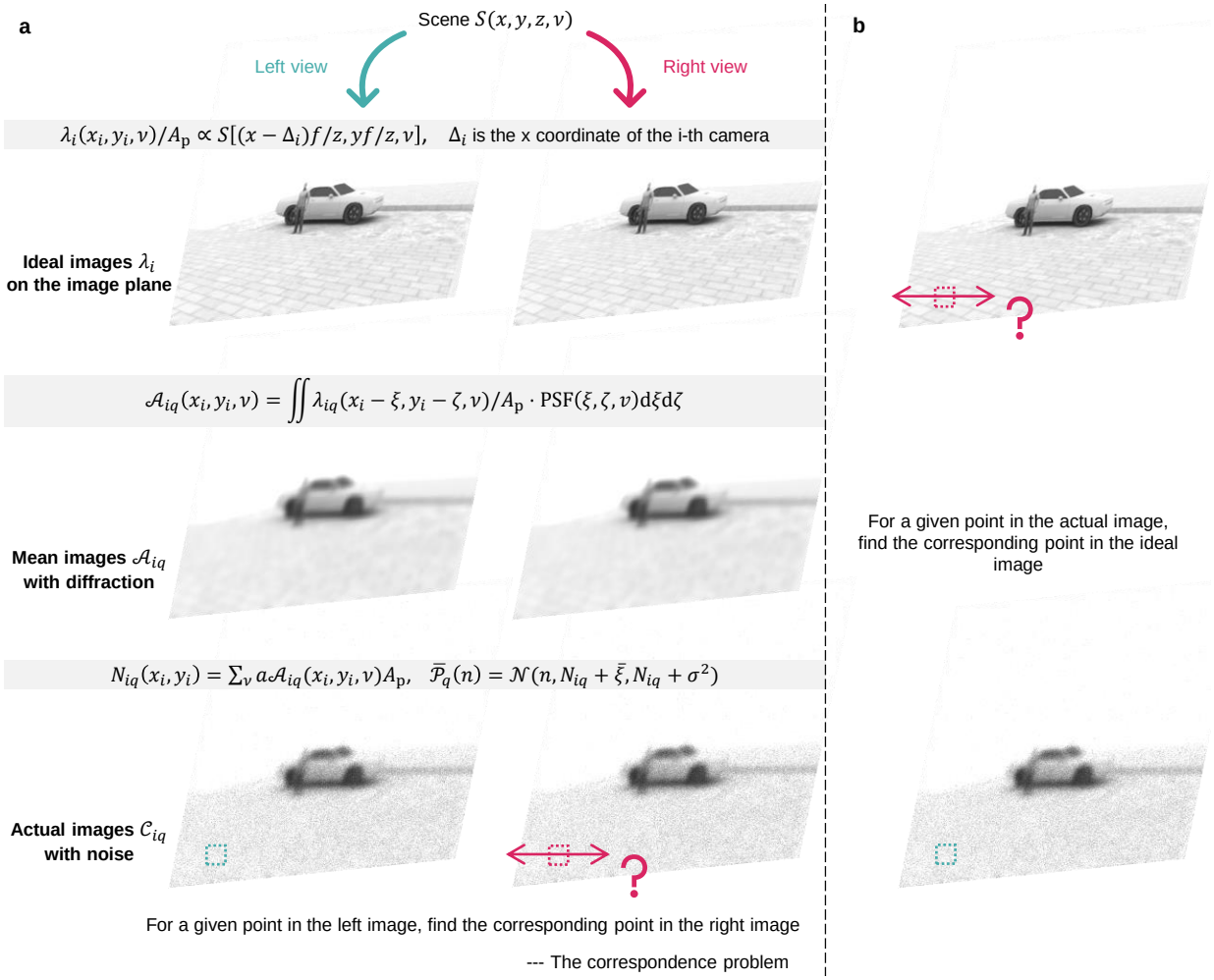


FIG. S9. a. Schematic of the correspondence problem in computer vision. b. We re-interpret the correspondence problem as a position estimation problem in estimation theory.

the diffraction. To focus on fundamental limits, we assume the binocular stereo systems are well calibrated so that $y_\mathrm{L} = y_\mathrm{R}$ and stereo matching is merely along the x axis on the image plane. This is exactly to ignore the practical sources of disparity error. We note that locating a window of the left image in the right image plane can be mapped to locating the window of few observed photons in its ideal image, as shown in Fig. S9b. This is to re-interpret the correspondence problem as a window-position estimation problem in estimation theory. Here, the ideal images are assumed to be noiseless without diffraction and free of occlusion, where exact solution to the correspondence problem exists. For the q-th filter or band, the Fisher information about the x-axis position of a window is given by

$$J_x = \langle \partial_x \log\bar{\mathcal{P}}(n) \cdot \partial_x \log\bar{\mathcal{P}}(n) \rangle = \frac{(\partial_x N_{iq})^2}{N_{iq}+\sigma^2}. \tag{S37}$$

Here, $\bar{\mathcal{P}}(n)$ is given in Fig. S9a, and we have used the same notation as previous subsections except that now the unknown parameter is x and the scene S depending on T, e and X is the given ground truth in the estimation theory. The above Fisher information captures the textures in realistic images that is useful for stereo matching. When photon flux is expressed in image intensity in realistic images (large-N limit), the above denominator is the total variance of noise, and the above Fisher information is consistent with the classical Cramér-Rao bound in [14] which is the inverse of the Fisher information. We emphasize that the classical Cramér-Rao bound in [14] is based on a deterministic signal model. However, imaging with few photons, in principle, is a probabilistic problem that cannot be described by a deterministic model. Therefore, the re-interpretation in Fig. S9 is crucial to derive the shot-noise limit.

In order to get the fundamental limit of ranging error that is independent of optical components, we optimize the Fisher information over optical components to be

$$J_x = \frac{1}{1+\gamma} \sum_{\nu} \frac{(\partial_x N_{x\nu})^2}{N_{x\nu}}, \tag{S38}$$

following the same procedures as Eq. (S29). Here, $N_{x\nu} = a\mathcal{A}_{x\nu}A_{\mathrm{p}}$, and $\mathcal{A}_{x\nu}$ is the convolution of the point-spread function (PSF) with λ_i instead of λ_{iq} as shown in Fig. S9a,

$$\mathcal{A}_{x\nu} \equiv \mathcal{A}_i(x_i, y_i, \nu) = \int \lambda_i(x_i - \xi, y_i - \zeta, \nu)/A_{\mathrm{p}} \cdot \mathrm{PSF}(\xi, \zeta, \nu)\, \mathrm{d}\xi \mathrm{d}\zeta. \tag{S39}$$

The above Fisher information is defined on a window area A_{p}, within measurement time t, and for a given filter or spectral band. It follows that we can retrieve the spectral Fisher information flux $j_{x\nu}$ as the Fisher information per unit wave number, per area on the image plane, and per coherence time. It is obtained by taking such limits of Eq. (S38), and is given by

$$j_{x\nu} = \frac{1}{1+\gamma} \frac{(\partial_x \mathcal{A}_{x\nu})^2}{\mathcal{A}_{x\nu}}. \tag{S40}$$

The above spectral Fisher information flux describes the (spectral, spatial, and temporal) density of incoming information collected by the sensor that is useful for window-position estimation and stereo matching. The maximum Fisher information over a feature/block window, with or without spectral and spatial resolution, can be obtained by taking corresponding limits of Eq. (S38), and are summarized in Tab. S4. Now, the window-position estimation error for either the left or the right view is lower bounded by $\delta x_{\mathrm{L/R}} \geq \sqrt{1/J_x}$,

Fisher information	Without spatial resolution in the window	With spatial resolution in the window
Without spectral resolution	$J_x = \frac{a}{1+\gamma} \frac{(\iint_\Omega \partial_x \mathcal{A}_{x\nu} \,\mathrm{d}s\mathrm{d}\nu)^2}{\iint_\Omega \mathcal{A}_{x\nu} \,\mathrm{d}s\mathrm{d}\nu}$	$J_x = \frac{a}{1+\gamma} \int_\Omega \frac{(\int \partial_x \mathcal{A}_{x\nu} \,\mathrm{d}\nu)^2}{\int \mathcal{A}_{x\nu} \,\mathrm{d}\nu} \,\mathrm{d}s$
With spectral resolution	$J_x = \frac{a}{1+\gamma} \int \frac{(\int_\Omega \partial_x \mathcal{A}_{x\nu} \,\mathrm{d}s)^2}{\int_\Omega \mathcal{A}_{x\nu} \,\mathrm{d}s} \,\mathrm{d}\nu$	$J_x = \frac{a}{1+\gamma} \iint_\Omega \frac{(\partial_x \mathcal{A}_{x\nu})^2}{\mathcal{A}_{x\nu}} \,\mathrm{d}s\mathrm{d}\nu$

TABLE S4. Fisher information about the window position in stereo matching.

with J_x given in Tab. S4. It follows that the fundamental limit of the photonic disparity error is

$$\delta d = \sqrt{\delta x_{\mathrm{L}}^2 + \delta x_{\mathrm{R}}^2} \geq \sqrt{2/J_x}. \tag{S41}$$

We can further simplify the expression for the above Cramér-Rao bound by decomposing the diffraction convolution in Eq. (S39). For convenience, we take the case with both spatial and spectral resolution (third row, third column in Tab. S4) as an example. The derivations apply to all other cases. Since $N = a \iint_\Omega \mathcal{A}_{x\nu} \,\mathrm{d}s\mathrm{d}\nu$ is the total input photon number to the selected window within the heat spectrum, we define $p_A(x,\nu) \equiv \mathcal{A}_{x\nu} / \iint_\Omega \mathcal{A}_{x\nu} \,\mathrm{d}s\mathrm{d}\nu$ as the spectral and spatial probability distribution of one photon. Now, the Fisher information can be rewritten as $J_x = \frac{N}{1+\gamma} \bar{J}_x^0 = \frac{N}{1+\gamma} \iint_\Omega \frac{(\partial_x p_A(x,\nu))^2}{p_A(x,\nu)} \,\mathrm{d}s\mathrm{d}\nu$, where $\bar{J}_x^0$ is the single-photon Fisher information about window position. Moreover, it is fair to approximate that the diffraction only affects the spatial distribution of photons but doesn't change the photon number. This is to assume that the diffraction pattern of the whole scene of interest is completely inside the sensor area. Consequently, we also have $N = a \iint_\Omega \lambda(x_i, y_i, \nu)/A_{\mathrm{p}} \,\mathrm{d}s\mathrm{d}\nu$. Denoting $\lambda_{x\nu} \equiv \lambda_i(x_i, y_i, \nu)/A_{\mathrm{p}}$ and $p_\lambda(x,\nu) \equiv \lambda_{x\nu} / \iint_\Omega \lambda_{x\nu} \,\mathrm{d}s\mathrm{d}\nu$, we can interpret Eq. (S39) as the probability convolution. This reveals that the random position of an observed photon on the image plane (say the left image, $\bar{x}_{\mathrm{L}}$) is a superposition of two independent and random variables, $\bar{x}_{\mathrm{L}} = x + \xi$. The first variable is the corresponding point source in the extended scene which shall generate an ideal image point at x, and the second variable ξ is the displacement of the photon position with respect to the ideal image point caused by diffraction. Now, we have

$$\delta \bar{x}_{\mathrm{L}}^2 = \delta x^2 + \sigma_{\mathrm{d}}^2. \tag{S42}$$

The photon-position uncertainty consists of photonic diffraction uncertainty (σ_{d}^2 given by the width of the PSF; caused by the finite aperture) and photonic correspondence uncertainty (δx^2; without diffraction; caused by the indistinguishability of the photon from different

Fisher information	Without spatial resolution in the window	With spatial resolution in the window
Without spectral resolution	$J_x^0 = \frac{(\iint_\Omega \partial_x p_{x\nu}\,\mathrm{d}s\mathrm{d}\nu)^2}{\iint_\Omega p_{x\nu}\,\mathrm{d}s\mathrm{d}\nu}$	$J_x^0 = \int_\Omega \frac{(\int \partial_x p_{x\nu}\,\mathrm{d}\nu)^2}{\int p_{x\nu}\,\mathrm{d}\nu}\,\mathrm{d}s$
With spectral resolution	$J_x^0 = \int \frac{(\int_\Omega \partial_x p_{x\nu}\,\mathrm{d}s)^2}{\int_\Omega p_{x\nu}\,\mathrm{d}s}\,\mathrm{d}\nu$	$J_x^0 = \iint_\Omega \frac{(\partial_x p_{x\nu})^2}{p_{x\nu}}\,\mathrm{d}s\mathrm{d}\nu$

TABLE S5. Single-photon Fisher information about the point-source location in stereo matching.

objects). The single-photon Cramér-Rao bound denotes the lower bound of the single-photon position uncertainty, $1/\bar{J}_x^0 = \min(\delta\bar{x}_\mathrm{L}^2)$. With the diffraction effect separated out, we can replace $p_A(x,\nu)$ with $p_{x\nu} \equiv p_\lambda(x,\nu)$ in $\bar{J}_x^0$ to get the lower bound of δx^2, $\sigma_\mathrm{c}^2 = \min(\delta x^2)$. Eventually, the fundamental limit of ranging error is

$$\sqrt{N}\delta z \geq \frac{z^2}{bf}\sqrt{2(1+\gamma)(\sigma_\mathrm{c}^2+\sigma_\mathrm{d}^2)}, \tag{S43}$$

where $\sigma_\mathrm{c}^2 = 1/J_x^0$, with the single-photon Fisher information about point-source location, J_x^0, summarized in Tab. S5. The above fundamental limit of ranging accuracy has been verified by Monte Carlo experiments with sub-pixel block matching in Fig. 4 of the main text. Ranging results based on machine learning shown in Fig. S10 also confirms that HADAR ranging is better than thermal ranging. We can further ignore the dispersion effect and assume identical diffraction for thermal radiation at different wave numbers. This is justified in practical stereo-vision applications where ranging error is mainly caused by photonic correspondence error. In the point-source limit, it can be shown that $p_{x\nu} \to \delta(x), \sigma_\mathrm{c} \to 0$, and Eq. (S43) recovers Rayleigh's limit. We now briefly prove that the Fisher information for HADAR ranging with spectral resolution is more than Fisher information for panchromatic thermal imaging. In the mathematical expression of the Fisher information for HADAR in Tab. S5, the spectral information is squared before integral, which prevents destruction of the spectrally resolved Fisher information from contributions of opposite signs. This leads to a larger Fisher information J_x^0 and a smaller photonic correspondence uncertainty σ_c. Mathematically, $\int \frac{(\partial_x p_{x\nu})^2}{p_{x\nu}}\,\mathrm{d}\nu - \frac{(\int \partial_x p_{x\nu}\,\mathrm{d}\nu)^2}{\int p_{x\nu}\,\mathrm{d}\nu}$ can be manipulated into a square form, $(*)^2 \geq 0$, $*$ being a certain expression, and hence it proves that the Fisher information is larger with spectral resolution. More importantly, by breaking the TeX degeneracy, HADAR can support sophisticated priors like sparsity or smoothness to further remove unknowns in the parameter set $\{m_\alpha, T_\alpha, V_\alpha\}$, suppressing ranging error toward a lower bound, $J_x^0 \leq \iint_\Omega \frac{(\partial_x b_{x\nu})^2}{b_{x\nu}} + \frac{(\partial_x k_{x\nu})^2}{k_{x\nu}}\mathrm{d}s\mathrm{d}\nu$, with $b_{x\nu} \equiv \tilde{S}_{x\nu}^0 / \iint_\Omega S_{x\nu}\mathrm{d}s\mathrm{d}\nu$ and $k_{x\nu} = p_{x\nu} - b_{x\nu}$. Here, $\tilde{S}_{x\nu}^0$ is

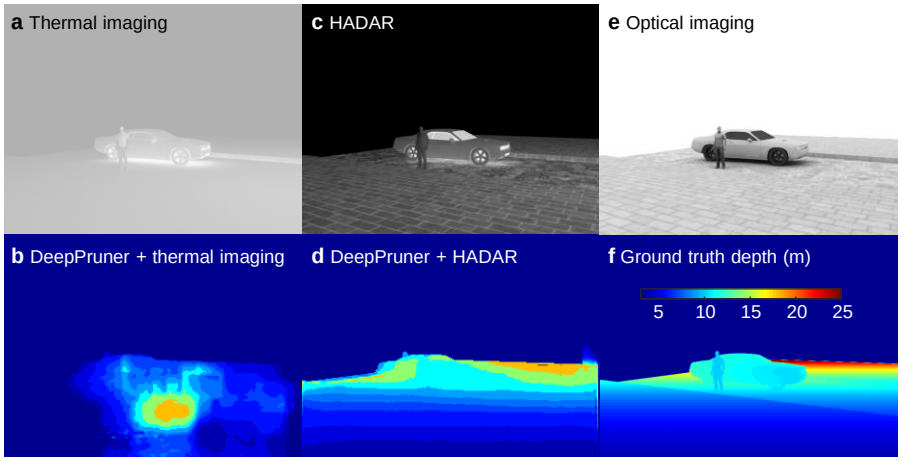


FIG. S10. HADAR ranging beats state-of-the-art thermal ranging. a, Thermal imaging with the ghosting effect. b, Thermal ranging with DeepPruner (pre-trained on the KITTI dataset) [15]. c, HADAR scattered signal with thermal textures. d, HADAR ranging based on DeepPruner. e, Optical imaging. f, Ground truth depth. DeepPruner gives improved performance for HADAR perception as compared to traditional thermal vision. The fundamental reason for this improved performance in HADAR is due to breaking of TeX degeneracy and overcoming the ghosting effect. In contrast, no texture information is collected in traditional thermal sensing, and hence post processing of AI algorithms cannot get accurate depth. HADAR works at the hardware level and enables the AI to provide accurate ranging comparable with the ground truth depth.

the direct emission.

The similar roles of photonic diffraction uncertainty and photonic correspondence uncertainty in Eq. (S43) can be further illustrated in Fig. S11. A point source with diffraction characterized by the point-spread function will generate the same mean image as an extended scene characterized by the point-spread function without diffraction. Therefore, they suffer from equal photonic disparity error, for given photon number and detector.

D. Texture quantification

The single-photon Fisher information about point-source location in last subsection, J_x^0, provides a metric to quantify the textures of the input light field. The numerator in J_x^0 is proportional to δS^2 mentioned in Sec. SID. The presence of its denominator is a consequence of inevitable photon shot noise. Particularly, when electronic noise and diffraction are taken

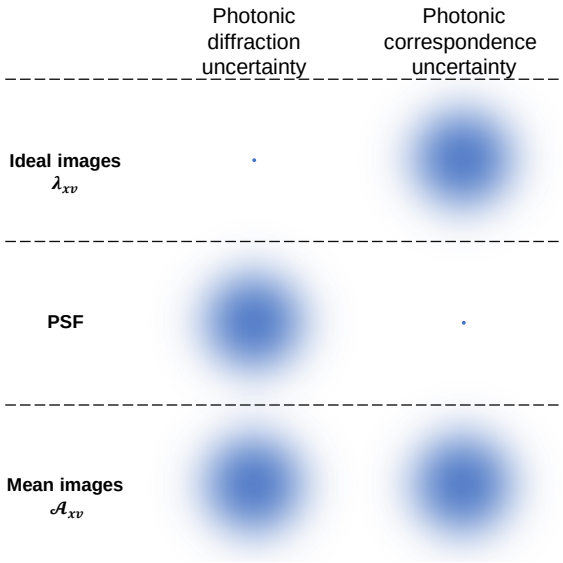


FIG. S11. Comparison of the photonic diffraction uncertainty and the photonic correspondence uncertainty. With one observed photon on the image plane whose position probability distribution is given by the mean image, if we want to infer the position of its emitting point source in the ideal image, error could arise either due to diffraction (photonic diffraction uncertainty) or because we don't know which point source in the extended object is actually emitting that photon (photonic correspondence uncertainty).

into account, Fisher information about window position in Eq. (S37) is the metric to quantify the local textures in realistic images that are useful for stereo matching. Such a Fisher information metric directly connects to the Cramér-Rao bound of ranging errors and has physical significance. Evaluation of the Fisher information metric J_x requires the ground truth $N_{iq} \sim \mathcal{A}_{x\nu} \sim S_{\alpha\nu}$ as well as the characterization of the detector. The requirement of ground truth in evaluating the Fisher information is common in estimation theory. In the literature, there are many other metrics to quantify textures, for example, the standard deviation (stdfilt, available in Matlab) and the local entropy (entropyfilt, available in Matlab) for local textures, or the global entropy for the entire image. These metrics are easy to compute and do not require the ground truth, but they cannot connect to the ranging accuracy. We point out that the finite-difference version of the numerator of J_x on a pixel array, $\sum_{s\in\Omega}[N_{iq}(s, y_i) - N_{iq}(x_i, y_i)]^2$, relates to the variance of the image and hence relates to the standard deviation metric. Important differences between the Fisher information metric and the standard deviation metric are two folds. Firstly, the Fisher information is computed

along the desired x axis, while the standard deviation is computed in square blocks (both x and y directions). The directional variation is crucial to capture the texture that is useful for stereo matching, since the variation along y direction cannot help stereo matching in the x direction. Secondly, the Fisher information removes the effect of noise from texture, while the standard deviation metric treats noise as texture too. All the above three metrics are illustrated in Fig. S12. The images are synthesized so that the ground truth is known for computing the Fisher information metric. In this paper, we will use the Fisher information metric to quantify textures whenever it is possible. When only the actual image is provided, we use the standard deviation in a window to define the local texture density.

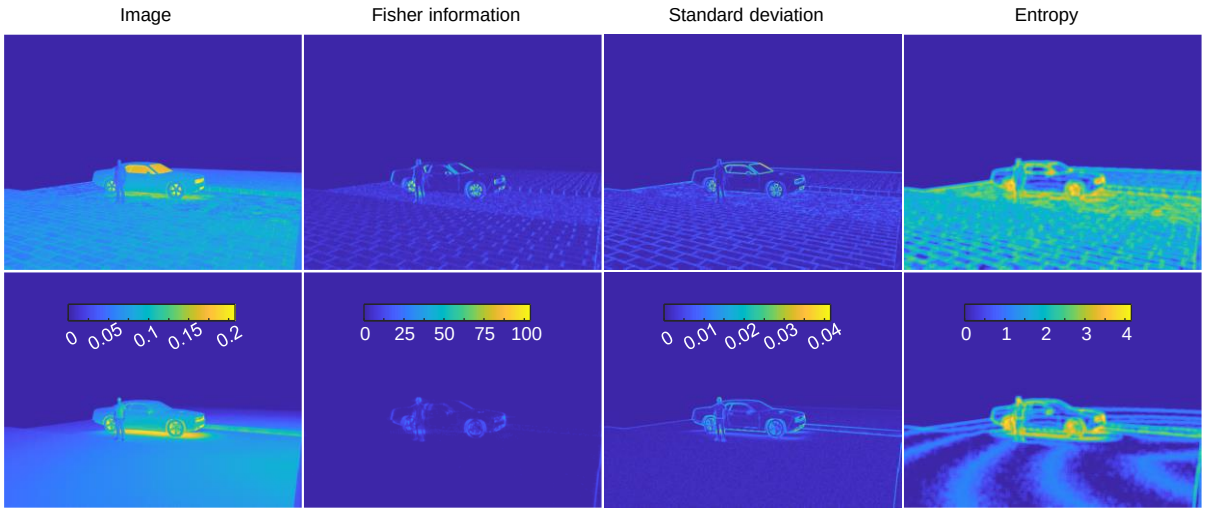


FIG. S12. Comparison of different metrics to quantify textures. The Fisher information metric connects to the Cramér-Rao bound of ranging errors and has physical significance. It captures the x-direction features and ignores the variation caused by noise. The standard deviation metric captures both x- and y-direction variations including noise. Particularly, it cannot connect to ranging errors. The entropy metric captures the local Shannon information but cannot capture the useful information specific to the estimation problem.

Now we use the metric of texture to quantitatively compare our TeX vision with state-of-the-art thermal imaging. Fig. S13 and Fig. S14 show the texture quantifications and the advantage of TeX vision in texture recovery, in winter and summer, respectively. Note that low resolution of textures is caused by low spatial resolution of the FLIR A325sc thermal camera (320×240). With higher spatial resolution, more details of the texture can be recovered, see Extended Data Figs. 2-5. For panchromatic thermal imaging without spec-

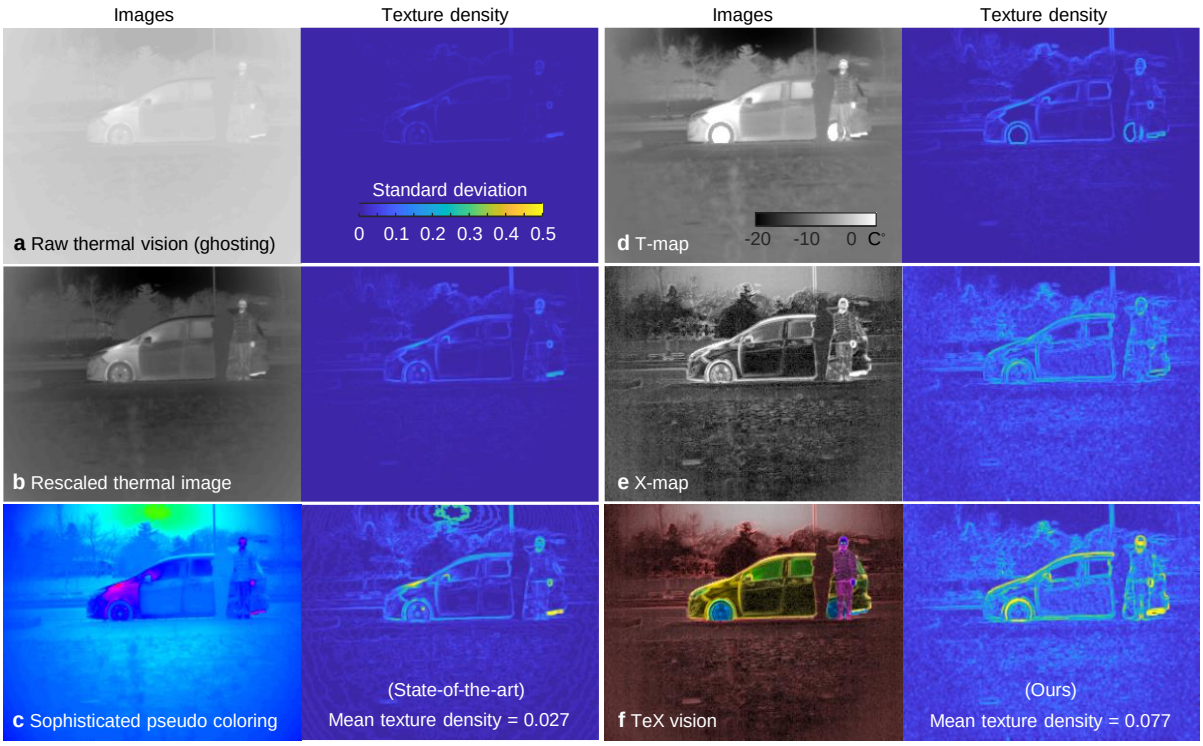


FIG. S13. Experimental HADAR TeX vision at a winter night (Dec. 2020, Indiana, USA) beats conventional thermal vision in textures. a, Raw thermal vision of a winter night scene taken by FLIR camera A325sc. b, Rescaled thermal image to improve the visual contrast. c, Empirical pseudo coloring (HSV, Matlab) adopted by modern thermal cameras to improve the visual contrast. d-f, HADAR T-map, X-map, and TeX vision taken by our HADAR prototype-1. The texture density value at each pixel is the standard deviation of the corresponding 3×3 neighboring pixel array. Our TeX vision gives a texture density of 0.077 on average over the whole image, better than the state-of-the-art pseudo coloring approach which has a texture density of 0.027 on average. Note that more textures means more details in images but means higher values in texture density, so the brighter the texture density figure, the better. Errors remaining in texture (e) are due to mismatch of the emissivity profiles in the real scene with the material library, as well as multi-object scattering contributions, and consequently they are obvious around boundaries. This can be overcome in the future by on-site calibrations of the material library and more advanced decomposition algorithms.

tral resolution, thermal vision becomes low contrast and textureless as shown in Fig. S13a, due to the texture-loss mechanisms explained in Sec. SID. State-of-the-art thermal sensing requiring better contrast uses AGC (automatic gain control) and empirically maps the

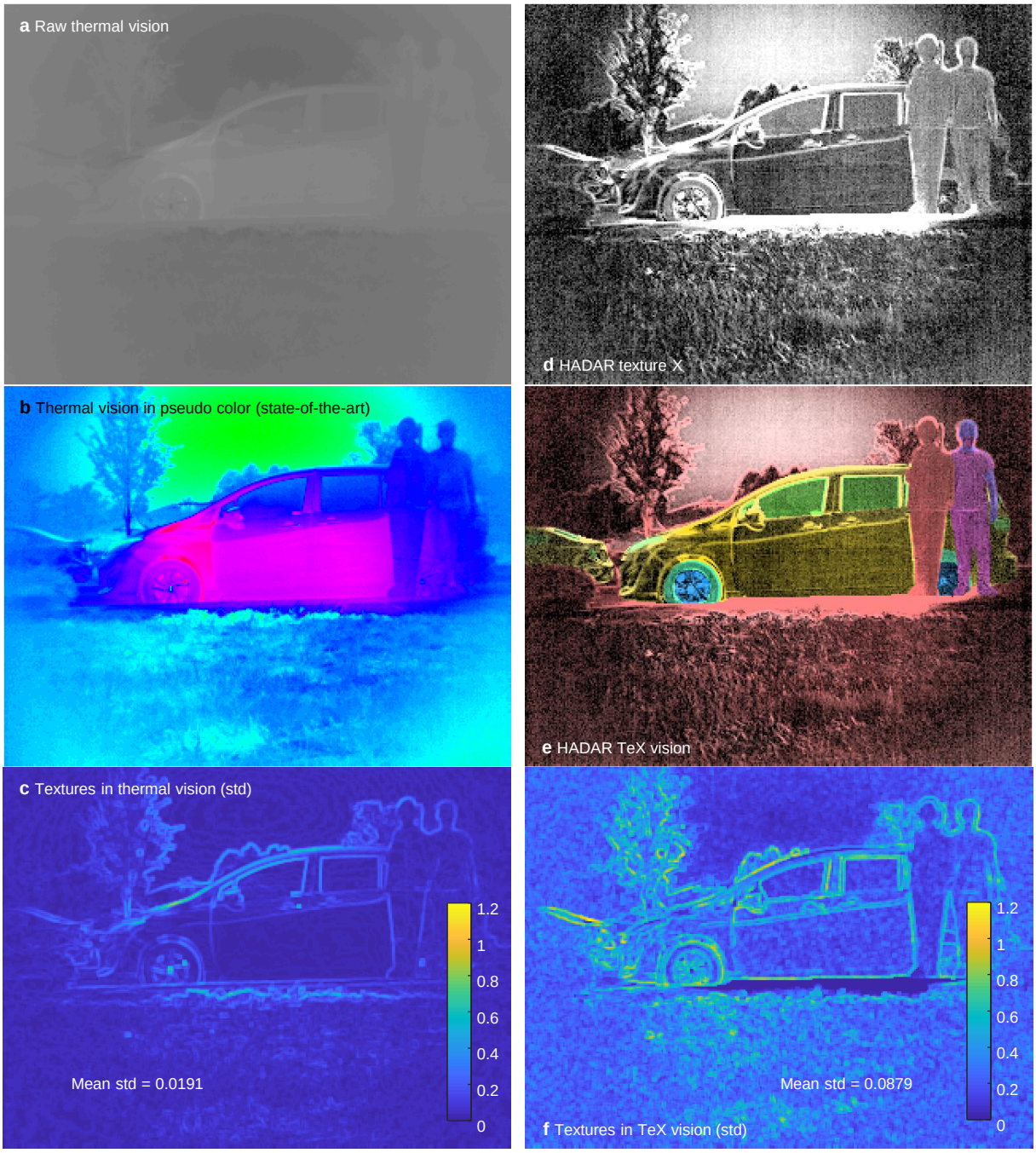


FIG. S14. Experimental HADAR TeX vision in summer daylight (Sep. 2021, Indiana, USA), in comparison with Fig. S13, showing the robustness of HADAR on different environment conditions. a, Raw thermal vision. b, State-of-the-art enhanced thermal vision in pseudo color. c, Texture density (in standard deviation metric) in enhanced thermal vision. d, HADAR texture X. e, TeX vision obtained by our prototype HADAR. f, Texture density in TeX vision. HADAR TeX vision is about 4.6 folds better in texture density than state-of-the-art thermal vision.

intensity with a nonlinear function into either the full grayscale range (b) or pseudo-colors (c). This empirical post-processing cannot block the 4 texture-loss channels in Sec. SID, and only visually and partially improves the contrast. Furthermore, AGC loses all quantitative information of thermal images that is useful for temperature estimation. We emphasize that post-processing can never add new information to the data. To be more rigorous, this empirical post-processing doesn't increase the Fisher information J_x in Eq. (S37). What the post-processing does is to present the data in a more suitable way, according to the visual response and acuity of human eyes to different colors and intensities. This leads to the fact that the pseudo coloring approach needs to be tuned per scene to give the optimal texture visualization. In practice, modern thermal cameras use multiple pre-designed color maps for experimentalists to choose during experiments. With the same FLIR camera, HADAR resolves TeX degeneracy, blocks texture-loss channels (1) and (2) in Sec. SID, and collects more information on the hardware level, which is particularly obvious in the human-robot identification and camouflage problems in Extended Data Figs. 7 and 9. On the visual level, HADAR temperature (d) is similar to the rescaled thermal image (b), but we remind that in (d) subtle geometrical X-type textures have been decoupled. The key to improving visual contrast is to subtract the strong signal floor and keep weak variations. FLIR AGC and pseudo-coloring are empirical approaches to subtract the signal floor. In contrast, HADAR measures temperature and emissivity to estimate the direct emission which is exactly the strong signal floor. Now, the resulting HADAR X-map (e) is mainly the scattered part of heat signal in Eq. (S2). We remind again that the scattered signal is indeed the carrier of rich textures in grayscale optical imaging in daylight, as explained in Sec. SID.

For hyperspectral thermal sensing, in order to visualize the hyperspectral data cube in RGB channels, researchers have been trying to pick out three most significant spectral bands or principal components to maximize textures. To do so for W spectral bands, the $W \times W$ mutual information (or covariance) matrix can be derived and the three bands with least mutual information (most independent) can be chosen to visualize the scene. Or alternatively, principal component analysis can be done to extract the first 3 components. In terms of the Fisher information metric of texture, as shown in the last row of Tab. S4, choosing 3 bands (components) and abandoning the rest is to restrict the spectral integral into 3 narrow bands (components). This decreases the Fisher information and degrades the ranging accuracy. In contrast, HADAR distills all physical information in the heat cube

into TeX parameters, according to the heat signal model. The natural TeX vision can be visualized in RGB (HSV) color space, without loss of information.

E. Bounds in the presence of scene flow

In most real-world applications, either the target or the intelligent agent equipped with HADAR will be moving. The relative motion of the scene with respect to the sensor is described by scene flow in the literature. Projected onto the image plane, scene flow is manifested as the motion blur in each individual image (heat cube in our case) and optical flow in sequential image frames (heat cubes).

The bounds of HADAR identifiability and ranging accuracy given in previous sub-sections are derived for stationary objects and they apply to non-stationary objects with negligible motion blur. Motion blur is usually negligible when the apparent motion Δ of a point source on the image plane is within one pixel, $\Delta < 1$. The apparent motion is given by $\Delta = (v \cdot t \cdot L)/(r \cdot \theta)$, where v is the relative transverse speed, t is the exposure time, L is the number of pixels in the horizontal direction, r is the distance of the target, and θ is the field of view. Motion blur is negligible when either the transverse speed is low or the exposure time is short. For example, a target at 30 m away captured by FLIR A325sc ($t < 12$ ms, $L = 320$, $\theta = 50$ deg) on a car driving at 30 mph [$v \leq 30\sin(\theta/2)$ mph] will have $\Delta \lesssim 0.8$ and hence the motion blur is negligible. To allow a higher travelling speed, the hyperspectral data cube acquisition rate of the used camera must be high so that the exposure time is sufficiently short to avoid motion blur, according to the criterion $\Delta < 1$. This criterion, $\Delta < 1$, constrains the applicability of our bounds. Within the criterion, TeX decomposition can be performed for each individual heat cube to get the TeX vision, and detection and ranging are based on TeX vision. Worth noting is that traditional optical flow, scene flow, semantic segmentation, *etc.*, can all be extensively explored based on TeX vision and depth, presenting a new research frontier. For example, the RGB-d flow in Ref. [16] can be formally transplanted on TeX vision and depth, to retrieve sequential information.

For stronger motion blur beyond the above criterion, if local motion field can be represented by linear convolutional kernels, there are multiple motion-blur removal algorithms available to estimate the motion field [17–19] and get the clean signal without motion blur out of the raw data. Consequently, TeX decomposition and TeX vision are applicable again

after the pre-processing of motion-blur removal.

In the limit of extremely long exposure time, the motion blur kernel is a complicated convolution depending on the velocity field of the scene flow. The algorithms to remove motion blur in the presence of such motion blur are still open questions and deserve future research.

However, in the presence of strong motion blur, the bound for ranging accuracy (Eq. (S43)) still holds, even though the photonic correspondence uncertainty now includes contributions from motion blur in a complicated form. In this scenario, we can directly use Eq. (S37), which is universal for all stereo images (including those with motion blur) and can be derived for given image pairs themselves.

SIII. HADAR ESTIMATION THEORY II: INVERSE MAPPING IN APPLICATIONS

Recall that the heat signal leaving object α is $S_{\alpha\nu} = e_{\alpha\nu}B_\nu(T_\alpha)+[1-e_{\alpha\nu}]X_{\alpha\nu}$, with $X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta}S_{\beta\nu}$. Starting with T_α, $e_{\alpha\nu}$, and $V_{\alpha\beta}$ for all compact and finite objects, Monte Carlo path tracing can solve $S_{\alpha\nu}$ asymptotically with the l-th order scattering-cutoff solution $\tilde{S}^l_{\alpha\nu}$. The residual error $\delta_{\alpha l} \equiv |\tilde{S}^l_{\alpha\nu}-S_{\alpha\nu}| \to 0$ when l increases. Let k denote the maximum number of significant environmental objects considered in the scene, whose spectral emissivity must be one out of M curves in the material library $\mathcal{M} = \{e_\nu(m)|m = 1, 2, \cdots, M\}$. The parameter set $\{klM\}$ determines the complexity of the inverse problem and also controls the accuracy of the solution of $T_\alpha, e_\nu(m_\alpha)$ and $X_{\alpha\nu}$ for given observed $S_{\alpha\nu}$. Note that $S_{\beta\nu}$ in texture $X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta}S_{\beta\nu}$ is partially observed as $S_{\alpha\nu}$. We down sample $S_{\alpha\nu}$ into k spectra to approximately describe k most significant environmental objects. For example, each heat cube in the HADAR-Street dataset has dimension of H $\times$ W $\times$ C $= 1080 \times 1920 \times 54$, H being height, W being width, and C being channel (number of wavenumbers). In our demonstration, we considered $k = 2$ environmental objects. To do so, we spatially split images H $\times$ W into 2×1 quadrants, each quadrant having dimension of $540 \times 1920 \times 54$. Then we spatially average each quadrant into a spectrum of length 54, *i.e.*, for each of 54 channels, we average the 540×1920 sub-image and get its mean value. These 2 spectra, denoted as $S_{1\nu}$ and $S_{2\nu}$, are equivalent objects of the environment, and now the texture is given by

$$X_{\alpha\nu} = V_{\alpha1}S_{1\nu} + V_{\alpha2}S_{2\nu} + \delta_{\alpha\nu,2}, \tag{S44}$$

where the residue $\delta_{\alpha\nu,k}$ is the summation of all sub-leading contributions,

$$\delta_{\alpha\nu,k} \equiv \sum_{\beta\neq1,2,\cdots,k} V_{\alpha\beta}S_{\beta\nu}, \tag{S45}$$

and we have $\delta_{\alpha\nu,k} \to 0$ as k increases. Spectral radiance of external objects beyond the view can also be provided in addition to the down-sampled spectra in Eq. (S44). For example, sky is usually a significant environmental object in open areas but may not be captured in the image.

As explained in Sec. SID, the part of scattered signal that people are familiar with in daily experience is originated only from sky illumination, and hence texture distillation is necessary to recast X. The distillation process is to turn off radiation of other environmental

objects than sky in Eq. (S44) and then evaluate the HADAR constitutive equation in the forward manner without direct emission. Due to the cutoff on number of environmental objects, $\delta_{\alpha\nu,k}$ also contains textures, and hence the final texture is a fusion of the distilled $\bar{X}$ and the residue $\delta_{\alpha\nu,k}$, see Extended Data Fig. 1b.

By substituting down-sampled $S_{\alpha\nu}$ into $S_{\beta\nu}$, we have taken into account infinite scattering ($l = \infty$). The number of environmental objects k is restricted by the number of channels C, $k - 1 + 2 \leq \mathrm{C}$, in order to have a determined solution (number of variables is no more than number of equations). With the texture model Eq. (S44) ignoring the residue, HADAR identifiability and material estimation theory in the last section can be readily generalized to any number of objects and infinite scattering bounces. The unknown parameter set to be estimated becomes $\{g, T_\alpha, V_{\alpha 1}, V_{\alpha 2}, \cdots, V_{\alpha k}\}$.

A. TeX-Net and machine learning

1. *Training data and training strategy*

Our TeX-Net was trained on the HADAR database (https://github.com/FanglinBao/HADAR). The HADAR database includes dissimilar scenes like Crowded Street, Highway, Suburb, Countryside, Indoor, Forest, Desert, etc., covering most common road conditions that HADAR may find applications in. The 11^{th} dataset is a real-world off-road scene with heat cube dimension Height $\times$ Width $\times$ Channel $= 260 \times 1500 \times 49$, while the first 10 scenes are synthetic with heat cube dimension Height $\times$ Width $\times$ Channel $= 1080 \times 1920 \times 54$. The channels in the real-world scene correspond to the $5^{\text{th}} \sim 53^{\text{rd}}$ channels of the synthetic scenes. The HADAR database mimics self-driving situations, with the HADAR sensor(s) either mounted at the positions of headlights, or on the top of the automated vehicles, or on robot helpers. Each scene has 5 frames for each camera, and there are 30 different kinds of materials in total in the HADAR database. For the Street, Suburb, Rocky Terrain, and the Real-World Off-Road scenes, TeX, RGB and IR images are provided for the purpose of ranging. The Street scene has a long animation version (100 frames, 12 channels). For the real-world experimental scene, HADAR sensor is a pushbroom hyperspectral imager that can produce 256 spectral bands. The heat cubes have been interpolated into 49 channels to match the channels in synthetic scenes. Only 49 channels of all the scenes are used

to train TeX-Net. Full technical details about the HADAR database, such as, ray depth, field of view, material properties, and so on, are available in the readme file along with the database.

We split the HADAR database (11 scenes) into training set (80% data) + validation set (20% data) to train the TeX-Net with 5-fold cross validation. Due to limited experimental data, we manually split the database, instead of randomly splitting, to ensure the same diversity of the validation set and training set. Explicitly, in each fold, one frame per view of each scene was selected for validation. We used a hybrid loss with half supervised loss and half physics loss, and we trained TeX-Net for 40K epochs. Since the real-world scene (260*1500) has a different image size with the synthetic scenes (1080*1920), we used random crop (256*256) in training. The network was trained using the number of workers of 8 and a batch size of 20. The learning rate started at 0.001 and dropped by a factor of 10 at 30000 and 37000 epochs. ADAM optimizer was used with the default momentum parameters. The used ResNet50 model was pre-trained on the ImageNet dataset. For synthetic scenes, ground truth temperature and material are synthesized along with the heat cubes. Thermal lighting factors are solved out with least-squares fitting as the ground truth. For the experimental scene, we first applied our proposed TeX-SGD (semi-global decomposition) to generate the TeX vision, as an estimation of the ground truth TeX vision. TeX-SGD results are then used together with synthetic data to train the TeX-Net. TeX-SGD is a non-machine-learning approach that decomposes TeX pixel per pixel based on the physics loss and a smoothness constraint. The hardware environment was Nvidia RTX A6000 48GB GPU. The TeX-Net codes, pre-trained weights and loss curves are available at https://github.com/FanglinBao/HADAR/tree/main/TeXNet.

2. Saliency maps

Saliency map shows the relevant region that is used to predict the desired quantity (material classification). The Saliency map for material classification $e(m)$ in TeX-Net is evaluated by Grad-CAM [20] and given in Fig. S15.

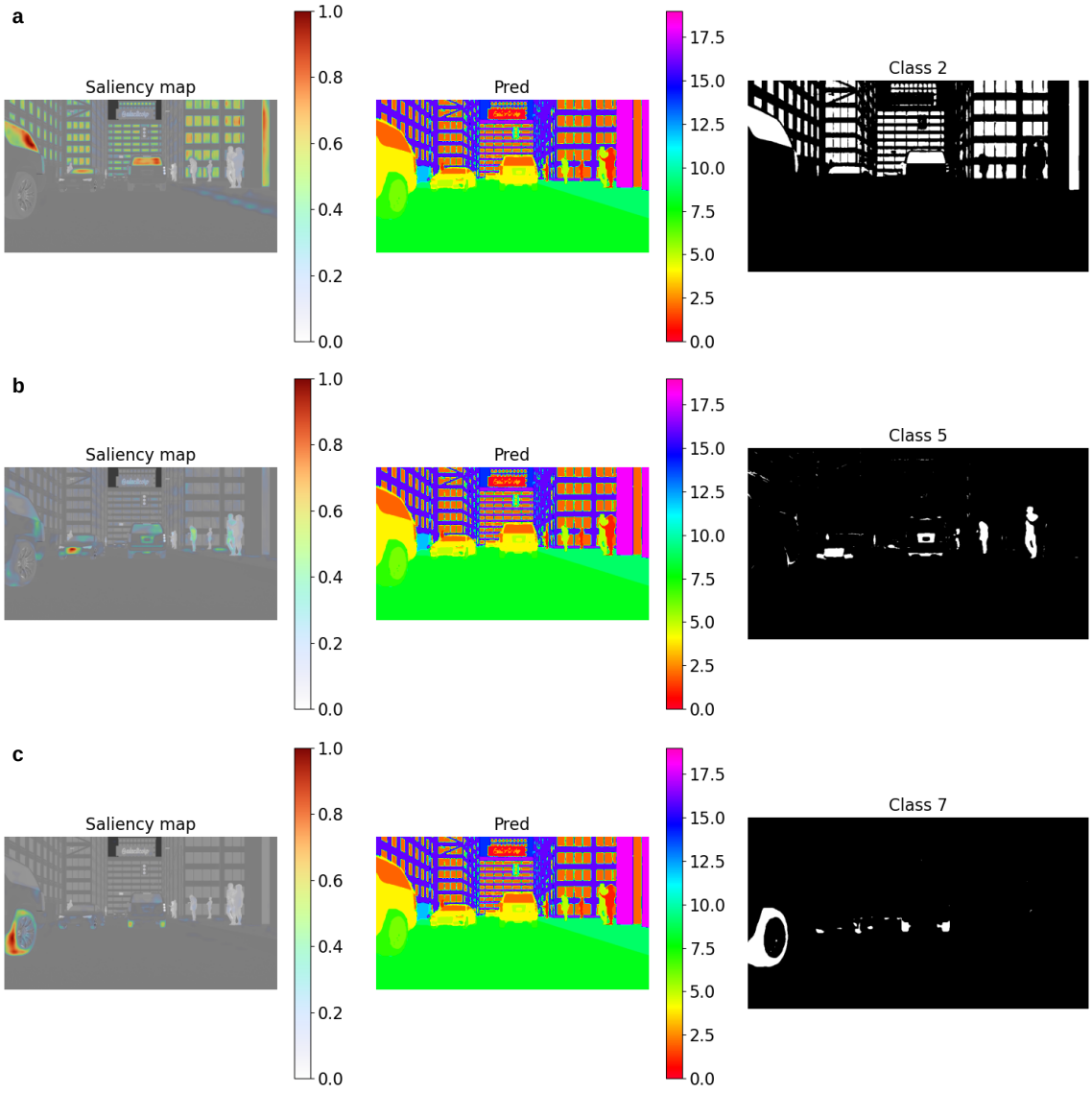


FIG. S15. Saliency map of TeX-Net in supervised learning. The active region in Saliency maps is localized and highly correlated with the corresponding material region (last column), indicating that TeX-Net has properly learnt spatial and spectral features for material classification. 3 samples out of 20 materials are shown. a, Saliency map for class 2, window glass. b, Saliency map for class 5, aluminum. c, Saliency map for class 7, tire. Pred: material index prediction of TeX-Net.

3. Performance and training loss

The supervised training loss and performance of TeX-Net on Street-Long-Animation are shown in Fig. S16.

In unsupervised learning with physics-based loss, TeX-Net searches the best matching

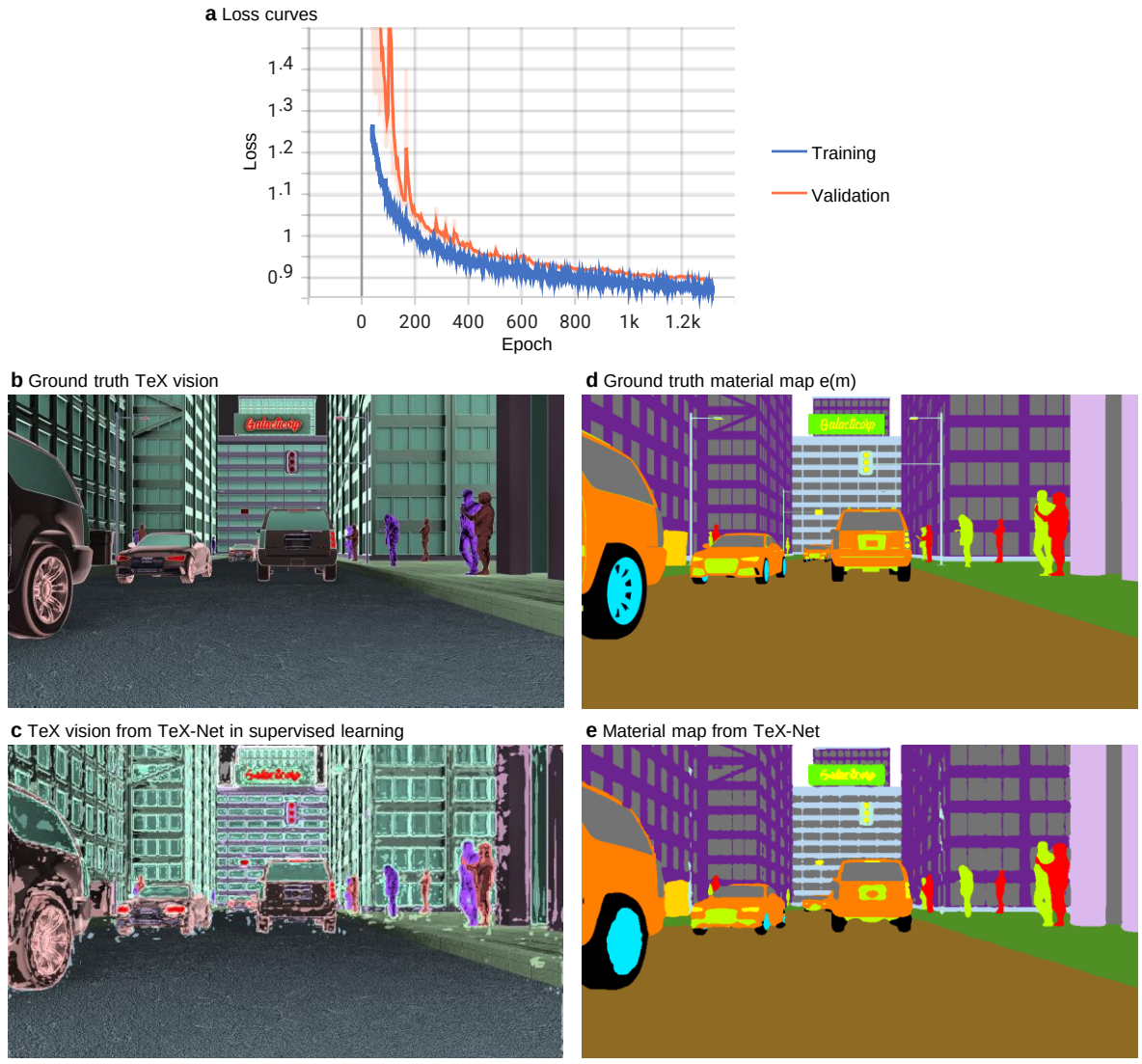


FIG. S16. a, Loss curves in supervised learning showing the convergence of TeX-Net training. b, Ground truth TeX vision. c, Output of TeX-Net. d, Ground truth material map. e, Material map from TeX-Net. The comparisons of TeX-Net output with the ground truth show that TeX-Net is indeed able to do TeX decomposition. Small prediction errors in temperature lead to texture error in brightness, and hence there are some noisy spots observed in c. This can be improved by imposing sophisticated smooth constraint on temperature and harder training in the future. This training was done on the Street Long-Animation dataset in the HADAR database.

TeX for a given signal S. As in practice, standard materials still bear small amount of variations in property, the material library is an approximation of the scene into several material classes. Therefore, the number of materials in the library affects the overall accuracy

of TeX decomposition. For the HADAR-Street dataset which consists of 20 materials, we show the role of material library, by approximating the scene into much fewer material classes and analyzing the overall physics-based loss. For example, in using 3 materials in the library, we only keep the most distinct emissivities of glass and brass, and approximate all other materials as blackbody. This approximation will surely lead to biased temperature and texture, but as the number of materials increases, the loss will decrease. The analysis is given in Fig. S17. With increasing materials, TeX-Net is trained from beginning, and training convergence is not significantly slower.

The TeX-Net performance on the HADAR database is shown in Fig. S18. Training loss curves and the TeX-Net codes are available along with the HADAR database.

While real-world experimental HADAR ranging is shown in Fig. 6 of the main text, generalized HADAR ranging performance over various scenes tested on the HADAR database is shown in Fig. S19. HADAR ranging with ground truth TeX vision shows the optimal performance (Fig. S19a-c). Practical HADAR ranging with predicted TeX vision from the TeX-Net is also shown in comparison with the optimal performance (Fig. S19d). Ranging with both ground truth and predicted TeX vision confirm our argument that 'HADAR sees depth through the darkness as if it were day'. We used DeepPruner (pre-trained on the SceneFlow dataset) for binocular stereovision.

B. Analytical inverse functions, Least-squares estimator, and the TeX-SGD (Semi-Global Decomposition)

Thermal infrared signatures of materials usually have spectral width around $10 \sim 50\,\mathrm{cm}^{-1}$, while temperature feature in the blackbody radiation spectrum has a spectral width around $300\,\mathrm{cm}^{-1}$. In the simplest nontrivial model ($k = l = 1$, the rest of the environment is deep space of zero radiation) of heat signal we demonstrate in this paper, $S_\nu = e_\nu(m)B_\nu(T) + [1 - e_\nu]V_0 B_\nu(T_0)$, approximately only S and e will survive when we take derivative with respect to wavenumber ν. Here, subscripts are occasionally suppressed without risk of confusion. Hence, we have

$$[S/(1-e)]' = [e/(1-e)]' B_\nu(T), \tag{S46}$$

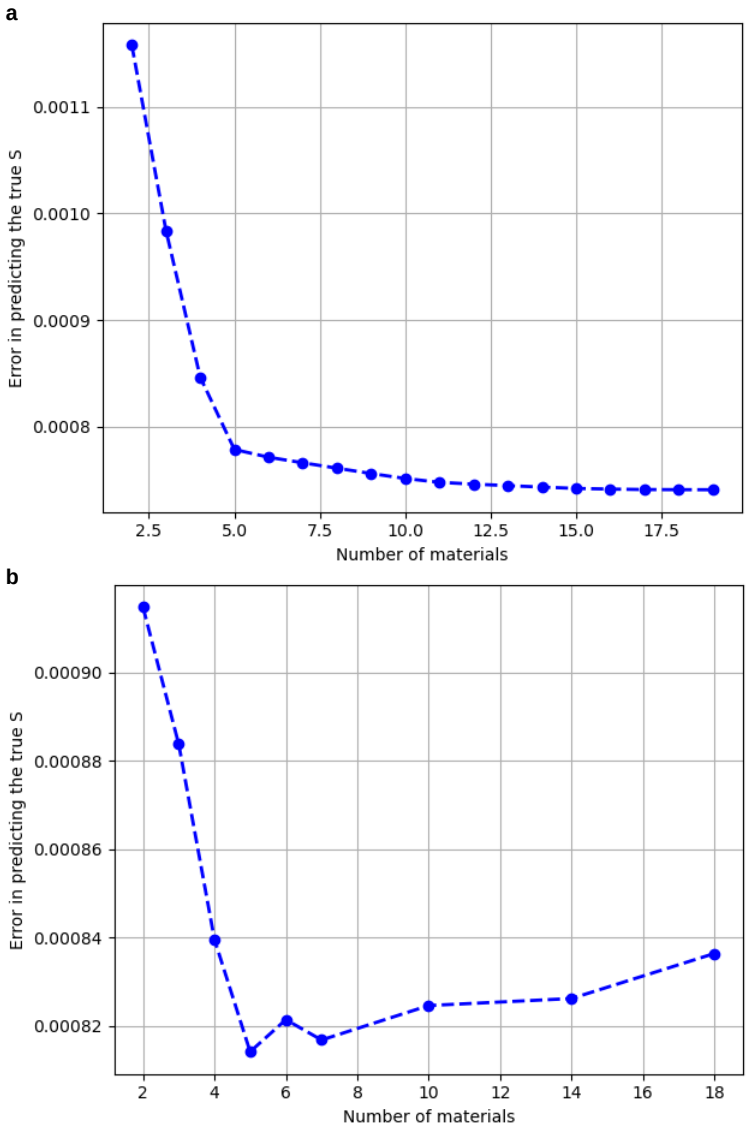


FIG. S17. Physics-based loss decreases as the number of materials in the library increases. a, materials are added into the library with a greedy approach, and pixels are classified into those material classes based on visual similarity. Temperature and thermal lighting factors are solved out accordingly. b, Pixels are classified into material classes with neural network (TeX-Net). TeX-Net finds more accurate TeX decomposition, and again, we can see that with more materials in the library the physics-based loss is lower. The error in (b) after 5 materials is noise.

where prime indicates derivative with respect to wavenumber. Since S is observed and e can be estimated with a material classifier, T can be solved out in the above equation and

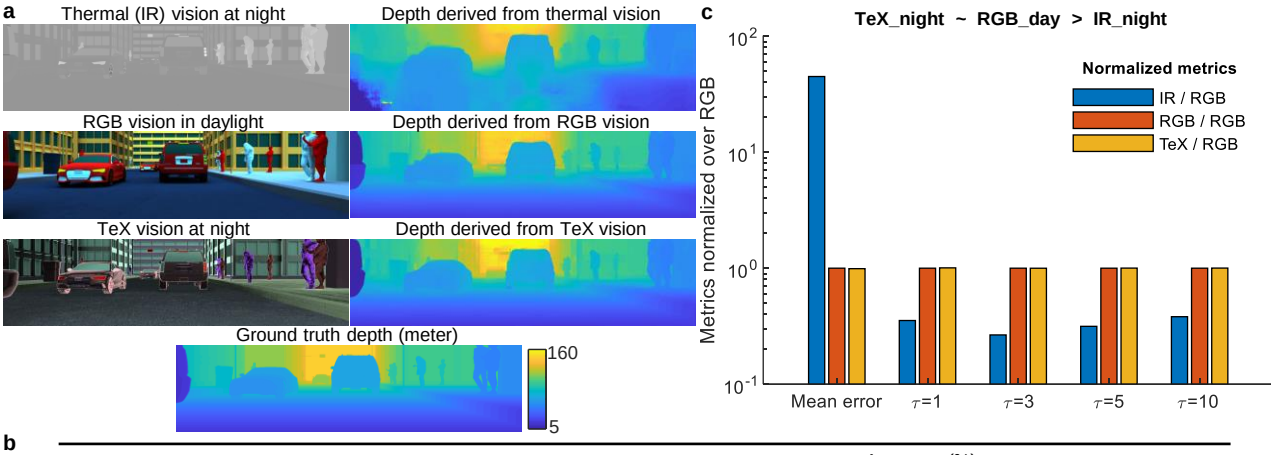
FIG. S18. TeX vision comparison between the ground truth and TeX-Net output. TeX-Net was trained with hybrid loss, an equal-weight combination of supervised loss and the physics-based loss. The HADAR database was split into training set (80% data) and the validation set(20% data) for 5-fold cross validation. The TeX-Net was trained with 40K epochs.

consequently V_0 can be solved in S_ν,

$$V_0 = \frac{S_\nu - e_\nu B_\nu(T)}{(1-e_\nu)B_\nu(T_0)}. \tag{S47}$$

At last, $X = \int V_0 B_\nu(T_0)\mathrm{d}\nu$ can be constructed. Analytical inverse functions are only valid for simple scenes ($k = l = 1$) with high signal-to-noise ratio. Generically, they suffer from noise due to differentiation. With multiple scattering, multiple environmental objects, or if the heat cube is only taken with broadband filters (not given in wavenumber), analytical inverse functions are not feasible.

We note that noise feature size is given by the spectral resolution. To keep the feature scale of thermal infrared signatures away from noise and temperature features, spectral resolution of $1\,\mathrm{cm}^{-1}$ or below is desired for a given noise level. If the measurement time is



b

Overall performance on various scenes (disparity)		Density	Mean error (px)	Accuracy (%)			
				$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
Thermal vision + AI	Ground	0.5	62.66	13.44	25.66	30.91	37.88
	Entire image	1	39.05	26.50	41.16	46.34	52.92
RGB vision + AI	Ground	0.5	1.39	37.98	96.54	98.04	99.24
	Entire image	1	6.12	43.36	83.88	86.27	88.79
TeX vision + AI	Ground	0.5	1.38	38.23	96.47	98.22	99.46
	Entire image	1	2.53	47.91	90.78	93.68	96.13

d

Practical vs. optimal performance		Density	Mean error (px)	Accuracy (%)			
				$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
TeX vision (GT)	Ground	0.5	1.24	55.09	95.87	97.53	99.20
	Entire image	1	3.08	55.05	87.28	89.97	93.22
TeX vision (TeX-Net)	Ground	0.5	1.33	52.43	95.91	97.55	98.85
	Entire image	1	5.15	44.29	77.98	81.23	85.39

FIG. S19. General HADAR ranging performance over various scenes. (c) corresponds to the ground in table (b). The metrics of TeX comparable with RGB and beating IR demonstrates that HADAR ranging at night beats thermal ranging and is comparable to RGB stereovision in daylight. Table (d) shows the comparison of practical HADAR ranging (based on TeX-Net outputs) against the optimal HADAR ranging (based on ground truth TeX vision). Practical HADAR ranging shows a near-optimal ranging performance. Table (b) is based on the Street-Long-Animation, Suburb, and Rocky Terrain scenes in the HADAR ranging dataset in the HADAR database. Table (d) is based on the Suburb and Rocky Terrain scenes, as Street-Long-Animation has less spectral bands and is not included in training TeX-Net. TeX-Net statistics were done with 5-fold cross validation. Ground: bottom half of the image. Density: fraction of the overall image area for which statistics is analyzed. Mean error: the mean absolute per-pixel disparity error with respect to the ground truth. Accuracy: fraction of pixels for which the estimated disparity is within τ pixels of the ground truth values.

fixed, optimal bandwidth for TeX decomposition is a trade off between spectral resolution and signal-to-noise ratio and is problem specific.

In our HADAR prototype-1 experiments, we consider two significant environmental objects ($k = 2, l = 1$),

$$S_\nu = e_\nu(m)B_\nu(T) + [1 - e_\nu][V_1 e_1 B_\nu(T_1) + V_2 e_2 B_\nu(T_2)], \tag{S48}$$

where e_1 and T_1 are emissivity and temperature of the cloudy sky, and e_2 and T_2 are emissivity and temperature of the ground. Subscript α is suppressed without risk of confusion. The unknown parameter set is $\{m, T, V_1\}$ with $V_2 = 1 - V_1$. From S_ν we construct the expected signal on the sensor, $\mathcal{C}'(x_i, y_i, q)$. The mathematical relation from S_ν to $\mathcal{C}'(x_i, y_i, q)$ is given in Sec.SIVA. Heat cube of 10 filters are observed, $\mathcal{C}(x_i, y_i, q), q = 1, 2, \cdots, 10$. Least-squares error, $||\mathcal{C}'(x_i, y_i, q) - \mathcal{C}(x_i, y_i, q)||$, is used to search the unknown parameter set. With the Least-squares estimator, we verified that TeX decomposition is crucial for vision applications and goes beyond the traditional TE (temperature-emissivity) separation approach. TE separation completely ignores the environmental radiation processes or assumes spatially uniform environmental heat signal. In stark contrast, TeX decomposition captures the interplay between the complex real-world scene and its non-uniform environment through the HADAR constitutive equation. Fig. S20 shows that TeX decomposition not only captures local surface normals of objects in the scene, but also gives more accurate material classification than the TE model.

In our HADAR prototype-2 experiments, we consider two significant environmental objects ($k = 2, l = \infty$). We have proposed a non-machine-learning algorithm for TeX decomposition — the TeX-SGD (Semi-Global Decomposition). For any given parameter set, we reconstruct the heat signal, $\tilde{S}_{\alpha\nu}$, to define the local cost, $\mathrm{C}_{\mathrm{local}} = ||\tilde{S}_{\alpha\nu} - S_{\alpha\nu}||$. Furthermore, we impose a global cost to ensure smoothness of the temperature map, $\mathrm{C}_{\mathrm{global}}(x, y) = p \times |t(x, y) - median(N(x, y))|/std(N(x, y))$, where $p = 0.1$ is the global penalty and $N(x, y)$ is a 3×3 neighbouring window of the temperature map around pixel (x, y). The total cost function $\mathrm{C} = \mathrm{C}_{\mathrm{local}} + \mathrm{C}_{\mathrm{global}}$ is used to search the unknown parameter set and decompose T, e and X.

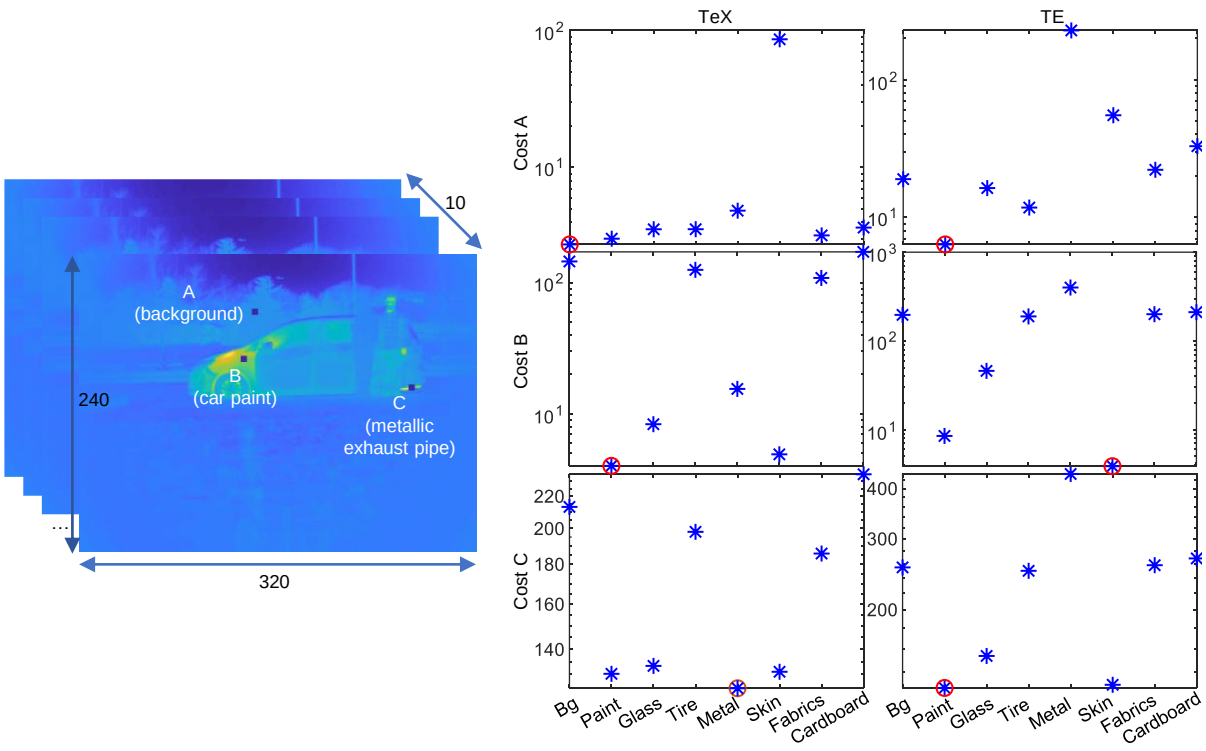


FIG. S20. Least-squares estimator verifies that TeX decomposition is a more accurate heat signal model than the traditional TE separation. TeX model gives textures while TE model ignores textures. Moreover, TeX model can predict correct material classifications, while the TE model returns wrong results. Comparisons are made for some typical pixels in the HADAR prototype-1 outdoor experiment at winter night. Blue stars: cost values for each possible material candidate. Red circles: predictions given by minimum costs.

C. AGC on TeX vision

Originally, automatic gain control (AGC) is a nonlinear mapping of raw thermal data to grayscale values between 0-255 that can be visualized on a computer screen. For more details, see FLIR AGC at https://flir.netx.net/file/asset/15755/original. The spirit of AGC can be adapted to HADAR to improve visual effects of TeX vision, and nonlinear mapping can be performed on individual channels of T, e, and X. In TeX vision, hue is the material label normalized inbetween 0 and 1. One can manually assign discrete values to hue for each material, to customize the color representation. For temperature, it usually has lower and upper bounds in the real world. For example, the lowest temperature and highest temperature of a given city are usually on record. Therefore, one can trim

temperature with the bounds and linearly transform it inbetween 0 and 1. The distilled texture (see Extended Data Fig. 1b) is based the thermal lighting factors in Eq. (S44) and the k-residue $\delta_{\alpha\nu,k}$ given in Eq. (S45). We take the logarithm of X, rescale it inbetween 0 and 1, and apply a nonlinear AGC for visualization. When sky is not specified as one environmental object, only the residue remains in the distilled texture. We observed that when the sensor is of low noise or when using TeX-Net, residue alone gives vivid textures. However, when the sensor is noisy and when using TeX-SGD, thermal lighting factor of the sky is usually needed for fusion.

Explicitly, here we show how a TeX vision image is formed from the $T(x,y), e[m(x,y)]$, and $X(x,y)$ output by TeX decomposition (through TeX-Net or TeX-SGD). (1), The material library $\mathcal{M}$ comes with a hue library, $\mathcal{H}$. For each material in $\mathcal{M}$, we've assigned to it a hue value corresponding to its typical color as can be seen in daylight, so that the TeX vision will be similar to the familiar RGB image. For example, a hue value 90/255 corresponding to 'Green' is assigned to the material 'Vegetation', a hue value 160/255 corresponding to 'Blue' is assigned to the material 'Water', and a hue value 40/255 corresponding to 'Yellow' is assigned to the material 'Sand'. The hue channel of the TeX vision image is given by the following matlab pseudo code, $\mathrm{H} = \mathrm{reshape}(\mathcal{H}(\mathrm{m}), \mathrm{size}(\mathrm{m}))$. (2), The saturation channel of the TeX vision image is given by the following matlab pseudo code, $\mathrm{S} = \mathrm{rescale}(\mathrm{T}, 0, 1, \text{'InputMax'}, \mathrm{tMax}, \text{'InputMin'}, \mathrm{tMin})$, with user customized temperature range. (3), The Brightness channel of the TeX vision image is given by the following matlab pseudo code, $\mathrm{V} = \mathrm{adapthisteq}(\mathrm{rescale}(\mathrm{X}, 0, 1))$. (4), Finally, the TeX vision image is given by a color space transform, $\mathrm{texIMG} = \mathrm{hsv2rgb}(\mathrm{cat}(3, \mathrm{H}, \mathrm{S}, \mathrm{V}))$. Sample data and codes are available along with the HADAR database at https://github.com/FanglinBao/HADAR. Fig. S21 further shows an example of TeX vision image, in comparison with thermal vision and RGB vision.

Traditional AGC tries to extract signal variations by subtracting a dominant local signal floor, since it is the dominant and uniform signal floor that visually fades signal variations and renders thermal images textureless. The dominant signal floor is the unknown direct emission and unwanted scattering which change smoothly over the image. Traditional AGC or pseudo coloring are empirical approaches to estimate the signal floor, while HADAR is to estimate the signal floor according to the heat signal model. Furthermore, using the distilled texture $\bar{X}$ as X is also consistent with the spirit of AGC. In this point of view, it

is not surprising that HADAR TeX vision can give better visual contrast than traditional AGC. In current TeX-Net training, we observed that T and V are not so accurately learnt for experimental scenes, and hence textures all remain in the residue. Without texture distillation, the current TeX-Net amounts to be a learning-based approach to estimate the signal floor and extract textures.

D. Pseudo-TeX vision

As TeX vision requires the input of hyperspectral heat cubes, we also propose pseudo-TeX vision to extend its applications to common thermal datasets without spectral resolution.

Existing thermal datasets provide panchromatic thermal images, $S_\alpha = \int S_{\alpha\nu} \mathrm{d}\nu$. Firstly, for near-black objects, $e_\nu \to 1$, $S_{\alpha\nu} = e_{\alpha\nu} B_\nu(T_\alpha) + [1 - e_{\alpha\nu}] X_{\alpha\nu} \approx B_\nu(T_\alpha)$, and hence thermal image is widely taken as the temperature contrast. Standard thermal cameras can do the inverse transform and estimate the temperature. Therefore, we use the thermal image itself to approximate temperature T. Secondly, existing semantic segmentation based on thermal vision can extract spatial patterns from thermal images and estimate semantic categories. Even though this segmentation is vision-driven, we can use it to approximate material category $e(m)$. Thirdly, AGC (automatic gain control) can improve visual contrast, maximizing the usage of residual texture in sensor data. We use it to approximate texture X. Putting them together, we formally get TeX vision, see Fig. S22, though the information contained in this pseudo-TeX vision is no more than the original thermal vision. Pseudo-TeX vision uses information of different levels (spatial pattern, rough temperature, and weak variation) to extrapolate the material and geometry information and might find applications, *e.g.*, in practical ranging, see Fig. S23.

E. Physics-driven semantic segmentation, object detection and visual object tracking

Here, we use a customized non-machine-learning algorithm, in addition to existing pre-trained neural network models, to demonstrate physics-driven semantic segmentation, object detection, and visual object tracking. We emphasize that machine-learning semantic segmentation and detection based on TeX vision (instead of traditional RGB vision, thermal

vision, or point cloud) present a new research frontier and deserve future studies.

For object detection in TeX vision, we can extract the region corresponding to the desired material according to the material map in the hue channel. Existing algorithms can be applied to the specific region for detection, instead of the entire image, see Fig. S24. This approach combines intrinsic material signatures with spatial patterns for detection and can distinguish human vs. robot, which is otherwise impossible. Here, Fig. S24 demonstrates sequential detection by performing detection on each individual material region. We believe that simultaneous detection can be achieved in the following approaches. (1) TeX vision images can be used as input to train a neural network for simultaneous detection. (2) Our TeX-Net with the physics model can be utilized as a backbone to design and train novel end-to-end networks for simultaneous detection, and it is not necessary to explicitly output TeX vision. These approaches deserve future studies.

The material map $e(m)$ itself is not a semantic segmentation, but each semantic category is usually a combination of several materials. For example, in the following Fig. S25a, material 'road', 'pavement', 'sky', and 'human' can be directly mapped to corresponding semantic categories. The semantic category 'car' is a combination of materials of 'car paint', 'window glass', 'headlights', 'rubber tire', 'wheel', and 'aluminum (logo)'. The semantic category 'robot' corresponds to material 'aluminum'. And semantic category 'building' consists of 'window glass', 'brick', 'concrete', 'steel', and so on. We can use the following heuristic algorithm 4 to transform a HADAR material map into a semantic segmentation in Fig. S25b. Worth noting is that material 'aluminum' in the car is enclosed by other components like 'car paint', but 'aluminum' in robot is open (not enclosed). This pixel interaction among neighboring pixel arrays is used to heuristically transform a material map into a semantic segmentation. More sophisticated algorithms to transform material map into semantic segmentation deserves future studies.

Visual object tracking based on TeX vision and corresponding semantic segmentations has been tested for a car and a pedestrian on the Street-Long-Animation scene. pytracking implementation (https://github.com/visionml/pytracking) of the ECO [21] method was used in the test. Robust tracking results show the applicability of TeX vision for visual object tracking. Tracking videos are available along with TeX videos at https://github.com/FanglinBao/HADAR.

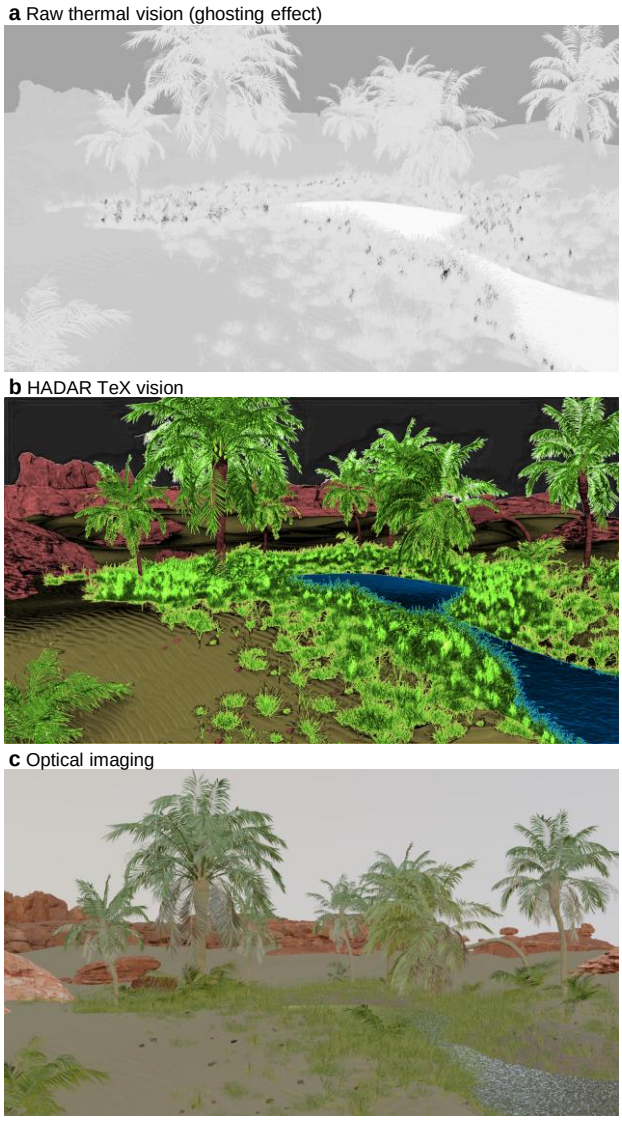


FIG. S21. HADAR sees through the darkness as if it were day. a, Ghosting thermal vision of a desert scene at night, synthesized by path tracing. b, HADAR TeX vision of the desert scene, with a library of 6 materials, sky: black, rock: dark red, sand: yellow, vegetation: green, bark: brown, water: blue. Color hue to present these 6 materials in TeX vision are determined according to daily experience. c, Optical RGB vision of the desert scene in daylight. HADAR TeX vision recovers textures and distinguishes different materials. It can be clearly seen that TeX vision captures the scene as if it were day.

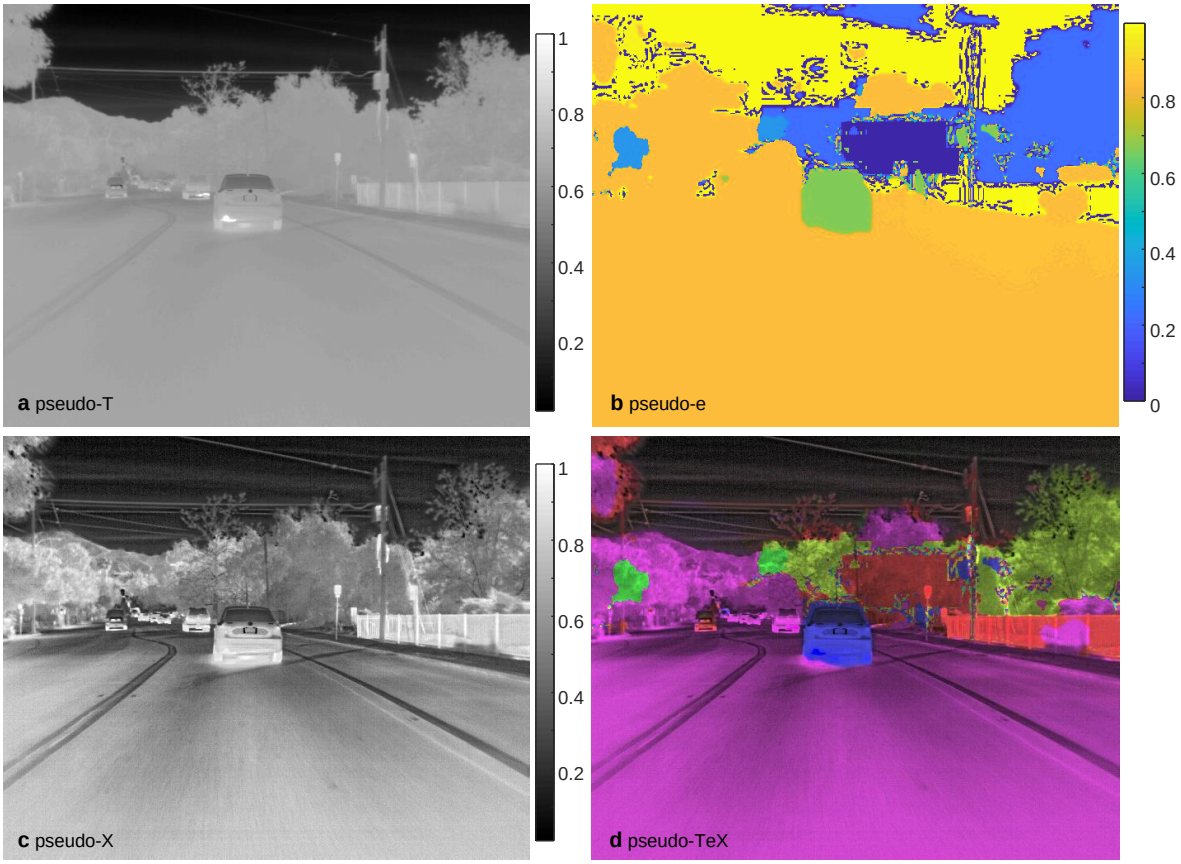


FIG. S22. Pseudo-TeX vision of the a sample thermal image from FLIR thermal dataset https://www.flir.com/oem/adas/adas-dataset-form/. Pseudo-TeX vision extracts information from multiple levels, such as, spatial pattern, absolute value, and spatial variation, and represents them in a compact and reader-friendly form.

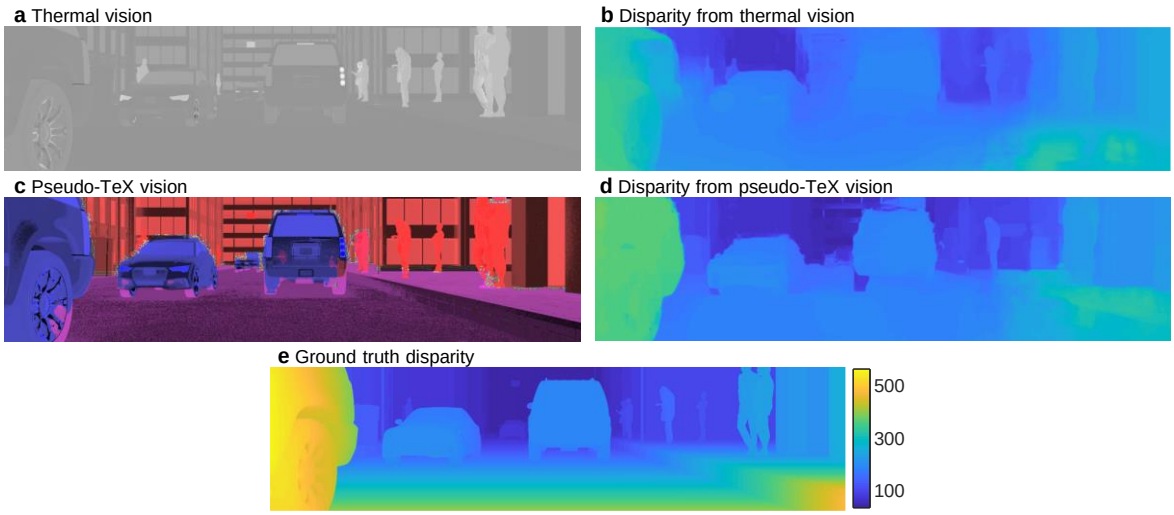


FIG. S23. Disparity in stereo matching based on raw thermal vision and pseudo-TeX vision. Pseudo-TeX vision gives more accurate disparity estimation than raw thermal vision closer to the ground truth. Disparity estimation is done with DeepPruner. Even though pseudo-TeX vision is derived from the same data as the raw thermal vision, a more sensitive representation of features can yield better performance.

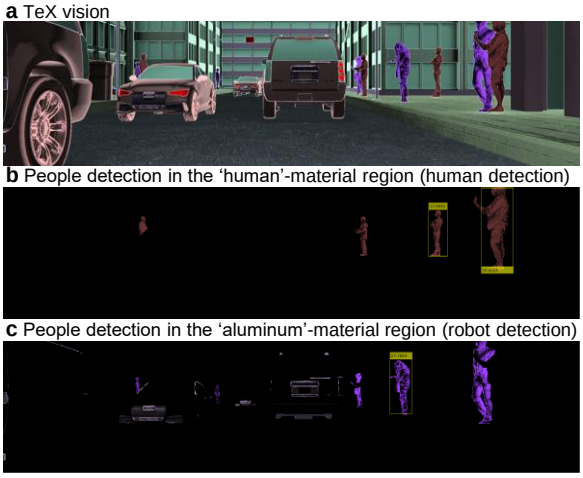


FIG. S24. Demonstration of physics-driven people detection. Instead of detecting people over the entire image, we extract the region corresponding to the desired material and then perform detection with existing algorithms in the literature.

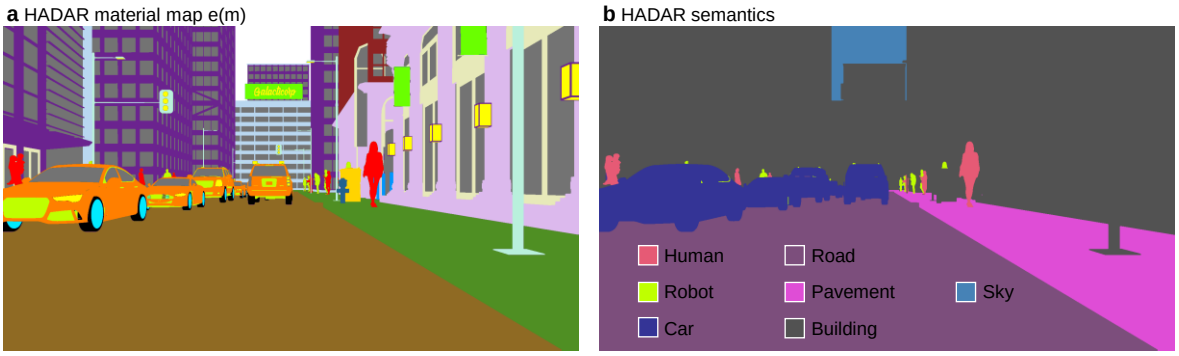


FIG. S25. Demonstration of physics-driven semantic segmentation. HADAR material map (a) is transformed into a semantic segmentation (b) using algorithm 4.

Algorithm 4: Physics-driven semantic segmentation

Input: The material map m_{xy}, the semantic categories $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, the material combination for each semantic category $\omega_i \leftarrow \{m_i\}$, and in each semantic category, the property of each material η indicating it is enclosed ($\eta = 1$) or not ($\eta = 0$).

for *each pixel* (x, y)
  Initialize the probability vector, $\mathbf{P}_{n\times 1} = 1/n$;
  /* $\mathbf{P}(i)$ is the probability in semantic category $i, i = 1, 2, \cdots, n$ */
  for *each semantic category* ω_i
    if $m_{xy} \notin \omega_i$ **:**
      $\mathbf{P}(i) = 0$.
    end
  end
  Normalize the probability vector, $\mathbf{P} = \mathbf{P}/\mathrm{Sum}(\mathbf{P})$;
  if $\max(\mathbf{P}) = 1$ **:**
    Semantic segmentation $S_{xy} = \mathrm{argmax}_i(\mathbf{P})$.
  else:
    Get left and/or up neighbor-pixel category S';
    if $\mathbf{P}(S') \neq 0$ **:**
      $S_{xy} = S'$.
    else:
      $S_{xy} = i$, where in ω_i, $\eta = 0$ for m_{xy};
    end
  end
end

/* The following procedure is for smoothing */

for *each pixel* (x, y)
  $S_{xy} = S'$, where S' is the mode of the local 5×5 pixel array. Symmetric padding is used for boundaries;
end

Output: Semantic segmentation map S_{xy}.

SIV. HADAR PROTOTYPE-1: EXPERIMENTS

The HADAR prototype-1 is based on the FLIR A325sc thermal camera and a set of filters in the long-wave infrared (LWIR). While the experimental setup with schematic diagram is shown in Extended Data Fig. 10, the corresponding heat signal model for each detector is shown in Fig. S26. The collected signal consists of multiple contributions. The rest of this section shows the calibration to get the signal from the scene out of the total collected signal.

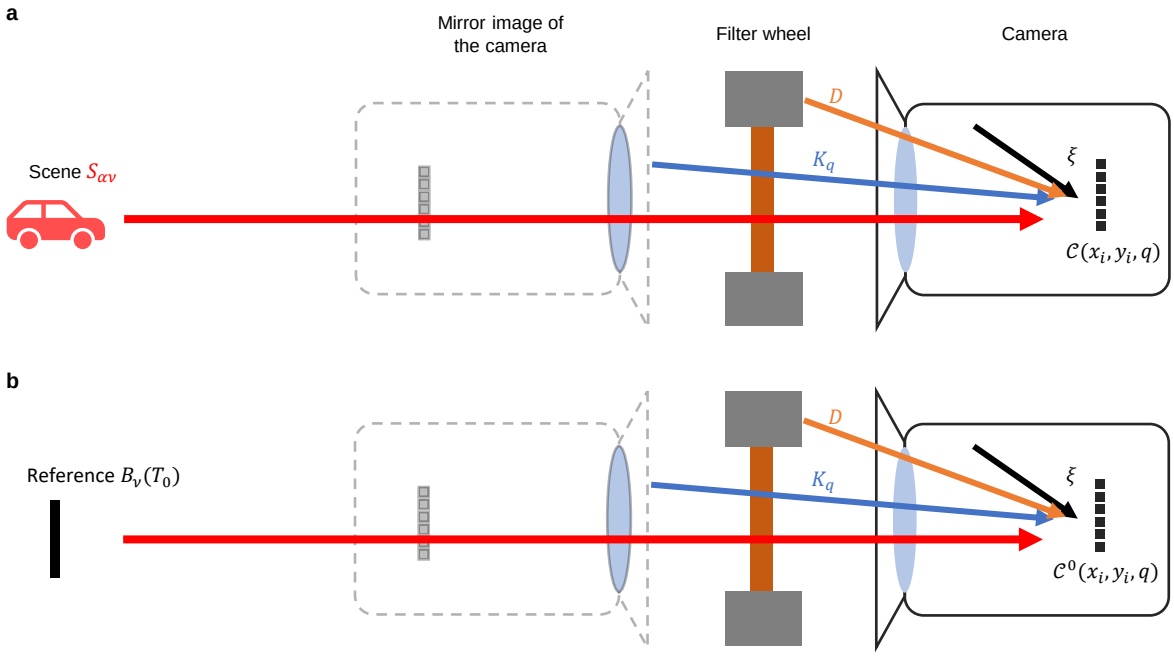


FIG. S26. Heat signal models of the HADAR prototype-1 in experiments (a) and calibration (b). $\mathcal{C}(x_i, y_i, q)$: Camera signal with the q-th filter. $\xi(x_i, y_i)$: dark noise pattern thermally excited or from the radiation of the camera box. $D(x_i, y_i)$: spatial radiation distribution of the filter wheel. $K_q(x_i, y_i)$: back reflection distribution of camera's self emission. The reference object is a standard extended blackbody (EOI. Inc. DCN1000N7) at temperature T_0. B_ν is the blackbody radiation. The mirror image of the camera is formed by filter reflections.

A. Dark noise calibration

To remove the dark noise contributions of $D(x_i, y_i), K_q(x_i, y_i)$ and $\xi(x_i, y_i)$, we use a uniform reference object. On one hand, the total collected signal for a scene $S_{\alpha\nu}$ is, on

average, given by

$$\mathcal{C}(x_i, y_i, q) = \int R_\nu \eta(x_i, y_i) \mathcal{T}_{q\nu} S_{\alpha\nu} \, \mathrm{d}\nu + K_q(x_i, y_i) + [1 - \eta(x_i, y_i)] \, D(x_i, y_i) + \xi(x_i, y_i). \quad \text{(S49)}$$

Here, we have followed the heat signal in Sec. SI and let the measurement time (a) and camera collection coefficient (λ_ν / S_ν) be absorbed in R_ν. The filter wheel in front of the camera amounts to be an out-of-focus optical diaphragm affecting the transmittance, as observed in experiments. We use the relative transmittance $\eta(x_i, y_i)$ normalized by the central pixel to describe the out-of-focus diaphragm effect. On the other hand, the total collected signal for the reference object is given by

$$\mathcal{C}^0(x_i, y_i, q) = \int R_\nu \eta(x_i, y_i) \mathcal{T}_{q\nu} B_\nu(T_0) \, \mathrm{d}\nu + K_q(x_i, y_i) + [1 - \eta(x_i, y_i)] \, D(x_i, y_i) + \xi(x_i, y_i). \quad \text{(S50)}$$

Subtracting Eq. (S50) from Eq. (S49) immediately gives

$$\mathcal{C}(x_i, y_i, q) - \mathcal{C}^0(x_i, y_i, q) = \eta(x_i, y_i) \int R_\nu \mathcal{T}_{q\nu} \left[S_{\alpha\nu} - B_\nu(T_0) \right] \mathrm{d}\nu. \quad \text{(S51)}$$

Note that, in principle, dark noise contributions can be characterized individually. However, due to the uncooled micro-bolometer used in FLIR A325sc and the heat exchange of the detector with the scene, $D(x_i, y_i)$, $K_q(x_i, y_i)$ and $\xi(x_i, y_i)$ change from scene to scene. Therefore, using a reference object proves more convenient in our experiments. Once the detector is under temperature control in future researches, the reference object is unnecessary. Also, the reference object could be any calibrated object other than a standard blackbody source. To ensure the dark noise is stable in $\mathcal{C}(x_i, y_i, q)$ and $\mathcal{C}^0(x_i, y_i, q)$, a gold mirror is mounted on the filter wheel to monitor the status of the detector in real time. Data is taken as valid only when the mirror signal is stable.

B. Characterization of filters and the optical diaphragm effect of the filter wheel

The transmittance curves $\mathcal{T}_{q\nu}$ of 10 used filters in our HADAR prototype-1 experiments are characterized by iS50 FTIR spectrometer and shown in Fig. S27. The optical diaphragm effect of the filter wheel, $\eta(x_i, y_i)$, is characterized by imaging the extended blackbody at various temperatures. When $S_{\alpha\nu}$ becomes spatially uniform, $\eta(x_i, y_i)$ can be obtained by normalizing $\mathcal{C}(x_i, y_i, q) - \mathcal{C}^0(x_i, y_i, q)$ with respect to the central pixel. The characterized $\eta(x_i, y_i)$ is shown in Fig. S28.

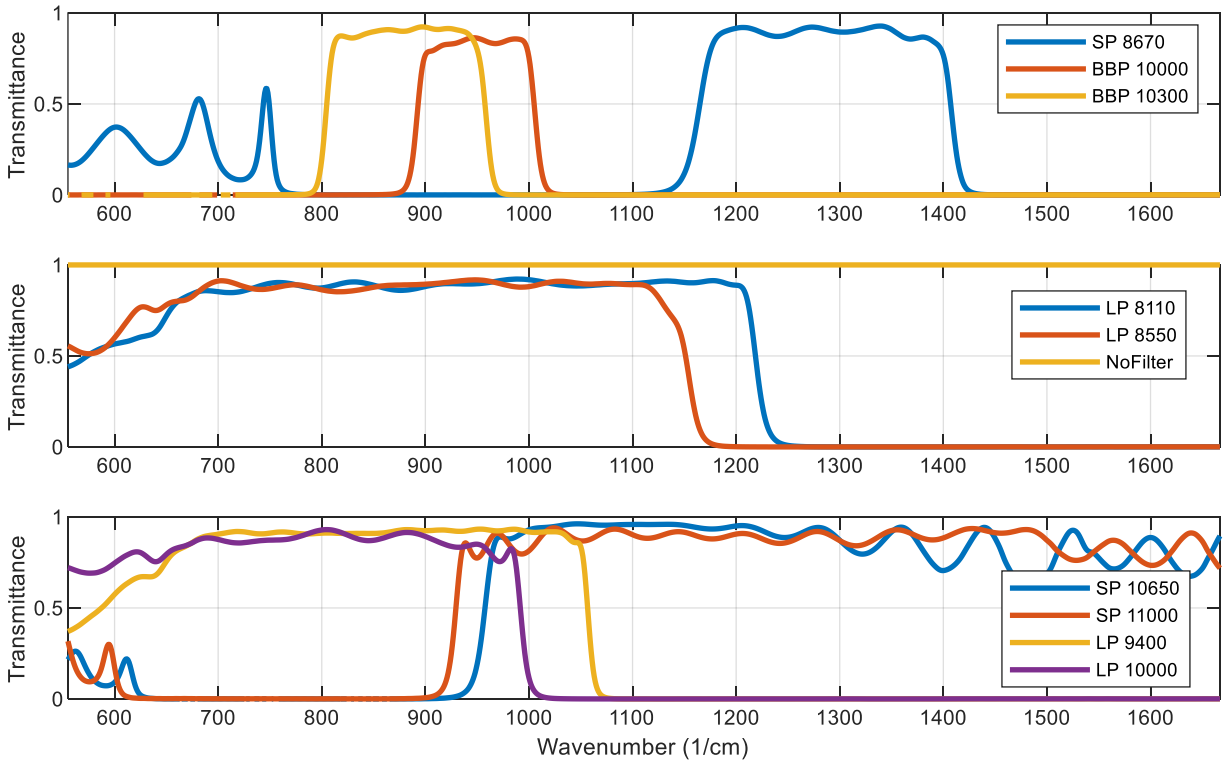


FIG. S27. Filter transmittance characterization. Filters are from Spectrogon.

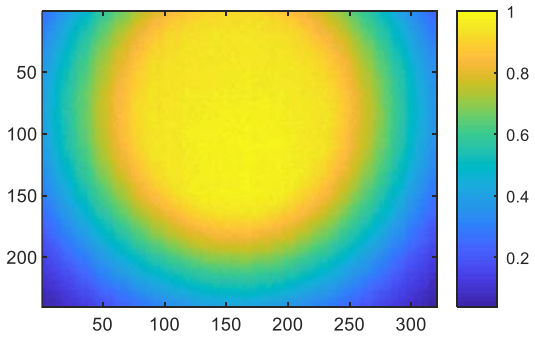


FIG. S28. Characterization of the optical diaphragm effect of the filter wheel. The transmittance is asymmetric because of the imperfect alignment between the filter wheel and the camera.

C. FLIR-A325sc response curve calibration

The spectral response curve R_ν of the FLIR-A325sc is calibrated around the central pixel with the extended blackbody. We scanned the temperature of the blackbody from 35 C° to 110 C° at the step of 2.5 C°, and we recorded corresponding 31 heat cubes $\mathcal{C}(x_i, y_i, q)$. Now, with the above calibrations, R_ν becomes the only variable in Eq. (S51). In total, we

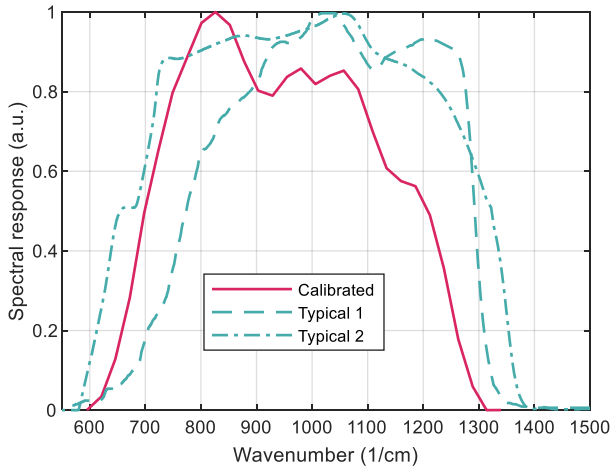


FIG. S29. Calibration of the spectral response curve of the FLIR A325sc camera.

have $(31-1)\times 10 = 310$ equations like Eq. (S51), and we further average them over x_i and y_i. From there, we solve 30 R_ν values in the spectral range $595 \sim 1340\,\mathrm{cm}^{-1}$ by the least-squares linear regression, with the typical response curve provided by FLIR as the initial solution. An additional smooth-curve constraint is used to ensure that the response curve is a smooth curve. The calibrated spectral response curve is normalized and shown in Fig. S29, in comparison with typical response curves. The normalization constant is calibrated to be 25010.

D. Spectrum reconstruction

First, we emphasize that spectrum reconstruction is not essential for TeX vision nor HADAR. It is useful when the explicit spectral resolution of radiance is desired, *e.g.*, to help estimate the material library or environmental radiance in real-world experiments.

When sufficient filters are available, reconstruction of $S_{\alpha\nu}$ from $\mathcal{C}(x_i, y_i, q)$ is similar to the calibration of R_ν from $\mathcal{C}(x_i, y_i, q)$ mentioned above. One can use the least-squares linear regression to solve $S_{\alpha\nu}$ [22], which is more robust than the direct matrix-inversion approach. Since the commercially available filters in the LWIR are under-developed and only 10 significantly-independent filters are found for our proof-of-concept experiments, we do not evaluate $S_{\alpha\nu}$ directly from 10 equations. Instead, we estimate the unknown parameter set $\{mTV\}$ according to the heat signal model in Sec. SI, with the help of a material library and the least-squares regression. Fig. S30 shows the library of spectral emissivity used in

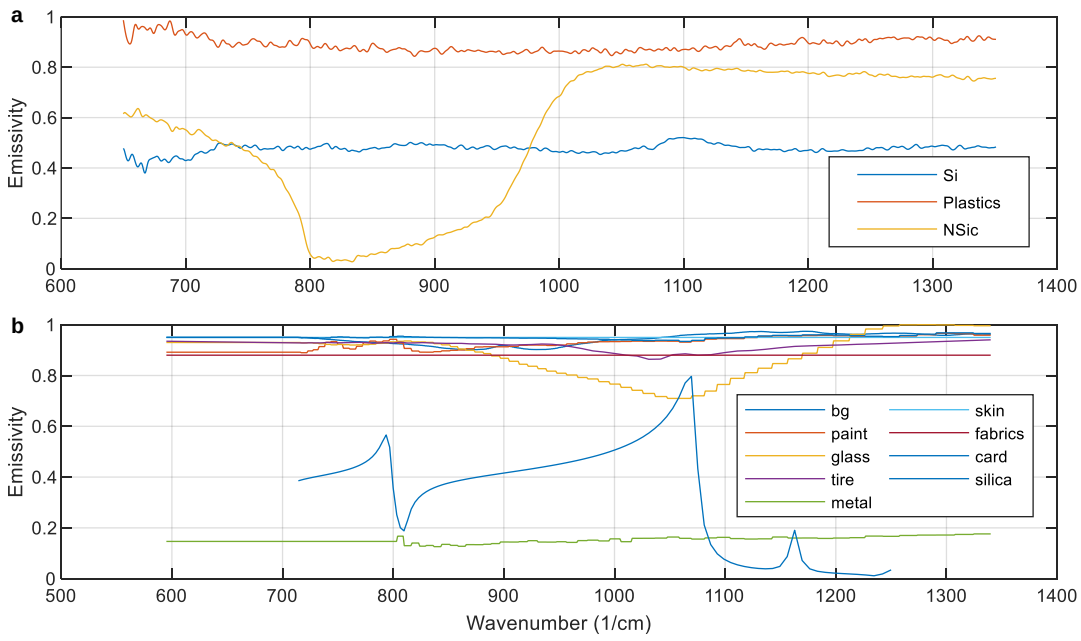


FIG. S30. Material library used in our experiments. a, Emissivities characterized by Nicolet iS50. b, Emissivities drawn from the NASA JPL ECOSTRESS spectral library. bg: modelled by dry grass to represent the whole background; paint: modelled by black paint on aluminum to represent the car; card: modelled by black spray; metal: modelled by weathered aluminum; skin: modelled by a constant 0.95; fabrics: modelled by a constant 0.88. silica in b is generated according to fluctuation-dissipation theorem with tabular refraction index data [23].

our experiments in this paper. The least-squares estimator is described in Sec. SIIID.

E. Stereo calibration

A checkerboard was used as our calibration pattern, see Fig. S31. The checkerboard is 3D printed with alternating square holes, placed in front of an extended blackbody source. The stereo pairs of images of this checkerboard are taken from multiple orientations to get HADAR stereo calibrations.

Considering such a calibration target, stereo images with multiple orientations were taken. These images were passed to *MATLAB Stereo Camera Calibrator* App in order to calibrate the stereo cameras. The results of the calibration are shown in Tabs. S6 and S7. The *MeanReprojectionError* from the calibration was 0.0746 pixels.

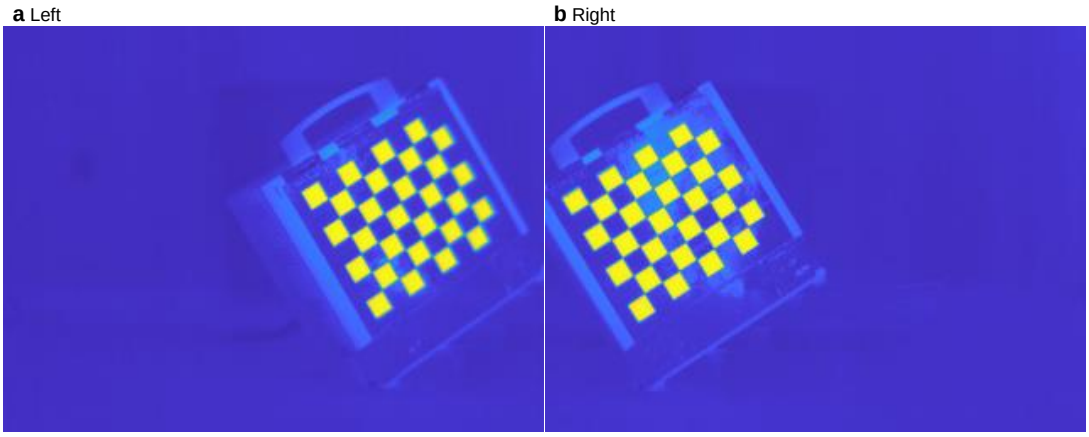


FIG. S31. Customized checkerboard for HADAR stereo calibration.

Camera index	Intrinsic Matrix
Camera 1	$\begin{bmatrix} 737.8141 & 0 & 0 \\ 0 & 737.6437 & 0 \\ 174.3212 & 121.0815 & 1 \end{bmatrix}$
Camera 2	$\begin{bmatrix} 739.4045 & 0 & 0 \\ 0 & 745.8156 & 0 \\ 133.6238 & 163.8408 & 1 \end{bmatrix}$

TABLE S6. Intrinsic Parameters of stereo cameras.

Camera index	R	T
Camera 2	$\begin{bmatrix} 0.9989 & 6.766e-4 & -0.0479 \\ 0.0016 & 0.9989 & 0.0478 \\ 0.0479 & -0.0478 & 0.9977 \end{bmatrix}$	$\begin{bmatrix} -278.8470 \\ -9.3732 \\ 17.1767 \end{bmatrix}$

TABLE S7. Extrinsic Parameters of stereo cameras. R and T denote the Rotation and Translation of Camera 2 with respect to Camera 1.

SV. HADAR PROTOTYPE-2: EXPERIMENTS

The HADAR prototype-2 is based an LWIR hyperspectral imager that is adopted in the DARPA (The Defense Advanced Research Projects Agency) IH (Invisible Headlights) project. The sensor has been calibrated, and the experimental data is available to the authors through the project.

A. Denoise

This HADAR sensor is a pushbroom sensor, which suffers from horizontal streak noise. This is due to dynamically drifting gain and offset of the sensor pixels. We remove the steaks by first detecting streaks and then performing a linear transform to correct gain and offset. The code for such a denoising algorithm is available along with the HADAR database at https://github.com/FanglinBao/HADAR, see TeX.destriper in the TeX matlab class. After denoising, the amount of remaining noise is introduced to RGB images as well, to define a fair comparison of HADAR ranging.

B. Extrinsic calibration between LiDAR and imaging sensors

In IH test experiments, square checkerboards have been adopted for calibrations. However, due to multiple factors, such as, symmetric and sparse checkerboard grids (2×2), the long distance between the checkerboards and the sensors, checkerboards being partially blocked by off-road objects, etc., automatic detection of checkerboard corners doesn't work in practice. Instead, we manually selected checkerboard corners for RGB images and hyperspectral data. For high-resolution LiDAR point cloud, we automatically detected checkerboard planes within manually refined regions, manually removed obvious outlier plane points, extracted corners, and then only kept the corners that are reasonably accurate. The rigid transform was computed with the RANSAC + P3P algorithm in matlab LiDAR toolbox. The calibration results are shown in Fig. S32. Note that the calibration error will lead to inaccurate ground truth LiDAR depth and hence decrease the depth accuracy enhancement of TeX vs. IR in Fig. 6 of the main text.

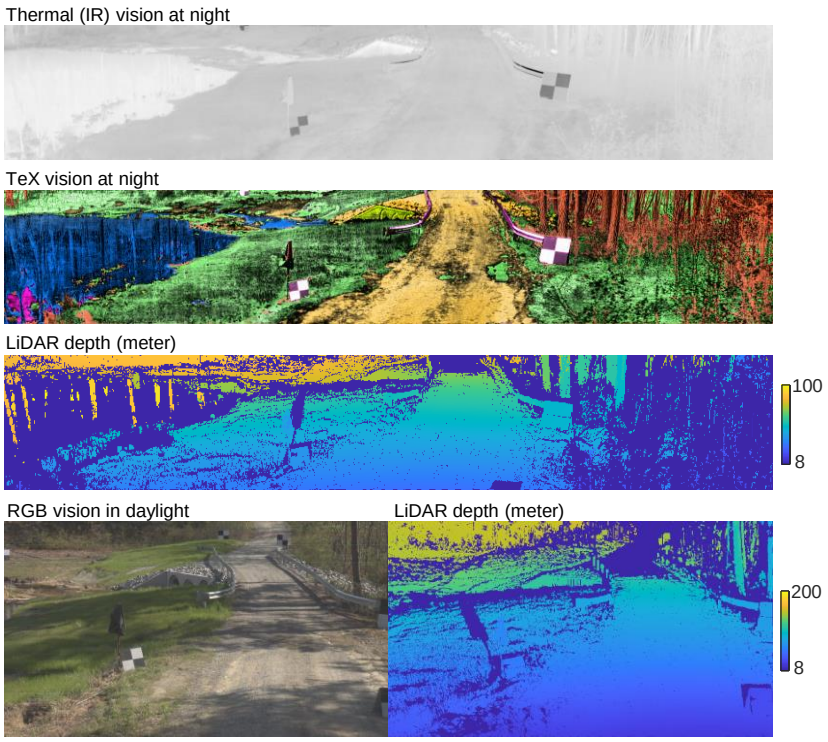


FIG. S32. LiDAR-Sensor calibration to get the ground truth depth in off-road experiments.

C. Semantic library estimation

In real-world off-road scenes, two or more materials may mix together with a smooth-varying mixing ratio. The exact material library may be difficult or even impossible to collect. Instead of using the material library, here we generalize our HADAR theory with a semantic library that can be estimated on-the-fly from the heat cube. Within the semantic library, each curve represents an approximated/averaged emissivity for several similar materials described by the same semantic class. For example, a gravel road by a grass lawn may consist of soil, sand, or little stones that cannot be spatially resolved by sensor pixels. In this case, each road pixel may exhibit a slightly different spectral emissivity curve, but their emissivity curves are still distinct with the grass. Therefore, using averaged emissivity curves can capture the semantics of road vs. grass, while the deviation from the exact emissivity will become a perturbation to the temperature and thermal lighting factors, or remain in the physics-based loss, res. This error will diminish as the number of semantic categories in the library increases.

In IH test experiments, the exact material library was not collected. We used a custom-

designed TES (temperature emissivity separation) algorithm to estimate emissivity per pixel. We adopted the NEM and RAT modules from the original TES algorithm that can be found in Ref. [24]. These modules output the relative profile of the spectral emissivity, leaving one parameter – the absolute magnitude – unfixed. After that, the original TES algorithm uses an empirical formula to determine temperature and the magnitude of spectral emissivity. The empirical formula is based on big data from space/air-based applications and hence not applicable to our current HADAR experiments. Instead, we then used the K-means clustering to categorize materials and derive the averaged emissivity profile for each cluster. We note that multiple pixels of the same cluster share the same spectral emissivity magnitude. Therefore, we estimated the emissivity magnitude and temperature by least-squares fitting according to the HADAR constitutional equation. The averaged spectral emissivity for each cluster form the semantic library of the scene. The resulting semantic library is available along with the HADAR database. In this work, we manually chose the K parameter for K-means clustering. This impacts the categorization process and will eventually lead to some errors in the material and semantic maps. To minimize this error, a potential approach deserving future investigations is to scan different K parameters, estimate semantic library and TeX vision for each K, and then choose the solution with lowest physics loss.

The sky radiance was not collected in IH experiments as well. We read the heat signal off the reflecting checkerboard (which was facing the sky) to approximate the sky radiance. Since the pushbroom sensor was used along with multiple other sensors (irrelevant to this work) in the IH project, the data collection took so long that we observed significant changes of the estimated sky radiance throughout the experiment. The inaccurate sky radiance estimation causes performance fluctuations of TeX vision. We emphasize that this practical restriction can be relieved with a proper on-site experimental characterization of the sky radiance.

D. Texture comparison and analysis between TeX vision and RGB vision in experiments

As explained in Sec. SID, textures in RGB vision and TeX vision images will be different, due to different working wavelength and different material responses in these two spectral

ranges. Furthermore, the following factors will also lead to different textures in two different imaging modalities. (1) Related to the working wavelength, the pixel size of a thermal infrared sensor is of the order of 20μm, while the pixel size of an RGB camera is below 2μm. The $\sim 10\times$ difference in working wavelength and pixel size leads to a poorer spatial resolution and less fine textures in TeX vision. (2) The electronic noise (NEP) of state-of-the-art thermal sensors is much higher than that of state-of-the-art RGB cameras. This means RGB images usually have a higher signal-to-noise ratio and more subtle textures. Especially in hyperspectral imaging, there is systematic noise like horizontal streaks in images for 'pushbroom' sensors, which will pollute real textures. (3) The state-of-the-art hyperspectral imagers are much slower than regular RGB cameras. The former takes several seconds to form one image, while the latter takes only milliseconds. Motion blur in real-world scenes becomes severer in current TeX vision than RGB vision. (4) The state-of-the-art hyperspectral imagers are usually focal-plane arrays, which means the sensor is focusing at infinity. While RGB cameras can focus on the surrounding scenes, focus blur becomes severer in current TeX vision than RGB vision. All these factors have been observed in the TeX vision obtained in real-world experiments, as shown in Fig. S33.

To define a fair comparison between HADAR ranging and RGB stereovision in Fig. 6 of the main text, we have introduced the same amount of noise of the HADAR data into RGB images, and we have down-sampled the RGB images to match the spatial resolution of HADAR sensor.

It can be seen in Fig. S33c-d that TeX vision can even have a larger texture density than original RGB vision. When the same amount of noise of the HADAR data is introduced into RGB images, and when the RGB images are down-sampled to match HADAR spatial resolution, the mean texture density of the RGB images changes to 0.0975, 0.0624 and 0.1553, respectively. The observation of 'TeX vision has more textures than RGB vision' still holds. Possible reasons for more textures in TeX vision include that (1) remaining HADAR sensor noise in the heat cube gets amplified in generating the TeX vision. (2) Poor ambient illumination (shadow) exists in RGB images. (3) HADAR sensor and RGB cameras have different field of view. And (4) More textures may come from more spectral bands in HADAR than RGB cameras. The last case suggests that it may be possible for HADAR ranging at night to even beat RGB stereovision in daylight. Deeper analysis and verification deserve extensive future studies.

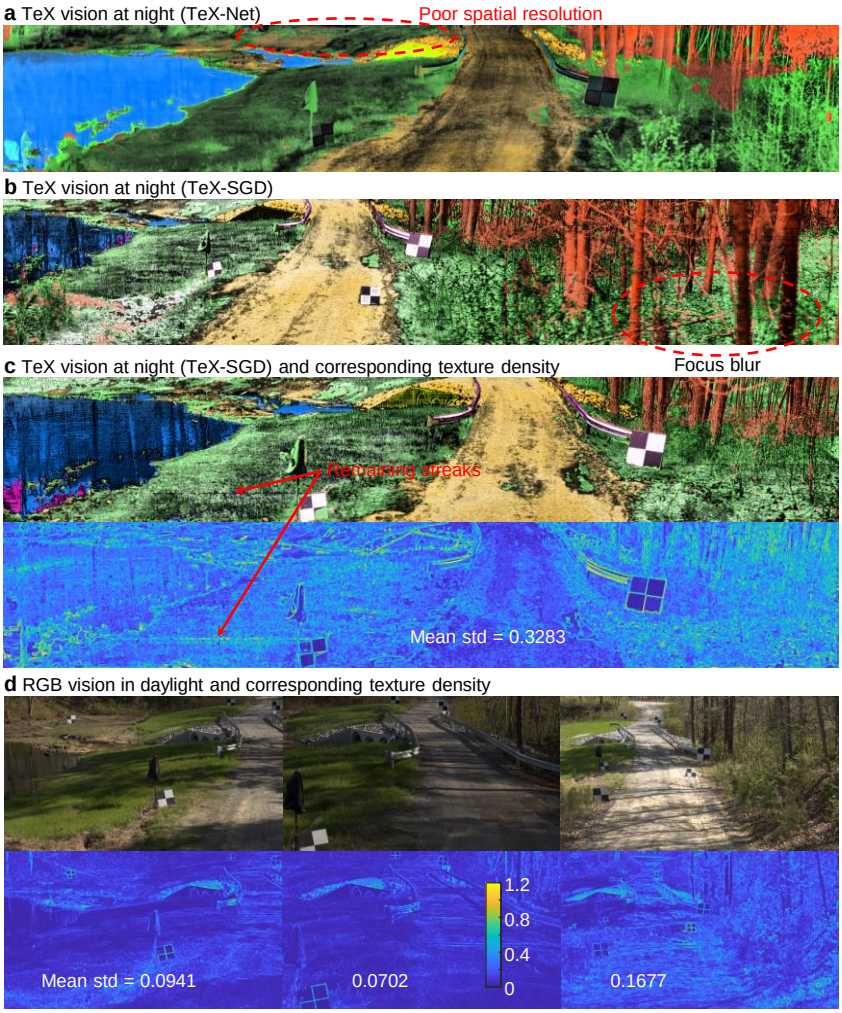


FIG. S33. Texture comparison between TeX vision and RGB vision. a-c show the typical sensor influences on observed textures. c-d show the texture quantification (with the standard deviation metric) and comparison between TeX vision and RGB vision.

E. TeX-RGB image fusion in comparison with IR-RGB image fusion

The fusion of thermal images and optical images has been a common practice in real scenes. In the literature, thermal images have been fused with optical images with poor ambient illuminations for night-vision enhancement, and the goal is to integrate complementary information from different sensors, *i.e.*, the detailed textures in optical images and target highlighting in thermal images. The multi-sensor fusion approach may have comprehensive information as HADAR does but cannot be fully passive. Explicitly, the visible-infrared image fusion is generically pseudo-passive as visible images rely on ambient illumination. For

completely dark scenes, such as our real-world off-road scene at night, there is absolutely no information in visible images and hence the visible-infrared image fusion is not better than the thermal vision. In our work, we focused more to demonstrate the advantage of HADAR vs. thermal sensing which are both based on heat signal, since our work aims to demonstrate that we can get rich information out of heat signal which was previously thought to be impossible. When full passivity or scalability is not required and more sensors can be considered in a multi-sensor fusion approach, HADAR can replace the traditional infrared sensor and work together with other sensors like the visible RGB camera. Explicitly, to compile RGB with TeX, one can following the procedures below. (1) Convert RGB images to grayscale. That is, keep the textures from material response in the visible-light range (which is complementary to textures in the infrared range), and discard the color. The color from TeX vision will be adopted since that is more meaningful. (2) Fuse X channel with grayscale optical images. (3) Use the fused image to replace the original X and, together with T and e channels, form the new 'enhanced' TeX vision images. The TeX-RGB fusion is shown in Fig. S34, in comparison with IR-RGB image fusion. Other images like degree-of-linear-polarization (DoLP) can also be fused with X to form an 'enhanced' TeX vision.

F. TeX vision comparison between two HADAR prototypes

Here, Fig. S35 shows the visual comparison of TeX vision obtained by our two HADAR prototypes for night scenes. This provides the intuitive understanding of TeX vision with different sensor performance and cost settings.

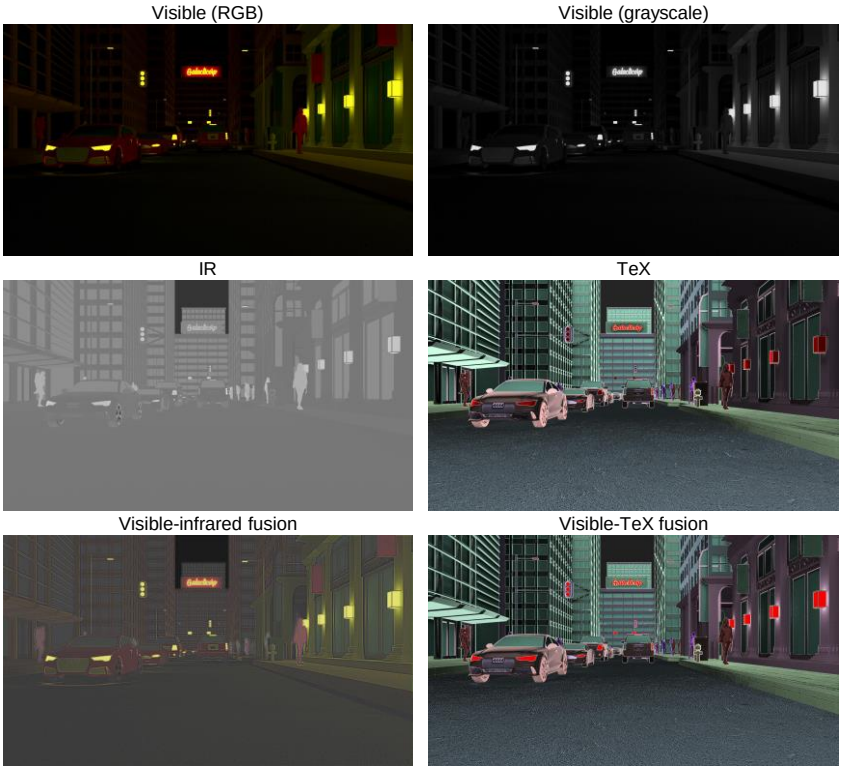


FIG. S34. Image fusion of TeX + RGB, in comparison with IR + RGB.

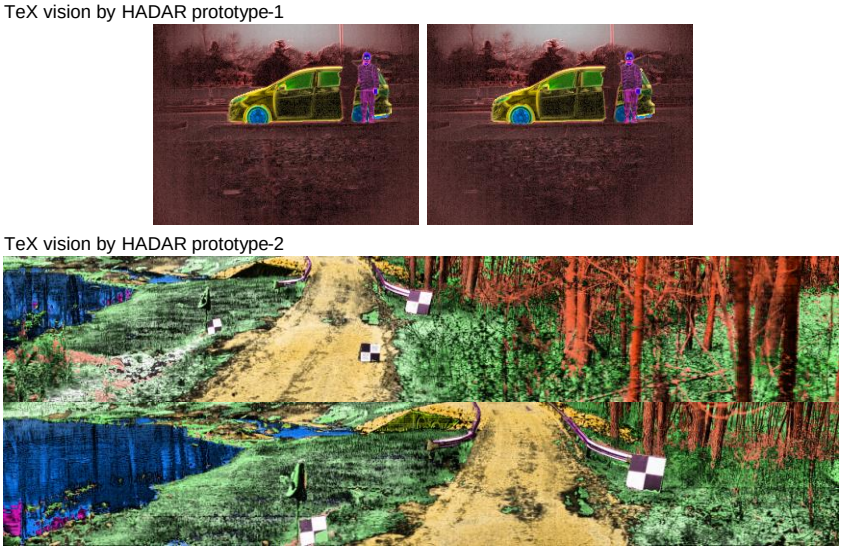


FIG. S35. TeX vision comparison between two HADAR prototypes.

SVI. FTIR SPECTROMETER CALIBRATION

According to the heat signal in Eq. (S9), the electronic signal is on average given by

$$\mathcal{C}_q = \int \left[\mathcal{T}_{q\nu}(R_\nu S_\nu - R_\nu K_\nu) + R_\nu K_\nu\right] \mathrm{d}\nu + \xi. \tag{S52}$$

Here, R_ν has absorbed the measurement time (a), optical efficiency η_ν^{o}, and camera collection coefficient (λ_ν/S_ν). FTIR uses interferometer, and $\mathcal{T}_{q\nu}$ amounts to be a Fourier transform. By inverse Fourier transform, FTIR outputs $(R_\nu S_\nu - R_\nu K_\nu)$ as the measured 'spectrum' which consists of a negative component from the back reflection of the detector's self radiation K_ν. Conventionally, one usually rewrites $-R_\nu K_\nu$ as the effective dark noise ξ_ν. Therefore, the FTIR output has the following form

$$\mathcal{C}_\nu = R_\nu \left\{e_\nu(m) B_\nu(T) + [1 - e_\nu(m)] X_\nu\right\} + \xi_\nu. \tag{S53}$$

A. System response and dark noise

For blackbodies, $e_\nu = 1$, Eq. (S53) reduces to

$$\mathcal{C}_\nu(T) = R_\nu B_\nu(T) + \xi_\nu, \tag{S54}$$

whose two unknown factors R_ν and ξ_ν can be fitted out by data for a series of T:

$$R_\nu = \frac{\mathcal{C}_\nu(T_1) - \mathcal{C}_\nu(T_2)}{B_\nu(T_1) - B_\nu(T_2)}, \tag{S55}$$

$$\xi_\nu = \mathcal{C}_\nu(T_1) - R_\nu B_\nu(T_1). \tag{S56}$$

Here, T_1 and T_2 are not restricted and $T_1 - T_2$ could be very large as long as the blackbody assumption holds true. Linear regression with more temperature data would be helpful.

B. Environment radiation

The previous subsection obtains S_ν from $\mathcal{C}_\nu$, $S_\nu = (\mathcal{C}_\nu - \xi_\nu)/R_\nu$. For non-blackbodies, assuming emissivity e_ν is slow-varying with respect to T, we can deduce e_ν from two closely spaced temperatures T_1 and T_2, $0 < T_1 - T_2 \ll T_2$,

$$e_\nu \approx \frac{S_\nu(T_1) - S_\nu(T_2)}{B_\nu(T_1) - B_\nu(T_2)}. \tag{S57}$$

Here, $T_1 - T_2$ should be small because otherwise $e_\nu(T_1)$ might be significantly different with $e_\nu(T_2)$. It follows that

$$X_\nu = \frac{S_\nu(m, T_2) - e_\nu(m) B_\nu(T_2)}{1 - e_\nu(m)}. \tag{S58}$$

Again, linear regression with more temperature data would be helpful.

---

[1] Y. Xiao, A. Shahsafi, C. Wan, P. J. Roney, G. Joe, Z. Yu, J. Salman, and M. A. Kats, Measuring thermal emission near room temperature using fourier-transform infrared spectroscopy, Phys. Rev. Applied **11**, 014026 (2019).

[2] F. Marsili, V. B. Verma, J. A. Stern, S. Harrington, A. E. Lita, T. Gerrits, I. Vayshenker, B. Baek, M. D. Shaw, R. P. Mirin, and S. W. Nam, Detecting single infrared photons with 93% system efficiency, Nat. Photonics **7**, 210 (2013).

[3] L. Chen, D. Schwarzer, V. B. Verma, M. J. Stevens, F. Marsili, R. P. Mirin, S. W. Nam, and A. M. Wodtke, Mid-infrared laser-induced fluorescence with nanosecond time resolution using a superconducting nanowire single-photon detector: New technology for molecular science, Acc. Chem. Res. **50**, 1400 (2017).

[4] D. F. Walls and G. J. Milburn, *Quantum optics* (Springer Science & Business Media, 2007).

[5] Z. H. Ye, P. Zhang, Y. Li, Y. Y. Chen, S. M. Zhou, C. H. Sun, Y. Huang, C. Lin, X. N. Hu, R. J. Ding, and L. He, Photon trapping photodiode design in hgcdte mid-wavelength infrared focal plane array detectors, Opt. Quantum. Electron. **46**, 1385 (2014).

[6] N. J. Beaudry and R. Renner, An intuitive proof of the data processing inequality, Quantum Information & Computation **12**, 432 (2012).

[7] J. Nowakowski, Fundamental limits in temperature estimation, in *Characterization, Propagation, and Simulation of Infrared Scenes*, Vol. 1311, edited by M. J. Triplett, W. R. Watkins, and F. H. Zegel, International Society for Optics and Photonics (SPIE, 1990) pp. 95 – 108.

[8] A. Baldridge, S. Hook, C. Grove, and G. Rivera, The aster spectral library version 2.0, Remote Sens. Environ. **113**, 711 (2009).

[9] P. Stoica and B. C. Ng, On the cramer-rao bound under parametric constraints, IEEE Signal Process. Lett. **5**, 177 (1998).

[10] C. W. Helstrom, Quantum detection and estimation theory, J. Stat. Phys. **1**, 231 (1969).

[11] C.-I. Chang, Unconstrained mixed pixel classification: Least-squares subspace projection, in *Hyperspectral Imaging: Techniques for Spectral Detection and Classification* (Springer US, Boston, MA, 2003) pp. 141–159.

[12] V. Saragadam and A. C. Sankaranarayanan, Programmable spectrometry: Per-pixel material classification using learned spectral filters, in *2020 IEEE International Conference on Computational Photography (ICCP)* (2020) pp. 1–10.

[13] Y. Xiong and L. Matthies, Error analysis of a real-time stereo system, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1997) pp. 1087–1093.

[14] L. Matthies, Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation, Int. J. Comput. Vis. **8**, 71 (1992).

[15] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, Deeppruner: Learning efficient stereo matching via differentiable patchmatch, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

[16] E. Herbst, X. Ren, and D. Fox, Rgb-d flow: Dense 3-d motion estimation using color and depth, in *2013 IEEE international conference on robotics and automation* (IEEE, 2013) pp. 2276–2282.

[17] J. Sun, W. Cao, Z. Xu, and J. Ponce, Learning a convolutional neural network for non-uniform motion blur removal, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) pp. 769–777.

[18] K. Gupta, B. Bhowmick, and A. Majumdar, Motion blur removal via coupled autoencoder, in *2017 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2017) pp. 480–484.

[19] T. Portz, L. Zhang, and H. Jiang, Optical flow in the presence of spatially-varying motion blur, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012) pp. 1752–1759.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE international conference on computer vision* (2017) pp. 618–626.

[21] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, Eco: Efficient convolution operators for tracking, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017) pp. 6638–6646.

[22] J. Bao and M. G. Bawendi, A colloidal quantum dot spectrometer, Nature **523**, 67 (2015).

[23] E. D. Palik, *Handbook of optical constants of solids*, Vol. 3 (Academic press, 1998).

[24] A. Gillespie, S. Rokugawa, T. Matsunaga, J. S. Cothern, S. Hook, and A. B. Kahle, A temperature and emissivity separation algorithm for advanced spaceborne thermal emission and reflection radiometer (aster) images, IEEE Trans. Geosci. Remote Sens. **36**, 1113 (1998).

[25] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, People detection and tracking from aerial thermal views, in *2014 IEEE international conference on robotics and automation (ICRA)* (IEEE, 2014) pp. 1794–1800.

[26] M. Krišto, M. Ivasic-Kos, and M. Pobar, Thermal object detection in difficult weather conditions using yolo, IEEE Access **8**, 125459 (2020).

[27] Y. Socarrás, S. Ramos, D. Vázquez, A. M. López, and T. Gevers, Adapting pedestrian detection from synthetic to far infrared images, in *ICCV Workshops*, Vol. 3 (2013).

[28] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. M. López, Pedestrian detection at day/night time with visible and fir cameras: A comparison, Sensors **16**, 820 (2016).

[29] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 1037–1045.

[30] A. Khellal, H. Ma, and Q. Fei, Pedestrian classification and detection in far infrared images, in *International Conference on Intelligent Robotics and Applications* (Springer, 2015) pp. 511–522.

[31] Z. Wu, N. Fuller, D. Theriault, and M. Betke, A thermal infrared video benchmark for visual analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2014) pp. 201–208.

[32] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, Thermal–visible registration of human silhouettes: A similarity measure performance evaluation, Infrared Physics & Technology **64**, 79 (2014).

[33] J. W. Davis and M. A. Keck, A two-stage template approach to person detection in thermal imagery, in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, Vol. 1 (IEEE, 2005) pp. 364–369.

[34] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Cehovin, G. Fernandez, T. Vojir, *et al.*, The thermal infrared visual object tracking vot-tir2015 challenge results, in *Proceedings of the ieee international conference on computer vision workshops* (2015) pp. 76–88.

[35] M. Krišto and M. Ivašić-Kos, Thermal imaging dataset for person detection, in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (IEEE, 2019) pp. 1126–1131.

[36] https://www.flir.com/oem/adas/adas-dataset-form/.

[37] E. Chen, O. Haik, and Y. Yitzhaky, Classification of moving objects in atmospherically degraded video, Optical Engineering **51**, 101710 (2012).

[38] R. Miezianko and D. Pokrajac, People detection in low resolution infrared videos, in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2008) pp. 1–6.

[39] B. Besbes, A. Rogozan, A.-M. Rus, A. Bensrhair, and A. Broggi, Pedestrian detection in far-infrared daytime images using a hierarchical codebook of surf, Sensors **15**, 8570 (2015).

[40] I. Riaz, J. Piao, and H. Shin, Human detection by using centrist features for thermal images, in *International Conference Computer Graphics, Visualization, Computer Vision and Image Processing* (Citeseer, 2013).

[41] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu, Cats: A color and thermal stereo benchmark, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) pp. 2961–2969.